A tale of two clades: genome evolution of oomycetes and fungi.

A thesis submitted to Maynooth University for the degree of **Doctor of Philosophy**.

Charley McCarthy, B.Sc.

October 2019

Supervisor
Dr. David Fitzpatrick
Department of Biology
Maynooth University
Co. Kildare
Ireland

Head of Department
Prof. Paul Moynagh
Department of Biology
Maynooth University
Co. Kildare
Ireland

For Patricia McCarthy

Table of Contents ii

Index of Figures viii

Index of Tables x

Abbreviations xi

Acknowledgements xiii

Declaration xiv

Publications and Presentations xv

Abstract xviii

Chapter 1 – Introduction 1

Chapter outline 2
1.1 Microbial genome evolution 3
1.1.1 Microbial genome sequencing: a brief history 3
1.1.2 Genome evolution in microbial eukaryotes: analysis and technique 5
1.1.2.1 How eukaryote genomes evolve 5
1.1.2.2 Comparative genomics 6
1.1.2.3 Phylogenetics and phylogenomics 7
1.1.2.4 Pangenomics 9
1.2 The oomycetes 11
1.2.1 The ecology of the oomycetes 11
1.2.1.1 Oomycete marine pathogens of algae, plants and animals 12
1.2.1.2 Oomycete terrestrial pathogens of important food crops 12
1.2.1.3 Oomycete terrestrial pathogens of forestry 13
1.2.2 The taxonomy of the oomycetes 13
1.2.3.1 The advent of the "egg fungi" 14
1.2.3.2 The oomycetes in the eukaryotic tree of life: from Chromista to SAR 16
1.2.3.3 The class-level phylogeny of the oomycetes 17
1.2.3 The oomycetes in the genomics era 18
1.2.3.1 Genome sequencing of oomycetes 18
1.2.3.2 Trends in oomycete genome evolution 19
1.3 The fungi 21
1.3.1 The ecology of the fungi 22
1.3.1.1 Fungi in food and biotechnology 22
1.3.1.2 Fungal pathogens of animals and plants: established and emerging diseases 23

1.3.2 The taxonomy of the fungi 24
1.3.2.1 Early-diverging fungi 25
1.3.2.2 The Dikarya: yeasts, lichens and mushrooms 26
1.3.3 Fungi in the genomics era 27
1.3.3.1 Genome sequencing of fungi: from yeast to the 1000 Fungal Genomes Project 28
1.3.3.2 Trends in fungal genome evolution 29
1.4 Thesis aims and overview 31
1.4.1 Thesis format and structure 31
1.4.2 Oomycete genome evolution: interdomain HGT and phylogenomics 31
1.4.3 Fungal genome evolution: kingdom-level phylogenomics 32
1.4.4 Fungal genome evolution: pangenomics of model and non-model fungi 32
1.4.5 Discussion and future perspective of microbial eukaryote genomics 33
Chapter 2 – Systematic search for evidence of inter-domain horizontal gene transfer from prokaryotes to oomycota lineages 34
Chapter outline 35
2.1 Introduction 36
2.1.1 Horizontal gene transfer in eukaryotes 36
2.1.2 Diversity and ecological roles of the oomycetes 36
2.1.3 Interdomain HGT in oomycetes 38
2.2 Materials and Methods 40
2.2.1 Dataset assembly 40
2.2.2 Identification of putative bacteria-oomycete HGT events 40
2.2.3 Phylogenetic reconstruction of putative bacteria-oomycete HGT events 43
2.2.4 Analysis of bacterial contamination and taxon sampling 43
2.2.5 Characterization and functional annotation of putative bacteria-oomycete HGT families 44
2.3 Results and Discussion 45
2.3.1 Analysis of bacterial HGT into *Phytophthora* and *Pythium* 45
2.3.2 A putative class II fumarase distinct from *Rickettsia* class II fumarase in *Phytopythium vexans* and *Pythium* spp. originates from bacteria 49
2.3.3 A putative proteobacterial NmrA-like oxidoreductase is present in multiple *Pythium* species 53
2.3.4 SnoaL-like proteins from soil-dwelling bacteria are putative members of the secretome of multiple *Pythium* species 56
2.3.5 A putative hydrolase from xenobiotic-degrading rhizosphere Proteobacteria is present in *Phytophthora capsici* 59
2.3.6 Sphingomonadale alcohol dehydrogenase is present in five *Phytophthora* species 62
2.3.7 Impact and extent of bacterial genes in oomycete evolution 63
2.4 Conclusions 65

Chapter 3 – Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes 66

Chapter outline 67
3.1 Introduction 68
3.1.1 Evolutionary history of the oomycetes 68
3.1.2 Taxonomy of *Phytophthora*, *Pythium* and other oomycete taxa 72
3.1.3 Phylogenetic and phylogenomic reconstructions of the oomycetes 74
3.2 Materials and Methods 78
3.2.1 Dataset assembly 78
3.2.2 Identification and reconstruction of gene phylogenies in oomycete and SAR genomes 78
3.2.3 Supertree analyses of single-copy and paralogous gene phylogenies 79
3.2.4 Identification and supermatrix analysis of ubiquituous oomycete gene phylogenies 79
3.2.5 Identification and supermatrix analysis of ubiquitous Peronosporales gene phylogenies 80
3.3 Results and Discussion 81
3.3.1 Identification of gene families 81
3.3.2 Supertree phylogenies fully resolve oomycete class and order phylogenies 82
3.3.3 Supermatrix approach based on ubiquitous Peronosporales gene phylogenies supports single-copy supertree phylogeny 86
3.3.4 Resolution of the Peronosporales order in phylogenomic analysis 88
3.3.5 The use of supertree and phylogenomic methods in oomycete systematics 90
3.4 Conclusions 92

Chapter 4 – Multiple approaches to phylogenomic reconstruction of the fungal kingdom 93

Chapter outline 94
4.1 Introduction 95
4.1.1 The phylogeny of the fungal kingdom 95
4.1.2 *Saccharomyces cerevisiae* and the origin of modern fungal genomics 96
4.1.3 Fungal genomics and phylogenomics beyond the yeast genome 97
4.1.4 The 1000 Fungal Genomes Project 98
4.2 Phylogenomic reconstructions of the fungal kingdom 100
4.2.1 Supermatrix phylogenomic analysis of fungi 107
4.2.1.1 Fungal phylogenomics using the supermatrix approach 108
4.2.1.2 Phylogenomic reconstruction of 84 fungal species from 72 ubiquitous gene families using Maximum Likelihood and Bayesian supermatrix analysis 109
4.2.1.3 Supermatrix analyses of 84 fungal species accurately reconstructs the fungal kingdom 110
4.2.1.3.1 Basal fungi 111
4.2.1.3.2 Basidiomycota 113

4.2.1.3.3 Ascomycota 114
4.2.2 Parsimony supertree phylogenomic analysis of fungi 115
4.2.2.1 Matrix representation with parsimony analysis in fungal phylogenomics 116
4.2.2.2 Phylogenomic reconstruction of 84 fungal species from 8,110 source phylogenies using MRP and AV supertree methods 117
4.2.2.3 MRP phylogenomic analysis of 84 fungal species is highly congruent with supermatrix phylogenomic analyses 118
4.2.2.3.1 Basal fungi 119
4.2.2.3.2 Basidiomycota 120
4.2.2.3.3 Ascomycota 120
4.2.2.4 Average Consensus phylogenomic reconstruction of 84 fungal species is affected by long-branch attraction artefacts 121
4.2.3 Bayesian supertree phylogenomic analysis of fungi 124
4.2.3.1 Heuristic MCMC Bayesian supertree reconstruction of 84 fungal genomes from 8,050 source phylogenies 125
4.2.3.2 Supertree reconstruction with a heuristic MCMC Bayesian method highly congruent with MRP and supermatrix phylogenies 125
4.2.3.2.1 Basal fungi 125
4.2.3.2.2 Basidiomycota 127
4.2.3.2.3 Ascomycota 128
4.2.4 Phylogenomics of fungi based on gene content 129
4.2.4.1 Gene content approaches to phylogenomics in fungi 129
4.2.4.2 Phylogenomic reconstruction of 84 fungal species based on COG presence-absence matrix 130
4.2.4.3 COG presence-absence matrix approach displays erroneous placement of branches within Dikarya 130
4.2.5 Alignment-free phylogenomic analysis of fungi 132
4.2.5.1 Composition vector method phylogenomics of fungi 133
4.2.5.2 Phylogenomic reconstruction of 84 fungal species using the CV approach 134
4.2.5.3 Composition vector phylogenomic reconstruction of 84 fungal species is congruent with alignment-based methods 135
4.2.5.3.1 Basal fungi 135
4.2.5.3.2 Basidiomycota 136
4.2.5.3.3 Ascomycota 137
4.3 A genome-scale phylogeny of 84 fungal species from seven phylogenomic methods 139
4.3.1 Higher-level genome phylogeny of the fungal kingdom 140
4.3.2 Multiple phylogenomic methods show moderate support for the modern designations of Mucoromycota and Zoopagomycota 141
4.3.3 Pezizomycotina as a benchmark for phylogenomic methodologies 142
4.3.4 The use of phylogenomics methods in fungal systematics 145
4.4 Conclusions 148
Chapter 5 – Pan-genome analysis of model fungal species 149
Chapter outline 150
5.1 Introduction 151
5.2 Materials and Methods 155
5.2.1 Dataset assembly 155
5.2.1.1 Saccharomyces cerervisiae 156

5.2.1.2 Candida albicans 157
5.2.1.3 Cryptococcus neoformans var. *grubii* 157
5.2.1.4 Aspergillus fumigatus 158
5.2.2 Pan-genome analysis of fungal species 158
5.2.3 Phylogenomic reconstruction of intraspecific phylogenies 160
5.2.4 Functional annotation and GO enrichment analysis of fungal species pan-genomes 161
5.2.5 Putative ancestral history of fungal core and accessory genomes 161
5.2.6 Extent of horizontal gene transfer into fungal accessory genomes 162
5.2.7 Chromosomal location of core and accessory gene models in species reference genomes 162
5.2.8 Distribution of knockout viability phenotypes in *Saccharomyces cerevisiae* S288C 163
5.2.9 Distribution of dispensable pathway genes in *Saccharomyces cerevisiae* pan-genome 163
5.2.10 Distribution of biosynthetic gene clusters in *Aspergillus fumigatus* pan-genome 163
5.3 Results 165
5.3.1 Analysis of the *Saccharomyces cerevisiae* pan-genome 166
5.3.2 Analysis of the *Candida albicans* pan-genome 169
5.3.3 Analysis of the *Cryptococcus neoformans* var. *grubii* pan-genome 171
5.3.4 Analysis of the *Aspergillus fumigatus* pan-genome 172
5.3.5 Functional analyses of fungal species pan-genomes 174
5.3.5.1 Gene ontology enrichment in fungal core and accessory genomes 174
5.3.5.2 Ancestral origin of fungal core and accessory genomes 175
5.3.5.3 Interdomain and intrakingdom HGT into fungal accessory genomes 176
5.3.5.4 Chromosomal location of core and accessory genomes in fungal reference genomes 176
5.3.5.5 Knockout viability of core and accessory genes in Saccharomyces cerevisiae S288C 177
5.3.5.6 Dispensable pathway gene clusters in the Saccharomyceres cerevisiae pan-genome 178
5.3.5.7 Biosynthetic gene clusters in the Aspergillus fumigatus pan-genome 178
5.4 Discussion 180
5.4.1 Applying genomic context in eukaryotic pan-genome analysis 180
5.4.2 The pan-genomes of four model fungi 180
5.4.3 Broad trends across fungal pan-genomes 184
5.4.3.1 Fungal core and accessory genomes enriched for potential infection and survival processes 184
5.4.3.2 The fungal core genome is more ancient in origin than the fungal accessory genome 185
5.4.3.3 Horizontal gene transfer may only play a limited role in fungal pan-genome evolution 186
5.4.3.4 Eukaryotic processes such as gene duplication may influence fungal pan-genome evolution 187
5.4.3.5 Subterminal regions of fungal genomes may be harbours of accessory genome content 188

5.4.3.6 Fungal core and accessory genomes encompass various biological pathways and phenotypes 189
5.4.4 Other remarks 192
5.5 Conclusions 194
Chapter 6 – Pangloss: a tool for pan-genome analysis of microbial eukaryotes 195
Chapter outline 196
6.1 Introduction 197
6.2 Materials and Methods 200
6.2.1 Implementation 200
6.2.1.1 Gene model prediction and annotation 201
6.2.1.2 BLASTp and PanOCT analysis 202
6.2.1.3 Refinement of pan-genome construction based on reciprocal sequence similarity 203
6.2.1.4 Functional annotation and characterization of pan-genome components 203
6.2.1.5 Selection analysis of pan-genome using yn00 203
6.2.1.6 Visualization of pan-genome data 204
6.2.2 Dataset assembly 205
6.2.2.1 Yarrowia lipolytica 205
6.2.2.2 Aspergillus fumigatus 205
6.2.3 Pangenome analysis 206
6.2.3.1 Yarrowia lipolytica 206
6.2.3.2 Aspergillus fumigatus 206
6.3 Results 208
6.3.1 Analysis of the *Yarrowia lipolytica* pan-genome 208
6.3.2 Characterization of the *Yarrowia lipolytica* pan-genome 211
6.3.3 Reanalysis of the *Aspergillus fumigatus* pan-genome 212
6.4 Discussion 214
6.5 Conclusions 216
Chapter 7 – Future work and perspectives 217
Chapter outline 218
7.1 Oomycete genomics: future perspectives 219
7.1.1 Oomycete evolutionary history: resolving problem taxa 219
7.1.2 The molecular evolution and diversity of oomycete species 220
7.2 Fungal genomics: future perspectives 221
7.2.1 Mapping major events in the fungal tree of life 221
7.2.2 Exploiting large-scale fungal genomics data 221
7.3 The future of microbial eukaryote genomics 223
Bibliography 224
Supplementary material 266

Index of Figures

Chapter 1

Figure 1.1: Cumulative plot of genomes deposited on NCBI Genbank. 5
Figure 1.2: Comparison of supermatrix and supertree phylogenies. 8
Figure 1.3: Illustrative example of a species pan-genome. 10
Figure 1.4: Simplified phylogeny of the oomycete class. 16
Figure 1.5: Simplified phylogeny of the fungal kingdom. 25

Chapter 2

Figure 2.1: HGT of class II fumarase into *Pythium* and *Phytopythium* species. 48
Figure 2.2: HGT of NmrA-like quinone oxidoreductase into *Pythium* species. 52
Figure 2.3: HGT of SnoaL-like polyketide synthase into *Pythium* species. 55
Figure 2.4: HGT of epoxide hydrolase into *Phytophthora capsici*. 58
Figure 2.5: HGT of alcohol dehydrogenase into *Phytophthora* species. 61

Chapter 3

Figure 3.1: Consensus phylogeny and biological information of SAR groups. 70
Figure 3.2: Comparison of 4 published multigene Peronosporales phylogenies. 71
Figure 3.3: MRP supertree phylogeny of 37 oomycete and 6 SAR species. 82
Figure 3.4: Congruence of 3 Peronosporales phylogenies. 85
Figure 3.5: GTP supertree phylogeny of 37 oomycete and 6 SAR species. 86
Figure 3.6: ML supermatrix phylogeny of 22 Peronosporales species. 88

Chapter 4

Figure 4.1: Illustrative comparison of phylogenomic methods. 102
Figure 4.2: Summary of study's methodology. 104
Figure 4.3: Maximum-likelihood supermatrix phylogeny of 84 fungal species. 112
Figure 4.4: Bayesian supermatrix phylogeny of 84 fungal species. 114
Figure 4.5: MRP supertree phylogeny of 84 fungal species. 120
Figure 4.6: Average consensus supertree phylogeny of 84 fungal species. 124
Figure 4.7: MCMC Bayesian supertree phylogeny of 84 fungal species. 128
Figure 4.8: Maximum parsimony phylogeny of 84 fungal species. 133
Figure 4.9: Composition vector phylogeny of 84 fungal species. 138
Figure 4.10: Congruence of 8 fungal phyla in 5 phylogenies. 141
Figure 4.11: Congruence of Pezizomycotina in 7 phylogenies. 145

Chapter 5

Figure 5.1: Illustrative example of a species pan-genome. 153
Figure 5.2: Pan-genomes of four model fungi. 166
Figure 5.3: Phylogeny of *Saccharomyces cerevisiae* pan-genome. 169
Figure 5.4: Phylogeny of *Candida albicans* pan-genome. 171
Figure 5.5: Phylogeny of *Cryptococcus neoformans* pan-genome. 173
Figure 5.6: Phylogeny of *Aspergillus fumigatus* pan-genome. 175

Chapter 6

Figure 6.1: Workflow for Pangloss. 201

Figure 6.2: Pan-genome of *Yarrowia lipolytica.* 210
Figure 6.3: Cluster sizes within *Yarrowia lipolytica* pan-genome. 211
Figure 6.4: Distribution of genes in *Yarrowia lipolytica* accessory genome. 212
Figure 6.5: Karyotype plot of core and accessory *Yarrowia lipolytica* genes. 214

Index of Tables

Chapter 1

Table 1.1: Genome sizes and number of genes of selected oomycetes species. 19

Chapter 2

Table 2.1: Host ranges of fourteen plant pathogenic oomycete species. 37
Table 2.2: Identification of interdomain HGT events in oomycete genomes. 42
Table 2.3: Identification of interdomain HGT events in three oomycete genera. 42
Table 2.4: Summary of five potential bacterial-oomycete HGT events. 47

Chapter 3

Table 3.1: Genomic information for 43 oomycete and SAR species. 76

Chapter 4

Table 4.1: 84 fungal genomes used for phylogenomics analysis. 105

Chapter 5

Table 5.1: Pan-genomes of four model fungal species. 167
Table 5.2: Gene models with ≥1 annotation term per annotation type. 167

Chapter 6

Table 6.1: Dependencies of Pangloss. 202
Table 6.2: Pan-genome of *Yarrowia lipolytica*. 210

Abbreviations

1KFG	1,000 Fungal Genomes Project
18S	18S ribosomal RNA
28S	28S ribosomal RNA
+F	Estimated amino acid character frequencies
+G	Estimated gamma shape parameter
+I	Estimated proportion of invariant sites
aa	Amino acid
AV	Average consensus
BGC	Biosynthetic gene cluster
BLASTn	Nucleotide-nucleotide Basic Local Alignment Search Tool
BLASTp	Protein-protein Basic Local Alignment Search Tool
BP	Bootstrap support
BUSCO	Benchmarking Universal Single-Copy Orthologs
Bya	Billion years ago
C-terminal	Carboxyl-terminal
CGD	Candida Genome Database
CGN	Conserved Gene Neighbourhood
CO1	Cytochrome oxidase 1 protein
COG	Conserved orthologous genes
CV	Composition vector
CYPs	Cytochrome P450 proteins
DDE	1,1-*bis*-(4-chlorophenyl)-2,2-dichloroethene
DDT	1,1,1-trichloro-2,2-*bis*(p-chlorophenyl)ethane
DNA	Deoxyribonucleic acid
DP	Dispensable pathway
e-value	Expect value
EC	Enzyme Commission
EST	Expressed sequence tag
FET	Fischer's exact test
FGI	Fungal Genome Initiative
Gb	Gigabase
GC-content	Guanine/Cytosine content
GCUA	General Codon Usage Analysis
GO	Gene ontology
GTP	Gene tree parsimony
GTR	Generalised time-reversible
HGT	Horizontal gene transfer
HMM	Hidden Markov Model
Hsp	Heat shock protein
iTOL	Interactive Tree of Life
ITS	Internal transcribed spacer
JGI	Joint Genome Institute
JTT	Jones-Taylor-Thorton amino acid replacement matrix
KOG	Eukaryotic conserved orthologous genes
LBA	Long branch attraction
LG	Le-Gascuel amino acid replacement matrix
LSU	Ribosomal large subunit

Mb	Megabase
MCL	Markov Clustering Algorithm
MCMC	Markov Chain Monte Carlo
MFS	Major facilitator superfamily
ME	Minimum evolution
ML	Maximum likelihood
MP	Maximum parsimony
MPI	Message Passing Interface
MRP	Matrix representation with parsimony
MSSA	Most similar supertree analysis
MUSCLE	Multiple Sequence Comparison by Log-Expectation
Mya	Million years ago
N-terminal	Amine-terminal
nt	Nucleotide
NAD(PH)	Nicotinamide adenine dinucleotide (phosphate)
NCBI	National Centre for Biotechnology Information
NHGRI	National Human Genome Research Institute
NMR	Nitrogen metabolite repression protein
p-value	Probability value
PAM	Presence-absence matrix
PAML	Phylogenetic Analysis by Maximum Likelihood
PANTHER	Protein Analysis Through Evolutionary Relationships
PanOCT	Pangenome Ortholog Clustering Tool
PAUP*	Phylogenetic Analysis Using Parsimony (*and other methods)
PhyML	Phylogenetic Maximum Likelihood
PP	Posterior probability
PTP	Permutation-tail probability
PWM	Position weight matrix
RF	Robinson-Foulds distances
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SAR	Stramenopiles-Alveolates-Rhizaria
SGD	Saccharomyces Genome Database
SMRT	Single-molecule real-time sequencing
SnoaL	Nogalonic acid methyl ester cyclase
SNV / P	Single nucleotide variant / polymorphism
SSU	Ribosomal small subunit
ST	Steel & Rodrigo method of likelihood estimation
tBLASTn	Translated protein-nucleotide Basic Local Alignment Search Tool
tRNA	Transfer RNA
WAG	Whelan-Goldman amino acid replacement matrix
WGD	Whole genome duplication

Acknowledgements

First off, I want to thank David for being the best supervisor a Ph.D. student could have, and I am grateful to all of the guidance and support he has given me over the last four years – and for letting me going off and do my own thing from time to time! I would also like to thank my Ph.D. advisory team, Prof. Sean Doyle and Dr. Paul Dowling, for their helpful advice throughout my Ph.D., and for their assistance in observing my demonstrating for academic modules or providing references for grants I cheekily applied for despite having no chance of getting.

Thank you to my two favourite girlos in the world Eoin and Jamie, for all the fun and games that were had as we all slowly lost our minds, together. It was nice to be able to chat about serious scientific or stupid bullshit (or both) in equal measure at any point in the day. Also apologies to Mark for staying on his couch so often over the last four years (across the various locations where that couch may have been). Thanks as well to Rob, Sarah and Steve for their help in the early days when I was just starting out.

Thank you to everyone in the Department, particularly the other postgrads and especially everyone in the Callan Building – Niall, Anatte, Gerard, Felipe, Sarah, Merissa, Rose, Sandra, Nadine, Dean, Ciara, Dejana, Peter, Dearbhlaith and anybody else I'm forgetting! You didn't see me as much because I was burrowed away in the downstairs lab but I really appreciated all the times we spent together as a group, like getting a 19 in bowling that one time and nearly breaking both my arms in an ice-skating rink in Blanchardstown. Also thanks to Trish and rest of the technical staff for helping me get through demonstrating, particularly when I wasn't familiar with the experiments I was teaching.

Thanks to all my friends, even if I haven't kept in contact as much I should have over the last couple years it's still nice to bump into people from time to time! Thanks in particular to the Gang of Stooges (Preet, Ali and Ciarán) for tolerating me whinging and moaning whenever we'd meet up when they were back in the old country or I was over in their new countries. I would also like to thank my family, especially Margaret and Granny, who put up with me in the last few months of my work in Maynooth and everyone else who was also supportive of me even though they didn't really know what I was doing or why I was still in school in my mid-20s.

Finally, a thanks to everyone who I didn't get around to mentioning here – just know that I'm grateful!

Declaration

I have read and understood the Departmental policy on plagiarism.

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education.

Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Signature:

.....

.....

Date:

.....

Publications and Presentations

Publications collected in this thesis:

McCarthy CGP and Fitzpatrick DA (2016). "Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages". *mSphere*, 1(5):e00195-16.

McCarthy CGP and Fitzpatrick DA (2017). "Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes". *mSphere*, 2(2):e00095-17.

McCarthy CGP and Fitzpatrick DA (2017). "Multiple approaches to phylogenomic reconstruction of the fungal kingdom". In *Advances in Genetics*, 100, pp. 211-266.

McCarthy CGP and Fitzpatrick DA (2019). "Pan-genome analysis of model fungal species". *Microbial Genomics*, 5(2).

McCarthy CGP and Fitzpatrick DA (2019). "Pangloss: a tool for pan-genome analysis of microbial eukaryotes". *Genes*, 10(7):521.

Publications not collected in this thesis:

Waldron R, McGowan J, Gordon N, **McCarthy C**, Mitchell EB, Doyle S and Fitzpatrick DA (2017). "Draft genome sequence of *Dermatophagoides pteronyssinus*, the European house dust mite". *Genome Announcements*, 5(32):e00789-17.

O'Brien CE, **McCarthy CGP**, Walshe, AE, Shaw DR, Sumski DA, Krassowski T, Fitzpatrick DA and Butler, G (2018). "Genome analysis of the yeast *Diutina catenulata*, a member of the Debaryomycetaceae/Metschnikowaiceae (CTG-Ser) clade". *PLOS One*, 13(6):e0198957.

Rahman F, Hassan M, Hanano A, Fitzpatrick DA, **McCarthy CGP** and Murphy DJ (2018). "Evolutionary, structural and functional analysis of the caleosin/peroxygenase gene family in the fungi". *BMC Genomics*, 19(1):976.

Waldron R, McGowan J, Gordon N, **McCarthy C**, Mitchell EB, and Fitzpatrick DA (2019). "Proteome and allergenome of the European house dust mite *Dermatophagoides pteronyssinus*". *PLOS One*, 14(5):e0216171.

Joyce A, **McCarthy CGP**, Murphy S & Walsh F (2019). "Antibiotic resistomes of healthy pig fecal metagenomes". *Microbial Genomics*, 5(5).

O'Connor E, McGowan J, **McCarthy CGP**, Amini A, Grogan H & Fitzpatrick DA (2019). "Whole genome sequence of the commercially relevant mushroom strain

Agaricus bisporus var. *bisporus* ARP23". *G3: Genes|Genomes|Genetics*, pii: g3.400563.2019.

Oral presentations:

McCarthy CGP and Fitzpatrick DA. "Horizontal gene transfer from bacteria into the oomycetes". 2016 Maynooth University Department of Biology Research Day. Maynooth University, 13th June 2016 (awarded 2nd prize).

McCarthy CGP and Fitzpatrick DA. "Evidence of interdomain HGT in oomycete lineages". 2016 Virtual Institute of Bioinfomatics & Evolution Conference. Trinity College Dublin, 1st September 2016.

McCarthy CGP and Fitzpatrick DA. "Whole-genome phylogenetic analysis of the oomycetes". 2016 Young Systematists' Forum. Natural History Museum London, 23rd November 2016.

McCarthy CGP and Fitzpatrick DA. "HGT analysis and phylogenomic reconstruction of the oomycetes". 2017 Maynooth University Department of Biology Research Day. Maynooth University, 6th June 2017.

McCarthy CGP and Fitzpatrick DA. "Genome phylogeny of the oomycetes". 2017 Irish Fungal Society Meeting. Limerick Institute of Technology, 16th June 2017.

McCarthy CGP and Fitzpatrick DA. "Examining pan-genomic structure in exemplar fungal species". 2018 Microbiology Society Conference. ICC Birmingham, 13th April 2018 (awarded travel grant).

McCarthy CGP and Fitzpatrick DA. "The pangenome of *Aspergillus fumigatus* and other important fungi". 2018 Maynooth University Department of Biology Research Day. Maynooth University, 5th June 2018.

McCarthy CGP and Fitzpatrick DA. "Analysis of the fungal pangenome". 2018 Irish Fungal Society Meeting. Maynooth University, 19th June 2018.

McCarthy CGP and Fitzpatrick DA. "One from many: pangenomes of model fungal organisms". 2019 Early Career Microbiologists' Forum Summer Conference. Trinity College Dublin, 20th June 2019 (awarded travel grant).

Poster presentations:

McCarthy CGP and Fitzpatrick DA. "Reconstruction of the oomycete phylogeny using 37 complete genomes". 2017 Society for Molecular Biology and Evolution Conference. J.W. Marriot Austin, 4th July 2018.

McCarthy CGP and Fitzpatrick DA. "The pangenome of *Aspergillus fumigatus*". 2018 Early Career Microbiologists' Forum Summer Conference. University of Birmingham, 14th June 2018.

McCarthy CGP and Fitzpatrick DA. "Visual interpretation of BLAST+ data using Wyndham". 2018 Molecular and Computational Biology Symposium. University College Dublin, 25th November 2018.

McCarthy CGP and Fitzpatrick DA. "The syntenic pangenome of *Saccharomyces cerevisiae*". 2019 Genetics Society of America Fungal Genetics Conference. Asilomar Conference Grounds, 15th March 2019 (awarded travel grant).

McCarthy CGP and Fitzpatrick DA. "Investigating the pangenomes of microbial eukaryotes". 2019 Microbiology Society Conference. ICC Belfast, 10th April 2019 (awarded travel grant).

McCarthy CGP and Fitzpatrick DA. "The yeast pangenome". 2019 Maynooth University Department of Biology Research Day. Maynooth University, June 5th 2019.

McCarthy CGP and Fitzpatrick DA. "Analyses of fungal pangenomes". 2019 Society for Molecular Biology and Evolution Conference. Manchester Central, July 2019.

Abstract

Some of the most ecologically-significant pathogens of plants, animals and marine life come from two groups of filamentous eukaryotes; the oomycetes and the fungi. Although similar in morphology and ecological niche, the two groups are only very-distantly related in terms of evolutionary history. The oomycetes are under-researched in evolutionary science, despite their historical and contemporary impact on food and environmental security. In contrast, fungi themselves are probably the most densely studied and sequenced group of organisms in evolutionary science outside of bacteria. This thesis is a collection of five published computational studies of the evolutionary biology of oomycetes and fungi. The first study is a systematic investigation of bacterial horizontal gene transfer into plant pathogenic oomycete species, which identifies 5 potential HGT events from prokaryotes into multiple oomycetes. The second study is a reconstruction of the evolutionary history of the oomycetes using whole-genome data from 37 species, which supports the larger groups within the oomycetes class but suggests that some exemplar oomycete genera are paraphyletic. Taking advantage of the abundance of genomics data available for all major fungal phyla, the third study reconstructs the evolutionary history of 84 fungal species using seven different phylogenomic techniques and critically evaluates each technique for accuracy, speed and other criteria. The fourth study looks at the pangenomes of four model fungal species, and compares the evolution of genomic variation, virulence and environmental adaptation within each species. The final study presents a refined iteration of the methodology used in the previous pangenome study as a self-contained software package and demonstrates the software's capabilities through pangenome analysis and re-analysis of both model and non-model fungal species. Together, these studies cover a breadth of molecular evolution, comparative genomics, phylogenomics and pangenomics research for two similar, but evolutionarily-distinct groups of important microscopic eukaryotes.

Chapter 1 – Introduction

Chapter outline

As this chapter reviews the entirety of my postgraduate work, I first give a broad introduction to microbial genomics and some of the different areas of evolutionary biology encompassed in this thesis. I then introduce the reader to the two groups of eukaryotic organisms I have researched in this doctoral work: the oomycetes and the fungi. For the oomycetes, I review 1) the ecological roles of the oomycetes, 2) their taxonomy and placement in the eukaryotic tree of life and 3) the genomics and genome evolution of the oomycetes. Moving onto the fungi, I review; 1) the role fungi play in human health and lifestyle, 2) the taxonomy and diversity of the fungi and 3) the history of fungal genomics. Finally, I briefly introduce the studies that form the body of this thesis in terms of their rationale, how research was conducted for each study, the findings of each study and any subsequent conclusions that may be drawn from them.

1.1 Microbial genome evolution

Microbes and in particular microbial eukaryotes have played an important role in the development of many technologies and methodological applications that are now commonplace in genomics and bioinformatics research. In this section, I briefly summarize the history and development of genomics and specifically microbial eukaryotic genomics from the mid-1990s to the present day. I then define and summarize some of the standard genomics and bioinformatics analysis and procedures which are typical in microbial eukaryotic genomics studies, many of which are performed in the studies which make up **Chapters 2-6** of this thesis.

1.1.1 Microbial genome sequencing: a brief history

On July 28 1995, the genomics era began with the public release of the *Haemophilus influenzae* genome, the first genome sequenced from a cellular organism (Fleischmann *et al.*, 1995). It was the culmination of over three decades of nucleic acid sequencing research beginning from the first tRNA sequenced from baker's yeast (*Saccharomyces cerevisiae*) in 1965 (Holley *et al.*, 1965), through to the sequencing of the first bacteriophage and organellar genomes in the 1970s and 1980s (Fiers *et al.*, 1976; Anderson *et al.*, 1981; Bibb *et al.*, 1981), and the concurrent development of first-generation Sanger sequencing and polymerase chain reaction DNA amplification techniques (Sanger, Nicklen and Coulson, 1977; Mullis *et al.*, 1986; Heather and Chain, 2016). For eukaryotes the first genome to be sequenced was that of *S. cerevisiae*, the exemplar model eukaryote (Goffeau *et al.*, 1996). The *S. cerevisiae* sequencing project began in 1991 was led by a consortium of over 94 laboratories and sequencing centres from 19 different countries sequencing individual chromosomes using a variety of sequencing approaches and automated "factory" or lab-based "network" strategies (Goffeau and Vassarotti, 1991; Vassarotti *et al.*, 1995; Goffeau *et al.*, 1996; Engel *et al.*, 2014). The publication of the yeast genome was followed by a number of different model multicellular eukaryote genome sequences and most notably the two draft sequences of the human genome in 2001 (The *C. elegans* Sequencing Consortium, 1998; Adams *et al.*, 2000; Kaul *et al.*, 2000; Craig Venter *et al.*, 2001; Lander *et al.*, 2001; McPherson *et al.*, 2001). It would take another few years for unicellular microbial eukaryotes to catch up to their multicellular counterparts, with genome sequences from other fungi emerging from 2002 onwards starting with the publication of the fisson yeast *Schizosaccharomyces*

pombe genome (Wood *et al.*, 2002) and followed in short by genome sequences from "Protistan" groups like algae, alveolates and oomycetes (**Figure 1.1**) (Gardner *et al.*, 2002; Matsuzaki *et al.*, 2004; Tyler *et al.*, 2006).

Genome sequencing at the dawn of the genomics era was something of an arduous process. Sequencing projects were headed by dedicated institutes which were often competing to sequence the same organism and sequencing projects took years to complete (Loman and Pallen, 2015). Most sequencing technologies and strategies of the 1990s relied on manual labour and early "automated" sequencers were considered somewhat unreliable (Hutchison, 2007), and even basic annotation of bacterial genomes took days at a time (Casari *et al.*, 1995). Initial assembly of more complex genomes such as the diploid *Candida albicans* genome also proved a challenge as new methodologies of genome assembly and data analysis had to be thought up on-the-fly (Jones *et al.*, 2004; Costanzo *et al.*, 2006). Over the course of the 2000s the improvement of sequencing technologies and improvements in assembly and analysis software and computational infrastructure enabled the first large-scale sequencing projects to commence (van Dijk *et al.*, 2018). The Fungal Genome Initiative (FGI) was launched by the Broad Institute to sequence many model non-yeast fungal organisms, while the 1,000 Genomes Project sequenced and analysed variation within 1000 human genomes (Cuomo and Birren, 2010; Auton *et al.*, 2015). By the time average sequencing costs had plummeted to ~$100,000 in 2009 (a >20-fold decrease from the $2.7 billion spent on the publicly-funded human genome sequencing project), approximately 100 eukaryotic genomes had been sequenced to draft-or-better quality (Liolios *et al.*, 2009; Sboner *et al.*, 2011). In recent years, third-generation long read sequencing technologies like PacBio SMRT and Nanopore have seen increasing application in genomics (van Dijk *et al.*, 2018). This has led to a further increase in major community and collaborative genomics projects between laboratories and agencies from different countries, like the 1000 Fungal Genomes Project or the 3000 Rice Genomes Project, which look to sequence and analyse diverse genomes within and across taxa (Li *et al.*, 2014; Stajich, 2017).

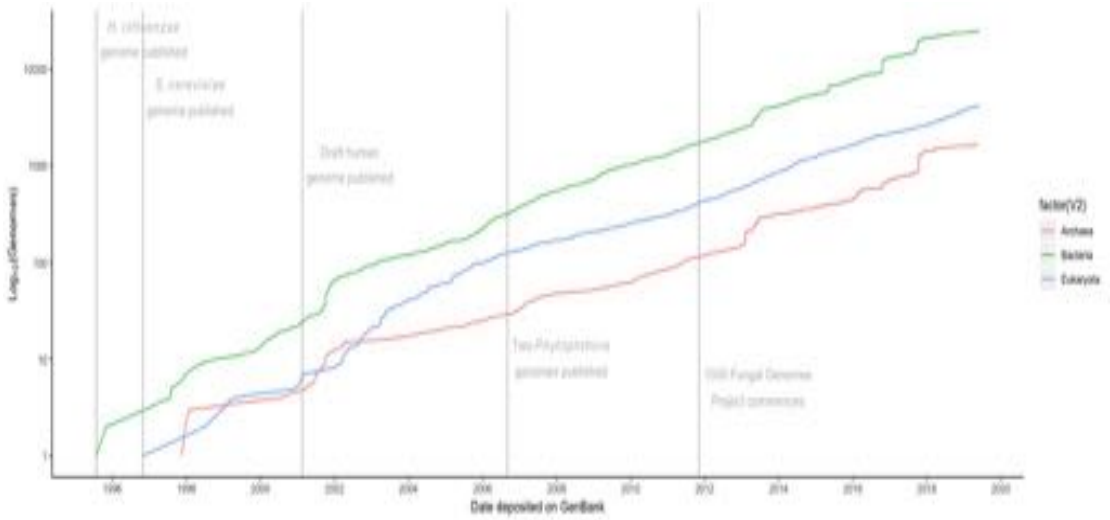


Figure 1.1 Cumulative plot of genomes deposited on NCBI Genbank from 1995 to present, categorized by taxonomic domain. Figure generated using R package rentrez.

1.1.2 Genome evolution in microbial eukaryotes: analysis and technique

Genome evolution as a term encompasses many of the processes by which genomes change and evolve over time, including sexual reproduction, point mutations and horizontal transfer of genetic material (HGT). Genome evolution as a field of study includes a variety of broad evolutionary analysis arising from genome sequence data including comparative genomics, phylogenomics, and the emerging field of pangenomics (Eisen and Fraser, 2003; Tettelin *et al.*, 2005). In this section I briefly discuss the mechanisms by which genomes evolve in eukaryotes and briefly touch upon the two fields of comparative genomics that I predominantly utilize in this thesis; phylogenomics and pangenomics.

1.1.2.1 How eukaryote genomes evolve

Eukaryote nuclear genomes are typically larger and more complex than prokaryote genomes and can vary substantially between and even within the major eukaryotic kingdoms and subgroups. Fungal genomes range from ~10 to ~175Mb in size with an average genome size of ~38Mb, whereas mammalian genomes have an average size of ~3.5Gb with the human genome slightly below that average at ~3.2Gb (Craig Venter *et al.*, 2001; Evans *et al.*, 2017; Stajich, 2017). The smallest known eukaryotic genomes belong to the parasitic microsporidians, with the ~2.3Mb *Encephalitozoon intestinalis* genome smaller than many prokaryote genomes (Corradi *et al.*, 2010). In

contrast, many plants have genomes in excess of ~20Gb in size which can prove a challenge to sequencing (Pellicer, Fay and Leitch, 2010; Li and Harkess, 2018; Pellicer *et al.*, 2018). Unlike prokaryotes, rates of HGT are relatively low between eukaryotes (Keeling and Palmer, 2008). Genomic size, content and complexity in eukaryotes are instead influenced by a number of different processes. Gene duplication is known to play a leading role in eukaryotic gene family evolution, and thus the evolution of eukaryotic genomes themselves (Treangen and Rocha, 2011; Yang, Hulse and Cai, 2012). A number of yeasts, stramenopiles and plants have also undergone at least one whole genome duplication or hybridisation event in their history (Wolfe and Shields, 1997; Kaul *et al.*, 2000; Aury *et al.*, 2006; Martens and Van de Peer, 2010; Marcet-Houben and Gabaldón, 2015). Ploidy variation arising from circumstances such as WGD have been an important factor in plant genome evolution and has led to the large genome sizes observed in many plants (Pellicer, Fay and Leitch, 2010; Brenchley *et al.*, 2012; Lavania, 2015; Guan *et al.*, 2016; Wendel *et al.*, 2016; Li and Harkess, 2018). Expansion of non-coding repetitive genomic regions and evolution of "genomic islands" under extensive purifying selection also been seen in a diverse array of eukaryotic genomes including humans, plants and plant pathogens (Venter *et al.*, 2001; McPherson *et al.*, 2001; Haas *et al.*, 2009; Li and Harkess, 2018; Plissonneau, Hartmann and Croll, 2018).

1.1.2.2 Comparative genomics

Comparative genomics is a field of biological research in which the features of different genomes from or within different species are compared for their similarities or differences so as to make some inference about the biology of those species, such as evolutionary history of genes and species or evolution of phenotype (Alföldi and Lindblad-Toh, 2013). The exact features of genomes that can be compared depends on what the researcher seeks to answer and on the type of analysis that will be performed; comparative genomics studies can be carried out using genome sequence data, individual gene or protein family data or molecular features such as single nucleotide variants (SNVs). In comparative studies of genome evolution in microbial eukaryotes, common types of analyses can include (but are not limited to):

1. Comparing content of genomes in terms of encoded gene or protein sequences to assess which functions/phenotypes are shared by different species, or by different strains of the same species.

2. Comparing the order (synteny) in which genes appear, which can be used to infer evolutionary relationships between/within species.
3. Using phylogenetics (discussed below) to attempt to identify cases of non-vertical inheritances of genes within a species, otherwise known as horizontal gene transfer (HGT).
4. Analysing rates of nucleotide substitution between orthologous genes from different species to identify instances of directional selection.

1.1.2.3 Phylogenetics and phylogenomics

Phylogenetics is the study of the evolutionary history of a group of organisms by way of representing evolutionary relationships within a group using a tree diagram. This method of organizing and visualizing relationships between organisms had its origins in Linnaean classification of life and was popularized by Darwin in *On the Origin of Species* (Darwin, 1859; Teichmann and Mitchison, 1999; Mindell, 2013). Phylogenetic trees, or phylogenies, are constructed by grouping organisms (e.g. strains, species) together based on shared and different characteristics. Traditionally, these characteristics were morphological – for example, a morphological phylogeny of eukaryotes could place bats and birds together based on both possessing wings and being capable of powered flight. However, as convergent evolution often produces similar phenotypes in distantly-related groups of organisms phylogenetics gradually shifted to using molecular characteristics to determine evolutionary history – in other words, grouping organisms by similarities or differences in nucleotide or amino acid content in biological sequences (Bryson and Vogel, 1965; Doolittle, 1999). This approach was used to identify endosymbiotic events in eukaryotes (Schwartz and Dayhoff, 1978), and was the approach that Carl Woese and George Fox adopted when they first proposed that cellular life could be organized into three kingdoms based on SSU rRNA sequence data (Woese and Fox, 1977; Pace, Sapp and Goldenfeld, 2012).

Reliance on single genes or small numbers of genes in early molecular phylogenies meant that inference of relationships between species would often vary depending on what gene(s) a phylogeny was constructed with (D'erchia *et al.*, 1996; Doolittle, 1999; Huynen, 1999). As more completed genome sequences were made available for different lineages of prokaryotes and eukaryotes, researchers began in turn constructing phylogenies based on as much phylogenetically-relevant data from genome

sequences as possible (**Figure 1.2**) (Snel, Bork and Huynen, 1999; Eisen and Fraser, 2003; Snel, Huynen and Dutilh, 2005). Although not without its own caveats this approach, commonly referred to as "phylogenomics", generally leads to phylogenies with more consistent topologies upon subsequent replication (Eisen and Fraser, 2003). Phylogenomics analyses have seen extensive application in reconstruction of eukaryote evolutionary history, with much of our current understanding of the major kingdoms and "super-kingdoms" within the eukaryote domain coming from large-scale phylogenomic analyses (discussed elsewhere in the sections below and **Chapters 3-4**) (Burki *et al.*, 2007; Pisani, Cotton and McInerney, 2007; Holton and Pisani, 2010; Burki, 2014; Spatafora *et al.*, 2016).

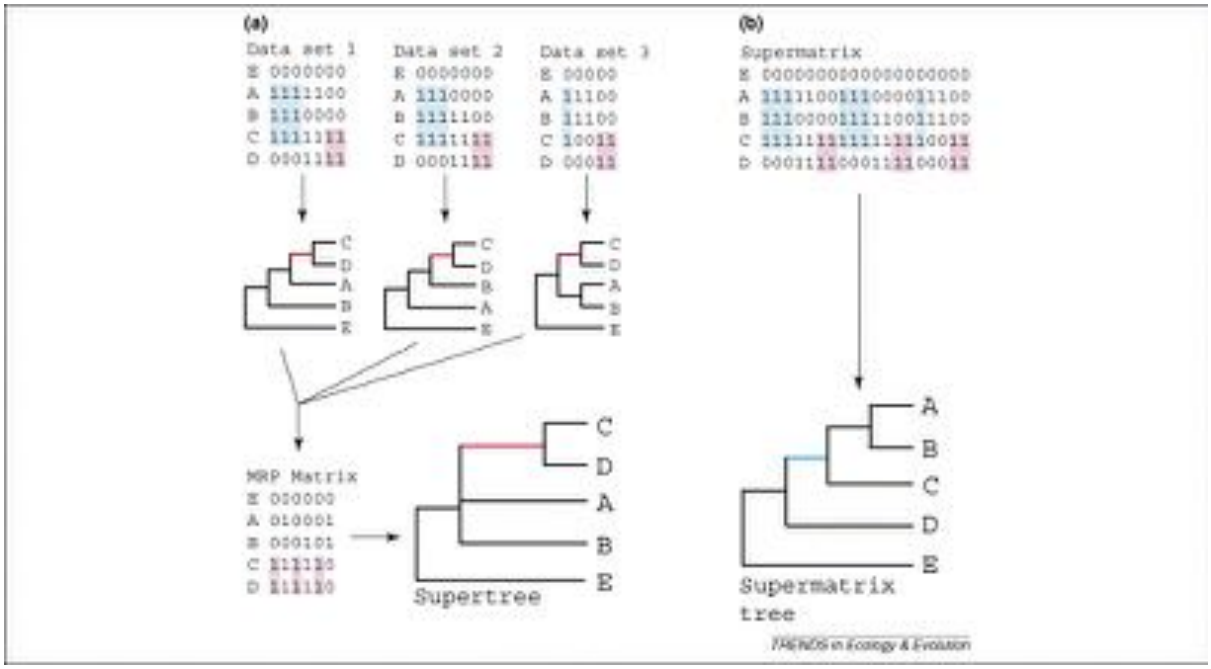


Figure 1.2 Illustration of two established methods of phylogenomic reconstruction; (a) MRP supertree reconstruction and (b) supermatrix reconstruction. Refer to text and Chapter 4 for more detail. Figure taken from de Queiroz and Casey, 2007.

To the present day, a number of different phylogenomic reconstruction methodologies have been described and implemented in various software packages (Guindon and Gascuel, 2003; Lartillot and Philippe, 2004; Qi, Luo and Hao, 2004; Creevey and McInerney, 2009; Lartillot *et al.*, 2013; Akanni *et al.*, 2015). As many of these are discussed in greater detail in **Chapter 4**, I will focus in brief on the two most common approaches: supertree and supermatrix phylogenomics (Baum, 1992; Ragan, 1992; de Queiroz and Gatesy, 2007; McCarthy and Fitzpatrick, 2017a) (**Figure 1.2**). A

supertree phylogeny is a consensus phylogeny that is constructed from individual gene phylogenies using a parsimony method, with different methods for orthologous and paralogous phylogenies (Baum, 1992; Ragan, 1992; Wehe *et al.*, 2008; Creevey and McInerney, 2009). Supermatrix phylogenies are constructed by identifying ubiquitous or near-ubiquitous gene families within a dataset, concatenating all gene families together by taxa into a "superalignment" and performing phylogenomic analysis directly on the concatenated sequence data using statistical methods (de Queiroz and Gatesy, 2007). Supertrees and supermatrices are generally considered robust and accurate approaches of reconstructing evolutionary history but there are some caveats to either approach which are discussed in greater detail in **Chapter 4** (Wilkinson *et al.*, 2004; de Queiroz and Gatesy, 2007; Lartillot, Brinkmann and Philippe, 2007). Other phylogenomics approaches are seeing increasing use, such as statistically-based supertrees or applying coalescent theory to phylogenomic reconstruction (Steel and Rodrigo, 2008; Liu *et al.*, 2009; Akanni *et al.*, 2014, 2015).

1.1.2.4 Pangenomics

Initial genomics studies of prokaryotes and eukaryotes focused on "reference genomes" of species – typically these were the genomes of strains that were well-studied within the research community for a given species, usually due to ease of culturing or breeding. Some prokaryote species including *Escherichia coli* had multiple strains sequenced in the early days of the sequencing era (Alm *et al.*, 1999; Hayashi *et al.*, 2001; Loman and Pallen, 2015). Comparative studies often noted pronounced differences in genomic content between strains of the same species; for example a comparison of a haemorrhagic strain of *E. coli* with the non-pathogenic reference strain found the genome of the former was 1.4Mb larger and encoded >800 more genes than the genome of the latter (Hayashi *et al.*, 2001). Similar genomic variation was also observed in early comparative studies of yeast strain genomes (Wei *et al.*, 2007). In 2005 Hervé Tettelin and researchers sequenced eight strains of *Streptococcus agalactiae*, a urogenital pathogen, and compared the shared and unique gene content in each strain genome (Tettelin *et al.*, 2005). In their analysis, Tettelin *et al.* introduced the concept of a species "pan-genome", which they defined as the set of all genes observed across all strain or isolate genomes within a given species (Tettelin *et al.*, 2005). A species pan-genome is often defined by its components; a species "core" genome and species "accessory"

genome (sometimes referred to as “dispensable” or a “shell” genome) (Medini *et al.*, 2005; Vernikos *et al.*, 2015) (**Figure 1.3**). The core genome contains all genes which are conserved across all strain or isolate genomes, and the accessory genome contains all genes which are variably distributed across strains within a species (Vernikos *et al.*, 2015) (**Figure 1.3**). The functionality and diversity of a species pan-genome can have important implications for evolution within a species. Core genes are typically involved in important housekeeping or survival processes and may be targets for potential therapeutics, whereas accessory genes are typically genes which confer specific phenotypes to individual strains within a species including potential antimicrobial resistance genes, disease-causing genes or genes associated with specific environmental adaptations (Tettelin *et al.*, 2005; Vernikos *et al.*, 2015). Pan-genome analyses have been performed for a variety of different prokaryotic and eukaryotic species, which are discussed in greater detail in **Chapters 5 and 6** (McCarthy and Fitzpatrick, 2019a, 2019b).

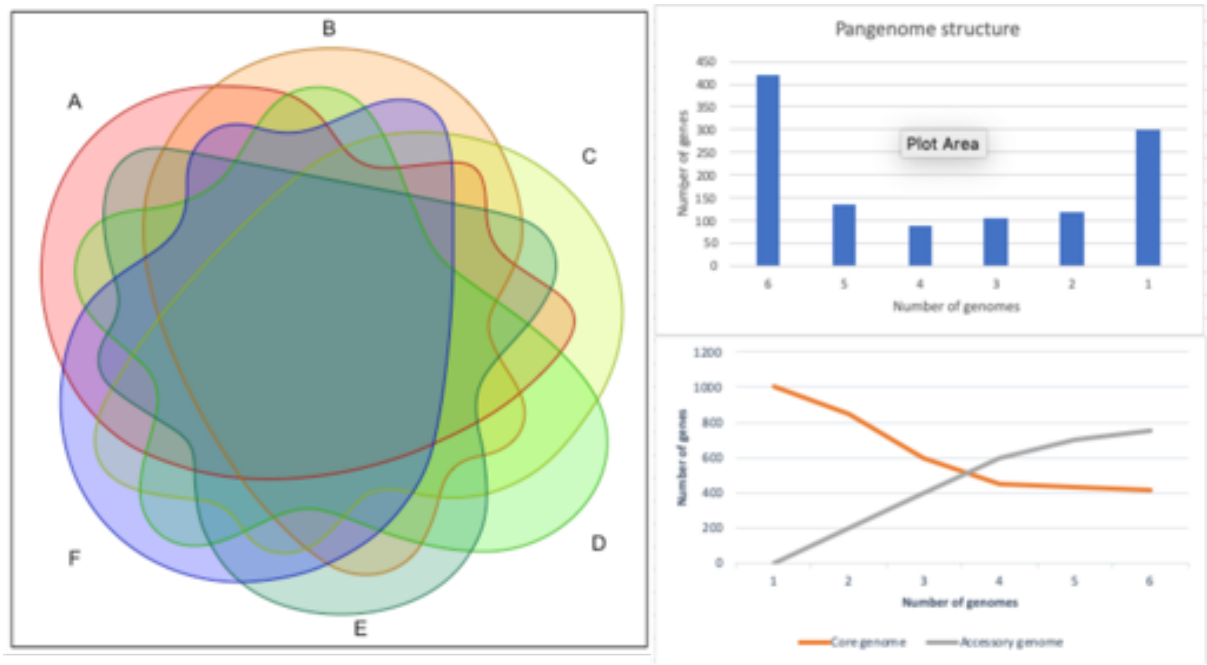


Figure 1.3 Simplified example of a 6-strain pangenome. **Left**: Venn diagram representing overlapping gene content between strains of a species. **Top right**: distribution of genes appearing across *n* strain genomes, ranging from core genes (*n* = 6) to singleton genes (*n* = 1). **Bottom right**: sizes of core (orange) and accessory (grey) genomes as number of input genomes is increased. Figure after Plissonneau, Hartmann & Croll, 2018.

1.2 The oomycetes

The oomycetes (Oomycota) are a class of microscopic filamentous eukaryotes that are ubiquitous in marine and terrestrial environments as pathogens and symbionts. Similar to the other major group of filamentous eukaryotes, the fungi, oomycetes acquire nutrients *via* osmotrophy by secreting an array of enzymes which break down complex macromolecules in the environment (Richards *et al.*, 2011). Like fungi, oomycetes display filamentous growth and a number of oomycetes are capable of both sexual and asexual reproduction. However, despite their macroscopic morphological similarities oomycetes and fungi have many discrete differences in morphology and biochemistry and very distantly related in their evolutionary history (Gunderson *et al.*, 1987; Forster *et al.*, 1990; Baldauf *et al.*, 2000). Those differences are discussed in greater detail in a number of sections below as well as **Chapters 2** and **3** (McCarthy and Fitzpatrick, 2016, 2017b). The oomycetes are a member of the diverse stramenopiles phylum (Stramenopila), with close relatives including brown algae (Phaeophyceae) and diatoms (Bacillariophyceae) (Beakes, Glockling and Sekimoto, 2012). Unlike their algal relatives the oomycetes lack plastids and thus the ability to photosynthesize, and have lost many of the genes derived from endosymbiosis found in photosynthetic stramenopiles (Martens, Vandepoele and Van de Peer, 2008; Beakes, Glockling and Sekimoto, 2012; Leonard *et al.*, 2018). The oomycetes are thought to have diverged from diatoms between 400 to 600 million years ago (mya), and later terrestrialization of oomycetes potentially coincided with early land colonization by plants (Matari and Blair, 2014; Morris *et al.*, 2018). The earliest emergence of oomycetes or oomycete-like organisms in the fossil record can be found in the Rhynie chert, a well-preserved fossil bed dated to the Early Devonian period approximately ~408 mya (Taylor, Krings and Kerp, 2006).

1.2.1 The ecology of the oomycetes

As they are ubiquitous in both marine and terrestrial habitats, it is unsurprising that oomycetes display a large variety of ecological roles and lifestyles. In this subsection, I briefly introduce some of the more important ecological roles that oomycetes play within various habitats and their effects on human activity and food security.

1.2.1.1 Oomycete marine pathogens of algae, plants and animals

Marine oomycetes have a diverse range of potential hosts within marine ecosystems. *Eurychasma dicksonii* is a basal oomycete which infects >40 different species of brown algae, and is widespread throughout temperate and cold seas (Gachon *et al.*, 2009; Li *et al.*, 2010; Grenville-Briggs *et al.*, 2011). As brown algae make up >70% of the total biomass of temperate seashores, and species such as *Saccharina japonica* (sugar kelp) are extensively cultivated for human consumption and alginate production, pathogens like *E. dicksonii* can have a significant impact on ecological diversity and human activity. Many saprolegniales, such as *Aphanomyces invadans* and *Saprolegnia parasitica*, are necrotrophic pathogens of fish and crustaceans – including farmed fishes such as salmon and tilapia (Jiang *et al.*, 2013). *S. parasitica* in particular kills at least 10% of unhatched or juvenile salmon per breeding season, and the only known treatments against infection are banned substances like malachite green (Earle and Hintz, 2014).

1.2.1.2 Oomycete terrestrial pathogens of important food crops

The most infamous member of the oomycetes is probably the hemibiotrophic species *Phytophthora infestans*, the causative agent of late blight in potatoes (*Solanum tuberosum*) and tomatoes (*Solanum lycopersicum*). *Ph. infestans* completes its life cycle in the host in less than a week, meaning that a seemingly-healthy potato crop can completely fail in a very short space of time (Arora, Sharma and Singh, 2014). Although late blight is most commonly associated with historical events like the Great Famine in Ireland in the 1840s (discussed in **Section 1.2.2**), it is still a major threat to food security in both developing nations and the Western Hemisphere (Arora, Sharma and Singh, 2014; McGowan, Byrne and Fitzpatrick, 2019). Treatment and management of *Ph. infestans* outbreaks in potato crop is thought to cost as much as €6 billion per annum worldwide (Haverkort *et al.*, 2008). In the United States, annual losses to the agri-food industry arising from late blight alone are estimated at over $17 billion (Fry and Mizubuti, 1998). Other economically-relevant pathogens from the *Phytophthora* genus include the soybean pathogen *Phytophthora sojae*, which is estimated to cost up to $2 billion in losses per annum and the cocoa tree pathogen *Phytophthora megakarya*, which is capable of causing almost total crop failure in Western and Central Africa (Opoku *et al.*, 2011; Ploetz, 2016; Ali *et al.*, 2017). Other oomycete pathogens of crops include many of the *Albugo* species

which parasitise brassicas and opportunistic root rot pathogens from the *Pythium* genus (Links *et al.*, 2011; Wakelin *et al.*, 2016).

1.2.1.3 Oomycete terrestrial pathogens of forestry

While most associated with agriculture, phytopathogenic oomycetes also have a significant impact on forestry. *Phytophthora ramorum*, known as "sudden oak death", emerged in the 1990s devastating many oak populations along the West Coast of the United States (Goheen *et al.*, 2002; Rizzo *et al.*, 2002). It was subsequently discovered in *Rhododendron* in Europe in the early 2000s, and by 2010 had spread to beech and larch forests in Ireland and the UK largely due to horticultural trading (Werres *et al.*, 2001; Rizzo, Garbelotto and Hansen, 2005; Grünwald, Goss and Press, 2008; Brasier and Webber, 2010). *Ph. ramorum* manifests as "bleeding" or resinous cankers on tree trunks, lesions on leaves and stems and extensive dieback of twigs and branches (Rizzo *et al.*, 2002; Rizzo, Garbelotto and Hansen, 2005). Despite its sobriquet, *Ph. ramorum* is not limited to oak and is thought to infect upwards of 40 forest species (Grünwald, Goss and Press, 2008; Brasier and Webber, 2010; Grünwald *et al.*, 2012). The invasive species *Rhododendron ponticum* is thought to be a vector of *Ph. ramorum* in Ireland (Frankel, Kliejunas and Palmieri, 2008; O'Hanlon, McCracken and Cooke, 2016), and partial removal of *R. ponticum* from Irish forests between 2005-2015 cost the Irish government approximately €3 million (Griffin, 2015). Other forest pathogens such as *Ph. cinnamomi* and *Ph. kernoviae* also pose significant risks to forest populations across many countries (Tomlinson, Dickinson and Boonham, 2010; Hardham and Blackman, 2018).

1.2.2 The taxonomy of the oomycetes

Although for practical purposes (i.e. similar ecological niches) the oomycetes are still studied alongside fungi under the broad field of mycology, their phylogenetic separation from the fungi has been repeatedly confirmed by many different analyses over the last 30 years. However, some aspects of their relationship to other eukaryotic groupings and the taxonomy of oomycete species themselves remain problematic or have only recently been resolved. Here, I summarize the placement of the oomycetes as a class within the eukaryotic tree of life and some of the issues and analyses of phylogenetic classification within the oomycete class.

1.2.3.1 The advent of the "egg fungi"

The first oomycete species to be studied and described in detail was the causal agent of potato late blight, *Phytophthora infestans*. *Ph. infestans* arose in Mexico during the first millennium, probably diverging from a close relative such as *Ph. mirabilis*, but does not appear to have accompanied the introduction of potatoes into Europe during the initial colonization of the New World (Crosby, 1972; Goss *et al.*, 2014). The probable introduction of the HERB-1 strain of *Ph. infestans* into Europe from North America in the 1840s coincided with increased reports of blight on both sides of the Atlantic (Matta, 2010; Yoshida *et al.*, 2013; Saville, Martin and Ristaino, 2016). During this time the potato had become a staple food in many European countries including Ireland, and cultivation among tenant farmers in Ireland was dominated by the "Irish Lumper" variety leading to a severe lack of genetic diversity in the Irish potato population (Kinealy, 1997; Choiseul, Doherty and Roe, 2008; Iomaire and Gallagher, 2009). When a *Ph. infestans* outbreak did eventually occur in Europe its impact was swift and brutal; in mainland Europe countries tens of thousands died as annual potato yields plummeted by up to 88% in 1845 and by Autumn 1846 the potato crop had almost entirely failed in both Ireland and Scotland (O'Grada, 1999; Vanhaute, Paping and O'Grada, 2007). In Ireland the early months of 1847 (so-named "Black '47") were the height of the Great Famine (Vanhaute, Paping and O'Grada, 2007). The complete failure of the potato crop along with an exceptionally cold winter lead to the scenes of widespread destitution most commonly associated with Ireland's Famine years (O'Grada, 1999, 2006). Although the worst of the Famine appeared to be over by 1848, in some areas of Ireland famine conditions were still reported into the early 1850s (O'Grada, 1999, 2006). The Famine had an enormous impact on the demographics of Ireland. It is thought that over a million people died in the island of Ireland between 1845 and 1851 as a result of the famine, although a precise estimate is nearly impossible due to a lack of recorded data outside of public institutions such as workhouses (O'Grada, 2006). The urban population of Ireland had increased by nearly 7% over the years 1841-1851, but in that same period the rural population had decreased by nearly a quarter (O'Grada, 2006).

Ph. infestans was proposed as the causative agent of late blight around 1845-1846 (Matta, 2010). At the time a "fungal hypothesis" for the cause of blight was controversial and different theories abounded as to what caused blight, ranging from a lack of outbreeding in potato crops to more nebulous "atmospheric influences" (Turner, 2005). It was the work of the developmental biologist Heinrich Anton de Bary, who first

described the life cycle of *Phytophthora infestans* in infected potatoes, that clearly established a link between *Ph. infestans* and the degenerative effect of blight on potatoes (Matta, 2010). It was de Bary who also coined the genus name *Phytophthora* ("plant destroyer" in Latin) for *Ph. infestans* in 1876 (Turner, 2005; Matta, 2010). Over time more and more *Phytophthora, Pythium* and *Saprolegnia* species were described and the grouping of these organisms was formally classified into the class Oomycota ("egg fungi") in the 1960s (Tucker, 1931; André Lévesque, 2011; Ribeiro, 2013). Even from the time of Berkeley and de Bary's work, naturalists were uncertain as to the true relationship between these novel oomycete "fungi" and other fungal plant pathogens (André Lévesque, 2011). As molecular and morphological research of the oomycetes grew more sophisticated in the 1960s and 1970s, it soon became clear that a more divergent relationship existed between fungi and oomycetes than had been previously understood. Fungi and oomycetes were to shown to have diverged substantially in important biochemical pathways and cell wall composition (discussed in André Lévesque, 2011), and in the latter case the predominantly cellulose and glucan-rich oomycete cell walls were shown to have similar composition to those of the aforementioned *Vaucheria* algae rather than the chitinous cell walls in fungi (Parker, Preston and Fogg, 1963). Definitive evidence of the divergent relationship between fungi and oomycetes came with the advent of molecular sequencing: two eukaryotic SSU rRNA phylogenies published in 1987 and 1990 placed the oomycetes closer to either planktonic or multicellular algae than to any fungus (Gunderson *et al.*, 1987; Forster *et al.*, 1990) (**Figure 1.4**).

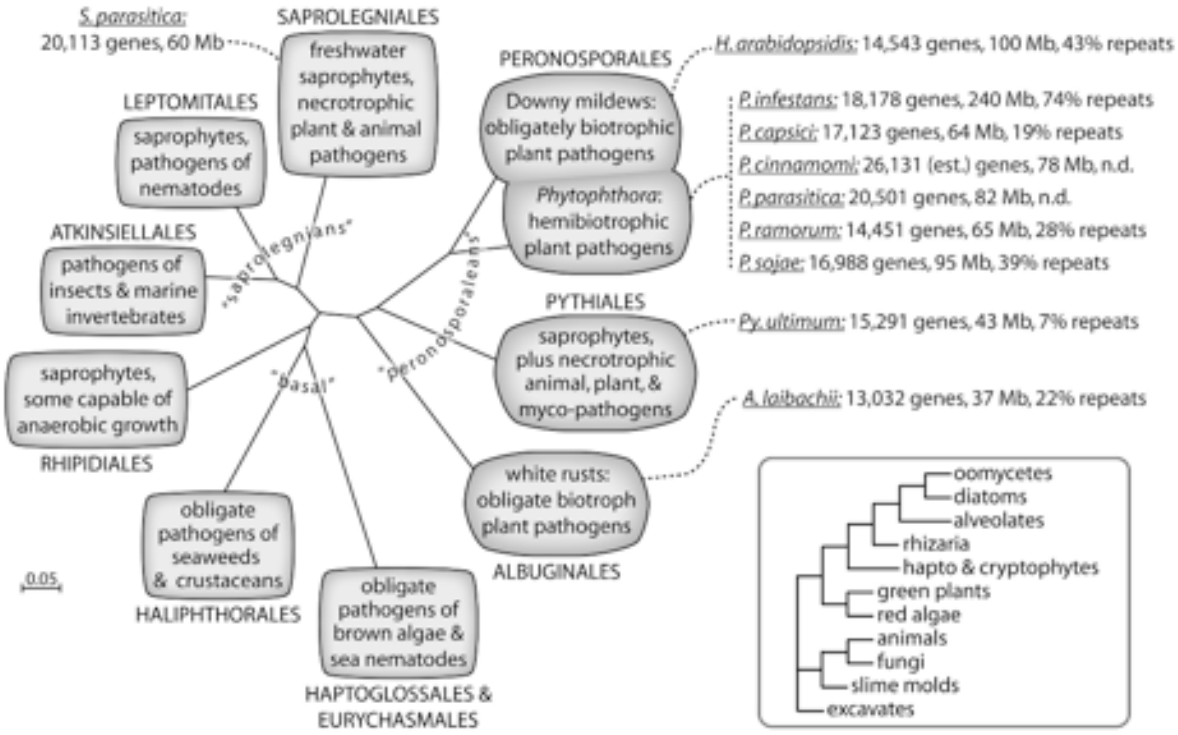


Figure 1.4 Simplified phylogeny of the oomycetes and some information on host ranges and habitats. Figure taken from Judelson (2012).

1.2.3.2 The oomycetes in the eukaryotic tree of life: from Chromista to SAR

During the 1980s, as molecular phylogenetics was beginning to disentangle the relationship between fungi and oomycetes, Thomas Cavalier-Smith proposed the "Chromista" kingdom, which encompassed all algae whose last common ancestor possessed a chloroplast containing both chlorophyll *a* and *c* (Cavalier-Smith, 1981). This kingdom included the oomycetes (who lost their chloroplasts as they evolved a non-photosynthetic lifestyle) within the stramenopiles phylum along with multicellular brown algae such as kelp, unicellular planktonic diatoms and human pathogens like *Blastocystis* (Cavalier-Smith, 1981). This "Chromista" proposal would later be expanded to encompass various other "Protistan" eukaryotes in the broad "chromalveolates" grouping (Cavalier-Smith, 1998; Yoon *et al.*, 2002). Later studies did not support this grouping as monophyletic, but did support a monophyletic grouping for the stramenopiles, phylum Alveolata and phylum Rhizaria into the "SAR" supergroup (Burki *et al.*, 2007; Hackett *et al.*, 2007; Beakes *et al.*, 2014; Burki, 2014).

1.2.3.3 The class-level phylogeny of the oomycetes

Oomycetes as a class diverged from diatoms approximately ~0.5 billion years ago, with their closest ancestors the similarly non-photsynthetic Hyphochytriomycetes (Judelson, 2012; Leonard *et al.*, 2018) (**Figure 1.4**). The basal oomycetes orders are predominantly marine in environment, and are parasites of seaweed, crustaceans and some nematodes (Li *et al.*, 2010; Beakes, Glockling and Sekimoto, 2012) (**Figure 1.4**). Most basal oomycetes lack sexual reproduction or do not perform oogamous sex, with the potential exception of some freshwater *Olpidiopsis* species (Sekimoto *et al.*, 2008; Beakes, Glockling and Sekimoto, 2012). The four "crown orders" of the oomycetes are the Saprolegniales, the Albuginales, the Pythiales and the Peronosporales (Beakes, Glockling and Sekimoto, 2012; Judelson, 2012; Beakes *et al.*, 2014) (**Figure 1.4**). The Saprolegniales are predominantly marine and freshwater saprophytes of animals like the cotton mould *Saprolegnia parasitica* or of plants like some *Aphanomyces* species (Hulvey, Padgett and Bailey, 2007; Beakes *et al.*, 2014) (**Figure 1.4**). The Albuginales are obligate biotrophs of terrestrial plants including white rust pathogen *Albugo candida* (Kemen *et al.*, 2011; Beakes *et al.*, 2014) (**Figure 1.4**). The Pythiales include the diverse *Pythium* genus, and broad host range pathogens from the *Lagenidium* genus (Riethmüller *et al.*, 2002; Beakes *et al.*, 2014) (**Figure 1.4**). The Peronosporales are predominantly hemibiotrophic soil-borne plant pathogens which include many specialized and broad pathogens from the *Phytophthora* and *Phytopythium* genera as well as obligate biotrophs such as *Hyaloperonospora* and *Plasmopara* species (usually grouped into the "downy mildews") (Cooke *et al.*, 2000; Beakes, Glockling and Sekimoto, 2012; Bourret *et al.*, 2018) (**Figure 1.4**).

The "crown orders" of the oomycetes are broadly supported by molecular phylogenies and phylogenomics, but the placement of taxa within the two most-densely studied orders (Pythiales and Peronosporales) has been more problematic, particularly within the exemplar *Pythium* and *Phytophthora* genera (Beakes, Glockling and Sekimoto, 2012). *Pythium* consists of over 120 species divided into 10 clades (Clades A-J) (Lévesque and de Cock, 2004). A former *Pythium* clade, *Pythium* Clade K, was formally reclassified as *Phytopythium* on the basis of phylogenomic analyses and comprises a genus of morphological intermediates between *Pythium* and *Phytophthora* (Lévesque and de Cock, 2004; de Cock *et al.*, 2015). Issues in the *Pythium* genus tree surrounding monophyly have led some researchers to propose that *Pythium* should be split into five new genera based on molecular data (Uzuhashi, Tojo and Kakishima, 2010; Ascunce *et*

al., 2017). The first large-scale molecular phylogenies for *Phytophthora* resolved the *Phytophthora* genus tree into 10 clades (named Clade 1-10) containing >150 species (Cooke *et al.*, 2000). While the clades themselves are generally supported as monophyletic in most molecular phylogenies, resolution of the relationships between clades has remained somewhat unclear (Cooke *et al.*, 2000; Blair *et al.*, 2008; Runge *et al.*, 2011). Additionally, some molecular phylogenies have placed downy mildew species like *H. arabidopsidis* within *Phytophthora*, which would in turn imply that *Phytophthora* itself is paraphyletic while others place downy mildews outside *Phytophthora* but within the Peronosporales order (Riethmüller *et al.*, 2002; Runge *et al.*, 2011; Bourret *et al.*, 2018).

1.2.3 The oomycetes in the genomics era

Although not as widely-targeted for genome sequencing as other eukaryotes (particularly fungi), oomycete genomics is a burgeoning field of eukaryote genomics. In this section, I review oomycete sequencing projects since the first oomycete genomes were released in 2006, summarize some of the broad features of oomycete genomes and introduce some of the comparative genomics studies which have been carried out for the oomycetes in recent years.

1.2.3.1 Genome sequencing of oomycetes

The first oomycetes to have their genome sequenced were two *Phytophthora* species; *Ph. sojae* and *Ph. ramorum* (**Table 1.1**) (Tyler *et al.*, 2006). *Ph. sojae*, a soybean pathogen first described in the 1950s, was selected due to its status as a model oomycete species while *Ph. ramorum* had been recently identified as the agent of the then-emerging "sudden oak death" disease in Californian oak (Govers and Gijzen, 2006; Grünwald *et al.*, 2012). This was followed by genome of *Phytophthora infestans* in 2009, the *Pythium ultimum* genome in 2010, the *Albugo laibachii* genome in 2011 and the *Saprolegnia parasitica* genome in 2013 (Haas *et al.*, 2009; Lévesque *et al.*, 2010; Kemen *et al.*, 2011; Jiang *et al.*, 2013). At the time of writing over 60 oomycete species have genome assembly data available on NCBI, including over 30 *Phytophthora* species and 11 Pythiales species. The vast majority of these sequenced species are plant pathogens, particularly pathogens of important crops and forest species. A small number of species have had multiple strains sequenced but as of writing, only *Phytophthora ramorum* has

been analysed for intraspecific genomic variation (Dale *et al.*, 2019). Although still far lower than that of fungi the number of genome sequencing projects for the oomycetes is expected to increase over the coming years, particularly with the recent formation of the *Phytophthora* Sequencing Consortium by a number of American universities (Dale *et al.*, 2019) and international "moonshot" initiatives such as the Earth BioGenome and the Darwin Tree of Life Projects which seek to sequence eukaryotic life on a national and/or international scale (Lewin *et al.*, 2018).

Table 1.1 Genome size, gene of a number of select oomycete genomes. Adapted from McGowan & Fitzpatrick (2018).

Species	Order	Genome size (Mb)	Genes	Reference
Phytophthora infestans	Peronosporales	228	17,797	Haas et al. (2009)
Plasmopara halstedii	Peronosporales	75	15,469	Sharma et al. (2015)
Hyaloperonospora arabidopsidis	Peronosporales	78	14,321	Baxter et al. (2010)
Phytophthora sojae	Peronosporales	82	26,584	Tyler et al. (2006)
Phytophthora ramorum	Peronosporales	66	15,743	Tyler et al. (2006)
Pythium ultimum	Pythiales	44	15,290	Lévesque et al. (2010)
Albugo candida	Albuginales	34	10,698	Links et al. (2011)
Saprolegnia parasitica	Saprolegniales	53	20,088	Jiang et al. (2013)

1.2.3.2 Trends in oomycete genome evolution

While oomycete genome assemblies are quite fragmented relative to some fungi due to repetitive genomic content, oomycete chromosome numbers are estimated to range from 8 to 14 in some *Phytophthora* and *Pythium* species. Some oomycetes such as *Ph. ramorum* can undergo extensive chromosomal rearrangements and aneuploidy upon host infection. Oomycetes have a larger average genome size (~75 Mb) than fungi (~38 Mb), with genomes ranging between 34 to 240Mb in size (**Table 1.1**) (Judelson, 2012; Tavares *et al.*, 2014). Despite the large variation in genome size among the oomycetes there is no particularly strong correlation between genome size and total gene content. *Ph. infestans* and *Albugo candida* have relatively similar numbers of predicted genes (~13-17,000 each) despite the former having a genome almost 200 Mb larger than the latter (Links *et al.*, 2011; Judelson, 2012). There does not appear to be a correlation between genome size or gene content and lifestyle, but some obligate biotrophs like *A. candida* appear to have undergone a reduction in the size of their genome size and total gene content relative to hemibiotrophic or necrotrophic oomycetes (Links *et al.*, 2011). Genome size differences between oomycetes are largely determined by repetitive DNA content: as much as 74% of the *Phytophthora infestans* genome consists of repetitive DNA compared to 17% of the *Albugo candida* genome (Haas *et al.*, 2009; Links *et al.*, 2011). Oomycete

genomes such as that of *Ph. infestans* are arranged into both gene-dense and gene-sparse regions, with genome expansion of the latter driven by repeat expansion and a proliferation of transposons and transposable elements (Haas *et al.*, 2009). Comparative analyses of oomycete genomes from various "crown" orders have shown extensive expansions of effector families in many *Phytophthora* species relative to other oomycetes (Adhikari *et al.*, 2013; Jiang *et al.*, 2013; McGowan and Fitzpatrick, 2017; McGowan, Byrne and Fitzpatrick, 2019) . Between 28% and 63% of genes in oomycete genomes belong to multi-gene families, with a lower proportion of duplicated genes in obligate biotrophs and an expanded proportion in highly-infective species like *Saprolegnia parasitica* and *Ph. ramroum* (McGowan, Byrne and Fitzpatrick, 2019). A number of oomycete genome papers have included some information as to the extent of putative HGT-derived genes in a given species' genome and some dedicated investigations into the extent of HGT into oomycetes genomes have been carried out (Richards *et al.*, 2006, 2011; Savory, Leonard and Richards, 2015; McCarthy and Fitzpatrick, 2016). HGT from fungi into oomycetes is thought to be one potential source of the convergent evolution of the two groups and two papers from Richards and collaborators in 2006 and 2011 show evidence for substantial transfer of genes from fungi to oomycetes, particularly genes related to carbohydrate metabolism and plant cell wall degradation (Richards *et al.*, 2006, 2011). A similar analysis of HGT from prokaryotes into *Phytophthora* species (found in **Chapter 2** of this thesis) found lower levels of putative HGT events from bacteria, largely from soil-based or rhizosphere-associated species, with transfer genes themselves involved in carbohydrate metabolism and xenobiotics degradation (McCarthy and Fitzpatrick, 2016).

1.3 The fungi

The fungal kingdom is probably the most diverse eukaryotic kingdom, with over 100,000 species described and an estimated 1 million extant species ubiquitous across all environments (Blackwell, 2011; Hibbett and Glotzer, 2011). Fungi are generally distinguished from the other kingdoms of eukaryotes by their chitinous cell walls, filamentous growth and their acquisition of nutrients *via* osmotrophy (although some of these traits have evolved independently in other eukaryotes, e.g. the oomycetes) (Jones, Forn, *et al.*, 2011; Richards *et al.*, 2011). Many soil-borne fungi are primary decomposers of dead and decayed organic materials and litters, such as lignocellulolytic or hemicellulolytic white and brown rot fungi, or secondary decomposers of soils and composts such as the edible mushroom *Agaricus bisporus* (Berg *et al.*, 2003; Morin *et al.*, 2012). Some fungi are cultivated for use as food or intoxicants, while yeasts and many filamentous fungi including *Aspergillus* species are used in the production of many foods, drinks, and condiments. Other fungi are sources for many industrial and pharmaceutical compounds including antimicrobials, organic acids, biofuels and recombinant proteins. Fungal pathogens cause considerable disruption to human health and activity, including opportunistic invasive human pathogens like *Candida albicans* and *Aspergillus fumigatus*, crop pathogens like the wheat blotch fungus *Zymoseptoria tritici* and even ecological damage from environmental pathogens like ash dieback (*Hymenoscyphus fraxineus*) (Odds, Brown and Gow, 2004; Nierman *et al.*, 2005; McMullan *et al.*, 2018; Plissonneau, Hartmann and Croll, 2018). Due to their ubiquity across different aspects of existence and relative ease of culture and analysis, fungi have been intensively-studied in evolutionary biology and modern-day genomics, second only to bacteria. A number of fungi, especially the baking and brewing yeast *Saccharomyces cerevisiae*, are model organisms for eukaryote cell biology and evolution at large. Fungi are closely-related to animals, and the greater Holomycota (fungi, nucleariids and related groups) and Holozoa (animals, choanoflagellates and related groups) groupings are sister branches of the opisthokonts clade (Moreira *et al.*, 2007; Jones, Richards, *et al.*, 2011; Burki, 2014). Fungi are estimated to have diverged from their closest unicellular ancestors approximately 1 billion years ago, and appear to have colonized terrestrial environments along with the oomycetes concomitant to the early colonization of land by plants (Dotzler *et al.*, 2009; Lücking *et al.*, 2009).

1.3.1 The ecology of the fungi

Fungi play a number of diverse roles in human lifestyle and human health. In this section, I focus on the various agricultural and biotechnological applications of fungi and on the role of fungi in disease in humans, animals and plants.

1.3.1.1 Fungi in food and biotechnology

Fungi are an important source of food and an important component in the production of food and drink. A number of fungi are cultivated as food, chief among which is the edible mushroom *Agaricus bisporus* which has been cultivated in Western Europe and the Americas since the 18th century (Morin *et al.*, 2012). In other regions of the world, oyster mushrooms (*Pleurotus ostreatus*) and shiitake mushrooms (*Lentinula edodes*) are widely cultivated for human consumption (Fernández-Fueyo *et al.*, 2014; Chen *et al.*, 2016). The edible mushroom industry is worth an estimated $42 billion to the global economy, with the Irish mushroom industry alone worth approximately $1 billion annually (Chang, 2006; O'Connor *et al.*, 2019). Yeasts, molds and filamentous fungi are commonly-used in food and drink production. *S. cerevisiae* and closely-related yeast species are most notably used as leavening or fermenting agents by converting sugars like glucose or maltose in a substrate (i.e. dough, wort, must) into ethanol and carbon dioxide. Other fermented drinks or condiments made from starchier substrates, such as soy sauce or sake, are brewed using amylase-rich filamentous fungi such as *Aspergillus oryzae* (Nout and Aidoo, 2002). *Penicillium* molds are used to produce blue cheeses such as Roquefort (*Penicillium roqueforti*) and fermented meats like salami (*Penicillium nalgiovense*) (Laich, Fierro and Martín, 2002). *Fusarium venenatum*, a non-pathogenic member of the *Fusarium* genus of filamentous fungi, is an industrial producer of mycoproteins including the meat substitute Quorn (King *et al.*, 2018).

Fungi also play an important role in biotechnology sectors including industrial production of additives and metabolites, pharmaceutical compounds and fuel sources. Industrial-scale production of compounds like citric acid and other organic acids utilizes lignocellulolytic enzymes from *Aspergillus* species, particularly *Aspergillus niger* (Cairns, Nai and Meyer, 2018). Genetically-modified yeasts like *Komagataella phaffi* are used extensively for recombinant protein production of insulins, vaccine compounds and animal feed additives (Cereghino and Cregg, 2000). Many of the current crop of antimicrobial compounds on the market are derivatives of antimicrobial compounds

produced by soil-borne and endophytic fungi, such as the penicillin and cephalosporin families of β-lactam antibiotics (Gao *et al.*, 2017). Fungi have also seen an increasing amount of research as sources of hydrocarbons and other organic compounds with potential use as biofuels. Oleaginous yeasts, such as *Yarrowia lipolytica*, can break down hydrocarbon substrates and accumulate lipids in the form of triacylglycerols in specialized organelles known as lipid bodies up to >40% of its total mass, making them potential hosts for industrial-scale biofuel production (Thevenieau *et al.*, 2009; Adrio, 2017).

1.3.1.2 Fungal pathogens of animals and plants: established and emerging diseases

Many fungi are pathogens of a wide variety of targets including humans, animals, plants and other microbes. Fungi are a much smaller component of the human and animal microbiome than bacteria or archaea, and it remains unclear whether their presence is of much benefit to the human host (Huffnagle and Noverr, 2013; Nash *et al.*, 2017). The most common types of fungal diseases in humans and animals are generally small-scale localized or subcutaneous infections. Dandruff and facial dermatitis are generally caused by basidiomycete *Malassezia* yeasts, and dermatophytosis (ringworm) is caused by various keratinophilic fungi (Rivera, Losada and Nierman, 2012; Saunders, Scheynius and Heitman, 2012; Nash *et al.*, 2017). More serious fungal diseases can occur in humans when the host immune system is weakened due to other diseases or treatment regimes, particularly in hospital settings. Infections by opportunistic *Candida* species can manifest as superficial or localized infections in body cavities such as the mouth or vagina (thrush), or it as systemic candidiasis with significant co-morbidity in AIDS and cancer patients (Palmer, Askew and Williamson, 2008; Butler *et al.*, 2009; Kabir and Ahmad, 2013). Another significant hospital-acquired fungal infection is aspergillosis, caused by the filamentous fungi *A. fumigatus* and *Aspergillus flavus* which both produce toxic secondary metabolites (Nierman *et al.*, 2005; McDonagh *et al.*, 2008; Kousha, Tadi and Soubani, 2011; Kosmidis and Denning, 2015). Chronic aspergillosis is a respiratory and pulmonary disease and can disseminate throughout the blood stream in immunocompromised hosts (Kousha, Tadi and Soubani, 2011; Kosmidis and Denning, 2015). Neglected tropical fungal diseases include subcutaneous mycetoma, caused by the ascomycete *Madurella mycetomatis*, and mucoromycosis caused by a number of "lower" fungi (Ahmed *et al.*, 2004; Riley *et al.*, 2016).

Fungal plant pathogens can have an enormous impact on agriculture and global food security. *Magnaporthe oryzae* is a filamentous fungus responsible for rice blast disease in grasses like rice and wheat, which can destroy up to 30% of total yields (Nalley *et al.*, 2016; Fernandez and Orth, 2018). Wheat stem rust (*Puccinia graminis* f. sp. *tritici*) is a problem for wheat and barley production worldwide. One particular lineage of wheat stem rust known as Ug99 (or TTKSK) is highly virulent against many common plant resistance genes and is spreading across Africa and the Middle East, where it has caused substantial and sometimes total crop loss (Singh *et al.*, 2008, 2011). Outbreaks of *Zymoseptoria tritici*, which is resistant to many common antifungal compounds and fungicides, can result in up to 50% crop losses in wheat (Eyal *et al.*, 1997; Dean *et al.*, 2012; Steinberg, 2015). *Z. tritici* has a highly plastic genome consisting of 21 chromosomes, 8 of which are thought to be entirely dispensable to the fungus, and commonly undergoes repeat-induced mutations which in turn produces a large accessory genome of adaptive genetic material (Möller *et al.*, 2018; Plissonneau, Hartmann and Croll, 2018). Other fungal plant pathogens have a significant impact on forestry and horticulture worldwide. *Armillaria ostoyae* is a pathogen of hardwood and conifer trees in the Pacific Northwest in the US, and is a causative agent of Armillaria root rot alongside other *Armillaria* species like the honey fungus (*Ar. mellea*) (Collins *et al.*, 2013; Sipos *et al.*, 2017; Coetzee, Wingfield and Wingfield, 2018). *Hymenoscyphus fraxineus,* the causative agent of ash dieback, is a fungal pathogen that has spread from Asia to Europe within the last 15 years and poses a significant threat to the largely dieback-susceptible ash populations of the UK and Western Europe (Mitchell *et al.*, 2014; McMullan *et al.*, 2018; Sollars and Buggs, 2018).

1.3.2 The taxonomy of the fungi

Traditionally, the fungi were classified into four groups: the Ascomycetes, Basidiomycetes, Zygomycetes and Chytridiomycetes (James, Kauff, *et al.*, 2006; Hibbett *et al.*, 2007). The first two groups (grouped together into the Dikarya) encompass many macrofungi and yeasts, and the latter two groups traditionally encompassed many of the so-called “lower fungi” (Hibbett *et al.*, 2007). With the advent of molecular phylogenies and phylogenomics a clearer picture of fungal evolution has formed, particularly for the lower fungi (James, Kauff, *et al.*, 2006; Hibbett *et al.*, 2007; Chang *et al.*, 2015; Spatafora *et al.*, 2016) (**Figure 1.5**). The current fungal taxonomy is divided into (generally) 7-8

well-supported phyla with either Cryptomycota or Microsporidia as sister to all other fungi (Jones, Forn, *et al.*, 2011; Capella-Gutiérrez, Marcet-Houben and Gabaldón, 2012; Spatafora *et al.*, 2016). The "lower fungi" are now categorized into Chytridiomycota, Blastocladiomycota, Neocallimastigomycota, Zoopagomycota and Mucoromycota (Hibbett *et al.*, 2007; Spatafora *et al.*, 2016) (**Figure 1.5**). These phyla include many rusts and animal pathogens. The dikarya contain Ascomycota (yeasts, filamentous fungi) and Basidiomycota (mushrooms, smuts) (Hibbett *et al.*, 2007; Chang *et al.*, 2015) (**Figure 1.5**). In this section, I briefly review the taxonomy of the fungal kingdom and list examples of some well-known fungi in each major phylum.

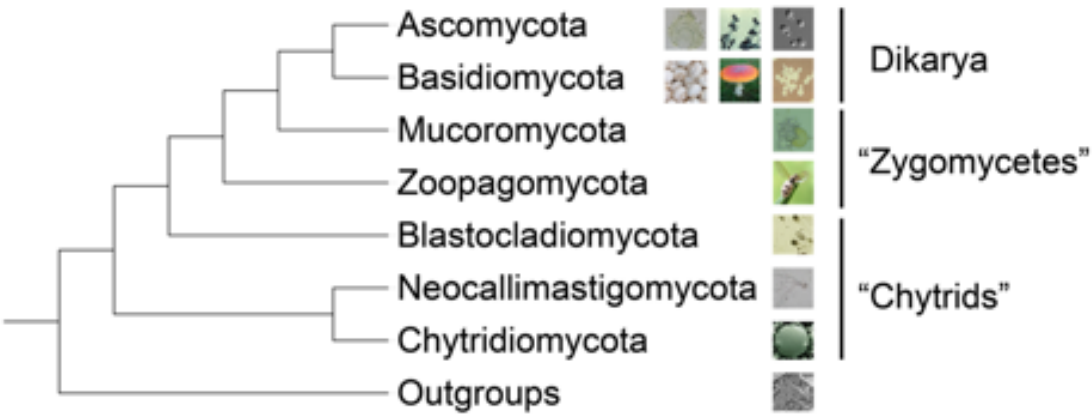


Figure 1.5 Simplified fungal tree of life, with example species included. Topology taken from Spatafora et al., 2016.

1.3.2.1 Early-diverging fungi

It is generally thought that the ancestor of "true" fungi diverged from other holozoans (most likely nucleariid amoebae) approximately 900 mya (Liu *et al.*, 2009; Berbee, James and Strullu-Derrien, 2017). Potential sister groups to the fungi include the Cryptomycota (or rozellids) or Microsporidia, two phyla of endoparasitic eukaryotes (Corradi *et al.*, 2010; Jones, Forn, *et al.*, 2011; Capella-Gutiérrez, Marcet-Houben and Gabaldón, 2012; Spatafora *et al.*, 2016; Berbee, James and Strullu-Derrien, 2017) . Species from both phyla have undergone extensive genome reduction accompanying a parasitic lifestyle – the microsporidian animal parasite *Encephalitozoon cuniculi* has one of the smallest eukaryotic genomes at a mere 2.9Mb – but some fungal innovations such as fungal-specific chitin synthase classes have been retained in Cryptomycota such as *Rozella allomycis* (Katinka *et al.*, 2001; Corradi *et al.*, 2010; Torruella *et al.*, 2015;

Berbee, James and Strullu-Derrien, 2017). This suggests the last common ancestor of all fungi and their closest holozoan relatives possessed both chitinous cell walls and an osmotrophic lifestyle (Berbee, James and Strullu-Derrien, 2017).

The remaining "lower" fungi all occupy varying environments and ecological niches (Hibbett *et al.*, 2007; Spatafora *et al.*, 2016) (**Figure 1.5**). Chytridiomycota are best known through the incredibly destructive *Batrachochytridium* genus of amphibian pathogens, most notably the chytridiomycosis agent *Ba. dendrobatidis*, which have resulted in major decline in amphibian populations worldwide as part of ongoing vertebrate extinction (Tanabe, Watanabe and Sugiyama, 2005; Fisher, Garner and Walker, 2009; Van Rooij *et al.*, 2015; Ceballos, Ehrlich and Dirzo, 2017). Their close relatives the Blastocladiomycota encompass both saprotrophic fungi and obligate parasites of plants and animals, and include model organisms for early fungi such as *Allomyces macrogynus* (James, Letcher, *et al.*, 2006; Porter *et al.*, 2011). The Neocallimastigomycota are strictly anaerobic fungi unique to the gut flora of ruminants and other herbivores, where they play a crucial role in plant degradation (Mackie *et al.*, 2004; Liggenstoffer *et al.*, 2010; Youssef *et al.*, 2013; Wang, Liu and Groenewald, 2017). The Zoopagomycota consists of saprobes, mycoparasites and pathogens of insects and other invertebrates (McLaughlin and Spatafora, 2014; Spatafora *et al.*, 2016). The Mucoromycota are the closest relatives to the Dikarya within the "lower" fungi, and include common bread molds and agents of opportunistic mucoromycosis infections from the *Rhizopus* genus as well as potential sources of lipids such as the oleaginous saprobe *Umbelopsis isabellina* (Riley *et al.*, 2016; Spatafora *et al.*, 2016; Gryganskyi *et al.*, 2018; Kosa *et al.*, 2018) (**Figure 1.5**).

1.3.2.2 The Dikarya: yeasts, lichens and mushrooms

The Dikarya subkingdom encompasses over 95% of all described fungi, and is divided into the phyla Ascomycota and Basidiomycota (Hibbett *et al.*, 2007; Stajich *et al.*, 2009) (**Figure 1.5**). The hallmark trait of the Dikarya is the evolution of dikaryotic cells which contain two unfused haploid nuclei, allowing for greater genetic diversity within species, and other traits including multicellularity have evolved independently multiple times within the subkingdom (Hibbett *et al.*, 2007; Stajich *et al.*, 2009). The Ascomycota encompasses many familiar fungi, including notable yeasts and pathogenic fungi (**Figure 1.5**). Most ascomycetes can be distinguished by the formation of asci –

sexual structures which contain multiples of two or four ascospores (Stajich *et al.*, 2009). The phylum is divided into three subphyla; Taphrinomycotina, Saccharomycotina, and Pezizomycotina. Taphrinomycotina encompass the fission yeasts, including the model fungus *Schizosaccharomyces pombe*, dimorphic plant pathogens such as *Taphrina* species and the Pneumocystidales – a group of yeasts which can cause serious pneumonia in humans (Wood *et al.*, 2002; Cissé *et al.*, 2013; Gigliotti, Limper and Wright, 2014). Saccharomycotina contains the *Saccharomyces* yeasts, along with pathogenic yeasts such as *Candida albicans* and oleaginous yeasts like *Y. lipolytica* (Butler *et al.*, 2009; Liti *et al.*, 2009; Magnan *et al.*, 2016). Pezizomycotina is the largest ascomycete subphylum with over 30,000 described species, including many filamentous fungi such as the aspergilli, other molds such as the model fungus *Neurospora crassa*, and the majority of lichenized fungi (Galagan *et al.*, 2003; Galagan, Calvo, *et al.*, 2005; Lücking *et al.*, 2009; Schoch *et al.*, 2009).

The Basidiomycota contain many familiar macrofungi such as mushrooms, as well as a number of plant and human pathogens (**Figure 1.5**). Most basidiomycetes can be distinguished by their production of "fruiting bodies" that possess sporangia known as basidia, which themselves bear between two to eight basidiospores (Stajich *et al.*, 2009). There are three subphyla in the Basidiomycota phylum; Pucciniomycotina, Ustilaginomycotina and Agaricomycotina. The Pucciniomycotina are the earliest-diverging subphylum and contains many plant pathogenic yeasts and rusts including the stem rust fungus *Puccinia graminis* f. sp. *tritici* (Singh *et al.*, 2008; Wang *et al.*, 2015; Oberwinkler, 2017). The Ustilaginomycotina are mostly dimorphic plant pathogens, with some exceptions such as *Malassezia* species (dandruff and seborrhoeic dermatitis in animals) (Begerow, Stoll and Bauer, 2006; Saunders, Scheynius and Heitman, 2012). Agaricomycotina contains edible and poisonous mushrooms, jelly fungi and a number of non-ascomycete yeast species such as *Cryptococcus neoformans* (Fraser *et al.*, 2005; Stajich *et al.*, 2009; Medina, Jones and Fitzpatrick, 2011; Morin *et al.*, 2012).

1.3.3 Fungi in the genomics era

Fungi are probably the most broadly-sampled branch of the eukaryotic tree of life for genome sequencing and comparative genomics. In this section, I provide a short recap of the history of fungal genome sequencing from the publication of the *Saccharomyces cerevisiae* genome in 1996 to large-scale community-led sequencing projects of the

present, and briefly discussed some observed trends within genomes across the fungal kingdom.

1.3.3.1 Genome sequencing of fungi: from yeast to the 1000 Fungal Genomes Project

As recounted in **Chapter 1.1** above, the first eukaryote genome sequenced was that of *S. cerevisiae* between 1989 and 1996. Sequencing the complete genome of *S. cerevisiae* with the available technology in the early 90s required the work of approximately 600 scientists across 19 countries – the sequencing of chromosome III alone involved collaborators from 35 European laboratories (Goffeau and Vassarotti, 1991; Oliver *et al.*, 1992; Goffeau *et al.*, 1996; Engel *et al.*, 2014). This was followed in due course by the sequencing of the fission yeast *S. pombe* and other model species such as *N. crassa* but around the turn of the millennium there was a lull in fungal genome sequencing relative to other eukaryote taxa (Wood *et al.*, 2002; Galagan *et al.*, 2003; Hofmann, McIntyre and Nielsen, 2003; Galagan, Henn, *et al.*, 2005; Cuomo and Birren, 2010). Around this time two fungal sequencing and comparative genomics initiatives were set up in the US and France, the Fungal Genome Initiative (FGI) and the Génolevures consortium (Stajich *et al.*, 2009; Cuomo and Birren, 2010; Souciet, 2011). The Génolevures project was organized by a number of laboratories within the French Centre national de la recherche scientifique (CNRS) to perform sequencing and large-scale comparative genomics of yeasts including pathogenic species like *Candida glabrata* and biotech-relevant species like *Y. lipolytica* (Souciet *et al.*, 2000; Souciet, 2011). The FGI was initially set up by fungal researchers and organizations in order to obtain greater funding for fungal genome sequencing from public bodies such as the National Human Genome Research Institute (NHGRI) in the US (Pennisi, 2001, 2002; Hofmann, McIntyre and Nielsen, 2003; Cuomo and Birren, 2010). The first white paper published by the initiative in 2002 proposed the sequencing of 15 important fungi divided into three categories; clinically-relevant species such as *Cryptococcus neoformans* (fungal meningitis), commercially-relevant species such as *Magnaporthe grisea* (rice blast) and important model species for evolutionary and population biology such as the gray shag mushroom *Coprinopsis cinerea* (Birren, Fink and Lander, 2002). Gradually the number of genome sequencing projects involved increased such that by the informal end of the FGI around the end of the 2000s, over 80 fungal species had their genomes sequenced (Cuomo and Birren, 2010).

Despite this greater genomic sampling of the fungal kingdom, 66 of the ~80 fungal genome sequences available to researchers around 2010 were from the Ascomycota and all but 5 were from the Dikarya (Cuomo and Birren, 2010). To broaden the amount of data available across the fungal tree of life, the Joint Genome Institute (JGI) in the US initiated the 1000 Fungal Genomes project (1KFG) from 2013 onwards (Grigoriev, Nordberg, *et al.*, 2011; Grigoriev, 2013; Grigoriev *et al.*, 2014). The 1KFG project is a community-led effort to sequence over 1000 genomes with a particular focus on covering the full diversity of fungi by "sequencing at least two reference genomes from the more than 500 recognized families of Fungi" according to the JGI's MycoCosm web portal (https://mycocosm.jgi.doe.gov/mycocosm/home/1000-fungal-genomes) (Grigoriev *et al.*, 2014). To date the project has seen an incredible increase in the amount of genomic data available for the fungal kingdom; there are over 1,400 fungal genome sequences available from MycoCosm as of October 2019, over a thousand more than there were five years ago and of which 529 have been sequenced as part of the 1KFG (Grigoriev *et al.*, 2014). Additionally, over 100 genome sequences from phyla outside the Dikarya are available from MycoCosm. Other large-scale sequencing projects for the fungi have limited their range to deeper parts of the fungal tree of life, such as the Y1000+ project based at the University of Wisconsin-Madison, which is currently sequencing over 1,000 yeast species from across the Saccharomycotina (Shen *et al.*, 2018) and the 1,002 Yeast Genome project which has sequenced over 1,000 individual strains of *S. cerevisiae* sampled from diverse global locations and ecological sources (Hittinger *et al.*, 2015; Peter and Schacherer, 2016; Peter *et al.*, 2018).

1.3.3.2 Trends in fungal genome evolution

As a broad kingdom containing potentially close to a million extant species, it is unsurprising that genome size and architecture varies substantially within the fungi (Blackwell, 2011; Hibbett and Glotzer, 2011). The average fungal genome is approximately 38Mb in size with an average number of protein-coding genes is approximately 11,000 genes, based on all successfully sequenced fungi to date (Stajich, 2017). The largest gene count observed in the fungi to date has been seen in the genome for *Sphaerobolus stellatus*, the cannonball fungus, which is estimated to encode over 35,000 genes (Kohler *et al.*, 2015). Other mycorhizzal fungi such as *Gymnopus luxurians* also possess genomes which encode over 20,000 genes, partially as a result of increased

gene duplication and greater evolution of multi-gene families (Kohler *et al.*, 2015). Genome expansion has been observed in some basidiomycete fungi such as the rust fungi (e.g. flow cytometry estimates the *Gymnosporangium confusum* genome around 800Mb in size) and the Entomophthoromycotina (DNA staining techniques estimate the vertebrate gut fungus *Basidiobolus ranarum* has a likely genome size of ~700Mb) (Henk and Fisher, 2012; Tavares *et al.*, 2014). Within the Entomophthoromycotina, genome expansions appears to have occurred due to the increased presence of transposable elements, similar to genome expansions in oomycetes like *Ph. infestans* (Haas *et al.*, 2009; De Fine Licht, Jensen and Eilenberg, 2017). Some obligate parasites from the Microsporidia have genomes as small as 3Mb (smaller than the 4.6Mb genome of *Escherichia coli* K-12) and the genome of *Encephalitozoon cuniculi* encodes less than 2,000 genes (Blattner *et al.*, 1997; Katinka *et al.*, 2001; Corradi *et al.*, 2010). Both ascomycete and basidiomycete yeasts have small genomes around 9-20Mb encoding between ~5,100 (in the case of *Schizosaccharomyces pombe*) and ~7,000 (in the case of *Cryptococcus neoformans*) protein-coding genes (Wood *et al.*, 2002; Fraser *et al.*, 2005; Stajich, 2017). The reduced gene count and genome size of both ascomycete and basidiomycete yeasts relative to their closest multicellular relatives appears to be the result of independent extensive gene reduction during the evolution of unicellular growth (Dujon *et al.*, 2004; Kohler *et al.*, 2015). The effects of hybridization and introgression on fungal genome evolution have been extensively-studied in different *Saccharomyces* species (De Barros Lopes *et al.*, 2002; Morales and Dujon, 2012; Marsit and Dequin, 2015; Dujon and Louis, 2017). Ancestral whole genome duplication events have also been studied in *S. cerevisiae* and closely-related yeast lineages, and potential WGD events have also been identified in other fungi such as *Rhizopus oryzae* (Wolfe and Shields, 1997; Ma *et al.*, 2009; Marcet-Houben and Gabaldón, 2015). Extensive HGT within fungi has been observed across and within different branches of the fungal tree of life including Neocallimastigomycota, plant pathogenic and human pathogenic ascomycetes, and hallucinogenic mushrooms (Szöllősi *et al.*, 2015; Qiu *et al.*, 2016; Reynolds *et al.*, 2018; Murphy *et al.*, 2019).

1.4 Thesis aims and overview

The bulk of this thesis consists of five separate studies of the genome evolution of two microbial eukaryote groups: the oomycetes and the fungi (**Chapters 2-6**). They cover a breadth of different comparative genomics studies that can be carried out for both undersampled and densely-sampled microbial eukaryotes as I have described in this chapter. Below, I briefly explain the format of the remaining chapters in this thesis, and give an overview of the aims and findings of each chapter.

1.4.1 Thesis format and structure

Each study in **Chapters 2-6** has been peer-reviewed and published in scientific journals. The text in each of these chapters appears as it was in the last revised version prior to publication, formatted to conform to the expected thesis standards. Therefore this is a PhD thesis by publication. As a collection of works rather than a monograph, each chapter reflects the current scientific knowledge (or knowledge of the author) at the time of writing and is written for a general scientific audience with an assumed level of expertise in the given subject area of each chapter. Significant terminology not otherwise discussed in this chapter is usually explained in the text of each chapter where relevant. Each study chapter contains its own chapter outline and discussion of relevant findings and other literature. The final chapter (**Chapter 7**) is a discussion of future perspectives based on the work in this thesis .

1.4.2 Oomycete genome evolution: interdomain HGT and phylogenomics

For the oomycetes, who are undersampled at the genomics level, I performed two "pioneer" genomics studies: the first an analysis of inter-domain HGT into plant pathogenic oomycete genomes (**Chapter 2**) and the second a phylogenomic reconstruction of oomycete evolutionary history based on the range of genomics data available at the time (**Chapter 3**). Both studies are "pioneer" in the sense that they 1) establish the incidences of transfer of bacterially or archaeally-inherited genes into oomycete genomes as low, but present and 2) represent the first phylogeny for the oomycete class using genome-level data, as opposed to single or multi-gene data, lending a greater degree of clarity to our understanding of oomycete evolutionary history.

1.4.3 Fungal genome evolution: kingdom-level phylogenomics

For the fungi themselves, I took advantage of the greater amount of genomics data both for the kingdom at large and in terms of genomes sampled within strains to perform several large-scale genomics analyses. The first of these was a critical review and benchmarking of seven different methods of phylogenomic analysis using 84 genomes taken from across the fungal kingdom as a test case (**Chapter 4**). For each method, I review its previous implementation in fungal phylogenomics (if any), then perform phylogenomic reconstruction of the fungal kingdom using that method and comparing the resultant phylogeny to the literature. In this study, we found that established methods of phylogenomic reconsturction (MRP supertree, ML/Bayesian supermatrix) generated phylogenies which were consistent with the established view of fungal phylogeny. A contemporary method of phylogenomic reconstruction (ST-RF supertree) also generated a phylogeny consistent with the literature, suggesting that ST-RF supertree reconstruction could become a useful comparison with other methods in the future. Other methods (e.g. Average Consensus, Maximum Parsimony) produce more aberrant phylogenies and have other disadvantages in terms of computational time, which is discussed in greater detail in **Chapter 4**.

1.4.4 Fungal genome evolution: pangenomics of model and non-model fungi

The second fungal study was a large-scale comparative analysis of the evolution, function and structure of the pangenomes of four model fungal organisms: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* and *Aspergillus fumigatus* (**Chapter 5**). Our analysis showed evidence for a syntenic "core" genome of 80%-90% of all gene content within each species, consistent with analyses of other eukaryotes. Preliminary analysis suggests that fungal pangenomes evolves *via* gene duplication as opposed to HGT as seen in prokaryotes. I also perform a number of characterization analyses of fungal core and accessory genomes to establish their functional and structural diversity among different fungi. The third and final study was a refinement and reimplementation of the methodology of **Chapter 5** intended for general release as the pangenomics pipeline "**Pangloss**", which included a reanalysis of the *Aspergillus fumigatus* pangenome data from **Chapter 5** and analysis of the pangenome of the oleaginous yeast *Yarrowia lipolytica* (**Chapter 6**). Compared to our *ad hoc*

methodology in **Chapter 5**, Pangloss features improvements in gene prediction methodology and data visualization capabilities in addition to greater ease-of-use for pangenome analysis.

1.4.5 Discussion and future perspective of microbial eukaryote genomics

The final chapter of this thesis, **Chapter 7,** consists of a short discussion of what future research may emerge within some of the topics covered in my Ph.D. research. For the oomycetes, I discuss what future oomycete genome evolution research may be conducted building on some of the observations in this thesis; including resolving the problematic branches of the oomycete evolutionary tree and broader studies of molecular evolution and diversity of oomycetes both for individual species and the class as a whole. For the fungi, I discuss the increasing abundance of genomics data available for the kingdom and how this abundance of data will affect fungal genome evolution research along similar lines.

Chapter 2 – Systematic search for evidence of inter-domain horizontal gene transfer from prokaryotes to oomycota lineages

This chapter was previously published in *mSphere* in September 2016.

McCarthy C. G. P. & Fitzpatrick D. A. (2016). Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages. *mSphere*, 1(5):e00195-16.

Chapter outline

While most commonly associated with prokaryotes, HGT can also have a significant influence of the evolution of eukaryotes. Systematic analysis of HGT in the genomes of the oomycetes, filamentous eukaryotic microorganisms in the SAR supergroup, has to date focused mainly on intra-domain transfer events between oomycetes and fungi. Using systematic whole genome analysis followed by phylogenetic reconstruction, we have investigated the extent of inter-domain HGT between bacteria and plant pathogenic oomycetes. We report five putative instances of HGT from bacteria into the oomycetes. Two transfers are found in *Phytophthora* species, including one unique to the cucurbit pathogen *Phytophthora capsici*. Two are found in *Pythium* species only and the final transfer event is present in *Phytopythium* and *Pythium* species, the first reported bacteria-inherited genes in these genera. Our putative transfers include one protein that appears to be a member of the *Pythium* secretome, metabolic proteins, and enzymes that could potentially break down xenobiotics within the cell. Our findings complement both previous reports of bacterial genes in oomycete and SAR genomes, and the growing body of evidence that inter-domain transfer from prokaryotes into eukaryotes occurs more frequently than previously thought.

2.1 Introduction

2.1.1 Horizontal gene transfer in eukaryotes

Horizontal gene transfer, "the non-genealogical transfer of genetic material from one organism to another" (Goldenfeld and Woese, 2007), is most closely associated with antimicrobial resistance in bacteria. The cumulative effect of transfer events has had a significant impact on overall prokaryotic genome evolution. For example it is estimated up to 80% of genes in some prokaryote genomes have undergone intra-domain HGT at some point in their history (Dagan, Artzy-Randrup and Martin, 2008). Inter-domain transfer of genetic material between prokaryotes and eukaryotes has previously been understood in the context of endosymbiotic gene transfer, which has made a significant contribution to the evolution of eukaryotic genomes (Keeling and Palmer, 2008), most notably in the evolution of the mitochondrion in eukaryotes through an ancestral primary endosymbiosis event with a *Rickettsia*-like α-proteobacterium, and the evolution of the plastid in the Archaeplastida through ancestral primary endosymbiosis with a cyanobacterium (Soucy, Huang and Gogarten, 2015). However, there is a growing body of literature supporting the existence of HGT between prokaryotes and eukaryotes, and many non-endosymbiotic horizontal inter-domain gene transfer events between bacteria and eukaryotes have been described (Dunning Hotopp, 2011). Numerous metabolic genes have been transferred into the genomes of parasitic microbial eukaryotes (Alsmark *et al.*, 2013; Hirt, Alsmark and Embley, 2015). Over 700 bacterial genes are present across fungi with particular concentration in Pezizomycotina (Marcet-Houben and Gabaldón, 2010), 71 putative bacterial genes have been identified in *Hydra vulgaris* (Chapman *et al.*, 2010), and the plant parasitic nematode *Meloidogyne incognita* secretes cell wall-degrading enzymes inherited from soil-dwelling Actinomycetales and the β-proteobacterium *Ralstonia solanacearum* (Danchin *et al.*, 2010).

2.1.2 Diversity and ecological roles of the oomycetes

The oomycetes are a class of microscopic eukaryotes placed in the diverse stramenopile (or heterokont) lineage within the Stramenopiles-Alveolata-Rhizaria eukaryotic supergroup (Burki, 2014). Historically classified as fungi due to their filamentous growth and similar ecological roles, oomycetes can be distinguished from "true" fungi by a number of structural, metabolic and reproductive differences (Beakes, Glockling and Sekimoto, 2012). The present placement of the oomycetes within the

stramenopile lineage, and by extension within the SAR supergroup, is supported by phylogenomic analyses of 18S rRNA, conserved protein and EST data, which also support the supergroup's monophyly over previous configurations such as "chromalveolates" (Burki *et al.*, 2007; Shalchian-Tabrizi *et al.*, 2007; Hampl *et al.*, 2009; Gaston and Roger, 2013).

The most ecologically destructive orders within the oomycetes are the Saprolegniales order, known as "cotton moulds", which include marine and freshwater pathogens of fish, and the closely related and predominantly terrestrial plant pathogenic orders Peronosporales and Pythiales (Jiang and Tyler, 2012). The Pythiales order includes members of the marine and terrestrial genus *Pythium*, necrotrophic generalistic causative agents of root rot and damping off in many terrestrial plants (**Table 2.1**). Some species (*Pythium aphanidermatum* and *Pythium ultimum*) are found in high-temperature or greenhouse conditions, while others (*Pythium irregulare* and *Pythium iwayami*) are most virulent at lower temperatures (Adhikari *et al.*, 2013). *Pythium ultimum* and *Pythium irregulare* have broad ecological host ranges, while *P. iwayami* and *Pythium arrhenomanes* display some preference for monocots (Lévesque and de Cock, 2004; Adhikari *et al.*, 2013).

Table 2.1. Summary of host ranges or optimum environments of oomycete species analysed in this study.

Species	**Host(s)**
Phytophthora capsici	Curcubits (e.g. *Cucurbita pepo*)
Phytophthora infestans	Solanaceae (e.g. *Solanum tuberosum*)
Phytophthora kernoviae	*Fagus sylvatica*, *Rhododendron*
Phytophthora lateralis	*Chamaecyparis lawsoniana*
Phytophthora parasitica	Broad range, incl. *Nicotiana tabacum*
Phytophthora ramorum	Broad range, incl. *Quercus*, *Rhododendron*
Phytophthora sojae	*Glycine max*
Phytopythium vexans	Tropical forest species
Pythium aphanidermatum	Broad range, virulent at higher temperatures
Pythium arrhenomanes	Monocots
Pythium irregulare	Broad range, virulent at lower temperatures
Pythium iwayami	Monocots, virulent at lower temperatures
Pythium ultimum var. *sporangiiferum*	Broad range
Pythium ultimum var. *ultimum*	Broad range, virulent at higher temperatures

The Peronosporales order includes the paraphyletic hemibiotrophic genus *Phytophthora*, whose member species exhibit both broad and highly specialized host ranges (**Table 2.1**). Generalistic *Pythophthora* species include *Phytophthora ramorum* and *Phytophthora kernoviae* (sudden oak death and dieback in many other plant species, particularly *Rhododendron*), *Phytophthora parasitica* (black shank disease in a diverse

range of plants) and *Phytophthora capsici* (blight and root rot in Cucurbitaceae, Solanaceae and Fabaceae). Species with more specialized host ranges include *Phytophthora sojae* and *Phytophthora lateralis* (root rot in soybean and Port Orford cedar, respectively), and *Phytophthora infestans* (late blight in some Solanaceae, most notoriously in potato). The tropical plant pathogen *Phytopythium vexans* was previously classified in *Pythium* clade K (Lévesque and de Cock, 2004), but that clade has since been reclassified into *Phytopythium*, a morphological and phylogenetic intermediate genus between *Phytopthora* and *Pythium* (de Cock *et al.*, 2015).

2.1.3 Interdomain HGT in oomycetes

To date, large scale systematic analysis of the influence of HGT on oomycete genome evolution has focused on intra-domain transfer between fungi and oomycetes (Judelson, 2012; Savory, Leonard and Richards, 2015). The most extensive study revealed up to 34 putative transfers from fungi to oomycetes, many of which were enzymes involved in carbohydrate metabolism (Richards *et al.*, 2011). Three of these genes had previously been transferred from bacteria to fungi (Richards *et al.*, 2006). The number of HGT events between bacteria and oomycetes described in the literature is sparse and most incidents of inter-domain HGT have been discovered within the context of fungi-focused studies. However, recent analyses have shown Actinobacterial cutinase has orthologs in a number of *Phytophthora* species (Belbahri *et al.*, 2008) with subsequent copy expansion in *Phytophthora sojae*. Disintegrins and endonucleases secreted by *Saprolegnia parasitica* appear to be bacterial in origin (Jiang *et al.*, 2013), and studies of the secretomes of the Saprolegniales species *Achlya hypogyna* and *Thraustotheca clavata* revealed one ancestral endoglucanase and three genes specific to the Saprolegniales order which had been transferred from bacteria (Misner *et al.*, 2015). As with other unicellular eukaryotes, some genes in *Phytophtora* involved in amino acid metabolism have been obtained *via* horizontal transfer from bacteria (Whitaker, McConkey and Westhead, 2009). Other studies have identified ancestral bacterial HGT events within other stramenopile genomes (Bowler *et al.*, 2008) or in other lineages within the SAR supergroup (Nosenko and Bhattacharya, 2007; Martens, Vandepoele and Van de Peer, 2008; Morris *et al.*, 2009).

In light of these previous studies of the influence of HGT in the evolution of the oomycetes, we undertook a systematic investigation focusing on the extent of bacterial

transfer into the oomycetes. We analysed 13 species from the plant pathogenic genera *Pythium* and *Phytophthora*, as well as the recently reclassified species *Phytopythium vexans*, for genes with sufficient evidence for non-vertical inheritance from bacteria. Here, we report five recent transfers from bacteria into individual oomycete lineages, including what we believe to be the first descriptions of inter-domain HGT involving *Pythium*.

2.2 Materials and Methods

2.2.1 Dataset assembly

The predicted proteomes for seven *Phytophthora* species (*P. capsici*, *P. infestans*, *P. kernoviae*, *P. lateralis*, *P. parasitica*, *P. ramorum* and *P. sojae*), *Phytopythium vexans*, and six *Pythium* species (*P. aphanidermatum*, *P. arrhenomanes*, *P. irregulare*, *P. iwayami*, *P. ultimum* var. *sporangiiferum* and *P. ultimum* var. *ultimum*) were analysed for possible bacterial-oomycete HGT events. To ensure a broad taxon sampling for the oomycetes as a whole, we downloaded all available oomycete genome data from public databases. The predicted proteomes of the Peronosporales species *Hyaloperonospora arabidopsidis* (Baxter *et al.*, 2010) and *Albugo laibachii* (Kemen *et al.*, 2011), the predicted proteomes of the Saprolegniales species *Saprolegnia parasitica* (Jiang *et al.*, 2013), *Saprolegnia diclina*, *Aphanomyces invadans* and *Aphanomyces astaci* (Broad Institute), and the secretomes of the Saprolegniales species *Achyla hypogyna* and *Thraustotheca clavata* (Misner *et al.*, 2015) were included in our local database. To cover taxon sampling of the stramenopiles, the predicted proteomes of the two diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (Armbrust *et al.*, 2004; Bowler *et al.*, 2008), and the alga *Aureococcus anophagefferens* (Gobler *et al.*, 2011) were also included. In addition to our oomycete and stramenopile data, our database contained all non-redundant prokaryotic protein data available. To construct this portion and reduce redundancy a representative genome from each prokaryotic species in the full NCBI GenBank database (Benson *et al.*, 2013) was included. In total, just under 5 million protein sequences from 1486 prokaryotic genomes were retained. More than 3 million sequences from 212 eukaryotic nuclear genomes were included, sampling a diverse range of animal, plant and fungal lineages (**Table S2.1**).

2.2.2 Identification of putative bacteria-oomycete HGT events

Our methods for identifying candidate bacterial HGT genes followed those of Richards et al. (Richards *et al.*, 2011) in their analysis of fungal HGT genes in the oomycetes. Repetitive and transposable elements were identified and removed from each *Phytophthora* and *Phytopythium*/*Pythium* proteome by performing homology searches against Repbase (Jurka *et al.*, 2005), using tBLASTn (Ramsay *et al.*, 2000; Camacho *et*

al., 2009) with an e-value cutoff of 10^{-20} (**Table 2.2**). The remaining protein sequences in each oomycete proteome were then further filtered and clustered into groups of paralogs using OrthoMCL (Li, Stoeckert and Roos, 2003), with an e-value cutoff of 10^{-20} and an inflation value of 1.5 (**Table 2.2**). Representative sequences from each group of paralogs, along with unclustered singleton sequences, were retrieved from their respective proteomes. These sequences were then queried against our local database using BLASTp with an e-value cutoff of 10^{-20}.

Using bespoke Python scripting we identified 106 genes whose homology supported a bacterial transfer into an individual oomycete lineage (proteins whose first hit outside their own genus was bacterial) and retrieved them for a second round of OrthoMCL clustering to remove redundancy in our datasets for each genus (**Table 2.2**). All retrieved protein sequences were clustered into groups of orthologs using OrthoMCL with an e-value cutoff of 10^{-20} and an inflation value of 1.5 (**Table 2.3**). 64 representative and singleton sequences from these datasets were then queried against our local database using BLASTp with an e-value cutoff of 10^{-20} and an arbitrary limit for maximum hits per query sequence. The corresponding gene family for each candidate HGT gene was constructed from our BLASTp results.

Table 2.2. Identification of sequences with high bacterial homology as candidate HGT events within oomycete genomes.

Proteome	Initial size	After Repbase filtering	OrthoMCL clusters (# of sequences)	OrthoMCL unclustered sequences	Intergenic bacterial hits
Phytophthora capsici	19,805	16,169	1,732 (8,879)	7,290	6
Phytophthora infestans	18,140	17,013	2,032 (9,459)	7,553	2
Phytophthora kernoviae	10,650	10,435	750 (3,244)	7,016	0
Phytophthora lateralis	11,635	10,539	880 (4,110)	6,337	14
Phytophthora parasitica	20,822	18,640	2,084 (10,153)	8,437	2
Phytophthora ramorum	15,743	13,403	1,639 (7,839)	5,564	5
Phytophthora sojae	26,584	22,210	2,418 (13,544)	8,666	2
Phytopythium vexans	11,958	11,634	1,097 (4,932)	6,702	7
Pythium aphanidermatum	12,312	12,002	1,144 (5,129)	6,873	11
Pythium arrhenomanes	13,805	13,224	1,221 (5,647)	7,577	18
Pythium irregulare	13,805	13,297	1,214 (5,888)	7,409	6
Pythium iwayami	14,875	14,279	1,303 (6,185)	8,094	6
Pythium ultimum var. *sporangiiferum*	14,096	13,915	917 (4,208)	9,707	13
Pythium ultimum var. *ultimum*	15,323	14,780	1,305 (6,661)	8,119	14

Table 2.3. Identification of putative bacterial HGT sequences in *Phytophthora*, *Pythium* and *Phytopythium*.

Genus	Intergenic bacterial hits	OrthoMCL clusters (# of sequences)	OrthoMCL unclustered sequences	Maximum-likelihood phylogenies	Putative HGT sequences
Phytophthora	31	22 (28)	3	25	3
Phytopythium / *Pythium*	75	16 (59)	23	39	2

2.2.3 Phylogenetic reconstruction of putative bacteria-oomycete HGT events

64 candidate HGT gene families were aligned using MUSCLE (Edgar, 2004) and best-fit amino acid replacement models were selected for each alignment using ProtTest (Darriba *et al.*, 2011). Maximum likelihood phylogenetic reconstruction for each alignment was carried out using PhyML (Guindon *et al.*, 2010) with 100 bootstrap replicates. Each phylogenetic tree was visualized and annotated with GenBank data using bespoke Python scripting and iTOL (Letunic and Bork, 2016). Additional phylogenetic analysis using consensus network methods was carried out using SplitsTree (Huson and Bryant, 2006).

2.2.4 Analysis of bacterial contamination and taxon sampling

Seed genes and their directly adjacent gene were examined for their particular homology; to determine whether candidate HGT genes were not simply the result of bacterial contamination of genomes along particular contigs or scaffolds. For each seed gene arising from *P. capsici*, the genomic location of that gene was identified by querying its corresponding protein sequence against the JGI *P. capsici* database (http://genome.jgi.doe.gov/PhycaF7) using tBLASTn with an e-value cutoff of 10^{-4}. Homology data for each seed gene and their adjacent genes were provided by the JGI *P. capsici* genome browser (**Table S2.2**). For each *Pythium* seed gene, the genomic location of the gene was identified by querying the corresponding protein sequence against the genomic scaffolds of the source species using tBLASTn with an e-value cutoff of 10^{-4}, and then the seed gene's corresponding protein sequence and its two adjacent protein sequences were queried against the NCBI's non-redundant protein sequence database using BLASTp with an e-value cutoff of 10^{-20} (**Table S2.2**).

For studies of HGT in eukaryotes, particularly transfer between prokaryotes and eukaryotes, it is essential that genomic data covers as broad a range of taxa, to prevent as much as possible the introduction of bias into analysis and thus reduce the likelihood of obtaining false positive for transfer events (Huang, 2013; Gluck-Thaler and Slot, 2015). Comparison of the taxon sampling in our database with the NCBI data was performed by searching each seed gene's protein sequence against the NCBI non-redundant protein sequence database using BLASTp with an e-value cutoff of 10^{-20}. The seed sequence and its homologs were aligned in MUSCLE and neighbour-joining trees were constructed in

QuickTree (Howe, Bateman and Durbin, 2002) using 100 bootstrap replicates, and each tree was annotated with GenBank data. Maximum-likelihood HGT phylogenies whose topology conflicted substantially with their corresponding neighbour-joining tree due to differences in taxon sampling were excluded from further analysis.

2.2.5 Characterization and functional annotation of putative bacteria-oomycete HGT families

For the remaining putative HGT families, bespoke Python scripting was used to calculate the sequence length, GC-content and exon number of each oomycete gene present. The average sequence length, GC-content and exon number for each *Phytophthora*, *Phytopythium* and *Pythium* genome was also calculated (**Table S2.3**). Additionally, the sequence length and GC-content of one or more bacterial sister genes were calculated using bespoke Python scripting (**Table S2.4**). Optimal local alignments of each seed protein sequence against a representative bacterial sister gene was generated using CLUSTAL Omega (Rice, Longden and Bleasby, 2000) (**Table S2.5**). Putative function of each putative HGT family was annotated by performing initial PFAM homology searches of each seed protein sequence (Finn *et al.*, 2015) (**Table S2.6**) with an e-value cutoff of 10^{-4} and BLAST homology searches against the NCBI's non-redundant protein database with an e-value cutoff of 10^{-20}. To complement these initial annotations, each seed protein sequence was then analysed in InterProScan (Jones *et al.*, 2014). Signal peptide and subcellular localization prediction analysis for each seed protein sequence was carried out using SignalP and TargetP, respectively (Emanuelsson *et al.*, 2000; Petersen *et al.*, 2011), with the default parameters. Multivariate codon usage analysis of each genome was carried out using GCUA (McInerney, 1998), and each (**Table S2.7**).

2.3 Results and Discussion

2.3.1 Analysis of bacterial HGT into *Phytophthora* and *Pythium*

To investigate the extent of bacterial HGT into the oomycetes, we generated gene phylogenies for every oomycete protein sequence whose bidirectional homology analysis supported a recent transfer from bacteria to an oomycete species. Such phylogenies were generated with techniques that have previously identified multiple intra-domain HGT events between fungi and oomycetes (18); using OrthoMCL (Li, Stoeckert and Roos, 2003) to generate clusters of orthologous proteins, searching representative proteins against a large database using BLASTp (Altschul *et al.*, 1997), and generation of maximum-likelihood phylogenetic reconstructions using PhyML (Guindon and Gascuel, 2003). To reduce the chances of false positive identification of putative HGT genes due to poor taxon sampling (Huang, 2013; Gluck-Thaler and Slot, 2015), oomycete protein sequences were queried against a local database using BLASTp, with broad taxon sampling in the database across prokaryotes and eukaryotes (**Table S2.1**). 106 oomycete proteins were found to have a top database hit with a bacterial protein. Filtering for redundancy (due to multiple homologs in a single species for example), 64 unique candidate maximum-likelihood HGT phylogenies with 100 bootstrap replicates (**Table 2.2**) were generated using PhyML with the best-fit model for each phylogeny chosen by ProtTest (Darriba *et al.*, 2011). Of these 64 phylogenies, 59 were ultimately discarded due to poor bootstrap support, inadequate taxon sampling or irresolvable topology (**Table 2.3**). Our phylogenies infer three types of bacteria-oomycete HGT within our candidate HGT phylogenies;

1) Recent bacterial transfer into the *Pythium* / *Phytopythium* lineage (1 individual incidences)
2) Recent bacterial transfer into the *Phytophthora* lineage (2 individual incidences).
3) Recent bacterial transfer into the *Pythium* lineage (2 individual incidences).

To help ensure that none of our putative HGT families were in fact the product of bacterial contamination, the homology of each seed gene and its adjacent genes were investigated. In each of our five putative HGT families we found that there was no obvious evidence of bacterial contamination along a source contig resulting in false positives for bacterial-oomycete HGT events (**Table S2.2**). As we were also conscious of

the risk of poor taxon sampling giving us false positives, we also compared the taxon sampling in our local database with the NCBI protein data. We queried each seed protein sequence against the NCBI's non-redundant protein sequence database using BLASTp with an e-value cutoff of 10^{-20}, aligned homologs and generated neighbour-joining phylogenies for each seed gene (not shown). Where the BLASTp data retrieved from NCBI mirrored our own local searches, and the corresponding neighbour-joining phylogeny showed the seed gene clearly grouped within an oomycete clade or a bacterial clade we were satisfied that our taxon sampling had sufficiently covered all available protein data. All of our 5 candidate HGT genes satisfy these criteria. Each phylogeny was evaluated for other characteristics that may have reinforced or rejected our hypothesis that HGT had occurred. Gene characteristics such as GC-content, exon number and sequence length of each oomycete gene arising from transfer in our phylogenies was calculated (**Table S2.3**) and compared to the average of their corresponding genome. Gene characteristics of bacterial homologs in potential donor species were also calculated (**Table S2.4**). Sequence similarity and identity at the amino acid level between each seed HGT protein and a sister homolog from a potential bacterial donor was also investigated (**Table S2.5**). Similarly, the codon usage patterns of the seed genes used to generate each phylogeny were also compared with the codon usage patterns of their potential bacterial donors (**Table S2.7**). No codon usage pattern analysis was conclusive in proving or disproving that horizontal inheritance of these genes had occurred. However, this is not uncommon for codon usage analyses as the codon usage of transferred genes is known to ameliorate to match that of the recipient genome (Koski, Morton and Golding, 2001).

We have identified five well supported phylogenies that show putative HGT events from bacterial species into the oomycetes. Three display topologies supporting a recent transfer into the *Pythium* or *Phytopythium* lineage, (**Figures 2.1-2.3**), while the remaining two support a recent HGT into the *Phythophthora* lineage (**Figures 2.4 & 2.5**). Below, we present and discuss each recent transfer individually; describing both the hypothesis for horizontal inheritance in each phylogenetic reconstruction and the functional characterization of each transferred gene family. We also compare the placement of the oomycete homologs in each of the five phylogenies with those of other eukaryotic homologs, particularly fungi, so as to illustrate that these genes were not inherited vertically through a shared eukaryotic lineage. Each transfer is also summarized in **Table 2.4**.

Table 2.4. Summary of each putative bacterial-oomycete HGT event.

Tree	Seed species	Potential donor(s)	Identity (%)	Putative function	Secreted
Figure 2.1	*P. ultimum*	*C. aerophila*	56.5	Class II fumarase	No
Figure 2.2	*P. aphanidermatum*	Proteobacteria	54.0	NmrA-like quinone oxidoreductase	No
Figure 2.3	*P. aphanidermatum*	Actinobacteria	58.6	SnoaL-like polyketide cyclase	Yes
Figure 2.4	*Ph. capsici*	*M. radiotolerans*	68.2	Epoxide hydrolase	No
Figure 2.5	*Ph. capsici*	*Sphingomonas*	59.1	Alcohol dehydrogenase	No

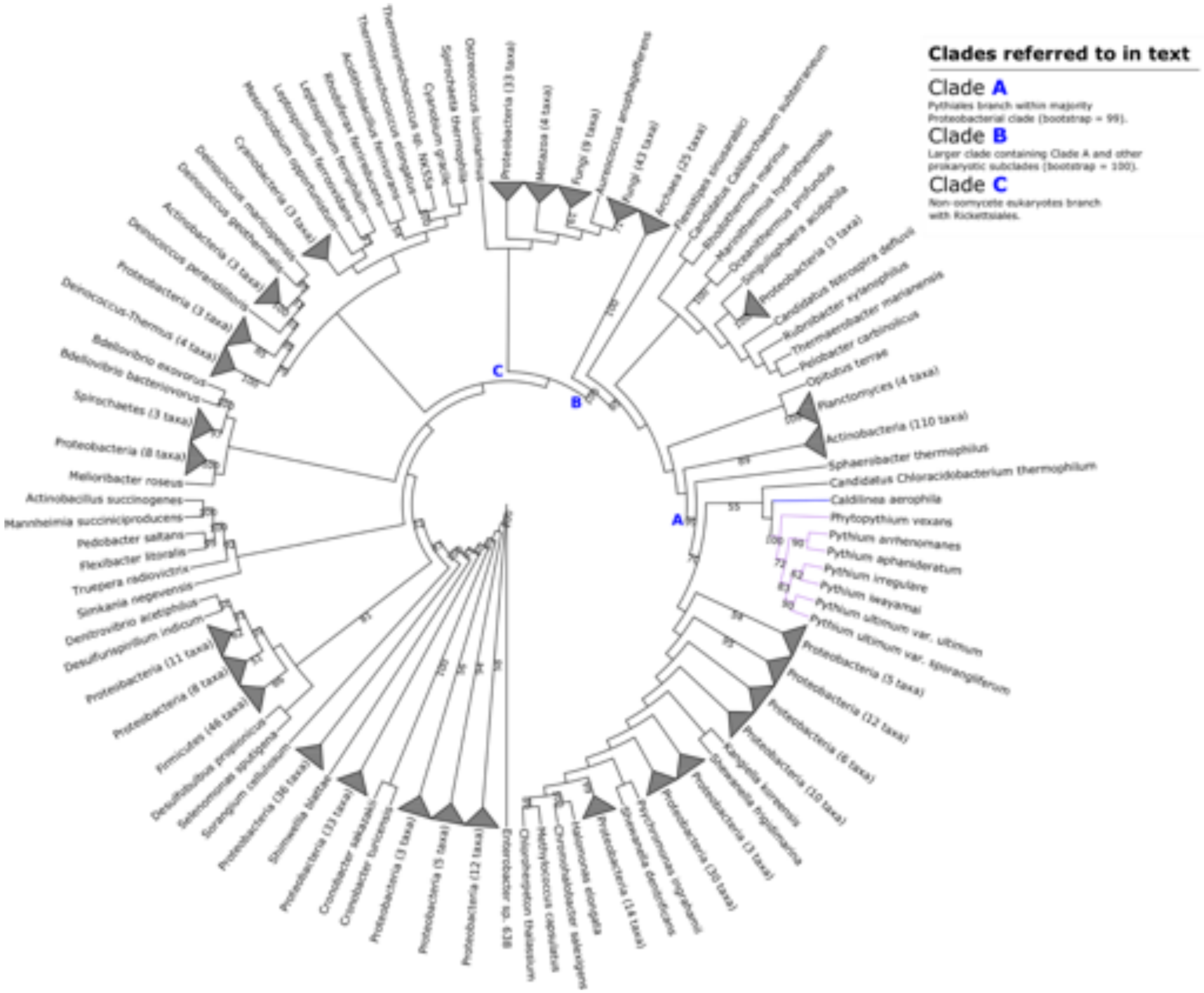


Figure 2.1. Maximum likelihood phylogeny illustrating putative transfer of class II fumarase from *Caldilinea aerophila* into *Phytopythium* / *Pythium* lineage. Clades A, B & C referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. See **Figure S2.1** in **Supplementary Material** for full phylogenetic tree.

2.3.2 A putative class II fumarase distinct from *Rickettsia* class II fumarase in *Phytopythium vexans* and *Pythium* spp. originates from bacteria

A protein in *Pythium ultimum* var. *sporangiiferum* (**Table 2.4**) was identified in our BLASTp homology searches as a candidate for an inter-domain HGT event into oomycete species. The maximum-likelihood phylogeny of this protein family was generated from a family containing 550 homologs, with a LG+I+G+F substitution model (**Figure 2.1**). 16 bacterial phyla are present in this reconstruction, of which Proteobacteria and Actinobacteria are by far the most represented. Twenty-six archaeal homologs are also present, of which all bar a *Candidatus* Caldiarchaeum subterraneum sequence form a monophyletic clade. Across the eukaryotes, homologs are present in fungi, animals, green algae and the stramenopiles.

Our phylogenetic reconstruction shows a monophyletic *Pythium*/*Phytopythium* clade within a large, predominantly Proteobacterial clade with 99% bootstrap support, adjacent to a homolog from the filamentous Chloroflexi species *Caldilinea aerophila* (**Figure 2.1**, **Clade A**). Further back along the tree, this greater subclade branches deep within a large prokaryotic clade with 100% bootstrap support, containing three major subclades; the aforementioned majority-Proteobacterial subclade containing *Pythium* and *Phytopythium* orthologs, a halophilic archaeal subclade, and a large Actinobacterial subclade containing 110 homologs (**Figure 2.1**, **Clade B**). Elsewhere, all non-oomycete eukaryote homologs (with the exception of an adjacent sequence from the microscopic green alga *Ostreococcus lucimarinus*) place in a monophyletic eukaryote clade containing 52 fungal homologs, 4 animal homologs and a homolog from the stramenopile alga *Aureococcus anophagefferns* adjacent to a clade containing 19 homologs from the α-Proteobacterial *Rickettsia* genus (**Figure 2.1**, **Clade C**). The neighbour-joining tree constructed from the BLAST homology search of the seed sequence against the NCBI's database places the seed deep within a large prokaryotic clade containing Proteobacteria, Actinobacteria and halophilic and methanogenic archaea, in a γ-Proteobacterial subclade similar to what we observe in our phylogenetic reconstruction (not shown).

Sequence analysis of the seed gene and its flanking genes in the *P. ultimum* var. *sporangiiferum* genome did not return any obvious evidence of bacterial contamination; the seed protein sequence's top hit against the NCBI database was a *C. aerophila* sequence, but the top hits of both flanking protein sequences were *Phytophthora*

parasitica homologs (**Table S2.2**). BLAST homology searches against the NCBI database found the seed sequence shared sequence similarity with many bacterial class II fumarases, and PFAM analysis of the sequence identified two lyase domains and the characteristic *FumC* C-terminal of a class II fumarase-like sequence (**Table S2.6**). InterProScan analysis identified further fumarase protein sequence signatures (**Table S2.6**). Fumarase, also known as fumarate hydratase (E.C. 4.2.1.2), is an enzyme that catalyses the reversible hydration of fumarate to (S)-malate in the mitochondrion in eukaryotes, as a component of the tricarboxylic acid cycle (Yogev *et al.*, 2010), and promotion of histone H3 methylation and DNA repair in the cytosol (Jiang *et al.*, 2015). There are two classes of fumarase; the heat-labile dimeric class I fumarases *FumA* and *FumB* found in prokaryotes and the heat-stable tetrameric class II fumarase *FumC* found in both prokaryotes and eukaryotes (Estévez *et al.*, 2002). While associated with mitochondrial function in eukaryotes, class II fumarases with distinct evolutionary histories have been detected in amitochondriate trichomonads (Gerbod *et al.*, 2001).

The nature of class II fumarase's conserved function in eukaryotic respiration would suggest that this gene had arisen in the nuclear genome of *Pythium* and *Phytopythium* gene by endosymbiotic gene transfer from the mitochondrial genome (Timmis *et al.*, 2004), and hence was not a product of recent transfer. To investigate the relationship between this putative horizontally-transferred fumarase and other potential fumarase orthologs in the oomycetes, we aligned the seed *Pythium ultimum* var. *sporangiiferum* sequence against 20 known oomycete and 230 other eukaryote and prokaryote class II fumarase sequences. Sequence and phylogenetic analysis show it branches as an outgroup in the corresponding phylogeny (not shown), suggesting that it is not an ortholog of the endosymbiotic oomycete class II fumarase. It seems most parsimonious to suggest therefore that this fumarase protein in *Pythium* and *Phytopythium vexans* is a class II fumarase distinct from endosymbiotic class II fumarase, and has arisen by a completely separate transfer event, possibly with *C. aerophila* or another Chloroflexi species (*Sphaerobacter thermophilus* for example) (**Figure 2.1**). An interesting aspect of this phylogeny is the presence of a homolog from *Phytopythium vexans* branching with *Pythium* species and the absence of *Phytophthora* homologs in the phylogeny. *Phytopythium vexans*, along with other members of what was once *Pythium* clade K, were reclassified to the morphological intermediate genus *Phytopythium,* based on molecular evidence from ribosomal large subunit (LSU), internal transcribed spacer (ITS) and mitochondrial cytochrome oxidase 1 (CO1) data. Furthermore the resultant phylogenetic

data grouped *Phytopythium* and *Phytophthora* as sister taxa with strong bootstrap support (de Cock *et al.*, 2015). This would suggest that the ancestor of the *Phytophthora, Phytopythium* and *Pythium* species obtained a bacterial copy of the class II fumarase and it was subsequently lost in the *Phytophthora* clade. Alternatively if we assume that rare HGT events can act as phylogenetic markers (Keeling and Palmer, 2008), it is plausible that in fact *Phytopythium* and *Pythium* are more closely related to one another to the exclusion of *Phytophthora* species. This observation challenges the phylogeny derived from traditional phylogenetic markers (de Cock *et al.*, 2015) and we suggest the relationships between these groups warrants further examination.

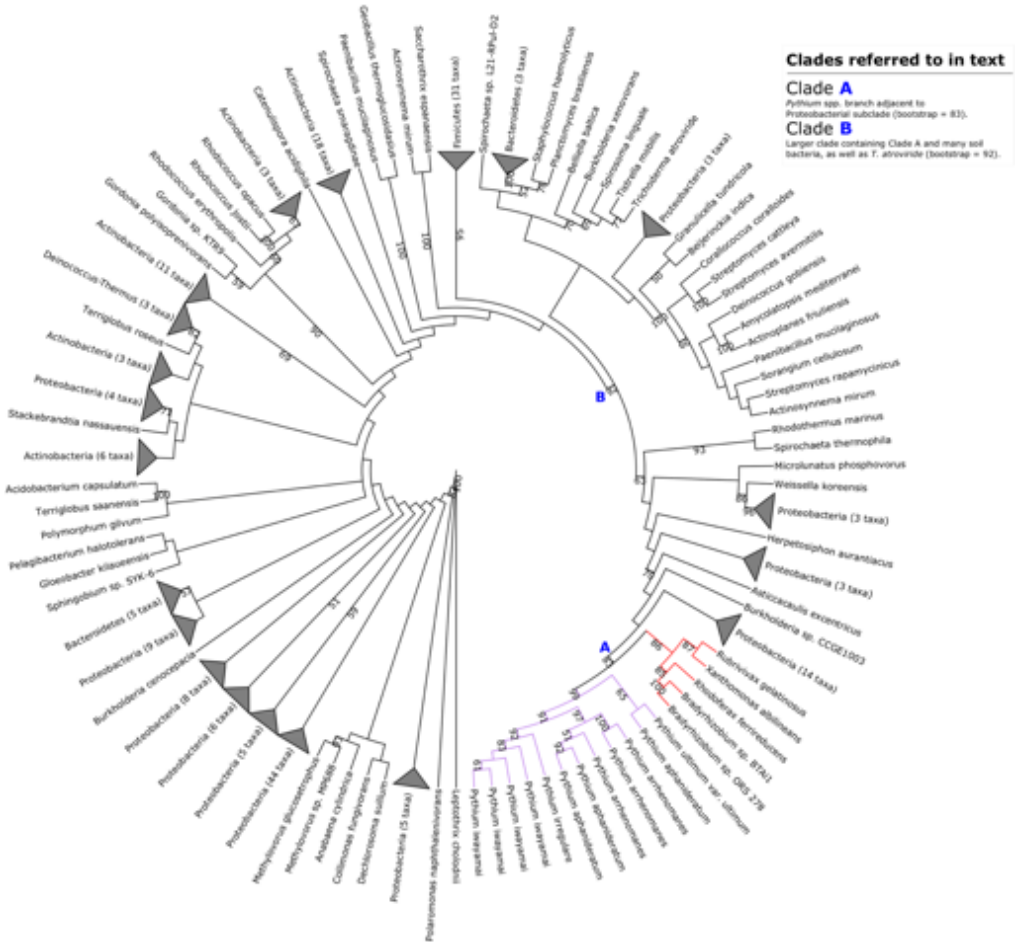


Figure 2.2. Maximum likelihood phylogeny illustrating putative transfer of NmrA-like quinone oxidoreductase from Proteobacteria into *Pythium* spp. Clades A & B referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. See **Figure S2.2** in **Supplementary Material** for full phylogenetic tree. *T. atroviride*: *Trichoderma atroviride*.

2.3.3 A putative proteobacterial NmrA-like oxidoreductase is present in multiple *Pythium* species

A *Pythium aphanidermatum* gene (**Table 2.4**) was identified in our homology searches as a candidate for bacterial HGT into an oomycete species. The maximum-likelihood phylogeny of this gene was constructed from a gene family containing 258 homologs, with a LG+I+G+F substitution model (**Figure 2.2**). 95% (245 of 258) of these homologs are bacterial, representing 10 different phyla. The majority of bacterial homologs are from either Proteobacteria, Actinobacteria or Firmicutes species. Of the 13 eukaryote homologs present, 12 are from the oomycetes and one is from the fungal species *Trichoderma viride* (**Figure 2.2**).

In our reconstruction, homologs (12 in total) from each *Pythium* species except *P. ultimum* var. *sporangiiferum* form a monophyletic subclade (99% bootstrap support) within a 70-member clade with 92% bootstrap support. Every other member of this clade except *Trichoderma viride* is bacterial. Around 30 members of this clade are proteobacterial, many of which are soil dwelling Rhizobales (**Figure 2.2**, **Clade B**). The *Pythium* subclade branches with 83% bootstrap support beside a small Proteobacterial subclade that includes two nitrogen-fixing species in *Bradyrhizobium* and *Xanthomonas albilineans*, the causative agent of leaf scald disease in sugarcane (Pieretti *et al.*, 2015) (**Figure 2.2**, **Clade A**). Homology analysis of the seed sequence and its flanking sequences in the *P. aphanidermatum* genome found no obvious evidence of bacterial contamination; the seed sequence was most closely related to a *Rubrivivax gelatinosus* sequence, however flanking genes have top hits from *Phytophthora infestans* (**Table S2.2**). The neighbour-joining phylogeny generated from BLAST homology searches of the seed sequence against the NCBI's protein database also placed the seed sequence adjacent to a large Proteobacterial clade.

BLAST homology searches against the NCBI database found the seed sequence shared homology with a bacterial nucleotide-sugar epimerases and NAD(P)-binding proteins. PFAM analysis of the sequence found the characteristic Rossmann fold of NAD(P)-binding proteins (**Table S2.2**), while InterProScan analysis found NmrA-like family and quinone oxidoreductase 2 subfamily PANTHER signatures (**Table S2.2**). NmrA is a NAD(P)-binding negative transcriptional regulator, involved in the regulation of nitrogen metabolite repression (NMR) genes in fungi, which suppress metabolic pathways for secondary nitrogen sources when preferred sources like ammonium and

glutamine are available (Stammers *et al.*, 2001). Such a metabolic system has not been described in oomycetes to date. The PANTHER quinone oxidoreductae subfamily (Thomas *et al.*, 2003) to which this transferred gene belongs (PTHR14194:SF73) includes eukaryotic orthologs from Pezizomycotina, *Monosiga brevicollis* and *Dictyostelium*, *Phytophthora infestans* and *Physcomitrella patens* and bacterial orthologs from multiple lineages. Among these orthologs is *qorB* in *Escherichia coli* K-12, which has redox activity on NAD(P)H using quinone as an acceptor (Kim *et al.*, 2008).

Our phyogenetic reconstruction of this *P. aphanidermatum* gene supports the transfer of this gene into *Pythium* spp. from a soil-dwelling Proteobacterium (**Figure 2.2**), either the phototrophic β-proteobacterial species *Rhodoferax ferrireducens* and *Rubrivivax gelatinosus*, or the phytopathogenic γ-proteobacterium *Xanthomonas albilineans*. Species related to *X. albilineans* and *R. ferrireducens*, within Xanthomonadales and Comamonadaceae respectively, have been identified in previous studies as endohyphal bacteria, hyphae-dwelling endosymbionts of endophytic fungi (Hoffman and Arnold, 2010; Hoffman *et al.*, 2013). It is not currently known whether such bacteria can also inhabit the hyphae of oomycetes, and consequently provide favourable conditions for potential inter-domain HGT. This transferred gene may be a NAD(P)H-binding quinone oxidoreductase (EC 1.6.5.2), and potentially have cytosolic redox activity in *Pythium* spp.

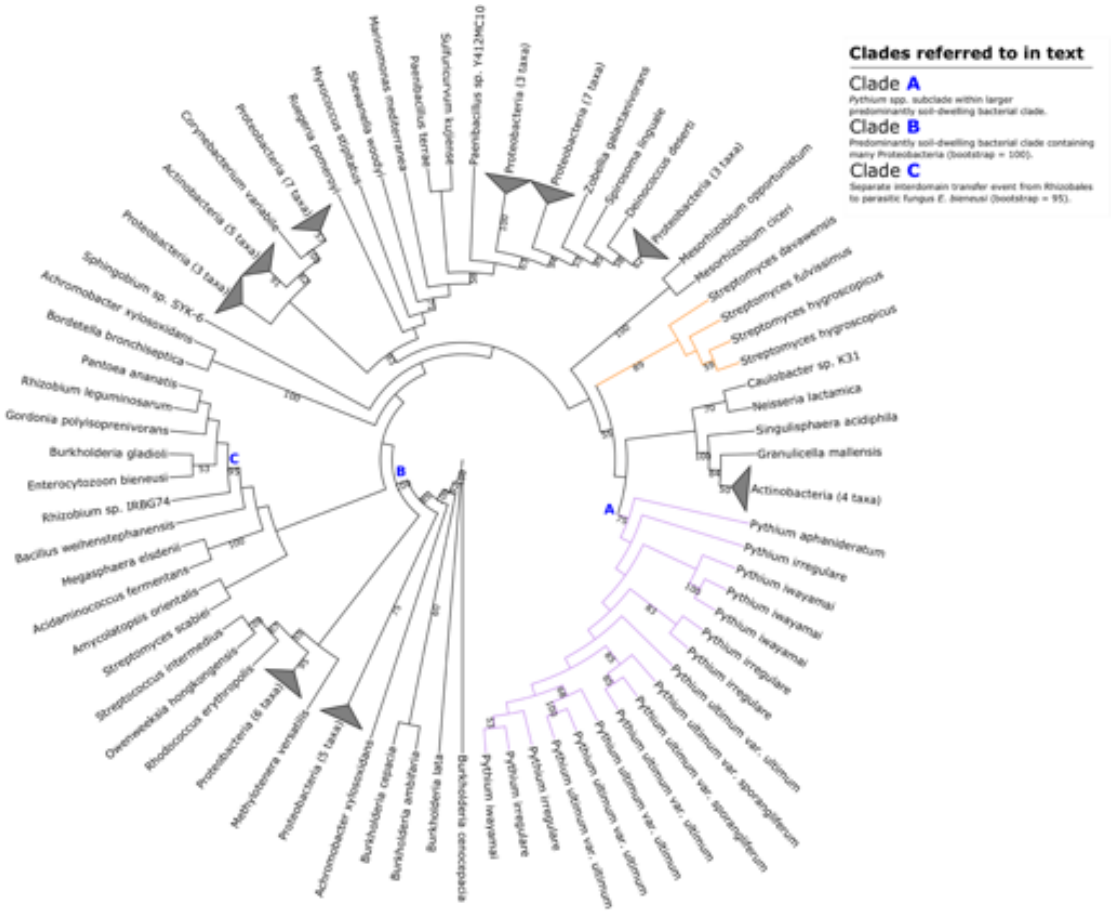


Figure 2.3. Maximum likelihood phylogeny illustrating putative transfer of SnoaL-like polyketide cyclase from Actinobacteria into *Pythium* spp. Clades A, B & C referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. See **Figure S2.3** in **Supplementary Material** for full phylogenetic tree. *T. atroviride*: *Trichoderma atroviride*.

2.3.4 SnoaL-like proteins from soil-dwelling bacteria are putative members of the secretome of multiple *Pythium* species

A second gene from *P. aphanidermatum* (**Table 2.4**) was identified in our BLASTp homology searches as a candidate for bacterial HGT into an oomycete species. The maximum-likelihood phylogeny of this gene was generated from a gene family containing 103 homologs constructed with a WAG+I+G substitution model (**Figure 2.3**). Seven bacterial phyla are present in this reconstruction, along with *Pythium* and the microsporidian parasite *Enterocytozoon bieneusi*, 53% of the homologs (55 of 103) come from Proteobacterial species.

The maximum-likelihood phylogenetic reconstruction places 17 *Pythium* homologs (with multiple paralogs in each species except *P. aphanidermatum* and no homolog in *P. arrhenomanes*) deep within a 93-member clade containing many typical soil-dwelling Proteobacterial and Actinobacterial species (**Figure 2.3**, **Clade B**) with 100% bootstrap support. The *Pythium* subclade (**Figure 2.3**, **Clade A**) is adjacent to a clade containing four orthologs from *Mycobacterium smegmatis*. The only other eukaryote homolog in our analysis (*E. bieneusi*) places in a separate subclade containing Rhizobales species with 95% bootstrap support indicative of a separate independent HGT event (**Figure 2.3**, **Clade C**). Homology analysis of the seed sequence and its adjacent sequences returned no evidence of bacterial contamination. Both flanking genes sequence are homologous to sequences in other oomycetes, and the seed sequence's highest degree of homology was with a *Streptomyces yerevanensis* sequence (**Table S2.2**).

BLAST homology searches of the seed sequence found numerous instances of homology with bacterial SnoaL-like polyketide cyclases. PFAM and InterProScan analysis of the sequence identified two SnoaL-like domains, and a number of signal peptide signatures within the N-terminal domain (**Table S2.6**). Polyketide cyclases are enzymatic components of the synthesis of aromatic polyketide compounds from carboxylic acids in bacteria and fungi. Polyketides are best characterized by the medicinally useful secondary metabolites produced by various Actinobacteria genera, such as the antitumourigenic anthracyclines from *Streptomyces* species (Strohl, 2001). Biochemically, polyketide cyclases catalyse the intramolecular cyclization of poly-β-ketone chain intermediates to form the core planar polyaromatic structures of polyketides, which are then subject to later functionalization. In the biosynthesis of the anthracycline nogalamycin in *Streptomyces nogalater*, the polyketide cyclase SnoaL (EC 5.5.1.26)

catalyses ring closure of a polyaromatic nogalamycin precursor through aldol condensation (Sultana *et al.*, 2004).

The maximum-likelihood phylogenetic reconstruction of this transfer event appears to support the transfer of this putative SnoaL-like protein into an ancestral *Pythium* from a Proteobacterial or Actinobacterial donor (**Figure 2.3**). Similarly, the neighbour-joining tree generated from the homology search against NCBI's non-redundant database places the *P. aphanidermatum* seed sequence within a large Proteobacterial and Actinobacterial clade (not shown). The SignalP (Petersen *et al.*, 2011) and TargetP (Emanuelsson *et al.*, 2000) analyses both predict that the protein contains a 25-reside long signal peptide sequence at its N-terminus with a discrimination score (used to distinguish between signal and non-signal peptides) well above the default cutoff, and thus identify the protein as part of the secretome of *P. aphanidermatum*. Therefore, this putative SnoaL-like protein may have arisen in *Pythium* species through horizontal transfer from an Actinobacteria species and may be a putative component of the secretome of *Pythium* species. It is worth noting that no polyketide synthase genes have been detected in model *Phytophthora* genomes, and in general oomycetes rely more on toxic effector proteins than toxic small-molecule secondary metabolites for necrotrophic growth (Tyler *et al.*, 2006; Soanes, Richards and Talbot, 2007). The presence of this putative SnoaL-like protein in multiple copies in most of the *Pythium* species we investigated, suggests an additional method of phytopathogenic infection which may be novel to *Pythium*, or which may have been subsequently lost in *Phytophthora*.

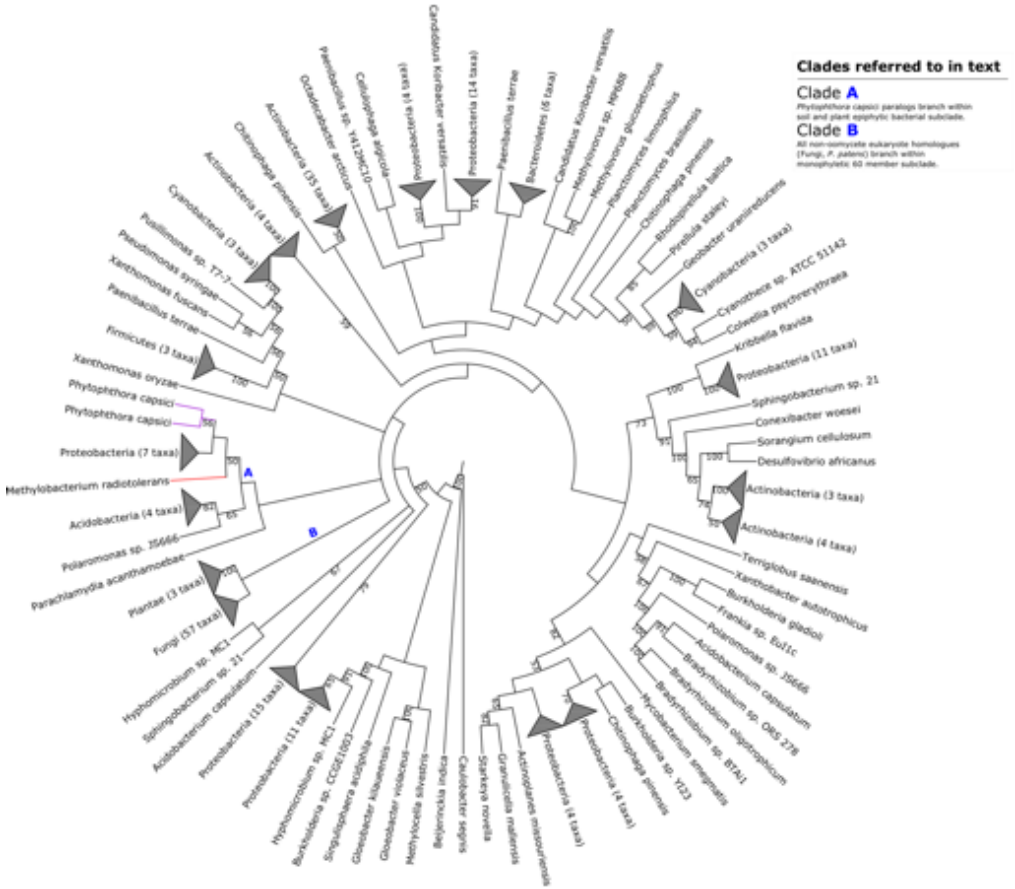


Figure 2.4. Maximum likelihood phylogeny illustrating putative transfer of epoxide hydrolase from *Methylobacterium radiotolerans* into *Phytophthora capsici*. Clades A & B referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. See **Figure S2.4a** in **Supplementary Material** for full phylogenetic tree.

2.3.5 A putative hydrolase from xenobiotic-degrading rhizosphere Proteobacteria is present in *Phytophthora capsici*

A gene from *Phytophthora capsici* (**Table 2.4**) was identified in our BLASTp homology searches as a candidate for bacterial HGT. A maximum likelihood phylogeny was generated from 253 homologs using a WAG+G substitution model. 8 bacterial phyla are represented in our reconstruction, with the majority of homologs coming from either Proteobacterial or Actinobacterial species. 57 fungal homologs and 3 paralogs from *Physcomitrella patens* (earthmoss) form a monophyletic eukaryotic clade (**Figure 2.4**, **Clade B**). Our maximum-likelihood phylogenetic tree placed two homologs from *P. capsici* adjacent to a homolog from the α-Proteobacterium *Methylobacterium radiotolerans* within a bacterial clade containing Acidobacteria and a number of soil-borne or plant epiphytic Proteobacteria (**Figure 2.4**, **Clade A**). BLASTp analysis aligned the seed sequence with an ortholog from the nitrogen-fixing Proteobacterium *Azotobacter vinelandii*. As there is only one *Phytophthora* species represented in this phylogeny, we carefully examined the sequence of the contig to rule out a bacterial contamination artefact in the *P. capsici* genome. All flanking genes were *Phytophthora* in origin thereby giving us confidence that this is a *bona fide* HGT event (**Table S2.2**). Furthermore, the phylogeny generated after homology searches against the NCBI database place the seed sequence within a large Proteobacterial clade (not shown).

As the bootstrap support for many of the more derived branches and clades in our phylogeny including the bacterial clade containing *P. capsici* homologs were weak (<50%), we generated a median phylogenetic network of all splits in the set of individual bootstrap trees generated by PhyML in our reconstruction using a consensus network method in SplitsTree (Huson and Bryant, 2006). This consensus network (**Figure S2.5**) places the two *P. capsici* homologs at the base of the large monophyletic bacterial clade, clearly separate from the fungal and plant homologs. With this analysis, we were satisfied that the phylogeny represented a *bona fide* bacteria-oomycete HGT event.

BLAST homology searches of the seed sequence against the NCBI database indicated that the sequence was homologous to bacterial hydrolases. PFAM analysis found a large α/β hydrolase fold domain present in the sequence, and InterProScan analysis returned a number of α/β hydrolase family PANTHER signatures, as well as epoxide hydrolase PRINTS (Attwood *et al.*, 2012) signatures across the sequence (**Table S2.6**). Epoxide hydrolases (E.C. 3.3.2.3) catalyse the dihydroxylation of epoxide residues

to diols, and are one of a number of protein families that contain an α/β hydrolase fold (Ollis *et al.*, 1992). Bacterial epoxide hydrolases are capable of degradation of xenobiotic organic compounds (van der Werf, Overkamp and de Bont, 1998; van Loo *et al.*, 2006). The structurally related haloalkane dehalogenases (E.C. 3.8.1.5), which can hydrolyse toxic haloalkanes into their corresponding alcohol and organic halide components in the cytosol, are widespread in soil bacteria (Janssen, 2004). It is interesting to note that strains of *M. radiotolerans* isolated from *Cucurbita pepo* roots, which is also a target for *P. capsici*, are capable of degrading xenobiotic 1,1-*bis*-(4-chlorophenyl)-2,2-dichloroethene or DDE (Eevers *et al.*, 2015). DDE is a highly toxic and highly recalcitrant major metabolite of the degradation of the toxic organochloride pesticide 1,1,1-trichloro-2,2-bis(p-chlorophenyl)ethane, or DDT, which saw widespread use for most of the 20th century (Thomas, Ou and Al-Agely, 2008).

Our maximum-likelihood phylogenetic reconstruction suggests that this putative hydrolase gene, which has two copies in *P. capsici*, has arisen through horizontal transfer from soil-dwelling bacteria, potentially from *M. radiotolerans* (**Figure 2.4**). Homology and functional analysis of the seed HGT gene indicates that these two paralogs contain hydrolase folds. The two paralogs in *P. capsici* are somewhat dissimilar at the nucleotide level; one appears to contain both peptidase and α/β hydrolase domains and is far more exonic than the seed HGT gene (**Table S2.3**). This putative transferred gene may have a potential cytosolic role in the degradation of toxic xenobiotic compounds in *P. capsici*. To date, descriptions of xenobiotic degradation or resistance in oomycetes are sparse in the literature; what is known is that few oomycete cytochrome P450 proteins (CYPs) appear to be involved in xenobiotic degradation compared with fungal CYPs (Moktali *et al.*, 2012; Sello *et al.*, 2015), and that *Phytophthora infestans* has far a lower proportion of major facilitator superfamily (MFS) transport proteins involved in efflux than many fungal type species do (Barabote *et al.*, 2011). As such, this acquisition may be a novel event in the context of plant parasitic oomycete genome evolution.

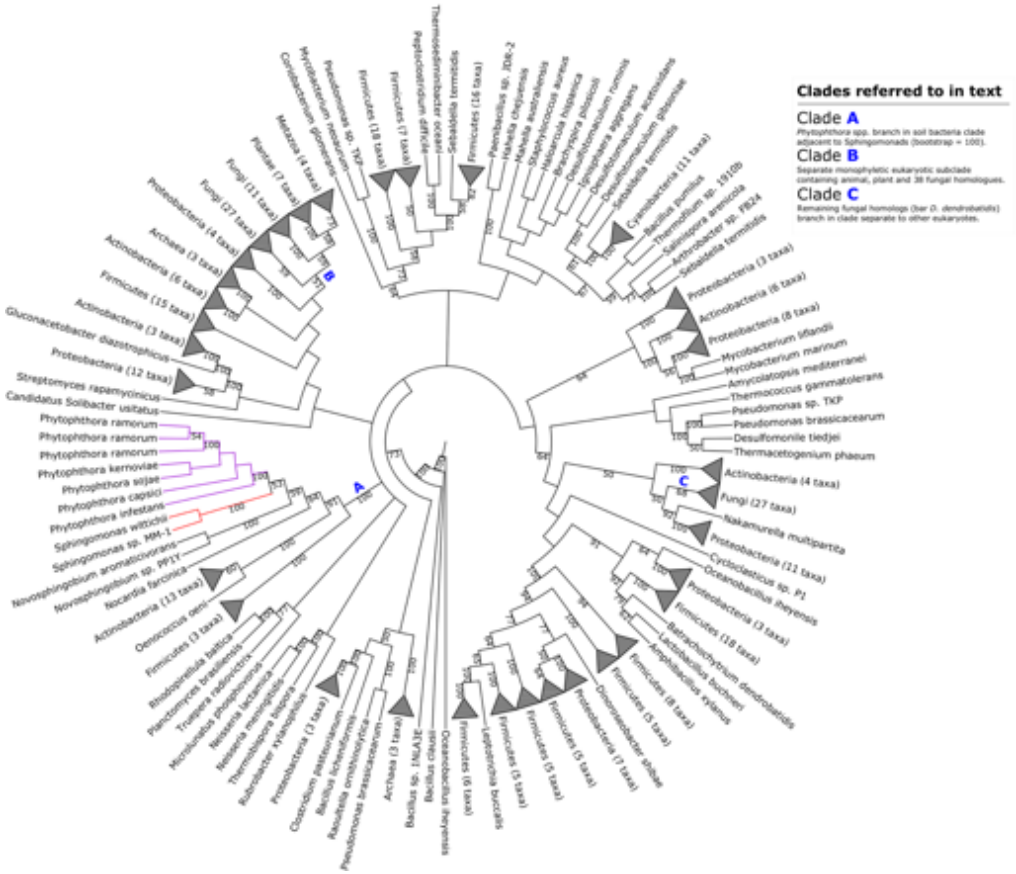


Figure 2.5. Maximum likelihood phylogeny illustrating putative transfer of alcohol dehydrogenase from Sphinomondales into *Phytophthora* spp. Clades A, B & C referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. See **Figure S2.4b** in **Supplementary Material** for full phylogenetic tree.

2.3.6 Sphingomonadale alcohol dehydrogenase is present in five *Phytophthora* species

A second *P. capsici* gene (**Table 2.4**) was identified in our BLASTp homology searches as a candidate for inter-domain HGT. Our phylogenetic reconstruction used 358 homologs with a LG+I+G substitution model (**Figure 2.5**). 9 bacterial phyla are represented in this reconstruction, the majority of which are homologs from Firmicutes species, 23% (84 of 358) of the homologs are of eukaryotic origin. Animal, plant and 38 fungal homologs form a eukaryote monophyletic clade (**Figure 2.5**, **Clade B**). 27 of the remaining 28 fungal homologs from a separate subclade (**Figure 2.5**, **Clade C**) almost entirely comprised of homologs from ascomycotes except for two paralogs from the Basidiomycota species *Phlebiopsis gigantea*, while *Batrachochytrium dendrobatidis* places within an adjacent Firmicutes subclade.

Our maximum-likelihood phylogeny inferred a monophyletic *Phytophthora* subclade with seven homologs from five species (excluding *P. lateralis* and *P. parasitica*) within a α-Proteobacterial Sphingomonadale subclade with 100% bootstrap support (**Figure 2.5**, **Clade A**). Homology data for the seed sequence and its adjacent sequences within the *P. capsici* genome from JGI showed no obvious evidence of bacterial contamination at the genomic level as neither of the flanking genes were bacterial in origin (**Table S2.2**).

BLAST homology searches of the seed sequence returned hits from many bacterial alcohol dehydrogenase proteins. PFAM and InterProScan analysis of the seed sequence found that it contained the hallmark signatures of a medium-chain Zn^{2+}-containing alcohol dehydrogenase; an N-terminus containing the conserved Zn^{2+} active site, the conserved GroES-like fold and the NAD(P)-binding Rossmann fold (**Table S2.7**). Alcohol dehydrogenases (EC 1.1.1.1) catalyse the NAD(P)-dependent reversible oxidation of alcohols to aldehydes or ketones. In most prokaryotes, fungi and plants, alcohol dehydrogenase is responsible for the reversed regeneration of NAD^+ in fermentation for glycolysis from the reduction of NADH and acetaldehyde to NAD^+ and ethanol. The high concentration of Firmicutes and fungal homologs in our reconstruction underlies the enzyme's important role in anaerobic Clostridia and fungi. Previous EST analysis of *P. sojae* infection of soybean found abundant matches for alcohol dehydrogenase amongst other intermediary metabolic genes differently expressed in host

tissue, suggesting alcohol fermentation is an important part of the catabolism of *P. sojae* in the early stages of growth inside host tissue (Qutob *et al.*, 2000).

The maximum-likelihood phylogenetic reconstruction for these putative *Phytophthora* alcohol dehydrogenase proteins supports a putative transfer from the α-Proteobacterial Sphingomonadales (**Figure 2.5**). Similarly, the phylogeny generated querying the seed sequence against the NCBI's non-redundant protein database placed the seed sequence within a small *Phytophthora* subclade that was found within a larger *Sphingobium* and *Novosphingobium* clade (not shown). We therefore propose that this alcohol dehydrogenase, found in a number of *Phytophthora* species has arisen in these species *via* recent transfer of the gene from Sphingomonadales.

2.3.7 Impact and extent of bacterial genes in oomycete evolution

Using stringent criteria, our analysis has found five putative gene families in oomycete species that have been acquired through horizontal transfer from bacteria. All five transfer events involve genes coding for proteins with putative enzymatic functions in their respective species; some of our findings complement those of other analyses of HGT in oomycete genomes, particularly the fumarase and alcohol dehydrogenase families. Many of the inter- and intra-domain HGT gene families identified in oomycete genomes to date are proteins with putative carbohydrate metabolism function; in the most extensive study of HGT into oomycete genomes to date, Richards *et al.* (2011) found 13 secreted proteins out of the 34 putative fungal HGT events in oomycetes that could be assigned with such function. Of the seven bacterial HGT events identified in oomycete species prior to our analysis, most were found in analyses of Saprolegniales species, and where function could be assigned were thought to be involved in carbohydrate metabolism also.

The bacterial-derived enzymes identified in oomycete species could have potentially found themselves more amenable to transfer and subsequent retention in oomycete genomes due to their relative low connectivity within a protein-protein interaction network, a significant factor in the influence of the "complexity hypothesis" on HGT (Jain, Rivera and Lake, 1999; Cohen, Gophna and Pupko, 2011). The relatively low number of bacterial-oomycete HGT events identified in this study and elsewhere in the literature, in comparison with other such studies of inter-domain HGT in fungi for example, may be partially explained by the paucity of oomycete genomic data overall and

lack of data for more basal lineages in particular (Beakes, Glockling and Sekimoto, 2012). Furthermore, our analysis was designed specifically to identify recent HGT events into individual plant parasitic oomycete lineages, as opposed to ancient transfers into the class as whole or even the greater stramenopiles group. Future analyses, facilitated by a greater amount of oomycete genomic data, may identify more instances of bacteria-oomycete HGT, either into specific lineages or ancient transfers into the class.

2.4 Conclusions

Using methods similar to those that have previously identified intra-domain HGT between fungi and *Phytophthora* (Richards *et al.*, 2011), we have identified five inter-domain HGT events between bacteria and plant pathogenic oomycetes. Of the five putative bacteria-oomycete HGT genes we have identified, one has signal peptide signatures and subcellular localization matches that indicate it is part of the oomycete secretome. The putative SnoaL-like protein may be a secreted transport protein or involved in production of other components of the *Pythium* secretome. A class II fumarase distinct from the endosymbiosis-derived fumarase is present in *Pythium* and *Phytopythium*, and a proteobacterial alcohol dehydrogenase gene is present in multiple *Phytophthora* species. The remaining two transferred genes may have more regulatory cytosolic roles in their respective oomycetes species, such as regulation of redox activity and neutralization of toxic xenobiotics. Our analysis shows that the transfer of genetic material from bacteria into oomycete lineages is rare, but has occurred, and is another example of cases of HGT between prokaryotes and eukaryotes.

Chapter 3 – Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes

This chapter was previously published in *mSphere* in April 2017.

McCarthy C. G. P. & Fitzpatrick D. A. (2017). Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *mSphere*, 2(2):e00095-17. DOI:10.1128/mSphere.00095-17

Chapter outline

The oomycetes are a class of microscopic, filamentous eukaryotes within the Stramenopiles-Alveolate-Rhizaria (SAR) supergroup which includes ecologically significant animal and plant pathogens, most infamously the causative agent of potato blight *Phytophthora infestans*. Single-gene and concatenated phylogenetic studies of both individual oomycete genera and the larger class have drawn conflicting conclusions for species phylogenies within the oomycetes, particularly for the large *Phytophthora* genus. Genome-scale phylogenetic studies have successfully resolved many eukaryotic relationships by using supertree methods, which combine large numbers of potentially disparate trees to determine evolutionary relationships that cannot be inferred from individual phylogenies alone. With a sufficient amount of genomic data now available, we have undertaken the first whole-genome phylogenetic analysis of the oomycetes using data from 37 oomycete species and six SAR species. In our analysis, we used established supertree methods to generate phylogenies from 8,355 homologous oomycete and SAR gene families, and have complemented those analyses with both phylogenomic network and concatenated supermatrix analyses. Our results show that a genome-scale approach to oomycete phylogeny resolves oomycete classes and individual Clades within the problematic *Phytophthora* genus. The resolution of the inferred relationships between individual *Phytophthora* Clades varies in support depending on the methodology used. Our analysis represents an important first step in large-scale phylogenomic analysis of the oomycetes.

3.1 Introduction

3.1.1 Evolutionary history of the oomycetes

The oomycetes are a class of microscopic eukaryotes which include some of the most ecologically destructive marine and terrestrial eukaryotic species (Beakes, Glockling and Sekimoto, 2012). Oomycete species display very similar filamentous morphology and ecological roles to fungi, and were historically regarded as a basal fungal lineage (Lévesque, 2011). As morphological and molecular studies have improved since the latter half of the 20th century to present, the oomycetes have come to be understood as very distant relations of "true" fungi which have evolved similar morphology and lifestyles through convergent evolution and horizontal gene transfer (HGT) (Richards *et al.*, 2006, 2011; Lévesque, 2011; Savory, Leonard and Richards, 2015). Present phylogenomic studies place the oomycetes in the diverse stramenopiles lineage within the Stramenopiles-Alveolata-Rhizaria (SAR) eukaryotic supergroup (Cavalier-Smith and Chao, 2006; Riisberg *et al.*, 2009; Tsui *et al.*, 2009; Judelson, 2012; Burki, 2014) (**Figure 3.1**). The stramenopiles were previously placed within Chromista (Cavalier-Smith, 1981) and then the "chromalveolates" supergroup (Chromista + Alveolata) (Cavalier-Smith, 1999; Keeling, 2009). While early phylogenetic analyses supported this "chromalveolates" hypothesis (Yoon *et al.*, 2002; Bachvaroff, Sanchez Puerta and Delwiche, 2005), later phylogenetic and HGT analyses have consistently failed to support a monophyletic chromalevolate grouping (Keeling, 2001; Harper, Waanders and Keeling, 2005; Rice and Palmer, 2006; Hackett *et al.*, 2007; Janouskovec *et al.*, 2010; Gaston and Roger, 2013). In contrast, molecular evidence for the monophyly of the current SAR supergroup has been demonstrated in multiple phylogenetic analyses (Baldauf *et al.*, 2000; Burki *et al.*, 2007; Hackett *et al.*, 2007; Moreira *et al.*, 2007; Shalchian-Tabrizi *et al.*, 2007; Hampl *et al.*, 2009; Gaston and Roger, 2013).

The oomycetes are thought to have diverged from diatoms between the late Proterozoic and the mid-Paleozoic eras (~0.4-0.6 bya) (Dick, 2001; Matari and Blair, 2014), and are present as early as the Devonian period (~400 mya) in the fossil record (Taylor, Krings and Kerp, 2006). Though many described species are phytopathogens, oomycete phytopathogenicity is thought to be a derived trait which has evolved independently in many lineages (Thines and Kamoun, 2010). Many species are still yet un-sampled and the class phylogeny of the oomycetes is still subject to revision; but with current data the oomycetes can be split into the earliest diverging clades and the later

"crown" taxa (Hakariya, Hirose and Tokumasu, 2007; Sekimoto *et al.*, 2008; Beakes *et al.*, 2014) (**Figure 3.1**). With the exception of some species infecting terrestrial nematodes (Hakariya, Hirose and Tokumasu, 2007), the earliest diverging oomycete clades are otherwise exclusively marine in habitat (Beakes, Glockling and Sekimoto, 2012). The remaining "crown" oomycetes can be subdivided into the predominantly marine and freshwater "saprolegnian" branches and the predominantly terrestrial "peronosporalean" branches, which diverged in the early Mesozoic era (Riethmüller *et al.*, 2002; Beakes, Glockling and Sekimoto, 2012; Jiang and Tyler, 2012; Matari and Blair, 2014; Thines, 2014). The "saprolegnian" branches include the fish pathogens *Saprolegnia*, also known as "cotton moulds" (Hulvey, Padgett and Bailey, 2007), and the animal and plant pathogenic *Aphanomyces* genus (Kamoun, 2003; Jiang and Tyler, 2012). The "peronosporalean" branches include the best characterized oomycete taxa, *Phytophthora* and *Pythium*, and the more basal Albuginales order (Beakes, Glockling and Sekimoto, 2012; Thines, 2014). The majority of "peronosporalean" oomycetes are phytopathogens, although *Pythium* includes species capable of infecting animals or acting as mycoparasitic biocontrol agents (Gaastra *et al.*, 2010; Benhamou *et al.*, 2012) (**Figure 3.1**).

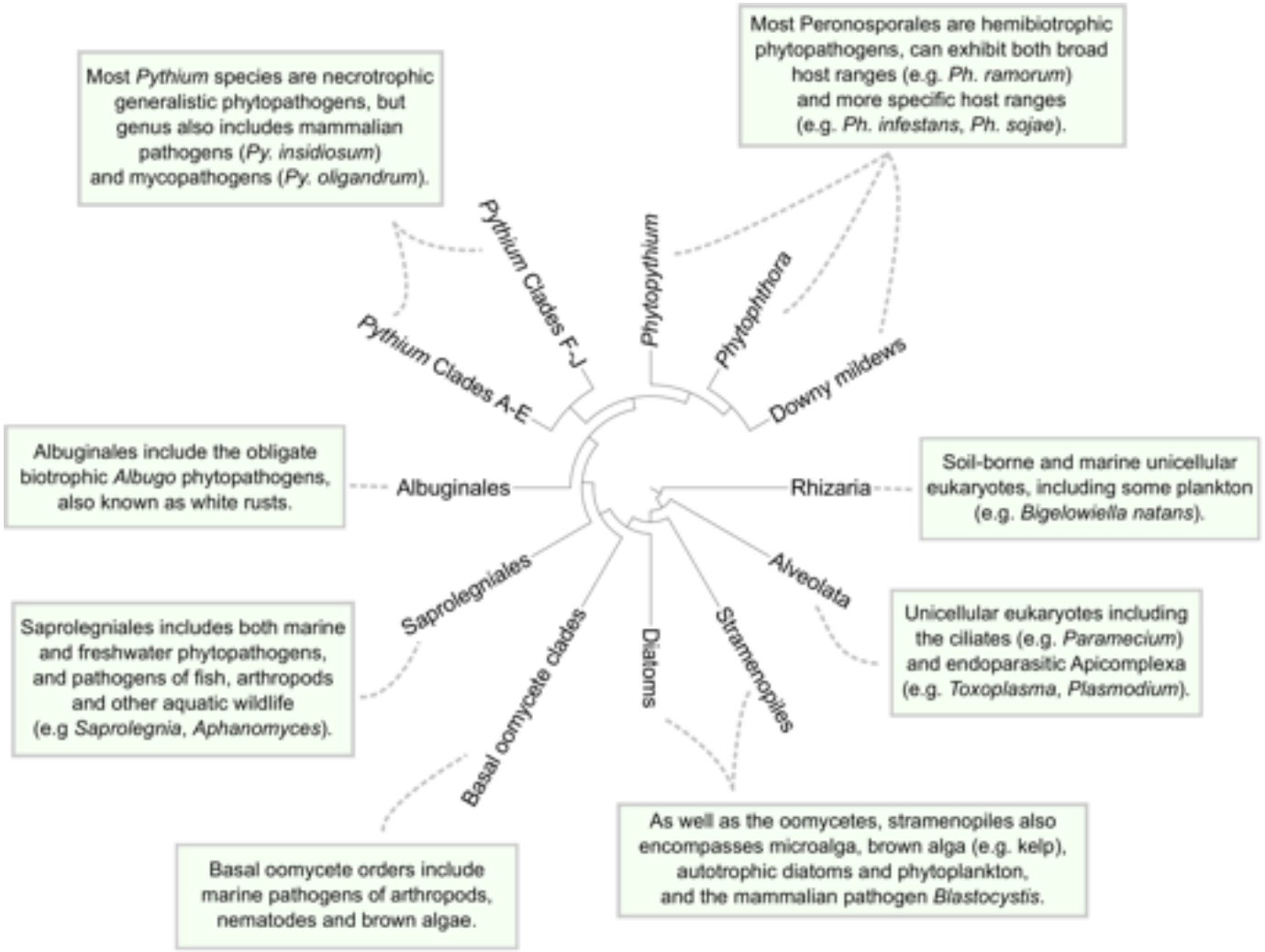


Figure 3.1. Consensus phylogeny of the oomycetes class within the SAR superkingdom, with information for various groups. Cladogram adapated from Judelson (2012).

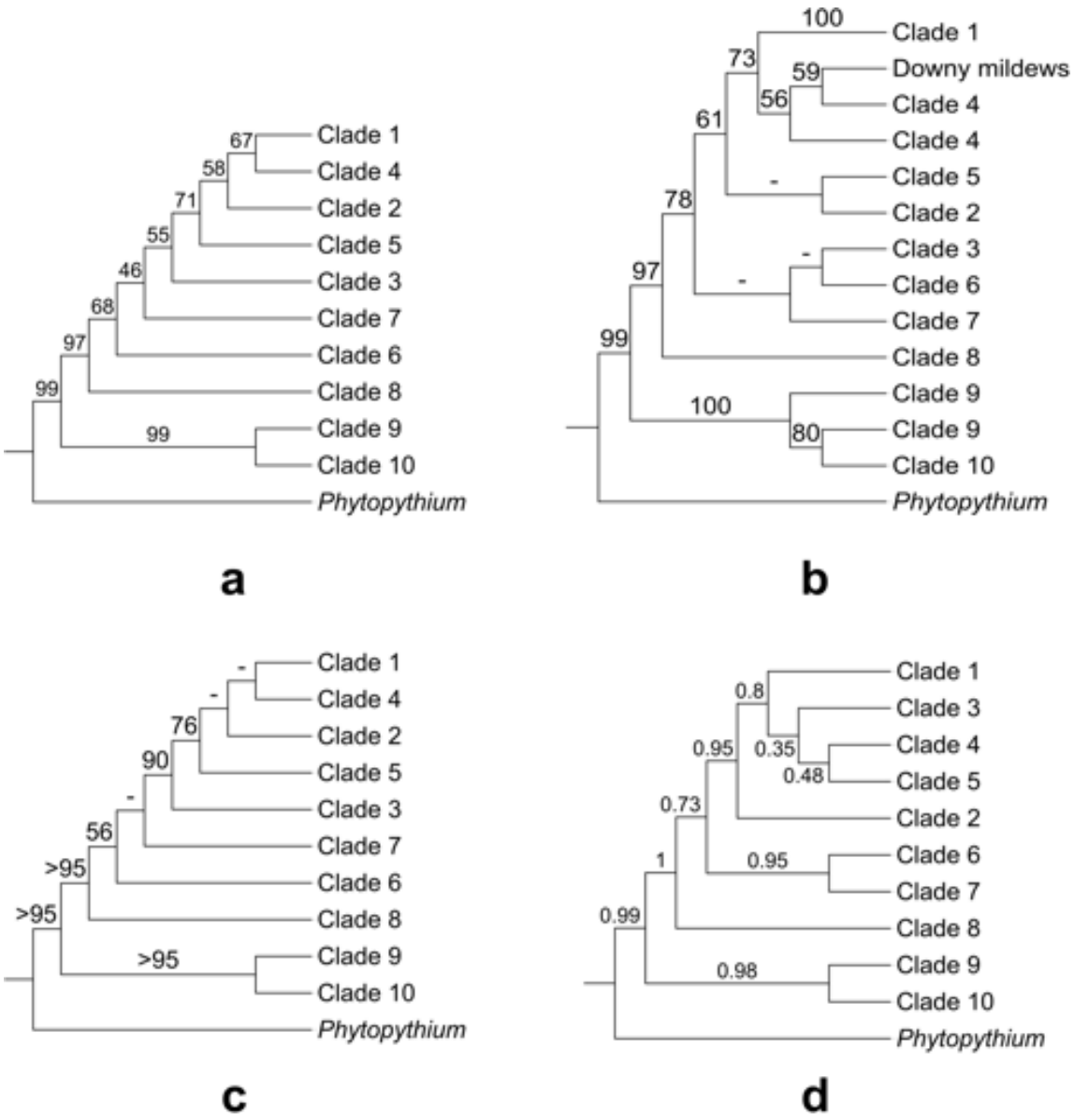


Figure 3.2. Congruence of the Peronosporales order among recent phylogenetic analyses. **(a)** Seven-locus maximum likelihood 7 loci ML/MP/Bayesian phylogeny of *Phytophthora* by Blair *et al*. (2008), **(b)** ME/ML/Bayesian phylogeny of *Phytophthora* and downy mildews by Runge *et al*. (2012), **(c)** 11-locus ML/MP/Bayesian phylogeny of *Phytophthora* by Martin et al., **(d)** Six-locus coalescent phylogeny of *Phytophthora* by Martin et al. Support values, where given, represent maximum-likelihood bootstrap supports except for **Figure 3.2d)**, where Bayseian posterior probabilities are given instead.

3.1.2 Taxonomy of *Phytophthora*, *Pythium* and other oomycete taxa

Phytophthora is the largest genus (>120 described species) within the order Peronosporales and is divided into 10 phylogenetic clades on the basis of initial ITS analysis and later combined nuclear and mitochondrial analysis (Cooke *et al.*, 2000; Blair *et al.*, 2008) (**Figure 3.2a**). The largest clades (clades 1, 2, 7 and 8) are further divided into subclades, while the smallest clades (clades 5, 10) contain fewer than 5 described species at present (Kroon *et al.*, 2004; Weir *et al.*, 2015). Initial ITS phylogeny by Cooke et al. (Cooke *et al.*, 2000) suggested that *Phytophthora* was paraphyletic with respect to the basal clades 9 and 10, however later multi-gene and combined nuclear and mitochondrial studies have placed these clades within *Phytophthora* (Martin and Tooley, 2003; Kroon *et al.*, 2004; Blair *et al.*, 2008). Generally, species within *Phytophthora* clades do not share consistent morphological features or reproductive strategies, although clades 6 to 8 form a distinct branch of terrestrial species with predominantly non-papillate sporangia within the genus tree (Kroon *et al.*, 2004). While many recent phylogenetic analyses have supported the current designation by Blair et al. (Blair *et al.*, 2008) of 10 distinct phylogenetic clades within *Phytophthora*, many of the same analyses draw conflicting conclusions as to the relationships between these clades. In their analysis, Blair *et al*. (Blair *et al.*, 2008) found strong support under maximum-likelihood, maximum-parsimony and Bayesian methods for the 10 phylogenetic clades using data from seven highly-conserved nuclear loci (including markers from 28S rDNA, Hsp90 and β-tubulin) from 82 *Phytophthora* species (**Figure 3.2a**). The relationship between the clades in Blair et al. (Blair *et al.*, 2008) was mostly upheld in a follow-up analysis by Runge et al. (Runge *et al.*, 2011) which included homologous data from an additional 39 *Phytophthora* species and other Peronosporales. One noticeable difference was that their analysis placed clades 3, 6 and 7 as sister clades within a monophyletic clade with strong support under minimum evolution, maximum-likelihood and Bayesian methods, while the clades were more distantly related in Blair et al. (Blair *et al.*, 2008) (**Figures 3.2a & 3.2b**). The addition of four mitochondrial markers (*cox2*, *nad9*, *rps10*, *secY*) in a later 11-loci analysis by Martin et al. (Martin, Blair and Coffey, 2014), while topologically supporting Blair et al. (Blair *et al.*, 2008) displayed poor resolution for many inter-clade relationships (particularly for more derived clades such as Clades 1-5) within *Phytophthora* under maximum-likelihood, maximum parsimony and Bayesian methods (**Figure 3.2c**). A coalescent approach on a similar dataset by the same authors showed

improved Bayesian support between some *Phytophthora* clades (e.g. Clades 1-5), but weaker supports for other clades and a conflicting topology from the 11-loci analysis (Martin, Blair and Coffey, 2014) (**Figure 3.2d**).

Placement of other taxa within the Peronosporales order, namely the "downy mildews", and the phylogeny of *Pythium* and the Pythiales order has also been difficult to resolve. The inclusion of two downy mildews species (*Hyaloperonospora arabidopsidis* and *Pseudoperonospora cubensis*) in an analysis conducted by Runge et al. placed the two species within *Phytophthora* Clade 4 and sister to Clade 1 species such as *Phytophthora infestans*, implying a paraphyletic *Phytophthora* genus (Runge *et al.*, 2011) (**Figure 3.2b**). However, a subsequent tree reconciliation analysis, inferred on a class phylogeny of 189 oomycete clusters of orthologous groups (COGs) placed *H. arabidopsidis* as sister to the *Phytophthora* genus (Seidl *et al.*, 2012). Another downy mildew species, *Plasmopara halstedii*, is placed sister to *Phytophthora* Clade 1 when included in similar phylogenetic analyses (Riethmüller *et al.*, 2002; Robideau, Rodrigue and André Lévesque, 2014). *Phytopythium*, a morphological intermediate between *Phytophthora* and *Pythium*, was reclassified from *Pythium* clade K to its own genus within the Peronosporales order based on recent multi-gene phylogenetic analysis which placed the genus sister to *Phytophthora* (de Cock *et al.*, 2015). *Pythium* itself is divided into 10 clades, labelled A to J, which were initially circumscribed with ITS data and consistent with mitochondrial data (Lévesque and de Cock, 2004). The main morphological difference between clades within *Pythium* is the development of the filamentous sporangium in species within clades A-C from the ancestral globose sporangium observed in the basal clades and *Phytopythium* (Lévesque and de Cock, 2004; Villa *et al.*, 2006), with an intermediate contiguous sporangium developing in species within clade D (Lévesque and de Cock, 2004), and an elongated sporangium in species within clade H (Uzuhashi, Tojo and Kakishima, 2010). Otherwise, as in *Phytophthora*, phylogenetic clades generally do not correlate with distinct morphological characters in *Pythium* (Lévesque and de Cock, 2004). A number of phylogenetic analyses suggest that *Pythium* is polyphyletic (Riethmüller *et al.*, 2002; Rahman *et al.*, 2003; Villa *et al.*, 2006; Uzuhashi, Tojo and Kakishima, 2010; Hyde *et al.*, 2014; Robideau, Rodrigue and André Lévesque, 2014), and there has been recent suggestion that it be amended entirely into at least five new genera (Uzuhashi, Tojo and Kakishima, 2010; Huang *et al.*, 2013).

3.1.3 Phylogenetic and phylogenomic reconstructions of the oomycetes

Many of the aforementioned phylogenetic analyses of the oomycetes are based upon a small number of highly-conserved nuclear and/or mitochondrial markers, either through consensus analysis or concatenated analysis. The selection of such markers, while usually robust, may unintentionally ignore other types of potential phylogenetic markers that may resolve conflicting analyses, such as lineages which include gene duplication events (Hackett *et al.*, 2007). One solution to the possible limitations of single gene or small-scale gene phylogenies is to assemble a consensus phylogeny for a given set of taxa using many source single gene phylogenies through supertree analysis, which enables the inclusion of phylogenies with missing or duplicated taxa (Bininda-Emonds, 2004). Matrix Representation using Parsimony (MRP), in which character matrices are generated for each source phylogeny and merged into a single binary character matrix for maximum-parsimony alignment (Baum, 1992; Ragan, 1992), is one of the most commonly-used supertree methods and has seen successful application in a number of eukaryotic phylogenomic studies (Beck *et al.*, 2006; Fitzpatrick *et al.*, 2006; Pisani, Cotton and McInerney, 2007). Other methods have been developed for inferring species phylogeny from paralogous gene phylogenies, the most successful of which has been Gene Tree Parsimony (GTP) (Cotton and Page, 2003). GTP attempts to find the most parsimonious species tree from a set of source phylogenies with the fewest number of events required to explain incongruences (i.e. gene duplication events) between the source phylogenies, and has seen application in large-scale phylogenetic analysis (Casewell *et al.*, 2011). Another method of large-scale phylogenetic analysis is the supermatrix approach of concatenating multiple character datasets for simultaneous analysis (de Queiroz and Gatesy, 2007).

Since the publication of the genome sequences of *Phytophthora sojae* and *Phytophthora ramorum* in 2006 (Tyler *et al.*, 2006), the quantity of oomycete genomic data has steadily increased; currently 37 oomycete species now have publicly-available genomic data at the assembly level or higher (**Table 3.1**). With this in mind we have conducted the first whole-genome phylogenetic analysis for the oomycetes as a class, using a variety of supertree and supermatrix approaches which have previously been used in fungal whole-genome phylogenetic analysis (Fitzpatrick *et al.*, 2006). In our analysis we utilized protein data from 37 complete oomycete genomes and 6 complete SAR genomes. This represents all extant genomic data from the four “crown” oomycete orders,

and covers 8 of the 10 phylogenetic clades within *Phytophthora* and 7 of the 10 phylogenetic clades within *Pythium* (**Table 3.1**). Our whole-genome phylogenetic analysis of the oomycetes supports the four oomycete orders, the placement of *Phytopythium* within the Peronosporales, and individual clades within *Phytophthora* and *Pythium*. The resolution of the Peronosporales as an order varied under different methods, probably due to missing data from clades 4 and 9 within *Phytophthora*, however the overall order phylogeny is relatively congruent between our different methods. This analysis will provide a useful backbone to future genome phylogenies of the oomycetes utilizing more taxonomically extensive datasets.

Table 3.1. Taxonomic and genomic information for the 43 oomycete and SAR species in this analysis. Protein counts generated in this study from assembly data highlighted with an asterisk (*). References are to the genome publications where possible, or NCBI BioProject identifier or sequencing organization(s) otherwise.

Species Name	Clade	Order	Class	Reference	Genes
Albugo candida	n/a	Albuginales	Oomycota	Links *et al.* (2011)	13310
Albugo labiachii	n/a	Albuginales	Oomycota	Kemen *et al.* (2011)	13804
Hyaloperonospora arabidopsidis	n/a	Peronosporales	Oomycota	Baxter *et al.* (2010)	14321
Phytophthora agathidicida	Clade 5	Peronosporales	Oomycota	Studholme *et al.* (2016)	14110 *
Phytophthora capsici	Clade 2	Peronosporales	Oomycota	Lamour *et al.* (2012)	19805
Phytophthora cinnamomi	Clade 7	Peronosporales	Oomycota	Studholme *et al.* (2016)	12942 *
Phytophthora cryptogea	Clade 8	Peronosporales	Oomycota	Feau *et al.* (2016)	11876 *
Phytophthora fragariae	Clade 7	Peronosporales	Oomycota	Gao *et al.* (2015)	13361 *
Phytophthora infestans	Clade 1	Peronosporales	Oomycota	Hass *et al.* (2009)	17797
Phytophthora kernoviae	Clade 10	Peronosporales	Oomycota	Sambles *et al.* (2015)	10650
Phytophthora lateralis	Clade 8	Peronosporales	Oomycota	Quinn *et al*. (2013)	11635
Phytophthora multivora	Clade 2	Peronosporales	Oomycota	Studholme *et al.* (2016)	15006 *
Phytophthora nicotianae	Clade 1	Peronosporales	Oomycota	Liu *et al.* (2016)	10521
Phytophthora parasitica	Clade 1	Peronosporales	Oomycota	Broad Institute	27942
Phytophthora pinifolia	Clade 6	Peronosporales	Oomycota	Feau *et al.* (2016)	19533 *
Phytophthora pluvialis	Clade 3	Peronosporales	Oomycota	Studholme *et al.* (2016)	18426 *
Phytophthora pisi	Clade 7	Peronosporales	Oomycota	PRJEB6298	15495 *
Phytophthora ramorum	Clade 8	Peronosporales	Oomycota	Tyler *et al.* (2006)	15743
Phytophthora rubi	Clade 7	Peronosporales	Oomycota	PRJNA244739	15462 *
Phytophthora sojae	Clade 7	Peronosporales	Oomycota	Tyler *et al.* (2006)	26584
Phytophthora taxon totara	Clade 3	Peronosporales	Oomycota	Studholme *et al.* (2016)	16691 *
Plasmopara halstedii	n/a	Peronosporales	Oomycota	Sharma *et al.* (2015)	15469
Plasmopara viticola	n/a	Peronosporales	Oomycota	PRJNA329579	12048 *
Phytopythium vexans	n/a	Peronosporales	Oomycota	Adhikari *et al.* (2013)	11958

Species Name	Clade	Order	Class	Reference	Genes
Pilasporangium apinafurcum	n/a	Pythiales	Oomycota	PRJDB3797	13184*
Pythium aphanidermatum	Clade A	Pythiales	Oomycota	Adhikari *et al.* (2013)	12312
Pythium arrhenomanes	Clade B	Pythiales	Oomycota	Adhikari *et al.* (2013)	13805
Pythium insidiosum	Clade C	Pythiales	Oomycota	Rujirawat et al. (2015)	19290*
Pythium irregulare	Clade F	Pythiales	Oomycota	Adhikari *et al.* (2013)	13805
Pythium iwayami	Clade G	Pythiales	Oomycota	Adhikari *et al.* (2013)	14875
Pythium oligandrum	Clade D	Pythiales	Oomycota	Berger *et al.* (2016)	14292*
Pythium ultimum var. *sporangiiferum*	Clade I	Pythiales	Oomycota	Adhikari *et al.* (2013)	14096
Pythium ultimum var. *ultimum*	Clade I	Pythiales	Oomycota	Lévesque *et al.* (2010)	15323
Aphanomyces astaci	n/a	Saprolegniales	Oomycota	Broad Institute	26259
Aphanomyces invadans	n/a	Saprolegniales	Oomycota	Broad Institute	20816
Saprolegnia diclina	n/a	Saprolegniales	Oomycota	Jiang *et al.* (2011)	18229
Saprolegnia parasitica	n/a	Saprolegniales	Oomycota	Broad Institute	20121
Aureococcus anophagefferns	n/a	Pelagomonadales	Pelagophyceae	Gobler *et al.* (2011)	11501
Ectocarpus siliculosus	n/a	Ectocarpales	Phaeophyceae	Cock *et al.* (2012)	16269
Phaeodactylum tricornutum	n/a	Naviculales	Bacillariophyceae	Bowler *et al.* (2008)	10402
Thalassiosira psuedonana	n/a	Thalassiosirales	Coscinodiscophyceae	Armbrust *et al.* (2004)	11776
Paramecium tetraurelia	n/a	Peniculida	Oligohymenophorea	Aury *et al.* (2006)	39580
Bigelowiella natans	n/a	Chlorarachniophyceae	Cercozoa	Curtis *et al.* (2012)	21708

3.2 Materials and Methods

3.2.1 Dataset assembly

The predicted proteomes for 29 SAR species (23 oomycete species, four other stramenopile species, the alveolate species *Paramecium tetraurelia* and the rhizarian species *Bigelowiella natans*) were obtained from public databases (**Table 3.1**). Predicted proteomes for a further 14 oomycete species (10 *Phythophthora* species, two *Pythium* species, *Plasmopara viticola* and *Pilasporangium apinafurcum*) were generated from publicly-available assembly data using AUGUSTUS (Stanke *et al.*, 2004). Templates for *ab initio* protein prediction with AUGUSTUS were generated from assembly and EST data from a number of reference oomycete species (*Phytophthora sojae*, *Phytophthora capsici*, *Pythium ultimum* var. *ultimum* and *Plasmopara halstedii*) (**Table S3.1**). *Ph. sojae* was used as a reference for *Phytophthora* species from clades 1-5, while *Ph. sojae* was used as a reference for *Phytophthora* species from clades 6-10. *Py. ultimum* var. *ultimum* was used as a reference for two *Pythium* species and *Pi. apinafurcum*. *Pl. halstedii* was used as a reference for *Pl. viticola*. GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008) was used in conjunction with AUGUSTUS for protein prediction for *Pi. apinafurcum*. The taxonomy, assembly and prediction statistics for each of the 14 assemblies included in this study are summarized in **Table S3.1**. Our final dataset contained 702,132 protein sequences from 37 oomycete genomes and 6 SAR genomes (**Table 3.1**, **Table S3.1**).

3.2.2 Identification and reconstruction of gene phylogenies in oomycete and SAR genomes

All 702,132 protein sequences in our dataset were filtered and clustered into 56,638 orthologous gene families using OrthoMCL (Li, Stoeckert and Roos, 2003), with a BLASTp e-value cutoff of 10^{-20} (Ramsay *et al.*, 2000) and an inflation value of 1.5. Using bespoke Python scripting, we identified and retrieved two types of gene family containing 200 sequences or fewer from the 56,638 families within our dataset:

1) 2,853 single-copy gene families: single-copy orthologs present in ≥5 species,
2) 11,158 multi-copy gene families: ≥1 paralog(s) present in ≥5 species.

Each of these gene families was retrieved and aligned in MUSCLE (Edgar, 2004), and highly conserved regions of these alignments were sampled using Gblocks (Castresana, 2000) with the default parameters. 266 single-copy gene families and 4,928 multi-copy

gene families did not retain alignment data after Gblocks sampling and were discarded. Permutation-tail probability (PTP) tests (Faith and Cranston, 1991) were carried out for every remaining sampled gene family in PAUP* (Swofford, 2002) using 100 replicates, to determine whether a given sampled gene family had phylogenetic signal. Those sampled gene families whose PTP test result had a result of $p \leq 0.05$ were considered to have signal and retained. 2,280 single-copy sampled gene families (containing 35,622 genes in total) and 6,055 multi-copy sampled gene families (containing 174,282 genes in total) ultimately satisfied our filtering process. Best-fit amino acid replacement models were selected for every remaining sampled gene family using ProtTest (**Table S3.2**), and maximum-likelihood phylogenetic reconstruction was carried out using PhyML with 100 bootstrap replicates.

3.2.3 Supertree analyses of single-copy and paralogous gene phylogenies

Maximum-parsimony supertree analysis of 2,280 single-copy gene phylogenies (containing 35,622 genes in total) was carried out using CLANN, by performing a subtree prune and regraft (SPR) heuristic search with 100 bootstrap replicates (Creevey and McInerney, 2005). This phylogeny was visualized and annotated as a cladogram using the Interactive Tree of Life (iTOL) website (Letunic and Bork, 2007) (**Figure 3.3**). As an additional analysis, a consensus super-network of phylogenetic multifurcations within the 2,280 individual gene phylogenies was generated in SplitsTree (Huson and Bryant, 2006) (**Fig S3.1**). Gene tree parsimony (GTP) supertree analyses of all 8,335 gene phylogenies (containing 209,904 genes in total) was carried out using DupTree (Wehe *et al.*, 2008), using a rooted SPR heuristic search over 100 bootstrapped replicates of each phylogeny. A consensus phylogeny was generated from all individual replicates using Consense and was visualized and annotated as a cladogram using iTOL (**Figure 3.5**).

3.2.4 Identification and supermatrix analysis of ubiquituous oomycete gene phylogenies

A reciprocal BLASTp search was carried out with an e-value cutoff of 10^{-10} between all 37 oomycetes proteomes in our dataset (590,896 protein sequences in total) and 458 core orthologous genes (COGs) in *Saccharomyces cerevisiae* from the CEGMA dataset (Ramsay *et al.*, 2000; Parra, Bradnam and Korf, 2007). 443 oomycete gene families representing oomycete top hits to *S. cerevisiae* COGs were retrieved, of which

144 families contained an ortholog from all 37 oomycete species in our dataset. Each of these 144 families was aligned in MUSCLE, and sampled for highly conserved regions using Gblocks with the default parameters. After removing 13 families which failed to retain alignment data after Gblocks sampling, the remaining 131 sampled alignments (containing 4,847 genes in total) were concatenated into a superalignment of 16,934 aligned positions. This superalignment was bootstrapped 100 times using SeqBoot, and maximum-likelihood phylogenetic trees were generated for each individual replicate using PhyML, with a LG+I+G+F amino acid substitution model as selected by ProtTest. A consensus tree was generated from these replicate trees using Consense and the consensus tree was visualized and annotated as a cladogram using iTOL (**Figure S3.2**). A neighbour-joining network of phylogenetic splits in the original superalignment was generated in SplitsTree (**Figure S3.3**).

3.2.5 Identification and supermatrix analysis of ubiquitous Peronosporales gene phylogenies

347,375 protein sequences from the 22 Peronosporales proteomes in our dataset were filtered and clustered into 22,803 orthologous gene families using OrthoMCL, with a BLASTp e-value cutoff of 10^{-20} and an inflation value of 1.5. Using bespoke Python scripting we identified 352 ubiquitous Peronosporales gene families, which we defined as any family which had exactly one representative ortholog from all 22 Peronosporales species in our dataset. Each of these families was aligned in MUSCLE and sampled for highly conserved regions using Gblocks with the default parameters. After removing 39 gene families which did not retain alignment data after sampling, the remaining 313 sampled alignments (containing 6,886 genes in total) were concatenated into a single superalignment of 47,365 aligned positions. This superalignment was bootstrapped 100 times using SeqBoot, and maximum-likelihood phylogenetic trees were generated for each individual replicate using PhyML with a JTT+I+G+F amino acid substitution model, as selected by ProtTest. A consensus tree was generated from these replicate trees using Consense and the consensus tree was visualized and annotated as a cladogram using iTOL (**Figure 3.6**).

3.3 Results and Discussion

3.3.1 Identification of gene families

For our supertree analyses, we constructed a dataset containing 43 complete genomes, 37 from oomycete species and 6 from other species within the SAR supergroup (**Table S3.1**). Of these 37 oomycete genomes, 26 were from either *Phytophthora* or *Pythium* species representing the majority of clades within both genera, and the remainder were sampled from all 4 of the "crown" orders. We downloaded proteomes for 23 oomycete species which were available from public databases, and we generated corresponding proteomes for the remaining 14 species from publicly-available assembly data using bespoke oomycete reference templates with AUGUSTUS and GeneMark-ES (Stanke *et al.*, 2004; Ter-Hovhannisyan *et al.*, 2008) (**Table S3.1**). In total, our final dataset contained 702,132 protein sequences from 37 complete oomycete genomes and 6 complete SAR genomes (**Table 3.1**).

The initial step in determining the phylogeny of the 43 oomycete and SAR genomes in our dataset through supertree methods was to identify groups of closely related orthologs or paralogs within our dataset, which we termed gene families, and to use these groups to generate gene phylogenies to use as source data for our methods. To identify families of orthologous and paralogous genes in our dataset, we set the following criteria;

1) A single-copy gene family contained no more than 1 orthologous gene per species in 4 or more species,
2) A multi-copy gene family contained at least more than 1 orthologous gene (i.e. one or more paralogs) in at least 1 species in 4 or more species.

Using OrthoMCL (Li, Stoeckert and Roos, 2003), with an inflation value of 1.5 and a strict BLASTp cutoff value of 10^{-20} (Ramsay *et al.*, 2000), and bespoke Python scripting we identified over 56,000 orthologous oomycete and SAR gene families in our dataset. Of these, 2,853 families matched our criterion for single-copy families and 11,158 families matched our criterion for multi-copy families. By aligning each of these gene families in MUSCLE (Robert C. Edgar, 2004) and sampling for highly conserved regions using Gblocks (Castresana, 2000), both using the default parameters, and then carrying out a permutation-tail possibility (PTP) tests for every remaining sampled alignment using PAUP* (Faith and Cranston, 1991; Swofford, 2002), we were able to remove 576 single-copy gene families and 5,103 multi-copy gene families with poor phylogenetic

signal from our data. All remaining gene families had their evolutionary model estimated using ProtTest (Darriba *et al.*, 2011) (**Table S3.2**), and maximum-likelihood gene phylogenies were generated using PhyML with 100 bootstrap replicates (Guindon *et al.*, 2010). We generated phylogenetic reconstructions for 2,280 orthologous gene families (containing 35,622 genes) and 6,055 paralogous gene families (containing 174,282 genes). In total, from our 43 genome dataset we identified 8,335 individual gene phylogenies, containing 209,904 oomycete and SAR genes.

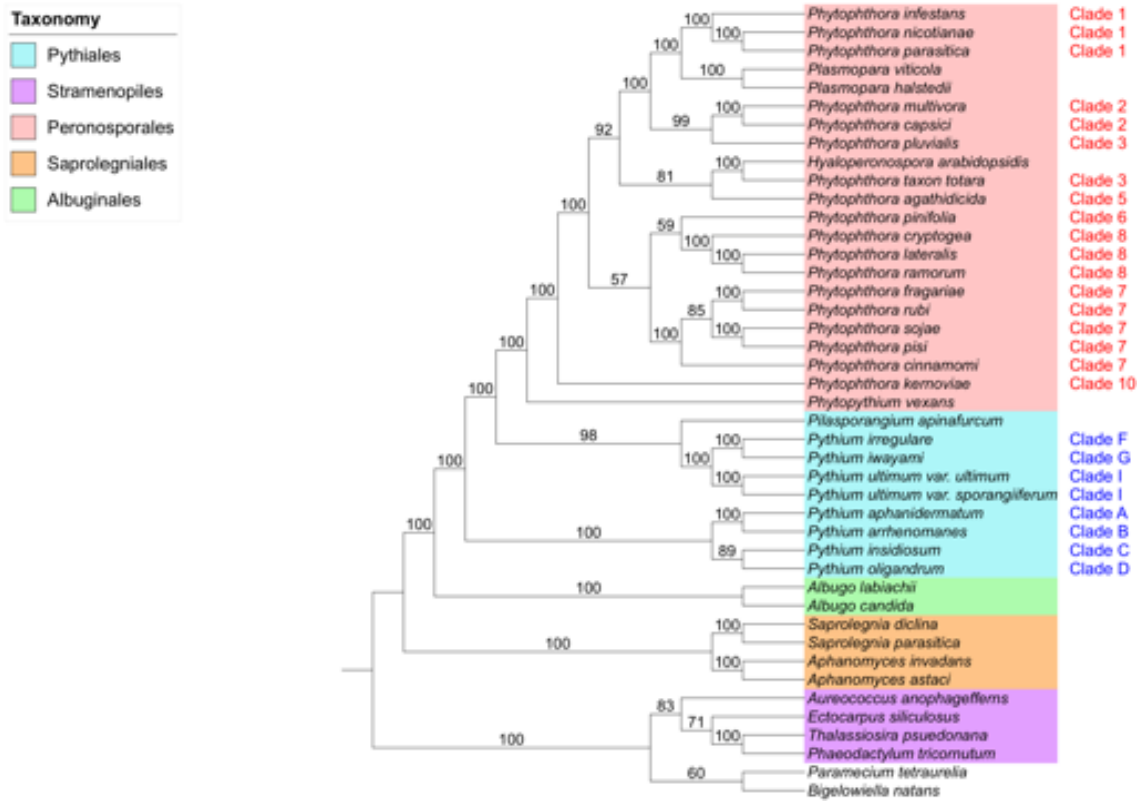


Figure 3.3. Matrix representation with parsimony (MRP) supertree of 37 oomycete species and 6 SAR species (2,280 source phylogenies). Supertree generated in CLANN. Phylogeny rooted at branch separating oomycetes and SAR. No colour: *P. tetraurelia* (Alveolata) and *B. natans* (Rhizaria).

3.3.2 Supertree phylogenies fully resolve oomycete class and order phylogenies

All 2,280 orthologous single-copy gene phylogenies (35,622 genes in total) were used as input for CLANN (Creevey and McInerney, 2005), which uses a Matrix

Representation using Parsimony (MRP) method to determine consensus phylogeny for many source phylogenies with overlapping taxa or missing taxa. An MRP supertree phylogeny was generated in CLANN using a heuristic search with 100 bootstrap replicates. The supertree was visualized and annotated within the Interactive Tree of Life (iTOL) website (Letunic and Bork, 2007) and rooted at the branch containing *Paramecium tetraurelia*, *Bigelowiella natans* and four Stramenopiles species (**Figure 3.3**).

MRP supertree analysis of 2,280 orthologous single-copy oomycete gene phylogenies supports the four "crown" oomycete orders; Saprolegniales, Albuginales, Pythiales and Peronosporales, with maximum bootstrap support (**Figure 3.3**). The MRP supertree reflects the consensus phylogeny of the oomycetes (Hakariya, Hirose and Tokumasu, 2007; Sekimoto *et al.*, 2008; Beakes *et al.*, 2014) (**Figure 3.1**). The Saprolegniales are the most basal "crown" order and the Albuginales is a sister order to the Pythiales. Within the Pythiales themselves a highly supported split between *Pythium* Clades A-D and Clades F-I is observed, matching similar splits seen in small-scale analyses (Lévesque and de Cock, 2004; Villa *et al.*, 2006) (**Figure 3.3**). *Pilasporangium apinafurcum*, a Pythiales species, is placed sister to *Pythium* Clades F-I. The placement of *Phytopythium vexans* as an basal taxa within the Peronosporales has maximum bootstrap support, supporting the recent reclassification of the *Phytopythium* genus from the Pythiales (de Cock *et al.*, 2015). Many individual *Phytophthora* clades within the Peronosporales are well-supported, In addition, the "downy mildews" species in our dataset (*Hyaloperonospora arabidopsidis* and two *Plasmopara* species) place as derived taxa within the Peronosporales order rather than as basal to *Phytophthora*. The overall phylogeny of the Peronosporales in our MRP supertree is summarized in **Figure 3.4a** and discussed in greater detail later in the text. As an additional analysis, a consensus super network of the phylogenetic splits within the 2,280 single-copy gene phylogenies was generated in SplitsTree (Huson and Bryant, 2006) (**Figure S3.1**). The network further highlights support for the four "crown" oomycete orders and the division of the Pythiales order as in the supertree phylogeny, it also recapitulates many of individual *Phytophthora* clades and intra-order relationships within the Peronosporales (**Figures 3.3-4a, Figure S3.1**).

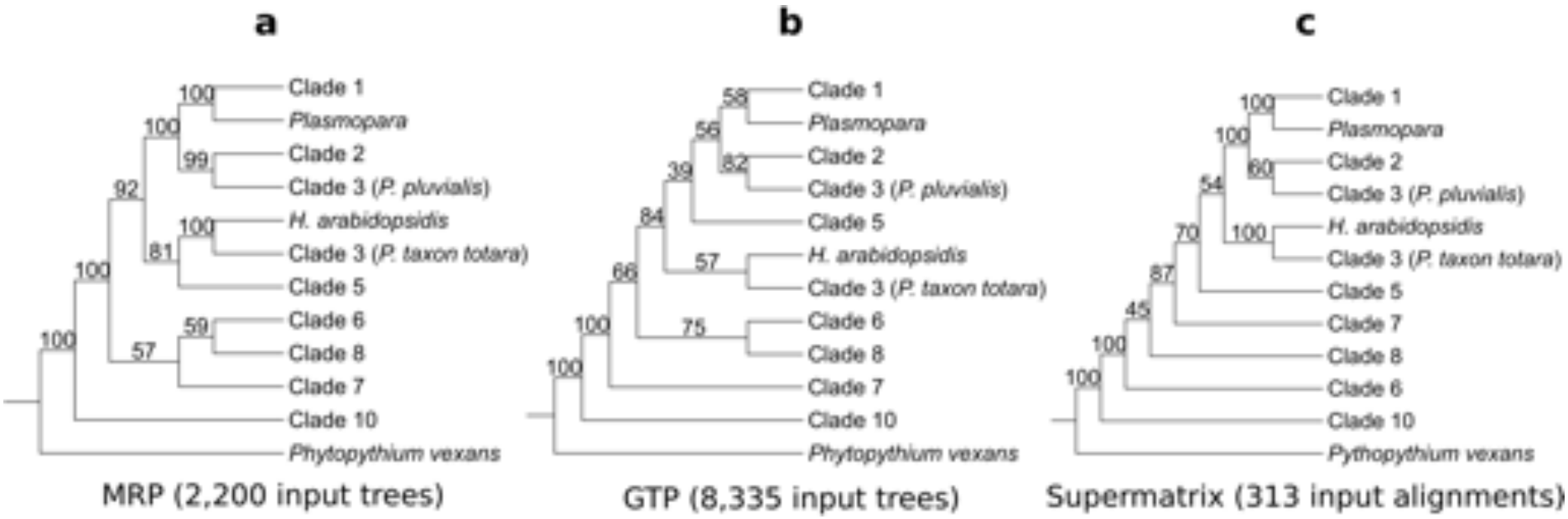


Figure 3.4. Congruence of the Peronosporales order between our supertree and supermatrix methods. **(a)** MRP analysis, **(b)** GTP analysis, **(c)** concatenated supermatrix analysis. For full phylogenies, refer to **Figures 3.3, 3.5 and 3.6** respectively

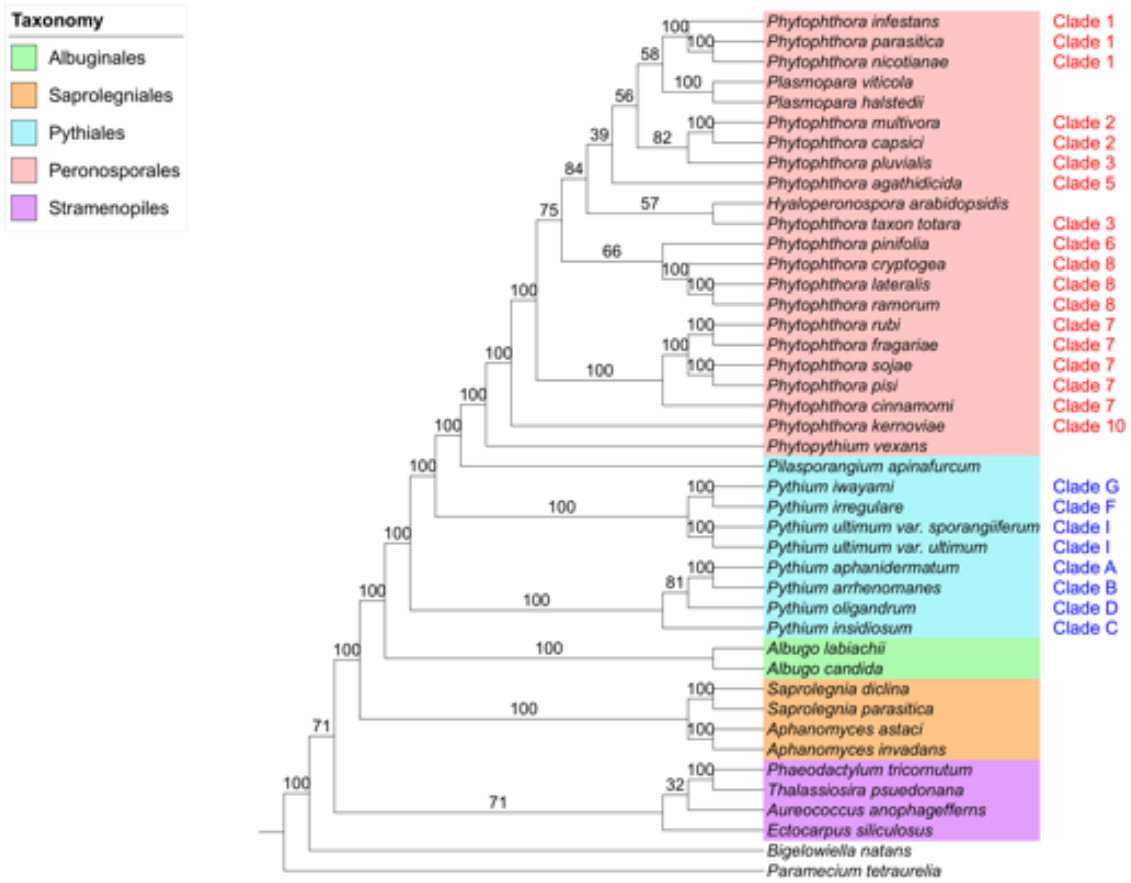


Figure 3.5. Gene tree parsimony (GTP) supertree of 37 oomycete species and 6 SAR species (8,335 source phylogenies). Supertree generated in DupTree. Phylogeny rooted at branch separating oomycetes and SAR. No colour: *P. tetraurelia* (Alveolata) and *B. natans* (Rhizaria).

Both the 2,280 single-copy phylogenies and the 6,055 multi-copy phylogenies (209,904 genes in total) were used as input for DupTree (Wehe *et al.*, 2008), which uses a gene tree parsimony (GTP) method to determine consensus phylogeny for many source phylogenies that may include gene duplication events. The source data was bootstrapped with 100 replicates, and the resultant consensus GTP supertree was rooted at the branch containing *Paramecium tetraurelia*, *Bigelowiella natans* and the other Stramenopiles species (**Figure 3.5**). As in the MRP supertree, all four individual crown oomycete orders and the oomycete class phylogeny are highly supported. The Pythiales order is once again split into highly-supported sister branches containing Clades A-D and Clades F-I respectively, and *Pi. apinafurcum* appears as a sister taxa to *Phytopythium vexans* (**Figure 3.5**). The Peronosporales order is fully supported again, as is the placement of *Phytopythium vexans* as a basal member of the order (**Figures 3.4b & 3.5**). The downy mildews also place as highly derived taxa within the order, with weaker bootstrap supports in more derived branches than in the MRP supertree (**Figures 3.3, 3.4a-b & 3.5**). Overall, the phylogeny of the Peronosporales order in the GTP supertree displays weaker bootstrap support at some branches than the MRP supertree, but there is relatively good taxonomic congruence between the two supertree approaches for the Peronosporales (**Figures 3.3, 3.4a-b & 3.5**).

3.3.3 Supermatrix approach based on ubiquitous Peronosporales gene phylogenies supports single-copy supertree phylogeny

As a complement to our supertree method phylogenies, we undertook a supermatrix approach to try to infer oomycete phylogeny by using oomycete homologs of known Clusters of Orthologous Groups (COG) proteins as markers (Parra, Bradnam and Korf, 2007). To identify oomycete COGs, we performed a reciprocal BLASTp of all 37 oomycete proteomes in our full dataset (590,896 protein sequences in total) against 458 *Saccharomyces cerevisiae* COGs with an e-value of 10^{-10}. 443 oomycete gene families representing oomycete reciprocal top hits with *S. cerevisiae* COGs were retrieved, of 144 families contained an ortholog from all 37 oomycete species. A superalignment of 16,934 characters was generated by concatenating 131 aligned families which retained alignment data after Gblocks sampling with FASconCAT (Kück and Meusemann, 2010). The maximum-likelihood phylogeny of this superalignment was reconstructed in PhyML with 100 bootstrap replicates and a LG+I+G+F amino acid

substitution model as selected by ProtTest, and the resultant consensus phylogeny was rooted at the Saprolegniales branch (**Figure S3.2**). This initial supermatrix phylogeny supported the four "crown" orders similar to our supertree phylogenies, however poor resolution and inconsistent phylogeny was observed within the Peronosporales, particularly the placement of species from *Phytophthora* Clades 7 and 8 (**Figure S3.2**). To attempt to tease apart the poor resolution of the Peronosporales in our maximum-likelihood phylogeny, a neighbour-joining network was generated for the OCOG superalignment in SplitsTree to visualize the bifurcations within the superalignment (**Figure S3.3**). As can be seen in the network, a significant amount of phylogenetic conflict displayed as alternative splits exist within Peronosporales between clades, matching the poor bootstrap supports and inconsistent topology throughout the Peronosporales in this class-level supermatrix phylogeny (**Figures S3.2-S3.3**).

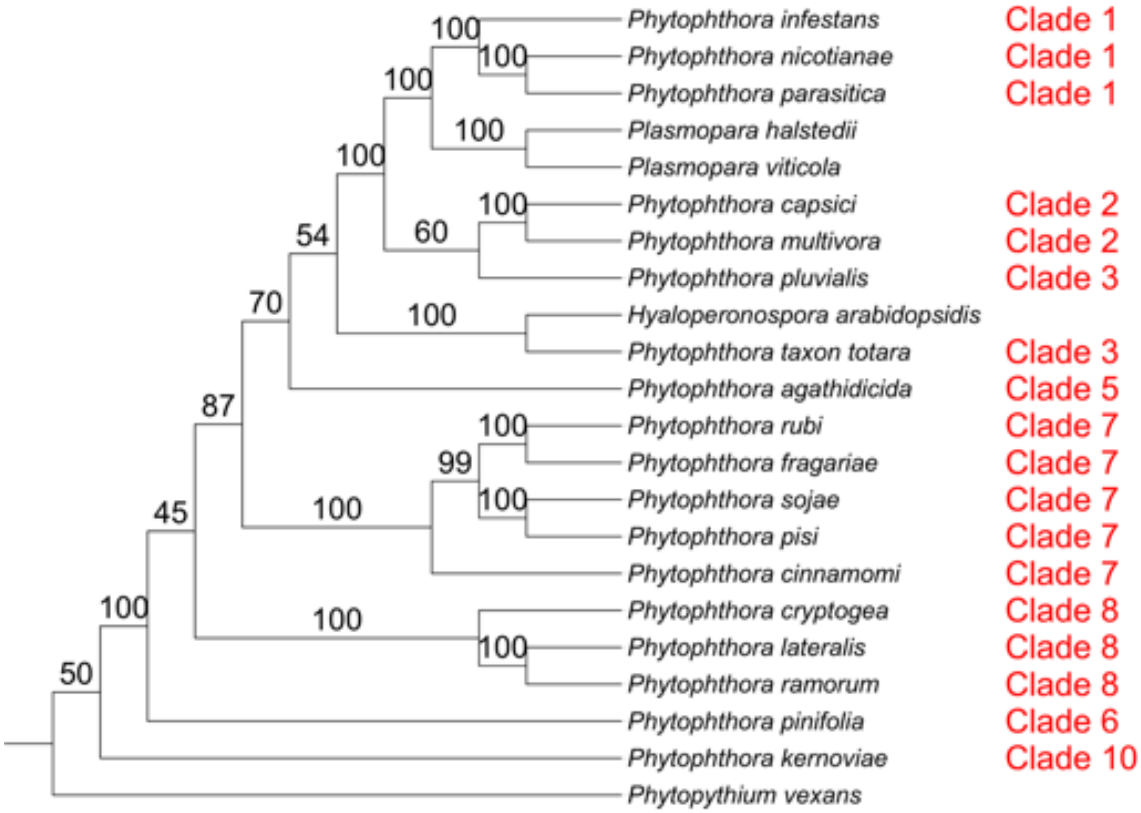


Figure 3.6. Maximum-likelihood (ML) supermatrix phylogeny of 22 Peronosporales species (313 ubiquitous Pernosporales gene families, 47,635 characters). Supermatrix phylogeny generated in PhyML with a JTT+I+G+F amino acid substitution model. Cladogram rooted at *Phytopythium vexans*.

To extend our OCOG supermatrix phylogeny, we took the approach of generating a supermatrix from ubiquitous gene families within the 22 Peronosporales species in our dataset. Using this approach, we hoped extend the amount of available alignment data for species solely within Peronosporales to improve resolution of the order. We defined a ubiquitous Peronosporales gene family as containing exactly one ortholog from all 22 Peronosporales species in our dataset. Using OrthoMCL, with a strict BLASTp e-value of 10^{-20} and an inflation value of 1.5, we identified over 20,000 orthologous gene families in the 22 Peronosporales proteomes in our dataset. From these families we identified 352 ubiquitous gene families within Peronosporales using bespoke Python scripting, each family was then aligned in MUSCLE and sampled in Gblocks. After removing families which did not retain alignment data after Gblocks, we concatenated the remaining 313 gene families into a superalignment of 47,365 amino acids in length. The maximum-likelihood phylogeny for this superalignment was generated with 100 bootstrap replicates and a JTT+I+G+F evolutionary model. The resultant consensus phylogeny was rooted at *Phytopythium vexans* (**Figure 3.6**). While resolution of relationships between clades is still weak at some branches, the higher support seen on many other branches as well as the overall topology of the ubiquitous supermatrix phylogeny represent a substantial improvement over the OCOG supermatrix. *Phytophthora* Clades 1, 2, 7 and 8 are now individually all monophyletic with 100% bootstrap support. The order is split between the basal lineages (*Phytopythium* and *Phytophthora* Clades 6-10), and the more derived *Phytophthora* Clades 1-5 and the downy mildews (70% bootstrap support) (**Figures 3.4c & 3.6**), matching the phylogeny of the order as seen in our supertree methods with greatest congruence to the MRP supertree (**Figures 3.4a-b**).

3.3.4 Resolution of the Peronosporales order in phylogenomic analysis

All three of our whole-genome supertree and supermatrix phylogenies support the Peronosporales order (**Figures 3.4a-c**) and display relative congruence with each other. Each phylogeny also supports the recent reclassification of *Phytopythium* from the Pythiales to the Peronosporales as a basal taxon (de Cock *et al.*, 2015). All three phylogenies also show varying but strong bootstrap support (70-92%) for the divergence of *Phytophthora* Clades 1-5 and the downy mildews (*Plasmopara* spp., *H. arabidopsidis*) from the remaining *Phytophthora* clades and *Phytopythium* at a single point (**Figures 3.4-**

c). The relationship between these taxa across our phylogenies can be summarized as follows:

1) *Phytophthora* Clade 3 is split in each phylogeny (**Figures 3.4a-c**).
2) The downy mildews species *Hyaloperonospora arabidopsidis* and *Phytophthora taxon totara* (*Phytophthora* Clade 3) are sister taxa, with maximum support in both MRP and supermatrix analysis (**Figures 3.4a & 3.4c**).
3) A close relationship between *Phytophthora* Clades 1 and 2, the Clade 3 species *Phytophthora pluvialis* and the downy mildew species *Plasmopara viticola* and *Plasmopara halstedii* is observed in each phylogeny, with maximum support in both MRP and supermatrix analysis (**Figures 3.4a & 3.4c**).

The placement of the Clade 5 species, *Phytophthora agathidcida*, varies in each phylogeny but it appears that the species is most closely related to *P. taxon totara* and *H. arabidopsidis* within the Peronosporales, as is most apparent in MRP analysis (81% bootstrap support) (**Figure 3.4a**). As for the more basal clades, both the MRP and GTP phylogenies show some support for the Clade 6 species *Phythophthora pinifolia* being sister to *Phytophthora* Clade 8, with highest bootstrap support (75%) seen in the latter (**Figures 3.4a-4b**).

In our analysis, we set out to resolve relationships within the oomycetes where conflicts has arisen in different analyses, particularly in the Peronosporales order (**Figures 3.2a-d**). In respect to the divergence of *Phytophthora* Clades 1-5 and the downy mildews from the remaining basal taxa in the Peronosporales (i.e. *Phytophthora* Clades 6-10 and *Phytopythium*), our results are congruent with the small-scale analyses performed by Blair et al. and Martin et al. (Blair *et al.*, 2008; Martin, Blair and Coffey, 2014) (**Figures 3.2a, 3.2c-2d**) with closest topological similarity to the latter authors' 6-loci coalescent method phylogeny (**Figure 3.2d**), despite a lack of data from *H. arabidopsidis* and *Plasmopara* species in both analyses and the inclusion of *H. arabidopsidis* data in the analysis carried out by Runge et al. (Runge *et al.*, 2011) (**Figure 3.2b**). Our own analysis lacks genomic data from any species in *Phytophthora* Clade 4, which is still un-sampled in terms of genome sequencing. In Runge et al., *H. arabidopsidis* branches within a paraphyletic *Phytophthora* Clade 4; were there a representative species from Clade 4 available a greater degree of resolution for the relationships between *Phytophthora* Clades 3-5 and *Hyaloperonospora* would likely be

observed. However, it is not clear whether the placement of *H. arabidopsidis* relative to *Phytophthora* Clade 1 would then recapitulate that of Runge et al. (Runge *et al.*, 2011). Similarly, with regards to the basal taxa our results are relatively congruent with the linearized relationships seen in previous analysis (**Figures 3.2a-d**), although the close relationship of the Clade 6 species *Phytophthora pinifolia* to *Phytophthora* Clade 7 seen in our two supertree methods is not reflected (**Figures 3.4a-b**) in any of the multi-locus phylogenies. The resolution of the relationships between *Phytophthora* Clades 6, 7 and 8 varies both in support and topology between our analyses (**Figures 3.4a-c**), however similar variation can be observed between the highlighted multi-locus phylogenies (Blair *et al.*, 2008; Runge *et al.*, 2011; Martin, Blair and Coffey, 2014) (**Figures 3.2a-d**). The lack of genomic data from *Phytophthora* Clade 9 available also prevents any conclusions regards its placement in a whole-genome phylogeny, however it is likely that it would branch sister to Clade 10 species such as *Phytophthora kernoviae*, as the relationship between Clades 9 and 10 has been highly supported in multi-locus analyses (Blair *et al.*, 2008; Runge *et al.*, 2011; Martin, Blair and Coffey, 2014).

3.3.5 The use of supertree and phylogenomic methods in oomycete systematics

Our analysis is the first large-scale genome phylogeny of the oomycetes as a class, using all extant genomic data from 37 oomycete species. Our analysis has recapitulated the four crown orders of the oomycetes and many relationships within the two largest-sampled orders, the Pythiales and the Peronosporales. During our analysis, we were conscious of potential characteristics of oomycete genomes that could obfuscate phylogenomic analysis. Intraspecific hybridization within the *Phytophthora* genus has been increasingly reported in the literature, and usually occurs in nature between *Phytophthora* species within the same phylogenetic clade (Burgess, 2015). A number of hybrid species or hybridization events have been described in *Phytophthora* Clades 6 to 8 (Bertier *et al.*, 2013; Burgess, 2015; Husson *et al.*, 2015), however none of these species are present in our dataset. The role of HGT in affecting the quality of our analyses must also be considered; supertree and supermatrix analyses are thought to be susceptible to misleading signal in datasets where a large degree of HGT has occurred, particularly MRP analysis (Lapierre, Lasek-Nesselquist and Gogarten, 2014). While HGT from other microbial eukaryotes, fungi and prokaryotes have been identified within oomycete

genomes, the majority of these events are thought to be ancestral or have not occurred in proportions large enough that we feel it may have affected our results (Richards *et al.*, 2011; Savory, Leonard and Richards, 2015; McCarthy and Fitzpatrick, 2016). Other factors, such as such as fast evolving regions of genomes or ancestral gene loss or duplication events within the oomycetes are not likely to have impacted on our analysis, given our genome-wide scale of data acquisition and our strict filtering of gene families with poor phylogenetic signal (Faith and Cranston, 1991; Judelson, 2012; Seidl *et al.*, 2012).

Compared with fungi, particularly in light of the ongoing 1000 fungal genomes project (http://1000.fungalgenomes.org), there is a relative dearth of genomic data for both the earliest diverging lineages and the "crown" taxa within the oomycetes. With a greater sampling of genomic sequencing of the oomycetes in the future, subsequent genome phylogenies of the oomycetes will hopefully match the success of genome phylogenies elsewhere in the eukaryotes at resolving individual species and clades (Beck *et al.*, 2006; Fitzpatrick *et al.*, 2006). It is possible that with a broader sampling of all *Phytophthora* clades and downy mildew species, we would see better resolution of the Peronosporales within any subsequent oomycete genome phylogenies. Similar approaches with other oomycete taxa, such as *Pythium*, may disentangle some of the phylogenetic conflicts seen in recent analyses (Uzuhashi, Tojo and Kakishima, 2010; Robideau, Rodrigue and André Lévesque, 2014). Similarly, sequencing of more Saprolegniales species or basal oomycete species and their inclusion in similar analyses will potentially help uncover further aspects of oomycete evolution, including the evolution of phytopathogenicity. Such analysis, of which ours is a first step, would also provide the benefit of establishing a robust phylogeny for an eukaryotic group with such devastating ecological impact, and hopefully encourage further genomics and phylogenomics research into the oomycetes.

3.4 Conclusions

Using 37 oomycete genomes and 6 SAR genomes, we have carried out the first whole-genome phylogenetic analysis of the oomycetes as a class. The different methods we used in our analysis support the four "crown" oomycete orders, and many individual phylogenetic clades within genera. Our analysis also generally supports the placement of *Phytopythium* within the Peronosporales, the placement of the downy mildews within the *Phytophthora* genus, and the topology of clades within the Pythiales order. However, resolution of the Peronosporales as an order remains weak at some branches, possibly due to a lack of genomic data for some phylogenetic clades within *Phytophthora*. As the amount of genomic data available for the oomycetes increases, future genome phylogenies of the class should resolve these branches, as well as those within currently unsampled basal lineages or under sampled taxa such as *Saprolegnia*. Our analysis represents an important backbone for oomycete phylogenetics, upon which future analyses can be compared.

Chapter 4 – Multiple approaches to phylogenomic reconstruction of the fungal kingdom

This chapter was previously published in *Advances in Genetics* in December 2017.

McCarthy C. G. P. & Fitzpatrick D. A. (2017). Multiple approaches to phylogenomic reconstruction of the fungal kingdom. *Advances in Genetics*, 100, pp. 211-266.

Chapter outline

Fungi are possibly the most diverse eukaryotic kingdom, with over a million-member species and an evolutionary history dating back a billion years. Fungi have been at the forefront of eukaryotic genomics and owing to initiatives like the 1000 Fungal Genomes Project the amount of fungal genomic data has considerably increased over the last five years, enabling large-scale comparative genomics of species across the kingdom. In this chapter, we first review fungal evolution and the history of fungal genomics. We then review in detail seven phylogenomic methods and reconstruct the phylogeny of 84 fungal species from 8 phyla using each method. Six methods have seen extensive use in previous fungal studies, while a Bayesian supertree method is novel to fungal phylogenomics. We find that both established and novel phylogenomic methods can accurately reconstruct the fungal kingdom. Finally, we discuss the accuracy and suitability of each phylogenomic method utilised.

4.1 Introduction

4.1.1 The phylogeny of the fungal kingdom

The fungi are one of the six kingdoms of life *sensu* Cavalier-Smith, sister to the animal kingdom, and encompasses millions of species found across a broad range of ecosystems (Berbee and Taylor, 1992; Baldauf and Palmer, 1993; Nikoh *et al.*, 1994; Cavalier-Smith, 1998; Hawksworth, 2001). While the overall fossil record of the fungi is poor due to their simple morphology, fungal fossils have been identified dating back to the Ordovician period approximately 400 million years ago (Redecker, 2000) and molecular clock analyses suggest that the fungi originated in the Precambrian eon approximately 0.76—1.06 billion years ago (Berbee and Taylor, 2010). Classic studies into fungal evolution were based on the comparison of morphological or biochemical characteristics, however the broad range of diversity within the fungal kingdom had limited the efficacy of some of these studies (Léjohn, 1974; Taylor, 1978; Heath, 1980; Berbee and Taylor, 1992). Since the development of phylogenetic approaches within systematics and the incorporation of molecular data into phylogenetic analyses our understanding of the evolution of fungi has improved substantially (Guarro, Gené and Stchigel, 1999).

Initial phylogenetic analyses of fungal species had revealed that there were four distinct phyla within the fungal kingdom; the early-diverging Chytridiomycota and Zygomycota, and the Ascomycota and Basidiomycota. The Chytridiomycota grouping was later subject to revision (James, Kauff, *et al.*, 2006), and in their comprehensive classification of the fungal kingdom in 2007, Hibbet et al. formally abandoned the phylum Zygomycota (Hibbett *et al.*, 2007). Instead Hibbet et al. treated zygomycete species as four *incertae sedis* subphyla (Entomophthoromycotina, Kickellomycotina, Mucoromycotina and Zoopagomycotina) and subsequently described one subkingdom (the Dikarya) and seven phyla namely Chytridiomycota, Neocallimastigomycota, Blastocladiomycota, Microsporidia, Glomeromycota, Ascomycota & Basidiomycota (Hibbett *et al.*, 2007). More recent phylogenetic classification of the zygomycetes has led to the circumscription of the Mucoromycota and Zoopagomycota phyla (Spatafora *et al.*, 2016). Furthermore, recent phylogenetic analyses have shown that *Rozella* species occupy a deep branching position in the fungal kingdom (James, Kauff, *et al.*, 2006; Jones, Forn, *et al.*, 2011), the clade containing these species are now termed the Cryptomycota phylum (Jones, Forn, *et al.*, 2011; Jones, Richards, *et al.*, 2011).

4.1.2 *Saccharomyces cerevisiae* and the origin of modern fungal genomics

In terms of genomic data fungi are by far the highest sampled eukaryotic kingdom, with assembly data available for over 1,000 fungal species on the NCBI's GenBank facility as of May 2017. Many of these species also have multiple strains sequenced (the most extreme example being *Saccharomyces cerevisiae*, which has over 400 strain assemblies available on GenBank). This reflects both the ubiquity of fungi in many areas of biological and medical study, and the relative simplicity of sequencing fungal genomes with modern sequencing technology. Fungi have been the exemplar group in eukaryote genetics and genomics, from the first determination of a nucleic acid sequence taken from *S. cerevisiae* by Holley and company in the late 1960s to the sequencing of the first eukaryotic genome in the mid-1990s (Holley *et al.*, 1965; Goffeau *et al.*, 1996). The genome of *S. cerevisiae* was sequenced through a massive international collaboration that grew to involve approximately 600 scientists in 94 laboratories and sequencing centres from across 19 countries between 1989 and 1996 (Goffeau and Vassarotti, 1991; Goffeau *et al.*, 1996; Engel *et al.*, 2014). Throughout the early 1990s each of the standard 16 nuclear chromosomes of *S. cerevisiae*, sourced from the common laboratory strain 288C and its isogenic derivative strains AB972 and FY1679, were individually sequenced and published by participating researchers. Engel et al. (2014) briefly summarizes each of these sequencing projects. The initial publication of chromosome III involving 35 European laboratories on its own (Oliver *et al.*, 1992). The complete genome sequence of *S. cerevisiae* 288C was finally published in 1996, with 5,885 putative protein-coding genes and 275 transfer RNA genes identified across the genome's ~12 million base pairs (Goffeau *et al.*, 1996).

In the intervening years the *S. cerevisiae* 288C reference genome has been constantly updated and refined as individual genes or chromosomes have been reanalysed or even resequenced, and all of these revisions have been recorded and maintained by the Saccharomyces Genome Database (Fisk *et al.*, 2006). It is worth noting however, that such was the attention paid to the original sequencing project by its contributors that the most recent major update of the *S. cerevisiae* 288C reference genome, a full resequencing of the derivative AB972 strain using far less labour-intensive modern sequencing and annotation techniques, made only minor alterations to the original genome annotation

overall (Engel *et al.*, 2014). Much of our understanding regarding the processes of genome evolution in eukaryotes since 1996 have also been derived from the study of the *S. cerevisae* 288C genome; including the confirmation that the *S. cerevisiae* genome had undergone a whole genome duplication (WGD) event (Wolfe and Shields, 1997; Kellis, Birren and Lander, 2004), the effect of interspecific hybridization on genome complexity (De Barros Lopes *et al.*, 2002), evidence that inter-domain horizontal gene transfer (HGT) from prokaryotes into eukaryotes has occurred (Hall and Dietrich, 2007), to the ongoing development of an entirely synthetic genome through the Sc2.0 project (Annaluru *et al.*, 2014).

4.1.3 Fungal genomics and phylogenomics beyond the yeast genome

As more model organisms from other eukaryotic kingdoms had their genomes sequenced, *S. cerevisiae* 288C provided a useful comparison as the reference fungal genome, even for more complex eukaryotes like *Drosophila melanogaster*. However, the later sequencing of other model fungal species *Schizosaccharomyces pombe* and *Neurospora crassa* showed the limits of relying solely on *S. cerevisiae* as a reference for the entire fungal kingdom, particularly the latter; *N. crassa* was found to have a far larger genome than either *S. cerevisiae* or *S. pombe* and over 57% of genes predicted in *N. crassa* had no homolog in either of the other two sequenced fungal genomes (Wood *et al.*, 2002; Galagan *et al.*, 2003; Galagan, Henn, *et al.*, 2005). Borne out of a lull in fungal genomic advances and the increasing sophistication of sequencing technology, the Fungal Genome Initiative (FGI) was set up by a number of research organizations in the early 2000s, under the aegis of the Broad Institute (Cuomo and Birren, 2010). Collaborators within the FGI were tasked with the sequencing and annotating the genomes of over 40 species from across the fungal kingdom, with a broad scope of species selected for analysis; medically significant human fungal pathogens like *Candida albicans* and *Aspergillus fumigatus*, commercially important species such as *Penicillium chrysogenum* and *Sclerotinia sclerotiorum*, as well as basal fungal species such as *Phycomyces blakesleeanus* (Cuomo and Birren, 2010). Between 2004 and 2012, in approximately the same amount of time it had taken to sequence each individual chromosome of *S. cerevisae* 288C in the 1990s, over 100 fungal genomes were sequenced and made publicly-available on facilities like GenBank and the Joint Genome Institute's Genome Portal website (Grigoriev, Nordberg, *et al.*, 2011; Benson *et al.*, 2013).

The steady increase in genomic data available for fungi from the first decade of this century on, while still sampled mainly from the Ascomycota and Basidiomycota phyla, allowed for a greater range of fungal genomic analyses to be conducted. This included phylogenomic analyses of the fungal kingdom using a variety of different methods (which we will discuss in detail in the following section) and comparative investigations such as analysis of the evolution of pathogenicity in genera like *Candida* or *Aspergillus* (Galagan, Calvo, *et al.*, 2005; Butler *et al.*, 2009; Jackson *et al.*, 2009), the extent of inter/intra-kingdom HGT both to and from fungal genomes (Fitzpatrick, Logue and Butler, 2008; Marcet-Houben and Gabaldón, 2010; Richards *et al.*, 2011; Szöllősi *et al.*, 2015), identification of clusters of secondary metabolites (Keller, Turner and Bennett, 2005; Khaldi *et al.*, 2010) and syntenic relationships across *Saccharomyces* and *Candida* (Byrne and Wolfe, 2005; Fitzpatrick *et al.*, 2010). The wealth of genomic data available for some fungal orders or classes has allowed for easier automation of the sequencing and annotation of novel related species, through the development of reference transcriptomic or proteomic data for gene prediction software such as AUGUSTUS or quality assessment software for genome assembly such as BUSCO (Stanke *et al.*, 2004; Simão *et al.*, 2015).

4.1.4 The 1000 Fungal Genomes Project

The recent deluge of genomic data available for the fungal kingdom comes as a result of the 1000 Fungal Genomes Project, an initiative headed by the Joint Genome Institute (JGI). The project (which can be found at http://genome.jgi.doe.gov/pages/fungi-1000-projects.jsf) aims to provide genomic sequence data from at least one species from every circumscribed fungal family, either from projects headed by the JGI, projects which have been incorporated into the MycoCosm database or through community-led nomination and provision of sequencing material. The project has an inbuilt preference for sequencing projects arising from families with no sequenced species to date, or only one other reference genome at the time of nomination. Assembly and annotation data is then hosted at the JGI's MycoCosm facility as well as other publicly-available databases (Grigoriev *et al.*, 2014). This community-wide effort has led to a staggering increase in the number of fungal genomes available within the last 5 years; Grigoriev et al. (2014) quoted the number of genomes present in MycoCosm at over 250 at the end of 2013, as of May 2017 there are 772 fungal genomes available to download from the facility, with another 500 species nominated for sequencing. The project has seen a large increase

particularly in the amount of data available from fungal phyla outside of the Dikarya, with 58 genomes currently available from the zygomycetes, the Chytridiomycota, Neocallimastigomycota and Blastocladiomycota. There are many other fungal families with species yet to be nominated for sequencing, including many families from the Pezizomycotina subphylum within Ascomycota and the Chytridiomycota phylum. It is hoped that the wealth of fungal genomic data arising from the 1000 Fungal Genomes will help, amongst countless other scenarios, to fuel the search for novel biosynthetic products and to better understand the ecological effects of different families within the fungal kingdom (Grigoriev, Cullen, *et al.*, 2011). The initiative will also enable the large-scale comparative analysis of hundreds of fungal species from across the fungal kingdom, including kingdom-level phylogenomic reconstructions.

4.2 Phylogenomic reconstructions of the fungal kingdom

Phylogenetic inference arising from molecular data has, in the past, predominately relied on single genes or small numbers of highly conserved genes or nuclear markers. While usually these markers make for robust individual phylogenies, potential conflicts can occur between individual phylogenies depending on the marker(s) used. The selection of such markers may also overlook other gene families which may be phylogenetically informative, such as gene duplication events or HGT events (Bininda-Emonds, 2004). With the advent of genome sequencing and the increasing sophistication of bioinformatics software and techniques, it has become common practice to reconstruct the evolutionary relationships of species by utilizing large amounts of phylogenetically informative genomic data. Such data can include ubiquitous or conserved genes, individual orthologous and paralogous gene phylogenies, shared genomic content or compositional signatures of genomes (**Figure 4.1**). Methods of phylogenomic analysis, in other words phylogenetic reconstruction of species using genome-scale data, have all been developed for each of these types of potential phylogenetic marker and each comes with their advantages and disadvantages. Many phylogenomic analyses of the fungal kingdom have been carried out using these methods.

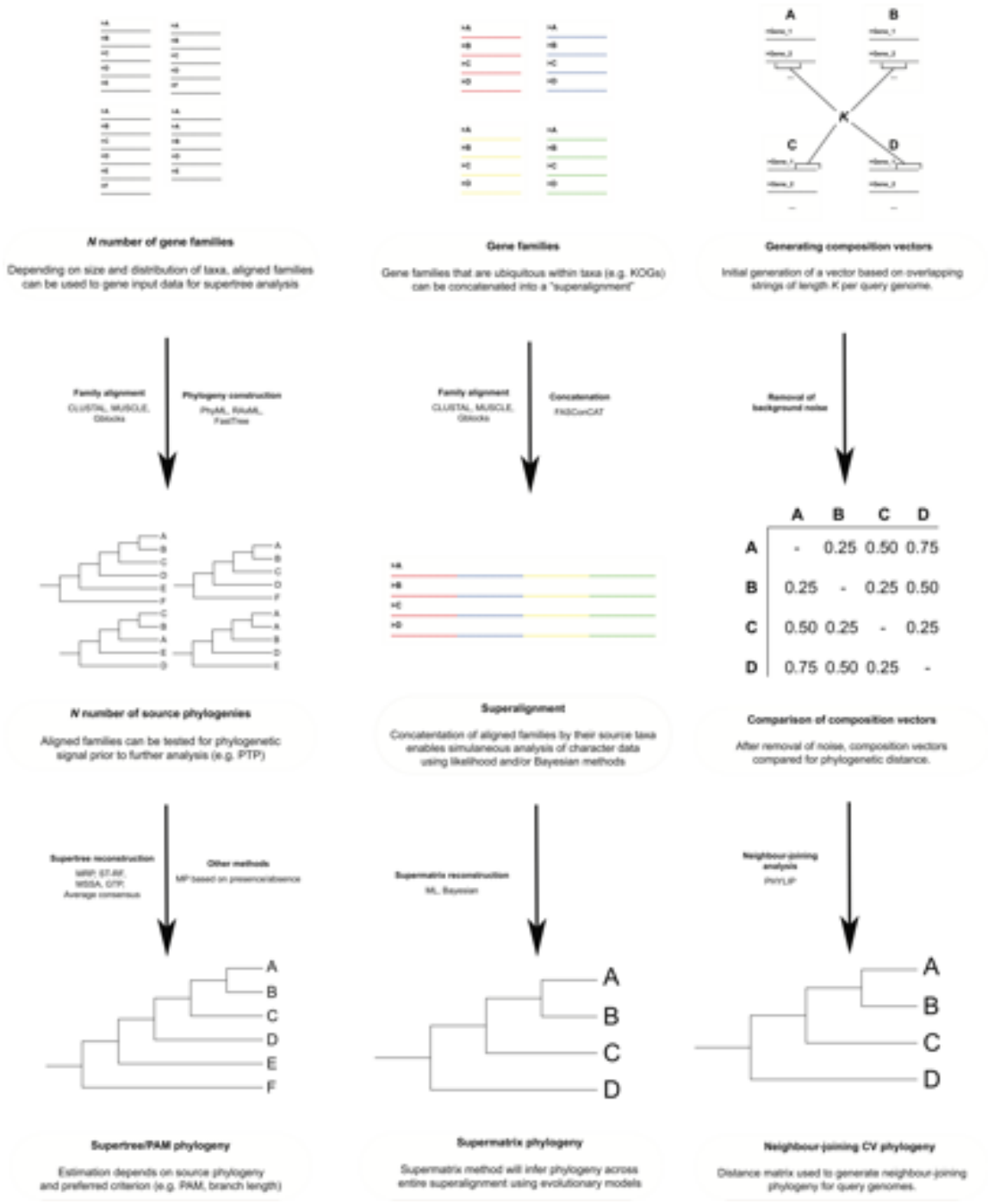


Figure 4.1. Illustrative comparison of common phylogenomic methods. Top: supertree and presence-absence methods, middle: supermatrix methods, bottom: composition vector methods.

In this section, we review in turn each established approach to phylogenomic reconstruction from molecular data present in the literature, and review each approach's application in previous fungal phylogenomic analyses. To demonstrate both the application and accuracy of all of these approaches to reconstructing phylogeny from

genome-scale data, we have conducted our own phylogenomic analyses of the fungal kingdom using each method (**Figure 4.2**). We have carried out such analyses to take advantage of both the greater coverage of the fungal kingdom arising from the 1000 Fungal Genomes Project, and the advances in phylogenetic methodologies in the years following many of the analyses that we review below. In total, 84 fungal genomes from across 8 phyla (**Table 4.1**) were selected for our large-scale phylogenomic reconstructions of the fungal kingdom. Where possible, we included at least one published representative genome from each order covered by the 1000 Fungal Genomes Project in our dataset. All genomic data was ultimately obtained from the Joint Genome Institute's Mycocosm facility (Grigoriev *et al.*, 2014). Our analyses include the first phylogenomic reconstruction of fungi carried out using a Bayesian supertree approach, and the first large-scale gene content approach to fungal phylogenomics that has been conducted in at least a decade. We discuss in brief, the methodology and the results of each reconstruction and their accuracy (or otherwise) in reconstructing the phylogeny of both basal fungal lineages and the Dikarya. In Section 3, we discuss the overall phylogeny of the fungal kingdom arising from our analyses and compare with previous literature.

Figure 4.2. Summary of the methodology of all seven phylogenomic analyses of 84 fungal species carried out in this review. Refer to text for acronyms.

Table 4.1. List of species used in phylogenomic analysis. Genome data from MycoCosm (http://genome.jgi.doe.gov/programs/fungi/index.jsf) has previously been published and MycoCosm ID is given in final golumn. Genbank accessions given for *Allomyces macrogynus* and *Batrachochytrium dendrobatidis*.

Species	Phylum	Subphylum	Class	MycoCosm ID
Bipolaris maydis	Ascomycota	Pezizomycotina	Dothideomycetes	CocheC4_1
Cenococcum geophilum	Ascomycota	Pezizomycotina	Dothideomycetes	Cenge3
Hysterium pulicare	Ascomycota	Pezizomycotina	Dothideomycetes	Hyspu1_1
Zymoseptoria tritici	Ascomycota	Pezizomycotina	Dothideomycetes	Mycgr3
Aspergillus niger	Ascomycota	Pezizomycotina	Eurotiomycetes	Aspni7
Coccidioides immitis	Ascomycota	Pezizomycotina	Eurotiomycetes	Cocim1
Endocarpon pusillum	Ascomycota	Pezizomycotina	Eurotiomycetes	EndpusZ1
Exophiala dermatitidis	Ascomycota	Pezizomycotina	Eurotiomycetes	Exode1
Phaeomoniella chlamydospora	Ascomycota	Pezizomycotina	Eurotiomycetes	Phach1
Blumeria graminis	Ascomycota	Pezizomycotina	Leotiomycetes	Blugr1
Botrytis cinerea	Ascomycota	Pezizomycotina	Leotiomycetes	Botci1
Arthrobotrys oligospora	Ascomycota	Pezizomycotina	Orbiliomycetes	Artol1
Dactylellina haptotyla	Ascomycota	Pezizomycotina	Orbiliomycetes	Monha1
Pyronema omphalodes	Ascomycota	Pezizomycotina	Pezizomycetes	Pyrco1
Tuber melanosporum	Ascomycota	Pezizomycotina	Pezizomycetes	Tubme1
Coniochaeta ligniaria	Ascomycota	Pezizomycotina	Sordariomycetes	Conli1
Hypoxylon sp. EC38	Ascomycota	Pezizomycotina	Sordariomycetes	HypEC38_3
Magnaporthe grisea	Ascomycota	Pezizomycotina	Sordariomycetes	Maggr1
Neurospora crassa	Ascomycota	Pezizomycotina	Sordariomycetes	Neucr_trp3_1
Ophiostoma piceae	Ascomycota	Pezizomycotina	Sordariomycetes	Ophpic1
Phaeoacremonium minimum	Ascomycota	Pezizomycotina	Sordariomycetes	Phaal1
Xylona heveae	Ascomycota	Pezizomycotina	Xylonomycetes	Xylhe1
Candida albicans	Ascomycota	Saccharomycotina	Saccharomycetes	Canalb1
Lipomyces starkeyi	Ascomycota	Saccharomycotina	Saccharomycetes	Lipst1_1
Ogataea polymorpha	Ascomycota	Saccharomycotina	Saccharomycetes	Hanpo2
Saccharomyces cerevisiae	Ascomycota	Saccharomycotina	Saccharomycetes	SacceM3707_1
Saitoella complicata	Ascomycota	Taphrinomycotina	N/A	Saico1
Pneumocystis jirovecii	Ascomycota	Taphrinomycotina	Pneumocystidomycetes	Pneji1

Schizosaccharomyces cryophilus	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schcy1
Schizosaccharomyces japonicus	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schja1
Schizosaccharomyces octosporus	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schoc1
Schizosaccharomyces pombe	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schpo1
Protomyces lactucaedebilis	Ascomycota	Taphrinomycotina	Taphrinomycetes	Prola1
Taphrina deformans	Ascomycota	Taphrinomycotina	Taphrinomycetes	Tapde1_1
Agaricus bisporus	Basidiomycota	Agaricomycotina	Agaricomycetes	Agabi_varbur_1
Auricularia subglabra	Basidiomycota	Agaricomycotina	Agaricomycetes	Aurde3_1
Botryobasidium botryosum	Basidiomycota	Agaricomycotina	Agaricomycetes	Botbo1
Fibulorhizoctonia	Basidiomycota	Agaricomycotina	Agaricomycetes	Fibsp1
Gloeophyllum trabeum	Basidiomycota	Agaricomycotina	Agaricomycetes	Glotr1_1
Heterobasidion annosum	Basidiomycota	Agaricomycotina	Agaricomycetes	Hetan2
Jaapia argillacea	Basidiomycota	Agaricomycotina	Agaricomycetes	Jaaar1
Punctularia strigosozonata	Basidiomycota	Agaricomycotina	Agaricomycetes	Punst1
Serendipita indica	Basidiomycota	Agaricomycotina	Agaricomycetes	Pirin1
Serpula lacrymans	Basidiomycota	Agaricomycotina	Agaricomycetes	SerlaS7_3_2
Sistotremastrum suecicum	Basidiomycota	Agaricomycotina	Agaricomycetes	Sissu1
Sphaerobolus stellatus	Basidiomycota	Agaricomycotina	Agaricomycetes	Sphst1
Wolfiporia cocos	Basidiomycota	Agaricomycotina	Agaricomycetes	Wolco1
Calocera cornea	Basidiomycota	Agaricomycotina	Dacrymycetes	Calco1
Dacryopinax primogenitus	Basidiomycota	Agaricomycotina	Dacrymycetes	Dacsp1
Basidioascus undulates	Basidiomycota	Agaricomycotina	Geminibasidiomycetes	Basun1
Cryptococcus neoformans	Basidiomycota	Agaricomycotina	Tremellomycetes	Cryne_JEC21_1
Cutaneotrichosporon oleaginosus	Basidiomycota	Agaricomycotina	Tremellomycetes	Triol1
Wallemia sebi	Basidiomycota	Agaricomycotina	Wallemiomycetes	Walse1
Leucosporidium creatinivorum	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Leucr1
Microbotryum lychnidis-dioicae	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Micld1
Rhodotorula graminis	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Rhoba1_1
Mixia osmundae	Basidiomycota	Pucciniomycotina	Mixiomycetes	Mixos1
Puccinia graminis	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucgr2
Tilletiaria anomala	Basidiomycota	Ustilaginomycotina	Exobasidiomycetes	Tilan2

Malassezia sympodialis	Basidiomycota	Ustilaginomycotina	Malasseziomycetes	Malsy1_1
Sporisorium reilianum	Basidiomycota	Ustilaginomycotina	Ustilaginomycetes	Spore1
Ustilago maydis	Basidiomycota	Ustilaginomycotina	Ustilaginomycetes	Ustma1
Allomyces macrogynus	Blastocladiomycota	N/A	Blastocladiomycetes	GCA_000151295.1
Catenaria anguillulae	Blastocladiomycota	N/A	Blastocladiomycetes	Catan2
Batrachochytrium dendrobatidis	Chytridiomycota	N/A	Chytridiomycetes	GCA_000149865.1
Rhizoclosmatium globosum	Chytridiomycota	N/A	Chytridiomycetes	Rhihy1
Spizellomyces punctatus	Chytridiomycota	N/A	Chytridiomycetes	Spipu1
Gonapodya prolifera	Chytridiomycota	N/A	Monoblepharidomycetes	Ganpr1
Rozella allomycis	Cryptomycota	N/A	N/A	Rozal1_1
Rhizophagus irregularis	Mucoromycota	Glomeromycotina	Glomeromycetes	Gloin1
Mortierella elongate	Mucoromycota	Mortierellomycotina	N/A	Morel2
Phycomyces blakesleeanus	Mucoromycota	Mucoromycotina	N/A	Phybl2
Rhizopus oryzae	Mucoromycota	Mucoromycotina	N/A	Rhior3
Umbelopsis ramanniana	Mucoromycota	Mucoromycotina	N/A	Umbra1
Anaeromyces robustus	Neocallimastigomycota	N/A	Neocallimastigomycetes	Anasp1
Neocallimastix californiae	Neocallimastigomycota	N/A	Neocallimastigomycetes	Neosp1
Orpinomyces sp. C1A	Neocallimastigomycota	N/A	Neocallimastigomycetes	Orpsp1_1
Piromyces finnis	Neocallimastigomycota	N/A	Neocallimastigomycetes	Pirfi3
Basidiobolus meristosporus	Zoopagomycota	Entomophthoromycotina	Basidiobolomycetes	Basme2finSC
Conidiobolus thromboides	Zoopagomycota	Entomophthoromycotina	Entomophthoromycetes	Conth1
Coemansia reversa	Zoopagomycota	Kickxellomycotina	N/A	Coere1
Linderina pennispora	Zoopagomycota	Kickxellomycotina	N/A	Linpe1
Martensiomyces pterosporus	Zoopagomycota	Kickxellomycotina	N/A	Marpt1
Ramicandelaber brevisporus	Zoopagomycota	Kickxellomycotina	N/A	Rambr1

4.2.1 Supermatrix phylogenomic analysis of fungi

The two best-established alignment-based approaches to reconstructing phylogeny on a genomic scale are the "supertree" method, in which a consensus phylogeny is derived from many individual gene phylogenies (discussed in Chapter 4.2.2), and the "supermatrix" method which we discuss here. Supermatrix method phylogeny is the simultaneous analysis of a phylogenetic matrix, also referred to as a "superalignment", constructed from all available character data from a given set of taxa. Generally supermatrices are constructed from concatenating highly-conserved markers (e.g. rRNA genes, mitochondrial markers) for small-scale multi-gene phylogenies, and from homologs of conserved orthologous genes (known as COGs, or sometimes KOGs in eukaryotes) for genome-scale phylogenies (Koonin *et al.*, 2004; de Queiroz and Gatesy, 2007). Supermatrix approaches can also incorporate statistically-powerful maximum-likelihood and Bayesian methods of phylogenomic analysis. Described in simple terms, given an alignment of sequences and a suitable evolutionary model maximum-likelihood phylogenetic analysis examines all possible trees by their possible parameters (e.g. topology, site support, branch length) and returns the most likely phylogenetic tree for the alignment (Page and Holmes, 1998). Similarly, Bayesian analysis incorporates phylogenetic likelihoods to calculate the posterior probability of a phylogeny, which is the probability of that phylogeny given the alignment data (Huelsenbeck *et al.*, 2001).

One advantage of a supermatrix approach to phylogenomic analysis over a supertree approach is the retention of character evidence in analysis in the former approach; most supertree methods can be considered estimations using individual trees based on summarized character data, at least two steps removed from any actual sequence data, whereas a supermatrix approach entails direct analysis of combined character data (de Queiroz and Gatesy, 2007; Creevey and McInerney, 2009). Supermatrix methods also have the potential resolve deep branches and reveal so-called "hidden supports" within phylogenies that supertree methods may overlook (de Queiroz and Gatesy, 2007). However supermatrix analysis requires ubiquitous sequences from all taxa being investigated, which restricts the available pool of character data and may overlook important phylogenetic information from phylogenies with gene deletion, gene duplication or horizontal gene transfer events that supertrees methods can utilize (Creevey and McInerney, 2009). Compositional biases may also have an effect on

supermatrix methods, though phylogenetic models have been developed which can ameliorate errors that these biases may induce during analysis (Lartillot and Philippe, 2004; Lartillot, Brinkmann and Philippe, 2007). In practice, many phylogenomic analyses utilize both supertree and supermatrix methods in tandem to reconstruct phylogeny in a "total evidence" approach (Kluge, 1989), and will often comment on the topological congruence (or otherwise) of the resulting phylogenies.

4.2.1.1 Fungal phylogenomics using the supermatrix approach

Supermatrix analysis has been widely-used in fungal phylogenomics. One of the initial comparisons of individual gene phylogenies with genome-scale species phylogenies used a maximum-parsimony analysis amongst other methods to reconstruct the phylogeny of 7 *Saccharomyces* species and *Candida albicans*; the authors showed that incongruence amongst individual gene phylogenies could be resolved with high support using a concatenated alignment (Rokas *et al.*, 2003). Initial genome-based phylogenies of Ascomycota using 17 genomes and both supertree and supermatrix methods resolved both Pezizomycotina and Saccharomycotina, as well as placing *Schizosaccharomyces pombe* as an early diverging branch within Ascomycota (Robbertse *et al.*, 2006). Robbertse *et al.* (2006) generated a superalignment of 195,664 amino acid characters in length derived from 781 gene families, which produced identical topologies under both neighbour-joining and maximum-likelihood criteria. The first large-scale phylogenomic analysis of fungi used a 67,101-character superalignment derived from 531 eukaryotic COGs found in 21 fungal genomes, all of which were sampled from Ascomycota and Basidiomycota (Kuramae *et al.*, 2006). A more extensive phylogenomic analysis from the same year produced two highly congruent genome phylogenies from 42 fungal genomes using two methods; a matrix representation with parsimony (MRP) supertree derived from 4,805 single-copy gene families (which we discuss in greater detail in Section 4.2.2.1), and a 38,000-character superalignment derived from 153 ubiquitous gene families (Fitzpatrick *et al.*, 2006).

Most of the relationships resolved in Fitzpatrick *et al.* (2006) were further supported by a 31,123-character superalignment from 69 proteins conserved in up to 60 fungal genomes generated by Marcet-Houben *et al.* (2009a), although they found a large degree of topological conflict within a 21-species Saccharomycotina clade (Marcet-Houben and Gabaldón, 2009; Marcet-Houben, Marceddu and Gabaldón, 2009). A later

follow-up analysis to Fitzpatrick *et al.* (2006) by Medina *et al.* (2011) reconstructed the phylogeny of 103 fungal species by performing Bayesian analysis on a 12,267-site superalignment derived from 87 gene families with a phyletic range of over half of their dataset, in addition to supertree analysis (Medina, Jones and Fitzpatrick, 2011). Medina et al. (2011) used this supermatrix phylogeny along with supertree phylogenies as a scaffold to investigate the distribution of yeast prion homologs throughout the fungal kingdom. A recent phylogenomic analysis of 46 fungal genomes, including 25 zygomycetes species, reconstructed the phylogeny of the early-diverging fungal lineages using a 60,383-character superalignment (Spatafora *et al.*, 2016). Another recent phylogenomic analysis used a 28,807-site superalignment derived from 136 gene families from 40 eukaryotic genomes to investigate the evolution of sourcing carbon from algal and plant pectin in early-diverging fungi (Chang *et al.*, 2015). Finally, an analysis of the dynamics of genome evolution within 28 Dikarya species used a supermatrix phylogeny of 24,514 amino acid characters from 529 fungal gene families with large phyletic range to infer rates of intra-kingdom HGT within Dikarya (Szöllősi *et al.*, 2015).

To extend the analyses above, we carried out supermartrix analysis using maximum-likelihood and Bayesian methods on a superalignment constructed from orthologous genes conserved throughout 84 species from 8 phyla within the fungal kingdom. We describe our methodology and the resulting phylogenies in detail below.

4.2.1.2 Phylogenomic reconstruction of 84 fungal species from 72 ubiquitous gene families using Maximum Likelihood and Bayesian supermatrix analysis

A reciprocal BLASTp search was carried out between all protein sequences from our 84-genome dataset and 458 core orthologous genes (COGs) from *Saccharomyces cerevisiae* obtained from the CEGMA dataset, with an e-value cutoff of 10^{-10} (Parra, Bradnam and Korf, 2007; Camacho *et al.*, 2009), from which 456 COG families were retrieved (two *S. cerevisiae* COGs did not return any homologs). From these, 86 ubiquitous fungal COG families, i.e. families containing a homolog from all 84-input species, were identified. Each ubiquitous fungal COG family was aligned in MUSCLE, and conserved regions of each alignment were sampled in Gblocks using the default parameters (Castresana, 2000; Edgar, 2004). Fourteen alignments did not retain any character data after Gblocks filtering, and were removed from further analysis. The

remaining 72 sampled alignments were concatenated into a superalignment of 8,529 aligned positions using the Perl program FASconCat (Kück and Meusemann, 2010). This superalignment was bootstrapped 100 times using Seqboot (Felsenstein, 1989), and maximum-likelihood phylogenetic trees were generated for each individual replicate using PhyML with an LG+I+G amino acid substitution model as selected by ProtTest (Guindon *et al.*, 2010; Darriba *et al.*, 2011). A consensus phylogeny was generated from all 100 individual replicate phylogenies using CLANN (Creevey and McInerney, 2005). Markov Chain Monte Carlo (MCMC) Bayesian phylogenetic inference was carried out on the same superalignment using PhyloBayes MPI with the default CAT+GTR amino acid substitution model, running 2 chains for 1,000,000 iterations and sampling every 100 iterations (Lartillot and Philippe, 2004; Lartillot *et al.*, 2013). Both chains were judged to have converged after 100,000 iterations and a consensus Bayesian phylogeny was generated with a burn-in of 1,000 trees. Both supermatrix phylogenies were visualized using the Interactive Tree of Life (iTOL) website and annotated according to the NCBI's taxonomy database (Federhen, 2012; Letunic and Bork, 2016). Both supermatrix phylogenies were rooted at *Rozella allomycis*, which is the most basal species in evolutionary terms in our dataset (Jones, Forn, *et al.*, 2011), and is the root for all the phylogenies we present hereafter (**Figures 4.3 & 4.4**).

4.2.1.3 Supermatrix analyses of 84 fungal species accurately reconstructs the fungal kingdom

We reconstructed the phylogeny of the fungal kingdom by generating a superalignment of 72 concatenated ubiquitous gene families and performing ML analysis using PhyML and Bayesian analysis using a parallelized version of PhyloBayes. Both ML and Bayesian analysis reconstruct the phylogeny of our fungal dataset with a high degree of accuracy relative to other kingdom phylogenies in the literature and in most cases recover the 8 fungal phyla in our dataset (**Figures 4.3 & 4.4**). Here, we discuss the results of both our analyses with regards to the basal fungal lineages, and the two Dikarya phyla. Further in this chapter, we use these supermatrix analyses as the point of comparison for our other phylogenomic methods.

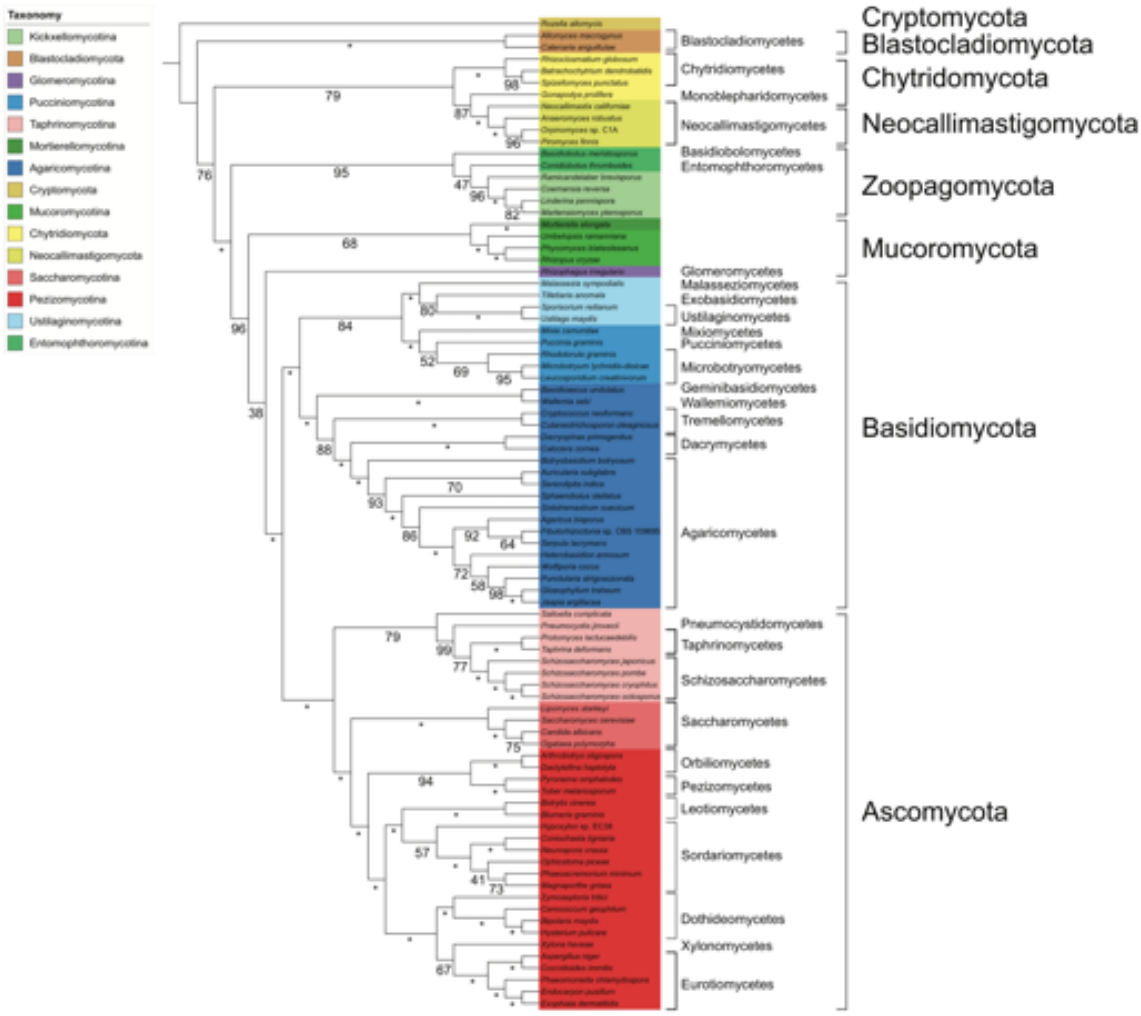


Figure 4.3. ML phylogeny of 84 fungal species from a 8,529-character superalignment derived from 72 ubiquitous fungal COG families sampled in Gblocks using PhyML a LG+I+G model. Bootstrap supports shown on branches. Maximum bootstrap support designated with an asterisk (*).

4.2.1.3.1 Basal fungi

In our ML supermatrix phylogeny, Blastocladiomycota emerge as the earliest-diverging fungi with maximum bootstrap support (henceforth abbreviated to BP) after rooting at *Rozella allomycis* (**Figure 4.3**). Chytridiomycota and Neocallimastigomycota are placed as sister clades with 79% BP, surprisingly the Chytrioiomycota species *Gonapodya prolifera* branches as sister to Neocallimastigomycota (87% BP). The Chytridiomycetes class is monophyletic with maximum bootstrap support, as is the Neocallimastiomycetes class (**Figure 4.3**). The former zygomycetes phylum Zoopagomycota is strongly supported as a monophyletic clade with 95% BP (**Figure 4.3**). The other former zygomycetes phylum Mucoromycota is paraphyletic and split between a clade containing the Mucoromycotina and Mortierellomycotina species *Mortierella*

elongata that has 68% BP, and the Glomeromycotina species *Rhizophagus irregularis* branching basal to Dikarya with lower support (38% BP). The placement of Mucoromycota as the closest phyla to Dikarya has near-maximum support (96% BP) which matches other analysis (Spatafora *et al.*, 2016).

The Bayesian supermatrix phylogeny is in near-total agreement with the ML phylogeny in resolving the relationships of the basal fungi in our dataset (**Figure 4.4**). The relationship between Chytridiomycota and Neocallimastigomycota in the Bayesian phylogeny mirrors that seen in the ML phylogeny, with all branches receiving maximum support as monophyletic with a Bayesian posterior probability (henceforth abbreviated to PP) equal to 1 (**Figure 4.4**). The Zoopagomycota are monophyletic with full support, with a topology matching the ML phylogeny with strong branch support throughout (**Figure 4.4**). There is also a close association between the three Mucoromycota subphyla; Glomeromycota branches earlier in the Bayesian phylogeny than in the ML phylogeny, which receives maximum support in the Bayesian phylogeny, and the sister relationship between Mucoromycotina and *Mortierella elongata* receives strong support (0.94 PP) in the Bayesian phylogeny (**Figure 4.4**). Both the ML and Bayesian place the Mucoromycota as the basal phylum that is most closely related to Dikarya (**Figure 4.4**).

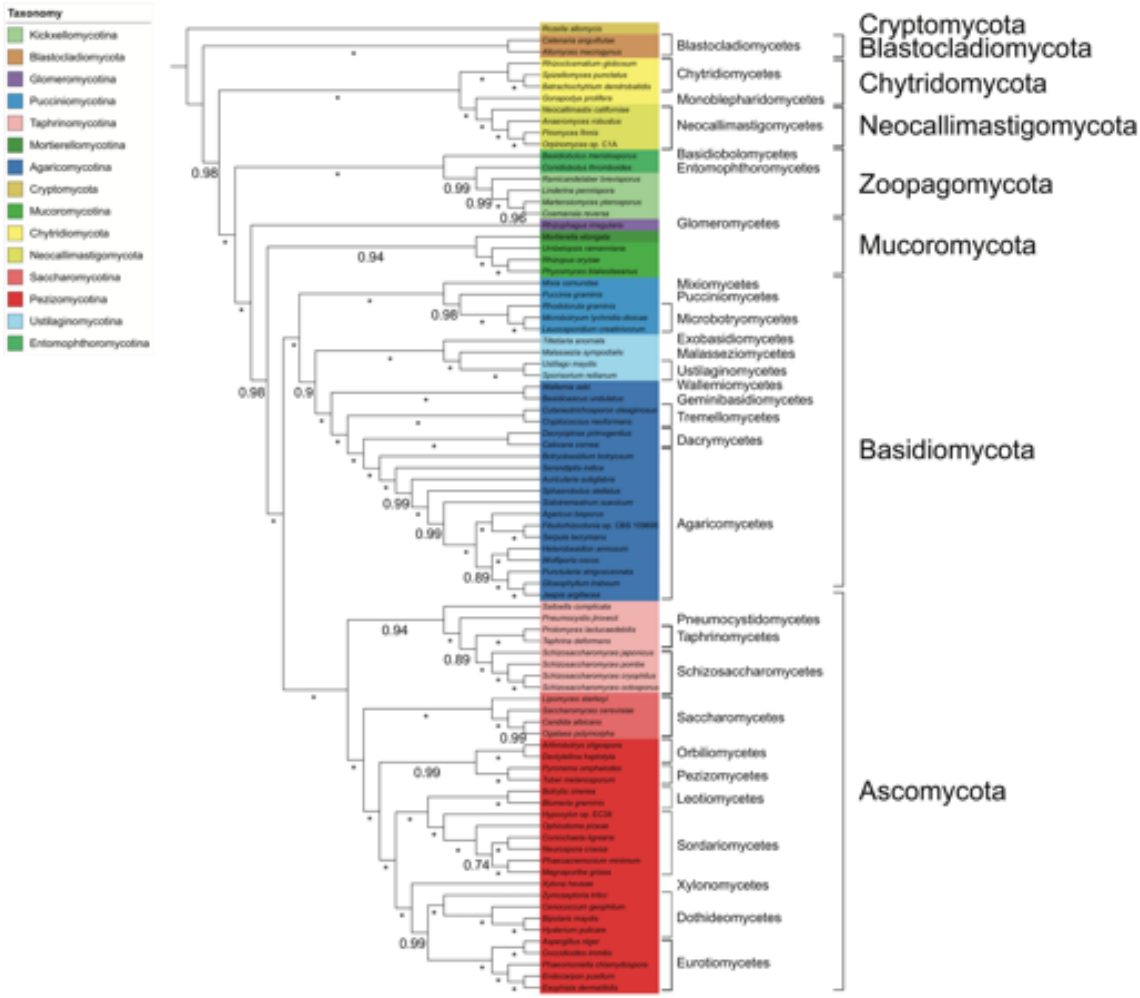


Figure 4.4. Bayesian phylogeny of 84 fungal species from a 8,529-character superalignment derived from 72 ubiquitous fungal COG families sampled in Gblocks using PhyloBayes MPI with a CAT+GTR model. Posterior probabilities shown on branches with a burn-in of 1,000 trees. Maximum posterior probability support designated with an asterisk (*).

4.2.1.3.2 Basidiomycota

In the ML phylogeny, the 3 subphyla within Basidiomycota are fully resolved with maximum BP, with 84% BP for the placement of Ustilagomycotina and Pucciniomycotina as sister clades (**Figure 4.3**). *Basidioascus undulatus* and *Wallemia sebi* branch at the base of Agaricomycotina with maximum BP, while the other classes with the subphyla are all fully supported. There is also high support (88% BP) for the placement of Tremellomycetes as sister to Dacrymycetes and Agaricomycetes (**Figure 4.3**). The Tremellomycetes, including *Cryptococcus neoformans*, are monophyletic. The Dacrymycetes are also monophyletic with maximum BP. The forest saprophyte *Botryobasidium botryosum* is placed at the base of the Agaricomycetes, which has some strong intra-clade resolution with weaker branch supports towards the tips of the clade

(**Figure 4.3**). *Malassezia sympodialis*, a commensal fungus of humans and animals, is placed at the base of the Ustilagomycotina. The Exobasidiomycetes species *Tilletiaria anomala* branches between *Malassezia sympodialis* and the Ustilagomycetes. The Pucciniomycotina are monophyletic with full support (**Figure 4.3**). The most highly represented Pucciniomycotina class, the Microbotryomycetes, are monphyletic with 69% BP (**Figure 4.3**).

The Bayesian phylogeny reflects the ML phylogeny in its resolution of the Basidiomycota as monophyletic with full support (**Figure 4.4**). The phylogeny places Pucciniomycotina at the base of the phylum with maximum support. Resolution of branches within Pucciniomycotina are substantially improved under Bayesian phylogeny (**Figure 4.4**). There is high support (0.9 PP) for a sister relationship between Ustilagomycotina and Agaricomycotina (**Figure 4.4**). The Exobasidiomycetes species *Tilletiaria anomala* now branches at the base of the Ustilagomycotina, which is resolved with maximum PP. There is maximum support for the placement of *Malassezia sympodialis* as sister to the Ustilagomycetes, which are monphyletic (**Figure 4.4**). As in the ML phylogeny, *Basidioascus undulatus* and *Wallemia sebi* branch at the base of Agaricomycotina with maximum support, while the other classes with the subphyla all have maximum support and have similar topology under Bayesian analysis. There is a large improvement in the support of branches in the Agaricomycotina in the Bayesian phylogeny relative to the ML phylogeny (**Figure 4.4**).

4.2.1.3.3 Ascomycota

Both the ML and Bayesian supermatrix phylogenies display near-identical topologies for the Ascomycota, and Bayesian analysis shows stronger support for some branches towards the tips of the phylogeny than the ML phylogeny does (**Figures 4. & 4.4**). The 3 subphyla within Ascomycota are fully resolved, with maximum BP support for Saccharomycotina and Pezizomycotina and 79% BP for the monophyly of Taphrinomycotina in the ML phylogeny (contrast with 0.94 PP for the monophyly of Taphrinomycotina in the Bayesian phylogeny; **Figures 4.3 & 4.4**). The placement of Taphrinomycotina as an ancestral clade within Ascomycota is fully supported, and within Taphrinomycotina there is high support (77% BP / 0.89 PP) for a sister relationship between Schizosaccharomycetes and Taphrinomycetes. 6 of the 7 classes within Pezizomycotina in our dataset with 2 or more representatives (i.e. all bar Xylonomycetes) are monophyletic, most of which receive maximum BP and/or PP support. Many of the

relationships between classes are also well-supported in both phylogenies, with lower support (67% BP) for a sister relationship between the Xylonomycetes species *Xylono heveae* and the Eurotiomycetes class in the ML phylogeny; in the Bayesian phylogeny *Xylono heveae* branches sister to a clade containing Dothideomycetes and Eurotiomycetes with maximum PP support (**Figures 4.3 & 4.4**). The Dothideomycetes are monophyletic in both phylogenies and branch into two clades with high support under both ML and Bayesian reconstruction (**Figures 4.3 & 4.4**). The Orbiliomycetes and Pezizomycetes are placed as the most basal Pezizomycotina classes; with strong support (94% BP / 0.99 BP) for a sister relationship (**Figures 4.3 & 4.4**). The Leotiomycetes and Sordariomycetes are also placed as a sister clades with maximum support in both phylogenies. The major difference in the resolution of the Sordariomycetes between the supermatrix phylogenies is the stronger branch supports within the order under Bayesian analysis (**Figures 4.3 & 4.4**).

4.2.2 Parsimony supertree phylogenomic analysis of fungi

The most common supertree methods for reconstructing genome phylogenies are grounded in parsimony methods, in which changes to character states (i.e. evolutionary events such as presence of a given taxon in a tree or even a tree branch) are calculated and phylogeny is reconstructed using as little state changes as possible. The first supertree construction method to see widespread use in large-scale phylogenetic and phylogenomic analysis was the matrix representation with parsimony (MRP) method. MRP, which was developed independently by Baum (1992) and Ragan (1992), enables the use of source phylogenies with overlapping or missing taxa in generating a consensus phylogeny (Baum, 1992; Ragan, 1992). The method generates a matrix (referred to as a Baum-Ragan matrix) where each column represents one internal branch in each given source phylogeny such that the number of columns within the matrix is equal to the number of internal branches across all source phylogenies, and assigns a score of 1 to taxa from a given source phylogeny *P* which are present in the clade defined by internal branch *A*, 0 to taxa present in *P* but not within the clade defined by *A*, and ? to taxa that are not present in *P* (Creevey and McInerney, 2009). The Baum-Ragan matrix is then subject to parsimony analysis, with equal weighting given to each source phylogeny, and reconstructs the supertree phylogeny with the minimum of evolutionary changes required which includes all taxa represented across all source phylogenies. Similar parsimony methods, most

notably gene tree parsimony (Slowinski and Page, 1999), extend MRP to include source phylogenies containing duplicated taxa, however we do not cover such methods in this subsection. Parsimony-based supertree methods like MRP are generally quite accurate in reconstructing phylogeny for large datasets, although some issues have been observed (which we discuss in later sections of this chapter).

4.2.2.1 Matrix representation with parsimony analysis in fungal phylogenomics

Many phylogenomic analyses of fungi have used parsimony methods. The first large-scale phylogenomic analysis of fungi to use MRP in supertree reconstruction was by Fitzpatrick *et al.* (2006), who carried out a phylogenomic reconstruction of fungi using 42 genomes from Dikarya and the zygomycete *Rhizopus oryzae* using both supertree and supermatrix methods (Fitzpatrick *et al.*, 2006). Using a random BLASTp approach to identify homologous gene families, where randomly selected query sequences are sequentially searched against a full database and then both query sequences and homologs (if any) are sequentially removed from the database, Fitzpatrick *et al.* (2006) utilized 4,805 single-copy gene phylogenies for MRP supertree reconstruction using the software package CLANN (Creevey and McInerney, 2005, 2009). The MRP phylogeny resolved the Pezizomycotina and Saccharomycotina subphyla within Ascomycota and inferred the Sordariomycetes and the Leotiomycetes as sister classes within Pezizomycotina. The MRP phylogeny also resolved two major clades within the Saccharomycotina; a monophyletic clade of species that translate the codon CTG as serine instead of leucine (the "CTG clade"), and a grouping of species that have undergone whole genome duplication (the "WGD clade") and their closest relatives. The authors compared the MRP phylogeny with a maximum-likelihood supermatrix phylogeny reconstructed using 38,000 characters from 153 gene families (as detailed in the previous subsection); both were highly congruent with conflict only in the placement of the sole Doethideomycetes species represented, *Stanonospora nodurum*. The authors also complemented their MRP phylogeny with two other supertree methods implemented in CLANN; a most similar supertree analysis (MSSA) method phylogeny which was identical to the MRP supertree (Creevey *et al.*, 2004) and an average consensus (AV) method phylogeny based on branch lengths (Lapointe and Cucumel, 1997), which the authors believed to suffer from long-branch attraction in the erroneous placement of some species within the WGD clade in

Saccharomycotina (Fitzpatrick *et al.*, 2006). A follow-up analysis to Fitzpatrick *et al.* (2006) by Medina *et al.* (2011) using 103 genomes was extended to include multi-copy gene families using the gene tree parsimony (GTP) method, and successfully resolved the major groupings within the fungal kingdom (Medina, Jones and Fitzpatrick, 2011). Using both a random BLASTp and a Markov Clustering Algorithm (MCL)-based approach with varying inflation values to identify orthologous gene families, the authors used as many as 30,012 single and paralogous gene phylogenies as input for supertree reconstruction.

As a follow-up to the supertree reconstructions of the fungal kingdom by Fitzpatrick *et al.* (2006) and Medina *et al.* (2011), we ran supertree analysis for 84 fungal species using MRP and AV methods and source phylogenies identifited *via* a random BLASTp approach described below.

4.2.2.2 Phylogenomic reconstruction of 84 fungal species from 8,110 source phylogenies using MRP and AV supertree methods

Following Fitzpatrick *et al.* (2006), families of homologous protein sequences within our 84-genome dataset were identified using BLASTp with an e-value cutoff of 10^{-20} by randomly selecting a query sequences from our database, finding all homologous sequences *via* BLASTp (Camacho *et al.*, 2009), and removing the entire family from the database before reformatting and repeating. 12,964 single-copy gene families, which contained no more than one homolog from 4 or more taxa, were identified. Each single-copy gene family was aligned in MUSCLE, and conserved regions of each alignment were sampled using Gblocks with the default parameters (Castresana, 2000; Edgar, 2004). Sampled alignments were tested for phylogenetic signal using the PTP test as implemented in PAUP* with 100 replicates (Faith and Cranston, 1991; Swofford, 2002). 8,110 sampled alignments which retained character data after Gblocks filtering and passed the PTP test were retained for phylogenomic reconstruction. 8,110 approximately-maximum-likelihood gene phylogenies were generated with FastTree, using the default JTT+CAT protein evolutionary model (Price, Dehal and Arkin, 2010). All 8,110 single-copy gene phylogenies were used to generate a matrix representation with parsimony (MRP) supertree using CLANN, with 100 bootstrap replicates (Creevey and McInerney, 2005). To complement the MRP supertree, an average consensus (AV) supertree was generated from the same input dataset in CLANN, with 100 bootstrap replicates. Both

supertrees were visualized in iTOL and annotated according to the NCBI's taxonomy database. Both supertrees were rooted at *Rozella allomycis* (**Figures 4.5 & 4.6**).

4.2.2.3 MRP phylogenomic analysis of 84 fungal species is highly congruent with supermatrix phylogenomic analyses

We reconstructed the overall phylogeny of 8,110 single-copy source phylogenies from our 84-genome dataset using an MRP supertree method analysis as implemented in CLANN (**Figure 4.5**). MRP supertree reconstruction of the fungal kingdom recovers the majority of the eight fungal phyla in our dataset and is effective in resolving the Dikarya. However, there is poorer resolution of some of the basal phyla due to smaller taxon sampling perhaps having a negative influence on the distribution of basal taxa within our source phylogenies (we return to this in Section 3). Overall our MRP analysis is highly congruent with our supermatrix phylogenies detailed above, with some variation in the placement and resolution in some branches. We discuss the results of our MRP analysis for the basal fungal lineages and both Dikarya phyla and note some of the congruences and incongruences where noteworthy with our supermatrix phylogenies (**Figures 4.3–4.5**).

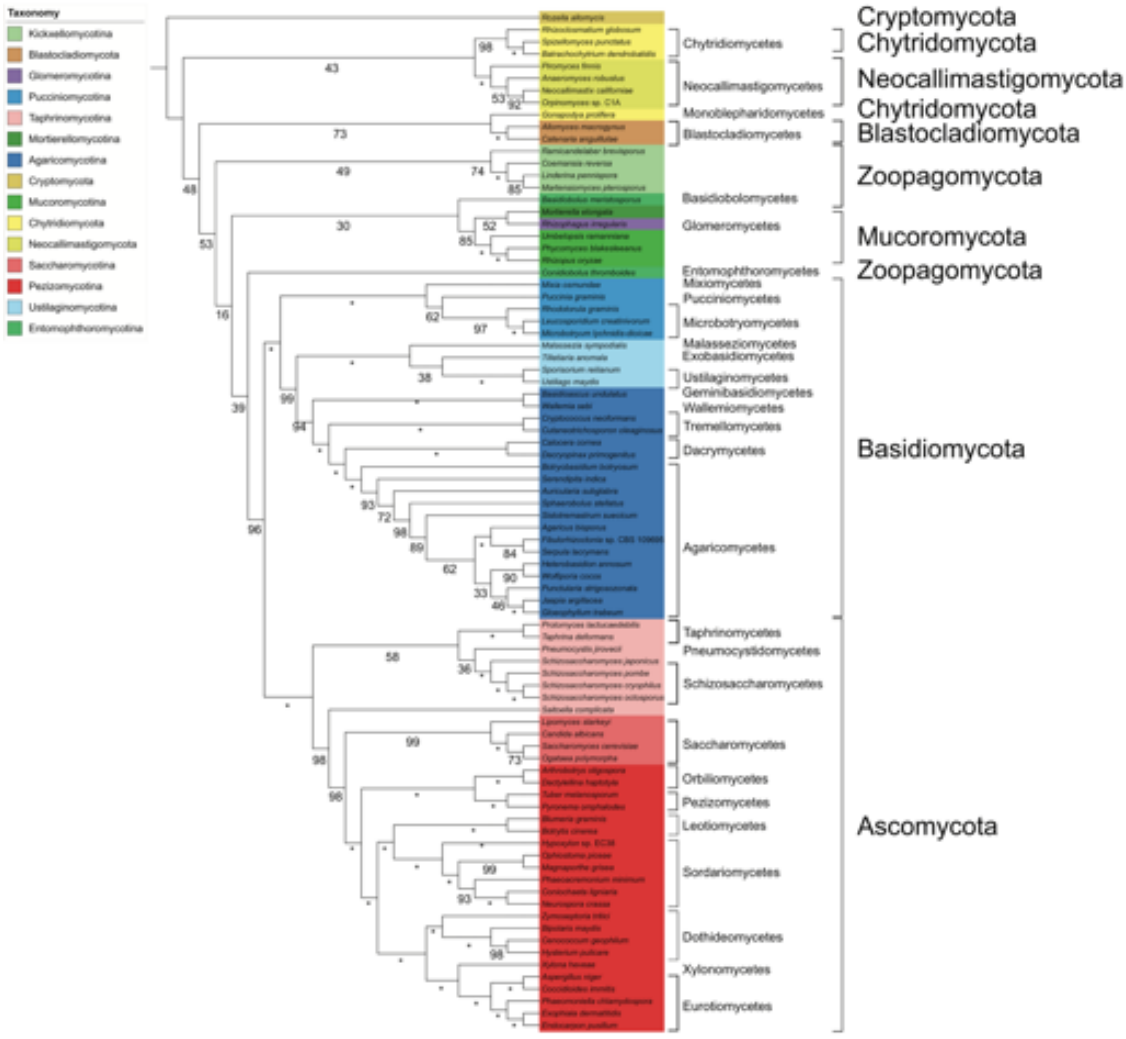


Figure 4.5. Matrix Representation with Parsimony (MRP) phylogeny of 84 fungal species derived from 8,110 source phylogenies. Bootstrap supports shown on branches. Maximum bootstrap support designated with an asterisk (*).

4.2.2.3.1 Basal fungi

After rooting at *Rozella allomycis*, the Neocallimastigomycota and Chytridiomycota (bar *Gondpodya prolifera*) emerge as the earliest-diverging fungal lineages. *G. prolifera* branches basal to the Blastocladiomycota with 73% BP (**Figure 4.5**). This arrangement of the Neocallimastigomycota, Chytridiomycota and Blastocladiomycota has poor support in general (43% BP for a sister relationship between Neocallimastigomycotina and 4 Chytridiomycota species), however with the exception of the aforementioned placement of *G. prolifera* the individual phyla receive maximum or near-maximum support as monophyletic (**Figure 4.5**). Zoopagomycota is paraphyletic in our MRP phylogeny; a monophyletic Kickxellomycotina clade receives 74% BP support (**Figure 4.5**), while as in the supermatrix phylogenies (**Figures 4.3 & 4.4**)

Entomophthoromycotina is paraphyletic. In our MRP analysis, *Basidiobolus meristosporus* branches at the base of Mucoromycota and *Conidiobolus thromboides* branches at the base of Dikarya, but those relationships are poorly supported (30% and 39% BP, respectively; **Figure 4.5**). The Glomeromycotina species *Rhizopagus irregularis* branches sister to the Mortierellomycota representative *Morteriella elongata* with weak support (52% BP), but Murocomycota (the placement of Glomeromycotina, Mortierellomycota and Mucoromycotina) receives higher support (85% BP). The monophyly of Mucoromycotina is also fully supported (**Figure 4.5**). Overall many of the associations between basal phyla we observed in our supermatrix phylogenies are present in our MRP analysis as well, however the overall placement of the basal fungal lineages varies between supermatrix and MRP analyses, such as the placement of Blastocladiomycota as a later-diverging clade than either Chytridiomycota or Neocallimastigomycota under MRP supertree analysis (**Figures 4.3–4.5**).

4.2.2.3.2 Basidiomycota

The Basidiomycota are recovered with maximum support in our MRP phylogeny (**Figure 4.5**). The Pucciniomycotina emerge as the most basal subphylum with maximum support, with *Mixia osmundae* branching at the base of the subphylum and *Puccinia graminis* placed as sister to the Microbotryomycetes (who are monophyletic with 97% BP). This reflects the topology of Pucciniomycotina seen in our supermatrix phylogenies (**Figures 4.3–4.5**). The Ustilagomycotina and Agaricomycotina branch as sister subphyla with 99% BP and both are monophyletic; the former is fully supported at the branch level and the latter has 94% BP. *Malassezia sympodialis* is placed at the base of Ustilagomycotina, reflecting the resolution of the Ustilagomycotina under ML supermatrix analysis (**Figures 4.3 & 4.5**). In the Agaricomycotina, *Wallemia sebi* and *Basidioascus undulatus* branch at the base of the subphylum with maximum support. The three larger classes from Agaricomycotina in our dataset (Agaricomycetes, Dacrymycetes, Tremellomycetes) are all monophyletic and are recovered with maximum support (**Figure 4.5**). The MRP phylogeny of the Basidiomycota is highly congruent overall with the supermatrix phylogenies, with comparable branch support (**Figures 4.3-4.5**).

4.2.2.3.3 Ascomycota

Our MRP phylogeny supports the Ascomycota as a monophyletic group with maximum BP (**Figure 4.5**). There is greater support along many deeper branches in the Ascomycota in our MRP phylogeny than in our ML supermatrix phylogeny and support is comparable with our Bayesian phylogeny; we ascribe this to a larger abundance of smaller source phylogenies containing closely-related Ascomycotina species in our dataset (**Figure 4.3-4.5**). Taphrinomycotina emerge as the earliest-diverging lineage but is paraphyletic; *Saitoella complicata* branches as an intermediate between Taphrinomycotina and a Saccharomycotina-Pezizomycotina clade with 98% BP, while the remaining members are monophyletic with weak support (58% BP). *Pneumocystis jirovecii* is placed as a sister taxon to Schizosaccharomycetes in our MRP analysis with weak support (36% BP); in the supermatrix phylogenies it was sister to Taphrinomycetes. The Taphrinomycetes and Schizosaccharomycetes themselves are monophyletic with maximum BP (**Figure 4.5**). The Saccharomycotina are monophyletic with 99% BP (**Figure 4.5**). The six larger classes (i.e. all bar Xylonomycetes) in our dataset from Pezizomycotina are all supported as monophyletic and receive maximum BP, with Pezizomycetes and Orbiliomycetes branching as the basal sister clades (**Figure 4.5**). The MRP phylogeny mirrors Bayesian supermatrix reconstruction in placing a single origin for three classes (Xylonomycetes, Eurotiomycetes and Dothideomycetes) with maximum support (**Figures 4.4 & 4.5**). As in both supermatrix phylogenies, Dothideomycetes are split into two clades with high or maximum support. In the Sordariomycetes, MRP analysis reflects the ML supermatrix phylogeny in placing *Hypoxylon* sp. EC58 at the base of the class (**Figures 4.3 & 4.5**). The MRP phylogeny of the Ascomycota is highly congruent with both of our supermatrix phylogenies with comparable branch supports, which is aided by the broad range of genomic data available for the phylum (**Figures 4.3–4.5**).

4.2.2.4 Average Consensus phylogenomic reconstruction of 84 fungal species is affected by long-branch attraction artefacts

To complement our MRP phylogeny, we generated an average consensus (AV) method supertree phylogeny (**Figure 4.6**) using the same set of input phylogenies as implemented in CLANN following Fitzpatrick *et al.* (2006). AV phylogeny infers phylogeny based on the branch lengths of source phylogenies, by computing the average value of the path-length matrices associated with said source phylogenies, and then using

a least-squares method to find the source matrix closest to this average value (Lapointe and Cucumel, 1997). The tree that is associated with this source matrix is the average consensus phylogeny for the total set of source phylogenies, and the method is thought to work best with a set of source phylogenies of similar size (Lapointe and Cucumel, 1997). Our AV phylogeny was rooted at *Rozella allomycis* (**Figure 4.6**). Given the results we obtained from our AV phylogeny we believe that the method is susceptible to long-branch attraction (Felsenstein, 1978), as reported by Fitzpatrick *et al.* (2006). Long-branch attraction occurs when two very divergent taxa or clades with long branch lengths (i.e. many character changes occurring over time) are inferred as each other's closest relative due to convergent evolution of a given character (e.g. amino acid substitution), and is a common problem in parsimony and distance-based methods (Felsenstein, 1978; Bergsten, 2005). In the AV phylogeny we recovered the two Blastocladiomycota species in our dataset within a large paraphyletic Pezizomycotina clade (**Figure 4.6**). Additionally, the Ascomycota are paraphyletic; one clade containing two Pezizomycotina classes (Pezizomycetes and Orbiliomycetes), the Taphrinomycotina and the Saccharomycotina species *Lipomyces starkeyi* places at the base of Dikarya, while three Saccharomycotina species (including *Saccharomyces cerevisiae*) appear as a sister clade to Pucciniomycotina (**Figure 4.6**). The Agaricomycotina are also paraphyletic; Tremellomycetes and two basal Basidiomycota species (*Basidioascus undulatas* and *Wallemia sebi*) appear closer to Ustilagomycota (**Figure 4.6**). Many of the supports throughout the tree are extremely poor (almost all of the incongruences we highlighted all have <40% BP), which seems to be another effect of long-branch attraction (**Figure 4.6**). Due to the breadth of fungal taxa we have sampled for our multiple analyses, and the time-scale of the evolution of the fungal kingdom being approximately 1 billion years old, it is unsurprising that a method using branch lengths to infer a close relationship between actually distantly-related species that both have long branches, a classic example of the "Felsenstein Zone" (Bergsten, 2005; Huelsenbeck & Hillis, 1993). Ultimately, our AV phylogeny (**Figure 4.6**) seems to confirm one of the concerns of Fitzpatrick et al. (2006) in much more stark fashion; that the AV method is not appropriate for large-scale phylogenomic reconstructions containing taxa sampled from across many phyla without prior predictive analysis of the potential for long-branch attraction in such datasets (Su and Townsend, 2015).

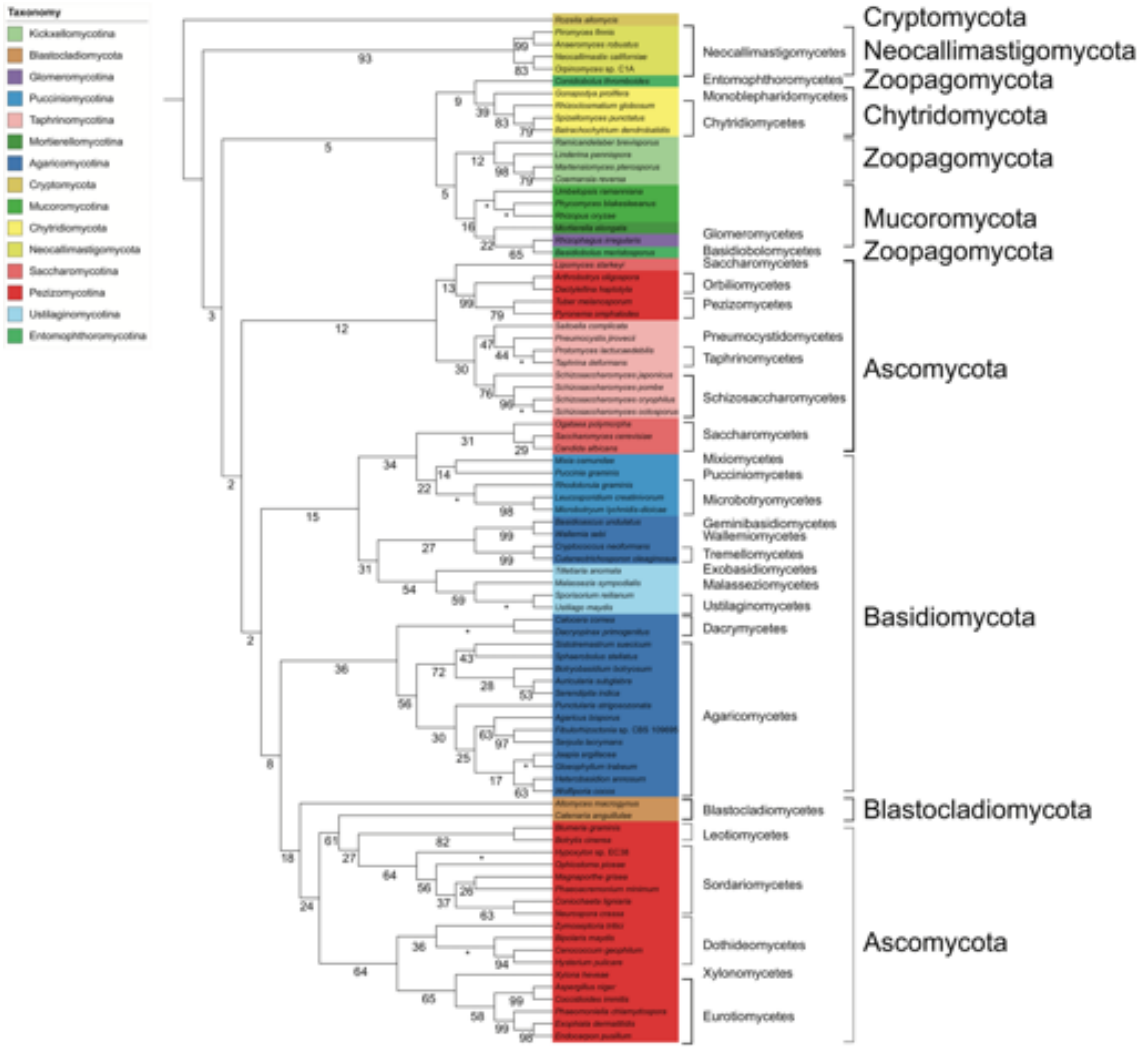


Figure 4.6. Average Consensus (AV) phylogeny of 84 fungal species derived from 8,110 source phylogenies. Bootstrap supports shown on branches. Maximum bootstrap support designated with an asterisk (*).

4.2.3 Bayesian supertree phylogenomic analysis of fungi

While parsimony-based supertree reconstructions are generally reliable, concerns have been raised in the past as to some of the underlying methodology of MRP reconstruction and the effects that factors like input tree sizes (Pisani and Wilkinson, 2002; Wilkinson *et al.*, 2004). There has long been the desire for a supertree method that infers phylogeny from source trees with more statistical rigour like Bayesian and maximum-likelihood inference methods. While Bayesian and ML analysis are the standard for supermatrix reconstruction, such methods have been difficult to implement in the past for supertree analysis due to computational limitations, most of which is down to the necessity of tree searching for the best supertree (i.e. calculating likelihoods for all possible supertrees given a set of source phylogenies).

It is only in recent years that phylogenomic inference based on ML and Bayesian methods have been implemented for supertree analysis; one such model for supertree likelihood estimation was first described by Steel & Rodrigo (2008) and then refined the following year (Steel and Rodrigo, 2008; Bryant and Steel, 2009). The Steel & Rodrigo method of likelihood estimation (henceforth referred to as ST-RF) is based on modelling the incongruences between input gene phylogenies and a corresponding unknown or provided supertree phylogeny. Two recent implementations of ST-RF ML analysis have been reported; the first a heuristic method of estimating approximate ML supertrees based on subtree pruning and regrafting implemented in the Python software L.U.St. by Akanni *et al.* (2014), and the second a heuristic Bayesian MCMC criterion by Akanni *et al.* (2015) implemented in the Python software package p4 (Foster, 2004; Akanni *et al.*, 2014, 2015). Akanni *et al.* (2015) tested the Bayesian MCMC implementation on both a large kingdom-wide metazoan dataset and a smaller Carnivora dataset; notably the analysis produced a Bayesian supertree in full agreement with both the literature on metazoan relationships and a previous MRP supertree analysis on the same dataset (Holton and Pisani, 2010).

No parametric supertree reconstruction has been carried out for the fungal kingdom to date, and with that in mind we reconstructed the phylogeny of our 84-genome dataset with the MCMC Bayesian criterion developed by Akanni *et al.* (2015) using a slightly amended gene phylogeny dataset from our MRP and AV supertree phylogenies.

4.2.3.1 Heuristic MCMC Bayesian supertree reconstruction of 84 fungal genomes from 8,050 source phylogenies

MCMC Bayesian supertree analysis was carried out on the single-copy phylogeny dataset using the ST-RF model as implemented in p4 (Foster, 2004; Steel and Rodrigo, 2008; Akanni *et al.*, 2015). As ST-RF analysis is currently only implemented in p4 for fully bifurcating phylogenies, 60 phylogenies were removed from the total single-copy phylogeny dataset, for an input dataset of 8,050 gene phylogenies. Two separate MCMC analyses with 4 chains each were ran for 30,000 generations with $\beta = 1$, sampling every 20 generations. The analyses converged after 30,000 generations, and a consensus phylogeny based on posterior probability of splits was generated from 150 supertrees sampled after convergence following Akanni et al. (2015). This consensus phylogeny was visualized in iTOL and annotated according to the NCBI's taxonomy database and rooted at *Rozella allomycis* (**Figure 4.7**).

4.2.3.2 Supertree reconstruction with a heuristic MCMC Bayesian method highly congruent with MRP and supermatrix phylogenies

Using 8,050 of the 8,110 individual gene phylogenies which we identified in our MRP supertree analysis, we have reconstructed the first parametric supertree of the fungal kingdom (**Figure 4.7**). We selected the ST-RF MCMC Bayesian supertree reconstruction method implemented in p4 for reconstruction over the heuristic method implemented in L.U.St. due to tractability issues regarding large datasets in the latter method (Akanni *et al.*, 2014, 2015). Two ST-RF analyses were carried out for 30,000 generations, and the analyses were adjudged to have converged after 20,000 generations. To construct a phylogeny from our MCMC analysis we sampled 150 trees generated after convergence and built a consensus tree in p4, where branch support values are the estimated posterior probabilities of a given split (i.e. bipartition) within a phylogeny (**Figure 4.7**). Our ST-RF MCMC analysis is highly congruent with both our MRP supertree phylogeny and supermatrix phylogenies, and supports the monophyly of the majority of the 8 fungal phyla in our dataset (**Figure 4.7**). Below, we detail the resolution of the basal and Dikarya lineages under ST-RF analysis.

4.2.3.2.1 Basal fungi

After rooting at *Rozella allomycis*, the Neocallimastiogmycota and Chytridiomycota (except *Gonapodya prolifera*) form a sister group relationship with maximum PP (**Figure 4.7**). The Blastocladiomycota emerge after this branch, and the Chytridiomycota species *Gonapodya prolifera* branches as sister to the phylum with maximum PP (**Figure 4.7**). There is weak support (0.51 PP) for a monophyletic clade containing both former zygomycetes phyla Zoopagomycota and Mucoromycota as sister clades (**Figure 4.7**). Notably, unlike MRP and supermatrix analysis, ST-RF phylogeny places the Entomophthoromycotina as monophyletic but with very weak support (0.38 PP). There is also weak support for the placement the Entomophthoromycotina as basal within Zoopagomycota. Kickxellomycotina are monophyletic with maximum support. The monophyly of Mucoromycota is fully supported, with *Rhizophagus irregularis* (Glomeromycotina) and *Mortierella elongata* (Mortierellomycotina) branching as sister taxa (**Figure 4.7**).

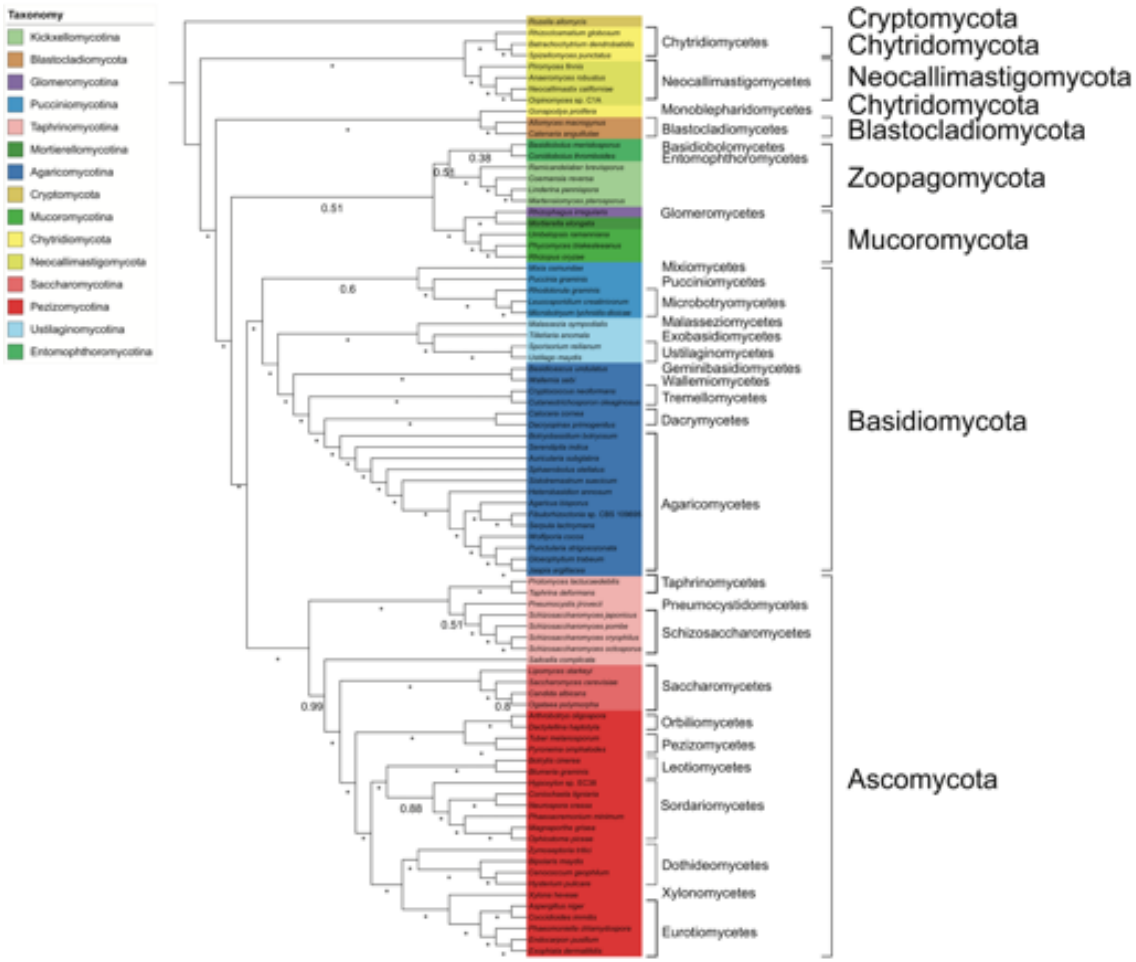


Figure 4.7. MCMC Bayesian supertree phylogeny of 84 fungal species derived from 8,060 fully bifurcating source phylogenies. Phylogeny generated in p4 using ST-RF model of maximum-likelihood supertree estimation running for 30,000 generations with $\beta = 1$. Posterior probabilities of bipartition(s) within 150 trees sampled after convergence shown on branches. Maximum posterior probability support designated with an asterisk (*).

4.2.3.2.2 Basidiomycota

The Basidiomycota are supported as a monophyletic group with maximum PP (**Figure 4.7**). There is weak support for the monophyly of Pucciniomycotina (0.6 PP), however the deeper branches within the subphyla are all fully supported and their topology reflects both the MRP supertree and ML supermatrix phylogenies discussed above (**Figures 4.3, 4.5, 4.7**). There is full support for a sister relationship between Ustilaginomycotina and Agaricomycotina, and both these subphyla are fully supported. In Ustilagoinomycotina, *Malassezia sympodialis* is the basal species with maximum support (**Figure 4.7**), as in our supermatrix and MRP supertree phylogenies. The topology of the Agaricomycotina is nearly identical on the class level to both the MRP and

supermatrix phylogenies; with *Basidioascus undulatus* and *Wallemia sebi* branching as basal species, the Tremellomycetes forming a monophyletic intermediate clade, and a fully-supported sister relationship between the Dacrymycetes and the Agaricomycetes (**Figure 4.7**).

4.2.3.2.3 Ascomycota

The monophyly of the Ascomycota is supported with maximum PP, as is the monophyly of two of the three subphyla in Ascomycota (**Figure 4.7**). Taphrinomycotina is paraphyletic as in the MRP phylogeny, with *Saitoella complicata* branching sister to Saccharomycota with near-maximum support (0.99 PP) and the remaining Taphrinomycotina species are placed as a monophyletic clade with maximum PP (**Figures 4.5 & 4.7**). The Taphrinomycetes branch at the base of the Taphrinomycotina clade, and there is weak support (0.51 PP) for the placement of *Pneumocystis jirovecii* as sister to the Schizosaccharomycotina (**Figure 4.7**). The Saccharomycotina are fully supported as monophyletic (1.0 PP) with *Lipomyces starkyei* placed at the base of the subphyla. The monophyly of the Pezizomycotina is also fully supported and there is maximum support for the monophyly of the six larger-represented classes within the subphylum (**Figure 4.7**). Additionally, the relationships between the individual classes within Pezizomycotina is identical to the topology seen in both the MRP supertree phylogeny and the ML supermatrix phylogeny (**Figures 4.3, 4.5 & 4.7**). The Orbiliomycetes and Pezizomycetes branch as the earliest-diverging clades within Pezizomycotina with maximum PP, the Sordariomycetes and Leotiomycetes are sister classes with maximum PP and a monophyletic Dothideiomycetes-Xylonomycetes-Eurotiomycetes clade receives maximum PP (**Figure 4.7**)

4.2.4 Phylogenomics of fungi based on gene content

A common alternative to phylogenomic reconstruction using gene phylogenies is to take a "gene content" approach in which evolutionary relationships between species are derived from shared genomic content, such as the presence or absence of conserved orthologous genes (COGs) or the overall proportion of shared genes between two species, working under the assumption that species that share more of their genome are closely related (Snel, Bork and Huynen, 1999; Snel, Huynen and Dutilh, 2005). In the case of presence-absence analyses, a matrix can be constructed for the species under investigation which can then have their phylogeny reconstructed *via* parsimony methods. Analyses based on proportions of shared genes can entail the construction of distance matrices for all input species, with values equal to the inverse ratio of shared genes (i.e. if two species share 75% of their genes, their distance is 0.25), which is then used to construct a neighbour-joining phylogeny. The advantages of such approaches is the relative tractability of parsimony or distance-based gene content methods, and their potential to use more information from genomes rather than the sourcing of data from smaller sets of gene families required by supertree or supermatrix approaches (Creevey and McInerney, 2009). However the gene content approach is by its very nature a "broad strokes" approach and can ignore potentially important phylogenetic information from individual gene phylogenies such as HGT events, and assumes the same evolutionary history for missing orthologs or genomic content among species (Page and Holmes, 1998).

4.2.4.1 Gene content approaches to phylogenomics in fungi

Gene content approaches to phylogenomic reconstruction have seen application in a number of phylogenomics studies, although its greatest use predated many of the now common supertree and supermatrix methods. One of the earliest phylogenomic studies used a distance-based approach based on shared gene content to reconstruct the phylogeny of 13 unicellular species, including *S. cerevisiae* (Snel, Bork and Huynen, 1999). Another study used a weighted distance matrix approach to reconstruct the phylogeny of 23 prokaryote and eukaryote species, including *Saccharomyces cerevisiae* and partial genomic data from *Schizosaccharomyces pombe* (Tekaia, Lazcano and Dujon, 1999). The most extensive gene content-based phylogenomic reconstruction of fungi was an analysis of 21 fungal genomes and 4 other eukaryote genomes in 2006 (Kuramae *et*

al., 2006). In their analysis, the authors generated a presence-absence matrix (PAM) of 4,852 COGs in fungal genomes as a complement to a supermatrix phylogeny using 531 concatenated proteins which was reconstructed using four different methods (MP, ML, neighbour-joining and Bayesian inference). The authors reconstructed the phylogeny of all 25 genomes using this presence-absence matrix and found that the PAM phylogeny differ most in the placement of *Schizosaccharomyces pombe* within Saccharomycetes as opposed to its basal position in Ascomycetes as seen in their supermatrix reconstructions (Kuramae *et al.*, 2006).

To test the accuracy of inferring the phylogeny of a large genomic dataset using simple parsimony methods based on shared genomic content, we carried out a simple parsimony-based presence-absence matrix (PAM) phylogenomic reconstruction of 84 fungal species based on the presence of orthologs from single-copy gene families.

4.2.4.2 Phylogenomic reconstruction of 84 fungal species based on COG presence-absence matrix

A simple presence-absence matrix (PAM) was generated for 84 fungal genomes based on their representation across 12,964 single-copy gene families identified *via* the random BLASTp approach detailed in Section 2.2. Parsimony analysis of this matrix was carried out using PAUP* with 100 bootstrap replicates. The resultant consensus phylogeny generated by PAUP* was visualized using iTOL and annotated according to the NCBI's taxonomy database. The phylogeny was rooted at *Rozella allomycis* (**Figure 4.8**).

4.2.4.3 COG presence-absence matrix approach displays erroneous placement of branches within Dikarya

We generated a simple presence-absence matrix (PAM) phylogeny for the 84 fungal genomes in our dataset by checking for the presence or absence of all 84 species across the 12,964 single-copy phylogenies we generated during our supertree analyses *via* random BLASTp searches and using the PAM as input for parsimony analysis (**Figure 4.8**). The simple PAM phylogeny shows some level of congruence with the other phylogenomic analyses described here along certain branches (**Figure 4.8**). The monophyly of Neocallimastigomycota, Chytridiomycota and Blastocladiomycota all display maximum or near-maximum BP, and there is 72% BP for a sister relationship

between Chytridiomycota and Neocallimastigomycota (**Figure 4.8**). The Zoopagomycota and Mucoromycota are placed in one monophyletic clade with 82% BP, with the two Entomophthoromycotina species in our dataset branching as closely related to the Mucoromycota (**Figure 4.8**). However, some glaring conflicts with the other phylogenomic methods we carried out can be observed within the Dikarya lineage. Most notably, the Agaricomycotina and Saccharomycotina are both paraphyletic in our single copy PAM approach; for the former, *Wallemia sebi* and *Basidioascus undulatus* branch at the base of the Basidiomycota adjacent to Ustilagomycotina, while in the latter 3 of the 4 Saccharomycotina (excluding *Lipomyces starkeyi*) species branch in our dataset at the base of the Ascomycota, implying that Taphrinomycotina diverged later than Saccharomycotina (**Figure 4.8**). There is uncertain placement of clades within the Basidiomycota subphyla in particular. In the Ascomycota, the Taphrinomycotina are paraphyletic and *Saitoella complicata* branches adjacent to *L. starkeyi*. The monophyly of all six larger Pezizomycotina classes are supported, many with relatively high or even maximum BP, however there is poorer resolution of many relationships within these classes with the clearest examples being the Sordariomycetes and Eurotiomycetes (**Figure 4.8**). In short, our PAM phylogeny is able to retrieve relationships with some level of accuracy within the fungal kingdom, but the method lacks the ability to resolve some of the more divergent relationships within fungi to the degree that some of our supermatrix or supertree phylogenies have illustrated.

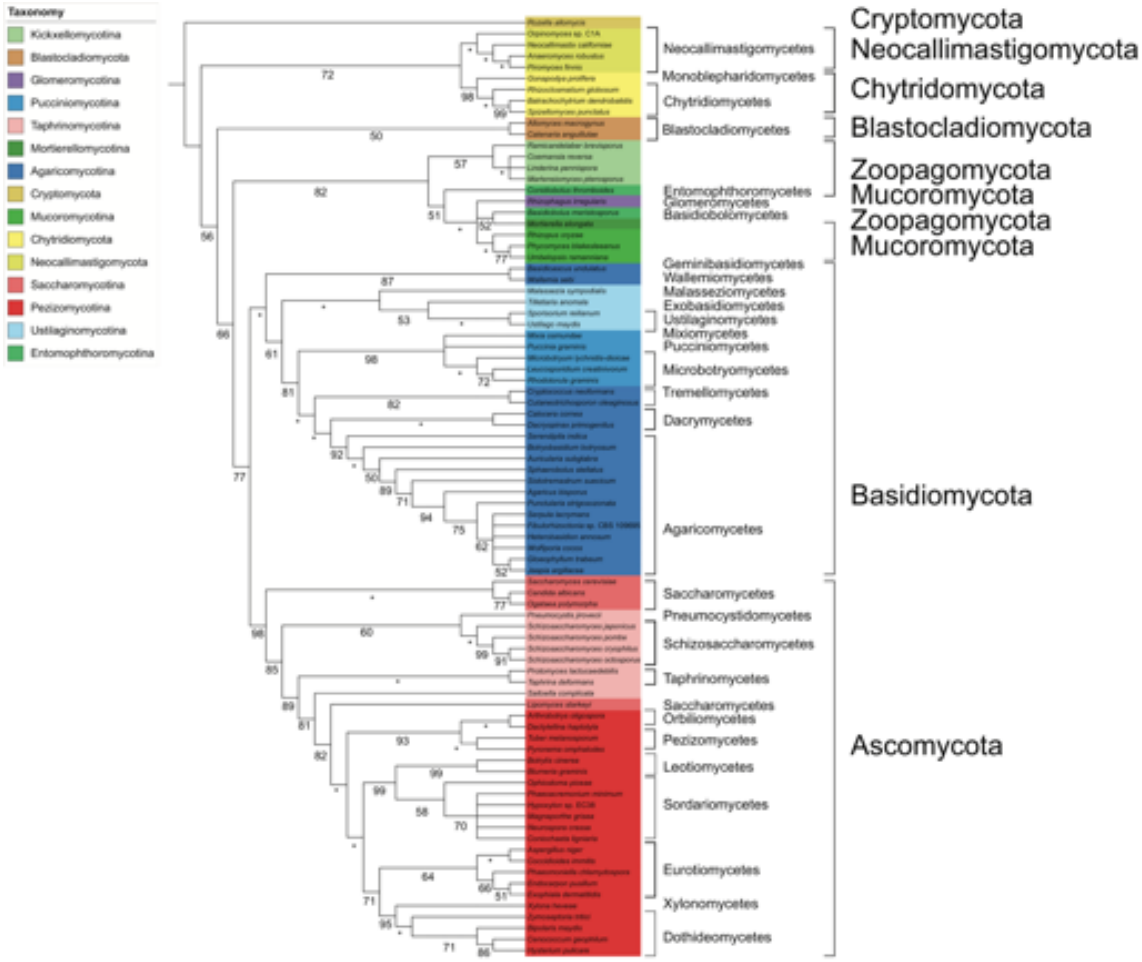


Figure 4.8. Maximum-parsimony (MP) phylogeny of 84 fungal species based on the presence of homologs from 12,964 single-copy gene families identified *via* random BLASTp searches. Bootstrap supports shown on branches. Maximum bootstrap support designated with an asterisk (*).

4.2.5 Alignment-free phylogenomic analysis of fungi

Another alternative to the alignment-based methods of phylogenomic reconstruction we have detailed above is the use of a string-based comparison of genomes to infer phylogeny, based on the assumption that under such comparisons each species should have a characteristic genomic signature that can act as a phylogenetic marker (Delsuc, Brinkmann and Philippe, 2005). Some analyses have thus used signatures such as distribution of protein folds or frequency of oligonucleotides from genetic and genomic data to infer phylogeny (Campbell, Mrázek and Karlin, 1999; Lin and Gerstein, 2000; Pride *et al.*, 2003). The most widely-used alignment-free phylogenomic method, the composition vector (CV) approach, was first implemented by Qi *et al.* (2004) who used the approach to reconstruct the phylogeny of 87 prokaryote species from 11 bacterial and

2 archaeal phyla (Qi, Wang and Hao, 2004). In their analysis, the authors detail the CV method for reconstructing phylogeny using genome-scale data, which we recount as follows:

1) Given a nucleic acid or amino acid sequence of length L in a genome, count the appearances of overlapping strings (i.e. oligonucleotides or oligopeptides) of a length K and construct a frequency vector of length 4^K for nucleic acid sequences and 20^K for amino acid sequences.
2) Subtract background noise, to account for random mutation at the molecular level, from each frequency vector to generate an overall composition vector for a given genome.
3) Calculate a distance matrix for the set of composition vectors corresponding to the set of input genomes.
4) Generate a neighbour-joining phylogeny from the distance matrix using software such as Neighbor or PAUP*.

The main advantages of the composition vector approach over traditional alignment-based methods of inferring phylogeny are the removal of artificial selection of phylogenetic markers from the process of reconstruction (the only variable in the method is K, the length of overlapping oligopeptides), and the relative speed with which the approach can infer phylogeny for large datasets over alignment-based supertree or supermatrix methods. Hence, it may be useful for quick phylogenomic identification of newly sequenced genomes against published data and as an independent verification step of previous alignment-based phylogenetic or phylogenomic analysis (Wang *et al.*, 2009). On that point however, interpreting the accuracy or otherwise of CV phylogenomic reconstructions is generally dependent on prior knowledge of the phylogeny of given taxa derived from alignment-based phylogenetic or phylogenomic analyses. An approach to inferring phylogeny based on nucleotide or amino acid composition may also be susceptible to compositional biases, and there has not been to the best of our knowledge a rigorous analysis of the potential effect these may have on accuracy of phylogenomic inference, as there have been for the supertree or supermatrix methods referred to above.

4.2.5.1 Composition vector method phylogenomics of fungi

Many of the phylogenomic analyses using the CV method have analysed large prokaryotic datasets or broad global datasets sampled from many phyla or kingdoms

across the three domains of life, whose phylogenies were recovered with quality comparative to alignment-based phylogenomic analyses. The most extensive application of the composition vector approach in fungal phylogenomics was an 85-genome analysis by Hao *et al.* (2009) using a CV implementation in the software program CVTree (Qi, Luo and Hao, 2004; Wang *et al.*, 2009). For their analysis, Wang *et al.* (2009) reconstructed the phylogeny of the fungal kingdom using 81 genomes from 4 fungal phyla (Basidiomycota, Ascomycota, Chytridiomycota and Mucoromycota) as well as the microsporidian *Encephalitozoon cuniculi* and three eukaryotic outgroup taxa. The authors described the resolution of both the Basidiomycota and Ascomycota in detail in their analysis; the three subphyla within Basidiomycota were recovered but with poor bootstrap support due to issues with taxon sampling (only 12 Basiomycota species had genomic data at the time of the analysis), while the main focus of the authors analysis was on the resolution of 65 Ascomycota species. Within the Ascomycota the Taphrinomycota (represented by three *Schizosaccharomyces* species) were fully resolved and in the Saccharomycotina the two clades described by Fitzpatrick *et al.* (2006), the CTG clade and the WGD clade, were also recovered. CV reconstruction recovered 4 classes within Pezizomycotina; the Dothideomycetes and Eurotiomycetes were placed as sister taxa with maximum support, as were the Sordariomycetes and Leotiomycetes.

To complement our phylogenomic analyses based on source gene phylogenies or identification of shared orthologs, we carried out alignment-free analysis of 84 fungal species using the composition vector method as implemented in CVTree.

4.2.5.2 Phylogenomic reconstruction of 84 fungal species using the CV approach

Composition vector analysis was carried out on 84 genomes using CVTree with $K = 5$ (Qi, Luo and Hao, 2004). We selected $K = 5$ as the best compromise of both computational requirements and resolution power. As the CV method does not generate bootstrapped phylogenies, we generated 100 bootstrap replicates of our 84-genome representative dataset using bespoke Python scripting, and ran composition vector analysis on each replicate dataset (Zuo *et al.*, 2010). 100 replicate neighbour-joining phylogenies were calculated from their corresponding CVTree output distance matrices using Neighbor (Felsenstein, 1989). The majority-rule consensus phylogeny for all 100 composition vector replicate trees was generated using Consense (Felsenstein, 1989), and

was visualized in iTOL, and annotated according to the NCBI's taxonomy database. The phylogeny was rooted at *Rozella allomycis* (**Figure 4.9**).

4.2.5.3 Composition vector phylogenomic reconstruction of 84 fungal species is congruent with alignment-based methods

We carried out composition vector method phylogenomic reconstruction of our 84-genome dataset to complement the alignment-based and genomic content methods we detailed above (**Figure 4.9**). Our composition vector analysis displays adequate levels of taxonomic congruence with our supermatrix and supertree analyses detailed in previous sections, supporting all the monophyly of each major fungal phylum and many of the subphyla within (**Figure 4.9**). There are however some variations in topology and support between the basal lineages and within the Pezizomycotina subphylum in our CV phylogeny compared to our supermatrix and supertree phylogenies.

4.2.5.3.1 Basal fungi

After rooting at *Rozella allomycis*, the Neocallimastigomycota emerge as the earliest-diverging fungal lineage (**Figure 4.9**). The monophyly of Neocallimasigomycetes is also fully supported. Monophyletic Blastocladiomycota and Chytridomycota clades branch as sister phyla with 62% BP. The monophyly of Blastocladiomycota receives maximum support, and notably unlike our MRP and supermatrix phylogenies *Gonapodya prolifera* branches within the Chrytridomycota with 86% BP (**Figures 4.3–4.5, 4.9**). In contrast to both supermatrix phylogenies and the MRP and ST-RF phylogenies, and like the AV and PAM phylogenies the two zygomycetes fungal phyla (Mucoromycota, Zoopagmycota) are placed within one monophyletic clade with 79% BP (**Figures 4.3–4.9)**. Kickxellomycotina are monophyletic with 95% BP, and branch at the base of this Zoopagomycota-Mucoromycota clade. Resolution of the relationship between the rest of the former zygomycetes subphyla is harder to ascertain and has weaker support; the two Entomophthoromycotina species branch distant from each other with *Basidiobolus meristosporus* branching within Mucoromycota adjacent to Mortierellomycotina and *Conidiobolus thromboides* branching beside the Glomeromycotina species *Rhizophagus irregularis*, similar to what is seen under PAM phylogenomic analysis (**Figures 4.8–4.9**). Like the MRP phylogeny (**Figure 4.5**),

Rhizopus irregularis is within a paraphyletic Mucoromycota clade instead of at the base of the Dikarya as seen in the supermatrix phylogenies (**Figures 4.3, 4.4 & 4.9**).

4.2.5.3.2 Basidiomycota

Pucciniomycotina is placed as the earliest-diverging subphylum within Basidiomycota with 52% BP, and the Ustilagomycotina and Agaricomycotina subphyla are sister clades with 95% BP (**Figure 4.9**). The most-represented class within the Pucciniomycotina, the Microbotryomycetes, are monophyletic with 65% BP (**Figure 4.9**), while unlike the rest of our phylogenies discussed above *Puccinia graminis* is placed as the most basal species within Pucciniomycotina. Within the Ustilaginomycotina, *Malassezia sympodialis* are placed as the basal lineage sister to the Exobasidiomycetes representative *Tilletieria anomala* similar to its position under ML supermatrix reconstruction and MRP reconstruction (**Figures 4.3, 4.5 & 4.9**). The Agaricomycetes are monophyletic with 84% BP, with varying support for relationships within the class but a topology identical to both supermatrix phylogenies and MRP phylogeny with the exception of the placement of Tremellomycetes within a monophyletic ancestral branch adjacent to *Basidioascus undulatus* and *Wallemia sebi* (**Figures 4.3–4.5, 4.9**).

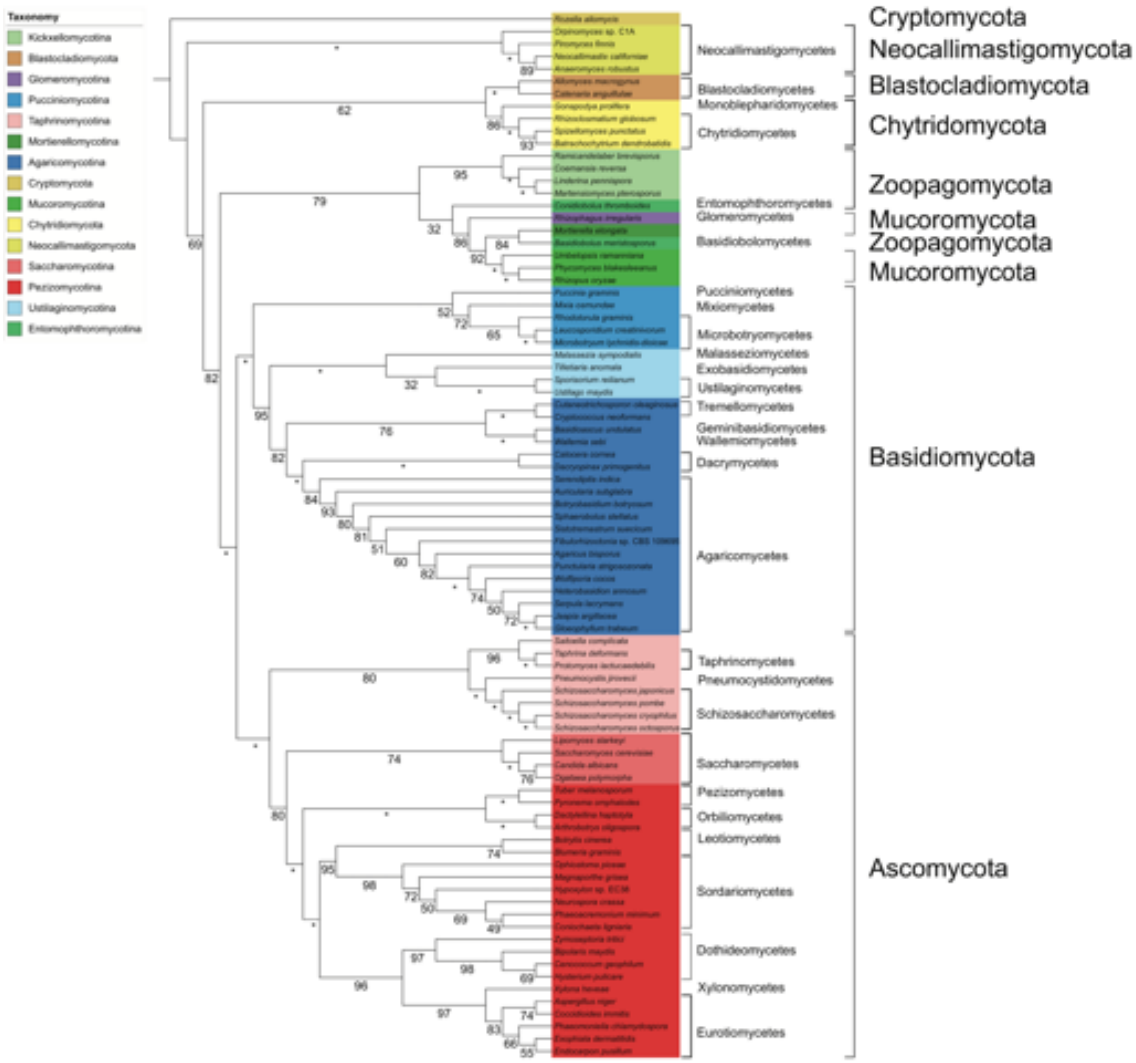


Figure 4.9. Composition vector (CV) method phylogeny of 84 fungal species generated from 100 bootstrapped replicates of an 84-genome dataset. Bootstrap supports shown on branches. Maximum bootstrap support designated with an asterisk (*).

4.2.5.3.3 Ascomycota

Within the Ascomycota, all three subphyla are resolved as monophyletic clades (**Figure 4.9**). Taphrinomycotina is placed as the most basal subphylum within Ascomycota with maximum support, while the Pezizomycotina and Saccharomycotina are sister subphyla with 80% BP (**Figure 4.9**). The Taphrinomycotina are monophyletic with 80% BP, and CV phylogeny displays maximum support for a sister relationship between *Pneumocystis jirovecii* and the Schizosaccharomycetes and near-maximum (96% BP) support for a similar relationship between *Saitoella complicata* and the two Taphrinomycetes representatives in our dataset (**Figure 4.9**). The Saccharomycotina are monophyletic with 74% support. (**Figure 4.9**) All 6 larger classes from the Pezizomycotina represented in our dataset are resolved as monophyletic. The

Orbiliomycetes and Pezizomycetes are placed as both sister subphyla and the earliest diverging Pezizomycotina clades, both with maximum BP. The Leotiomycetes and Sordariomycetes are also sister clades with 95% BP. As our MRP phylogeny, the Eurotiomycetes are placed as sister to the Xylonomycetes species *Xylona heveae* with 97% BP (**Figures 4.5 & 4.9**).

4.3 A genome-scale phylogeny of 84 fungal species from seven phylogenomic methods

There is a large degree of congruence in the resolution of the fungal kingdom in most of the phylogenomic analyses we've described in this analysis, which speaks to the quality of the genomic data we obtained from MycoCosm and the relative accuracy of the majority of the phylogenomic methods we utilized. In constructing a dataset for our analyses, we selected one representative from as many fungal orders as had been sequenced to date; this was to generate a phylogeny that was representative on the order level (though we do not focus on order phylogeny in this review) and to avoid over-representation of highly sampled taxa such as Eurotiomycetes or Saccharomycotina. Many of the best-known phylogenetic relationships within the fungal kingdom were recovered in our analyses, such as the monophyly of Dikarya as a whole (Hibbett *et al.*, 2007). However, our analyses also supports more recent studies that have attempted to resolve outstanding branches of the fungal tree of life (Spatafora *et al.*, 2016). In this section, we briefly describe the main trends seen across our seven phylogenomic reconstructions of the fungal kingdom and their congruence with previous studies, and comment on the reconstructions of both the well-studied and highly-represented Pezizomycotina subphylum and some of the newly-circumscribed basal phyla. Finally, we discuss the suitability of the phylogenomic methods we have described and applied in this review for future fungal systematics studies.

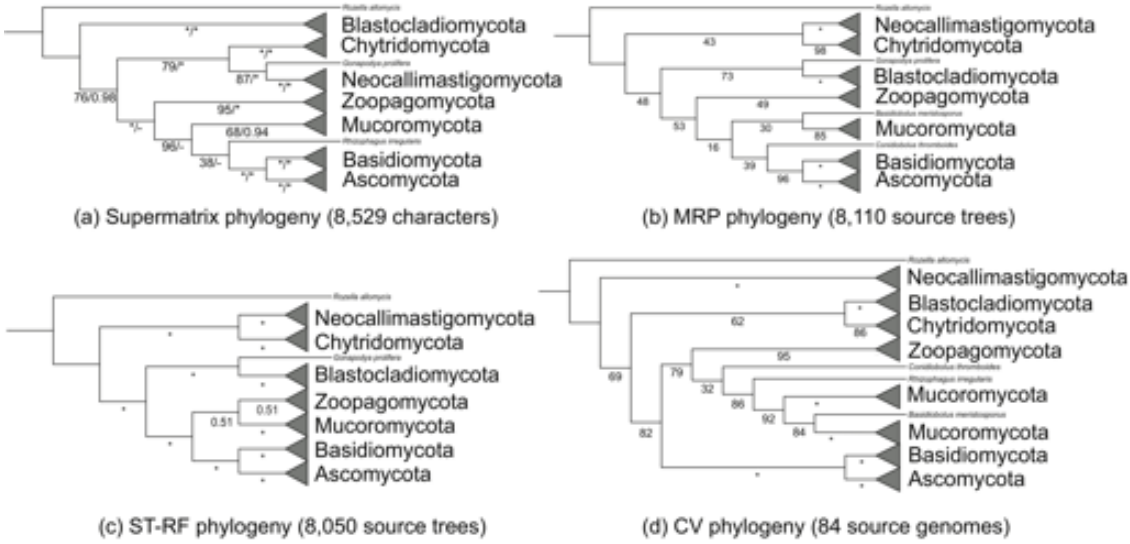


Figure 4.10. Congruence of 8 fungal phyla under 5 phylogenomic reconstructions. All clades bar Cryptomycota (represented *Rozella allomycis*) collapsed by phylum, paraphyletic species displayed as individual leaves. *Gonapodya prolifera* = Chytridiomycota, *Rhizophagus irregularis* = Mucoromycota, all other species except *R. allomycis* = Zoopagomycota. Refer to **Figures 4.3, 4.4, 4.5, 4.7 and 4.9** respectively for original phylogenies. **Figure 4.10a**. ML and Bayesian supermatrix phylogenies. Branch supports given as ML bootstrap supports and, where topology is identical, Bayesian posterior probabilities. Maximum bootstrap or posterior probability support designated with an asterisk (*). **Figure 4.10b.** MRP supertree phylogeny. Branch supports given as bootstrap supports. Maximum bootstrap support designated with an asterisk (*). **Figure 4.10c.** MCMC Bayesian supertree phylogeny using ST-RF ML method. Branch supports given as posterior probabilities of bipartition(s). Maximum posterior probability support designated with an asterisk (*). **Figure 4.10d.** CV phylogeny. Branch supports given as bootstrap supports. Maximum bootstrap support designated with an asterisk (*).

4.3.1 Higher-level genome phylogeny of the fungal kingdom

Despite variations in the resolution of some branches, there is a trend across the majority of phylogenies conducted of support or partial support for the eight phyla described in our dataset. **Figure 4.10** shows the congruence on the phylum level within the fungal kingdom in five of our seven phylogenetic reconstructions. We will refer to **Figure 4.10** and the subfigures (**Figures 4.10a–d**) in **Figure 4.10** when comparing the different reconstructions on the phylum level and to the corresponding full phylogenies themselves for comparisons at lower levels here and elsewhere (average consensus and gene content phylogenies are omitted from **Figure 4.10** on the basis of erroneous placement of taxa). Beginning with the Cryptomycota species *Rozella allomycis*, the next-earliest diverging clade within the fungal kingdom is the Blastocladiomycota under both

supermatrix analyses followed by Neocallimastigomycota and Chytridiomycota (**Figure 4.10a**). Other analyses place Neocallimastigomycota and Chytridiomycota (except *Gonapodya prolifera*) as closest to *R. allomycis* (**Figure 4.10b-d**).

We describe the resolution of the former zygomycetes in greater detail below, but in the five phylogenies in **Figure 4.10** all support at least a sister relationship between the two zygomycetes phyla Zoopagomycota and Mucoromycota. The placement of the Glomeromycotina species *Rhizophagus irregularis* varies, but Mucoromycota is generally placed as sister to the Dikarya (**Figure 4.10**). The Basidiomycota are fully supported as monophyletic in each of the five phylogenies represented in **Figure 4.10**, and all bar ML supermatrix reconstruction are in exact agreement with the two most extensive fungal genome phylogenies containing all three Basidiomycota subphyla (Wang *et al.*, 2009; Medina, Jones and Fitzpatrick, 2011). The Ascomycota are also fully supported as monophyletic in each of the five phylogenies represented in **Figure 4.10**, with the only major variation being the placement of *Saitoella complicata* within (or paraphyletic to) Taphrinomycotina (**Figure 4.10**). The Saccharomycotina are monophyletic in all five phylogenies (**Figure 4.10**). We discuss the class-level phylogeny within Pezizomycotina in greater detail in below (**Figure 4.11**), but to briefly summarize here we see strong-to-maximum support for all six of the larger classes that were present in our dataset, and support for the two unofficial "Sordariomyceta" and "Dothideomyceta" groupings within Pezizomycotina (Schoch *et al.*, 2009).

4.3.2 Multiple phylogenomic methods show moderate support for the modern designations of Mucoromycota and Zoopagomycota

There is moderate support for the recent designation of the zygomycetes phyla Zoopagomycota and Mucoromycota by Spatafora *et al.* (2016) across most of our phylogenomic methods (**Figure 4.10**). Previously the species within these two phyla were classified within Zygomycota, a phylum-level classification that had dated back to the 1950s until it was formally disputed by Hibbett *et al.* (2007). Six *incertae sedis* zygomycetes subphyla were later circumscribed (Hoffmann, Voigt and Kirk, 2011), and subsequent phylogenetic analyses informally classified the zygomycetes subphyla into two groups, which were later established as Mucoromycota and Zoopagomycota (Chang *et al.*, 2015; Spatafora *et al.*, 2016).

Our phylogenomic analyses included 11 species from the two zygomycetes phyla, with the best resolution found in the ST-RF phylogeny where Zoopagomycota and Mucoromycota are placed as sister phyla with 0.51 PP and branch sister to Dikarya (**Figure 4.10c**). Notably, our ST-RF phylogeny is the only phylogeny that resolves Entomophthoromycotina as a monophyletic clade (**Figure 4.7**), albeit with extremely weak posterior probability support (0.38 PP). Within Zoopagomycota in our ST-RF phylogeny, Entomophthoromycotina branch as the basal clade with 0.51 PP, sister to Kickxellomycotina (**Figure 4.7**). Our ST-RF phylogeny also places *Rhizophagus irregularis* (Glomeromycotina) adjacent to *Mortierella elongata* (Mortierellomycotina) within the Mucoromycota (**Figure 4.7**). Within Mucoromycota, Mortiellomycotina and Mucoromycotina are supported as sister subphyla throughout the majority of our phylogenies (e.g. Bayesian supermatrix analysis, **Figure 4.4**) with high to maximum support. Both of these phylum-level topologies are in agreement with Spatafora et al. (2016), though their phylogeny does not support a distinctive monophyletic branch containing both Zoopagomycota and Mucoromycota (**Figure 4.10c**). The majority of our remaining phylogenomic analysis all show some degree of support for both Zoopagomycota and Mucoromycota in relative agreement with Spatafora *et al.* (2016), however in each of these phylogenies there is some conflict in either subphylum-level topology or lower BP/PP support due to issues of taxon sampling or low gene tree coverage in our dataset (of our 8,110 source phylogenies for MRP analysis over 3,500 contain seven taxa or less**; Figure 4.10**). With greater sampling of species from these lineages we hope to see more consistent support of both the Zoopagomycota and Mucoromycota in future genome phylogenies using these methods, in line with what appears to be moderate-to-strong support for the new classification in our analyses based on total evidence (Kluge, 1989).

4.3.3 Pezizomycotina as a benchmark for phylogenomic methodologies

The Pezizomycotina are by far the most sampled subphylum within the fungal kingdom in terms of genome sequencing (375 Pezizomycotina species have genomic data available from MycoCosm as of May 2017). Reflecting this, 22 Pezizomycotina species representing 7 classes are present in our 84-genome dataset (>25% of our final dataset). As a well-represented clade within our dataset at both the subphylum and individual class level, we are able to see how multiple phylogenomic analyses conducted in a total

evidence approach (Kluge, 1989) are able to resolve a single clade of closely-related classes containing some important ecological and pathogenic fungi. In every phylogenomic reconstruction we attempted bar average consensus (AV) phylogeny, Pezizomycotina were monophyletic with maximum bootstrap or posterior probability branch support and every class within Pezizomycotina is monophyletic with high or maximum BP or PP support (**Figures 4.3–4.5, 4.7–4.9**). There is a consistent trend within each of these phylogenies in the resolution of relationships between Pezizomycetes classes:

1) The Orbiliomycetes and Pezizomycetes always branch as the basal classes within Pezizomycotina, and are always sister taxa (**Figures 4.3–4.5, 4.7–4.9**).
2) The relationship between Sordariomycetes and Leotiomycetes (within "Sordariomyceta" *sensu* Schoch *et al.* (2009)) is always present and is fully supported in each phylogeny (**Figures 4.3–4.5, 4.7–4.9**).
3) The relationship between Dothideomycetes, Xylonomycetes, and Eurotiomycetes (within "Dothideomyceta" *sensu* Schoch *et al.* (2009)) is always present and is fully supported in each phylogeny (**Figures 4.3–4.5, 4.7–4.9**).

Figure 4.11 displays on the left the topology of the Pezizomycotina classes supported under ML supermatrix reconstruction, MRP supertree reconstruction and ST-RF supertree reconstruction (**Figures 4.3, 4.5, 4.7**), and indicates the congruence (or otherwise) of Pezizomycotina under every phylogenomic analysis we attempted (**Figures 4.3–4.9**). All methods bar AV are highly congruent in their resolution of the Pezizomycotina subphylum, with placement of the Xylonomycetes class the most notable variation. Even within the highly aberrant AV phylogeny, sister relationships such as those between Orbiliomycetes and Pezizomycetes or the association of classes within Sodariomyceta or Dothideomyceta can still be observed, though with lower resolution and support (**Figure 4.6**). There is a high degree of congruence between our genome phylogenies of Pezizomycotina (**Figure 4.11**) and the most extensive molecular phylogenies of Pezizomycotina that we could find in the literature derived from either small concatenated sets or whole genomes (Spatafora *et al.*, 2006; Wang *et al.*, 2009; Medina, Jones and Fitzpatrick, 2011). The relative consistency of our analyses with both each other and with previous literature suggests that the resolution of Pezizomycotina could be considered a good benchmark for the accurary of novel or existing

phylogenomic methods (e.g. ST-RF analysis) when incorporated into a total evidence analysis, as the subphylum is large and diverse (the 10th edition of Ainsworth & Bisby's Dictionary of the Fungi estimates close to 70,000 Pezizomycetes species) but also densely-sampled in genomic terms and containing a number of genomes of reference quality (Kirk *et al.*, 2008).

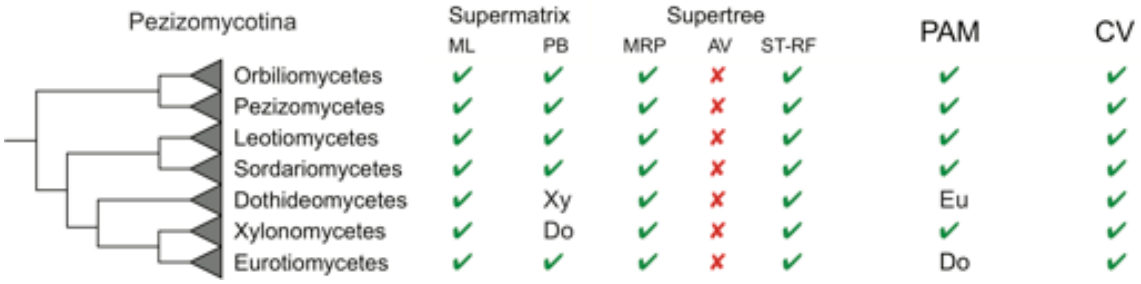


Figure 4.11. Congruence of Pezizomycotina under 7 phylogenomic methods. Placement of classes identical to topology on the left (see text) indicated with a tick, varying placement of classes indicated by the first two letters of a class. Average consensus (AV) phylogeny produced paraphyletic Pezizomycotina and so entire column labelled with crosses. Refer to text for discussion of topology of Pezizomycotina under AV phylogeny. Refer to **Figures 4.3-4.9** for original phylogenies.

4.3.4 The use of phylogenomics methods in fungal systematics

Phylogenomic analyses with larger datasets across a wider spectrum of taxa are becoming more and more computationally tractable as methods of identifying potential phylogenetic markers on a genome-wide scale (e.g. identification and reconstruction of orthologous gene phylogenies in supertree analysis) and genome-scale reconstruction improve. In as much as the majority of our multiple analyses strongly support the major phyla of the fungal kingdom, we can also treat our analyses as measures of the accuracy of each of these phylogenomic methods in the reconstruction of large datasets. Supermatrix, MRP and ST-RF supertree and CV method reconstructions all appear to arrive at relatively congruent results, and may be useful for approximating a total evidence style approach for phylogenomic analyses of fungi. Simplified parsimony methods like our PAM phylogeny or branch length-based methods like our average consensus phylogeny may be useful for the reconstruction of smaller but well-represented datasets (for example our PAM phylogeny does reconstruct the Pezizomycotina with support and topology close to supertree and supermatrix phylogenies) but for phylum or kingdom-wide analyses issues such as long-branch attraction begin to emerge (Bergsten, 2005). Long-branch attraction is thought to be an issue with MRP reconstruction as well, and while it is likely a factor in the weaker supports in some of the ancestral branches in our MRP phylogeny (for example, the weak supports in some of the internal branches grouping the basal phyla together), the MRP phylogeny seems to have been relatively immune to the topological effects of long-branch attraction that are very apparent in our branch-length dependent average consensus method phylogeny (Pisani and Wilkinson, 2002).

For our supertree analyses we identified groups of orthologous proteins using a sequential random BLASTp approach as implemented by Fitzpatrick *et al.* (2006), where a random sequence from a given database is searched against that entire database, and then the sequence and its homologs (if any) are removed and the database reformatted (Fitzpatrick *et al.*, 2006). Overall this *ad hoc* approach to identifying orthology within our dataset seems to have been sufficient as a first step to generating source gene phylogenies, however it may have had an impact downstream on resolution of internal branches within our MRP analysis. It is possible a random BLASTp approach is too conservative, in that the orthologous families it identifies are missing members or that

two “separate” orthologous families may in fact be one large orthologous family. Other established methods of identifying orthologous families, such as the OrthoMCL pipeline, have been used in phylogenomic analyses and can be tuned for granularity (i.e. orthologous cluster size) which may produce broader source phylogenies (Li, Stoeckert and Roos, 2003). However, the large SQL-dependent computational overheard required for the current implementation of OrthoMCL was not considered suitable for an analysis of this scale.

Most of the phylogenomic methods we attempted are relatively tractable even for a dataset as large as ours. Depending on computational resources and available data, some of the methods we have discussed may be more appropriate for future fungal phylogenomic analyses than others. The most common techniques like MRP analysis and both ML and Bayesian supermatrix analysis were both tractable and produced phylogenies with largely congruent topologies and supports on most branches (although we should note that we utilized the parallelized version of PhyloBayes for our Bayesian analysis). The heuristic MCMC Bayesian supertree reconstruction we attempted using the ST-RF model as implemented in p4 was also relatively tractable despite not being parallelized, and Akanni *et al.* (2015) note that the method is far more efficient than the approximate ML reconstruction implemented in L.U.St. (Akanni *et al.*, 2015). However, ST-RF analysis using either p4 or L.U.St. is currently only able to use fully resolved input phylogenies. While in our case this meant only 60 single-copy phylogenies (<1% of our total dataset) had to be removed before carrying out analysis, this may cause issues for more polytomous datasets. Bayesian and ML supertree reconstruction is certainly a promising development for phylogenomics, and hopefully methods like ST-RF should see more widespread use in future phylogenomic analysis as they mature.

Phylogenomic reconstruction using average consensus as implemented in CLANN was extremely inefficient time-wise and returned a severely erroneous phylogeny, so while it is certainly desirable for branch lengths to be incorporated in supertree reconstruction, a branch length-based method like AV is not appropriate for this kind of large-scale analysis. While PAM method reconstruction was straightforward to carry out, as we state above there were issues with erroneous placement of taxa and as such we do not recommend the method for large-scale datasets. Finally, composition vector method analysis produced a phylogeny relatively congruent to our alignment-based methods at $K = 5$. Other CV method analyses have recommended K-values between 5 and 7 for most datasets (Zuo, Li and Hao, 2014), however with the size of our dataset

and the increase in computational resources required for generating distance matrices for eukaryotic genomes at $K > 5$ in CVTree we felt that $K = 5$ was the best compromise between accuracy and computational tractability. We would recommend however that CV analysis should be used in conjunction with alignment-based methods for eukaryotic datasets, as interpretation of CV analysis requires *a priori* knowledge of the phylogeny of a given dataset.

4.4 Conclusions

Fungi make up one of the major eukaryotic kingdoms, with millions of member species inhabiting a diverse variety of ecological niches and an evolutionary history dating back over a billion years. It is imperative that evolutionary relationships within the fungal kingdom are well-understood by analysis of as much quality phylogenetic data as is available with the most accurate methodologies possible. In this chapter, we discussed the evolutionary diversity of the fungal kingdom and the important role that fungi have had in the area of genomics and phylogenomics. We have reviewed previous phylogenomic analyses of the fungal kingdom over the last decade, and using seven phylogenomic methods we have reconstructed the phylogeny of 84 fungal species across 8 fungal phyla. We found that established supermatrix and supertree methods produced relatively congruent phylogenies that were in large agreement with the literature. We also conducted the first analysis of the fungal kingdom using a heuristic MCMC Bayesian approach to supertree reconstruction previously used in Metazoa, and found that this novel supertree approach resolves the fungal kingdom with a high degree of accuracy. The majority of our analyses overall show moderate-to-strong support of the newly-assigned zygomycete phyla Mucoromycota and Zoopagomycota and strongly support the monophyly of Dikarya, while within the highly-sampled Pezizomycotina subphylum there is a large amount of congruence between different phylogenomic methods as to the resolution of class relationships within the subphylum. We also conclude that supermatrix and supertree analyses remain the exemplar methods of phylogenomic reconstruction for fungi, based on their accuracy and computational tractability. We believe through both our discussion of the ecological diversity of the fungal kingdom and the history of its study on the genomic level we have demonstrated the need for a robust fungal tree of life with a broad representation, and that through our multiple phylogenomic analysis we have generated an important backbone for future comparative genomic analysis of fungi, particularly with the constantly increasing amount of quality genomic data arising from the 1000 Fungal Genomes Project and its certain use in future studies.

Chapter 5 – Pan-genome analysis of model fungal species

This chapter was previously published in *Microbial Genomics* in February 2019.

McCarthy C. G. P. & Fitzpatrick D. A. (2019). Pan-genome analyses of model Fungal species. *Microbial Genomics*, 5(2).

Chapter outline

The concept of the species "pan-genome", the union of "core" conserved genes and all "accessory" non-conserved genes across all strains of a species, was first proposed in prokaryotes to account for intraspecific variability. Species pan-genomes have been extensively studied in prokaryotes, but evidence of species pan-genomes has also been demonstrated in eukaryotes such as plants and fungi. Using a previously-published methodology based on sequence homology and conserved microsynteny in addition to bespoke pipelines, we have investigated the pan-genomes of four model fungal species: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. Between 80-90% of gene models per strain in each of these species are core genes that are highly-conserved across all strains of that species, many of which are involved in housekeeping and conserved survival processes. In many of these species the remaining "accessory" gene models are clustered within subterminal regions and may be involved in pathogenesis and antimicrobial resistance. Analysis of the ancestry of species core and accessory genomes suggests that fungal pan-genomes evolve by strain-level innovations such as gene duplication as opposed to wide-scale horizontal gene transfer. Our findings lend further supporting evidence to the existence of species pan-genomes in eukaryote taxa.

5.1 Introduction

Many fields of eukaryote functional and comparative genomics rely on the use of curated reference genomes intended to be broadly representative of a given species. Regardless of their quality, reference genomes do not and cannot contain all genetic information for a species due to genetic and genomic variation between individuals within a species (Parfrey, Lahr and Katz, 2008). To account for such variation it has become increasingly common to refer to species with multiple genomes sequenced in terms of their "pan-genome", which is defined as the union of all genes observed across all isolates/strains of a species (2-4) (**Figure 5.1**). The pan-genome of a species is then usually subdivided into two components:

- The "core" genome, containing genes conserved across all observed genomes from a species. These genes are usually, but not always, essential for the viability of an individual organism (Rouli *et al.*, 2015).
- The "accessory" or "dispensable" genome, containing genes specific to sets of isolate genomes or individual isolate genomes within a species. These genes could influence phenotypic differences between isolates; for example, antibiotic-resistant and antibiotic-susceptible isolates of the same species may have different accessory genomes (Rouli *et al.*, 2015).

A species' pan-genome can evolve as a consequence of lifestyle: sympatric species may have large pan-genomes (and thus a large degree of intraspecific variation), while environmentally isolated or highly specialized species have smaller pan-genomes (Snipen, Almøy and Ussery, 2009; Lefebure *et al.*, 2010; Diene *et al.*, 2013; Rouli *et al.*, 2015). The existence of a species pan-genome in prokaryotes was first demonstrated across eight pathogenic strains of *Streptococcus agalactiae* in 2005 (Tettelin *et al.*, 2005), and was quickly confirmed by similar analysis of exemplar bacteria and archaea including *Haemophilus influenzae*, *Escherichia coli* and *Sulfolobus islandicus* (Young *et al.*, 2006; Hogg *et al.*, 2007; Rasko *et al.*, 2008; Reno *et al.*, 2009; Boissy *et al.*, 2011). Over 40 prokaryote species had their pan-genomes described in the literature by 2013 (Rouli *et al.*, 2015). Many tools for pan-genome analysis have been published in recent years, which utilize methods such as whole-genome alignment, read mapping, clustering algorithms or de Bruijn graph construction (Marcus, Lee and Schatz, 2014; Page *et al.*, 2015; Song *et al.*, 2015; Chaudhari, Gupta and Dutta, 2016; Jandrasits *et al.*, 2018).

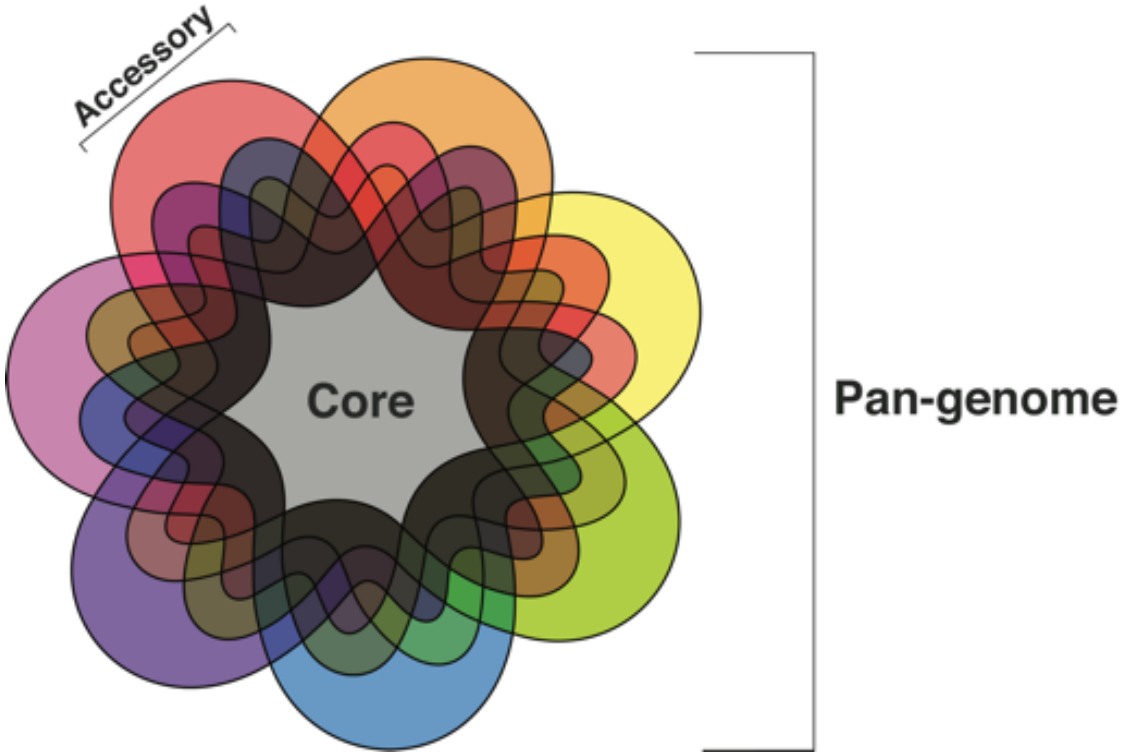


Figure 5.1. Seven-set Venn diagram representing a hypothetical species pan-genome. Each set represents genes/gene models conserved across strains of a given species. The "core" species genome (grey) is defined as the set of all genes/gene models conserved across all strains of a species, while the "accessory" genome consists of all genes/gene models not universally conserved within a species.

Although the concept of the species pan-genome is well-established in comparative prokaryote genomics, it has only recently been extended to comparative intraspecific studies of eukaryotes. This is despite repeated observation of intraspecific genomic content variation in eukaryotes dating back to the first intraspecific comparative analyses of *Saccharomyces cerevisiae* genomes in the mid-2000s (Gu *et al.*, 2005; Ronald, Tang and Brem, 2006; Wei *et al.*, 2007; Engel and Cherry, 2013). The relative dearth of eukaryotic pan-genome analysis in the literature is due in part to the relative difficulty of sequencing and analysing large eukaryotic genome datasets relative to prokaryotes (Golicz, Batley and Edwards, 2016). Additionally, while horizontal gene transfer (HGT) is thought to be the driving influence in prokaryotic gene family and pan-genome evolution, HGT occurs in far lower rates in eukaryotes and is more difficult to detect (Keeling and Palmer, 2008; Lapierre and Gogarten, 2009; Ku and Martin, 2016; Martin, 2017; McInerney, McNally and O'Connell, 2017). Despite these challenges, there have been a number of recent studies of intraspecific variation within diverse eukaryote

taxa that show strong evidence for the existence of a eukaryotic pan-genome in some form.

Comparative analysis of nine diverse cultivars of *Brassica oleracea* found that ~19% of all genes analysed were part of the *B. oleracea* accessory genome, with ~2% of these being cultivar-specific (Golicz *et al.*, 2016). A similar comparison of seven geographically diverse wild soybean (*Glycine soja*) strains found approximately the same 80:20 proportion of core to accessory gene content within the wild soybean pan-genome, while larger accessory genome sizes have been reported in wheat, maize, grasses and *Medicago* (Hirsch *et al.*, 2014; Y. H. Y. F. Li *et al.*, 2014; Gordon *et al.*, 2017; Montenegro *et al.*, 2017; Zhou *et al.*, 2017). Individual strains of the coccolitophore *Emiliania huxleyi* have an accessory complement of up to 30% of their total gene content which varies with geographical location (Read *et al.*, 2013). In fungi, a number of studies of the *Saccharomyces cerevisiae* pan-genome, including a recent large-scale analysis of genome evolution across 1,011 strains, have shown evidence for an accessory genome of varying size as well as large variation in subterminal regions across multiple *S. cerevisiae* strains (Dunn *et al.*, 2012; Bergström *et al.*, 2014; Song *et al.*, 2015; Peter *et al.*, 2018), and recent analysis of the *Zymoseptoria tritici* pan-genome found that up to 40% of genes in the total *Z. tritici* pan-genome were either lineage or strain-specific (Plissonneau, Hartmann and Croll, 2018).

The methods of pan-genome evolution within eukaryotes in the absence of rampant HGT appears to vary among species and can include genome rearrangement events or more discrete adaptive evolution processes. In plants accessory genomes may evolve as a result of varying levels of ploidy, heterozygosity and whole-genome duplication within species as well as adaptive changes and the evolution of phenotypic differences, such as in *Brassica oleracea* (Golicz *et al.*, 2016). Adaptive evolution has also influenced the evolution of the *Emiliania huxleyi* pan-genome, with strains containing varying amounts of nutrient acquisition and metabolism as a result of niche specialization (Read *et al.*, 2013). High levels of functionally-redundant accessory genome content can be observed within the *Z. tritici* species pan-genome, which is thought to arise from the species' own genome defence mechanisms inducing polymorphisms as opposed to gene duplication events (Plissonneau, Hartmann and Croll, 2018). Peter *et al.* (2018) observed a large proportion of accessory genes within *S. cerevisiae* appear to have arose *via* introgression from closely-related *Saccharomyces*

species, with a smaller number originating from HGT events with other yeasts (Peter *et al.*, 2018).

We have adapted a method of prokaryotic pan-genome analysis that identifies putative pan-genomic structure within species by accounting for conserved genomic neighbourhoods (CGNs) between strain genomes and applied it to eukaryote analysis (Fouts *et al.*, 2012) (**Figure S5.1**). We have used this method in tandem with bespoke pre- and post-processing pipelines which analyse the extent of gene duplication within species pan-genomes (available from https://github.com/chmccarthy/pangenome_pipelines) to construct and characterize the pan-genomes of four exemplar fungal species; *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. All four species are model organisms in eukaryotic genomics and play important roles in human health and lifestyles; *S. cerevisiae* is used extensively in biotechnology, *C. albicans* is an opportunistic invasive pathogen and the second-most common cause of fungal infection, *C. neoformans* var. *grubii* is an intracellular pathogen that causes meningitis in immunocompromised hosts, and *A. fumigatus* is an opportunistic respiratory pathogen (Goffeau *et al.*, 1996; Jones *et al.*, 2004; Nierman *et al.*, 2005; Cock *et al.*, 2009; Janbon *et al.*, 2014). We have found strong evidence for pan-genomic structure within all four fungal species. In line with previous analyses of other eukaryotes, we found that approximately 80-90% of fungal species' pan-genomes are composed of core genes while the remainder is composed of strain or lineage-specific accessory genes. Analysis of the origin of fungal pan-genomes suggests that fungal accessory genomes are enriched for genes of eukaryotic origin and arise *via* eukaryotic innovations such as gene duplication as opposed to large-scale HGT. Functionally, fungal core genomes are enriched for both housekeeping processes and essential survival processes in pathogenic species, whereas many fungal accessory gene models are found within clusters in the terminal and subterminal regions of genomes and are enriched for processes that may be implicated in fungal pathogenicity or antimicrobial resistance. Our findings complement the increasing amount of studies showing evidence for pan-genomic structure in eukaryote species.

5.2 Materials and Methods

5.2.1 Dataset assembly

For each of the four fungal species chosen, we obtained strain genome assemblies from the NCBI's GenBank facility (**Table S5.1**). Strains were selected based on geographic and environmental diversity where possible (**Table S5.1**). The predicted protein set from each species' reference genome was also obtained from GenBank. For each strain genome in each species dataset, translated gene model and gene model location prediction was performed using a bespoke prediction pipeline consisting of three parts (**Figure S5.2a**):

1. Reference proteins were queried against individual strain genomes using Exonerate with a heuristic protein2genome search model (Slater and Birney, 2005). Translated gene model top-hits whose sequence length was ≥50% of the query reference protein's sequence length were considered homologs and included in the strain gene model set. The genomic locations of these gene models were included in the strain genomic locations dataset.
2. *Ab initio* Hidden Markov Model (HMM)-dependent gene model prediction was carried out using GeneMark-ES, with self-training and a fungal-specific branch point site prediction model enabled (Ter-Hovhannisyan *et al.*, 2008). Predicted gene models whose genomic locations did not overlap with any gene models previously predicted *via* step 1 were included in the strain gene model set. The genomic locations of these gene model were also included in the strain genomic locations dataset.
3. Finally, position weight matrix (PWM)-dependent gene model prediction was carried out for all remaining non-coding regions of the genome using TransDecoder (Haas *et al.*, 2013). For *Saccharomyces cerevisiae* and *Candida albicans* strain genomes, these gene models were additionally screened against a dataset of known "dubious" pseudogenes in each species taken from their respective public repositories using BLASTp with an e-value cutoff of 10^{-4} (Camacho *et al.*, 2009; Skrzypek *et al.*, 2017). Predicted gene models whose top BLASTp hit against a known dubious pseudogene had a sequence

coverage of ≥70% were removed from further processing. All remaining predicted gene models with a length of ≥200 amino acids and a coding potential score of 100 or greater as assigned by TransDecoder were included in the final strain gene model set. Their corresponding genomic locations were also included in the strain genomic locations dataset.

Thus for each strain genome in a species dataset, a gene model set and corresponding genomic location set was constructed using two initial independent prediction methods; a search for gene models orthologous to the reference protein set and an *ab initio* prediction approach, followed by a "last resort" approach for predicting gene models in genomic regions for which gene models had not been previously called. We used this approach to ensure consistency in gene models calls between strains and to reduce the potential of poor heterogenous gene model calling within each species dataset, which would in turn reduce the number of false positives/negatives in our analysis. The completeness of each set of predicted gene models was assessed using BUSCO with the appropriate BUSCO dataset for each species (Simão *et al.*, 2015) (**Table S5.1**). For each species dataset, all strain genome gene model sets were combined and an all-vs.-all BLASTp search was carried out for all predicted gene models using an e-value cutoff of 10^{-4}. The results of the BLASTp search were used as input for PanOCT along with the combined genomic location data for each strain genome in a species dataset (Fouts *et al.*, 2012). Further information for each species dataset is detailed below.

5.2.1.1 Saccharomyces cerervisiae

Genomic data for 100 *Saccharomyces cerevisiae* strains were obtained from the NCBI's GenBank facility. Of these 100 genomes, 99 had previously been included in the geographically- and phenotypically-diverse "100-genomes strains" resource for *S. cerevisiae* (Strope *et al.*, 2015). For our analysis, we excluded the 100GS European vineyard strain M22 as its lower assembly quality prevented us from carrying out *ab initio* gene model prediction using GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008; Strope *et al.*, 2015). In its place we included the European commercial winemaking strain Lalvin EC118 (Novo *et al.*, 2009). The protein set for the reference *S. cerevisiae* strain S288C was also obtained from GenBank (Goffeau *et al.*, 1996). Construction of the *S. cerevisiae* pan-genome dataset was performed as detailed above, with potentially dubious gene

model predictions for each strain genome checked against a dataset of 689 known dubious *S. cerevisiae* gene models obtained from the Saccharomyces Genome Database (SGD) (Engel and Cherry, 2013). The completeness of each strain's gene model dataset was assessed using 1,711 *S. cerevisiae* BUSCOs from the Saccharomycetales dataset; on average ~1,677 BUSCOs (~98%) were retrieved as complete gene models in each strain (**Table S5.1**). In total, 576,578 gene models and corresponding unique genomic locations were predicted for 100 *S. cerevisiae* genomes (**Table S5.1**).

5.2.1.2 Candida albicans

Genomic data for 34 *Candida albicans* strains were obtained from the NCBI's GenBank facility, encompassing predominantly clinical or presumed-clinical strains isolated from North America, Europe and the Middle East (**Table S5.1**). The protein set for the reference *C. albicans* strain SC5314 was also obtained from GenBank (Jones *et al.*, 2004). Construction of the *C. albicans* pan-genome dataset was performed as detailed above, with potentially dubious gene model predictions for each genome checked against a dataset of 152 known dubious gene models from *C. albicans* SC5314 obtained from the Candida Genome Database (CGD) (Skrzypek *et al.*, 2017). The completeness of each strain's gene model dataset was assessed using 1,711 *S. cerevisiae* BUSCOs from the Saccharomycetales dataset; on average ~1,642 BUSCOs (~96%) were retrieved as complete gene models in each strain (**Table S5.1**). In total, 204,407 gene models and their corresponding unique genomic locations were predicted for 34 *C. albicans* genomes (**Table S5.1**).

5.2.1.3 Cryptococcus neoformans var. *grubii*

Genomic data for 25 *Cryptococcus neoformans* var. *grubii* strains was obtained from the NCBI's GenBank facility, encompassing both clinical and wild-type strains sampled from North America and Southern African regions (**Table S5.1**). The protein set for the reference *C. neoformans* var. *grubii* strain H99 was also obtained from GenBank (Janbon *et al.*, 2014). Construction of the *C. neoformans* var. *grubii* pan-genome dataset was performed as detailed above, with the exception that a check for known dubious gene models was not carried out as no such data were available for *C. neoformans* var. *grubii*. The completeness of each strain's gene model dataset was assessed using the 1,335 BUSCOs from the Basidiomycota dataset; on average ~987 BUSCOs (~74%) were retrieved as complete gene models in each strain (**Table S5.1**). In total, 172,105 gene

models and their corresponding genomic locations were predicted for 25 *Cryptococcus neoformans* var. *grubii* genomes (**Table S5.1**).

5.2.1.4 Aspergillus fumigatus

Genomic data for 12 *Aspergillus fumigatus* strains were obtained from the NCBI's GenBank facility, including both clinical and wild-type strains isolated from the Northern and Southern hemispheres and the International Space Station (**Table S5.1**). The protein set for the reference *A. fumigatus* strain AF293 was also obtained from GenBank (Nierman *et al.*, 2005). Construction of the *A. fumigatus* pan-genome dataset was performed as detailed above, with the exception that a check for known dubious gene models was not carried out as no such data was available for *A. fumigatus*. The completeness of each strain's gene model dataset was assessed using 4,046 *Aspergillus nidulans* BUSCOs from the Eurotiomycetes dataset; on average ~3,410 BUSCOs (~84%) were retrieved as complete gene models in each strain (**Table S5.1**). In total, 117,230 putative proteins and their corresponding unique genomic locations were predicted for 12 *A. fumigatus* genomes (**Table S5.1**).

5.2.2 Pan-genome analysis of fungal species

Analysis of the pan-genomes of the four fungal species in our study was performed using the Perl software PanOCT (Fouts *et al.*, 2012). PanOCT is a graph-based method that uses both BLAST score ratio (BSR) (Rasko, Myers and Ravel, 2005) and conserved gene neighbourhood (CGN) (Deniélou *et al.*, 2011) approaches to establish clusters of syntenically-conserved orthologs across multiple genomes for species pan-genome analysis (**Figure S5.1**). The use of genomic context in addition to sequence similarity in PanOCT allowed us to distinguish between multiple homologous sequences within any genome analysed (i.e. paralogs) (Fouts *et al.*, 2012). We used CGN (window size = 5, the default value) as our criterion for defining conserved gene evolution between strains of fungal species. In the sections below, we refer to gene models containing an ortholog from all strains present in a species dataset as "core" gene models (and thus part of the "core" genome) and those missing an ortholog from one or more strains as "accessory" clusters (and thus part of the "accessory" genome). After removing invalid or low-quality BLASTp hits in each species dataset (**Table S5.1**), the initial core and

accessory genomes for each species dataset were constructed using PanOCT with the default parameters.

To assess the influence of duplication and microsynteny loss on fungal pan-genomes we processed the results of the PanOCT analysis using a multi-step Python/R post-processing pipeline. This first step of this pipeline was an iterative search for independent syntenic clusters with the potential to be merged based on reciprocal sequence similarity. Starting with accessory clusters of size $n - 1$ (where n is the number of strains in a dataset), parallelized all-vs.-all BLASTp searches of all remaining gene models from accessory clusters ($e = 10^{-4}$) were performed, and this output was parsed to identify instances where two accessory clusters with no overlapping strain representation could be merged into one cluster based on the following criteria:

1. Each member gene model in a "query" cluster of size m had a reciprocal BLASTp strain top-hit with a sufficient number of member gene models in a "subject" cluster of size $n - m$ or smaller.
2. The size of the resulting "merged" cluster was $\leq n$.

This approach attempted to account for loss-of-synteny events such as rearrangements, or other artefacts arising from different genome sequencing and assembly methods. Merged accessory clusters that now had an orthologous gene model from each strain in a dataset (i.e. whose size = n) were recategorized as core clusters, although for this study such recategorizations were a rare occurrence.

The second step of our post-processing pipeline assessed the influence of gene duplication on fungal pan-genome evolution by analysing the proportion of accessory gene models that were potentially paralogous to the core genome. Gene models from accessory clusters were assessed for sequence similarity to core gene models from the initial all-vs.-all BLASTp search used as input for PanOCT. If accessory gene models were sufficiently similar to every gene model from a given core cluster (e-value cutoff of $1e^{-4}$), then that accessory cluster was classified as being a paralogous cluster or a cluster of duplicated core gene models. This approach attempted to account for duplication events followed by subsequent gene loss, rearrangement in strains or strain/lineage-specific expansions of gene families. Using a sequence-based approach of pan-genome analysis as opposed to genome alignment or other methods also facilitated the downstream application of systematic functional analysis of species pan-genomes; e.g. GO-slim enrichment, which are detailed below. We visualized the distribution of syntenic orthologs within fungal accessory genomes using the UpSet technique, an alternative to

Venn or Euler diagrams, which visualizes intersections of sets and their occurrences using a matrix representation (Lex *et al.*, 2014). This technique, implemented in the R package UpSetR, allowed us to see the number of shared syntenic orthologs (intersections) across different strains (sets) within a species dataset (R Core Team and R Development Core Team, 2013; Conway, Lex and Gehlenborg, 2017). Singleton gene models from each reference strain genome were functionally characterized by searching against their corresponding reference protein set using BLASTp ($e = 10^{-4}$). Statistics for each pangenome dataset is given in **Table 5.1** below.

5.2.3 Phylogenomic reconstruction of intraspecific phylogenies

Phylogenomic reconstruction of intraspecific lineages was carried out for all four fungal species using a supermatrix approach. For each fungal pan-genome dataset, all core ortholog clusters whose smallest gene model was at least 90% the length of the longest gene model were retrieved from the dataset. Each cluster was aligned in MUSCLE with the default parameters, and for each cluster alignment phylogenetically-informative character sites were extracted using PAUP* (Swofford, 2002; Edgar, 2004). Sampled alignments retaining character data were concatenated into a superalignment using FASConCAT (Kück and Meusemann, 2010). In total,

- 4,311 *S. cerevisiae* core clusters (431,100 gene models) passed the minimum sequence length criterion and retained alignment data after sampling, and were concatenated into a 100-genome superalignment containing 54,860 amino acid (aa) sites.
- 4,327 *C. albicans* core clusters (68,904 gene models) retained alignment data after sampling, and were concatenated into a 34-genome superalignment containing 31,999 aa sites.
- 4,512 *C. neoformans* var. *grubii* core clusters (112,800 gene models) retained alignment data after sampling, and were concatenated into a 25-genome superalignment containing 47,811 aa sites.
- 5,724 *A. fumigatus* core clusters (68,904 gene models) retained alignment data after sampling for phylogenetically-informative residues, and were concatenated into a 12-genome superalignment containing 20,760 aa sites.

Approximate maximum-likelihood phylogenomic reconstruction was performed for each superalignment using FastTree with the default JTT + CAT evolutionary model and

Shimodaira-Hasegawa local supports (Price, Dehal and Arkin, 2010). All phylogenomic trees were rooted at the midpoint and annotated using the iTOL website (Letunic and Bork, 2016) (**Figures 5.3-5,6**). A binary matrix was generated for the presence/absence of all ortholog clusters across all strains within each species accessory genome. Each species matrix was mapped onto the corresponding intraspecific supermatrix phylogeny and Dollo parsimony analysis was performed on each matrix using Count (**Figures 5.3-5.6**) (Farris, 1977; Csurös, 2010). Ortholog gain and loss events were manually annotated onto each intraspecific phylogeny.

5.2.4 Functional annotation and GO enrichment analysis of fungal species pan-genomes

Pfam, InterPro and gene ontology (GO) annotation for all four fungal datasets was carried out using InterProScan (Hunter *et al.*, 2012; Jones *et al.*, 2014; Finn *et al.*, 2015; Carbon *et al.*, 2017). The total numbers of proteins with at least one annotation per database from the original putative protein sets per species is given in **Table 5.2**. Enrichment analysis of GO terms was carried out for the core and accessory complements of each species' pan-genome by mapping all GO terms per species to their species GO-slim counterparts (or to the general GO-slim term basket for *C. neoformans* var. *grubii*) and performing a Fischer's exact test (FET) analysis with parent term propagation and false discovery rate (FDR) correction ($p < 0.05$) for all complements using the Python package GOAtools (**Table S5.2)** (Agresti, 2002; Carbon *et al.*, 2017; Klopfenstein *et al.*, 2018). FDR correction was applied for all FETs in GOAtools using a p-value distribution generated from 500 resampled p-values.

5.2.5 Putative ancestral history of fungal core and accessory genomes

The putative evolutionary history of fungal core and accessory genomes was analysed by querying all gene models per species against a >5-million protein dataset sampled from 1,109 bacterial and 488 archaeal genomes obtained from UniProt, using BLASTp with an e-value cutoff of 10^{-20} (Cotton and McInerney, 2010). Gene models were filtered by their ancestral history into three classifications using the following criteria:

- Gene models whose hits were exclusively from bacterial or archaeal sequences were classified as "bacterial" or "archaeal" in origin, respectively.

- Gene models whose hits contained both bacterial and archaeal sequences were classified as “undefined prokaryote” in origin.
- Gene models which did not hit any protein sequence in the dataset were classified as “eukaryotic” in origin (**Table S5.3)**.

Pearson’s χ^2 tests were carried out to determine the significance of prokaryote and eukaryote origin frequencies within the complements of each species pan-genome (Agresti, 2002) (**Table S5.3)**.

5.2.6 Extent of horizontal gene transfer into fungal accessory genomes

The extent of HGT in each fungal accessory genome was assessed by randomly selecting representative gene models from each accessory cluster and searching these using BLASTp with an e-value cutoff of $1e^{-20}$ against a dataset representative of fully sequenced prokaryotic and eukaryotic species. This dataset was composed of over 8 million protein sequences from 1,698 genomes sampled from all three domains of life which had been used in previous interdomain HGT analysis (McCarthy and Fitzpatrick, 2016), as well as all predicted gene models per species dataset. Putative interdomain HGT events were identified by locating gene models whose first top hit outside either the sequence’s source species or genus was prokaryotic in origin. Putative HGT events identified by either filter are given per species in **Table S5.3**. Putative intrakingdom fungal HGT events were identified by filtering the same BLASTp output for gene models whose first top hit outside the sequence’s source species was fungal in origin but not from the same genus (**Table S5.3)**.

5.2.7 Chromosomal location of core and accessory gene models in species reference genomes

Pearson’s χ^2 tests were carried out for the global frequencies of core and accessory gene models along the subterminal regions of chromosomes, which we defined as approximately the first and last 10% of each chromosome, in each reference genome (**Table S5.4)**. Pearson’s χ^2 tests were also carried out for the frequencies of core and accessory gene models per chromosome for each reference genome (**Table S5.4)** (Agresti, 2002). The chromosomal locations of core and accessory gene models along each reference genome were visualized using the Ruby software PhenoGram (Wolfe *et al.*, 2013).

5.2.8 Distribution of knockout viability phenotypes in *Saccharomyces cerevisiae* S288C

All available knockout phenotype data for *S. cerevisiae* S288C were obtained from the Saccharomyces Genome Database (Giaever and Nislow, 2014). A reciprocal BLASTp search was carried out between all 5,815 *S. cerevisiae* S288C gene models from our *S. cerevisiae* pan-genome dataset and the reference protein set for *S. cerevisiae* S288C with an e-value cutoff of 10^{-20} to match predicted proteins to orthologs from the reference protein set. Knockout phenotype viability data, if available, was then inferred for each of our *S. cerevisiae* S288C gene models that had a reciprocal reference ortholog. Pearson's χ^2 tests were carried out for the frequencies of knockout phenotype viability in both the core and accessory genomes of *S. cerevisiae* S288C (**Table S5.5)**.

5.2.9 Distribution of dispensable pathway genes in *Saccharomyces cerevisiae* pan-genome

Data for 14 "dispensable pathway" (DP) gene clusters containing 41 genes found in *S. cerevisiae* was taken from a previously published analysis of biotin reacquisition in yeast species (Hall and Dietrich, 2007). A total of 38 DP genes were extracted from the *S. cerevisiae* S288C reference protein set, encompassing 13 of the 14 DP clusters. A reciprocal BLASTp search was performed between these genes and all 5,815 *S. cerevisiae* S288C gene models from the *S. cerevisiae* pan-genome dataset with an e-value cutoff of 10^{-20} to identify DP genes in our predicted gene model set. All 38 DP genes had a unique reciprocal match with a predicted gene model in *S. cerevisiae* S288C. A binary matrix was generated for the presence/absence of syntenic orthologs of DP genes from *S. cerevisiae* S288C in the *S. cerevisiae* pan-genome dataset (**Table S5.5)**.

5.2.10 Distribution of biosynthetic gene clusters in *Aspergillus fumigatus* pan-genome

Data for 33 known biosynthetic gene clusters (BGCs) encompassing 307 genes in *Aspergillus fumigatus* Af293 was obtained from a previous analysis of secondary metabolism in *A. fumigatus* (Lind *et al.*, 2018). *Aspergillus fumigatus* Af293 gene models from the *A. fumigatus* pan-genome dataset were matched to their homologs from the

reference gene data set using a reciprocal BLASTp search with an e-value cutoff of 10^{-20}. A binary matrix was constructed for the presence/absence of syntenic orthologs of the 307 putative BGC genes from *A. fumigatus* Af293 within the *A. fumigatus* pan-genome dataset (**Table S5.5)**.

5.3 Results

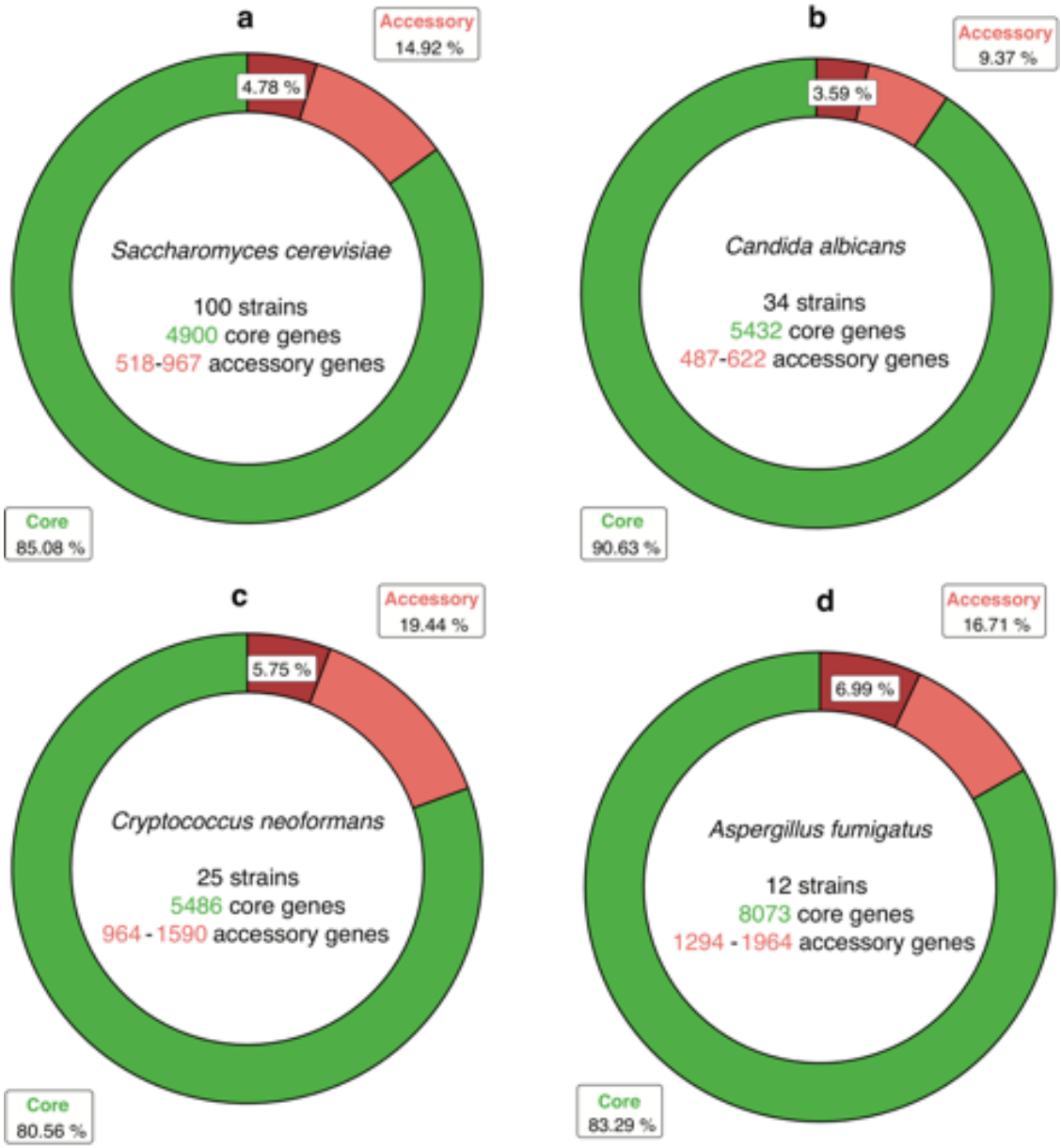


Figure 5.2. Pan-genomes of four fungal species. **A**: *Saccharomyces cerevisiae*, **B**: *Candida albicans*, **C**: *Cryptococcus neoformans* var. *grubii*, **D**: *Aspergillus fumigatus*. Ring charts represent the total number of gene models in pan-genome complements expressed as a proportion of total pan-genome size. Section in dark-red with unlabelled percentage represents duplicated core gene models in accessory genome.

Table 5.1. Pan-genomes of four fungal species. GMs: gene models. Duplicated core gene models and clusters in accessory genome given in parentheses.

Species	Strains	Core genome		Accessory genome		Pan-genome	
		GMs	Clusters	GMs	Clusters	GMs	Clusters
Saccharomyces cerevisiae	100	490,000	4,900	85,940 (27,511)	2,850 (776)	575,940	7,750
Candida albicans	34	184,688	5,432	19,098 (7,312)	1,893 (1,013)	203,786	7,325
Cryptococcus neoformans	25	137,150	5,486	33,091 (9,974)	2,698 (776)	170,241	8,193
Aspergillus fumigatus	12	96,876	8,073	19,435 (8,127)	3,002 (1,170)	116,311	11,075

Table 5.2. Number of gene models in our four fungal pan-genomes datasets with at least one annotation term per annotation type. Percentage of annotated gene models relative to pan-genomes datasets in parentheses

Species	Pfam	InterPro	GO
Saccharomyces cerevisiae	468,511 (81%)	455,582 (79%)	312,161 (54%)
Candida albicans	161,235 (79%)	155,271 (76%)	105,694 (52%)
Cryptococcus neoformans	111,305 (65%)	106,655 (63%)	72,243 (42%)
Aspergillus fumigatus	83,239 (71%)	79,231 (68%)	54,457 (46%)

5.3.1 Analysis of the *Saccharomyces cerevisiae* pan-genome

Overall, 575,940 gene models were predicted across all 100 *S. cerevisiae* strains with an average of 5,759 gene models predicted per strain (**Table 5.1, Table S5.1**). These 575,940 gene models were distributed across 7,750 unique syntenic ortholog clusters (**Table 5.1**). The core *S. cerevisiae* genome contained 4,900 gene models which were conserved across 100 *S. cerevisiae* strains (490,000 gene models in total, 85% of the total

species pan-genome). For individual strain genomes, this corresponded to between 83% to 90% of their total predicted gene model content (**Figure 5.2a**, **Table S5.1**). The remaining 85,940 predicted gene models were accessory gene models, distributed across 2,850 clusters, with strain accessory genome sizes ranging from 518 to 967 gene models per *S. cerevisiae* strain (average size = ~859 gene models). Further analysis of the *S. cerevisiae* species accessory genome identified that ~32% of accessory gene models (776 clusters, 4.77% of the total species pan-genome) were duplicates of core gene models conserved across one or more strains. This corresponded to an average of 275 gene models per *S. cerevisiae* strain, and 27,511 gene models in total (**Table 5.1**, **Table S5.1**). Overall, 455 syntenic clusters (encompassing 45,045 accessory gene models) were missing a syntenic ortholog in only one other strain and 1,416 accessory gene models were singletons. Analysis of the distribution of orthologs within the *S. cerevisiae* accessory genome using the R package UpSetR showed that the most frequent sets are singleton gene models or syntenic clusters missing a syntenic ortholog in one strain, with YPS163 having the most singleton genes (74 in total) (**Figure S5.3**). Other strains (e.g. YJM1477) lacked singleton gene models altogether (**Figure 5.3)**. There were 13,756 gene models (from 1,935 syntenic clusters) which did not have a syntenic ortholog in *S. cerevisiae* S288C. Of these non-reference gene models, 1,385 were singleton gene models found only in one strain. The widest-distributed non-reference gene model was present in 93 strains and there was no accessory gene model solely missing from *S. cerevisiae* S288C. YPS163 had the smallest accessory genome of the 100 yeast strains (518 gene models) and YJM271 had the largest (967 gene models) (**Figure 5.3)**.

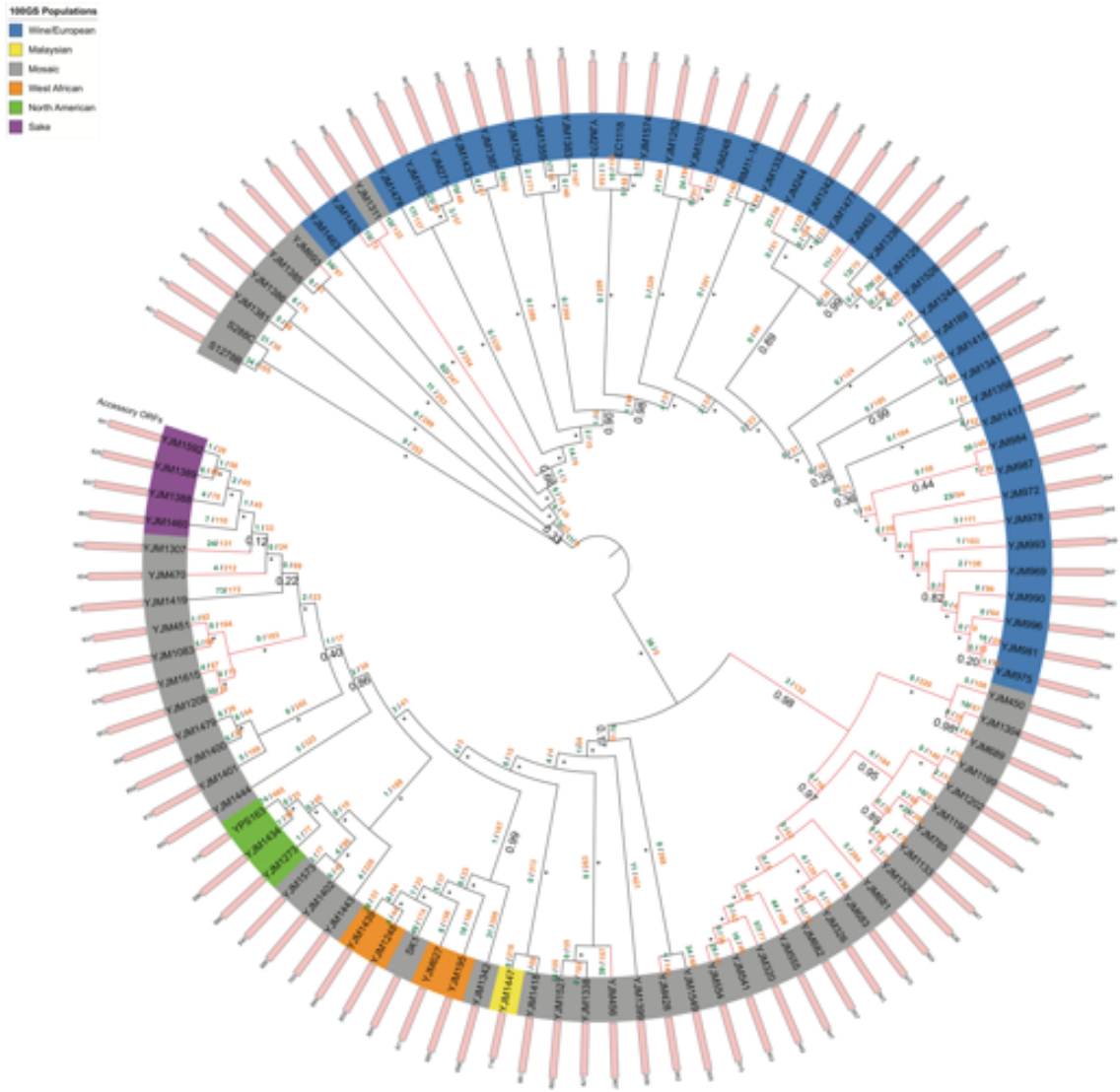


Figure 5.3. Approximate maximum-likelihood supermatrix phylogeny of *Saccharomyces cerevisiae* pan-genome dataset based on 4,311 core ortholog clusters. *S. cerevisiae* populations as assigned by Strope *et al.* (2015), clinical strains designated by red branches. Numbers below branches refer to Shimodaira-Hasegawa local supports, maximum supports indicated by an asterisk (*). Dollo parsimony analysis of gene model gain/loss events annotated above branches in green and orange, respectively.

Phylogenomic reconstruction of all 100 *S. cerevisiae* strains resolved two major groups; a clade containing strains and mosaics derived from Malaysian, West African, North American and sake populations and a clade containing strains and mosaics derived from wine/European populations (**Figure 5.3)**. Each of the non-mosaic populations as assigned by Strope *et al.* (2015) present in the dataset (except the singleton Malaysian strain YJM1447) resolved to a monophyletic geographical group (Strope *et al.*, 2015); the placement of the mosaic laboratory strain SK-1 in a West African clade is consistent with its West African origin (Warringer *et al.*, 2011), and the clinical mosaic strain

YJM1311 was of predominantly wine/European ancestry hence its placement at the base of the wine/European clade (Strope *et al.*, 2015) (**Figure 5.3)**. Many of the remaining mosaic strains branched close to non-mosaic clades which shared their dominant population fraction as determined by Strope *et al.* (2015); for example, many of the clinical mosaic strains placed adjacent to the sake clade had predominantly sake population ancestry (Strope *et al.*, 2015) (**Figure 5.3)**. Three strains (YJM248, YJM1252, YJM1078) identified by Strope *et al.* (2015) as having an higher relative proportion of introgressed genes than other *S. cerevisiae* strains (potentially arising from recent hybridization with *Saccharomyces paradoxus*) formed a monophyletic branch within the previously described wine/European clade (Strope *et al.*, 2015).

5.3.2 Analysis of the *Candida albicans* pan-genome

A total of 203,786 gene models were predicted across all 34 *C. albicans* strain genomes, with an average of 5,993 gene models predicted per strain, distributed across 7,325 unique syntenic ortholog clusters (**Table 5.1**, **Table S5.1**). The core *C. albicans* genome contained 5,432 gene models which were conserved across 34 *C. albicans* strains (184,688 in total, 90% of the total species pan-genome). This corresponded to between 89% and 91% of the total predicted gene models for each strain genome (**Figure 5.2b**, **Table S5.1**). The remaining 19,098 predicted gene models were accessory gene models, distributed across 1,893 clusters, with strain accessory genome sizes ranging from 487 to 622 gene models per *C. albicans* strain (average size = ~561 gene models) (**Table 5.1**, **Table S5.1**). Further analysis of the *C. albicans* species accessory genome identified that ~38% of accessory gene models (1,013 clusters, ~3.59% of the total species pan-genome) were duplicates of core gene models conserved across one or more strains. This corresponded to an average of 215 gene models per *C. albicans* strain, and 7,312 gene models in total (**Table 5.1**, **Table S5.1**). Of the 19,098 *C. albicans* accessory gene models identified, 3,624 accessory gene models (from 268 syntenic clusters) were missing a syntenic ortholog in only one other strain while 928 gene models were singletons. UpSet analysis of the distribution of orthologs within the *C. albicans* accessory genome showed that 1,056 gene models (32 syntenic clusters) from 33 *C. albicans* strains were missing an ortholog in *C. albicans* WO-1 and *C. albicans* 3153A had 53 putative gene models with no ortholog in any other strain (**Figure S5.4**). SC5314 had the smallest number of singleton gene models (nine in total). *C. albicans* A48 had the largest accessory genome

(622 gene models) and *C. albicans* Ca6 had the smallest (487 gene models) (**Figure 5.4)**. Phylogenomic reconstruction of all 34 *C. albicans* strains resolved two main groups when rooted at the midpoint; one containing the exemplar *MTL*-homozygous strain WO-1 and a ladderized group containing the reference strain SC5314 (**Figure 5.4)**.

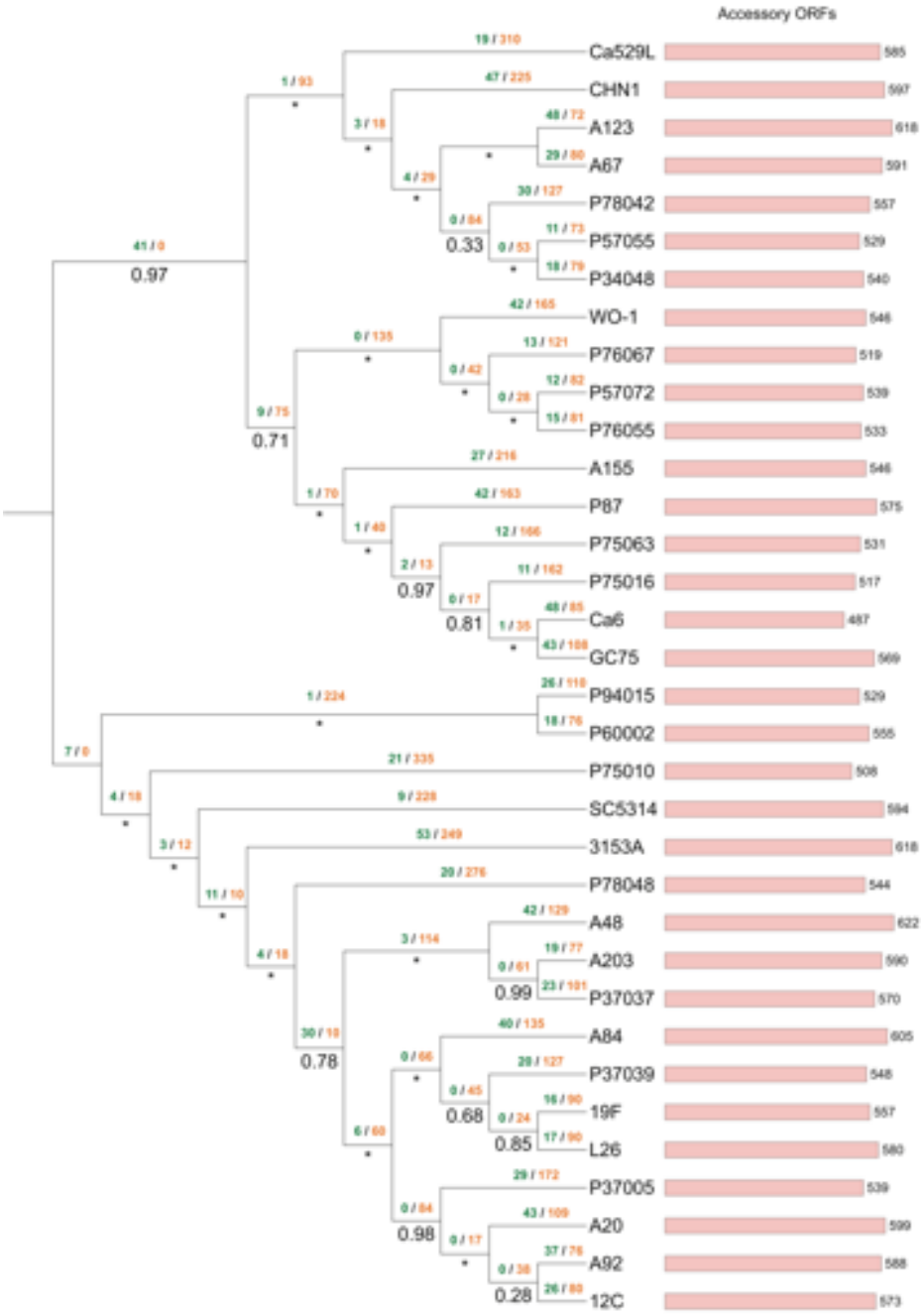


Figure 5.4. Approximate maximum-likelihood supermatrix phylogeny of *Candida albicans* pan-genome dataset based on 4,327 core ortholog clusters. Numbers below branches refer to Shimodaira-Hasegawa local supports, maximum supports indicated by an asterisk (*). Dollo parsimony analysis of gene model gain/loss events annotated above branches in green and orange, respectively.

5.3.3 Analysis of the *Cryptococcus neoformans* var. *grubii* pan-genome

A total of 170,241 gene models were predicted across all 25 *C. neoformans* var. *grubii* strain genomes, with an average of 6,809 gene models predicted per strain, distributed across 8,193 unique syntenic ortholog clusters (**Table 5.1**, **Table S5.1**). The core *C. neoformans* var. *grubii* genome contained 5,486 gene models which were conserved across 25 *C. neoformans* var. *grubii* strains (137,150 in total, 80% of the total species pan-genome). This corresponded to between 76% and 85% of the total predicted gene models for each strain genome (**Figure 5.2c**, **Table S5.1**). The remaining 33,091 predicted gene models were accessory gene models distributed across 2,698 clusters, with strain accessory genome sizes ranging from 964 to 1654 gene models per *C. neoformans* var. *grubii* strain (average size = ~1,334 gene models) (**Table S5.1**). Detailed analysis of the *C. neoformans* var. *grubii* species accessory genome identified that ~29% of accessory gene models (776 clusters, ~5.8% of the total species pan-genome) were duplicates of core gene models conserved across one or more strains. This corresponded to an average of ~391 gene models per *C. neoformans* var. *grubii* strain, and 9,794 gene models in total (**Table 5.1**, **Table S5.1**). Overall 674 *C. neoformans* var. *grubii* clusters (encompassing 16,032 accessory gene models) were missing a syntenic ortholog in only one other strain and 668 accessory gene models were singletons. UpSet analysis of the distribution of orthologs within the *C. neoformans* var. *grubii* accessory genome showed that 3,600 gene models (150 syntenic clusters) from 24 *C. neoformans* var. *grubii* strains were missing an ortholog in *C. neoformans* var. *grubii* MWRSA852, whereas the *C. neoformans* var. *grubii* A1358 genome had 49 putative gene models with no ortholog in any other strain (**Figure S5.5)**. KN99 had no singleton gene models, but it should be noted that that strain is an isogenic derivative of the reference H99 strain. *C. neoformans* var. *grubii* H99 itself had the largest accessory genome (1590 gene models) and *C. neoformans* var. *grubii* MW-RSA852 had the smallest (964 gene models) (**Figure 5.5)**. The most frequent sets found in the accessory genome include both singleton genes and clusters missing orthologs from one or two strains. Phylogenomic reconstruction of all 25 strains using a 47,811-site amino acid supermatrix derived from the core *C. neoformans* var. *grubii* genome resolved two monophyletic groups when rooted at the midpoint (**Figure 5.5)**.

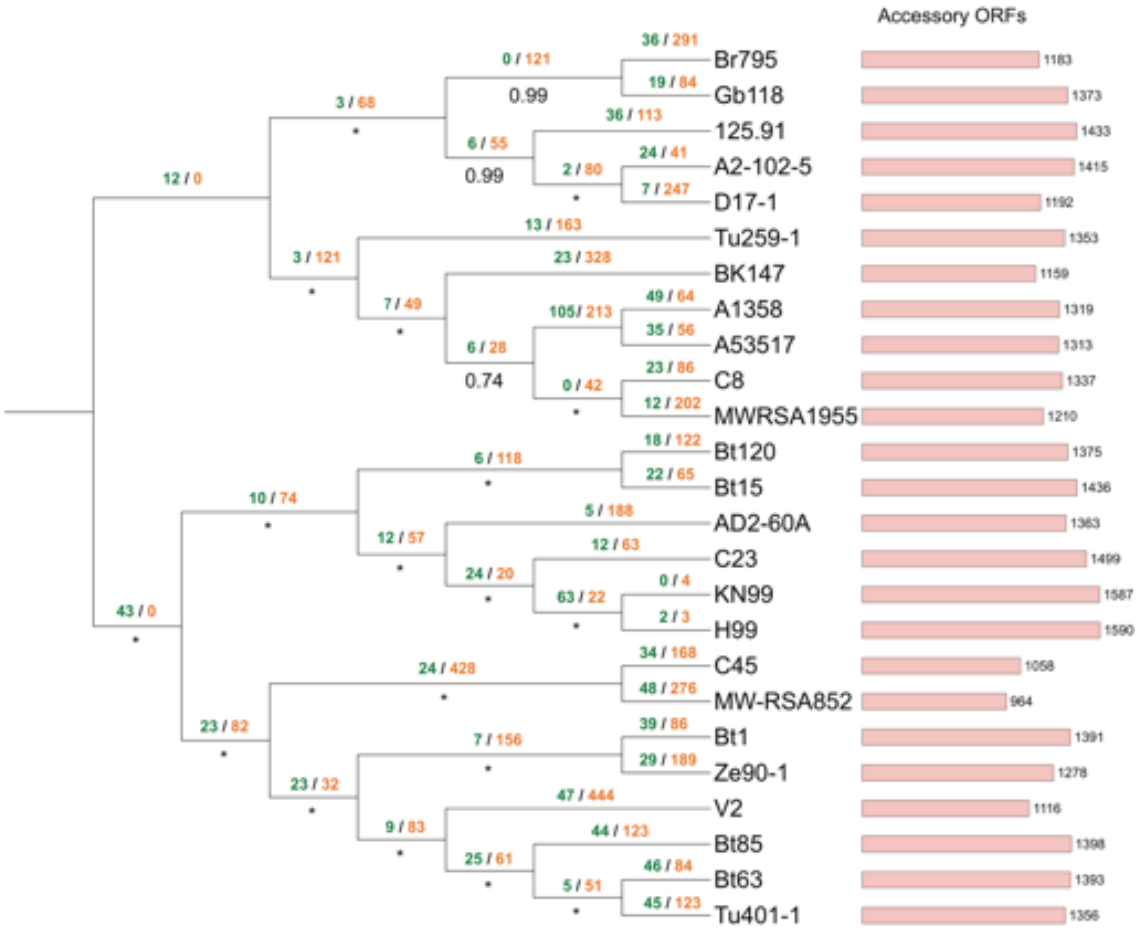


Figure 5.5. Approximate maximum-likelihood supermatrix phylogeny of *Cryptococcus neoformans* var. *grubii* pan-genome dataset based on 4,512 core ortholog clusters. Numbers below branches refer to Shimodaira-Hasegawa local supports, maximum supports indicated by an asterisk (*). Dollo parsimony analysis of gene model gain/loss events annotated above branches in green and orange, respectively.

5.3.4 Analysis of the *Aspergillus fumigatus* pan-genome

A total of 116,311 gene models were predicted across all 12 *A. fumigatus* strain genomes, distributed across 11,075 unique syntenic ortholog clusters, with an average of 9,692 gene models predicted per strain. The core *A. fumigatus* genome contained 8,073 core gene models which are conserved across 12 *A. fumigatus* strains (96,876 in total, 83% of the total species pan-genome). This corresponded to between 80% and 86% of the total predicted gene models for each strain genome (**Figure 5.2d**, **Table S5.1**). The remaining 19,435 predicted gene models were accessory gene models distributed across 3,002 clusters, with strain accessory genome sizes ranging from 1,294 to 1,964 gene models per *A. fumigatus* strain (average size = ~1,619 gene models) (**Table S5.1**). Detailed analysis of the *A. fumigatus* species accessory genome identified that ~41% of

accessory gene models (1,170 clusters, ~6.9% of the total species pan-genome) were duplicates of core gene models that were conserved across one or more strains. This corresponded to an average of 677 gene models per *A. fumigatus* strain, and 8127 gene models in total. Overall, 7,953 gene models (from 958 syntenic clusters) were missing a syntenic ortholog in only one other strain whereas 723 gene models were singletons.

UpSet analysis of the ortholog distribution in the *A. fumigatus* accessory genome found that 2,167 gene models (197 syntenic clusters) from 11 *A. fumigatus* strains were missing an ortholog in *A. fumigatus* IFISWF4 and the reference *A. fumigatus* Af293 genome has 150 putative gene models with no ortholog in any other strain (**Figure S5.6)**. The latter may be due to a lower degree of strain sampling within the *A. fumigatus* dataset or the reference genome having a higher-quality assembly than other strains of *A. fumigatus*. IFISWF4 has the smallest number of singleton gene models (nine in total). *A. fumigatus* Af293 has the largest accessory genome (1,964 gene models) and *A. fumigatus* HMRAF706 has the smallest (1,294 gene models) (**Figure S5.6)**. Phylogenomic reconstruction of all 12 strains using a 20,760-site amino acid supermatrix derived from the core *A. fumigatus* genome resolved two monophyletic groups when rooted at the midpoint, one containing both International Space Station strains and *A. fumigatus* Af10 and one containing all three environmental strains as well as *A. fumigatus* Af293 and Af210 (**Figure 5.6)**. The placement of the two ISS strains as well as the aforementioned individual clinical strains is in relative agreement with the most extensive intraspecific *A. fumigatus* phylogeny published (Knox *et al.*, 2016).

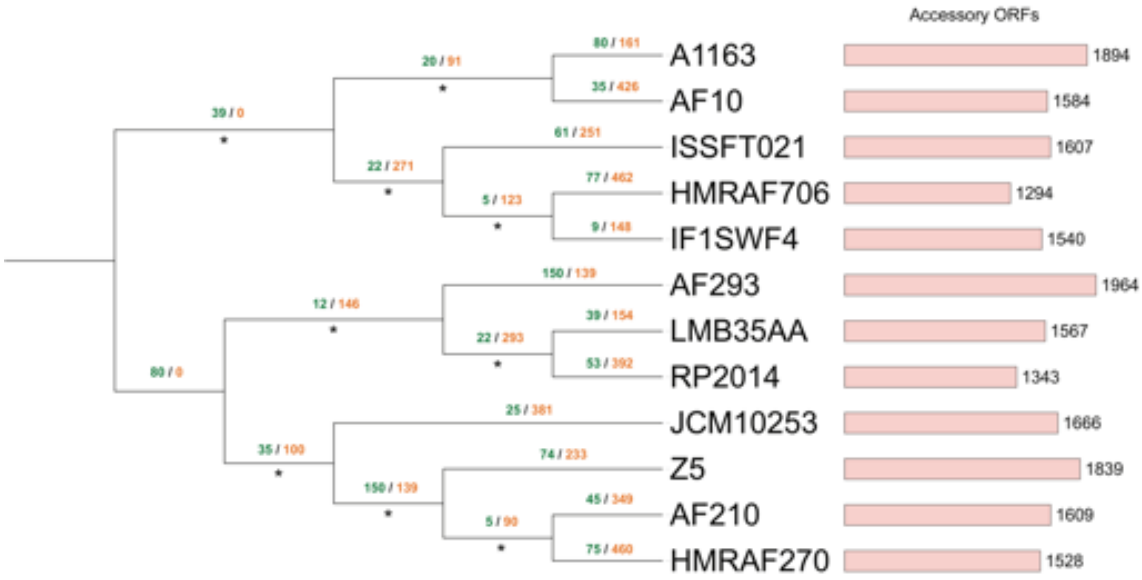


Figure 5.6. Approximate maximum-likelihood supermatrix phylogeny of *Aspergillus fumigatus* pan-genome dataset based on 5,724 core ortholog clusters. Numbers below branches refer to Shimodaira-Hasegawa local supports, maximum supports indicated by an asterisk (*). Dollo parsimony analysis of gene model gain/loss events annotated above branches in green and orange, respectively.

5.3.5 Functional analyses of fungal species pan-genomes

5.3.5.1 Gene ontology enrichment in fungal core and accessory genomes

Analysis of the distribution of GO terms in fungal core genomes shows that many housekeeping biological processes such as translation, nucleic acid metabolism and oligopeptide metabolism are significantly over-represented in each species ($p < 0.05$) (**Table S5.2)**. Furthermore, molecular function terms for enzymatic and nucleic acid binding activity are also significantly over-represented (**Table S5.2)**. In fungal accessory genomes terms relating to transport and localization of proteins, carbohydrate metabolism as well as protein modification and carboxyl acid metabolism are significantly over-represented in many species (**Table S5.2)**. Terms relating to housekeeping processes are significantly under-represented in fungal accessory genomes compared to core genomes. There are no common or synonymous cellular component or molecular function terms that are significantly under-represented across all four fungal accessory genomes in our analysis. However, terms relating to the functions of intracellular membrane-bound organelles are significantly over-represented in the accessory genomes of both *C. neoformans* var. *grubii* and *A. fumigatus* (**Table S5.2)**.

Many broad and granular housekeeping terms relating to nucleic acid and protein biological processes are significantly over-represented within the core genome of *S. cerevisiae* (**Table S5.2)**. In addition to transport processes, genes potentially involved in vitamin metabolism and protein dephosphorylation are significantly over-represented within the core genome of *S. cerevisiae*. Similar terms are also significantly over-represented within the core genome of *C. albicans* (**Table S5.2)**. The *C. neoformans* var. *grubii* core genome is significantly over-represented in some unique terms involved in regulation of homeostasis and biological quality, functional pathways such as the unfolded protein response (UPR) pathway as well as signal transduction (**Table S5.2)**. There are fewer terms that are significantly over-represented within the *C. neoformans* var. *grubii* accessory genome than in the other fungal accessory genomes in this study. Those terms that are significantly over-represented in the *C. neoformans* var. *grubii* accessory genome are also found elsewhere; e.g. transport. The core *A. fumigatus* genome is significantly over-represented in terms related to small molecule biosynthesis and other biosynthetic processes (**Table S5.2)**. Within the *A. fumigatus* core genome terms relating to vesicle-mediated transport and carboxylic acid metabolism are significantly over-represented, these terms are also significantly over-represented in the *S. cerevisiae* core genome.

5.3.5.2 Ancestral origin of fungal core and accessory genomes

The ancestral origin of fungal core and accessory genomes was inferred *via* BLASTp searches ($1e^{-20}$) of fungal gene models against >5 million prokaryotic sequences from >1,500 bacterial and archaeal genomes. Gene models which had hits with prokaryotic sequences exclusively were classified as having originated within the prokaryotes (broken down further by prokaryotic domain in **Table S5.3)**, and gene models that lacked a BLASTp hit against the prokaryotic database were classified as having originated within the eukaryotes. Using these criteria, for each fungal pan-genome dataset between 69-77% of all gene models were inferred as eukaryotic in origin. Similar proportions of gene models inferred as having originated within eukaryotes were also observed in fungal core genomes. Higher proportions of gene models with a putative origin within eukaryotes was observed in fungal accessory genomes (74-81% of all accessory gene models in each species). Statistical analysis of the ancestral history of each fungal species pan-genome found that each fungal accessory genome was

significantly enriched for genes of eukaryotic origin and each fungal core genome was significantly enriched for genes of prokaryotic origin ($p < 0.05$) (**Table S5.3)**.

5.3.5.3 Interdomain and intrakingdom HGT into fungal accessory genomes

Systematic screening for interdomain HGT events in each fungal accessory genome revealed small numbers of putative HGT events from prokaryote sources per species, ranging from a single event in the *C. albicans* accessory genome to 11 events in the *A. fumigatus* accessory genome (**Table S5.3)**. The distribution of these putative HGT genes in fungal accessory genomes varies from strain-unique singleton genes (particularly in *S. cerevisiae*) to more widely-distributed genes (as seen in *C. neoformans* and *A. fumigatus*) (**Table S5.3)**. The majority of potential prokaryote donors are soil-dwelling bacteria, such as *Clostridium pasteurianum* (a donor to the *A. fumigatus* accessory genome) and *Acinetobacter pittii* (a donor to the *S. cerevisiae* accessory genome). We then applied a similar screen for recent HGT from other fungal species, which suggested up to 8% of fungal accessory genomes may have arisen *via* intrakingdom HGT. The largest extent of such intradomain HGT appeared to have occurred into the accessory genomes of *C. neoformans* and *A. fumigatus* (420 and 391 potential events, respectively) (**Table S5.3)**. In each accessory genome, putative HGT-derived gene models appear to have been transferred mainly from closely-related species or species that share similar niches. For example, *A. fumigatus* is a potential donor of three *C. albicans* accessory gene models (**Table S5.3)**. However, further comprehensive investigations are required to confidently confirm that these HGT events are *bona fide*.

5.3.5.4 Chromosomal location of core and accessory genomes in fungal reference genomes

Between 17-21% of all predicted gene models for each fungal reference strain lie in the subterminal regions of that strain's genome. Approximately 15% of all core gene models in both *S. cerevisiae* S288C and *C. neoformans* var. *grubii* H99 are found in their subterminal regions, whereas this proportion is higher in *C. albicans* SC5314 and *A. fumigatus* Af293 (~21% and ~18% of all core gene models, respectively). *Candida albicans* SC5314 has a lower proportion of accessory gene models (115 of 594 gene models, ~19% of its total accessory genome) found in subterminal regions than the other three fungal species, where that proportion is ~28-33% of their total accessory genomes.

There is a statistically-significant bias ($p < 0.05$) towards accessory gene models in the subterminal regions of *Saccharomyces cerevisiae* S288c, *Cryptococcus neoformans* var. *grubii* H99 and *Aspergillus fumigatus* Af293 with a corresponding bias ($p < 0.05$) towards core gene models in the non-subterminal regions of each genome (**Table S5.4)**. In contrast, there is no significant pattern in the distribution of accessory gene models in *C. albicans* SC5314, and instead its subterminal regions are significantly enriched for core gene models ($p < 0.05$) (**Table S5.4)**. Statistical analysis of core and accessory gene model enrichment per chromosome in each reference genome found that at least one chromosome was significantly enriched for core gene models and another chromosome was significantly enriched for accessory gene models per genome ($p < 0.05$) (**Table S5.4)**. The number of chromosomes per genome that were significantly biased towards either core or accessory gene models ranged from two in *C. albicans* SC5314 (chromosomes 2 and 7) to six in *S. cerevisiae* S288C (chromosomes I-III, VI, VIII and XIII) (**Table S5.4)**. Visualizing chromosomal plots showed that clustering of accessory genes mostly occurred in subterminal regions of fungal genomes (**Figures S5.7a-d**). There are some exceptions: some chromosomes in *Saccharomyces cerevisiae* S288c, *Cryptococcus neoformans* var. *grubii* H99 and *Aspergillus fumigatus* Af293 had at least one larger accessory gene cluster closer to the chromosomal midpoint (**Figures S5.7a, c-d**). In contrast, there appeared to be no major clustering of accessory genes in any chromosome in *Candida albicans* SC5314 (**Figure S5.7b**).

5.3.5.5 Knockout viability of core and accessory genes in Saccharomyces cerevisiae S288C

A total of 5,343 predicted *S. cerevisiae* S288C gene models from the species pan-genome dataset, encompassing 4,730 core gene models and 613 accessory gene models, were assigned their reference homolog's corresponding knockout viability phenotype. The remaining 472 predicted gene models from *S. cerevisiae* S288C did not have a knockout viability phenotype assigned to them, either due to the lack of a unique reciprocal BLASTp hit or a lack of viability data for the reference homolog (**Table S5.5)**. Those *S. cerevisiae* S288C gene models that had knockout phenotype data were predominantly knockout-viable; ~79% of annotated core gene models and ~88% of annotated accessory gene models had a reciprocal reference homolog with a viable knockout phenotype (**Table S5.5)**. There was no significant bias in the distribution of

knockout viability within the core *S. cerevisiae* S288C genome; i.e. the core genome was enriched for neither knockout-viable or knockout-inviable gene models (of those which had knockout phenotype data available) (**Table S5.5)**. The *S. cerevisiae* S288C accessory genome however was over-represented with for knockout-viable gene models ($p < 0.05$) (**Table S5.5)**.

5.3.5.6 Dispensable pathway gene clusters in the Saccharomyceres cerevisiae pan-genome

All 38 reference DP genes had a unique reciprocal homolog within the set of predicted *S. cerevisiae* S288C gene models taken from our pan-genome dataset (**Table S5.5)**. One of the 13 reference DP clusters was syntentically-conserved within all strains in the *S. cerevisiae* pan-genome dataset; a three-member *GAL* cluster involved in galactose utilization. Some clusters are widely-conserved within the dataset, but are missing a member gene in a small number of strains; these include a three-member *BIO* cluster that mediates biotin uptake, a *SNO1-SNZ1* vitamin B6 metabolism cluster and a large six-member *DAL-DCG* cluster that enables utilization of allantoin as a nitrogen source (**Table S5.5)**. Other clusters had more patchy distribution within the species pan-genome, most notably a three-member *ARR* gene cluster which confers arsenic resistance was missing a member gene (*ARR3*) in 49 out of 100 strains (**Table S5.5)**. Some clusters, such as a four-member *FIT*/*FRE* iron uptake cluster, are completely missing in a small number of strains (**Table S5.5)**.

5.3.5.7 Biosynthetic gene clusters in the Aspergillus fumigatus pan-genome

A total of 307 known biosynthetic genes from 33 BGCs in *A. fumigatus* Af293 had a unique reciprocal homolog within the set of predicted *A. fumigatus* Af293 gene models from the *A. fumigatus* pan-genome (Lind *et al.*, 2018). A total of 240 of the 307 known biosynthetic genes were core genes found in all 12 *A. fumigatus* strains, none of which were unique to *A. fumigatus* Af293 alone (**Table S5.5)**. There were 14 *A. fumigatus* BGCs that were completely conserved (i.e. all genes within that cluster are core genes), which included known mycotoxin-producing BGCs such as fumagillin and gliotoxin clusters (**Table S5.5)**. Other BGCs were found to have one or two genes missing, potentially due to synteny loss or pseudogenization. Some BGCs showed far more variable distribution within the *A. fumigatus* pan-genome; for example, a polyketide

synthase (PKS) cluster was wholly conserved in 4 strains (Af293, Z5, HMRAF270 and JCM10253) and absent or translocated in the other 8, and a fusarielin-like cluster was completely absent from A1163 and only partially present in some strains but was wholly conserved in others (**Table S5.5)**.

5.4 Discussion

5.4.1 Applying genomic context in eukaryotic pan-genome analysis

To investigate pan-genomic structure within four fungal species, we adapted a method previously used in bacterial pan-genome analysis and implemented in PanOCT (**Pan-**genome **O**rtholog **C**lustering **T**ool) (Fouts *et al.*, 2012). Our rationale for using this method to construct species pan-genomes was that it allowed us to investigate intraspecific variability on a gene-to-gene level, as opposed to defining core and accessory genomes based on families of related gene models (e.g. a "core" gene family may be present in all strains of a species, but the number of genes belonging to that family will usually vary between strains). This allowed us to see which genes and biological functions are relatively conserved in their distribution and which have varying expansion and distribution in fungal species. A similar approach was used in a previous analysis of genome variation in *Saccharomyces* species, but was limited to assessing syntenic conservation of reference homologs using immediately-adjacent genes (Bergström *et al.*, 2014). To ensure consistency between strain genomes in each of our datasets we constructed a custom gene model prediction pipeline which used three different predictive methods to generate a unique set of predicted gene models and their genomic locations (i.e. no isoforms) per strain genome (**Figure S5.2a**) (Slater and Birney, 2005; Ter-Hovhannisyan *et al.*, 2008; Haas *et al.*, 2013). As our definition of what constitutes a "core" or "accessory" gene model is quite stringent compared to other pan-genome analyses, we also developed a post-processing pipeline which attempted to account for loss of microsynteny between fungal strain genomes and to also examine the extent of duplication of core genome content within fungal accessory genomes.

5.4.2 The pan-genomes of four model fungi

We chose to investigate the potential pan-genomic structure of four model fungal species; *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. In addition to their impact on human lifestyle each species chosen is a model organism for fungal evolutionary biology, genomics and comparative genomics. *Saccharomyces cerevisiae* was the first eukaryote to have its genome sequenced, and the other three species each had their genome sequenced during the initial wave of fungal genomics research in the early-to-mid 2000s (Goffeau *et al.*, 1996; Jones *et al.*, 2004; Nierman *et al.*, 2005; Cock *et al.*, 2009; Janbon *et al.*, 2014;

McCarthy and Fitzpatrick, 2017a). Our selection covers fungal species with different genomic characteristics; *S. cerevisiae* has undergone ancestral whole-genome duplication and *C. albicans* has an alternative genetic code (Santos and Tuite, 1995; Wolfe, 2015), whereas *Cr. neoformans* and *A. fumigatus* are more intron-dense than either *S. cerevisiae* or *C. albicans* and extensive alternative splicing occurs in *Cryptococcus* species (Stajich, Dietrich and Roy, 2007; Gonzalez-Hilarion *et al.*, 2016). Our selection also covers fungal species with different evolutionary histories. *S. cerevisiae*, *C. albicans* and *A. fumigatus* are members of the Ascomycota phylum of fungi; the former two are closely-related members of the Saccharomycotina subphylum which includes many typical commensal and pathogenic yeasts that reproduce by budding while *A. fumigatus* is a member of the large Pezizomycotina subphylum of filamentous fungi (McCarthy and Fitzpatrick, 2017a). *Cryptococcus neoformans* var. *grubii* superficially resembles many yeast species and also replicates by budding, but is a member of the Basidiomycota phylum and is more closely related to multicellular fungi within the Agaricomycotina subphylum than other yeast species (McCarthy and Fitzpatrick, 2017a). Genome assemblies available on GenBank for each species at the time of writing range from 12 for *A. fumigatus* to >400 for *S. cerevisiae* (Peter *et al.*, 2018).

Our species pan-genome for *Saccharomyces cerevisiae* was constructed using genomic data from 100 strains, 99 of which were previously included in the "100-genomes strains" (100GS) resource (**Table S5.1**) (Strope *et al.*, 2015). The resource includes 7 *S. cerevisiae* genomes sequenced prior to 2015 and 93 *S. cerevisiae* genomes sequenced *de novo* by the 100GS authors, taken from diverse genotypic and phenotypic backgrounds (populations referred to henceforth are as assigned by the 100GS authors after Liti et al. (2009)) (Liti *et al.*, 2009; Strope *et al.*, 2015). The resource covers strains from laboratory, biotech, clinical and wild populations, which makes it an excellent dataset for carrying out *S. cerevisiae* population genomics and pan-genomics studies of this kind. In their analysis, the 100GS authors screened *S. cerevisiae* strains for aneuploidy, introgressed genes, phenotypically-relevant single-nucleotide polymorphisms and non-reference genomic content (Strope *et al.*, 2015). The 100GS authors also assessed levels of resistance to environmental stresses such as sulphite and copper resistance, as well as fungicides such as ketoconazole (Strope *et al.*, 2015).

A more recent study of 1,011 *S. cerevisiae* genomes included an analysis of the pan-genome of *S. cerevisiae* in which the authors of that study detected non-reference genomic content by aligning strain genomes to the S288C genome using BLASTn and

extracting and annotating unique non-reference genes using an integrative multi-method procedure (Yue *et al.*, 2017; Peter *et al.*, 2018). Notably, despite a ten-fold difference in the number of input strains and different methods of identifying core and accessory genome content both their study (4,940 core genes) and our own (4,900 core gene models) predict a similar-sized core *S. cerevisiae* genome (Peter *et al.*, 2018). The 1,011-genome study predicted an almost identical accessory genome to our analysis also; they identified 2,856 accessory genes with varying distribution across 1,011 genomes (Peter *et al.*, 2018), whereas we identified an accessory genome of 2,850 genes for our pangenome dataset. These 1,011-genome study also observed a number of evolutionary and functional trends within the *S. cerevisiae* accessory genome; accessory genes were clustered within the subterminal regions of *S. cerevisiae* genomes and some accessory genes may have originated *via* HGT from divergent yeast species or other fungi (Peter *et al.*, 2018). We observe similar trends in our analysis of the *S. cerevisiae* accessory genome.

For the remaining three species, we constructed species pan-genome datasets based on strain genome assemblies that were available from GenBank at the start of our analyses. For each of these datasets, we attempted to sample strain genomes with as many diverse characteristics (e.g. geographical location, phenotype) as was possible with the genome assembly data available. Although there are a smaller number of strains sampled for these species pan-genomes, the sizes of these species' core and accessory genomes are in line with our analysis of *S. cerevisiae* as well as larger analyses of species pan-genomes in fungi and other taxa. The *Candida albicans* species pan-genome dataset was constructed using data from 34 strains, predominantly clinical in origin, including both homozygous and heterozygous *MTL* mating-type strains (**Table S5.1**) (Lockhart *et al.*, 2002). A substantial amount of genome assembly data available for *C. albicans* comes from strains isolated in hospitals; of the 34 strains in our dataset, 14 strains were clinical isolates from the US alone (**Table S5.1**). A number of other strains were isolated from European and Middle East sources, but for 13 strains no information was available on the isolate source for the genome from GenBank. Perhaps as a consequence of a lower degree of environmental diversity due to sampling primarily clinical strains, the *C. albicans* pan-genome has the smallest proportion of accessory gene content of the four species analysed in this study (~9% of the entire species pan-genome). The *C. albicans* pan-genome also has the lowest degree of variation in accessory genome size between individual strains of the four species analysed (**Figure 5.2b**, **Figure 5.4)**. The UpSet distribution of the *C. albicans* accessory genome illustrates this lower degree of variability within the *C.*

albicans pan-genome, as the most frequent sets are either singleton clusters or clusters that are missing an ortholog from one strain (**Figure S5.4**). Despite this caveat however, the *C. albicans* pan-genome otherwise exhibits many of the same functional and evolutionary trends seen in the other three species we have investigated (as detailed below). With a broader sampling of strains found outside of a clinical context, a more accurate picture of the size of the *C. albicans* accessory genome will be attained.

In contrast to *C. albicans*, both our *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus* pan-genome datasets were constructed using a diverse array of strain genomes taken from both clinical and wild environments. The *C. neoformans* var. *grubii* pan-genome dataset was constructed using clinical strain genomes isolated predominantly from HIV+ patients from the US and Botswana predominantly and wild-type strains sampled from Southern Africa sources (**Table S5.1**). *C. neoformans* var. *grubii* has the largest proportion of accessory genes of the four species analysed (~20% of the entire species pan-genome). As *C. neoformans* is an intracellular pathogen in humans, it has to adapt to extreme variations in environmental stresses in order to survive. This is thought to lead to the high level of genomic rearrangement and instability seen in *C. neoformans* (Fraser *et al.*, 2005). It is possible that this in turn creates more novel genetic content, which may explain the higher level of accessory genome content seen in *C. neoformans* var. *grubii*. Genomic instability as a result of pathogenic lifestyle fuelling pan-genome evolution has previously been observed in the wheat pathogen *Zymoseptoria tritici* (Plissonneau, Hartmann and Croll, 2018). The *A. fumigatus* pan-genome dataset was constructed using 12 strain genomes sampled from clinical environments in the UK, US and Canada, wild-type samples taken from China and from South American forest floors, and two strains isolated from surfaces within the International Space Station (Knox *et al.*, 2016) (**Table S5.1**), Approximately 15% of the *A. fumigatus* pan-genome is made up of accessory gene content, which is predominantly clustered in the subterminal regions of chromosomes (discussed below). There is a greater degree of variation in the accessory genome sizes of individual *A. fumigatus* strains than in the other species analysed, we believe that this is primarily an artefact of the smaller number of genomes in our dataset (at the time of writing our *A. fumigatus* dataset includes almost all strain genomes available as assembly data on GenBank).

5.4.3 Broad trends across fungal pan-genomes

5.4.3.1 Fungal core and accessory genomes enriched for potential infection and survival processes

Between 65-81% of gene models per species pan-genome had at least one Pfam domain, while the proportion of gene models with GO data was between 42-54% per species (**Table 5.2**). This variation is primarily down to a lack of human annotation for some species, and for *C. neoformans var. grubii* in particular the lack of a dedicated GO-slim dataset. This can be seen in our statistical analyses of the distribution of GO terms in individual species pan-genomes; *S. cerevisiae* currently has a far more detailed array of ontological terms than *A. fumigatus* for example (**Table S5.2)**. In spite of gaps in ontological data for some of our species of interest, there are a number of patterns we can observe across multiple species in our GO analyses of fungal core and accessory genomes as well as unique patterns of enrichment in some species. Many housekeeping terms such as translation, nucleic acid metabolism and oligopeptide metabolism are statistically over-represented in each fungal core genome we have analysed ($p < 0.05$) (**Table S5.2)**. There is an over-representation of similar cellular component terms in each of the three "yeast" core genomes (i.e. all excluding *A. fumigatus*) (**Table S5.2)**. This may reflect the morphological distinctions between these three species and *A. fumigatus*, however the lack of dedicated annotation data for *C. neoformans* var. *grubii* makes a definitive observation difficult. Terms relating to transport, localization and Crazy processes are statistically over-represented in fungal accessory genomes (**Table S5.2)**. In part this is to be expected, as many fungi have varying numbers of copies of genes involved in Crazy and transport processes (Wisecaver, Slot and Rokas, 2014). Terms relating to housekeeping processes are statistically under-represented in fungal accessory genomes, which may be due to potential gene dosage effects. The similar patterns of statistical over-representation for terms relating to intracellular membrane-bound organelles in the accessory genomes of *C. neoformans* var. *grubii* and *A. fumigatus* may reflect infection or in-host survival processes for both pathogenic species (**Table S5.2)**. Both the *C. albicans* core and accessory species genome share similarly over-represented terms to their *S. cerevisiae* counterparts, a reflection of the two species' relatively close evolutionary relationship (**Table S5.2)**.

Many of the terms that are over-represented in the *C. neoformans* var. *grubii* core genome may reflect the species' lifestyle as an intracellular pathogen (**Table S5.2)**. Such

terms include regulation of homeostasis and biological quality (e.g. cell mass), which are vital for *C. neoformans* var. *grubii* to survive the plethora of environmental stresses it encounters in the host. Similarly, unfolded protein response (UPR) is an over-represented molecular function in the *C. neoformans* var. *grubii* core genome; the UPR pathway is known to influence thermoregulation in *C. neoformans* var. *grubii* particularly during the initial infection period (Cheon *et al.*, 2014). Another over-represented term in the *C. neoformans* var. *grubii* core genome is signal transduction; many signal transduction pathways in *C. neoformans* var. *grubii* play an important role in cell differentiation as well as pathogenicity (**Table S5.2)** (Lengeler *et al.*, 2000). The core *A. fumigatus* genome is enriched for small molecule biosynthesis and other biosynthetic processes, which concurs with previous comparative studies of *Aspergillus* species (Khaldi *et al.*, 2010; Andersen *et al.*, 2013) (**Table S5.2)**. This also appears to agree with our findings of biosynthetic gene cluster conservation within the *A. fumigatus* species pan-genome (**Table S5.5)**. Both transport and localization processes are over-represented within the *A. fumigatus* accessory genome, which may have an indirect role in the infection processes of *A. fumigatus*. *Aspergillus fumigatus* strain pathogenesis may therefore be influenced by accessory genome evolution, particularly within subterminal regions (McDonagh *et al.*, 2008).

5.4.3.2 The fungal core genome is more ancient in origin than the fungal accessory genome

Our statistical analysis of the ancestral history of each fungal species pan-genome found that gene models of eukaryotic origin are statistically over-represented within fungal accessory genomes, while gene models of prokaryotic origin are statistically over-represented in fungal core genomes ($p < 0.05$) (**Table S5.3)**. In other words, genes of prokaryotic origin appear to be more likely to be syntenically-conserved and universally-retained within these fungal species (**Table S5.3)**. This appears consistent with the observation that prokaryote-derived genes in *Saccharomyces cerevisiae* are essential for survival (Cotton and McInerney, 2010). On the other hand, it appears that the accessory genome contains more genes which arose at some point during the evolution of eukaryotes and which may be more likely to be variably-retained or lost within strains of fungal species (**Table S5.3)**. This would concur with our analysis of the gains and losses

of syntenic orthologs in fungal accessory genomes, which are largely mediated at the strain level.

5.4.3.3 Horizontal gene transfer may only play a limited role in fungal pan-genome evolution

Given the extent of HGT in prokaryotes and its role in generating novel genetic content and in the evolution of prokaryotic gene families, it is likely that HGT plays a significant role in prokaryote pan-genome evolution. HGT in eukaryotes is known to be far less frequent than in prokaryotes however, so its impact on eukaryotic pan-genome evolution may be limited. We examined the extent of horizontal gene transfer into fungal accessory genomes from two potential sources of novel genetic content: prokaryotic species and other species within the fungal kingdom. A screen for interdomain HGT events in each fungal accessory genome following previous methodology (Richards *et al.*, 2011; McCarthy and Fitzpatrick, 2016), revealed low numbers of putative HGT events from prokaryote sources into fungal accessory genomes per species (**Table S5.3)**. Gene transfer between prokaryotes and eukaryotes is a subject of some controversy, with different studies suggesting that interdomain HGT is alternately non-existent or a rare but real occurrence (Marcet-Houben and Gabaldón, 2010; Ku and Martin, 2016; Martin, 2017). Regardless, from our analysis it appears that interdomain HGT is not an influencing factor on accessory genome evolution (and hence, pan-genome evolution) within fungi. We then applied a similar screen for HGT from other fungal species into fungal accessory genomes, and found that up to 8% of fungal accessory genomes may be derived from intrakingdom HGT. There are caveats to consider when interpreting this finding however; although some of these events may be genuine incidences of HGT it is equally plausible that these genes have undergone pseudogenization or have otherwise lost synteny in one or more strains/lineages. That the majority of potential donor species are close relatives in each analysis we performed may in part suggest this; for example 96 of the 102 putative HGT events into the *S. cerevisiae* accessory genome have a potential donor from the species in the same phylum (Saccharomycotina) and 379 of the 392 putative HGT events into the *A. fumigatus* accessory genome suggest transfer from other species in the Pezizomycotina subphylum (132 from *Penicillium* species alone) (**Table S5.3)**. Although there appears to be greater evidence for intrakingdom HGT having a role to play in fungal accessory genome evolution than interdomain HGT, it is

our opinion that a dedicated analysis of intrakingdom HGT in fungal accessory genomes using robust phylogenetic methods is required to test the true role of intrakingdom HGT in fungal pan-genome evolution.

5.4.3.4 Eukaryotic processes such as gene duplication may influence fungal pan-genome evolution

Between 29-41% of genes contained within fungal accessory genomes appear to be duplicates of core gene models that have undergone subsequent loss in some strains, possibly by pseudogenization, microsynteny loss, or expansion in other strains (**Table 5.1, Table S5.1**). *Cryptococcus neoformans* var. *grubii* has the smallest proportion of these duplicated core gene models (and consequently, the highest proportion of accessory gene models that have potentially arisen *via* other processes) and *A. fumigatus* has the largest (**Table S5.1**). This accounts for between 3-7% of the total size of fungal pangenomes, with the smallest proportion in *C. albicans* and the largest in *A. fumigatus* (**Figure 5.2**, **Table S5.1**). These results appear to indicate that gene duplication, which is the driving factor of gene family expansion in eukaryotes, does play an important role in the evolution of fungal accessory genomes (and pan-genomes as a whole) (Lynch and Conery, 2000; Treangen and Rocha, 2011). The larger proportion of duplicated core genes in *A. fumigatus* appears to reflect the greater extent of gene duplication and paralog diversity within that species relative to *C. neoformans* var. *grubii* and *S. cerevisiae* (Yang, Hulse and Cai, 2012). Preliminary annotation of these gene models shows that many have putative or known functions in transport and outer membrane processes, which are processes that are often mediated by expanded gene families in fungi.

Mapping the presence or absence of syntenic orthologs within fungal accessory genomes finds that for each species the majority of syntenic ortholog loss events, through chromosomal rearrangement or gene loss, or the gain of new genes has occurred within strains as opposed to more ancestral branches (**Figures 5.3-5.6**). We searched each set of singleton gene models from each reference genome against the reference protein set to assess the putative function(s) of some of these strain-unique genes. Many singleton gene models are homologous to membrane proteins, DNA/RNA-binding or transposition-related genes (e.g. *gag*/*pol* retrotransposons in *S. cerevisiae*, DDE1 transposases in *A. fumigatus*), which are usually independently expanded or redistributed within individual fungal genomes (Liti *et al.*, 2009; Perez-Nadales *et al.*, 2014). Between 30-60% of

singleton gene models within each species pan-genome dataset had at least one Pfam domain, a lower proportion than that seen in each species dataset (65-81%) as a whole, which may be another artefact of gaps in human annotation (**Table S5.2)**. Closely-related strains of many species also appear to have similar accessory genome sizes (e.g. many clades within the *S. cerevisiae* 100GS dataset, the reduced sizes of both *C. neoformans* var. *grubii* C45 and MW-RSA852 relative to most other strains) (**Figures 5.3-5.5**). There is greater variation in the sizes of strain accessory genomes in *A. fumigatus*, however this may be an artefact of taxon sampling (**Figure S5.6)**. *Saccharomyces cerevisiae* S288C itself had 31 singleton gene models not found in any other *S. cerevisiae* strain. By comparison, the 100GS authors located 108 genes present in ≥1 strains but not in S288C and 28 genes unique to S288C (Strope *et al.*, 2015). In total, these analyses suggest that fungal pan-genomes evolve by innovations originating within fungi on the strain level, such as gene duplication or rearrangement, as opposed to being influenced by factors such as HGT from prokaryotic sources or larger species-level events.

5.4.3.5 Subterminal regions of fungal genomes may be harbours of accessory genome content

Analysis of the global distribution of core and accessory gene models shows that there is a statistically-significant bias towards accessory gene models in the subterminal regions within three of the four reference genomes in our study and a statistically-significant bias towards core gene models outside these subterminal regions in the same genomes ($p < 0.05$) (**Figures S5.7a, c-d**, **Table S5.4)**. The sole exception is *C. albicans* SC5314, wherein there is a statistically-significant bias for core gene models within subterminal regions ($p < 0.05$) (**Figure S5.7b**, **Table S5.4)**. The subterminal regions of chromosomes are usually areas of genomic instability in eukaryotes, so it is unsurprising that we observe greater breakdown of synteny in these regions (Fedorova *et al.*, 2008). Terminal and subterminal regions of chromosomes (i.e. telomeres and subtelomeric regions) are also known hotspots of recombination in fungi, which can lead to the evolution of novel genetic content, and in some fungi such recombinatory hotspots are potentially enriched for secreted proteins (Croll *et al.*, 2015). All fungal reference genomes possess at least one chromosome that is enriched for accessory gene models; these chromosomes may have undergone recombination or translocation events that lead to the breakdown of synteny or the eventual evolution of novel genes (**Table S5.4)**. Such

translocation events are known to have occurred within some strains of *S. cerevisiae* and *A. fumigatus* in particular (Colson, Delneri and Oliver, 2004; Fraser *et al.*, 2005; Fedorova *et al.*, 2008; Schmidt *et al.*, 2010). In some reference genomes such as *A. fumigatus* Af293 large clusters of accessory genome content can be observed outside the subterminal regions, which may reflect instances of strain- or lineage-specific genomic rearrangement events (**Figure S5.7)**. Such rearrangements are linked to environmental adaptation and reproductive isolation in *S. cerevisiae* genomes (Hou *et al.*, 2014). In *C. neoformans* var. *grubii*, the greater degree of accessory genome content found outside subterminal regions may be a reflection of the role that genomic rearrangement plays in shaping the genomes of individual strains within the host (Fraser *et al.*, 2005).

5.4.3.6 Fungal core and accessory genomes encompass various biological pathways and phenotypes

Due to its position as arguably the most complete fungal model organism, there is a wealth of manually-annotated functional data available for *Saccharomyces cerevisiae* that is lacking for other species. One such collection is the systematic mutation set available from the SGD, which includes amongst other datasets a systematically-constructed genome-wide set of deletion phenotypes for many different strains of *S. cerevisiae* (Engel and Cherry, 2013; Giaever and Nislow, 2014). Using reciprocal BLASTp searches against the reference protein set as well as data from the systematic mutation set, we inferred the knockout viability of the core and accessory genomes of *S. cerevisiae* S288C. We found that the core *S. cerevisiae* S288C genome is not significantly over-represented for either knockout-viable or knockout-inviable genes (**Table S5.5)**. This may reflect the fact less than 20% of the genes encoded in the *S. cerevisiae* S288C genome are thought to be essential for growth and thus likely knockout-inviable (Giaever *et al.*, 2002). It is worth observing however that 962 of the 1,031 predicted gene models with an inviable knockout phenotype are within the core *S. cerevisiae* genome (**Table S5.5)**. In contrast, there is a significant proportion of gene models within the *S. cerevisiae* S288C accessory genome that are associated with a viable knockout phenotype ($p < 0.05$), which appears to reinforce the more variable nature of species accessory genomes relative to core genomes (**Table S5.5)**.

Unlike filamentous fungi such as *Aspergillus* species, many yeasts lack biosynthetic gene clusters (BGCs). Somewhat analogous to BGCs in *S. cerevisiae* are

small “dispensable pathway” (DP) gene clusters of functionally-related genes, which have been lost in other *Saccharomyces* and related species but were later regained in *S. cerevisiae via* HGT or neofunctionalization (Hall and Dietrich, 2007). Hall and Dietrich (2007) previously described 14 such clusters, encompassing 38 reference and another three non-reference genes, which are involved in many different metabolic processes (Hall and Dietrich, 2007). Our analysis of the distribution of 38 reference DP genes within the *S. cerevisiae* pan-genome found one DP cluster which appears to be completely conserved in the pan-genome; a cluster on chromosome II containing three *GAL* genes which mediates the degradation of galactose to galactose-1-phosphate within the glycolysis pathway (Slot and Rokas, 2010) (**Table S5.5)**. Other clusters were highly conserved across almost all strains but not universally-conserved in our dataset, i.e. a small number of strains. Such highly-conserved clusters include two clusters involved in the metabolism of B vitamins; a three-gene *BIO* biotin uptake cluster on chromosome XIV and a *SNO1-SNZ1* vitamin B6 metabolism cluster on chromosome XIII (**Table S5.5)** (Hall and Dietrich, 2007). Another highly-conserved six-gene *DAL-DCG* cluster found on chromosome IX, the largest DP cluster, allows *S. cerevisiae* to use allantonin as its sole nitrogen source through a pathway in which allantonin is converted to urea which is then converted into ammonium by *DUR1-2* (Naseeb and Delneri, 2012). A *SAM4-SAM3* cluster that enables the usage of S-adenosylmethionine as a sulphur source which has one of the two member genes missing in four strains (and is entirely absent in YJM969) (**Table S5.5)**. It is possible that some strains may simply be missing a syntenic ortholog of one or more genes in a cluster due to pseudogenization or synteny loss due to chromosomal rearrangement.

Other DP clusters have more patchy distribution within the *S. cerevisiae* species pan-genome, particularly those within subterminal regions in *S. cerevisiae* S288C, which may indicate a greater breakdown of synteny or gene loss within these clusters. For some clusters this may be due to functional redundancy; for example three DP clusters are involved in vitamin B1 and B6 metabolism, the aforementioned *SNO1-SNZ1* cluster is conserved across almost all 100 strains whereas the other two clusters have patchier distribution or are totally missing in some strains (e.g. in the Indonesian strain YJM1244, two clusters are completely-conserved but the other is absent) (**Table S5.5)**. Other potential causes for this varying distribution of DP clusters may include environmental adaptations. One DP cluster which confers arsenic resistance is prevalent in many wine/European strains, but has much patchier conservation in non-European strains or

strains with Malaysian or West African ancestry (such as SK1). One member gene of this cluster, *ARR3*, is absent in 49 out of the 100 strains in our dataset including many mosaic strains with wine/European and Malaysian ancestry. Increased arsenic resistance has been observed in strains of European ancestry, likely as a result of anthropogenic influence on soil composition, which may explain the *ARR* cluster's absence in some non-European strains (Warringer *et al.*, 2011; Bergström *et al.*, 2014). Additionally, the *ARR* cluster is located in the subterminal regions of chromosome XVI in *S. cerevisiae* S228C; this suggests gene loss or chromosomal rearrangements amongst other events may be responsible for the absence of *ARR3* in the *ARR* cluster of many strains (Maciaszczyk *et al.*, 2004; Bergström *et al.*, 2014).

Within the aspergilli and other fungi, functionally-related genes involved in secondary metabolism pathways are often arranged into BGCs within the subterminal regions of chromosomes. These BGCs are involved in a range of infection and survival processes in the aspergilli, and subterminal regions themselves are believed to mediate the infection process of *A. fumigatus* in the human host (Keller, Turner and Bennett, 2005; Fedorova *et al.*, 2008; McDonagh *et al.*, 2008; Andersen *et al.*, 2013). Our analysis of known BGCs in the *A. fumigatus* pan-genome found 14 BGCs that were completely conserved, a number of which are involved in the production of mycotoxins. Other BGCs have one or two syntenic orthologs that are missing in other strains, in these cases the majority of these genes may play more indirect roles in cluster function and therefore be less likely to be conserved within clusters, while some are only partially present or completely absent in some strains but are highly-conserved in others (**Table S5.5)**. An analysis of variation of *A. fumigatus* BGCs using short-read data by Lind et al. (2017) found similar patterns of BGC variation to our gene-level functional analysis (Lind *et al.*, 2017). Lind et al. (2017) observed some trends which explain the variation in BGCs within *A. fumigatus* in both their analysis and ours; for example a fusarielin-like cluster we identified as missing from A1163 and partially present in other strains has gained pseudogenizing mutations in some strains but not others, whereas variation in other accessory BGCs is due to factors such as transposable elements (as is the case in a 27-member PKS cluster) or lineage-specific gene acquisition/loss events (Lind *et al.*, 2017). This suggests that some BGCs are invariably conserved due to the importance of their function (such as gliotoxins), while others may be lost in particular strains due to environmental adaptations or other factors.

5.4.4 Other remarks

Compared to the volume of software designed to construct and characterize bacterial and archaeal pan-genomes, few dedicated pan-genome software exists for eukaryote taxa. Our overall method of analysis, bespoke gene model prediction followed by pan-genome construction using PanOCT as the anchor method, is *ad hoc* but may point towards a sufficiently-optimized syntenic method of pan-genome construction for eukaryotes in the future. On this point, it is worth noting that PanOCT's current implementation has an exponential memory usage curve per genome added, which makes analysis of prokaryotic or eukaryotic datasets of this scale difficult without dedicated high-performance computational facilities (Fouts *et al.*, 2012). The relative lack of GO information for some fungal species (e.g. *Cryptococcus neoformans* var. *grubii*, which currently lacks a dedicated GO-slim dataset) may have affected our functional characterization of fungal pan-genomes. We attempted to ameliorate this lack of data by using other sources of genomic information (e.g. knockout data from SGD for *S. cerevisiae*), though their efficacy is ultimately dependent on human annotation. One caveat of large-scale pan-genome analysis of this kind may be the usage of genomes assembled *via* a reference-based approach as opposed to *de novo* approaches, which may then lead to an underestimation of accessory genome sizes within species pan-genomes due to underestimation of sequence diversity or inheritance of assembly artefacts from the reference genome (Ekblom and Wolf, 2014). The majority of genomes used for each species dataset were assembled using *de novo* approaches, for example the 100GS dataset is predominantly *de novo* sequenced strains, so the potential effects of overreliance on reference-based assembly data may have been reduced in our study (Strope *et al.*, 2015).

The size of a species pan-genome and its complements are ultimately dependent on the amount and the geographical or phenotypical variety of genomic data sampled. Methodological differences notwithstanding, our 100-strain analysis of the *S. cerevisiae* pan-genome and the 1,011-strain analysis by Peter *et al.* (2018) predict similar-sized pan-genomes (Peter *et al.*, 2018). In contrast, our construction of the *C. albicans* pan-genome likely underestimates the true size of the *C. albicans* accessory genome due to a lack of non-clinical genomic data. The greater variation of accessory genome sizes between individual strains of *C. neoformans* var. *grubii* and *A. fumigatus* may be an artefact of there being fewer strain genomes available for both species, which would in turn affect the sizes of those species' pan-genomes. There have been attempts to estimate the "true"

size of bacterial pan-genomes from existing data using different mathematical models, which vary from inferring almost infinite pan-genomes which increase in size with each strain added to stricter models which infer a more finite structure for most bacterial species (Tettelin *et al.*, 2005; Hogg *et al.*, 2007; Snipen, Almøy and Ussery, 2009). Future analysis of fungal species pan-genomes should attempt to quantify their true size of using similar methods.

5.5 Conclusions

Evidence for the existence of pan-genomic structure has been demonstrated in eukaryotic taxa using a variety of methodologies. Using computational methods based on sequence similarity and conserved synteny between strains, we have constructed and characterized species pan-genomes for four model fungi; *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. Defining "core" genomes as containing gene models syntenically-conserved throughout species and "accessory" genomes as containing gene models of varying syntenic conversation and distribution throughout species, we find strong evidence for pan-genomic structure within fungi. Between 80-90% of all potential gene models in fungal species are core gene models, with the remainder being accessory gene models that are strain-specific or specific to individual groups of strains. Fungal core genomes are enriched for genes of ancient origin and facilitate many essential metabolic, regulatory and survival processes in both commensal and pathogenic species. Fungal accessory genomes are enriched for genes of more recent origin, appear to evolve and vary in size by processes like gene duplication and gain/loss events within strains, and are enriched for genes involved in molecule transport and carbohydrate metabolism amongst other functions. Our analysis supports the growing amount of evidence for pan-genomic structure in eukaryotes.

Chapter 6 – Pangloss: a tool for pan-genome analysis of microbial eukaryotes

This chapter was published in *Genes* in July 2019.

McCarthy CGP & Fitzpatrick DA (2019). Pangloss: a tool for pan-genome analysis of microbial eukaryotes. *Genes*, 10(7), 521.

Chapter outline

Although the pan-genome concept originated in prokaryote genomics, an increasing number of eukaryote species pan-genomes have also been analyzed. However, there is a relative lack of software intended for eukaryote pan-genome analysis compared to that available for prokaryotes. In a previous study, we analyzed the pan-genomes of four model fungi with a computational pipeline which constructed pan-genomes using the synteny-dependent PanOCT approach. Here, we present a modified and improved version of that pipeline which we have called Pangloss. Pangloss can perform gene prediction for a set of genomes from a given species that the user provides, constructs and optionally refines a species pan-genome from that set using PanOCT and can perform various functional characterization and visualization analyses of species pan-genome data. To demonstrate Pangloss's capabilities, we constructed and analyzed a species pan-genome for the oleaginous yeast *Yarrowia lipolytica* and also reconstructed a previously-published species pan-genome for the opportunistic respiratory pathogen *Aspergillus fumigatus*. Pangloss is available from http://github.com/chmccarthy/Pangloss.

6.1 Introduction

Species pan-genomes have been extensively studied in prokaryotes, where pan-genome evolution is primarily driven by rampant horizontal gene transfer (HGT) (Medini *et al.*, 2005; Tettelin *et al.*, 2005; Rouli *et al.*, 2015; Vernikos *et al.*, 2015). Pan-genome evolution in prokaryotes can also vary substantially as a result of lifestyle and environmental factors; opportunistic pathogens such as *Pseudomonas aeruginosa* have large "open" pan-genomes with large proportions of accessory genes, whereas obligate intracellular parasites such as *Chlamydia* species have smaller "closed" pan-genomes with larger proportions of conserved core genes and a smaller pool of novel genetic content (Lefebure *et al.*, 2010; Rouli *et al.*, 2015; Mosquera-Rendón *et al.*, 2016; Sigalova *et al.*, 2018). Studies of pan-genome evolution within eukaryotes has not been as extensive as that of prokaryotes to date, as eukaryote genomes are generally more difficult to sequence and assemble in large numbers relative to prokaryote genomes. However, consistent evidence for pan-genomic structure within eukaryotes has been demonstrated in plant, fungal and planktonic species (Read *et al.*, 2013; Golicz *et al.*, 2016; Peter *et al.*, 2018; Plissonneau, Hartmann and Croll, 2018; McCarthy and Fitzpatrick, 2019a). Unlike prokaryote pan-genomes, eukaryote pan-genomes evolve *via* a variety of processes besides HGT – these include variations in ploidy and heterozygosity within plants (Golicz *et al.*, 2016), and cases of introgression, gene duplication and repeat-induced point mutation in some fungi (Peter *et al.*, 2018; Plissonneau, Hartmann and Croll, 2018; McCarthy and Fitzpatrick, 2019a).

The majority of software and pipelines available for pan-genome analysis are explicitly or implicitly intended for prokaryote datasets. For example, the commonly-cited pipeline Roary is intended for use with genomic location data generated by the prokaryote genome annotation software Prokka (Seemann, 2014; Page *et al.*, 2015). A number of other methodologies such as seq-seq-pan or SplitMEM use genome alignment or de Bruijn graph-based approaches for pan-genome construction which are usually computationally impracticable for eukaryote analysis (Marcus, Lee and Schatz, 2014; Jandrasits *et al.*, 2018). Other common pan-genome approaches, such as LS-BSR or the MCL/MultiParanoid-dependent PGAP, may have potential application in eukaryote pan-genome analysis but as of writing no such application has occurred (Enright, Van Dongen and Ouzounis, 2002; Alexeyenko *et al.*, 2006; Zhao *et al.*, 2012; Sahl *et al.*, 2014). Of the eukaryote pan-genome analyses in the literature, some construct pan-genomes by

mapping and aligning sequence reads using pipelines such as EUPAN (Read *et al.*, 2013; Golicz *et al.*, 2016; Hu *et al.*, 2017), or have constructed and characterized eukaryote pan-genomes using bespoke BLAST-dependent or clustering algorithm-dependent sequence clustering approaches (Read *et al.*, 2013; Peter *et al.*, 2018; Plissonneau, Hartmann and Croll, 2018). In a previous article, we constructed and analyzed the species pan-genomes of four model fungi including *Saccharomyces cerevisiae* using the synteny-based PanOCT method in addition to our own prediction and analysis pipelines (Fouts *et al.*, 2012; McCarthy and Fitzpatrick, 2019a). PanOCT was initially developed for prokaryote pan-genome analysis, and constructs a pan-genome from a given dataset by clustering homologous sequences from different input genomes together into clusters of syntenic orthologs based on a measurement of local syntenic conservation between these sequences, referred to as a conserved gene neighbourhood (CGN) score, and BLAST score ratio (BSR) assessment of sequence similarity (Rasko, Myers and Ravel, 2005; Fouts *et al.*, 2012). Crucially, this synteny-based approach allows PanOCT to distinguish between paralogous sequences within the same genome when assessing orthologous sequences between genomes (McCarthy and Fitzpatrick, 2019a).

Here, we present a refined and improved version of our PanOCT-based pan-genome analysis pipeline which we have called Pangloss. Pangloss incorporates reference-based and *ab initio* gene model prediction methods, and synteny-based pan-genome construction using PanOCT with an optional refinement based on reciprocal sequence similarity between clusters of syntenic orthologs. Pangloss can also perform a number of downstream characterization analyses of eukaryote pan-genomes, including GO-slim term enrichment in core and accessory genomes, selection analyses in core and accessory genomes and visualization of pan-genomic data. To demonstrate the pipeline's capabilities we have constructed and analysed a species pan-genome for the oleaginous yeast *Yarrowia lipolytica* using Pangloss (Dujon *et al.*, 2004). *Y. lipolytica* is one of the earliest-diverging yeasts and has seen various applications as a non-conventional yeast model for protein secretion, regulation of dimorphism and lipid accumulation, and is a potential alternative source for biofuels and other oleochemicals (Nicaud, 2012; Friedlander *et al.*, 2016; Shen *et al.*, 2016; Zeng *et al.*, 2016; Adrio, 2017; Qiao *et al.*, 2017; O'Brien *et al.*, 2018). We have also reconstructed the species pan-genome of the opportunistic respiratory pathogen *Aspergillus fumigatus* from a previous study as a control (McCarthy and Fitzpatrick, 2019a). Pangloss is implemented primarily in Python

and R, and is freely available under an open source GPLv3 licence from http://github.com/chmccarthy/Pangloss.

6.2 Materials and Methods

6.2.1 Implementation

Pangloss is predominantly written in Python with some R and Perl components, and is compatible with macOS and Linux operating systems. Pangloss performs a series of gene prediction, gene annotation and functional analyses to characterize the pan-genomes of microbial eukaryotes. These analyses can be enabled by the user by invoking their corresponding flags on the command line, and many of the parameters of these analyses are controlled by Pangloss using a INI-like configuration file. The various dependencies for eukaryote pan-genome analysis using Pangloss are given in **Table 6.1** and the workflow of the pipeline is given in **Figure 6.1,** both are described in greater detail below (Robert C. Edgar, 2004; Slater and Birney, 2005; Yang, 2007; Ter-Hovhannisyan *et al.*, 2008; Camacho *et al.*, 2009; Cock *et al.*, 2009; Wickham, 2011; Haas *et al.*, 2013; Jones *et al.*, 2014; Obenchain *et al.*, 2015; Simão *et al.*, 2015; Conway, Lex and Gehlenborg, 2017; Gel and Serra, 2017; Klopfenstein *et al.*, 2018). Further installation instructions for all dependencies of Pangloss are available from http://github.com/chmccarthy/Pangloss/.

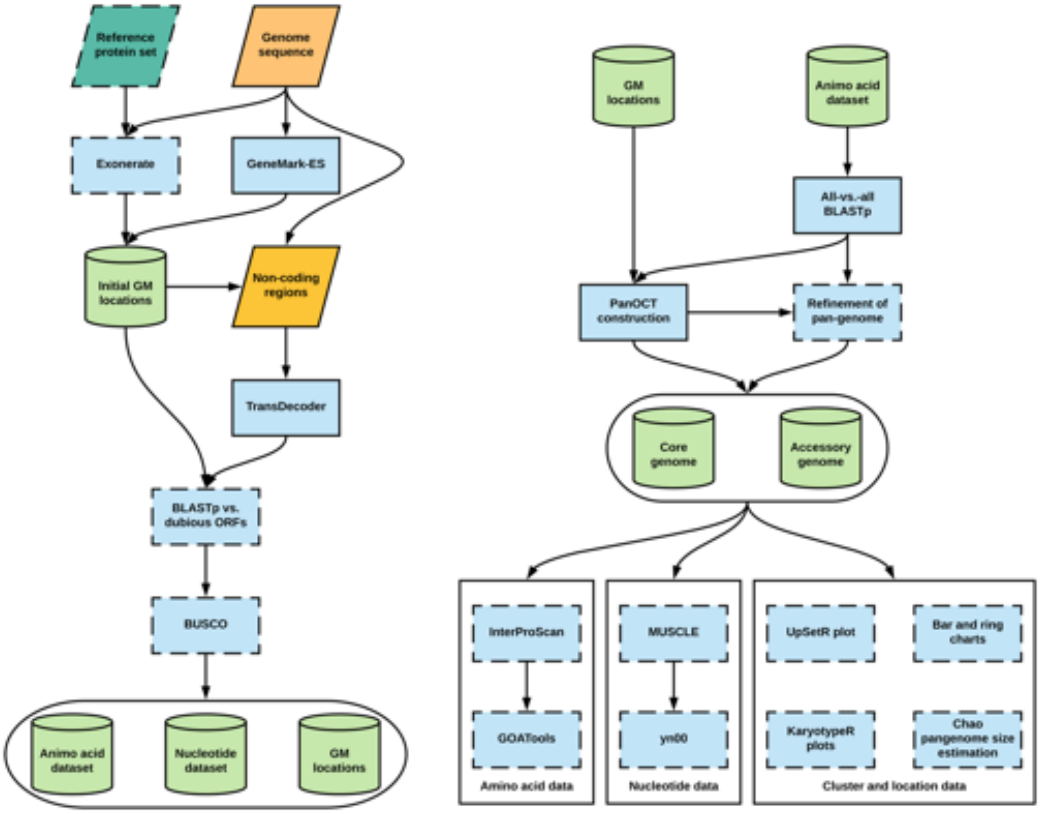


Figure 6.1. Workflow of Pangloss. Optional analyses represented with dotted borders.

Table 6.1. List of various dependencies for Pangloss. See http://github.com/chmccarthy/Pangloss/ for installation instructions for each dependency.

Dependencies	Function
Python*, BioPython (Cock *et al.*, 2009)	Base environment for Pangloss.
Exonerate (Slater and Birney, 2005), GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008), TransDecoder (Haas *et al.*, 2013)	Gene model prediction.
BLAST+ (Camacho *et al.*, 2009)	All-vs.-all sequence similarity search, dubious gene similarity search.
BUSCO (Simão *et al.*, 2015)	Gene model set completedness analysis.
MUSCLE (Robert C. Edgar, 2004), PAML (Yang, 2007)	Selection analysis of core/accessory cluster alignment using yn00.
InterProScan† (Jones *et al.*, 2014), GOATools (Klopfenstein *et al.*, 2018)	Functional classification and functional enrichment analysis of pan-genome.
R, ggplot (Wickham, 2011), ggrepel, UpSetR (Conway, Lex and Gehlenborg, 2017), Bioconductor (Obenchain *et al.*, 2015), KaryoploteR (Gel and Serra, 2017)	Visualization of pan-genome size and distributions across genomes.

*Required for all analyses. †InterProScan is only available for Linux distributions.

6.2.1.1 Gene model prediction and annotation

By default, Pangloss performs its own gene model prediction to generate nucleotide and protein sequence data for all gene models from each genome in a dataset (**Figure 6.1**). Pangloss also generates a set of PanOCT-compatible gene model location data for each genome. Gene model prediction can be skipped by including the flag --no_pred if such data has already been generated, or the user can solely run gene model prediction with no downstream analysis by including the flag --pred_only. For each genome in a dataset, up to three methods of prediction are used:

1. All predicted protein sequences from a user-provided reference genome are queried against each genome using Exonerate, with a heurisitic protein2genome search model (Slater and Birney, 2005). Translated gene model top-hits with an alignment score of ≥90% of the maximum possible alignment score as determined by Exonerate are retained as potential gene models. This search step is parallelized through Python's multiprocessing library, and can be optionally disabled by the user by including the flag --no_exonerate.
2. *Ab initio* Hidden Markov Model (HMM)-dependent gene model prediction is performed using GeneMark-ES with self-training enabled (Ter-Hovhannisyan *et al.*, 2008). If the species of interest is fungal, the user can enable a fungal-specific

branch point site prediction model in the configuration file. If the user has also predicted gene models *via* step 1, those gene models whose locations do not overlap with gene model predicted *via* GeneMark-ES are incorporated into the latter dataset.

3. All remaining non-coding regions of the genome are extracted and subjected to position weight matrix (PVM)-dependent gene model prediction using TransDecoder (Haas *et al.*, 2013). Any remaining predicted gene models with a length of ≥200 amino acids are included in the final gene model dataset.

There are a number of optional steps after that the user can take to assess the quality of gene model prediction within a dataset (**Figure 6.1**). The user can filter gene model sets for potential pseudogenes by querying a set of known dubious genes (either user-curated or from an appropriate resource such as the Saccharomyces Genome Database) against each gene model set using BLASTp (enabled *via* the --qc flag) (Altschul *et al.*, 1990; Engel and Cherry, 2013). Any gene models whose top BLASTp hit against a dubious gene has sequence coverage of ≥70% are removed from further analysis. The completeness of each gene model set can also be assessed using BUSCO (enabled *via* the --busco flag), with the appropriate dataset assigned by the user (Simão *et al.*, 2015).

6.2.1.2 BLASTp and PanOCT analysis

By default, all predicted gene models within a dataset are combined and an all-vs.-all BLASTp search is performed within Pangloss with a user-defined e-value cutoff (default = 10^{-4}) (**Figure 6.1**). However, if the user prefers to perform the all-vs.-all BLASTp step on their own HPC environment they can skip the search *via* the --no_blast flag. The BLASTp search data, along with all gene models and gene model location datasets combined, are used as input for PanOCT. For a pan-genome dataset of syntenic ortholog clusters as constructed by Pangloss, clusters that contain an ortholog from all input genomes are classified as "core" clusters (containing "core" gene models) and clusters missing an ortholog from ≥1 input genomes are classified as "accessory" clusters (containing "accessory" gene models) (McCarthy and Fitzpatrick, 2019a). Pangloss also generates nucleotide and amino acid datasets for every core and accessory cluster for further downstream analyses.

6.2.1.3 Refinement of pan-genome construction based on reciprocal sequence similarity

After construction of the initial pan-genome, the user has the option of refining the pan-genome with Pangloss *via* the --refine flag (**Figure 6.1**). This method attempts to refine the PanOCT-derived microsyntenic pan-genome by accounting for microsynteny loss due to genome assembly artefacts or genomic rearrangements. In this method, Pangloss first extracts all accessory clusters from the accessory genome and parses the previously-generated all-vs.-all BLASTp data used for PanOCT. For each accessory cluster *A*, Pangloss extracts the BLASTp data for each ortholog in *A* and generates a list of BLASTp top-hits to each strain genome not represented in *A* with ≥30% sequence identity. If this list matches another accessory cluster *B* in the accessory genome, Pangloss will then check if each ortholog in *B* has a reciprocal strain top-hit to each ortholog in *A*. If *A* and *B* satisfy this criterion they are merged into a new cluster *AB*, and *A* and *B* themselves are subsequently removed from the accessory genome. If this new cluster *AB* has an ortholog from every input strain genome in the dataset it is then reclassified as a core cluster (McCarthy and Fitzpatrick, 2019a).

6.*2.1.4 Functional annotation and characterization of pan-genome components*

There are optional arguments in Pangloss through which the user can characterize pan-genomes once they are constructed (**Figure 6.1**). If InterProScan is installed, the user can select to have the entire pan-genome dataset annotated with Pfam, InterPro and gene ontology (GO) information *via* the --ips flag (Jones *et al.*, 2014). Additionally, if GOAtools is installed the output from InterProScan can be used to perform GO-enrichment analysis of the core and accessory components of the pan-genome *via* the --go flag, using Fischer's exact test (FET) with parent term propagation and false discovery rate correction ($p < 0.05$) using a p-value distribution generated from 500 resampled p-values (Agresti, 2002; Klopfenstein *et al.*, 2018).

6.*2.1.5 Selection analysis of pan-genome using yn00*

The user can perform selection analysis on core and accessory gene model clusters using yn00 from the PAML package of phylogenetic software (enabled *via* the --yn00 flag) (**Figure 6.1**) (Yang, 2007). For each cluster in a pangenome dataset, an amino acid alignment is performed using MUSCLE with the default parameters. A corresponding

nucleotide alignment is then generated by Pangloss by transferring gaps in the amino acid alignment into the nucleotide data for the same cluster. yn00 selection analysis is handled by Biopython's Bio.Phylo.PAML module and is run with the default parameters (universal genetic code, equal weighting of pathways between codons and estimated codon frequencies). From each cluster alignment, Pangloss will report where available the estimated transition/transversion rate ratio of the cluster (κ) and the number of pairwise alignments within the cluster that show evidence of positive selection according to Yang & Nielsen's (2000) method where the d_N/d_S ratio (ω) is ≥ 1, if $\omega \neq \infty$ (Yang and Nielsen, 2000).

6.2.1.6 Visualization of pan-genome data

A number of optional methods of visualizing pan-genome data are incorporated into Pangloss (**Figure 6.1**). A simple ring chart of the proportion of core and accessory gene models in a pangenome dataset is generated in R using the --size flag. The same flag also generates a bar chart for the distribution of syntenic cluster sizes within a pangenome dataset and estimates the true size of the pan-genome using the Chao lower bound method in R, as previously implemented in the prokaryote pan-genome analysis package micropan (Chao, 1984; Snipen and Liland, 2015). The Chao lower bound method estimates the size of a population given a set of occurrence data for that population from singleton and doubleton occurrences (Chao, 1984). In the case of pan-genomic data we can estimate the true number of syntenic clusters within a pan-genome ($\widehat{N}$) given the observed number of clusters (N) from the numbers of 1-member and 2-member clusters in the pan-genome (y_1 and y_2, respectively), as given by the equation $\widehat{N} = N + \frac{y_1^2}{2y_2}$ (Chao, 1984). The Chao lower bound method is a conservative method of estimating true pan-genome size, but it is worth noting that this estimation may be skewed in cases of overabundance of singleton data (e.g. singleton genes arising from highly fragmented genomes) (Snipen and Liland, 2015; Böhning, Kaskasamkul and van der Heijden, 2019). The distribution of syntenic orthologous gene models within the species accessory genome can be visualized using the R package UpSetR *via* the --upset flag (Conway, Lex and Gehlenborg, 2017). This generates an ortholog distribution plot based on the UpSet technique of visualizing intersections of sets and their occurrences within a dataset using matrix representation, allowing for more input sets than similar Venn-based or Euler-based methods (Lex *et al.*, 2014). Finally, karyotype plots of the genomic locations of

core and accessory gene models along each chromosome/contig within a genome, coloured by either pan-genome component or by syntenic cluster size, can be generated for each genome in a dataset using the Bioconductor package KaryoploteR *via* the --karyo flag (Obenchain *et al.*, 2015; Gel and Serra, 2017).

6.2.2 Dataset assembly

6.2.2.1 Yarrowia lipolytica

Nuclear genome assembly data for seven *Yarrowia lipolytica* strains was obtained from GenBank. Each strain genome was selected based on geographic and environmental distribution, information on which is found in **Table S6.1** (Dujon *et al.*, 2004; Liu and Alper, 2014; Magnan *et al.*, 2016; Devillers and Neuvéglise, 2019). Gene model and gene model location prediction was carried out for all *Y. lipolytica* strain genomes using Pangloss (**Figure 6.1**). GeneMark-ES gene model prediction was performed with a fungal branching point model and TransDecoder gene model prediction was performed with an amino acid sequence length cutoff of ≥200aa. All predicted gene model sets were filtered against a set of 936 known pseudogenes or dubious ORFs from *Saccharomyces cerevisiae* and *Candida albicans* obtained from the *Saccharomyces* and *Candida* Genome Database websites respectively, with a BLASTp e-value cutoff of 10^{-4} (Engel and Cherry, 2013; Skrzypek *et al.*, 2017). Gene models with sequence coverage of ≥70% to a pseudogene/dubious ORF were removed from the dataset (**Table S6.1**). BUSCO analysis for each strain gene model set was performed using the Saccharomycetales dataset (**Table S6.1**). In total, 45,533 gene models were predicted across our entire *Y. lipolytica* pan-genome dataset, with an average of 6,504 gene models per strain and BUSCO completeness per gene model set ranging from approximately 83-89% (87.9% average) (**Table S6.1**).

6.2.2.2 Aspergillus fumigatus

Nuclear genome assembly data for 12 *Aspergillus fumigatus* strains was obtained from GenBank. Each strain genome was previously used to construct an initial *A. fumigatus* species pan-genome by McCarthy & Fitzpatrick (2019a), and strains were selected based on geographic and environmental distribution, including both clinical and wild-type strains (McCarthy and Fitzpatrick, 2019a) (**Table S6.1**). Gene model and gene model location prediction was carried out for all *A. fumigatus* genomes using Pangloss

(**Figure 6.1**). GeneMark-ES gene model prediction was performed with a fungal branching point model and TransDecoder gene model prediction was performed with an amino acid sequence length cutoff of ≥200aa. No filtering for pseudogenes or dubious ORFs was performed for the *A. fumigatus* dataset as no such data is available. BUSCO analysis for each strain gene model set was performed using the Eurotiomycetes dataset (**Table S6.1**). In total, 113,414 gene models were predicted across our entire *A. fumigatus* pan-genome dataset, with an average of 9,451 gene models per strain and BUSCO completeness per gene model set ranging from approximately 93-97% (96% average) (**Table S6.1**).

6.2.3 Pangenome analysis

6.*2.3.1 Yarrowia lipolytica*

An all-vs.-all BLASTp search for the entire *Y. lipolytica* dataset was performed within Pangloss with an e-value cutoff of 10^{-4}. PanOCT analysis for the *Y. lipolytica* dataset was performed within Pangloss using the default parameters for PanOCT (CGN window = 5, sequence identity cutoff = ≥35%). Pan-genome refinement was carried out within Pangloss (**Table S6.1**). Pfam, InterPro and Gene Ontology annotation of the dataset was performed using InterProScan with the default parameters (Hunter *et al.*, 2012; Jones *et al.*, 2014; Finn *et al.*, 2015; Carbon *et al.*, 2017). GO-slim enrichment analysis was carried out for both the core and accessory *Y. lipolytica* genomes using GOATools. GO terms were mapped to the general GO-slim term basket and a Fischer's exact test (FET) analysis with parent term propagation and false discovery rate (FDR) correction ($p < 0.05$) with a p-value distribution generated from 500 resampled p-values (Agresti, 2002; Carbon *et al.*, 2017; Klopfenstein *et al.*, 2018). yn00 analysis of the *Y. lipolytica* pan-genome dataset was performed within Pangloss with the default parameters (Yang and Nielsen, 2000; Yang, 2007). All plots were generated within Pangloss using its various R components as detailed above (**Figures 6.2-6.5**).

6.*2.3.2 Aspergillus fumigatus*

An all-vs.-all BLASTp search for the entire *A. fumigatus* dataset was performed within Pangloss with an e-value cutoff of 10^{-4}. PanOCT analysis for the *A. fumigatus* dataset was performed within Pangloss using the default parameters for PanOCT (CGN

window = 5, sequence identity cutoff = ≥35%). Pan-genome refinement was carried out within Pangloss (**Table S6.1**).

6.3 Results

6.3.1 Analysis of the *Yarrowia lipolytica* pan-genome

A *Y. lipolytica* species pan-genome was constructed with Pangloss *via* PanOCT using publicly-available assembly data from seven strains, including the reference CLIB122 strain and a number of other industrially-relevant strains (Dujon *et al.*, 2004; Liu and Alper, 2014; Magnan *et al.*, 2016; Devillers and Neuvéglise, 2019) (**Table 6.S1**). Strain genomes ranged in size from 19.7-21.3Mb, and the majority had been assembled to near-scaffold quality (**Table S6.1**). A total of 45,533 valid *Y. lipolytica* gene models were predicted by Pangloss after filtering for known pseudogenes from model yeasts, for an average of ~6,505 gene models per strain genome (**Table S6.1**). Pangloss constructed a refined species pan-genome for *Y. lipolytica* containing 6,042 core syntenic clusters (42,294 gene models in total) and 972 accessory syntenic clusters (3,239 gene models in total) (**Figure 6.2, Tables 6.2 and S6.1**). This gives a core:accessory proportion split of approximately 92:8 in terms of gene models and 87:13 in terms of unique syntenic clusters (**Figure 6.2, Table S6.1**). These core:accessory proportions were similar to our previous analyses of other yeasts such as *Saccharomyces cerevisiae* (85:15) and *Candida albicans* (91:9) (McCarthy and Fitzpatrick, 2019a). Accessory genome size in individual *Y. lipolytica* strains varied from 303 gene models in IBT446 to 583 gene models in H222 (**Table S6.1**). Using Chao's lower bound method, the size of the *Y. lipolytica* pan-genome was estimated to contain 7,970 syntenic clusters (**Figure 6.3**). 341 syntenic clusters were missing an ortholog in one strain, with 202 clusters missing an ortholog from IBT446 only, and 390 syntenic clusters consisted of a singleton gene model (**Figures 6.3-6.4**). The number of singleton gene models in individual strains varied from 23 gene models in WSH-Z06 and CBA6003 to 121 gene models in H222 (**Figure 6.4**). Karyotype plots were generated for each *Y. lipolytica* strain in our dataset and display varying amounts of accessory gene models distributed across the 6 chromosomes of *Y. lipolytica* (e.g. CLIB122 in **Figures 6.5a-b**). This is similar to our previous observation of accessory genome distribution within the *Candida albicans* pan-genome, which may have arisen due to a lack of non-clinical strain genomes for that species (McCarthy and Fitzpatrick, 2019a). A large accessory region in chromosome D in CLIB122 (NC_006070.1, **Figures 6.5a-b**) appears to be the result of a gapped region in the same chromosome in PO1f, presumably arising from sequencing artefacts (**Figures 6.5a-b**).

Table 6.2. Pan-genomes of *Yarrowia lipolytica* and *Aspergillus fumigatus*. Refer to **Table S6.1** for further information including strain assembly statistics, BUSCO completeness and links to relevant literature.

Species	**Strains**	**Core genome**		**Accessory genome**		**Pan-genome**	
		Gene models	**Clusters**	**Gene models**	**Clusters**	**Gene models**	**Clusters**
Yarrowia lipolytica	7	42,294	6,042	3,239	972	45,533	7,014
Aspergillus fumigatus	12	92,016	7,668	21,398	3,727	113,414	11,395

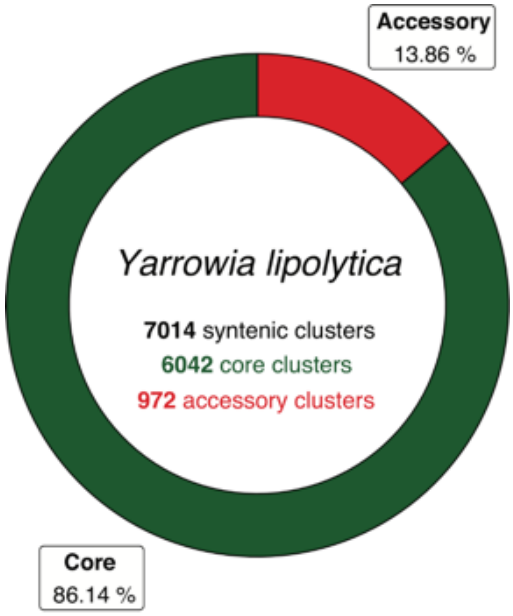


Figure 6.2. Pan-genome of *Yarrowia lipolytica* represented as a ring chart of proportions of core and accessory ortholog clusters within total dataset. Modified from original figure generated by Pangloss. Core proportions coloured in green, accessory proportions coloured in red.

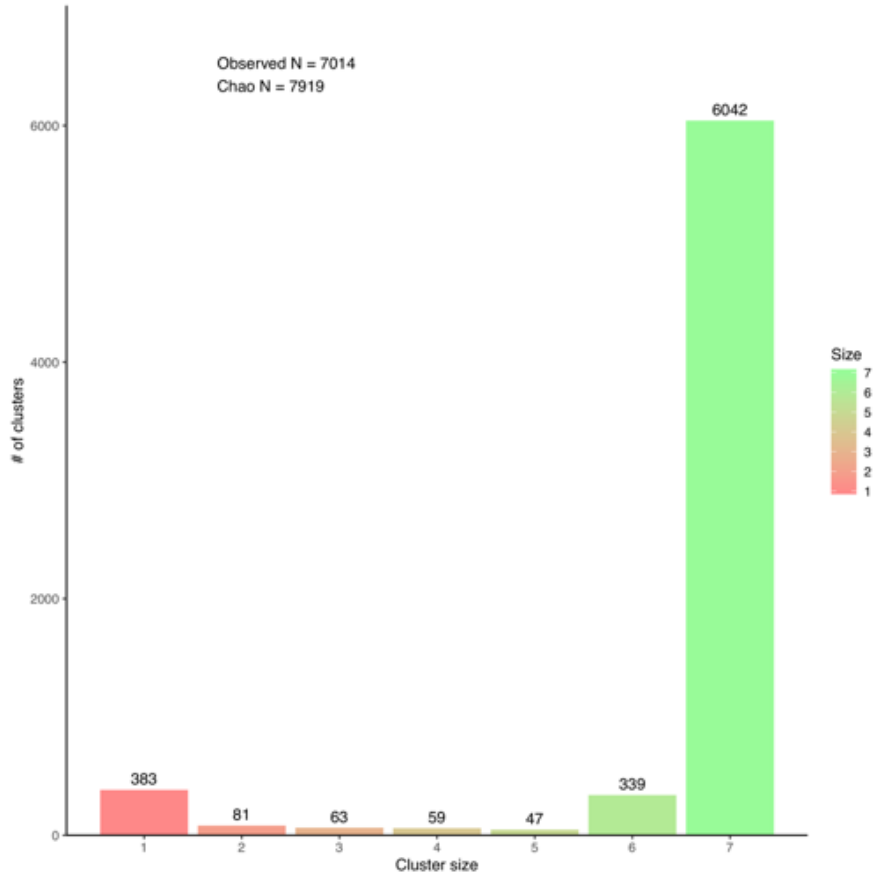


Figure 6.3. Bar chart representing the distribution of syntenic cluster sizes within *Yarrowia lipolytica* pan-genome and Chao's lower bound estimation of true pan-genome size. Figure generated by Pangloss.

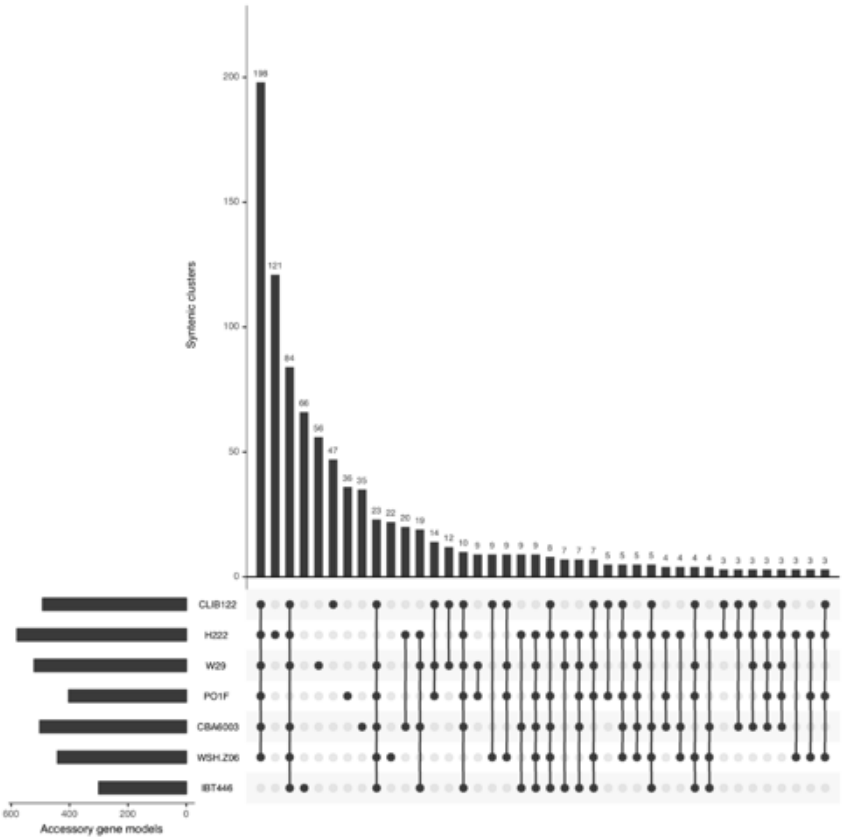


Figure 6.4. UpSet plot of the distribution of syntenic orthologs within the *Yarrowia lipolytica* accessory genome, ranked by syntenic cluster frequency. UpSet plots represent intersections between sets within data as a matrix, and give the number of occurrences of those intersections as a bar chart. In our case, the set intersection matrix represents clusters which contain a syntenic ortholog from 1-6 strains in our dataset and the number of their occurrences is given by the bar chart. Numbers of singleton clusters range from 22 in WSH-Z06 to 121 in H222. Figure generated by Pangloss.

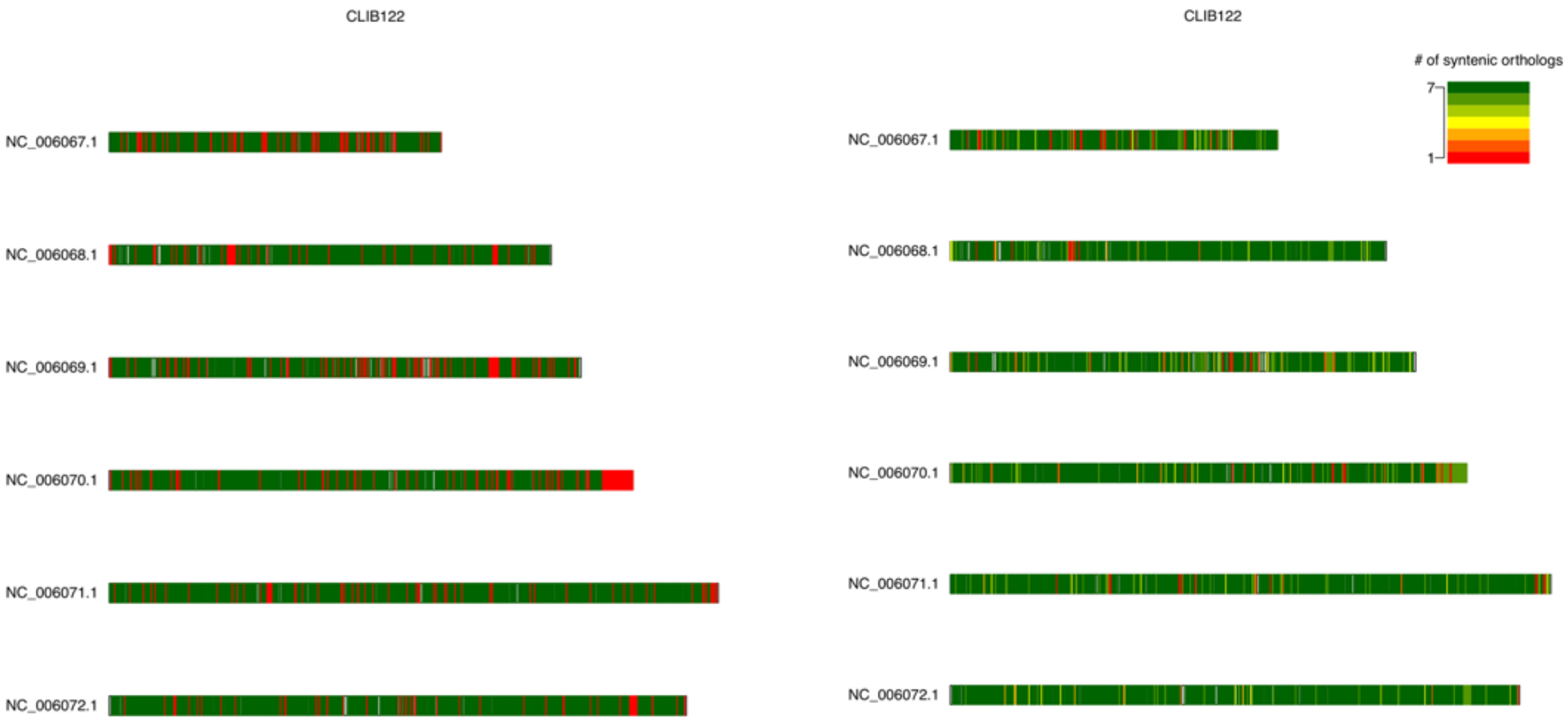
6.3.2 Characterization of the *Yarrowia lipolytica* pan-genome

Selection analysis was performed for all non-singleton clusters in the *Y. lipolytica* core and accessory genome using yn00, which estimates synonymous and non-synonymous rates of substitution within a gene family using pairwise comparisons (Yang, 2007). Of the 6,042 core clusters in the *Y. lipolytica* pan-genome dataset, 453 clusters had at least one pairwise alignment which had $\omega \geq 1$ (7% of all core clusters) whereas for the 582 non-singleton accessory clusters only 52 clusters had at least one pairwise alignment with $\omega \geq 1$ (9% of all non-singleton accessory clusters). It is possible that the low levels of positive selection (i.e. clusters with ≥1 pairwise alignment with $\omega \geq 1$) within the accessory genome reflects the potential lack of evolutionary distance between the strains in our *Y. lipolytica* dataset. The *Y. lipolytica* pangenome dataset was annotated with Pfam, InterPro and Gene Ontology data using InterProScan (Hunter *et al.*, 2012; Jones *et al.*,

2014; Finn *et al.*, 2015; Carbon *et al.*, 2017). Approximately 77% of the total dataset (35,139 gene models) contained at least one Pfam domain. GO-slim enrichment analysis was performed for both core and accessory genomes using GOATools with the default parameters as implemented in Pangloss (**Table S6.2**). Unlike our previous analysis of term enrichment in fungal pan-genomes, transport processes appear to be enriched within the core *Y. lipolytica* genome and processes relating to the production of organic and aromatic compounds are enriched within the accessory *Y. lipolytica* genome (**Table S6.2**) (McCarthy and Fitzpatrick, 2019a). The former may be due to the array of the lipid transport systems that *Y. lipolytica* uses to live in environments rich in hydrophobic substrates (Thevenieau *et al.*, 2009). Similarly, genes whose functions are related to intracellular organelle function are enriched in the *Y. lipolytica* core genome – this may encompass the accumulation of lipids and fatty acids within organelles and lipid body formation within the *Y. lipolytica* cell (**Table S6.2**) (Mlíčková *et al.*, 2004).

6.3.3 Reanalysis of the *Aspergillus fumigatus* pan-genome

As a way of assessing the quality of Pangloss's pan-genome construction we also reconstructed a species pan-genome for *Aspergillus fumigatus*, the opportunistic agent of invasive aspergillosis, using a previously-analyzed dataset containing both clinical and wild-type strains (Nierman *et al.*, 2005; McCarthy and Fitzpatrick, 2019a) (**Tables 6.2 & S6.1**). A total of 113,414 valid *A. fumigatus* gene models were predicted by Pangloss with an average of ~9,451 gene models per strain genome (**Tables 6.2 & S6.1**). Pangloss constructed a refined species pan-genome for *A. fumigatus* containing 7,668 core syntenic clusters (92,016 gene models in total) and 1,783 accessory syntenic clusters (21,398 gene models in total) (**Tables 6.2 & S6.1**). This gives a core:accessory proportion split of approximately 81:19 in terms of gene models and 67:33 in terms of unique syntenic clusters (**Tables 6.2 & S6.1**). These core:accessory proportions are relatively in line with our previous study of the same *A. fumigatus* pan-genome dataset, which found core:accessory proportion splits of 83:17 in terms of gene models and 73:27 in terms of unique syntenic clusters (McCarthy and Fitzpatrick, 2019a). Variation between the two *A. fumigatus* pan-genome analyses is a result of performing gene prediction using Exonerate in our initial analysis in McCarthy & Fitzpatrick (2019a), but not in our subsequent reanalysis (McCarthy and Fitzpatrick, 2019a).



Figures 6.5a and 6.5b. Karyotype plots of core and accessory gene model locations across the six chromosomes of *Yarrowia lipolytica* strain CLIB122. Left: **(a)** Gene model locations coloured by source pan-genome component (core: green, accessory: red). Right: **(b)** Gene model locations coloured by the size of their source syntenic cluster. Non-coding regions coloured in grey. Both figures generated by Pangloss.

6.4 Discussion

As pan-genome analysis of eukaryotes becomes more commonplace, ideally the amount of software to construct and characterize eukaryote pan-genome should begin to match that which is already available for prokaryotes. Our software pipeline Pangloss applies a sequence similarity and synteny-based approach from prokaryote pan-genome analysis, implemented as PanOCT by Fouts et al (2012), to eukaryote pan-genome analysis and allows the user to perform their own gene prediction and downstream characterization and visualization of pan-genome data from one self-contained script (Fouts *et al.*, 2012; McCarthy and Fitzpatrick, 2019a). Although our pipeline has been designed for eukaryote pan-genome analysis, as PanOCT is a prokaryote method in origin Pangloss should also support prokaryote datasets – albeit with some modifications to gene model prediction strategies by the user. Unlike other common gene clustering approaches such as MCL, PanOCT incorporates local synteny *via* assessing the CGN between potential orthologs as a criterion to clustering in addition to sequence similarity (Alexeyenko *et al.*, 2006; Fouts *et al.*, 2012). This makes PanOCT distinct from most clustering approaches in that it can distinguish orthologs from paralogs – i.e. if one assumes that "true" orthologs are more likely to be located in relatively-similar regions of their respective genomes they then should in turn be more likely to cluster together when syntenic conservation is taken into consideration. This is of particular relevance to eukaryote pan-genomes, as gene duplication plays a substantial role in eukaryote gene family and genome evolution (Friedman and Hughes, 2001; McCarthy and Fitzpatrick, 2019a). Although this approach is more stringent than clustering gene families based on approaches like MCL or BLAST searches alone, it is potentially more reflective of evolution on a gene-level basis within strains of the same species.

There are ways in which our approach can be improved upon in future methodologies, both in terms of prediction and analytic strategies. For example, Pangloss has an optional Exonerate-based gene model prediction strategy which searches input genomes for translated homologs of reference sequences (Slater and Birney, 2005). This is an exhaustive approach that may pick up potential gene models missed by GeneMark-ES and/or TransDecoder, but it is also time-inefficient. To search all 6,472 reference protein sequences from *Y. lipolytica* CLIB222 against a single *Y. lipolytica* genome takes on average four hours on three threads on a server running Ubuntu 18.04.2 LTS

(approximately 9 sequences per minute per thread), whereas both GeneMark-ES gene model prediction with fungal point branching and subsequent ORF prediction in non-coding regions with TransDecoder performed on the same genome with the same number of threads typically takes ~30-35 minutes. It is for this reason primarily that we have made the Exonerate-based strategy optional for any gene prediction that is performed by Pangloss. Furthermore, PanOCT's memory usage increases exponentially per strain added, notwithstanding the potentially complex distribution of gene models between strains themselves (Fouts *et al.*, 2012; McCarthy and Fitzpatrick, 2019a). Constructing a species pan-genome using PanOCT from a small and relatively well-conserved dataset such as that for our *Y. lipolytica* or *A. fumigatus* studies should be achievable on most standard hardware. For larger datasets, such as our previous pan-genome analysis of 100 *Saccharomyces cerevisiae* genomes however, it may be preferable to perform such analysis on a high-performance cluster environment or otherwise an alternative synteny-based method of pan-genome construction may be more appropriate (McCarthy and Fitzpatrick, 2019a). Finally, we would encourage users to interrogate and visualize the results of analysis using Pangloss and adjust the input parameters where appropriate for their data. In our case, the parameters which were chosen for use in Pangloss for this analysis (e.g. BLAST e-value cutoff, CGN window size) are largely based on those from our previous analysis of fungal pan-genomes or other studies using PanOCT (Fouts *et al.*, 2012; McCarthy and Fitzpatrick, 2019a). Depending on the size of a pan-genome dataset or the species of interest, different cutoffs may be more suitable – e.g. for species with longer average gene lengths a lower sequence identity cutoff for PanOCT clustering than the default (>35%) may be more appropriate. Many of these parameters can be adjusted in the INI-like configuration file provided with Pangloss.

6.5 Conclusions

Pan-genome analysis of eukaryotes has become more common, but many of the available software for pan-genome analysis are intended for use with prokaryote data. We have developed Pangloss, a pipeline that allows users to generate input data and construct species pan-genomes for microbial eukaryotes using the synteny-dependent PanOCT method and various downstream characterization analyses. To demonstrate the capabilities of our pipeline we constructed a species pan-genome for *Yarrowia lipolytica*, an oleaginous yeast with potential biotechnological applications, and performed various functional and data visualization analyses using Pangloss. The *Y. lipolytica* pangenome is similar in terms of core and accessory genome proportions to previously analyzed fungal pan-genomes but is unique in that biological processes such as transport are statistically-enriched in the core genome. We also used Pangloss to reconstruct a species pan-genome for the respiratory pathogen *Aspergillus fumigatus* using a previously-analyzed dataset and found that Pangloss generated a similar pan-genomic structure for *A. fumigatus* to that of our previous analysis. Building on our previous work on fungal pan-genomes, this study not only provides further evidence for pan-genomic structure within eukaryote species but also presents a methodological pipeline for future eukaryote pan-genome analysis.

Chapter 7 – Future work and perspectives

Chapter outline

In this chapter, I briefly discuss potential future work that may follow for both the oomycetes and fungi arising from genome sequencing data, and compare the current states of oomycete and fungal genomics with what both fields may look like in the near future.

7.1 Oomycete genomics: future perspectives

Oomycete genomics has come a long way since the publication of the genomes of *Phytophthora sojae* and *Phytophthora ramorum* in 2006. At the time of writing (October 2019), there are 61 oomycete species with genome assemblies that are publicly-available from NCBI – an increase of at least 30 from the start of 2015. Many of the species sequenced in recent years have come from outside the two major genera *Phytophthora* and *Pythium* – the "downy mildews" seem to be a particular target for oomycete genome sequencing projects due to their host range of economically-important plant species. In this section, I propose how future efforts in genome sequencing and comparative work may help us better answer some underlying questions of oomycete biology and evolutionary history.

7.1.1 Oomycete evolutionary history: resolving problem taxa

A greater amount of genome sequencing for as-yet unsampled or under-sampled *Phytophthora* and *Pythium* clades may allow researchers to address whether these clades are monophyletic under phylogenomic reconstruction as they have been in smaller multigene phylogenetics. In the case of *Phytophthora*, more genomic data for these clades should yield more accurate phylogenomic studies and help to clarify the relationships between the more derived clades (Clades 1-5). Some clades within the *Phytophthora* genus such as Clade 6 are known to contain species which undergo hybridization with other *Phytophthora* species – this may conflate phylogenetic inference if hybridization has occurred across clades and so selection of species for future phylogenomic studies of the oomycetes should be conscious of this issue. Additionally, the sequencing of more downy mildew genomes should help resolve the particularly relationships between the two groups of downy mildews and the *Phytophthora* genus as a whole – potentially earmarking a reclassification of sort for some members of *Phytophthora* or the downy mildews. For *Pythium* greater genomic data across the genus will allow us to determine whether the genus is truly monophyletic or should be reorganized into five different genera as per previous research has suggested. Broader sampling of other orders outside of Peronosporales, not only other "crown" orders like Saprolegniales and Albuginales but other intermediate and basal orders like Rhipidales, will afford us a greater picture of oomycete diversity outside of plant pathogenic *Phytophthora* and *Pythium* species. With such phylogenomic data, researchers will be able to investigate fundamental and applied

questions of oomycete evolutionary and molecular biology – this can include questions such as why plant pathogenicity has evolved independently multiple times within the oomycetes or the expansion of effector families in *Phytophthora* species relative to other oomycetes.

7.1.2 The molecular evolution and diversity of oomycete species

Oomycetes, unlike filamentous fungi like *Aspergillus* species for example, do not produce arrays of secondary metabolites for host infection. Instead, they produce "effector" proteins which attempt to control host immune response to enable colonization within the host. The hallmark effectors of the oomycetes - RXLR-motif and CRN-motif effectors - have been the subject of extensive genomic and phylogenetic research as more genomics data has become available for the oomycetes. With more data and refined analytic methodologies, we will be able to have a greater understanding of how these molecular features have evolved. Other trends such as the evolution of "pathogenicity islands" within oomycetes species and the evolution of so-called "two-speed genomes" (Dong, Raffaele and Kamoun, 2015) in plant pathogenic *Phytophthora* species may also be investigated in greater detail. As oomycete genomes are significantly more complex than fungal genomes (greater instances of repeat regions, segregation of genomic content into gene-rich and gene-sparse areas), generating a single high-quality reference genome sequence for an oomycete species has previously been a challenge in and of itself. With the advent of new sequencing technologies such as Oxford Nanopore and PacBio SMRT, which allow for longer sequencing reads and can be used in tandem with more established methods such as Illumina, it is now possible to quickly sequence multiple oomycete genomes across different species or within different species. This will enable analysis of variation within species, such as pangenome approaches or GWAS approaches. The expected increase of oomycete genomic data coming out of initiatives such as the *Phytophthora* Sequencing Consortium will help facilitate such research also.

7.2 Fungal genomics: future perspectives

Fungal genome sequence data has increased dramatically over the last ten years, and with more sophisticated sequencing technology that number will only increase further. In this section, I briefly discuss the importance in accurate and representative phylogenomics can be used to elucidate how important traits have evolved within fungi, and how the wealth of genomics data available can be exploited for various biotechnological applications.

7.2.1 Mapping major events in the fungal tree of life

With greater sampling of non-Dikaryan species we now have a greater understanding of the diversity of the fungal kingdom as a whole. There are however, a number of outstanding questions still to be addressed regarding how important traits in certain branches in the fungal kingdom have evolved. These include the multiple independent origins of multicellularity within the fungi (and the seemingly convergent evolution of filamentation in the otherwise unrelated oomycetes), the evolution of various parasitic and saprotrophic lifestyles across all branches of fungi, the true extent of HGT amongst fungi and the impact of gene remodelling events across the fungal kingdom. To accurately place these events however, a robust phylogeny generated from high-quality genomic data must be in place otherwise any inferences of where such traits (and their corresponding gene families) evolved may be conflated. While some branches of the fungal tree of life (e.g. Pezizomycotina) are highly-represented to at least the order level, many of the more early-diverging lineages are quite under-represented due to difficulties in culturing and detection. As genome sequencing and bioinformatics technologies improve, the numbers of early-diverging fungal genomes taken from cultures (or even metagenomics samples) should improve.

7.2.2 Exploiting large-scale fungal genomics data

As discussed in **Chapter 1**, fungi fulfil a broad range of roles not only in the environment but in human activity as well. More genome sequence data will allow greater predictive research into potential applications of fungi within clinical and biotechnological contexts, while also facilitating more proteomics and genetic engineering research into exploitable compounds and systems in fungi. With the

increasing prevalence of resistance to common antimicrobial compounds, it is of utmost importance that new sources of antimicrobials can identified and soil fungi – who are naturally in competition with many other pathogenic microbes – could have potential application in this area. Some recent analysis on this front has proved promising; e.g. novel antimicrobial compounds such as yanuthones have been identified in a number of *Aspergillus* and *Penicillium* species using a variety of genomics and spectroscopy approaches (Holm *et al.*, 2014; Banani *et al.*, 2016; Nielsen *et al.*, 2017). Similarly, genome sequence data will be useful in determining the suitability and application of potential biocontrol agents within the fungi - such as natural pesticides and mycorhizzal parasites of plant pathogenic bacteria and eukaryotes (Grigoriev, Cullen, *et al.*, 2011). A growing area of fungal research is the production of hydrocarbons and long-chain fatty acids using oleaginous fungi such as *Yarrowia lipolytica*, and greater genomic data for these species will aid the engeering of more sophisticated models for heterologous expression of biofuels and other important compounds (Shi *et al.*, 2018). Genomics data can also be used to guide gene editing and hybridization approaches for fungi used in food biotechnology, such as reducing the production of astrigent byproducts in brewing yeasts and optimizing production of endogenous and recombinant molecules in *Aspergillus niger* (Leynaud-Kieffer *et al.*, 2019; Mertens *et al.*, 2019).

7.3 The future of microbial eukaryote genomics

Genome sequencing has been such a fundamental paradigm shift in biology that its presence and the resultant genetic information it generates is often taken for granted by researchers, particularly those who study model organisms like yeast or *Drosophila melanogaster*. However, there are many branches of the eukaryotic tree of life – often less charismatic branches – which are still poorly represented in terms of available genomics data. Without such data answering fundamental and applied questions of eukaryote evolution, e.g. how certain taxa evolved multicellularity or how host pathogenicity range evolves within a genus, remains a challenge (Richards, 2015). Comparing the two groups of microbial eukaryotes I have studied in this thesis, the oomycetes and fungi, there is a great disparity in the volume of genomic data and genomic analyses performed for both groups. The oomycetes are still something of a niche area in terms of genome sequencing, partly due to their genomic complexity and partly due to not being an established field relative to fungal or animal genome sequencing and comparative genomics. However, their importance to food security and the environment cannot be overstated and so it is critical that researchers have a thorough understanding of their molecular and genomic evolution to ameliorate their effects amidst a booming world population and the advancing climate crisis. For the fungi, increased genomic data will facilitate greater molecular and biochemical research into fighting antimicrobial resistance as well as being able to treat neglected tropical diseases and environmental pathogens. Cutting-edge research in these areas will be of the utmost importance in confronting the challenges ahead that the planet faces.

Bibliography

Adams, M. D. *et al.* (2000) 'The genome sequence of Drosophila melanogaster.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 287(5461), pp. 2185–95. doi: 10.1126/SCIENCE.287.5461.2185.
Adhikari, B. N. *et al.* (2013) 'Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes', *PLoS ONE*. Public Library of Science, 8(10), p. e75072. doi: 10.1371/journal.pone.0075072.
Adrio, J. L. (2017) 'Oleaginous yeasts: Promising platforms for the production of oleochemicals and biofuels', *Biotechnology and Bioengineering*, 114(9), pp. 1915–1920. doi: 10.1002/bit.26337.
Agresti, A. (2002) *Categorical Data Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (Wiley Series in Probability and Statistics). doi: 10.1002/0471249688.
Ahmed, A. O. *et al.* (2004) 'Mycetoma caused by Madurella mycetomatis: a neglected infectious burden', *The Lancet Infectious Diseases*. Elsevier, 4(9), pp. 566–574. doi: 10.1016/S1473-3099(04)01131-4.
Akanni, W. A. *et al.* (2014) 'L.U.St: a tool for approximated maximum likelihood supertree reconstruction', *BMC Bioinformatics*, 15(1), p. 183. doi: 10.1186/1471-2105-15-183.
Akanni, W. A. *et al.* (2015) 'Implementing and testing Bayesian and maximum-likelihood supertree methods in phylogenetics.', *Royal Society open science*, 2(8), p. 140436. doi: 10.1098/rsos.140436.
Alexeyenko, A. *et al.* (2006) 'Automatic clustering of orthologs and inparalogs shared by multiple proteomes', *Bioinformatics*. Narnia, 22(14), pp. e9–e15. doi: 10.1093/bioinformatics/btl213.
Alföldi, J. and Lindblad-Toh, K. (2013) 'Comparative genomics as a tool to understand evolution and disease', *Genome Research*. Cold Spring Harbor Laboratory Press, 23(7), pp. 1063–1068. doi: 10.1101/GR.157503.113.
Ali, S. S. *et al.* (2017) 'Phytophthora megakarya and P. palmivora, Causal Agents of Black Pod Rot, Induce Similar Plant Defense Responses Late during Infection of Susceptible Cacao Pods', *Frontiers in Plant Science*. Frontiers, 8, p. 169. doi: 10.3389/fpls.2017.00169.
Alm, R. A. *et al.* (1999) 'Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori', *Nature*. Nature Publishing Group, 397(6715), pp. 176–180. doi: 10.1038/16495.
Alsmark, C. *et al.* (2013) 'Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes.', *Genome biology*. BioMed Central, 14(2), p. R19. doi: 10.1186/gb-2013-14-2-r19.
Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.
Altschul, S. F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.', *Nucleic acids research*. Oxford University Press, 25(17), pp. 3389–402. doi: 10.1093/nar/25.17.3389.
Andersen, M. R. *et al.* (2013) 'Accurate prediction of secondary metabolite gene clusters in filamentous fungi', *Proceedings of the National Academy of Sciences*,

110(1), pp. E99–E107. doi: 10.1073/pnas.1205532110.
Anderson, S. *et al.* (1981) 'Sequence and organization of the human mitochondrial genome', *Nature*. Nature Publishing Group, 290(5806), pp. 457–465. doi: 10.1038/290457a0.
André Lévesque, C. (2011) 'Fifty years of oomycetes-from consolidation to evolutionary and genomic exploration', *Fungal Diversity*. Springer Netherlands, 50(1), pp. 35–46. doi: 10.1007/s13225-011-0128-7.
Annaluru, N. *et al.* (2014) 'Total synthesis of a functional designer eukaryotic chromosome', *Science*, 344(6179), pp. 55–58. doi: 10.1126/science.1249252.
Armbrust, E. V. *et al.* (2004) 'The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 306(5693), pp. 79–86. doi: 10.1126/science.1101156.
Arora, R. K., Sharma, S. and Singh, B. P. (2014) 'Late blight disease of potato and its management', *Potato Journal*, 41(1), pp. 16–40.
Ascunce, M. S. *et al.* (2017) 'Phylogenomic analysis supports multiple instances of polyphyly in the oomycete peronosporalean lineage', *Molecular Phylogenetics and Evolution*. Academic Press, 114, pp. 199–211. doi: 10.1016/j.ympev.2017.06.013.
Attwood, T. K. *et al.* (2012) 'The PRINTS database: A fine-grained protein sequence annotation and analysis resource-its status in 2012', *Database*, 2012, p. bas019. doi: 10.1093/database/bas019.
Aury, J.-M. *et al.* (2006) 'Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia.', *Nature*. Nature Publishing Group, 444(7116), pp. 171–178. doi: 10.1038/nature05230.
Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.
Bachvaroff, T. R., Sanchez Puerta, M. V. and Delwiche, C. F. (2005) 'Chlorophyll c-containing plastid relationships based on analyses of a multigene data set with all four chromalveolate lineages', *Molecular Biology and Evolution*, 22(9), pp. 1772–1782. doi: 10.1093/molbev/msi172.
Baldauf, S. L. *et al.* (2000) 'A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data', *Science*. American Association for the Advancement of Science, 290(5493), pp. 972–977. doi: 10.1126/science.290.5493.972.
Baldauf, S. L. and Palmer, J. D. (1993) 'Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins.', *Proceedings of the National Academy of Sciences of the United States of America*, 90(24), pp. 11558–62. doi: 10.1073/pnas.90.24.11558.
Banani, H. *et al.* (2016) 'Genome sequencing and secondary metabolism of the postharvest pathogen Penicillium griseofulvum', *BMC Genomics*. BioMed Central, 17(1), p. 19. doi: 10.1186/s12864-015-2347-x.
Barabote, R. D. *et al.* (2011) 'Xenobiotic efflux in bacteria and fungi: a genomics update.', *Advances in enzymology and related areas of molecular biology*, 77, pp. 237–306. doi: 10.1007/s12671-013-0269-8.Moving.
De Barros Lopes, M. *et al.* (2002) 'Evidence for multiple interspecific hybridization in

Saccharomyces sensu stricto species', *FEMS Yeast Research*. Oxford University Press, 1(4), pp. 323–331. doi: 10.1016/S1567-1356(01)00051-4.
Baum, B. R. (1992) 'Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees', *Taxon*, 41(1), pp. 3–10. doi: 10.2307/1222480.
Baxter, L. *et al.* (2010) 'Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 330(6010), pp. 1549–51. doi: 10.1126/science.1195203.
Beakes, D. *et al.* (2014) 'Systematics of the Straminipila : Labyrinthulomycota , Hyphochytriomycota , and Oomycota', in *The Mycota*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 39–97. doi: 10.1007/978-3-642-55318-9_3.
Beakes, G. W., Glockling, S. L. and Sekimoto, S. (2012) 'The evolutionary phylogeny of the oomycete "fungi"', *Protoplasma*. Springer Vienna, 249(1), pp. 3–19. doi: 10.1007/s00709-011-0269-2.
Beck, R. M. D. *et al.* (2006) 'A higher-level MRP supertree of placental mammals.', *BMC evolutionary biology*. BioMed Central, 6(1), p. 93. doi: 10.1186/1471-2148-6-93.
Begerow, D., Stoll, M. and Bauer, R. (2006) 'A phylogenetic hypothesis of Ustilaginomycotina based on multiple gene analyses and morphological data', *Mycologia*. Taylor & Francis, 98(6), pp. 906–916. doi: 10.3852/mycologia.98.6.906.
Belbahri, L. *et al.* (2008) 'Evolution of the cutinase gene family: evidence for lateral gene transfer of a candidate Phytophthora virulence factor.', *Gene*, 408(1–2), pp. 1–8. doi: 10.1016/j.gene.2007.10.019.
Benhamou, N. *et al.* (2012) 'Pythium oligandrum: An example of opportunistic success', *Microbiology (United Kingdom)*. Microbiology Society, 158(11), pp. 2679–2694. doi: 10.1099/mic.0.061457-0.
Benson, D. A. *et al.* (2013) 'GenBank.', *Nucleic acids research*, 41(Database issue), pp. D36-42. doi: 10.1093/nar/gks1195.
Berbee, M. L., James, T. Y. and Strullu-Derrien, C. (2017) 'Early Diverging Fungi: Diversity and Impact at the Dawn of Terrestrial Life', *Annual Review of Microbiology*. BioMed Central, 71(1), pp. 41–60. doi: 10.1146/annurev-micro-030117-020324.
Berbee, M. L. and Taylor, J. W. (1992) 'Detecting Morphological Convergence in True Fungi, Using-18s Ribosomal-Rna Gene Sequence Data', *Biosystems*, 28(1–3), pp. 117–125.
Berbee, M. L. and Taylor, J. W. (2010) 'Dating the molecular clock in fungi - how close are we?', *Fungal Biology Reviews*, pp. 1–16. doi: 10.1016/j.fbr.2010.03.001.
Berg, B. *et al.* (2003) 'Decomposer organisms', in *Plant Litter*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 31–48. doi: 10.1007/978-3-662-05349-2_3.
Bergsten, J. (2005) 'A review of long-branch attraction', *Cladistics*, pp. 163–193. doi: 10.1111/j.1096-0031.2005.00059.x.
Bergström, A. *et al.* (2014) 'A high-definition view of functional genetic variation from natural yeast genomes', *Molecular Biology and Evolution*. Oxford University Press, 31(4), pp. 872–888. doi: 10.1093/molbev/msu037.

Bertier, L. *et al.* (2013) 'Host Adaptation and Speciation through Hybridization and Polyploidy in Phytophthora', *PLoS ONE*. Edited by W. J. Etges. Public Library of Science, 8(12), p. e85385. doi: 10.1371/journal.pone.0085385.
Bibb, M. J. *et al.* (1981) 'Sequence and gene organization of mouse mitochondrial DNA', *Cell*. Cell Press, 26(2 PART 2), pp. 167–180. doi: 10.1016/0092-8674(81)90300-7.
Bininda-Emonds, O. R. P. (2004) 'The evolution of supertrees', *Trends in Ecology and Evolution*, 19(6), pp. 315–322. doi: 10.1016/j.tree.2004.03.015.
Birren, B., Fink, G. and Lander, E. (2002) 'Fungal Genome Initiative: White Paper developed by the Fungal Research Community', *Cambridge, MA:Whitehead Institue Center for Genome Research*.
Blackwell, M. (2011) 'The fungi: 1, 2, 3 ... 5.1 million species?', *American Journal of Botany*, 98(3), pp. 426–438. doi: 10.3732/ajb.1000298.
Blair, J. E. *et al.* (2008) 'A multi-locus phylogeny for Phytophthora utilizing markers derived from complete genome sequences', *Fungal Genetics and Biology*, 45(3), pp. 266–277. doi: 10.1016/j.fgb.2007.10.010.
Blattner, F. R. *et al.* (1997) 'The complete genome sequence of Escherichia coli K-12', *Science*. American Association for the Advancement of Science, 277(5331), pp. 1453–1462. doi: 10.1126/science.277.5331.1453.
Böhning, D., Kaskasamkul, P. and van der Heijden, P. G. M. (2019) 'A modification of Chao's lower bound estimator in the case of one-inflation', *Metrika*. Springer Berlin Heidelberg, 82(3), pp. 361–384. doi: 10.1007/s00184-018-0689-5.
Boissy, R. *et al.* (2011) 'Comparative supragenomic analyses among the pathogens Staphylococcus aureus, Streptococcus pneumoniae, and Haemophilus influenzae Using a modification of the finite supragenome model', *BMC Genomics*, 12(1), p. 187. doi: 10.1186/1471-2164-12-187.
Bourret, T. B. *et al.* (2018) 'Multiple origins of downy mildews and mito-nuclear discordance within the paraphyletic genus Phytophthora', *PLoS ONE*. Edited by M. Gijzen. Public Library of Science, 13(3), p. e0192502. doi: 10.1371/journal.pone.0192502.
Bowler, C. *et al.* (2008) 'The Phaeodactylum genome reveals the evolutionary history of diatom genomes.', *Nature*. Macmillan Publishers Limited. All rights reserved, 456(7219), pp. 239–44. doi: 10.1038/nature07410.
Brasier, C. and Webber, J. (2010) 'Plant pathology: Sudden larch death', *Nature*. Nature Publishing Group, 466(7308), pp. 824–825. doi: 10.1038/466824a.
Brenchley, R. *et al.* (2012) 'Analysis of the bread wheat genome using whole-genome shotgun sequencing', *Nature*. Nature Publishing Group, 491(7426), pp. 705–710. doi: 10.1038/nature11650.
Bryant, D. and Steel, M. (2009) 'Computing the distribution of a tree metric', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3), pp. 420–426. doi: 10.1109/TCBB.2009.32.
Bryson, V. and Vogel, H. J. (1965) 'Evolving genes and proteins', in *Science*, pp. 68–71. doi: 10.1126/science.147.3653.68.

Burgess, T. I. (2015) ‘Molecular Characterization of Natural Hybrids Formed between Five Related Indigenous Clade 6 Phytophthora Species’, *PLoS ONE*. Edited by M. Gijzen. Public Library of Science, 10(8), p. e0134225. doi: 10.1371/journal.pone.0134225.
Burki, F. *et al.* (2007) ‘Phylogenomics Reshuffles the Eukaryotic Supergroups’, *PLoS ONE*. Edited by G. Butler, 2(8), p. e790. doi: 10.1371/journal.pone.0000790.
Burki, F. (2014) ‘The eukaryotic tree of life from a global phylogenomic perspective’, *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harbor Laboratory Press, 6(5), p. a016147. doi: 10.1101/cshperspect.a016147.
Butler, G. *et al.* (2009) ‘Evolution of pathogenicity and sexual reproduction in eight Candida genomes’, *Nature*. Nature Publishing Group, 459(7247), pp. 657–662. doi: 10.1038/nature08064.
Byrne, K. P. and Wolfe, K. H. (2005) ‘The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species’, *Genome Research*. Cold Spring Harbor Laboratory Press, 15(10), pp. 1456–1461. doi: 10.1101/gr.3672305.
Cairns, T. C., Nai, C. and Meyer, V. (2018) ‘How a fungus shapes biotechnology: 100 years of Aspergillus niger research’, *Fungal Biology and Biotechnology*. BioMed Central, 5(1), p. 13. doi: 10.1186/s40694-018-0054-5.
Camacho, C. *et al.* (2009) ‘BLAST+: Architecture and applications’, *BMC Bioinformatics*. BioMed Central, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.
Campbell, A., Mrázek, J. and Karlin, S. (1999) ‘Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 96(16), pp. 9184–9. doi: 10.1073/pnas.96.16.9184.
Capella-Gutiérrez, S., Marcet-Houben, M. and Gabaldón, T. (2012) ‘Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi’, *BMC Biology*. BioMed Central, 10(1), p. 47. doi: 10.1186/1741-7007-10-47.
Carbon, S. *et al.* (2017) ‘Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium’, *Nucleic Acids Research*, 45(D1), pp. D331–D338. doi: 10.1093/nar/gkw1108.
Casari, G. *et al.* (1995) ‘Challenging times for bioinformatics’, *Nature*. Nature Publishing Group, 376(6542), pp. 647–648. doi: 10.1038/376647a0.
Casewell, N. R. *et al.* (2011) ‘Gene tree parsimony of multilocus snake venom protein families reveals species tree conflict as a result of multiple parallel gene loss’, *Molecular Biology and Evolution*. Oxford University Press, 28(3), pp. 1157–1172. doi: 10.1093/molbev/msq302.
Castresana, J. (2000) ‘Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis’, *Molecular Biology and Evolution*. Oxford University Press, 17(4), pp. 540–552. doi: 10.1093/oxfordjournals.molbev.a026334.
Cavalier-Smith, T. (1981) ‘Eukaryote kingdoms: Seven or nine?’, *BioSystems*, 14(3–4), pp. 461–481. doi: 10.1016/0303-2647(81)90050-2.
Cavalier-Smith, T. (1998) ‘A revised six-kingdom system of life.’, *Biological reviews*

of the Cambridge Philosophical Society, 73(3), pp. 203–66.
Cavalier-Smith, T. (1999) 'Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree.', *The Journal of eukaryotic microbiology*, 46(4), pp. 347–366. doi: 10.1111/j.1550-7408.1999.tb04614.x.
Cavalier-Smith, T. and Chao, E. E. Y. (2006) 'Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista)', *Journal of Molecular Evolution*. Springer-Verlag, 62(4), pp. 388–420. doi: 10.1007/s00239-004-0353-8.
Ceballos, G., Ehrlich, P. R. and Dirzo, R. (2017) 'Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(30), pp. E6089–E6096. doi: 10.1073/pnas.1704949114.
Cereghino, J. L. and Cregg, J. M. (2000) 'Heterologous protein expression in the methylotrophic yeast Pichia pastoris', *FEMS Microbiology Reviews*. Narnia, 24(1), pp. 45–66. doi: 10.1016/S0168-6445(99)00029-7.
Chang, S. T. (2006) 'The world mushroom industry: Trends and technological development', *International Journal of Medicinal Mushrooms*. Begel House Inc., 8(4), pp. 297–314. doi: 10.1615/IntJMedMushr.v8.i4.10.
Chang, Y. *et al.* (2015) 'Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants', *Genome Biology and Evolution*. Oxford University Press, 7(6), pp. 1590–1601. doi: 10.1093/gbe/evv090.
Chao, A. (1984) 'Non-parametric estimation of the classes in a population', *Scandinavian Journal of Statistics*. WileyBoard of the Foundation of the Scandinavian Journal of Statistics, 11(4), pp. 265–270. doi: 10.2307/4615964.
Chapman, J. A. *et al.* (2010) 'The dynamic genome of Hydra.', *Nature*. Nature Publishing Group, 464(7288), pp. 592–6. doi: 10.1038/nature08830.
Chaudhari, N. M., Gupta, V. K. and Dutta, C. (2016) 'BPGA-an ultra-fast pan-genome analysis pipeline', *Scientific Reports*. Nature Publishing Group, 6, p. 24373. doi: 10.1038/srep24373.
Chen, L. *et al.* (2016) 'Genome sequence of the edible cultivated mushroom Lentinula edodes (shiitake) reveals insights into lignocellulose degradation', *PLoS ONE*, 11(8). doi: 10.1371/journal.pone.0160336.
Cheon, S. A. *et al.* (2014) 'The Unfolded Protein Response (UPR) pathway in Cryptococcus', *Virulence*. Taylor & Francis, pp. 341–350. doi: 10.4161/viru.26774.
Choiseul, J., Doherty, G. and Roe, G. (2008) *Potato Varieties of Historical Interest in Ireland*, *Department of Agriculture, Fisheries and Food*.
Cissé, O. H. *et al.* (2013) 'Genome sequencing of the plant pathogen Taphrina deformans, the causal agent of peach leaf curl', *mBio*. American Society for Microbiology, 4(3), pp. e00055-13. doi: 10.1128/mBio.00055-13.
de Cock, A. W. A. M. *et al.* (2015) 'Phytopythium : molecular phylogeny and systematics', *Persoonia - Molecular Phylogeny and Evolution of Fungi*. Naturalis Biodiversity Center, 34(1), pp. 25–39. doi: 10.3767/003158515X685382.

Cock, P. J. A. *et al.* (2009) 'Biopython: Freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*. Oxford University Press, 25(11), pp. 1422–1423. doi: 10.1093/bioinformatics/btp163.
Coetzee, M. P. A., Wingfield, B. D. and Wingfield, M. J. (2018) 'Armillaria root-rot pathogens: Species boundaries and global distribution', *Pathogens*. Multidisciplinary Digital Publishing Institute (MDPI), 7(4). doi: 10.3390/pathogens7040083.
Cohen, O., Gophna, U. and Pupko, T. (2011) 'The complexity hypothesis revisited: Connectivity Rather Than function constitutes a barrier to horizontal gene transfer', *Molecular Biology and Evolution*, 28(4), pp. 1481–1489. doi: 10.1093/molbev/msq333.
Collins, C. *et al.* (2013) 'Genomic and proteomic dissection of the ubiquitous plant pathogen, armillaria mellea: Toward a new infection model system', *Journal of Proteome Research*. American Chemical Society, 12(6), pp. 2552–2570. doi: 10.1021/pr301131t.
Colson, I., Delneri, D. and Oliver, S. G. (2004) 'Effects of reciprocal chromosomal translocations on the fitness of Saccharomyces cerevisiae', *EMBO Reports*. European Molecular Biology Organization, 5(4), pp. 392–398. doi: 10.1038/sj.embor.7400123.
Conway, J. R., Lex, A. and Gehlenborg, N. (2017) 'UpSetR: An R package for the visualization of intersecting sets and their properties', *Bioinformatics*. Oxford University Press, 33(18), pp. 2938–2940. doi: 10.1093/bioinformatics/btx364.
Cooke, D. E. *et al.* (2000) 'A molecular phylogeny of Phytophthora and related oomycetes.', *Fungal genetics and biology : FG & B*. Academic Press, 30(1), pp. 17–32. doi: 10.1006/fgbi.2000.1202.
Corradi, N. *et al.* (2010) 'The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis', *Nature Communications*. Nature Publishing Group, 1(6), p. 77. doi: 10.1038/ncomms1082.
Costanzo, M. C. *et al.* (2006) 'The Candida Genome Database: Facilitating research on Candida albicans molecular biology', *FEMS Yeast Research*. Narnia, 6(5), pp. 671–684. doi: 10.1111/j.1567-1364.2006.00074.x.
Cotton, J. A. and McInerney, J. O. (2010) 'Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 107(40), pp. 17252–17255. doi: 10.1073/pnas.1000265107.
Cotton, J. A. and Page, R. D. M. (2003) 'Gene tree parsimony vs. uninode coding for phylogenetic reconstruction', *Molecular Phylogenetics and Evolution*, 29(2), pp. 298–308. doi: 10.1016/S1055-7903(03)00109-X.
Craig Venter, J. *et al.* (2001) 'The sequence of the human genome', *Science*. American Association for the Advancement of Science, 291(5507), pp. 1304–1351. doi: 10.1126/science.1058040.
Creevey, C. J. *et al.* (2004) 'Does a tree-like phylogeny only exist at the tips in the prokaryotes?', *Proceedings. Biological sciences / The Royal Society*, 271(1557), pp. 2551–2558. doi: 10.1098/rspb.2004.2864.
Creevey, C. J. and McInerney, J. O. (2005) 'Clann: Investigating phylogenetic information through supertree analyses', *Bioinformatics*. Oxford University Press,

21(3), pp. 390–392. doi: 10.1093/bioinformatics/bti020.
Creevey, C. J. and McInerney, J. O. (2009) 'Trees from trees: Construction of phylogenetic supertrees using CLANN', *Methods in Molecular Biology*, 537, pp. 139–161. doi: 10.1007/978-1-59745-251-9_7.
Croll, D. *et al.* (2015) 'The impact of recombination hotspots on genome evolution of a fungal plant pathogen', *Genetics*. Genetics Society of America, 201(3), pp. 1213–1228. doi: 10.1534/genetics.115.180968.
Crosby, A. W. (1972) *The Columbian Exchange: Biological and Cultural Consequences of 1492*, *Contributions in American studies*. doi: EB MB CROSB.
Csurös, M. (2010) 'Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood', *Bioinformatics*. Oxford University Press, 26(15), pp. 1910–1912. doi: 10.1093/bioinformatics/btq315.
Cuomo, C. A. and Birren, B. W. (2010) 'The fungal genome initiative and lessons learned from genome sequencing', *Methods in Enzymology*, 470(C), pp. 833–855. doi: 10.1016/S0076-6879(10)70034-3.
D'erchia, A. M. *et al.* (1996) 'The guinea-pig is not a rodent', *Nature*, 381(6583), pp. 597–600. doi: 10.1038/381597a0.
Dagan, T., Artzy-Randrup, Y. and Martin, W. (2008) 'Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(29), pp. 10039–44. doi: 10.1073/pnas.0800679105.
Dale, A. L. *et al.* (2019) 'Mitotic recombination and rapid genome evolution in the invasive forest pathogen phytophthora ramorum', *mBio*, 10(2), pp. e02452-18. doi: 10.1128/mBio.02452-18.
Danchin, E. G. J. *et al.* (2010) 'Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes.', *Proceedings of the National Academy of Sciences of the United States of America*, 107(41), pp. 17651–6. doi: 10.1073/pnas.1008486107.
Darriba, D. *et al.* (2011) 'ProtTest 3: fast selection of best-fit models of protein evolution.', *Bioinformatics*, 27(8), pp. 1164–5. doi: 10.1093/bioinformatics/btr088.
Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection*, *D. Appleton and Company*. doi: 10.1007/s11664-006-0098-9.
Dean, R. *et al.* (2012) 'The Top 10 fungal pathogens in molecular plant pathology', *Molecular Plant Pathology*. John Wiley & Sons, Ltd (10.1111), 13(4), pp. 414–430. doi: 10.1111/j.1364-3703.2011.00783.x.
Delsuc, F., Brinkmann, H. and Philippe, H. (2005) 'Phylogenomics and the reconstruction of the tree of life.', *Nature reviews. Genetics*. Nature Publishing Group, 6(5), pp. 361–375. doi: 10.1038/nrg1603.
Deniélou, Y. P. *et al.* (2011) 'Bacterial syntenies: An exact approach with gene quorum', *BMC Bioinformatics*. BioMed Central, 12, p. 193. doi: 10.1186/1471-2105-12-193.
Devillers, H. and Neuvéglise, C. (2019) 'Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica H222', *Microbiology Resource Announcements*. American Society

for Microbiology (ASM), 8(4). doi: 10.1128/mra.01547-18.
Dick, M. W. (2001) *Straminipilous Fungi Kluwer Academic Publishers*, *The Netherlands*. Kluwer Academic Publishers. doi: 10.1007/978-94-015-9733-3.
Diene, S. M. *et al.* (2013) 'The rhizome of the multidrug-resistant enterobacter aerogenes genome reveals how new "Killer Bugs" are created because of a sympatric lifestyle', *Molecular Biology and Evolution*, 30(2), pp. 369–383. doi: 10.1093/molbev/mss236.
van Dijk, E. L. *et al.* (2018) 'The Third Revolution in Sequencing Technology', *Trends in Genetics*, 34(9), pp. 666–681. doi: 10.1016/j.tig.2018.05.008.
Dong, S., Raffaele, S. and Kamoun, S. (2015) 'The two-speed genomes of filamentous pathogens: Waltz with plants', *Current Opinion in Genetics and Development*, 35, pp. 57–65. doi: 10.1016/j.gde.2015.09.001.
Doolittle, W. F. (1999) 'Phylogenetic classification and the universal tree', *Science*. American Association for the Advancement of Science, 284(5423), pp. 2124–2128. doi: 10.1126/science.284.5423.2124.
Dotzler, N. *et al.* (2009) 'Acaulosporoid glomeromycotan spores with a germination shield from the 400-million-year-old Rhynie chert', *Mycological Progress*. Springer-Verlag, 8(1), pp. 9–18. doi: 10.1007/s11557-008-0573-1.
Dujon, B. *et al.* (2004) 'Genome evolution in yeasts', *Nature*, 430(6995), pp. 35–44. doi: 10.1038/nature02579.
Dujon, B. A. and Louis, E. J. (2017) 'Genome diversity and evolution in the budding yeasts (Saccharomycotina)', *Genetics*. Genetics Society of America, 206(2), pp. 717–750. doi: 10.1534/genetics.116.199216.
Dunn, B. *et al.* (2012) 'Analysis of the Saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments', *Genome Research*, 22(5), pp. 908–924. doi: 10.1101/gr.130310.111.
Dunning Hotopp, J. C. (2011) 'Horizontal gene transfer between bacteria and animals', *Trends in Genetics*, 27(4), pp. 157–163. doi: 10.1016/j.tig.2011.01.005.
Earle, G. and Hintz, W. (2014) 'New Approaches for Controlling Saprolegnia parasitica, the Causal Agent of a Devastating Fish Disease.', *Tropical life sciences research*. School of Medical Sciences, Universiti Sains Malaysia, 25(2), pp. 101–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27073602 (Accessed: 27 August 2019).
Edgar, Robert C (2004) 'MUSCLE: a multiple sequence alignment method with reduced time and space complexity.', *BMC bioinformatics*, 5, p. 113. doi: 10.1186/1471-2105-5-113.
Edgar, Robert C. (2004) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*. Oxford University Press, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.
Eevers, N. *et al.* (2015) 'Draft Genome Sequence of Methylobacterium radiotolerans, a DDE-Degrading and Plant Growth-Promoting Strain Isolated from Cucurbita pepo', *Genome Announcements*, 3(3), pp. e00488-15. doi: 10.1128/genomeA.00488-15.

Eisen, J. A. and Fraser, C. M. (2003) ‘Phylogenomics: Intersection of Evolution and Genomics’, *Science*, 300(5626), pp. 1706–1707. doi: 10.1126/science.1086292.
Ekblom, R. and Wolf, J. B. W. (2014) ‘A field guide to whole-genome sequencing, assembly and annotation’, *Evolutionary Applications*. Wiley-Blackwell, 7(9), pp. 1026–1042. doi: 10.1111/eva.12178.
Emanuelsson, O. *et al.* (2000) ‘Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence’, *Journal of Molecular Biology*, 300(4), pp. 1005–1016. doi: 10.1006/jmbi.2000.3903.
Engel, S. R. *et al.* (2014) ‘The reference genome sequence of Saccharomyces cerevisiae: then and now.’, *G3 (Bethesda)*. Genetics Society of America, 4(3), pp. 389–98. doi: 10.1534/g3.113.008995.
Engel, S. R. and Cherry, J. M. (2013) ‘The new modern era of yeast genomics: Community sequencing and the resulting annotation of multiple Saccharomyces cerevisiae strains at the Saccharomyces Genome Database’, *Database*. Oxford University Press, 2013, p. bat012. doi: 10.1093/database/bat012.
Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) ‘An efficient algorithm for large-scale detection of protein families’, *Nucleic Acids Research*. Narnia, 30(7), pp. 1575–1584. doi: 10.1093/nar/30.7.1575.
Estévez, M. *et al.* (2002) ‘X-ray crystallographic and kinetic correlation of a clinically observed human fumarase mutation.’, *Protein science : a publication of the Protein Society*, 11(6), pp. 1552–7. doi: 10.1110/ps.0201502.
Evans, B. J. *et al.* (2017) ‘Evolution of the largest mammalian genome’, *Genome Biology and Evolution*. Oxford University Press, 9(6), pp. 1711–1724. doi: 10.1093/gbe/evx113.
Eyal, Z. *et al.* (1997) *The Septoria Diseases of Wheat: Concepts and Methods of Disease Management*, *Rust Diseases of Wheat: Concepts and methods of disease management.*
Faith, D. P. and Cranston, P. S. (1991) ‘Could a Cladogram This Short Have Arisen By Chance Alone?: on Permutation Tests for Cladistic Structure’, *Cladistics*. Blackwell Publishing Ltd, 7(1), pp. 1–28. doi: 10.1111/j.1096-0031.1991.tb00020.x.
Farris, J. S. (1977) ‘Phylogenetic analysis under dollo’s law’, *Systematic Biology*, 26(1), pp. 77–88. doi: 10.1093/sysbio/26.1.77.
Federhen, S. (2012) ‘The NCBI Taxonomy database’, *Nucleic Acids Res.*, 40(D1), pp. D136–D143. doi: 10.1093/nar/gkr1178.
Fedorova, N. D. *et al.* (2008) ‘Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus’, *PLoS Genetics*. Edited by P. M. Richardson, 4(4), p. e1000046. doi: 10.1371/journal.pgen.1000046.
Felsenstein, J. (1978) ‘Cases in which Parsimony or Compatibility Methods Will be Positively Misleading’, *Systematic Zoology*, 27(4), p. 401. doi: 10.2307/2412923.
Felsenstein, J. (1989) ‘PHYLIP - Phylogeny inference package - v3.2’, *Cladistics*, 5(2), pp. 164–166. doi: 10.1111/j.1096-0031.1989.tb00562.x.
Fernández-Fueyo, E. *et al.* (2014) ‘Ligninolytic peroxidase genes in the oyster mushroom genome: Heterologous expression, molecular structure, catalytic and

stability properties, and lignin-degrading ability', *Biotechnology for Biofuels*, 7(1). doi: 10.1186/1754-6834-7-2.
Fernandez, J. and Orth, K. (2018) 'Rise of a Cereal Killer: The Biology of Magnaporthe oryzae Biotrophic Growth', *Trends in Microbiology*. Elsevier, 26(7), pp. 582–597. doi: 10.1016/j.tim.2017.12.007.
Fiers, W. *et al.* (1976) 'Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.', *Nature*. Nature Publishing Group, 260(5551), pp. 500–7. doi: 10.1038/260500a0.
De Fine Licht, H. H., Jensen, A. B. and Eilenberg, J. (2017) 'Comparative transcriptomics reveal host-specific nucleotide variation in entomophthoralean fungi', *Molecular Ecology*, 26(7), pp. 2092–2110. doi: 10.1111/mec.13863.
Finn, R. D. *et al.* (2015) 'The Pfam protein families database: Towards a more sustainable future', *Nucleic acids research*. Oxford University Press, 44(D1), pp. D279–D285. doi: 10.1093/nar/gkv1344.
Fisher, M. C., Garner, T. W. J. and Walker, S. F. (2009) 'Global Emergence of Batrachochytrium dendrobatidis and Amphibian Chytridiomycosis in Space, Time, and Host', *Annual Review of Microbiology*, 63(1), pp. 291–310. doi: 10.1146/annurev.micro.091208.073435.
Fisk, D. G. *et al.* (2006) 'Saccharomyces cerevisiae S288C genome annotation: A working hypothesis', *Yeast*. NIH Public Access, 23(12), pp. 857–865. doi: 10.1002/yea.1400.
Fitzpatrick, D. A. *et al.* (2006) 'A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis.', *BMC evolutionary biology*. BioMed Central, 6(1), p. 99. doi: 10.1186/1471-2148-6-99.
Fitzpatrick, D. A. *et al.* (2010) 'Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser.', *BMC genomics*, 11(1), p. 290. doi: 10.1186/1471-2164-11-290.
Fitzpatrick, D. A., Logue, M. E. and Butler, G. (2008) 'Evidence of recent interkingdom horizontal gene transfer between bacteria and Candida parapsilosis', *BMC Evolutionary Biology*, 8(1), p. 181. doi: 10.1186/1471-2148-8-181.
Fleischmann, R. D. *et al.* (1995) 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd', *Science*. American Association for the Advancement of Science, 269(5223), pp. 496–512. doi: 10.1126/science.7542800.
Forster, H. *et al.* (1990) 'Sequence Analysis of the Small Subunit Ribosomal RNAs of Three Zoosporic Fungi and Implications for Fungal Evolution', *Mycologia*. Taylor & Francis, 82(3), p. 306. doi: 10.2307/3759901.
Foster, P. G. (2004) 'Modeling compositional heterogeneity', *Systematic Biology*. Edited by T. Schultz. Narnia, 53(3), pp. 485–495. doi: 10.1080/10635150490445779.
Fouts, D. E. *et al.* (2012) 'PanOCT: Automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species', *Nucleic Acids Research*, 40(22), pp. e172–e172. doi: 10.1093/nar/gks757.
Frankel, S. J., Kliejunas, J. T. and Palmieri, K. M. (2008) *Proceedings of the sudden oak death third science symposium*, *Gen. Tech. Rep. PSW-GTR-214, Albany, CA:*

Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture. 491 p. doi: 10.2737/PSW-GTR-214.
Fraser, J. A. *et al.* (2005) 'Chromosomal translocation and segmental duplication in Cryptococcus neoformans', *Eukaryotic Cell*, 4(2), pp. 401–406. doi: 10.1128/EC.4.2.401-406.2005.
Friedlander, J. *et al.* (2016) 'Engineering of a high lipid producing Yarrowia lipolytica strain', *Biotechnology for Biofuels*. BioMed Central, 9(1), p. 77. doi: 10.1186/s13068-016-0492-3.
Friedman, R. and Hughes, A. L. (2001) 'Gene duplication and the structure of eukaryotic genomes.', *Genome research*. Cold Spring Harbor Laboratory Press, 11(3), pp. 373–81. doi: 10.1101/gr.155801.
Fry, W. E. and Mizubuti, E. S. (1998) 'Potato late blight', in *The Epidemiology of Plant Diseases*. Dordrecht: Springer Netherlands, pp. 371–388. doi: 10.1007/978-94-017-3302-1_18.
Gaastra, W. *et al.* (2010) 'Pythium insidiosum: An overview', *Veterinary Microbiology*, 146(1–2), pp. 1–16. doi: 10.1016/j.vetmic.2010.07.019.
Gachon, C. M. M. *et al.* (2009) 'Detection of differential host susceptibility to the marine oomycete pathogen Eurychasma dicksonii by real-time PCR: Not all algae are equal', *Applied and Environmental Microbiology*. American Society for Microbiology, 75(2), pp. 322–328. doi: 10.1128/AEM.01885-08.
Galagan, J. E. *et al.* (2003) 'The genome sequence of the filamentous fungus Neurospora crassa.', *Nature*, 422(6934), pp. 859–868. doi: 10.1038/nature01554.
Galagan, J. E., Henn, M. R., *et al.* (2005) 'Genomics of the fungal kingdom: Insights into eukaryotic biology', *Genome Research*. Cold Spring Harbor Laboratory Press, 15(12), pp. 1620–1631. doi: 10.1101/gr.3767105.
Galagan, J. E., Calvo, S. E., *et al.* (2005) 'Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae', *Nature*. Nature Publishing Group, 438(7071), pp. 1105–1115. doi: 10.1038/nature04341.
Gao, M. *et al.* (2017) 'Fungal lactamases: Their occurrence and function', *Frontiers in Microbiology*. Frontiers Media SA, 8(SEP), p. 1775. doi: 10.3389/fmicb.2017.01775.
Gardner, M. J. *et al.* (2002) 'Genome sequence of the human malaria parasite Plasmodium falciparum', *Nature*. Europe PMC Funders, 419(6906), pp. 498–511. doi: 10.1038/nature01097.
Gaston, D. and Roger, A. J. (2013) 'Functional Divergence and Convergent Evolution in the Plastid-Targeted Glyceraldehyde-3-Phosphate Dehydrogenases of Diverse Eukaryotic Algae', *PLoS ONE*. Edited by N. Nikolaidis. Public Library of Science, 8(7), p. e70396. doi: 10.1371/journal.pone.0070396.
Gel, B. and Serra, E. (2017) 'KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data', *Bioinformatics*. Edited by J. Hancock. Oxford University Press, 33(19), pp. 3088–3090. doi: 10.1093/bioinformatics/btx346.
Gerbod, D. *et al.* (2001) 'Phylogenetic Relationships of Class II Fumarase Genes from Trichomonad Species', *Molecular Biology and Evolution*, 18(8), pp. 1574–1584. doi: 10.1093/oxfordjournals.molbev.a003944.

Giaever, G. *et al.* (2002) 'Functional profiling of the Saccharomyces cerevisiae genome', *Nature*, 418(6896), pp. 387–391. doi: 10.1038/nature00935.
Giaever, G. and Nislow, C. (2014) 'The yeast deletion collection: A decade of functional genomics', *Genetics*. Genetics Society of America, 197(2), pp. 451–465. doi: 10.1534/genetics.114.161620.
Gigliotti, F., Limper, A. H. and Wright, T. (2014) 'Pneumocystis.', *Cold Spring Harbor perspectives in medicine*. Cold Spring Harbor Laboratory Press, 4(12), p. a019828. doi: 10.1101/cshperspect.a019828.
Gluck-Thaler, E. and Slot, J. C. (2015) 'Dimensions of Horizontal Gene Transfer in Eukaryotic Microbial Pathogens', *PLoS Pathogens*. Edited by D. A. Hogan. Public Library of Science, 11(10), p. e1005156. doi: 10.1371/journal.ppat.1005156.
Gobler, C. J. *et al.* (2011) 'Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(11), pp. 4352–7. doi: 10.1073/pnas.1016106108.
Goffeau, A. *et al.* (1996) 'Life with 6000 genes', *Science*, 274(5287), pp. 546–567. doi: 10.1126/science.274.5287.546.
Goffeau, A. and Vassarotti, A. (1991) 'The European project for sequencing the yeast genome', *Research in Microbiology*, 142(7–8), pp. 901–903. doi: 10.1016/0923-2508(91)90071-H.
Goheen, E. M. *et al.* (2002) 'Sudden Oak Death Caused by Phytophthora ramorum in Oregon', *Plant Disease*, 86(4), pp. 441–441. doi: 10.1094/pdis.2002.86.4.441c.
Goldenfeld, N. and Woese, C. (2007) 'Biology's next revolution.', *Nature*, 445(7126), p. 369. doi: 10.1038/445369a.
Golicz, A. A. *et al.* (2016) 'The pangenome of an agronomically important crop plant Brassica oleracea', *Nature Communications*. Nature Publishing Group, 7, p. 13390. doi: 10.1038/ncomms13390.
Golicz, A. A., Batley, J. and Edwards, D. (2016) 'Towards plant pangenomics', *Plant Biotechnology Journal*, pp. 1099–1105. doi: 10.1111/pbi.12499.
Gonzalez-Hilarion, S. *et al.* (2016) 'Intron retention-dependent gene regulation in Cryptococcus neoformans', *Scientific Reports*. Nature Publishing Group, 6(1), p. 32252. doi: 10.1038/srep32252.
Gordon, S. P. *et al.* (2017) 'Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure', *Nature Communications*. Nature Publishing Group, 8(1), p. 2184. doi: 10.1038/s41467-017-02292-8.
Goss, E. M. *et al.* (2014) 'The Irish potato famine pathogen Phytophthora infestans originated in central Mexico rather than the Andes', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 111(24), pp. 8791–8796. doi: 10.1073/pnas.1401884111.
Govers, F. and Gijzen, M. (2006) 'Phytophthora genomics: the plant destroyers' genome decoded.', *Molecular plant-microbe interactions : MPMI*, 19(12), pp. 1295–1301. doi: 10.1094/MPMI-19-1295.
Grenville-Briggs, L. *et al.* (2011) 'A molecular insight into algal-oomycete warfare: CDNA analysis of ectocarpus siliculosus infected with the basal oomycete eurychasma

dicksonii', *PLoS ONE*. Edited by D. A. Carter. Public Library of Science, 6(9), p. e24500. doi: 10.1371/journal.pone.0024500.
Griffin, D. (2015) 'The 30 years war: the fight against rhododendron', *The Irish Times*, 15 August. Available at: https://www.irishtimes.com/news/environment/the-30-years-war-the-fight-against-rhododendron-1.2317249 (Accessed: 27 August 2019).
Grigoriev, I. V., Cullen, D., *et al.* (2011) 'Fueling the future with fungal genomics', *Mycology*, 2(3), pp. 192–209. doi: 10.1080/21501203.2011.584577.
Grigoriev, I. V., Nordberg, H., *et al.* (2011) 'The Genome Portal of the Department of Energy Joint Genome Institute.', *Nucleic acids research*. Oxford University Press, 40(November 2011), pp. 1–7. doi: 10.1093/nar/gkr947.
Grigoriev, I. V. (2013) 'A Changing Landscape of Fungal Genomics', in *The Ecological Genomics of Fungi*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 3–20. doi: 10.1002/9781118735893.ch1.
Grigoriev, I. V. *et al.* (2014) 'MycoCosm portal: Gearing up for 1000 fungal genomes', *Nucleic Acids Research*. Oxford University Press, 42(D1), pp. D699-704. doi: 10.1093/nar/gkt1183.
Grünwald, N. J. *et al.* (2012) 'Emergence of the sudden oak death pathogen Phytophthora ramorum', *Trends in Microbiology*. Elsevier Current Trends, 20(3), pp. 131–138. doi: 10.1016/j.tim.2011.12.006.
Grünwald, N. J., Goss, E. M. and Press, C. M. (2008) 'Phytophthora ramorum: A pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals', *Molecular Plant Pathology*. John Wiley & Sons, Ltd (10.1111), 9(6), pp. 729–740. doi: 10.1111/j.1364-3703.2008.00500.x.
Gryganskyi, A. P. *et al.* (2018) 'Phylogenetic and phylogenomic definition of Rhizopus species', *G3: Genes, Genomes, Genetics*. Genetics Society of America, 8(6), pp. 2007–2018. doi: 10.1534/g3.118.200235.
Gu, Z. *et al.* (2005) 'Elevated evolutionary rates in the laboratory strain of Saccharomyces cerevisiae', *Proceedings of the National Academy of Sciences*, 102(4), pp. 1092–1097. doi: 10.1073/pnas.0409159102.
Guan, R. *et al.* (2016) 'Draft genome of the living fossil Ginkgo biloba', *GigaScience*, 5(1), p. 49. doi: 10.1186/s13742-016-0154-1.
Guarro, J., Gené, J. and Stchigel, A. M. (1999) 'Developments in fungal taxonomy', *Clinical Microbiology Reviews*, 12(3), pp. 454–500. doi: 0893-8512/99/$04.00?0.
Guindon, S. *et al.* (2010) 'New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.', *Systematic biology*, 59(3), pp. 307–21. doi: 10.1093/sysbio/syq010.
Guindon, S. and Gascuel, O. (2003) 'A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood', *Systematic Biology*, 52(5), pp. 696–704. doi: 10.1080/10635150390235520.
Gunderson, J. H. *et al.* (1987) 'Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes.', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 84(16), pp. 5823–5827. doi: 10.1073/pnas.84.16.5823.
Haas, B. J. *et al.* (2009) 'Genome sequence and analysis of the Irish potato famine

pathogen Phytophthora infestans.', *Nature*. Macmillan Publishers Limited. All rights reserved, 461(7262), pp. 393–8. doi: 10.1038/nature08358.
Haas, B. J. *et al.* (2013) 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature Protocols*. NIH Public Access, 8(8), pp. 1494–1512. doi: 10.1038/nprot.2013.084.
Hackett, J. D. *et al.* (2007) 'Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates', *Molecular Biology and Evolution*. Oxford University Press, 24(8), pp. 1702–1713. doi: 10.1093/molbev/msm089.
Hakariya, M., Hirose, D. and Tokumasu, S. (2007) 'A molecular phylogeny of Haptoglossa species, terrestrial peronosporomycetes (oomycetes) endoparasitic on nematodes', *Mycoscience*. Elsevier, 48(3), pp. 169–175. doi: 10.1007/s10267-007-0355-7.
Hall, C. and Dietrich, F. S. (2007) 'The reacquisition of biotin prototrophy in Saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering', *Genetics*. Genetics Society of America, 177(4), pp. 2293–2307. doi: 10.1534/genetics.107.074963.
Hampl, V. *et al.* (2009) 'Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups"', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 106(10), pp. 3859–64. doi: 10.1073/pnas.0807880106.
Hardham, A. R. and Blackman, L. M. (2018) 'Phytophthora cinnamomi', *Molecular Plant Pathology*, 19(2), pp. 260–285. doi: 10.1111/mpp.12568.
Harper, J. T., Waanders, E. and Keeling, P. J. (2005) 'On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes', *International Journal of Systematic and Evolutionary Microbiology*. Microbiology Society, 55(1), pp. 487–496. doi: 10.1099/ijs.0.63216-0.
Haverkort, A. J. *et al.* (2008) 'Societal costs of late blight in potato and prospects of durable resistance through cisgenic modification', *Potato Research*. Springer Netherlands, 51(1), pp. 47–57. doi: 10.1007/s11540-008-9089-y.
Hawksworth, D. L. (2001) 'The magnitude of fungal diversity: the 1.5 million species estimate revisited', *Mycological Research*, 105(12), pp. 1422–1432. doi: 10.1017/S0953756201004725.
Hayashi, T. *et al.* (2001) 'Complete genome sequence of enterohemorrhagic Eschelichia coli O157:H7 and genomic comparison with a laboratory strain K-12', *DNA Research*. Narnia, 8(1), pp. 11–22. doi: 10.1093/dnares/8.1.11.
Heath, I. B. (1980) 'Variant Mitoses in Lower Eukaryotes: Indicators of the Evolution of Mitosis?', *International Review of Cytology*, 64(C), pp. 1–80. doi: 10.1016/S0074-7696(08)60235-1.
Heather, J. M. and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*. Elsevier, 107(1), pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.
Henk, D. A. and Fisher, M. C. (2012) 'The gut fungus basidiobolus ranarum has a large

genome and different copy numbers of putatively functionally redundant elongation factor genes', *PLoS ONE*. Edited by J. E. Stajich, 7(2), p. e31268. doi: 10.1371/journal.pone.0031268.
Hibbett, D. and Glotzer, D. (2011) 'Where are all the undocumented fungal species? A study of Mortierella demonstrates the need for sequence-based classification', *New Phytologist*. John Wiley & Sons, Ltd (10.1111), 191(3), pp. 592–596. doi: 10.1111/j.1469-8137.2011.03819.x.
Hibbett, D. S. *et al.* (2007) 'A higher-level phylogenetic classification of the Fungi', *Mycological Research*, 111(5), pp. 509–547. doi: 10.1016/j.mycres.2007.03.004.
Hirsch, C. N. *et al.* (2014) 'Insights into the Maize Pan-Genome and Pan-Transcriptome', *The Plant Cell*. American Society of Plant Biologists, 26(1), pp. 121–135. doi: 10.1105/tpc.113.119982.
Hirt, R. P., Alsmark, C. and Embley, T. M. (2015) 'Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites.', *Current opinion in microbiology*, 23, pp. 155–62. doi: 10.1016/j.mib.2014.11.018.
Hittinger, C. T. *et al.* (2015) 'Genomics and the making of yeast biodiversity', *Current Opinion in Genetics and Development*, 35, pp. 100–109. doi: 10.1016/j.gde.2015.10.008.
Hoffman, M. T. *et al.* (2013) 'Endohyphal Bacterium Enhances Production of Indole-3-Acetic Acid by a Foliar Fungal Endophyte', *PLoS ONE*, 8(9). doi: 10.1371/journal.pone.0073132.
Hoffman, M. T. and Arnold, A. E. (2010) 'Diverse bacteria inhabit living hyphae of phylogenetically diverse fungal endophytes', *Applied and Environmental Microbiology*, 76(12), pp. 4063–4075. doi: 10.1128/AEM.02928-09.
Hoffmann, K., Voigt, K. and Kirk, P. M. (2011) 'Mortierellomycotina subphyl. nov., based on multi-gene genealogies', *Mycotaxon*, 115(1), pp. 353–363. doi: 10.5248/115.353.
Hofmann, G., McIntyre, M. and Nielsen, J. (2003) 'Fungal genomics beyond Saccharomyces cerevisiae?', *Current Opinion in Biotechnology*. Elsevier Current Trends, 14(2), pp. 226–231. doi: 10.1016/S0958-1669(03)00020-X.
Hogg, J. S. *et al.* (2007) 'Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains.', *Genome biology*. BioMed Central, 8(6), p. R103. doi: 10.1186/gb-2007-8-6-r103.
Holley, R. W. *et al.* (1965) 'Structure of a Ribonucleic Acid', *Science*, 147(3664), pp. 1462–5. doi: 10.1126/science.147.3664.1462.
Holm, D. K. *et al.* (2014) 'Molecular and chemical characterization of the biosynthesis of the 6-MSA-derived meroterpenoid yanuthone D in Aspergillus niger', *Chemistry and Biology*. Cell Press, 21(4), pp. 519–529. doi: 10.1016/j.chembiol.2014.01.013.
Holton, T. A. and Pisani, D. (2010) 'Deep genomic-scale analyses of the metazoa reject coelomata: Evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm', *Genome Biology and Evolution*, 2(1), pp. 310–324. doi: 10.1093/gbe/evq016.

Hou, J. *et al.* (2014) 'Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in saccharomyces cerevisiae', *Current Biology*. NIH Public Access, 24(10), pp. 1153–1159. doi: 10.1016/j.cub.2014.03.063.
Howe, K., Bateman, A. and Durbin, R. (2002) 'QuickTree: building huge Neighbour-Joining trees of protein sequences', *Bioinformatics*, 18(11), pp. 1546–1547. doi: 10.1093/bioinformatics/18.11.1546.
Hu, Z. *et al.* (2017) 'EUPAN enables pan-genome studies of a large number of eukaryotic genomes', *Bioinformatics*. Edited by O. Stegle. Narnia, 33(15), pp. 2408–2409. doi: 10.1093/bioinformatics/btx170.
Huang, J. (2013) 'Horizontal gene transfer in eukaryotes: The weak-link model', *BioEssays*, 35(10), pp. 868–875. doi: 10.1002/bies.201300007.
Huang, J. H. *et al.* (2013) 'Six new species of Pythiogeton in Taiwan, with an account of the molecular phylogeny of this genus', *Mycoscience*, 54(2), pp. 130–147. doi: 10.1016/j.myc.2012.09.007.
Huelsenbeck, J. P. *et al.* (2001) 'Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology', *Science*, 294(5550), pp. 2310–2314. doi: 10.1126/science.1065889.
Huelsenbeck, J. P. and Hillis, D. M. (1993) 'Success of phylogenetic methods in the four taxon case', *Systematic Biology*. Oxford University Press, 42(3), pp. 247–264. doi: 10.1093/sysbio/42.3.247.
Huffnagle, G. B. and Noverr, M. C. (2013) 'The emerging world of the fungal microbiome', *Trends in Microbiology*. NIH Public Access, 21(7), pp. 334–341. doi: 10.1016/j.tim.2013.04.002.
Hulvey, J. P., Padgett, D. E. and Bailey, J. C. (2007) 'Species boundaries within Saprolegnia (Saprolegniales, Oomycota) based on morphological and DNA sequence data.', *Mycologia*. Mycological Society of America, 99(3), pp. 421–429. doi: 10.3852/mycologia.99.3.421.
Hunter, S. *et al.* (2012) 'InterPro in 2011: New developments in the family and domain prediction database', *Nucleic Acids Research*, 40(D1), pp. D306–D312. doi: 10.1093/nar/gkr948.
Huson, D. H. and Bryant, D. (2006) 'Application of phylogenetic networks in evolutionary studies', *Molecular Biology and Evolution*, 23(2), pp. 254–267. doi: 10.1093/molbev/msj030.
Husson, C. *et al.* (2015) 'Evidence for homoploid speciation in Phytophthora alni supports taxonomic reclassification in this species complex', *Fungal Genetics and Biology*, 77, pp. 12–21. doi: 10.1016/j.fgb.2015.02.013.
Hutchison, C. A. (2007) 'DNA sequencing: Bench to bedside and beyond', *Nucleic Acids Research*. Oxford University Press, 35(18), pp. 6227–6237. doi: 10.1093/nar/gkm688.
Huynen, M. (1999) 'Lateral Gene Transfer, Genome Surveys, and the Phylogeny of Prokaryotes', *Science*. doi: 10.1126/science.286.5444.1443a.
Hyde, K. D. *et al.* (2014) 'One stop shop: backbones trees for important phytopathogenic genera: I (2014)', *Fungal Diversity*. Springer Netherlands, 67(1), pp.

21–125. doi: 10.1007/s13225-014-0298-1.
Iomaire, M. M. C. and Gallagher, P. Ó. (2009) 'The potato in Irish cuisine and culture', *Journal of Culinary Science and Technology*, 7(2–3), pp. 152–167. doi: 10.1080/15428050903313457.
Jackson, A. P. *et al.* (2009) 'Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans', *Genome Research*. Cold Spring Harbor Laboratory Press, 19(12), pp. 2231–2244. doi: 10.1101/gr.097501.109.
Jain, R., Rivera, M. C. and Lake, J. A. (1999) 'Horizontal gene transfer among genomes: The complexity hypothesis', *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp. 3801–3806. doi: 10.1073/pnas.96.7.3801.
James, T. Y., Letcher, P. M., *et al.* (2006) 'A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota)', *Mycologia*, 98(6), pp. 860–871. doi: 10.3852/mycologia.98.6.860.
James, T. Y., Kauff, F., *et al.* (2006) 'Reconstructing the early evolution of Fungi using a six-gene phylogeny', *Nature*, 443(7113), pp. 818–822. doi: 10.1038/nature05110.
Janbon, G. *et al.* (2014) 'Analysis of the Genome and Transcriptome of Cryptococcus neoformans var. grubii Reveals Complex RNA Expression and Microevolution Leading to Virulence Attenuation', *PLoS Genetics*. Edited by M. Freitag. Public Library of Science, 10(4), p. e1004261. doi: 10.1371/journal.pgen.1004261.
Jandrasits, C. *et al.* (2018) 'seq-seq-pan: building a computational pan-genome data structure on whole genome alignment', *BMC Genomics*. BioMed Central, 19(1), p. 47. doi: 10.1186/s12864-017-4401-3.
Janouskovec, J. *et al.* (2010) 'A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 107(24), pp. 10949–54. doi: 10.1073/pnas.1003335107.
Janssen, D. B. (2004) 'Evolving haloalkane dehalogenases', *Current Opinion in Chemical Biology*, 8(2), pp. 150–159. doi: 10.1016/j.cbpa.2004.02.012.
Jiang, R. H. Y. *et al.* (2013) 'Distinctive Expansion of Potential Virulence Genes in the Genome of the Oomycete Fish Pathogen Saprolegnia parasitica', *PLoS genetics*. Edited by J. M. McDowell. Public Library of Science, 9(6), p. e1003272. doi: 10.1371/journal.pgen.1003272.
Jiang, R. H. Y. and Tyler, B. M. (2012) 'Mechanisms and evolution of virulence in oomycetes.', *Annual review of phytopathology*. Annual Reviews, 50(1), pp. 295–318. doi: 10.1146/annurev-phyto-081211-172912.
Jiang, Y. *et al.* (2015) 'Local generation of fumarate promotes DNA repair through inhibition of histone H3 demethylation.', *Nature cell biology*. Nature Publishing Group, 17(9), pp. 1158–68. doi: 10.1038/ncb3209.
Jones, M. D. M., Forn, I., *et al.* (2011) 'Discovery of novel intermediate forms redefines the fungal tree of life', *Nature*. Nature Research, 474(7350), pp. 200–203. doi: 10.1038/nature09984.
Jones, M. D. M., Richards, T. A., *et al.* (2011) 'Validation and justification of the

phylum name Cryptomycota phyl. nov.', *IMA fungus*, 2(2), pp. 173–5. doi: 10.5598/imafungus.2011.02.02.08.
Jones, P. *et al.* (2014) 'InterProScan 5: Genome-scale protein function classification', *Bioinformatics*. Oxford University Press, 30(9), pp. 1236–1240. doi: 10.1093/bioinformatics/btu031.
Jones, T. *et al.* (2004) 'The diploid genome sequence of Candida albicans', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 101(19), pp. 7329–7334. doi: 10.1073/pnas.0401648101.
Judelson, H. S. (2012) 'Dynamics and innovations within oomycete genomes: Insights into biology, pathology, and evolution', *Eukaryotic Cell*. American Society for Microbiology, 11(11), pp. 1304–1312. doi: 10.1128/EC.00155-12.
Jurka, J. *et al.* (2005) 'Repbase Update, a database of eukaryotic repetitive elements', *Cytogenetic and Genome Research*, 110(1–4), pp. 462–467. doi: 10.1159/000084979.
Kabir, M. A. and Ahmad, Z. (2013) 'Candida Infections and Their Prevention', *ISRN Preventive Medicine*. Hindawi Limited, 2013, pp. 1–13. doi: 10.5402/2013/763628.
Kamoun, S. (2003) 'Molecular genetics of pathogenic oomycetes', *Eukaryotic Cell*. American Society for Microbiology, 2(2), pp. 191–199. doi: 10.1128/EC.2.2.191-199.2003.
Katinka, M. D. *et al.* (2001) 'Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi', *Nature*, 414(6862), pp. 450–453. doi: 10.1038/35106579.
Kaul, S. *et al.* (2000) 'Analysis of the genome sequence of the flowering plant Arabidopsis thaliana', *Nature*. Nature Publishing Group, 408(6814), pp. 796–815. doi: 10.1038/35048692.
Keeling, P. J. (2001) 'Foraminifera and Cercozoa are related in actin phylogeny: two orphans find a home?', *Molecular biology and evolution*, 18(8), pp. 1551–1557. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11470846 (Accessed: 21 September 2016).
Keeling, P. J. (2009) 'Chromalveolates and the evolution of plastids by secondary endosymbiosis', *Journal of Eukaryotic Microbiology*, 56(1), pp. 1–8. doi: 10.1111/j.1550-7408.2008.00371.x.
Keeling, P. J. and Palmer, J. D. (2008) 'Horizontal gene transfer in eukaryotic evolution.', *Nature reviews. Genetics*. Nature Publishing Group, 9(8), pp. 605–18. doi: 10.1038/nrg2386.
Keller, N. P., Turner, G. and Bennett, J. W. (2005) 'Fungal secondary metabolism — from biochemistry to genomics', *Nature Reviews Microbiology*. Nature Publishing Group, 3(12), pp. 937–947. doi: 10.1038/nrmicro1286.
Kellis, M., Birren, B. W. and Lander, E. S. (2004) 'Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.', *Nature*. Nature Publishing Group, 428 VN-(6983), pp. 617–624. doi: 10.1038/nature02424.
Kemen, E. *et al.* (2011) 'Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of Arabidopsis thaliana', *PLoS biology*. Edited by F. M. Ausubel. Cambridge University Press, 9(7), p. e1001094. doi:

10.1371/journal.pbio.1001094.
Khaldi, N. *et al.* (2010) 'SMURF: Genomic mapping of fungal secondary metabolite clusters', *Fungal Genetics and Biology*, 47(9), pp. 736–741. doi: 10.1016/j.fgb.2010.06.003.
Kim, I.-K. *et al.* (2008) 'Crystal structure of a new type of NADPH-dependent quinone oxidoreductase (QOR2) from Escherichia coli.', *Journal of molecular biology*, 379(2), pp. 372–84. doi: 10.1016/j.jmb.2008.04.003.
Kinealy, C. (1997) 'Food Exports from Ireland 1846-47', *History Ireland*, pp. 32–36. doi: 10.2307/27724428.
King, R. *et al.* (2018) 'Inter-genome comparison of the Quorn fungus Fusarium venenatum and the closely related plant infecting pathogen Fusarium graminearum', *BMC Genomics*. BioMed Central, 19(1), p. 269. doi: 10.1186/s12864-018-4612-2.
Kirk, P. M. *et al.* (2008) *Ainsworth & Bisby's dictionary of the fungi*. 10th edn. Wallingford, UK: CABI.
Klopfenstein, D. V. *et al.* (2018) 'GOATOOLS: A Python library for Gene Ontology analyses', *Scientific Reports*. Nature Publishing Group, 8(1), p. 10872. doi: 10.1038/s41598-018-28948-z.
Kluge, A. G. (1989) 'A concern for evidence and a phylogenetic hypothesis of relationships among epicrates (Boidae, serpentes)', *Systematic Biology*, 38(1), pp. 7–25. doi: 10.1093/sysbio/38.1.7.
Knox, B. P. *et al.* (2016) 'Characterization of Aspergillus fumigatus Isolates from Air and Surfaces of the International Space Station', *mSphere*. American Society for Microbiology Journals, 1(5), pp. e00227-16. doi: 10.1128/mSphere.00227-16.
Kohler, A. *et al.* (2015) 'Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists', *Nature Genetics*. Nature Publishing Group, 47(4), pp. 410–415. doi: 10.1038/ng.3223.
Koonin, E. V *et al.* (2004) 'A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.', *Genome biology*, 5(2), p. R7. doi: 10.1186/gb-2004-5-2-r7.
Kosa, G. *et al.* (2018) 'High-throughput screening of Mucoromycota fungi for production of low- and high-value lipids', *Biotechnology for Biofuels*. BioMed Central, 11(1), p. 66. doi: 10.1186/s13068-018-1070-7.
Koski, L. B., Morton, R. A. and Golding, G. B. (2001) 'Codon Bias and Base Composition Are Poor Indicators of Horizontally Transferred Genes', *Molecular Biology and Evolution*, 18(3), pp. 404–412. doi: 10.1093/oxfordjournals.molbev.a003816.
Kosmidis, C. and Denning, D. W. (2015) 'The clinical spectrum of pulmonary aspergillosis.', *Thorax*. BMJ Publishing Group Ltd, 70(3), pp. 270–7. doi: 10.1136/thoraxjnl-2014-206291.
Kousha, M., Tadi, R. and Soubani, A. O. (2011) 'Pulmonary aspergillosis: A clinical review', *European Respiratory Review*, 20(121), pp. 156–174. doi: 10.1183/09059180.00001011.
Kroon, L. P. N. M. *et al.* (2004) 'Phylogenetic analysis of Phytophthora species based

on mitochondrial and nuclear DNA sequences', *Fungal Genetics and Biology*, 41(8), pp. 766–782. doi: 10.1016/j.fgb.2004.03.007.
Ku, C. and Martin, W. F. (2016) 'A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70 % rule', *BMC Biology*. BioMed Central, 14(1), p. 89. doi: 10.1186/s12915-016-0315-9.
Kück, P. and Meusemann, K. (2010) 'FASconCAT: Convenient handling of data matrices', *Molecular Phylogenetics and Evolution*, 56(3), pp. 1115–1118. doi: 10.1016/j.ympev.2010.04.024.
Kuramae, E. E. *et al.* (2006) 'Phylogenomics reveal a robust fungal tree of life', *FEMS Yeast Research*. Blackwell Publishing Ltd, 6(8), pp. 1213–1220. doi: 10.1111/j.1567-1364.2006.00119.x.
Laich, F., Fierro, F. and Martín, J. F. (2002) 'Production of penicillin by fungi growing on food products: Identification of a complete penicillin gene cluster in Penicillium griseofulvum and a truncated cluster in Penicillium verrucosum', *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 68(3), pp. 1211–1219. doi: 10.1128/AEM.68.3.1211-1219.2002.
Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.
Lapierre, P. and Gogarten, J. P. (2009) 'Estimating the size of the bacterial pan-genome', *Trends in Genetics*. Elsevier Current Trends, pp. 107–110. doi: 10.1016/j.tig.2008.12.004.
Lapierre, P., Lasek-Nesselquist, E. and Gogarten, J. P. (2014) 'The impact of HGT on phylogenomic reconstruction methods', *Briefings in Bioinformatics*. Oxford University Press, 15(1), pp. 79–90. doi: 10.1093/bib/bbs050.
Lapointe, F.-J. and Cucumel, G. (1997) 'The Average Consensus Procedure: Combination of Weighted Trees Containing Identical or Overlapping Sets of Taxa', *Systematic Biology*. Oxford University Press, 46(2), pp. 306–312. doi: 10.1093/sysbio/46.2.306.
Lartillot, N. *et al.* (2013) 'Phylobayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment', *Systematic Biology*. Oxford University Press, 62(4), pp. 611–615. doi: 10.1093/sysbio/syt022.
Lartillot, N., Brinkmann, H. and Philippe, H. (2007) 'Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model.', *BMC evolutionary biology*, 7 Suppl 1(Suppl 1), p. S4. doi: 10.1186/1471-2148-7-S1-S4.
Lartillot, N. and Philippe, H. (2004) 'A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process', *Molecular Biology and Evolution*. Narnia, 21(6), pp. 1095–1109. doi: 10.1093/molbev/msh112.
Lavania, U. C. (2015) 'Emerging trends in polyploidy research', *Nucleus (India)*. Springer India, 58(1), pp. 1–2. doi: 10.1007/s13237-015-0136-1.
Lefebure, T. *et al.* (2010) 'Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept', *Genome Biology and Evolution*, 2(1), pp. 646–655. doi: 10.1093/gbe/evq048.
Léjohn, H. B. (1974) 'Biochemical Parameters of Fungal Phylogenetics', in

Dobzhansky, T., Hecht, M. K., and Steere, W. C. (eds) *Evolutionary Biology*. Boston, MA: Springer US, pp. 79–125. doi: 10.1007/978-1-4615-6944-2_3.
Lengeler, K. B. *et al.* (2000) 'Signal Transduction Cascades Regulating Fungal Development and Virulence', *Microbiology and Molecular Biology Reviews*. American Society for Microbiology (ASM), 64(4), pp. 746–785. doi: 10.1128/MMBR.64.4.746-785.2000.
Leonard, G. *et al.* (2018) 'Comparative genomic analysis of the "pseudofungus" Hyphochytrium catenoides', *Open Biology*, 8(1), p. 170184. doi: 10.1098/rsob.170184.
Letunic, I. and Bork, P. (2007) 'Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation', *Bioinformatics*. Oxford University Press, 23(1), pp. 127–128. doi: 10.1093/bioinformatics/btl529.
Letunic, I. and Bork, P. (2016) 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic Acids Research*. Oxford University Press, 44(W1), pp. W242–W245. doi: 10.1093/nar/gkw290.
Lévesque, C. A. *et al.* (2010) 'Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire.', *Genome biology*. BioMed Central, 11(7), p. R73. doi: 10.1186/gb-2010-11-7-r73.
Lévesque, C. A. and de Cock, A. W. a M. (2004) 'Molecular phylogeny and taxonomy of the genus Pythium.', *Mycological research*. Elsevier, 108(Pt 12), pp. 1363–1383. doi: 10.1017/s0953756204001431.
Lewin, H. A. *et al.* (2018) 'Earth BioGenome Project: Sequencing life for the future of life', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(17), pp. 4325–4333. doi: 10.1073/pnas.1720115115.
Lex, A. *et al.* (2014) 'UpSet: Visualization of intersecting sets', *IEEE Transactions on Visualization and Computer Graphics*, 20(12), pp. 1983–1992. doi: 10.1109/TVCG.2014.2346248.
Leynaud-Kieffer, L. M. C. *et al.* (2019) 'A new approach to Cas9-based genome editing in Aspergillus niger that is precise, efficient and selectable', *PLoS ONE*. Edited by K.-H. Han. Public Library of Science, 14(1), p. e0210243. doi: 10.1371/journal.pone.0210243.
Li, F.-W. and Harkess, A. (2018) 'A guide to sequence your favorite plant genomes', *Applications in Plant Sciences*. John Wiley & Sons, Ltd, 6(3), p. e1030. doi: 10.1002/aps3.1030.
Li, L., Stoeckert, C. J. and Roos, D. S. (2003) 'OrthoMCL: Identification of ortholog groups for eukaryotic genomes', *Genome Research*. Cold Spring Harbor Laboratory Press, 13(9), pp. 2178–2189. doi: 10.1101/gr.1224503.
Li, W. *et al.* (2010) 'Oomycetes and fungi: Important parasites on marine algae', *Acta Oceanologica Sinica*. The Chinese Society of Oceanography, 29(5), pp. 74–81. doi: 10.1007/s13131-010-0065-4.
Li, Y. H. Y. F. *et al.* (2014) 'De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits', *Nature Biotechnology*. Nature Publishing Group, 32(10), pp. 1045–1052. doi: 10.1038/nbt.2979.

Li, Z. *et al.* (2014) 'The 3,000 rice genomes project', *GigaScience*. BioMed Central, 3(1), p. 7. doi: 10.1186/2047-217X-3-7.
Liggenstoffer, A. S. *et al.* (2010) 'Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores', *ISME Journal*. Nature Publishing Group, 4(10), pp. 1225–1235. doi: 10.1038/ismej.2010.49.
Lin, J. and Gerstein, M. (2000) 'Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels', *Genome Research*. Cold Spring Harbor Laboratory Press, 10(6), pp. 808–818. doi: 10.1101/gr.10.6.808.
Lind, A. L. *et al.* (2017) 'Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species', *PLoS Biology*. Public Library of Science, 15(11), p. e2003583. doi: 10.1371/journal.pbio.2003583.
Lind, A. L. *et al.* (2018) 'An LaeA- and BrlA-Dependent Cellular Network Governs Tissue-Specific Secondary Metabolism in the Human Pathogen Aspergillus fumigatus', *mSphere*. American Society for Microbiology Journals, 3(2), pp. e00050-18. doi: 10.1128/mSphere.00050-18.
Links, M. G. *et al.* (2011) 'De novo sequence assembly of Albugo candida reveals a small genome relative to other biotrophic oomycetes', *BMC Genomics*. BioMed Central, 12(1), pp. 1–12. doi: 10.1186/1471-2164-12-503.
Liolios, K. *et al.* (2009) 'The Genomes On Line Database (GOLD) in 2009: Status of genomic and metagenomic projects and their associated metadata', *Nucleic Acids Research*. Narnia, 38(SUPPL.1), pp. D346–D354. doi: 10.1093/nar/gkp848.
Liti, G. *et al.* (2009) 'Population genomics of domestic and wild yeasts', *Nature*. Europe PMC Funders, 458(7236), pp. 337–341. doi: 10.1038/nature07743.
Liu, L. *et al.* (2009) 'Coalescent methods for estimating phylogenetic trees', *Molecular Phylogenetics and Evolution*. Academic Press, 53(1), pp. 320–328. doi: 10.1016/J.YMPEV.2009.05.033.
Liu, L. and Alper, H. S. (2014) 'Draft Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica PO1f, a Commonly Used Metabolic Engineering Host', *Genome Announcements*. American Society for Microbiology (ASM), 2(4). doi: 10.1128/genomea.00652-14.
Liu, Y. *et al.* (2009) 'Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support', *BMC Evolutionary Biology*. BioMed Central, 9(1), p. 272. doi: 10.1186/1471-2148-9-272.
Lockhart, S. R. *et al.* (2002) 'In Candida albicans, white-opaque switchers are homozygous for mating type', *Genetics*. Genetics Society of America, 162(2), pp. 737–745. doi: 10.1128/EC.00041-07.
Loman, N. J. and Pallen, M. J. (2015) 'Twenty years of bacterial genome sequencing', *Nature Reviews Microbiology*. Nature Publishing Group, 13(12), pp. 787–794. doi: 10.1038/nrmicro3565.
van Loo, B. *et al.* (2006) 'Diversity and biocatalytic potential of epoxide hydrolases identified by genome analysis.', *Applied and environmental microbiology*, 72(4), pp.

2905–17. doi: 10.1128/AEM.72.4.2905-2917.2006.
Lücking, R. *et al.* (2009) ‘Fungi evolved right on track’, *Mycologia*. Taylor & Francis, 101(6), pp. 810–822. doi: 10.3852/09-016.
Lynch, M. and Conery, J. S. (2000) ‘The evolutionary fate and consequences of duplicate genes’, *Science*. American Association for the Advancement of Science, 290(5494), pp. 1151–1155. doi: 10.1126/science.290.5494.1151.
Ma, L. J. *et al.* (2009) ‘Genomic analysis of the basal lineage fungus Rhizopus oryzae reveals a whole-genome duplication’, *PLoS Genetics*. Edited by H. D. Madhani. Public Library of Science, 5(7), p. e1000549. doi: 10.1371/journal.pgen.1000549.
Maciaszczyk, E. *et al.* (2004) ‘Arsenical resistance genes in Saccharomyces douglasii and other yeast species undergo rapid evolution involving genomic rearrangements and duplications’, *FEMS Yeast Research*. Oxford University Press, 4(8), pp. 821–832. doi: 10.1016/j.femsyr.2004.03.002.
Mackie, R. I. *et al.* (2004) ‘Biochemical and microbiological evidence for fermentative digestion in free-living land iguanas (Conolophus pallidus) and marine iguanas (Amblyrhynchus cristatus) on the Galápagos archipelago’, *Physiological and Biochemical Zoology*. The University of Chicago Press, 77(1), pp. 127–138. doi: 10.1086/383498.
Magnan, C. *et al.* (2016) ‘Sequence assembly of Yarrowia lipolytica strain W29/CLIB89 shows transposable element diversity’, *PLoS ONE*. Edited by J. Schacherer. Public Library of Science, 11(9), p. e0162363. doi: 10.1371/journal.pone.0162363.
Marcet-Houben, M. and Gabaldón, T. (2009) ‘The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome.’, *PLoS ONE*. Edited by C. d’Enfert. Public Library of Science, 4(2), p. e4357. doi: 10.1371/journal.pone.0004357.
Marcet-Houben, M. and Gabaldón, T. (2010) ‘Acquisition of prokaryotic genes by fungal genomes.’, *Trends in genetics : TIG*, 26(1), pp. 5–8. doi: 10.1016/j.tig.2009.11.007.
Marcet-Houben, M. and Gabaldón, T. (2015) ‘Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker’s yeast lineage’, *PLoS Biology*. Edited by L. D. Hurst. Public Library of Science, 13(8), p. e1002220. doi: 10.1371/journal.pbio.1002220.
Marcet-Houben, M., Marceddu, G. and Gabaldón, T. (2009) ‘Phylogenomics of the oxidative phosphorylation in fungi reveals extensive gene duplication followed by functional divergence.’, *BMC evolutionary biology*, 9(1), p. 295. doi: 10.1186/1471-2148-9-295.
Marcus, S., Lee, H. and Schatz, M. C. (2014) ‘SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips’, *Bioinformatics*. Oxford University Press, 30(24), pp. 3476–3483. doi: 10.1093/bioinformatics/btu756.
Marsit, S. and Dequin, S. (2015) ‘Diversity and adaptive evolution of Saccharomyces wine yeast: a review’, *FEMS Yeast Research*. Edited by J. Nielsen. Narnia, 15(7), p. fov067. doi: 10.1093/femsyr/fov067.

Martens, C. and Van de Peer, Y. (2010) 'The hidden duplication past of the plant pathogen Phytophthora and its consequences for infection.', *BMC genomics*. BioMed Central, 11(1), p. 353. doi: 10.1186/1471-2164-11-353.
Martens, C., Vandepoele, K. and Van de Peer, Y. (2008) 'Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species', *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp. 3427–3432. doi: 10.1073/pnas.0712248105.
Martin, F. N., Blair, J. E. and Coffey, M. D. (2014) 'A combined mitochondrial and nuclear multilocus phylogeny of the genus Phytophthora', *Fungal Genetics and Biology*, 66, pp. 19–32. doi: 10.1016/j.fgb.2014.02.006.
Martin, F. N. and Tooley, P. W. (2003) 'Phylogenetic relationships of Phytophthora ramorum, P. nemorosa, and P. pseudosyringae, three species recovered from areas in California with sudden oak death.', *Mycological research*. Mycological Society of America, 107(Pt 12), pp. 1379–1391. doi: 10.1017/S0953756203008785.
Martin, W. F. (2017) 'Too Much Eukaryote LGT', *BioEssays*, p. 1700115. doi: 10.1002/bies.201700115.
Matari, N. H. and Blair, J. E. (2014) 'A multilocus timescale for oomycete evolution estimated under three distinct molecular clock models', *BMC Evolutionary Biology*. BioMed Central, 14(1), p. 101. doi: 10.1186/1471-2148-14-101.
Matsuzaki, M. *et al.* (2004) 'Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D', *Nature*, 428(6983), pp. 653–657. doi: 10.1038/nature02398.
Matta, C. (2010) 'Spontaneous Generation and Disease Causation: Anton de Bary's Experiments with Phytophthora infestans and Late Blight of Potato', *Journal of the History of Biology*. Springer Netherlands, 43(3), pp. 459–491. doi: 10.1007/s10739-009-9220-1.
McCarthy, C. G. P. and Fitzpatrick, D. A. (2016) 'Systematic Search for Evidence of Interdomain Horizontal Gene Transfer from Prokaryotes to Oomycete Lineages', *mSphere*. Edited by A. P. Mitchell, 1(5), pp. e00195-16. doi: 10.1128/mSphere.00195-16.
McCarthy, C. G. P. and Fitzpatrick, D. A. (2017a) 'Multiple Approaches to Phylogenomic Reconstruction of the Fungal Kingdom', in *Advances in Genetics*, pp. 211–266. doi: 10.1016/bs.adgen.2017.09.006.
McCarthy, C. G. P. and Fitzpatrick, D. A. (2017b) 'Phylogenomic Reconstruction of the Oomycete Phylogeny Derived from 37 Genomes', *mSphere*. Edited by A. P. Mitchell, 2(2), pp. e00095-17. doi: 10.1128/mSphere.00095-17.
McCarthy, C. G. P. and Fitzpatrick, D. A. (2019a) 'Pan-genome analyses of model fungal species', *Microbial Genomics*, 5(2), pp. 1–23. doi: 10.1099/mgen.0.000243.
McCarthy, C. G. P. and Fitzpatrick, D. A. (2019b) 'Pangloss: A Tool for Pan-Genome Analysis of Microbial Eukaryotes', *Genes*. Multidisciplinary Digital Publishing Institute, 10(7), p. 521. doi: 10.3390/genes10070521.
McDonagh, A. *et al.* (2008) 'Sub-telomere directed gene expression during initiation of invasive aspergillosis', *PLoS Pathogens*. Edited by B. P. Cormack. Public Library of

Science, 4(9), p. e1000154. doi: 10.1371/journal.ppat.1000154.
McGowan, J., Byrne, K. P. and Fitzpatrick, D. A. (2019) 'Comparative analysis of oomycete genome evolution using the Oomycete gene order browser (OGOB)', *Genome Biology and Evolution*. Edited by S. Baldauf. Narnia, 11(1), pp. 189–206. doi: 10.1093/gbe/evy267.
McGowan, J. and Fitzpatrick, D. A. (2017) 'Genomic, Network, and Phylogenetic Analysis of the Oomycete Effector Arsenal', *mSphere*. American Society for Microbiology Journals, 2(6), pp. e00408-17. doi: 10.1128/msphere.00408-17.
McInerney, J. O. (1998) 'GCUA: general codon usage analysis', *Bioinformatics*, 14(4), pp. 372–373. doi: 10.1093/bioinformatics/14.4.372.
McInerney, J. O., McNally, A. and O'Connell, M. J. (2017) 'Why prokaryotes have pangenomes', *Nature Microbiology*. Nature Publishing Group, p. 17040. doi: 10.1038/nmicrobiol.2017.40.
McLaughlin, D. J. and Spatafora, J. W. (2014) *Systematics and evolution: Part A: Second edition*, *Systematics and Evolution: Part A: Second Edition*. doi: 10.1007/978-3-642-55318-9.
McMullan, M. *et al.* (2018) 'The ash dieback invasion of Europe was founded by two genetically divergent individuals', *Nature Ecology and Evolution*. Nature Publishing Group, 2(6), pp. 1000–1008. doi: 10.1038/s41559-018-0548-9.
McPherson, J. D. *et al.* (2001) 'A physical map of the human genome', *Nature*. Nature Publishing Group, 409(6822), pp. 934–941. doi: 10.1038/35057157.
Medina, E. M., Jones, G. W. and Fitzpatrick, D. A. (2011) 'Reconstructing the fungal tree of life using phylogenomics and a preliminary investigation of the distribution of yeast prion-like proteins in the fungal kingdom', *Journal of Molecular Evolution*. Springer-Verlag, 73(3–4), pp. 116–133. doi: 10.1007/s00239-011-9461-4.
Medini, D. *et al.* (2005) 'The microbial pan-genome', *Current Opinion in Genetics and Development*. Elsevier Current Trends, pp. 589–594. doi: 10.1016/j.gde.2005.09.006.
Mertens, S. *et al.* (2019) 'Reducing phenolic off-flavors through CRISPR-based gene editing of the FDC1 gene in Saccharomyces cerevisiae x Saccharomyces eubayanus hybrid lager beer yeasts', *PLoS ONE*. Edited by J. Schacherer. Public Library of Science, 14(1), p. e0209124. doi: 10.1371/journal.pone.0209124.
Mindell, D. P. (2013) 'The tree of life: Metaphor, model, and heuristic device', *Systematic Biology*. Oxford University Press, pp. 479–489. doi: 10.1093/sysbio/sys115.
Misner, I. *et al.* (2015) 'The secreted proteins of Achlya hypogyna and Thraustotheca clavata identify the ancestral oomycete secretome and reveal gene acquisitions by horizontal gene transfer.', *Genome biology and evolution*, 7(1), pp. 120–35. doi: 10.1093/gbe/evu276.
Mitchell, R. J. *et al.* (2014) 'Ash dieback in the UK: A review of the ecological and conservation implications and potential management options', *Biological Conservation*. Elsevier, 175, pp. 95–109. doi: 10.1016/j.biocon.2014.04.019.
Mlíčková, K. *et al.* (2004) 'Lipid accumulation, lipid body formation, and acyl coenzyme A oxidases of the yeast Yarrowia lipolytica', *Applied and Environmental Microbiology*. American Society for Microbiology, 70(7), pp. 3918–3924. doi:

10.1128/AEM.70.7.3918-3924.2004.
Moktali, V. *et al.* (2012) 'Systematic and searchable classification of cytochrome P450 proteins encoded by fungal and oomycete genomes', *BMC Genomics*, 13(1), p. 525. doi: 10.1186/1471-2164-13-525.
Möller, M. *et al.* (2018) 'Extraordinary genome instability and widespread chromosome rearrangements during vegetative growth', *Genetics*. Genetics Society of America, 210(2), pp. 517–529. doi: 10.1534/genetics.118.301050.
Montenegro, J. D. *et al.* (2017) 'The pangenome of hexaploid bread wheat', *Plant Journal*, 90(5), pp. 1007–1013. doi: 10.1111/tpj.13515.
Morales, L. and Dujon, B. (2012) 'Evolutionary Role of Interspecies Hybridization and Genetic Exchanges in Yeasts', *Microbiology and Molecular Biology Reviews*. American Society for Microbiology, 76(4), pp. 721–739. doi: 10.1128/mmbr.00022-12.
Moreira, D. *et al.* (2007) 'Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata', *Molecular Phylogenetics and Evolution*, 44(1), pp. 255–266. doi: 10.1016/j.ympev.2006.11.001.
Morin, E. *et al.* (2012) 'Genome sequence of the button mushroom Agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche', *Proceedings of the National Academy of Sciences of the United States of America*, 109(43), pp. 17501–17506. doi: 10.1073/pnas.1206847109.
Morris, J. L. *et al.* (2018) 'The timescale of early land plant evolution', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 115(10), p. 201719588. doi: 10.1073/pnas.1719588115.
Morris, P. F. *et al.* (2009) 'Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome.', *PloS one*. Public Library of Science, 4(7), p. e6133. doi: 10.1371/journal.pone.0006133.
Mosquera-Rendón, J. *et al.* (2016) 'Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species', *BMC Genomics*. BioMed Central, 17(1), p. 45. doi: 10.1186/s12864-016-2364-4.
Mullis, K. *et al.* (1986) 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.', *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1, pp. 263–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3472723 (Accessed: 17 June 2019).
Murphy, C. L. *et al.* (2019) 'Horizontal Gene Transfer as an Indispensable Driver for Evolution of Neocallimastigomycota into a Distinct Gut-Dwelling Fungal Lineage', *Applied and Environmental Microbiology*. American Society for Microbiology, 85(15), pp. e00988-19. doi: 10.1128/aem.00988-19.
Nalley, L. *et al.* (2016) 'Economic and environmental impact of rice blast pathogen (Magnaporthe oryzae) alleviation in the United States', *PLoS ONE*. Edited by Z. Wang. Public Library of Science, 11(12), p. e0167295. doi: 10.1371/journal.pone.0167295.
Naseeb, S. and Delneri, D. (2012) 'Impact of chromosomal inversions on the yeast dal cluster', *PLoS ONE*. Public Library of Science, 7(8), p. e42022. doi: 10.1371/journal.pone.0042022.

Nash, A. K. *et al.* (2017) 'The gut mycobiome of the Human Microbiome Project healthy cohort', *Microbiome*. BioMed Central, 5(1), p. 153. doi: 10.1186/s40168-017-0373-4.
Nicaud, J. M. (2012) 'Yarrowia lipolytica', *Yeast*. John Wiley & Sons, Ltd, 29(10), pp. 409–418. doi: 10.1002/yea.2921.
Nielsen, J. C. *et al.* (2017) 'Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species', *Nature Microbiology*. Nature Publishing Group, 2(6), p. 17044. doi: 10.1038/nmicrobiol.2017.44.
Nierman, W. C. *et al.* (2005) 'Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus', *Nature*, 438(7071), pp. 1151–1156. doi: 10.1038/nature04332.
Nikoh, N. *et al.* (1994) 'Phylogenetic relationship of the kingdoms Animalia, Plantae, and Fungi, inferred from 23 different protein species.', *Molecular biology and evolution*, 11(5), pp. 762–768. doi: 0737-4038/94f I lOS-0005$02.
Nosenko, T. and Bhattacharya, D. (2007) 'Horizontal gene transfer in chromalveolates.', *BMC evolutionary biology*. BioMed Central, 7(1), p. 173. doi: 10.1186/1471-2148-7-173.
Nout, M. J. R. and Aidoo, K. E. (2002) 'Asian Fungal Fermented Food', in *Industrial Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 23–47. doi: 10.1007/978-3-662-10378-4_2.
Novo, M. *et al.* (2009) 'Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 106(38), pp. 16333–16338. doi: 10.1073/pnas.0904673106.
O'Brien, C. E. *et al.* (2018) 'Genome analysis of the yeast Diutina catenulata, a member of the Debaryomycetaceae/Metschnikowiaceae (CTG-Ser) clade', *PLoS ONE*, 13(6). doi: 10.1371/journal.pone.0198957.
O'Connor, E. *et al.* (2019) ' Whole Genome Sequence of the Commercially Relevant Mushroom Strain Agaricus bisporus var. bisporus ARP23 ', *G3: Genes|Genomes|Genetics*. doi: 10.1534/g3.119.400563.
O'Grada, C. (1999) *Black '47 and beyond: the great Irish famine in history, economy, and memory*. Princeton University Press.
O'Grada, C. (2006) *Ireland's great famine: interdisciplinary perspectives*. University College Dublin Press.
O'Hanlon, R., McCracken, A. R. and Cooke, L. R. (2016) 'Diversity and ecology of Phytophthora species on the Island of Ireland', *Biology and Environment*, 116B(1), pp. 27–51. doi: 10.3318/BIOE.2016.03.
Obenchain, V. *et al.* (2015) 'Orchestrating high-throughput genomic analysis with Bioconductor', *Nature Methods*. Nature Publishing Group, 12(2), pp. 115–121. doi: 10.1038/nmeth.3252.
Oberwinkler, F. (2017) 'Yeasts in Pucciniomycotina', *Mycological Progress*. Springer Berlin Heidelberg, 16(9), pp. 831–856. doi: 10.1007/s11557-017-1327-8.

Odds, F. C., Brown, A. J. P. and Gow, N. A. R. (2004) 'Candida albicans genome sequence: A platform for genomics in the absence of genetics', *Genome Biology*. BioMed Central, 5(7), p. 230. doi: 10.1186/gb-2004-5-7-230.
Oliver, S. G. *et al.* (1992) 'The complete DNA sequence of yeast chromosome III', *Nature*. Nature Publishing Group, 357(6373), pp. 38–46. doi: 10.1038/357038a0.
Ollis, D. L. *et al.* (1992) 'The alpha/beta hydrolase fold.', *Protein engineering*, 5(3), pp. 197–211. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1409539 (Accessed: 16 May 2016).
Opoku, I. *et al.* (2011) 'Phytophthora megakarya: A potential threat to the cocoa industry in Ghana', *Ghana Journal of Agricultural Science*. Accra: Council for Scientific and Industrial Research, Ghana, 33(2), pp. 237–248. doi: 10.4314/gjas.v33i2.1876.
Pace, N. R., Sapp, J. and Goldenfeld, N. (2012) 'Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 109(4), pp. 1011–1018. doi: 10.1073/pnas.1109716109.
Page, A. J. *et al.* (2015) 'Roary: Rapid large-scale prokaryote pan genome analysis', *Bioinformatics*. Oxford University Press, 31(22), pp. 3691–3693. doi: 10.1093/bioinformatics/btv421.
Page, R. D. M. and Holmes, E. C. (1998) *Molecular evolution: A phylogenetic approach*. Oxford, UK: Blackwell Science.
Palmer, G. E., Askew, D. S. and Williamson, P. R. (2008) 'The diverse roles of autophagy in medically important fungi', *Autophagy*. Taylor & Francis, 4(8), pp. 982–988. doi: 10.4161/auto.7075.
Parfrey, L. W., Lahr, D. J. G. and Katz, L. A. (2008) 'The Dynamic Nature of Eukaryotic Genomes', *Molecular Biology and Evolution*. Oxford University Press, 25(4), pp. 787–794. doi: 10.1093/molbev/msn032.
Parker, B. C., Preston, R. D. and Fogg, G. E. (1963) 'Studies of the structure and chemical composition of the cell walls of Vaucheriaceae and Saprolegniaceae', *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 158(973), pp. 435–445. doi: 10.1098/rspb.1963.0056.
Parra, G., Bradnam, K. and Korf, I. (2007) 'CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes', *Bioinformatics*. Oxford University Press, 23(9), pp. 1061–1067. doi: 10.1093/bioinformatics/btm071.
Pellicer, J. *et al.* (2018) 'Genome Size Diversity and Its Impact on the Evolution of Land Plants.', *Genes*. Multidisciplinary Digital Publishing Institute (MDPI), 9(2). doi: 10.3390/genes9020088.
Pellicer, J., Fay, M. F. and Leitch, I. J. (2010) 'The largest eukaryotic genome of them all?', *Botanical Journal of the Linnean Society*. Narnia, 164(1), pp. 10–15. doi: 10.1111/j.1095-8339.2010.01072.x.
Pennisi, E. (2001) 'The push to pit genomics against fungal pathogens', *Science*, 292(5525), pp. 2273–2274. doi: 10.1126/science.292.5525.2273.
Pennisi, E. (2002) 'SEQUENCING: Chimps and Fungi Make Genome', *Science*.

American Association for the Advancement of Science, 296(5573), pp. 1589b – 1591. doi: 10.1126/science.296.5573.1589b.
Perez-Nadales, E. *et al.* (2014) 'Fungal model systems and the elucidation of pathogenicity determinants', *Fungal Genetics and Biology*. Elsevier, pp. 42–67. doi: 10.1016/j.fgb.2014.06.011.
Peter, J. *et al.* (2018) 'Genome evolution across 1,011 Saccharomyces cerevisiae isolates', *Nature*. Nature Publishing Group, 556(7701), pp. 339–344. doi: 10.1038/s41586-018-0030-5.
Peter, J. and Schacherer, J. (2016) 'Population genomics of yeasts: Towards a comprehensive view across a broad evolutionary scale', *Yeast*. John Wiley & Sons, Ltd, 33(3), pp. 73–81. doi: 10.1002/yea.3142.
Petersen, T. N. *et al.* (2011) 'SignalP 4.0: discriminating signal peptides from transmembrane regions.', *Nature methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 8(10), pp. 785–6. doi: 10.1038/nmeth.1701.
Pieretti, I. *et al.* (2015) 'What makes Xanthomonas albilineans unique amongst xanthomonads?', *Frontiers in plant science*, 6, p. 289. doi: 10.3389/fpls.2015.00289.
Pisani, D., Cotton, J. A. and McInerney, J. O. (2007) 'Supertrees disentangle the chimerical origin of eukaryotic genomes', *Molecular Biology and Evolution*. Oxford University Press, 24(8), pp. 1752–1760. doi: 10.1093/molbev/msm095.
Pisani, D. and Wilkinson, M. (2002) 'Matrix representation with parsimony, taxonomic congruence, and total evidence', *Systematic Biology*. [Oxford University Press, Society of Systematic Biologists], 51(1), pp. 151–155. doi: 10.1080/106351502753475925.
Plissonneau, C., Hartmann, F. E. and Croll, D. (2018) 'Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome', *BMC Biology*. BioMed Central, 16(1), p. 5. doi: 10.1186/s12915-017-0457-4.
Ploetz, R. (2016) 'The impact of diseases on cacao production: A global overview', in *Cacao Diseases: A History of Old Enemies and New Encounters*. Cham: Springer International Publishing, pp. 33–59. doi: 10.1007/978-3-319-24789-2_2.
Porter, T. M. *et al.* (2011) 'Molecular phylogeny of the Blastocladiomycota (Fungi) based on nuclear ribosomal DNA', *Fungal Biology*. Elsevier, 115(4–5), pp. 381–392. doi: 10.1016/j.funbio.2011.02.004.
Price, M. N., Dehal, P. S. and Arkin, A. P. (2010) 'FastTree 2 - Approximately maximum-likelihood trees for large alignments', *PLoS ONE*. Edited by A. F. Y. Poon. Public Library of Science, 5(3), p. e9490. doi: 10.1371/journal.pone.0009490.
Pride, D. T. *et al.* (2003) 'Evolutionary implications of microbial genome tetranucleotide frequency biases', *Genome Research*. Cold Spring Harbor Laboratory Press, 13(2), pp. 145–156. doi: 10.1101/gr.335003.
Qi, J., Luo, H. and Hao, B. (2004) 'CVTree: A phylogenetic tree reconstruction tool based on whole genomes', *Nucleic Acids Research*, 32(WEB SERVER ISS.), pp. W45-7. doi: 10.1093/nar/gkh362.
Qi, J., Wang, B. and Hao, B. I. (2004) 'Whole Proteome Prokaryote Phylogeny Without

Sequence Alignment: A K-String Composition Approach', *Journal of Molecular Evolution*. Springer-Verlag, 58(1), pp. 1–11. doi: 10.1007/s00239-003-2493-7.
Qiao, K. *et al.* (2017) 'Lipid production in Yarrowia lipolytica is maximized by engineering cytosolic redox metabolism', *Nature Biotechnology*. Nature Publishing Group, 35(2), pp. 173–177. doi: 10.1038/nbt.3763.
Qiu, H. *et al.* (2016) 'Extensive horizontal gene transfers between plant pathogenic fungi', *BMC Biology*. BioMed Central, 14(1), p. 41. doi: 10.1186/s12915-016-0264-3.
de Queiroz, A. and Gatesy, J. (2007) 'The supermatrix approach to systematics', *Trends in Ecology and Evolution*, 22(1), pp. 34–41. doi: 10.1016/j.tree.2006.10.002.
Qutob, D. *et al.* (2000) 'Comparative analysis of expressed sequences in Phytophthora sojae.', *Plant physiology*, 123(1), pp. 243–54. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=58998&tool=pmcentrez&rendertype=abstract (Accessed: 18 May 2016).
R Core Team and R Development Core Team (2013) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria, Austria. doi: 10.1007/978-3-540-74686-7.
Ragan, M. A. (1992) 'Phylogenetic inference based on matrix representation of trees', *Molecular Phylogenetics and Evolution*. Academic Press, 1(1), pp. 53–58. doi: 10.1016/1055-7903(92)90035-F.
Rahman, a K. M. S. *et al.* (2003) 'A role of xylanase, alpha-L-arabinofuranosidase, and xylosidase in xylan degradation.', *Canadian journal of microbiology*, 49(1), pp. 58–64. doi: 10.1139/w02-114.
Ramsay, L. *et al.* (2000) 'A simple sequence repeat-based linkage map of Barley', *Genetics*, 156(4), pp. 1997–2005. doi: 10.1093/nar/25.17.3389.
Rasko, D. A. *et al.* (2008) 'The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates', *Journal of Bacteriology*, 190(20), pp. 6881–6893. doi: 10.1128/JB.00619-08.
Rasko, D. A., Myers, G. S. A. and Ravel, J. (2005) 'Visualization of comparative genomic analyses by BLAST score ratio', *BMC Bioinformatics*, 6(1), p. 2. doi: 10.1186/1471-2105-6-2.
Read, B. A. *et al.* (2013) 'Pan genome of the phytoplankton Emiliania underpins its global distribution', *Nature*. Nature Publishing Group, 499(7457), pp. 209–213. doi: 10.1038/nature12221.
Redecker, D. (2000) 'Glomalean Fungi from the Ordovician', *Science*, 289(5486), pp. 1920–1921. doi: 10.1126/science.289.5486.1920.
Reno, M. L. *et al.* (2009) 'Biogeography of the Sulfolobus islandicus pan-genome', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 106(21), pp. 8605–8610. doi: 10.1073/pnas.0808945106.
Reynolds, H. T. *et al.* (2018) 'Horizontal gene cluster transfer increased hallucinogenic mushroom diversity', *Evolution Letters*. John Wiley & Sons, Ltd, 2(2), pp. 88–101. doi: 10.1002/evl3.42.
Ribeiro, O. K. (2013) 'A historical perspective of Phytophthora.', in *Phytophthora: a global perspective.* Wallingford: CABI, pp. 1–10. doi: 10.1079/9781780640938.0001.
Rice, D. W. and Palmer, J. D. (2006) 'An exceptional horizontal gene transfer in

plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters.', *BMC biology*. BioMed Central, 4, p. 31. doi: 10.1186/1741-7007-4-31.
Rice, P., Longden, I. and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics*, 16(6), pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.
Richards, S. (2015) 'It's more than stamp collecting: how genome sequencing can unify biological research.', *Trends in genetics : TIG*. NIH Public Access, 31(7), pp. 411–21. doi: 10.1016/j.tig.2015.04.007.
Richards, T. A. *et al.* (2006) 'Evolution of Filamentous Plant Pathogens: Gene Exchange across Eukaryotic Kingdoms', *Current Biology*, 16(18), pp. 1857–1864. doi: 10.1016/j.cub.2006.07.052.
Richards, T. A. *et al.* (2011) 'Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 108(37), pp. 15258–15263. doi: 10.1073/pnas.1105100108.
Riethmüller, A. *et al.* (2002) 'Phylogenetic relationships of the downy mildews (Peronosporales) and related groups based on nuclear large subunit ribosomal DNA sequences', *Mycologia*. Mycological Society of America, 94(5), pp. 834–849. doi: 10.2307/3761698.
Riisberg, I. *et al.* (2009) 'Seven Gene Phylogeny of Heterokonts', *Protist*, 160(2), pp. 191–204. doi: 10.1016/j.protis.2008.11.004.
Riley, T. T. *et al.* (2016) 'Breaking the Mold: A Review of Mucormycosis and Current Pharmacological Treatment Options', *Annals of Pharmacotherapy*, 50(9), pp. 747–757. doi: 10.1177/1060028016655425.
Rivera, Z. S., Losada, L. and Nierman, W. C. (2012) 'Back to the future for dermatophyte genomics', *mBio*. American Society for Microbiology, 3(6), pp. e00381-12. doi: 10.1128/mBio.00381-12.
Rizzo, D. M. *et al.* (2002) 'Phytophthora ramorum as the cause of extensive mortality of Quercus spp. and Lithocarpus densiflorus in California', *Plant Disease*, 86(3), pp. 205–214. doi: 10.1094/PDIS.2002.86.3.205.
Rizzo, D. M., Garbelotto, M. and Hansen, E. M. (2005) 'Phytophthora ramorum: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests', *Annual Review of Phytopathology*. Annual Reviews , 43(1), pp. 309–335. doi: 10.1146/annurev.phyto.42.040803.140418.
Robbertse, B. *et al.* (2006) 'A phylogenomic analysis of the Ascomycota', *Fungal Genetics and Biology*, 43(10), pp. 715–725. doi: 10.1016/j.fgb.2006.05.001.
Robideau, G. P., Rodrigue, N. and André Lévesque, C. (2014) 'Codon-based phylogenetics introduces novel flagellar gene markers to oomycete systematics', *Molecular Phylogenetics and Evolution*, 79(1), pp. 279–291. doi: 10.1016/j.ympev.2014.04.009.
Rokas, A. *et al.* (2003) 'Genome-scale approaches to resolving incongruence in molecular phylogenies.', *Nature*. Nature Publishing Group, 425(6960), pp. 798–804.

doi: 10.1038/nature02053.
Ronald, J., Tang, H. and Brem, R. B. (2006) 'Genomewide evolutionary rates in laboratory and wild yeast', *Genetics*, 174(1), pp. 541–544. doi: 10.1534/genetics.106.060863.
Van Rooij, P. *et al.* (2015) 'Amphibian chytridiomycosis: a review with focus on fungus-host interactions.', *Veterinary research*. BioMed Central, 46, p. 137. doi: 10.1186/s13567-015-0266-0.
Rouli, L. *et al.* (2015) 'The bacterial pangenome as a new tool for analysing pathogenic bacteria', *New Microbes and New Infections*. Elsevier, 7, pp. 72–85. doi: 10.1016/j.nmni.2015.06.005.
Runge, F. *et al.* (2011) 'The inclusion of downy mildews in a multi-locus-dataset and its reanalysis reveals a high degree of paraphyly in <I>Phytophthora</I>', *IMA Fungus*. CBS Fungal Biodiversity Centre, 2(2), pp. 163–171. doi: 10.5598/imafungus.2011.02.02.07.
Sahl, J. W. *et al.* (2014) 'The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes', *PeerJ*. PeerJ Inc., 2, p. e332. doi: 10.7717/peerj.332.
Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 74(12), pp. 5463–7. doi: 10.1073/pnas.74.12.5463.
Santos, M. A. S. and Tuite, M. F. (1995) 'The CUG codon is decoded in vivo as serine and not leucine in Candida albicans', *Nucleic Acids Research*, 23(9), pp. 1481–1486. doi: 10.1093/nar/23.9.1481.
Saunders, C. W., Scheynius, A. and Heitman, J. (2012) 'Malassezia fungi are specialized to live on skin and associated with dandruff, eczema, and other skin diseases', *PLoS Pathogens*. Public Library of Science, 8(6), p. e1002701. doi: 10.1371/journal.ppat.1002701.
Saville, A. C., Martin, M. D. and Ristaino, J. B. (2016) 'Historic late blight outbreaks caused by a widespread dominant lineage of Phytophthora infestans (Mont.) de Bary', *PLoS ONE*. Edited by M. Gijzen. Public Library of Science. doi: 10.1371/journal.pone.0168381.
Savory, F., Leonard, G. and Richards, T. A. (2015) 'The role of horizontal gene transfer in the evolution of the oomycetes.', *PLoS pathogens*. Public Library of Science, 11(5), p. e1004805. doi: 10.1371/journal.ppat.1004805.
Sboner, A. *et al.* (2011) 'The real cost of sequencing: Higher than you think!', *Genome Biology*. BioMed Central, 12(8), p. 125. doi: 10.1186/gb-2011-12-8-125.
Schmidt, K. H. *et al.* (2010) 'Formation of complex and unstable chromosomal translocations in yeast', *PLoS ONE*. Edited by A.-K. Bielinsky. Public Library of Science, 5(8), p. e12007. doi: 10.1371/journal.pone.0012007.
Schoch, C. L. *et al.* (2009) 'The ascomycota tree of life: A phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits', *Systematic Biology*. Oxford University Press, 58(2), pp. 224–239. doi:

10.1093/sysbio/syp020.
Schwartz, R. M. and Dayhoff, M. O. (1978) 'Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 199(4327), pp. 395–403. doi: 10.1126/SCIENCE.202030.
Seemann, T. (2014) 'Prokka: Rapid prokaryotic genome annotation', *Bioinformatics*, 30(14), pp. 2068–2069. doi: 10.1093/bioinformatics/btu153.
Seidl, M. F. *et al.* (2012) 'Reconstruction of Oomycete Genome Evolution Identifies Differences in Evolutionary Trajectories Leading to Present-Day Large Gene Families', *Genome Biology and Evolution*. Oxford University Press, 4(3), pp. 199–211. doi: 10.1093/gbe/evs003.
Sekimoto, S. *et al.* (2008) 'Taxonomy, molecular phylogeny, and ultrastructural morphology of Olpidiopsis porphyrae sp. nov. (Oomycetes, straminipiles), a unicellular obligate endoparasite of Bangia and Porphyra spp. (Bangiales, Rhodophyta)', *Mycological Research*, 112(3), pp. 361–374. doi: 10.1016/j.mycres.2007.11.002.
Sello, M. M. *et al.* (2015) 'Diversity and evolution of cytochrome P450 monooxygenases in Oomycetes.', *Scientific reports*, 5, p. 11572. doi: 10.1038/srep11572.
Shalchian-Tabrizi, K. *et al.* (2007) 'Analysis of Environmental 18S Ribosomal RNA Sequences reveals Unknown Diversity of the Cosmopolitan Phylum Telonemia', *Protist*, 158(2), pp. 173–180. doi: 10.1016/j.protis.2006.10.003.
Shen, X.-X. *et al.* (2016) 'Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data', *G3: Genes|Genomes|Genetics*. Genetics Society of America, 6(12), pp. 3927–3939. doi: 10.1534/g3.116.034744.
Shen, X. X. *et al.* (2018) 'Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum', *Cell*. Cell Press, 175(6), pp. 1533-1545.e20. doi: 10.1016/j.cell.2018.10.023.
Shi, T. Q. *et al.* (2018) 'Advancing metabolic engineering of Yarrowia lipolytica using the CRISPR/Cas system', *Applied Microbiology and Biotechnology*. Springer Berlin Heidelberg, 102(22), pp. 9541–9548. doi: 10.1007/s00253-018-9366-x.
Sigalova, O. *et al.* (2018) 'Chlamydia pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction.', *bioRxiv*. Cold Spring Harbor Laboratory, p. 506121. doi: 10.1101/506121.
Simão, F. A. *et al.* (2015) 'BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.
Singh, R. P. *et al.* (2008) 'Will Stem Rust Destroy the World's Wheat Crop?', *Advances in Agronomy*. Academic Press, 98, pp. 271–309. doi: 10.1016/S0065-2113(08)00205-8.
Singh, R. P. *et al.* (2011) 'The Emergence of Ug99 Races of the Stem Rust Fungus is a Threat to World Wheat Production', *Annual Review of Phytopathology*, 49(1), pp. 465–481. doi: 10.1146/annurev-phyto-072910-095423.
Sipos, G. *et al.* (2017) 'Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria', *Nature Ecology and Evolution*. Nature Publishing Group, 1(12), pp. 1931–1941. doi: 10.1038/s41559-017-0347-8.

Skrzypek, M. S. *et al.* (2017) 'The Candida Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D592–D596. doi: 10.1093/nar/gkw924.
Slater, G. S. C. and Birney, E. (2005) 'Automated generation of heuristics for biological sequence comparison', *BMC Bioinformatics*. BioMed Central, 6(1), p. 31. doi: 10.1186/1471-2105-6-31.
Slot, J. C. and Rokas, A. (2010) 'Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 107(22), pp. 10136–10141. doi: 10.1073/pnas.0914418107.
Slowinski, J. B. and Page, R. D. (1999) 'How should species phylogenies be inferred from sequence data?', *Systematic biology*. Oxford University Press, 48(4), pp. 814–825. doi: 10.1080/106351599260030.
Snel, B., Bork, P. and Huynen, M. A. (1999) 'Genome phylogeny based on gene content', *Nature Genetics*. Nature Publishing Group, 21(1), pp. 108–110. doi: 10.1038/5052.
Snel, B., Huynen, M. A. and Dutilh, B. E. (2005) 'Genome Trees and the Nature of Genome Evolution', *Annual Review of Microbiology*, 59(1), pp. 191–209. doi: 10.1146/annurev.micro.59.030804.121233.
Snipen, L., Almøy, T. and Ussery, D. W. (2009) 'Microbial comparative pan-genomics using binomial mixture models', *BMC Genomics*. BioMed Central, 10(1), p. 385. doi: 10.1186/1471-2164-10-385.
Snipen, L. and Liland, K. H. (2015) 'micropan: An R-package for microbial pan-genomics', *BMC Bioinformatics*. BioMed Central, 16(1), pp. 1–8. doi: 10.1186/s12859-015-0517-0.
Soanes, D. M., Richards, T. a and Talbot, N. J. (2007) 'Insights from Sequencing Fungal and Oomycete Genomes: What Can We Learn about Plant Disease and the Evolution of Pathogenicity?', *The Plant cell*, 19(11), pp. 3318–3326. doi: 10.1105/tpc.107.056663.
Sollars, E. S. A. and Buggs, R. J. A. (2018) 'Genome-wide epigenetic variation among ash trees differing in susceptibility to a fungal disease', *BMC Genomics*. BioMed Central, 19(1), p. 502. doi: 10.1186/s12864-018-4874-8.
Song, G. *et al.* (2015) 'AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of Saccharomyces cerevisiae', *PLoS ONE*. Edited by J. Schacherer. Public Library of Science, 10(3), p. e0120671. doi: 10.1371/journal.pone.0120671.
Souciet, J. L. *et al.* (2000) 'Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies', *FEBS Letters*. John Wiley & Sons, Ltd, 487(1), pp. 3–12. doi: 10.1016/S0014-5793(00)02272-9.
Souciet, J. L. (2011) 'Ten years of the Génolevures Consortium: A brief history', *Comptes Rendus - Biologies*. Elsevier Masson, 334(8–9), pp. 580–584. doi: 10.1016/j.crvi.2011.05.005.
Soucy, S. M., Huang, J. and Gogarten, J. P. (2015) 'Horizontal gene transfer: building

the web of life', *Nature Reviews Genetics*. Nature Publishing Group, 16(8), pp. 472–482. doi: 10.1038/nrg3962.
Spatafora, J. *et al.* (2006) 'A five-gene phylogeny of Pezizomycotina', *Mycologia*, 98(6), pp. 1018–1028. doi: 10.3852/mycologia.98.6.1018.
Spatafora, J. W. *et al.* (2016) 'A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data.', *Mycologia*. Taylor & Francis, 108(5), pp. 1028–1046. doi: 10.3852/16-042.
Stajich, J. E. *et al.* (2009) 'The fungi.', *Current biology : CB*. NIH Public Access, 19(18), pp. R840-5. doi: 10.1016/j.cub.2009.07.004.
Stajich, J. E. (2017) 'Fungal Genomes and Insights into the Evolution of the Kingdom', in *The Fungal Kingdom*. NIH Public Access, pp. 619–633. doi: 10.1128/microbiolspec.funk-0055-2016.
Stajich, J. E., Dietrich, F. S. and Roy, S. W. (2007) 'Comparative genomic analysis of fungal genomes reveals intron-rich ancestors', *Genome Biology*. BioMed Central, 8(10), p. R223. doi: 10.1186/gb-2007-8-10-r223.
Stammers, D. K. *et al.* (2001) 'The structure of the negative transcriptional regulator NmrA reveals a structural superfamily which includes the short-chain dehydrogenase/reductases', *EMBO Journal*. EMBO Press, 20(23), pp. 6619–6626. doi: 10.1093/emboj/20.23.6619.
Stanke, M. *et al.* (2004) 'AUGUSTUS: A web server for gene finding in eukaryotes', *Nucleic Acids Research*. Oxford University Press, 32(WEB SERVER ISS.), pp. W309-12. doi: 10.1093/nar/gkh379.
Steel, M. and Rodrigo, A. (2008) 'Maximum likelihood supertrees.', *Systematic biology*. Oxford University Press, 57(2), pp. 243–250. doi: 10.1080/10635150802033014.
Steinberg, G. (2015) 'Cell biology of Zymoseptoria tritici: Pathogen cell organization and wheat infection', *Fungal Genetics and Biology*. Elsevier, 79, pp. 17–23. doi: 10.1016/j.fgb.2015.04.002.
Strohl, W. R. (2001) 'Biochemical engineering of natural product biosynthesis pathways.', *Metabolic engineering*. Academic Press, 3(1), pp. 4–14. doi: 10.1006/mben.2000.0172.
Strope, P. K. *et al.* (2015) 'The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen', *Genome Research*. Cold Spring Harbor Laboratory Press, 125(5), pp. 762–774. doi: 10.1101/gr.185538.114.
Su, Z. and Townsend, J. P. (2015) 'Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: analytical predictions of long-branch effects', *BMC Evolutionary Biology*, 15(86), p. 86. doi: 10.1186/s12862-015-0364-7.
Sultana, A. *et al.* (2004) 'Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation.', *The EMBO journal*. EMBO Press, 23(9), pp. 1911–21. doi: 10.1038/sj.emboj.7600201.
Swofford, L. D. (2002) *PAUP*: phylogenetic analysis using parsimony (* and other*

methods), *version 4.0 beta. Sunderland (MA): Sinauer.*
Szöllősi, G. J. *et al.* (2015) 'Genome-scale phylogenetic analysis finds extensive gene transfer among fungi', *Philosophical Transactions of the Royal Society B: Biological Sciences*. The Royal Society, 370(1678), p. 20140335. doi: 10.1098/rstb.2014.0355.
Tanabe, Y., Watanabe, M. M. and Sugiyama, J. (2005) 'Evolutionary relationships among basal fungi (Chytridiomycota and Zygomycota): Insights from molecular phylogenetics.', *The Journal of general and applied microbiology*, 51(5), pp. 267–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16314681 (Accessed: 17 September 2019).
Tavares, S. S. *et al.* (2014) 'Genome size analyses of Pucciniales reveal the largest fungal genomes', *Frontiers in Plant Science*. Frontiers, 5(AUG), p. 422. doi: 10.3389/fpls.2014.00422.
Taylor, F. J. R. (1978) 'Problems in the development of an explicit hypothetical phylogeny of the lower eukaryotes', *BioSystems*, 10(1–2), pp. 67–89. doi: 10.1016/0303-2647(78)90031-X.
Taylor, T. N., Krings, M. and Kerp, H. (2006) 'Hassiella monospora gen. et sp. nov., a microfungus from the 400 million year old Rhynie chert', *Mycological Research*. Elsevier, 110(6), pp. 628–632. doi: 10.1016/j.mycres.2006.02.009.
Teichmann, S. A. and Mitchison, G. (1999) 'Making family trees from gene families', *Nature Genetics*. Nature Publishing Group, 21(1), pp. 66–67. doi: 10.1038/5001.
Tekaia, F., Lazcano, A. and Dujon, B. (1999) 'The genomic tree as revealed from whole proteome comparisons', *Genome Research*. Cold Spring Harbor Laboratory Press, 9(6), pp. 550–557. doi: 10.1101/gr.9.6.550.
Ter-Hovhannisyan, V. *et al.* (2008) 'Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training', *Genome Research*, 18(12), pp. 1979–1990. doi: 10.1101/gr.081612.108.
Tettelin, H. *et al.* (2005) 'Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome"', *Proceedings of the National Academy of Sciences*, 102(39), pp. 13950–13955. doi: 10.1073/pnas.0506758102.
The C. elegans Sequencing Consortium (1998) 'Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology', *Science*, 282(5396), pp. 2012–2018. doi: 10.1126/science.282.5396.2012.
Thevenieau, F. *et al.* (2009) 'Uptake and Assimilation of Hydrophobic Substrates by the Oleaginous Yeast Yarrowia lipolytica', in *Handbook of Hydrocarbon and Lipid Microbiology*. Berlin, Heidelberg, Heidelberg: Springer Berlin Heidelberg, pp. 1513–1527. doi: 10.1007/978-3-540-77587-4_104.
Thines, M. (2014) 'Phylogeny and evolution of plant pathogenic oomycetes-a global overview', *European Journal of Plant Pathology*. Springer Netherlands, 138(3), pp. 431–447. doi: 10.1007/s10658-013-0366-5.
Thines, M. and Kamoun, S. (2010) 'Oomycete–plant coevolution: recent advances and future prospects', *Current Opinion in Plant Biology*, 13(4), pp. 427–433. doi: 10.1016/j.pbi.2010.04.001.

Thomas, J. E., Ou, L. T. and Al-Agely, A. (2008) 'DDE remediation and degradation', *Reviews of Environmental Contamination and Toxicology*, 194, pp. 55–69. doi: 10.1007/978-0-387-74816-0_3.
Thomas, P. D. *et al.* (2003) 'PANTHER: A library of protein families and subfamilies indexed by function', *Genome Research*. Cold Spring Harbor Lab, 13(9), pp. 2129–2141. doi: 10.1101/gr.772403.
Timmis, J. N. *et al.* (2004) 'Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes.', *Nature reviews. Genetics*. Nature Publishing Group, 5(2), pp. 123–35. doi: 10.1038/nrg1271.
Tomlinson, J. A., Dickinson, M. J. and Boonham, N. (2010) 'Rapid detection of phytophthora ramorum and P. kernoviae by two-minute DNA extraction followed by isothermal amplification and amplicon detection by generic lateral flow device', *Phytopathology*, 100(2), pp. 143–149. doi: 10.1094/PHYTO-100-2-0143.
Torruella, G. *et al.* (2015) 'Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi', *Current Biology*. Elsevier, 25(18), pp. 2404–2410. doi: 10.1016/j.cub.2015.07.053.
Treangen, T. J. and Rocha, E. P. C. (2011) 'Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes', *PLoS Genetics*. Edited by N. A. Moran. Public Library of Science, 7(1), p. e1001284. doi: 10.1371/journal.pgen.1001284.
Tsui, C. K. M. *et al.* (2009) 'Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding', *Molecular Phylogenetics and Evolution*, 50(1), pp. 129–140. doi: 10.1016/j.ympev.2008.09.027.
Tucker, C. (1931) *Taxonomy of the genus Phytophthora de Bary*. Columbia Mo.: University of Missouri College of Agriculture Agricultural Experiment Station. Available at: http://www.worldcat.org/title/taxonomy-of-the-genus-phytophthora-de-bary/oclc/19841623 (Accessed: 21 August 2018).
Turner, R. S. (2005) 'After the famine: Plant pathology, Phytophthora infestans, and the late blight of potatoes, 1845–1960', *Historical Studies in the Physical and Biological Sciences*, 35(2), pp. 341–370. doi: 10.1525/hsps.2005.35.2.341.
Tyler, B. M. *et al.* (2006) 'Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis', *Science*. American Association for the Advancement of Science, 313(5791), pp. 1261–1266. doi: 10.1126/science.1128796.
Uzuhashi, S., Tojo, M. and Kakishima, M. (2010) 'Phylogeny of the genus pythium and description of new genera', *Mycoscience*. Springer Japan, 51(5), pp. 337–365. doi: 10.1007/s10267-010-0046-7.
Vanhaute, E., Paping, R. and O'Grada, C. (2007) 'The European subsistence crisis of 1845-1850: a comparative perspective', in *When the potato failed. Causes and effects of the' last' European subsistence crisis, 1845-1850*, pp. 1–31. doi: http://dx.doi.org/10.1680/geot.2008.T.003.
Vassarotti, A. *et al.* (1995) 'Structure and organization of the European Yeast Genome Sequencing Network', *Journal of Biotechnology*, 41(2–3), pp. 131–137. doi: 10.1016/0168-1656(95)00066-y.

Vernikos, G. *et al.* (2015) 'Ten years of pan-genome analyses', *Current Opinion in Microbiology*. Elsevier Current Trends, pp. 148–154. doi: 10.1016/j.mib.2014.11.016.
Villa, N. O. *et al.* (2006) 'Phylogenetic relationships of Pythium and Phytophthora species based on ITS rDNA, cytochrome oxidase II and beta-tubulin gene sequences.', *Mycologia*. Mycological Society of America, 98(3), pp. 410–422. doi: 10.3852/mycologia.98.3.410.
Wakelin, S. A. *et al.* (2016) 'Cost of root disease on white clover growth in New Zealand dairy pastures', *Australasian Plant Pathology*, 45(3), pp. 289–296. doi: 10.1007/s13313-016-0411-x.
Wang, H. *et al.* (2009) 'A fungal phylogeny based on 82 complete genomes using the composition vector method.', *BMC evolutionary biology*, 9(1), p. 195. doi: 10.1186/1471-2148-9-195.
Wang, Q. M. *et al.* (2015) 'Phylogeny of yeasts and related filamentous fungi within Pucciniomycotina determined from multigene sequence analyses', *Studies in Mycology*. CBS Fungal Biodiversity Centre, 81, pp. 27–53. doi: 10.1016/j.simyco.2015.08.002.
Wang, X., Liu, X. and Groenewald, J. Z. (2017) 'Phylogeny of anaerobic fungi (phylum Neocallimastigomycota), with contributions from yak in China', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*. Springer, 110(1), pp. 87–103. doi: 10.1007/s10482-016-0779-1.
Warringer, J. *et al.* (2011) 'Trait variation in yeast is defined by population history', *PLoS Genetics*. Edited by L. Kruglyak. Public Library of Science, 7(6), p. e1002111. doi: 10.1371/journal.pgen.1002111.
Wehe, A. *et al.* (2008) 'DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony', *Bioinformatics*, 24(13), pp. 1540–1541. doi: 10.1093/bioinformatics/btn230.
Wei, W. *et al.* (2007) 'Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 104(31), pp. 12825–12830. doi: 10.1073/pnas.0701291104.
Weir, B. S. *et al.* (2015) 'A taxonomic revision of phytophthora clade 5 including two new species, phytophthora agathidicida and P. Cocois', *Phytotaxa*, 205(1), pp. 21–38. doi: 10.11646/phytotaxa.205.1.2.
Wendel, J. F. *et al.* (2016) 'Evolution of plant genome architecture', *Genome Biology*. BioMed Central, 17(1), p. 37. doi: 10.1186/s13059-016-0908-1.
van der Werf, M. J., Overkamp, K. M. and de Bont, J. A. M. (1998) 'Limonene-1,2-Epoxide Hydrolase from Rhodococcus erythropolis DCL14 Belongs to a Novel Class of Epoxide Hydrolases', *J. Bacteriol.*, 180(19), pp. 5052–5057. Available at: http://jb.asm.org/content/180/19/5052.full (Accessed: 16 May 2016).
Werres, S. *et al.* (2001) 'Phytophthora ramorum sp. nov., a new pathogen on Rhododendron and Viburnum', *Mycological Research*. Elsevier, 105(10), pp. 1155–1165. doi: 10.1016/S0953-7562(08)61986-3.
Whitaker, J. W., McConkey, G. A. and Westhead, D. R. (2009) 'The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes.', *Genome biology*. BioMed Central, 10(4), p. R36. doi:

10.1186/gb-2009-10-4-r36.
Wickham, H. (2011) ‘ggplot2’, *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), pp. 180–185. doi: 10.1002/wics.147.
Wilkinson, M. *et al.* (2004) ‘Some desiderata for liberal supertrees’, in Bininda-Emonds, O. R. P. (ed.) *Phylogenetic supertrees: combining information to reveal the Tree of Life*. Dordrecht: Springer Netherlands, pp. 227–246. doi: 10.1007/978-1-4020-2330-9_11.
Wisecaver, J. H., Slot, J. C. and Rokas, A. (2014) ‘The Evolution of Fungal Metabolic Pathways’, *PLoS Genetics*. Public Library of Science, 10(12), p. e1004816. doi: 10.1371/journal.pgen.1004816.
Woese, C. R. and Fox, G. E. (1977) ‘Phylogenetic structure of the prokaryotic domain: the primary kingdoms.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 74(11), pp. 5088–90. doi: 10.1073/pnas.74.11.5088.
Wolfe, D. *et al.* (2013) ‘Visualizing genomic information across chromosomes with PhenoGram’, *BioData Mining*, 6(1), p. 18. doi: 10.1186/1756-0381-6-18.
Wolfe, K. H. (2015) ‘Origin of the yeast whole-genome duplication’, *PLoS Biology*. Public Library of Science, 13(8), p. e1002221. doi: 10.1371/journal.pbio.1002221.
Wolfe, K. H. and Shields, D. C. (1997) ‘Molecular evidence for an ancient duplication of the entire yeast genome.’, *Nature*, 387(6634), pp. 708–713. doi: 10.1038/42711.
Wood, V. *et al.* (2002) ‘The genome sequence of Schizosaccharomyces pombe’, *Nature*, 415(6874), pp. 871–880. doi: 10.1038/nature724.
Yang, E., Hulse, A. M. and Cai, J. J. (2012) ‘Evolutionary analysis of sequence divergence and diversity of duplicate genes in Aspergillus fumigatus’, *Evolutionary Bioinformatics*. SAGE Publications, 2012(8), pp. 623–644. doi: 10.4137/EBO.S10372.
Yang, Z. (2007) ‘PAML 4: Phylogenetic analysis by maximum likelihood’, *Molecular Biology and Evolution*. Oxford University Press, 24(8), pp. 1586–1591. doi: 10.1093/molbev/msm088.
Yang, Z. and Nielsen, R. (2000) ‘Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models’, *Molecular Biology and Evolution*, 17(1), pp. 32–43. doi: 10.1093/oxfordjournals.molbev.a026236.
Yogev, Ohad *et al.* (2010) ‘Fumarase: a mitochondrial metabolic enzyme and a cytosolic/nuclear component of the DNA damage response.’, *PLoS biology*. Public Library of Science, 8(3), p. e1000328. doi: 10.1371/journal.pbio.1000328.
Yoon, H. S. *et al.* (2002) ‘The single, ancient origin of chromist plastids.’, *Proceedings of the National Academy of Sciences of the United States of America*, 99(24), pp. 15507–15512. doi: 10.1073/pnas.242379899.
Yoshida, K. *et al.* (2013) ‘The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine’, *eLife*. eLife Sciences Publications, Ltd, 2013(2), p. e00731. doi: 10.7554/eLife.00731.
Young, J. P. W. *et al.* (2006) ‘The genome of Rhizobium leguminosarum has recognizable core and accessory components’, *Genome Biology*. BioMed Central, 7(4), p. R34. doi: 10.1186/gb-2006-7-4-r34.

Youssef, N. H. *et al.* (2013) 'The genome of the anaerobic fungus orpinomyces sp. strain c1a reveals the unique evolutionary history of a remarkable plant biomass degrader', *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 79(15), pp. 4620–4634. doi: 10.1128/AEM.00821-13.
Yue, J. X. *et al.* (2017) 'Contrasting evolutionary genome dynamics between domesticated and wild yeasts', *Nature Genetics*. Europe PMC Funders, 49(6), pp. 913–924. doi: 10.1038/ng.3847.
Zeng, W. *et al.* (2016) 'Comparative genomics analysis of a series of Yarrowia lipolytica WSH-Z06 mutants with varied capacity for α-ketoglutarate production', *Journal of Biotechnology*. Elsevier, 239, pp. 76–82. doi: 10.1016/j.jbiotec.2016.10.008.
Zhao, Y. *et al.* (2012) 'PGAP: Pan-genomes analysis pipeline', *Bioinformatics*. Narnia, 28(3), pp. 416–418. doi: 10.1093/bioinformatics/btr655.
Zhou, P. *et al.* (2017) 'Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes', *BMC Genomics*. BioMed Central, 18(1), p. 261. doi: 10.1186/s12864-017-3654-1.
Zuo, G. *et al.* (2010) 'Jackknife and Bootstrap Tests of the Composition Vector Trees', *Genomics, Proteomics and Bioinformatics*. Elsevier, 8(4), pp. 262–267. doi: 10.1016/S1672-0229(10)60028-9.
Zuo, G., Li, Q. and Hao, B. (2014) 'On K-peptide length in composition vector phylogeny of prokaryotes', *Computational Biology and Chemistry*, 53(PA), pp. 166–173. doi: 10.1016/j.compbiolchem.2014.08.021.

Supplementary material

Supplementary material is provided with the USB containing the electronic copy of this thesis, and at http://chmccarthy.github.io/thesis. An index is given below.

Chapter 1

No supplementary figures or tables included in Chapter 1.

Chapter 2

Supplementary figures

Figure S2.1: Detailed HGT of class II fumarase into *Pythium* and *Phytopythium* species.

Figure S2.2: Detailed HGT of NmrA-like quinone oxidoreductase into *Pythium* species.

Figure S2.3: Detailed HGT of SnoaL-like polyketide synthase into *Pythium* species.

Figure S2.4a: Detailed HGT of epoxide hydrolase into *Phytophthora capsici.*

Figure S2.4b: Detailed HGT of alcohol dehydrogenase into *Phytophthora* species.

Figure S2.5: Consensus method network of phylogenetic splits within *Phytophthora capsici* epoxide hydrolase phylogeny.

Supplementary tables

Table S2.1: Dataset of species genomes included in HGT analysis.

Table S2.2: Homology analysis of putative HGT genes and adjacent genes.

Table S2.3: Genetic characteristics of putative HGT genes in oomycete genomes.

Table S2.4: Characteristics of bacterial homologs to putative HGT genes.

Table S2.5: Local protein-protein alignments of putative HGT genes and bacterial homologs.

Table S2.6: InterPro analyses of putative HGT genes.

Table S2.7: GCUA analysis of putative HGT genes and donor bacterial genomes.

Chapter 3

Supplementary figures

Figure S3.1: Consensus method network of phylogenetic splits within oomycete supertree.

Figure S3.2: ML supermatrix phylogeny of 37 oomycete and 6 SAR species.

Figure S3.3: Neighbour-joining network of phylogenetic splits within oomycete supermatrix phylogeny.

Supplementary tables

Table S3.1: Taxonomy, statistics and gene prediction results for 14 oomycete species.

Table S3.2: Model selection in single- and multi-copy oomycete phylogenies.

Chapter 4

No supplementary figures or tables included in Chapter 4.

Chapter 5

Supplementary figures

Figure S5.1: Simplified illustration of the methodology of PanOCT.

Figure S5.2: Analysis workflow for **(a)** gene prediction and **(b)** pangenome analyses.

Figure S5.3: UpSetR plot of ortholog distribution in *S. cerevisiae* accessory genome.

Figure S5.4: UpSetR plot of ortholog distribution in *C. albicans* accessory genome.

Figure S5.5: UpSetR plot of ortholog distribution in *Cr. neoformans* accessory genome.

Figure S5.6: UpSetR plot of ortholog distribution in *A. fumigatus* accessory genome.

Figure S5.7: Karyotype plots of fungal core and accessory genomes in reference strains.

Supplementary tables

Table S5.1: Strain datasets for four fungal pan-genomes.

Table S5.2: GO-slim enrichment analysis for fungal core and accessory genomes.

Table S5.3: Putative inter- and intra-domain HGT events in fungal accessory genomes.

Table S5.4: Statistical analysis of chromosomal distributions of core and accessory genomes.

Table S5.5: Statistical and comparative analysis of gene clusters and phenotypes in fungi.

Chapter 6

Supplementary tables

Table S6.1: *Yarrowia lipolytica* pan-genome dataset.

Table S6.2: GO-slim enrichment analysis for *Yarrowia lipolytica* core and accessory genome.

Chapter 7

No supplementary figures or tables included in Chapter 7.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere™

RESEARCH ARTICLE
Ecological and Evolutionary Science

Systematic Search for Evidence of Interdomain Horizontal Gene Transfer from Prokaryotes to Oomycete Lineages

Charley G. P. McCarthy, David A. Fitzpatrick

Genome Evolution Laboratory, Department of Biology, Maynooth University, Maynooth, County Kildare, Ireland

ABSTRACT While most commonly associated with prokaryotes, horizontal gene transfer (HGT) can also have a significant influence on the evolution of microscopic eukaryotes. Systematic analysis of HGT in the genomes of the oomycetes, filamentous eukaryotic microorganisms in the *Stramenopiles-Alveolates-Rhizaria* (SAR) supergroup, has to date focused mainly on intradomain transfer events between oomycetes and fungi. Using systematic whole-genome analysis followed by phylogenetic reconstruction, we have investigated the extent of interdomain HGT between bacteria and plant-pathogenic oomycetes. We report five putative instances of HGT from bacteria into the oomycetes. Two transfers were found in *Phytophthora* species, including one unique to the cucurbit pathogen *Phytophthora capsici*. Two were found in *Pythium* species only, and the final transfer event was present in *Phytopythium* and *Pythium* species, the first reported bacterium-inherited genes in these genera. Our putative transfers included one protein that appears to be a member of the *Pythium* secretome, metabolic proteins, and enzymes that could potentially break down xenobiotics within the cell. Our findings complement both previous reports of bacterial genes in oomycete and SAR genomes and the growing body of evidence suggesting that interdomain transfer from prokaryotes into eukaryotes occurs more frequently than previously thought.

IMPORTANCE Horizontal gene transfer (HGT) is the nonvertical inheritance of genetic material by transfer between different species. HGT is an important evolutionary mechanism for prokaryotes and in some cases is responsible for the spread of antibiotic resistance from resistant to benign species. Genome analysis has shown that examples of HGT are not as frequent in eukaryotes, but when they do occur they may have important evolutionary consequences. For example, the acquisition of fungal genes by an ancestral *Phytophthora* (plant destroyer) species is responsible for the large repertoire of enzymes in the plant-degrading arsenal of modern-day *Phytophthora* species. In this analysis, we set out to systematically search oomycete genomes for evidence of interdomain HGT (transfer of bacterial genes into oomycete species). Our results show that interdomain HGT is rare in oomycetes but has occurred. We located five well-supported examples, including one that could potentially break down xenobiotics within the cell.

KEYWORDS: *Phytophthora*, *Phytopythium*, *Pythium*, interdomain HGT, oomycota

Horizontal gene transfer (HGT), "the nongenealogical transfer of genetic material from one organism to another" (1), is most closely associated with antimicrobial resistance in bacteria. The cumulative effect of transfer events has had a significant impact on overall prokaryotic genome evolution. For example, it is estimated that up to 80% of genes in some prokaryote genomes underwent intradomain HGT at some point in their history (2). Interdomain transfer of genetic material between prokaryotes and eukaryotes has previously been understood in the context of endosymbiotic gene transfer, which has made a significant contribution to the evolution of eukaryotic genomes (3), most notably in the evolution of the mitochondrion in eukaryotes through an ancestral primary endosymbiosis event with a *Rickettsia*-like alphaproteobacterium and the evolution of the plastid in the *Archaeplastida* through ancestral primary endosymbiosis with a cyanobacterium (4). However, there is a growing body of literature supporting the notion of the existence of HGT between prokaryotes and eukaryotes, and many nonendosymbiotic horizontal interdomain gene transfer events between bacteria and eukaryotes have been described (5). Numerous metabolic genes have been transferred into the genomes of parasitic microbial eukaryotes (6, 7). Over 700 bacterial genes are present across fungi, with a particular concentration in *Pezizomycotina* (8); 71 putative bacterial genes have been identified in *Hydra vulgaris* (9); and the plant-parasitic nematode *Meloidogyne incognita* secretes cell wall-degrading enzymes inherited from soil-dwelling *Actinomycetales* and the betaproteobacterium *Ralstonia solanacearum* (10).

The oomycetes are a class of microscopic eukaryotes placed in the diverse stramenopile (or heterokont) lineage within the *Stramenopiles-Alveolata-Rhizaria* (SAR) eukaryotic supergroup (11). Historically classified as fungi due to their filamentous growth and similar ecological roles, oomycetes can be distinguished from "true" fungi by a number of structural, metabolic, and reproductive differences (12). The present placement of the oomycetes within the stramenopile lineage, and, by extension, within the SAR supergroup, is supported by phylogenomic analyses of 18S rRNA and conserved protein and expressed sequence tag (EST) data, which also support the supergroup's monophyly over previous configurations such as "chromalveolates" (13–16).

The most ecologically destructive orders within the oomycetes are the *Saprolegniales* order, whose member species are known as "cotton molds," which includes marine and freshwater pathogens of fish, and the closely related and predominantly terrestrial plant-pathogenic orders *Peronosporales* and *Pythiales* (17). The *Pythiales* order includes members of the marine and terrestrial genus *Pythium*, necrotrophic generalistic causative agents of root rot and damping-off in many terrestrial plants (Table 1). Some species (*Pythium aphanidermatum* and *Pythium ultimum*) are found under high-temperature or greenhouse conditions, while others (*Pythium irregulare* and *Pythium iwayami*) are most virulent at lower temperatures (18). *Pythium ultimum* and *Pythium irregulare* have broad ecological host ranges, while *Pythium iwayami* and *Pythium arrhenomanes* display some preference for monocots (18, 19).

The *Peronosporales* order includes the paraphyletic hemibiotrophic genus *Phytophthora*, whose member species exhibit both broad and highly specialized host ranges (Table 1). Generalistic *Pythophthora* species include *Phytophthora ramorum* and *Phytophthora kernoviae* (causing sudden oak death and dieback in many other plant

Received 13 July 2016 **Accepted** 26 August 2016 **Published** 14 September 2016

Citation McCarthy CGP, Fitzpatrick DA. 2016. Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages. mSphere 1(5):e00195-16. doi:10.1128/mSphere.00195-16.

Editor Aaron P. Mitchell, Carnegie Mellon University

Address correspondence to David A. Fitzpatrick, david.fitzpatrick@nuim.ie.

TABLE 1 Summary of host ranges of plant-parasitic oomycete species analyzed in this study[a]

Species	Host(s)
Phytophthora capsici	Curcubits (e.g., *Cucurbita pepo*)
Phytophthora infestans	*Solanaceae* (e.g., *Solanum tuberosum*)
Phytophthora kernoviae	*Fagus sylvatica*, *Rhododendron*
Phytophthora lateralis	*Chamaecyparis lawsoniana*
Phytophthora parasitica	Broad range, including *Nicotiana tabacum*
Phytophthora ramorum	Broad range, including *Quercus*, *Rhododendron*
Phytophthora sojae	*Glycine max*
Phytopythium vexans	Tropical forest species
Pythium aphanidermatum	Broad range, virulent at higher temperatures
Pythium arrhenomanes	Monocots
Pythium irregulare	Broad range, virulent at lower temperatures
Pythium iwayami	Monocots, virulent at lower temperatures
Pythium ultimum var. *sporangiiferum*	Broad range
Pythium ultimum var. *ultimum*	Broad range, virulent at higher temperatures

[a]Refer to the introduction for references.

species, particularly *Rhododendron* spp.), *Phytophthora parasitica* (causing black shank disease in a diverse range of plants), and *Phytophthora capsici* (causing blight and root rot in *Cucurbitaceae*, *Solanaceae*, and *Fabaceae*). Species with more specialized host ranges include *Phytophthora sojae* and *Phytophthora lateralis* (causing root rot in soybean and Port Orford cedar, respectively), and *Phytophthora infestans* (causing late blight in some *Solanaceae* spp., most notoriously in potato). The tropical plant pathogen *Phytopythium vexans* was previously classified in *Pythium* clade K (19), but that clade has since been reclassified into *Phytopythium*, a morphological and phylogenetic genus intermediate between *Phytopthora* and *Pythium* (20).

To date, large-scale systematic analysis of the influence of HGT on oomycete genome evolution has focused on intradomain transfer between fungi and oomycetes (21, 22). The most extensive study revealed up to 34 putative transfers from fungi to oomycetes, many of which were associated with enzymes involved in carbohydrate metabolism (23). Three of these genes had previously been transferred from bacteria to fungi (24). Few events of HGT between bacteria and oomycetes have been described in the literature, and most incidents of interdomain HGT have been discovered within the context of fungus-focused studies. However, recent analyses have shown that actinobacterial cutinase has orthologs in a number of *Phytophthora* species (25), with subsequent copy expansion in *Phytophthora sojae*. Disintegrins and endonucleases secreted by *Saprolegnia parasitica* appear to be bacterial in origin (26), and studies of the secretomes of *Saprolegniales* species *Achlya hypogyna* and *Thraustotheca clavata* revealed one ancestral endoglucanase and three genes specific to the *Saprolegniales* order which had been transferred from bacteria (27). As with other unicellular eukaryotes, some genes in *Phytophtora* involved in amino acid metabolism have been obtained via horizontal transfer from bacteria (28). Other studies have identified ancestral bacterial events of HGT within other stramenopile genomes (29) or in other lineages within the SAR supergroup (30–32).

In light of these previous studies of the influence of HGT in the evolution of the oomycetes, we undertook a systematic investigation focusing on the extent of bacterial transfer into the oomycetes. We analyzed 13 species from the plant-pathogenic genera *Pythium* and *Phytophthora*, as well as the recently reclassified species *Phytopythium vexans*, for genes with sufficient evidence for nonvertical inheritance from bacteria. Here, we report five recent transfers from bacteria into individual oomycete lineages, including what we believe to be the first descriptions of interdomain HGT involving *Pythium*.

RESULTS AND DISCUSSION

Analysis of bacterial HGT into *Phytophthora* and *Pythium*. To investigate the extent of bacterial HGT into the oomycetes, we generated gene phylogenies for every oomycete protein sequence whose bidirectional homology analysis supported a recent transfer from bacteria to an oomycete species. Such phylogenies were generated with techniques that have previously identified multiple intradomain events of HGT between fungi and oomycetes (23): using OrthoMCL (33) to generate clusters of orthologous proteins, searching representative proteins against a large database using BLASTp (34), and generating maximum-likelihood phylogenetic reconstructions using PhyML (35). To reduce the chances of false-positive identification of putative HGT genes due to poor taxon sampling (36, 37), oomycete protein sequences were queried against a local database using BLASTp, with broad taxon sampling in the database across prokaryotes and eukaryotes (see Data Set S1 in the supplemental material). A total of 106 oomycete proteins were found to have a top database hit with a bacterial protein. Filtering for redundancy (due to multiple homologs in a single species, for example), 64 unique candidate maximum-likelihood HGT phylogenies with 100 bootstrap replicates (Table 2) were generated using PhyML with the best-fit model for each phylogeny chosen by ProtTest (38). Through our process of examination, we retained 25 phylogenies which satisfied our criteria (resolvable topology and adequate taxon sampling) (Table 2). Of these 25 phylogenies, 20 were ultimately discarded due to poor phylo-

TABLE 2 Identification of putative bacterial HGT sequences in *Phytophthora*, *Pythium*, and *Phytopythium*

Genus	No. of intergenic bacterial hits	No. of OrthoMCL clusters (no. of sequences)	No. of OrthoMCL unclustered sequences	No. of maximum likelihood phylogenies	Putative no. of HGT sequences
Phytophthora	31	22 (28)	3	25	3
Phytopythium/Pythium	75	16 (59)	23	39	2

genetic and bootstrap support or signal. Our phylogenies infer three types of bacterium-oomycete HGT within our candidate HGT phylogenies:

(i) Recent bacterial transfer into the *Pythium* or *Phytopythium* (*Pythium*/*Phytopythium*) lineage (1 individual example).

(ii) Recent bacterial transfer into the *Phytophthora* lineage (2 individual examples).

(iii) Recent bacterial transfer into the *Pythium* lineage (2 individual examples).

Each phylogeny was evaluated for other characteristics that might have led to reinforcement or rejection of our hypothesis that HGT had occurred. Gene characteristics such as GC content, exon number, and the sequence length of each oomycete gene arising from transfer in our phylogenies were calculated (see Table S1 in the supplemental material), and the results were compared to the average results determined for their corresponding genomes. Gene characteristics of bacterial homologs in potential donor species were also calculated (see Table S2). Similarly, the codon usage patterns of each *Phytophthora* and *Pythium*/*Phytopythium* genome were analyzed, and the patterns of each of the candidate genes potentially arising from HGT in each species were compared to the general pattern to see whether they were outliers. The codon usage patterns of the seed genes used to generate each phylogeny were also compared with the codon usage patterns of potential bacterial donors (not shown). None of these analyses were conclusive with respect to proving or disproving that horizontal inheritance of these genes had occurred. However, this is not uncommon for codon usage analyses as the codon usage of transferred genes is known to ameliorate to match that of the recipient genome (39). Sequence similarity and identity at the amino acid level between each seed HGT protein and a sister homolog from a potential bacterial donor were also investigated (see Table S3).

To help ensure that none of our putative HGT families were in fact the product of bacterial contamination, the homology of each seed gene to its adjacent genes was investigated. In each of our five putative HGT families, we found that there was no obvious evidence of bacterial contamination along a source contig that resulted in false positives for bacterium-oomycete events of HGT (see Table S4 in the supplemental material). As we were also conscious of the risk of poor taxon sampling giving us false positives, we also compared the taxon sampling in our local database with the NCBI protein data. We queried each seed protein sequence against the NCBI's nonredundant protein sequence database using BLASTp with an E value cutoff of 10^{-20}, aligned homologs, and generated neighbor-joining phylogenies for each seed gene (not shown). Where the BLASTp data retrieved from NCBI mirrored our own local searches and the corresponding neighbor-joining phylogeny showed that the seed gene clearly grouped within an oomycete clade or a bacterial clade, we were satisfied that our taxon sampling had sufficiently covered all available protein data. All 5 of our candidate HGT genes satisfy these criteria.

We have identified five well-supported phylogenies that show putative events of HGT from bacterial species into the oomycetes. Three display topologies supporting a recent transfer into the *Pythium* or *Phytopythium* lineage (Fig. 1, 2, and 3), while the remaining two support a recent HGT into the *Phythophthora* lineage (Fig. 4 and 5). Below, we present and discuss each recent transfer individually, describing both the hypothesis for horizontal inheritance in each phylogenetic reconstruction and the functional characterization of each

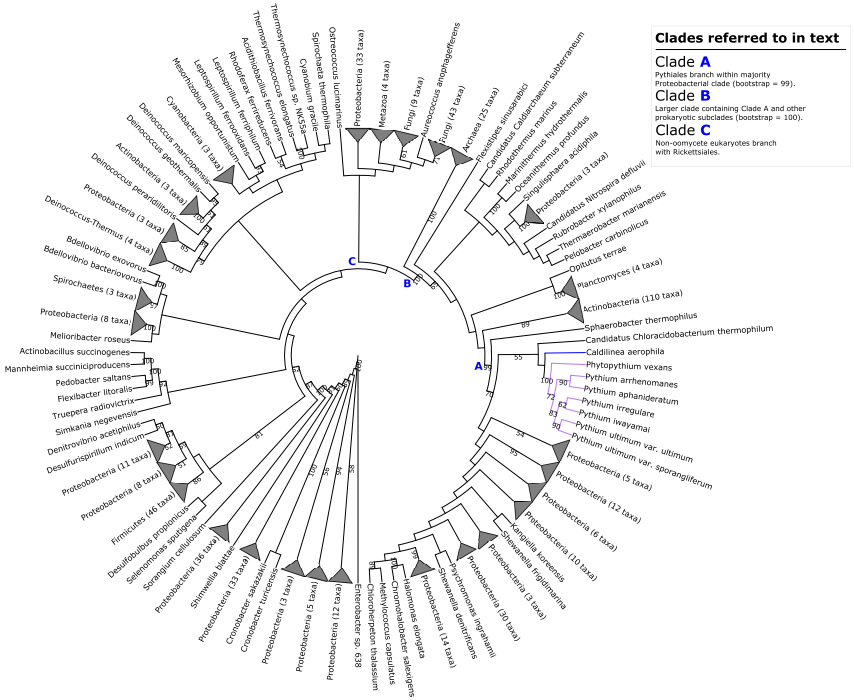


FIG 1 Maximum-likelihood phylogeny illustrating putative transfer of class II fumarase from *Caldilinea aerophila* into the *Phytopythium/Pythium* lineage. Clades A, B, and C referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. The corresponding full phylogenetic trees with detailed clades can be viewed in Fig. S1 in the supplemental material.

transferred gene family. We also compare the placement of the oomycete homologs in each of the five phylogenies with those of other eukaryotic homologs. This comparison is important as we expect transferred genes to violate the species phylogeny and transferred genes should form sister clades with bacterial species rather than their eukaryotic homologs. Each transfer is also summarized in Table 3.

A putative class II fumarase distinct from *Rickettsia* class II fumarase in *Phytopythium vexans* and *Pythium* spp. originates from bacteria. A protein in *Pythium ultimum* var. *sporangiiferum* (Table 3) was identified in our BLASTp homology

TABLE 3 Summary of each putative bacterium-oomycete HGT event

Tree	Seed species	Potential donor(s)	Identity (%)	Putative function	Secreted
Fig. 1	*Pythium ultimum*	*Caldilinea aerophila*	56.5	Class II fumarase	No
Fig. 2	*Pythium aphanidermatum*	*Proteobacteria*	54.0	NmrA-like quinone oxidoreductase	No
Fig. 3	*Pythium aphanidermatum*	*Actinobacteria*	58.6	SnoaL-like polyketide cyclase	Yes
Fig. 4	*Phytophthora capsici*	*Methylobacterium radiotolerans*	68.2	Epoxide hydrolase	No
Fig. 5	*Phytophthora capsici*	*Sphingomonas*	59.1	Alcohol dehydrogenase	No

searches as a candidate for an interdomain HGT event into oomycete species. The maximum-likelihood phylogeny of this protein family was generated from a family containing 550 homologs, with an LG+I+G+F substitution model (Fig. 1). A total of 16 bacterial phyla were present in this reconstruction, among which *Proteobacteria* and *Actinobacteria* were by far the most extensively represented. A total of 26 archaeal homologs were also present, of which all except a "*Candidatus* Caldiarchaeum subterraneum" sequence form a monophyletic clade. Across the eukaryotes, homologs are present in fungi, animals, green algae, and the stramenopiles.

Our phylogenetic reconstruction shows a monophyletic *Pythium/Phytopythium* clade within a large, predominantly proteobacterial clade with 99% bootstrap support, adjacent to a homolog from the filamentous *Chloroflexi* species *Caldilinea aerophila* (Fig. 1, clade A). Further back along the tree, this greater subclade branches deep within a large prokaryotic clade with 100% bootstrap support and contains three major subclades: the aformentioned majority-proteobacterial subclade containing *Pythium* and *Phytopythium* orthologs, a halophilic archaeal subclade, and a large actinobacterial subclade containing 110 homologs (Fig. 1, clade B). Elsewhere, all nonoomycete eukaryote homologs (with the exception of an adjacent sequence from the microscopic green alga *Ostreococcus lucimarinus*) are placed in a monophyletic eukaryote clade containing 52 fungal homologs, 4 animal homologs, and a homolog from the stramenopile alga *Aureococcus anophagefferns* adjacent to a clade containing 19 homologs from the alphaproteobacterial *Rickettsia* genus (Fig. 1, clade C). The neighbor-joining tree constructed from the BLAST homology search of the seed sequence against the NCBI's database places the seed deep within a large prokaryotic clade containing *Proteobacteria*, *Actinobacteria*, and halophilic and methanogenic archaea, in a gammaproteobacterial subclade similar to what we observed in our phylogenetic reconstruction (not shown).

Sequence analysis of the seed gene and its flanking genes in the *Pythium ultimum* var. *sporangiiferum* genome did not return any obvious evidence of bacterial contamination; the top hit of the seed protein sequence against the NCBI database was a *C. aerophila* sequence, but the top hits of both flanking protein sequences were *Phytophthora parasitica* homologs (see Table S4 in the supplemental material). BLAST homology searches against the NCBI database found that the seed sequence shared sequence similarity with many bacterial class II fumarases, and Pfam analysis of the sequence identified two lyase domains and the characteristic *fumC* C terminus of a class II fumarase-like sequence (see Data Set S1). InterProScan analysis identified further fumarase protein sequence signatures (see Data Set S1). Fumarase, also known as fumarate hydratase (EC 4.2.1.2), is an enzyme that catalyzes the reversible hydration of fumarate to (S)-malate in the mitochondrion in eukaryotes, as a component of the tricarboxylic acid cycle (40), and promotion of histone H3 methylation and DNA repair in the cytosol (41). There are two classes of fumarase: the heat-labile dimeric class I fumarases encoded by *fumA* and *fumB* found in prokaryotes and the heat-stable tetrameric class II fumarase encoded by *fumC* found in both prokaryotes and eukaryotes (42). While associated with mitochondrial function in eukaryotes, class II fumarases with distinct evolutionary histories have been detected in amitochondriate trichomonads (43).

The nature of the conserved function of the gene encoding class II fumarases in eukaryotic respiration would suggest that this gene had arisen in the nuclear genome of *Pythium* and *Phytopythium* by endosymbiotic gene transfer from the mitochondrial genome (44) and hence was not a product of recent transfer. To investigate the relationship between this putative horizontally transferred fumarase and other potential fumarase orthologs in the oomycetes, we aligned the seed *Pythium ultimum* var. *sporangiiferum* sequence against 20 known oomycete and 230 other eukaryote and prokaryote class II fumarase sequences. Sequence and phylogenetic analysis showed that it branches as an outgroup in the corresponding phylogeny (not shown), suggesting that it is not an ortholog of the endosymbiotic oomycete class II fumarase. It seems most parsimonious to suggest, therefore, that this fumarase protein in *Pythium* and

Phytopythium vexans is a class II fumarase distinct from endosymbiotic class II fumarase and arose by a completely separate transfer event, possibly with *C. aerophila* or another *Chloroflexi* species (*Sphaerobacter thermophilus*, for example) (Fig. 1). An interesting aspect of this phylogeny is the presence of a homolog from *Phytopythium vexans* branching with *Pythium* species and the absence of *Phytophthora* homologs in the phylogeny. *Phytopythium vexans*, along with other members of what was once *Pythium* clade K, was reclassified to the morphological intermediate genus *Phytopythium*, based on molecular evidence, with ribosomal large subunit (LSU), internal transcribed spacer (ITS), and mitochondrial cytochrome oxidase 1 (CO1). Furthermore, the resultant phylogenetic data grouped *Phytopythium* and *Phytophthora* as sister taxa with strong bootstrap support (20). This would suggest that the ancestor of the *Phytophthora*, *Phytopythium*, and *Pythium* species obtained a bacterial copy of the class II fumarase and that it was subsequently lost in the *Phytophthora* clade. Alternatively, if we assume that rare events of HGT can act as phylogenetic markers (3), it is plausible that *Phytopythium* and *Pythium* are in fact more closely related to one another, to the exclusion of *Phytophthora* species. This observation challenges the phylogeny derived from traditional phylogenetic markers (20), and we suggest that the relationships between these groups warrant further examination.

A putative proteobacterial NmrA-like oxidoreductase is present in multiple *Pythium* species. A *Pythium aphanidermatum* gene (Table 3) was identified in our homology searches as a candidate for bacterial HGT into an oomycete species. The maximum-likelihood phylogeny of this gene was constructed from a gene family containing 258 homologs, with an LG+I+G+F substitution model (Fig. 2). Among these homologs, 95% (245 of 258) were bacterial, representing 10 different phyla. The majority of bacterial homologs were from *Proteobacteria*, *Actinobacteria*, or *Firmicutes* species. Of the 13 eukaryote homologs present, 12 were from the oomycetes and 1 was from the fungal species *Trichoderma viride* (Fig. 2).

In our reconstruction, homologs (12 in total) from each *Pythium* species except *Pythium ultimum* var. *sporangiiferum* formed a monophyletic subclade (99% bootstrap support) within a 70-member clade with 92% bootstrap support. Every other member of this clade except *Trichoderma viride* was bacterial. Around 30 members of this clade, many of which were soil-dwelling *Rhizobales*, were proteobacterial (Fig. 2, clade B). The *Pythium* subclade branches with 83% bootstrap support beside a small proteobacterial subclade that includes two nitrogen-fixing species in *Bradyrhizobium* and *Xanthomonas albilineans*, the causative agent of leaf scald disease in sugarcane (45) (Fig. 2, clade A). Homology analysis of the seed sequence and its flanking sequences in the *P. aphanidermatum* genome found no obvious evidence of bacterial contamination, and the seed sequence was most closely related to a *Rubrivivax gelatinosus* sequence; however, flanking genes had top hits from *Phytophthora infestans* (see Table S4 in the supplemental material). The neighbor-joining phylogeny generated from BLAST homology searches of the seed sequence against the NCBI's protein database also placed the seed sequence adjacent to a large proteobacterial clade (not shown).

BLAST homology searches against the NCBI database found that the seed sequence shared homology with bacterial nucleotide-sugar epimerases and NAD(P)-binding proteins. Pfam analysis of the sequence found the characteristic Rossmann fold of NAD(P)-binding proteins (see Data Set S1 in the supplemental material), while InterProScan analysis found NmrA-like family and quinone oxidoreductase 2 subfamily PANTHER signatures (see Data Set S1). NmrA is a NAD(P)-binding negative transcriptional regulator, involved in the regulation of nitrogen metabolite repression (NMR) genes in fungi, which suppress metabolic pathways for secondary nitrogen sources when preferred sources like ammonium and glutamine are available (46). Such a metabolic system has not been described in oomycetes to date. The PANTHER quinone oxidoreductae subfamily (47) to which this transferred gene belongs (PTHR14194:SF73) includes eukaryotic orthologs from *Pezizomycotina*, *Monosiga brevicollis* and *Dictyostelium* spp., *Phytophthora infestans* and *Physcomitrella patens*, and bacterial orthologs

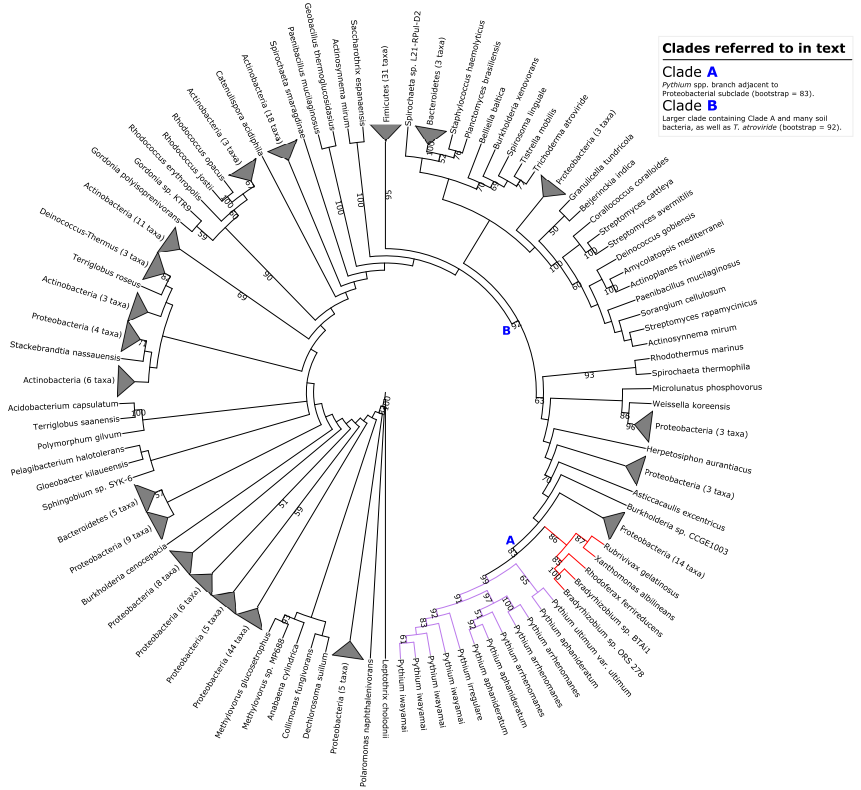


FIG 2 Maximum-likelihood phylogeny illustrating putative transfer of NmrA-like quinone oxidoreductase from *Proteobacteria* into *Pythium* spp. Clades A and B referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. The corresponding full phylogenetic trees with detailed clades can be viewed in Fig. S2 in the supplemental material. *T. atroviride*, *Trichoderma atroviride*.

from multiple lineages. Among these orthologs is *qorB* in *Escherichia coli* K-12, which has redox activity on NAD(P)H using quinone as an acceptor (48).

Our phyogenetic reconstruction of this *Pythium aphanidermatum* gene supports the hypothesis of the transfer of this gene into *Pythium* spp. from a soil-dwelling proteobacterium (Fig. 2), either the phototrophic betaproteobacterial species *Rhodoferax ferrireducens*/*Rubrivivax gelatinosus* or the phytopathogenic gammaproteobacterium *Xanthomonas albilineans*. Species related to *X. albilineans* and *R. ferrireducens*, within *Xanthomonadales* and *Comamonadaceae*, respectively, have been identified in previous studies as endohyphal bacteria, hypha-dwelling endosymbionts of endophytic fungi (49, 50). It is not currently known whether such bacteria can also inhabit the hyphae of oomycetes and thus consequently provide favorable conditions for potential interdomain HGT. This transferred gene may be a NAD(P)H-binding quinone oxidoreductase (EC 1.6.5.2) and potentially has cytosolic redox activity in *Pythium* spp.

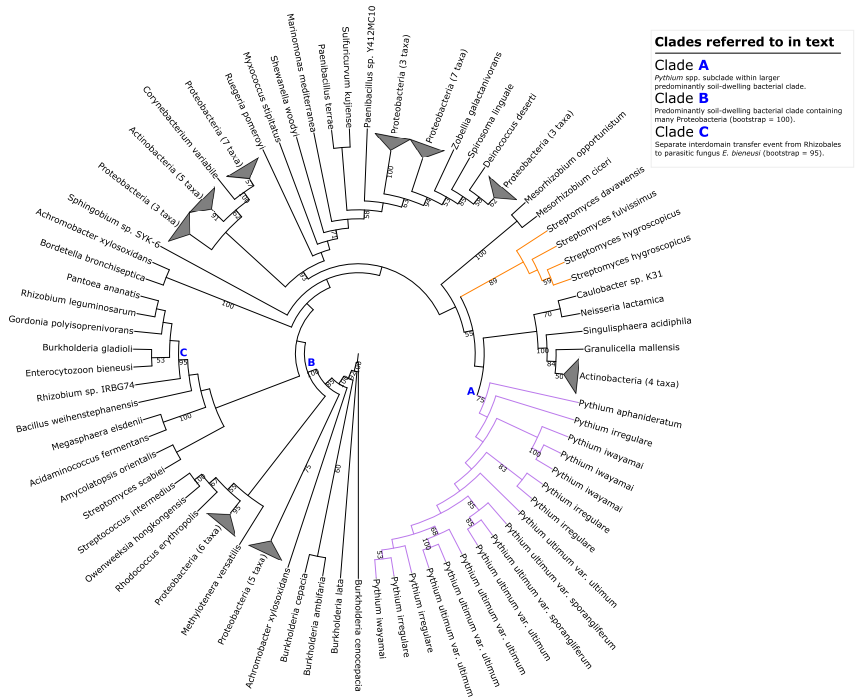


FIG 3 Maximum-likelihood phylogeny illustrating putative transfer of SnoaL-like polyketide cyclase from *Actinobacteria* into *Pythium* spp. Clades A, B, and C referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. The corresponding full phylogenetic trees with detailed clades can be viewed in Fig. S3 in the supplemental material.

SnoaL-like proteins from soil-dwelling bacteria are putative members of the secretome of multiple *Pythium* species. A second gene from *P. aphanidermatum* (Table 3) was identified in our BLASTp homology searches as a candidate for bacterial HGT into an oomycete species. The maximum-likelihood phylogeny of this gene was generated from a gene family containing 103 homologs constructed with a WAG+I+G substitution model (Fig. 3). Seven bacterial phyla are present in this reconstruction, along with *Pythium* and the fungal parasite *Enterocytozoon bieneusi*, and 53% of the homologs (55 of 103) come from proteobacterial species.

The maximum-likelihood phylogenetic reconstruction places 17 *Pythium* homologs (with multiple paralogs in each species except *P. aphanidermatum* and no homolog in *P. arrhenomanes*) deep within a 93-member clade containing many typical soil-dwelling proteobacterial and actinobacterial species (Fig. 3, clade B) with 100% bootstrap support. The *Pythium* subclade (Fig. 3, clade A) is adjacent to a clade containing four orthologs from *Mycobacterium smegmatis*. The only other eukaryote homolog in our analysis (*E. bieneusi*) is placed in a separate subclade containing *Rhizobales* species with 95% bootstrap support, indicative of a separate independent HGT event (Fig. 3, clade C). Homology analysis of the seed sequence and its adjacent sequences returned no

evidence of bacterial contamination. The sequences of both flanking genes are homologous to sequences in other oomycetes, and the seed sequence's highest degree of homology was with a *Streptomyces yerevanensis* sequence (see Table S4 in the supplemental material).

BLAST homology searches of the seed sequence found numerous instances of homology with bacterial SnoaL-like polyketide cyclases. Pfam and InterProScan analysis of the sequence identified two SnoaL-like domains and a number of signal peptide signatures within the N-terminal domain (see Data Set S1 in the supplemental material). Polyketide cyclases are enzymatic components of the synthesis of aromatic polyketide compounds from carboxylic acids in bacteria and fungi. Polyketides are best characterized by the medicinally useful secondary metabolites produced by various *Actinobacteria* genera, such as the antitumourigenic anthracyclines from *Streptomyces* species (51). Biochemically, polyketide cyclases catalyze the intramolecular cyclization of poly-β-ketone chain intermediates to form the core planar polyaromatic structures of polyketides, which are then subject to later functionalization. In the biosynthesis of the anthracycline nogalamycin in *Streptomyces nogalater*, the polyketide cyclase SnoaL (EC 5.5.1.26) catalyzes ring closure of a polyaromatic nogalamycin precursor through aldol condensation (52).

The maximum-likelihood phylogenetic reconstruction of this transfer event appears to support the transfer of this putative SnoaL-like protein into a *Pythium* ancestor from a proteobacterial or actinobacterial donor (Fig. 3). Similarly, the neighbor-joining tree generated from the homology search against NCBI's nonredundant database places the *P. aphanidermatum* seed sequence within a large proteobacterial and actinobacterial clade (not shown). The SignalP (53) and TargetP (54) analyses both indicated that the protein contains a 25-reside-long signal peptide sequence at its N terminus with a discrimination score (used to distinguish between signal and nonsignal peptides) well above the default cutoff value and thus identified the protein as part of the secretome of *P. aphanidermatum*. Therefore, this putative SnoaL-like protein may have arisen in *Pythium* species through horizontal transfer from an *Actinobacteria* species and may be a putative component of the secretome of *Pythium* species. It is worth noting that no polyketide synthase genes have been detected in model *Phytophthora* genomes and that, in general, oomycetes rely more on toxic effector proteins than on toxic small-molecule secondary metabolites for necrotrophic growth (55, 56). The presence of this putative SnoaL-like protein in multiple copies in most of the *Pythium* species that we investigated suggests an additional method of phytopathogenic infection which may be novel to *Pythium* or which may have been subsequently lost in *Phytophthora*.

A putative hydrolase from xenobiotic-degrading rhizosphere proteobacteria is present in *Phytophthora capsici*. A gene from *Phytophthora capsici* (Table 3) was identified in our BLASTp homology searches as a candidate for bacterial HGT. A maximum-likelihood phylogeny was generated from 253 homologs using a WAG+G substitution model. Eight bacterial phyla are represented in our reconstruction, with the majority of homologs coming from either proteobacterial or actinobacterial species. A total of 57 fungal homologs and 3 paralogs from *Physcomitrella patens* (earthmoss) form a monophyletic eukaryotic clade (Fig. 4, clade B). Our maximum-likelihood phylogenetic tree placed two homologs from *P. capsici* adjacent to a homolog from the alphaproteobacterium *Methylobacterium radiotolerans* within a bacterial clade containing *Acidobacteria* and a number of soil-borne or plant-epiphytic *Proteobacteria* (Fig. 4, clade A). BLASTp analysis aligned the seed sequence with an ortholog from the nitrogen-fixing proteobacterium *Azotobacter vinelandii*. As there is only one *Phytophthora* species represented in this phylogeny, we carefully examined the sequence of the contig to rule out a bacterial contamination artifact in the *P. capsici* genome. All flanking genes were from *Phytophthora* spp., thereby giving us confidence that this represents a bona fide HGT event (see Table S4 in the supplemental material). Furthermore, the phylogeny generated after homology searches against the NCBI database placed the seed sequence within a large proteobacterial clade (not shown).

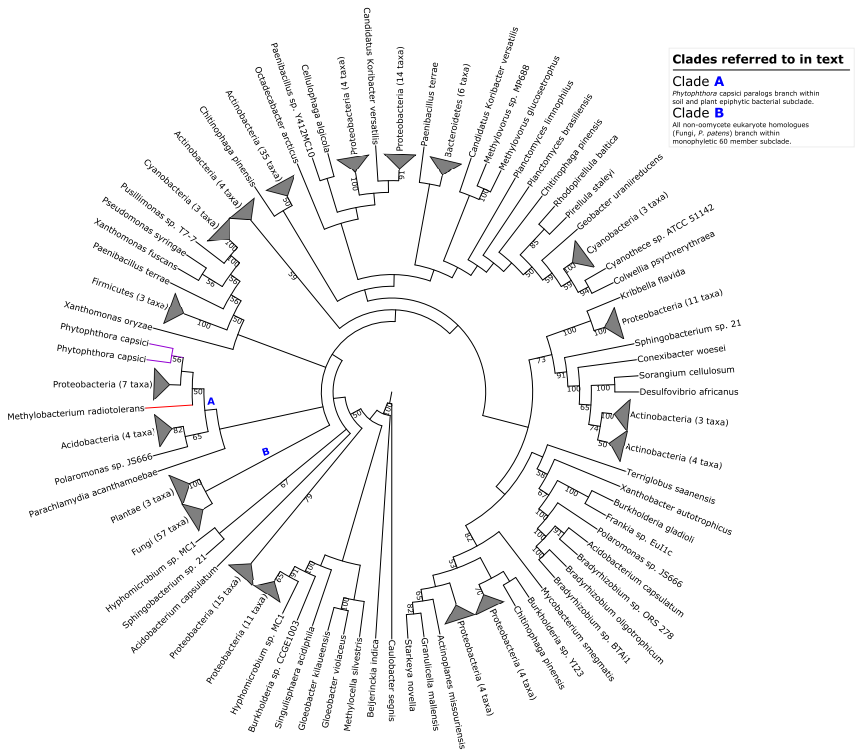


FIG 4 Maximum-likelihood phylogeny illustrating putative transfer of epoxide hydrolase from *Methylobacterium radiotolerans* into *Phytophthora capsici*. Clades A and B referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. The corresponding full phylogenetic trees with detailed clades can be viewed in Fig. S4A in the supplemental material.

As the levels of bootstrap support for many of the more derived branches and clades in our phylogeny, including the bacterial clade containing *P. capsici* homologs, were weak (<50%), we generated a median phylogenetic network of all splits in the set of individual bootstrap trees generated by PhyML in our reconstruction using a consensus network method in SplitsTree (57). This consensus network (see Fig. S5 in the supplemental material) places the two *P. capsici* homologs at the base of the large monophyletic bacterial clade, clearly separate from the fungal and plant homologs. With this analysis, we were satisfied that the phylogeny represented a bona fide bacterium-oomycete HGT event.

BLAST homology searches of the seed sequence against the NCBI database indicated that the sequence was homologous to those associated with bacterial hydrolases. Pfam analysis found a large α/β hydrolase fold domain present in the sequence, and InterProScan analysis returned a number of α/β hydrolase family PANTHER signatures, as well as epoxide hydrolase PRINTS (58) signatures, across the sequence (see Data

Set S1 in the supplemental material). Epoxide hydrolases (EC 3.3.2.3) catalyze the dihydroxylation of epoxide residues to diols and are among the members of a number of protein families that contain an α/β hydrolase fold (59). Bacterial epoxide hydrolases are capable of degradation of xenobiotic organic compounds (60, 61). The structurally related haloalkane dehalogenases (EC 3.8.1.5), which can hydrolyze toxic haloalkanes into their corresponding alcohol and organic halide components in the cytosol, are widespread in soil bacteria (62). It is interesting that strains of *M. radiotolerans* isolated from *Cucurbita pepo* roots, which is also a target for *P. capsici*, are capable of degrading xenobiotic 1,1-*bis*-(4-chlorophenyl)-2,2-dichloroethene (DDE) (63). DDE is a highly toxic and highly recalcitrant major metabolite of the degradation of the toxic organochloride pesticide 1,1,1-trichloro-2,2-bis(p-chlorophenyl)ethane (DDT), which saw widespread use for most of the 20th century (64).

Our maximum-likelihood phylogenetic reconstruction suggests that this putative hydrolase gene, which has two copies in *P. capsici*, arose through horizontal transfer from soil-dwelling bacteria, potentially from *M. radiotolerans* (Fig. 4). Homology and functional analysis of the seed HGT gene indicates that these two paralogs contain hydrolase folds. The two paralogs in *P. capsici* are somewhat dissimilar at the nucleotide level; one appears to contain both peptidase and α/β hydrolase domains and is far more exonic than the seed HGT gene (see Table S1 in the supplemental material). This putative transferred gene may have a potential cytosolic role in the degradation of toxic xenobiotic compounds in *P. capsici*. To date, descriptions of xenobiotic degradation or resistance in oomycetes have been sparse in the literature; what is known is that few oomycete cytochrome P450 proteins (CYPs) appear to be involved in xenobiotic degradation compared with fungal CYPs (65, 66) and that *Phytophthora infestans* has far a lower proportion of major facilitator superfamily (MFS) transport proteins involved in efflux than many fungal type species do (67). As such, this acquisition may be a novel event in the context of plant-parasitic oomycete genome evolution.

Sphingomonadale alcohol dehydrogenase is present in five *Phytophthora* species. A second *P. capsici* gene (Table 3) was identified in our BLASTp homology searches as a candidate for interdomain HGT. Our phylogenetic reconstruction used 358 homologs with an LG+I+G substitution model (Fig. 5). Nine bacterial phyla are represented in this reconstruction, the majority of which are homologs from *Firmicutes* species, and 23% (84 of 358) of the homologs are of eukaryotic origin. Animal, plant, and 38 fungal homologs form a eukaryote monophyletic clade (Fig. 5, clade B). A total of 27 of the remaining 28 fungal homologs form a separate subclade (Fig. 5, clade C) almost entirely comprised of homologs from *Ascomycotes* except for two paralogs from the *Basidiomycota* species *Phlebiopsis gigantea*, while *Batrachochytrium dendrobatidis* is placed within an adjacent *Firmicutes* subclade.

Our maximum-likelihood phylogeny inferred a monophyletic *Phytophthora* subclade with seven homologs from five species (excluding *P. lateralis* and *P. parasitica*) within an alphaproteobacterial *Sphingomonadale* subclade with 100% bootstrap support (Fig. 5, clade A). Homology data for the seed sequence and its adjacent sequences within the *P. capsici* genome from JGI showed no obvious evidence of bacterial contamination at the genomic level, as neither of the flanking genes was bacterial in origin (see Table S4 in the supplemental material).

BLAST homology searches of the seed sequence returned hits from many bacterial alcohol dehydrogenase proteins. Pfam and InterProScan analysis of the seed sequence found that it contained the hallmark signatures of a medium-chain Zn^{2+}-containing alcohol dehydrogenase: an N terminus containing the conserved Zn^{2+} active site, the conserved GroES-like fold, and the NAD(P)-binding Rossmann fold (see Data Set S1 in the supplemental material). Alcohol dehydrogenases (EC 1.1.1.1) catalyze the NAD(P)-dependent reversible oxidation of alcohols to aldehydes or ketones. In most prokaryotes, fungi, and plants, alcohol dehydrogenase is responsible for the reversed regeneration of NAD^+ in fermentation for glycolysis from the reduction of NADH and acetaldehyde to NAD^+ and ethanol. The high concentration of *Firmicutes* and fungal homologs in our reconstruction underlines the enzyme's important role in anaerobic *Clostridia* and fungi. Previous EST analysis of

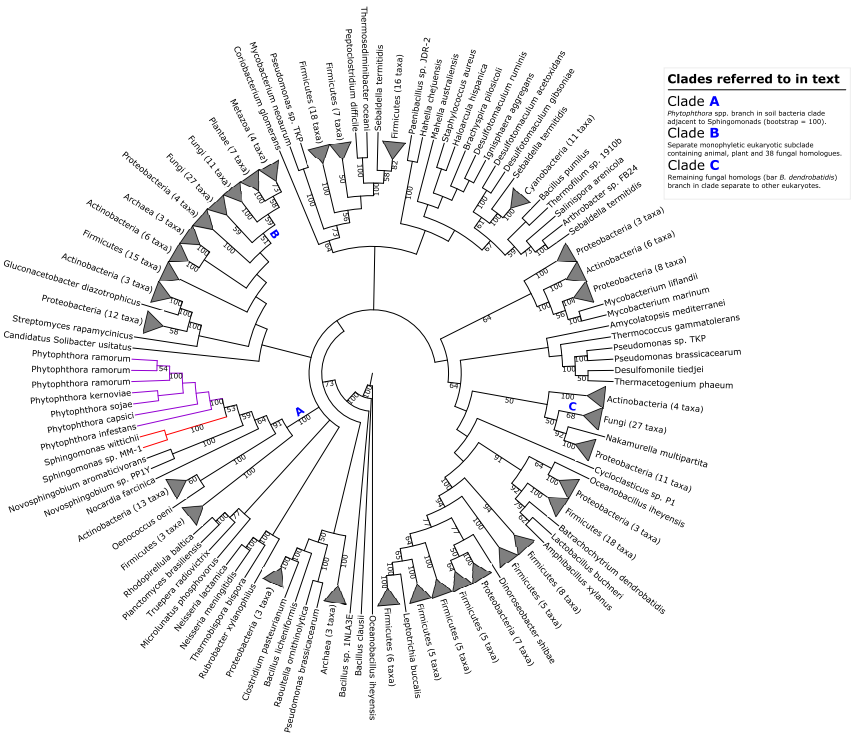


FIG 5 Maximum-likelihood phylogeny illustrating putative transfer of alcohol dehydrogenase from *Sphingomondales* into *Phytophthora* spp. Clades A, B, and C referred to in the main text are highlighted. Selected bootstrap support values are shown at nodes. The corresponding full phylogenetic trees with detailed clades can be viewed in Fig. S4B in the supplemental material.

P. sojae infection of soybean found abundant matches for alcohol dehydrogenase genes, among other intermediary metabolic genes differently expressed in host tissue, suggesting that alcohol fermentation is an important part of the catabolism of *P. sojae* in the early stages of growth inside host tissue (68).

The maximum-likelihood phylogenetic reconstruction performed for these putative *Phytophthora* alcohol dehydrogenase proteins supports the notion of a putative transfer from the alphaproteobacterial *Sphingomonadales* (Fig. 5). Similarly, the phylogeny generated in querying the seed sequence against the NCBI's nonredundant protein database placed the seed sequence within a small *Phytophthora* subclade that was found within a larger *Sphingobium* and *Novosphingobium* clade (not shown). We therefore propose that this alcohol dehydrogenase, found in a number of *Phytophthora* species, arose in these species via recent transfer of the gene from *Sphingomonadales*.

Impact and extent of bacterial genes in oomycete evolution. Using stringent criteria, our analysis has found five putative gene families in oomycete species that have been acquired through horizontal transfer from bacteria. All five transfer events involve

genes coding for proteins with putative enzymatic functions in their respective species; some of our findings, particularly with respect to the putative epoxide hydrolase gene in *Phytophthora capsici*, appear to represent novel evolutions, and some, particularly with respect to the fumarase and alcohol dehydrogenase families, complement those found in other analyses of HGT in oomycete genomes. Many of the inter- and intradomain HGT gene families identified in oomycete genomes to date are proteins with a putative carbohydrate metabolism function (16); in the most extensive study of HGT into oomycete genomes to date, Richards et al. (23) found 13 secreted proteins among the 34 potential fungal events of HGT in oomycetes that could be assigned with such a function. Of the seven bacterial events of HGT identified in oomycete species prior to our analysis (16), most were found in analyses of *Saprolegniales* species (21, 22) and, where function could be assigned, were thought to be involved in carbohydrate metabolism also.

The bacterially derived enzymes identified in oomycete species could have potentially found themselves more amenable to transfer and subsequent retention in oomycete genomes due to their relative low connectivity within a protein-protein interaction network, a significant factor in the influence of the "complexity hypothesis" on HGT (69, 70). The relatively low number of bacterium-oomycete events of HGT identified in this study and elsewhere in the literature, in comparison with other such studies of interdomain HGT, in fungi (8), for example, may be partially explained by the paucity of oomycete genomic data overall and the lack of data for more basal lineages in particular (12). Furthermore, our analysis was designed specifically to identify recent events of HGT in individual plant-parasitic oomycete lineages, as opposed to ancient transfers into the class as a whole or even into the greater stramenopiles group. Future analyses, facilitated by a greater amount of oomycete genomic data, may identify more instances of either bacterium-oomycete HGT to specific lineages or ancient transfers into the class.

Conclusions. Using methods similar to those that have previously identified intradomain HGT between fungi and *Phytophthora* (23), we have identified five interdomain events of HGT between bacteria and plant-pathogenic oomycetes (Table 3). Of the five putative bacterium-oomycete HGT genes that we have identified (Table 3), one has signal peptide signatures and subcellular localization matches that indicate that it is part of the oomycete secretome. The putative SnoaL-like protein may be a secreted transport protein or involved in production of other components of the *Pythium* secretome. A class II fumarase distinct from the endosymbiosis-derived fumarase is present in *Pythium* and *Phytopythium*, and a proteobacterial alcohol dehydrogenase gene is present in multiple *Phytophthora* species (see Table S1 in the supplemental material). The remaining two transferred genes may have more regulatory cytosolic roles in their respective oomycetes species (Table 3), such as regulation of redox activity and neutralization of toxic xenobiotics. Our analysis shows that the transfer of genetic material from bacteria into oomycete lineages is rare but has occurred and that it is another example of cases of HGT between prokaryotes and eukaryotes.

MATERIALS AND METHODS

Data set assembly. The predicted proteomes for seven *Phytophthora* species (*P. capsici*, *P. infestans*, *P. kernoviae*, *P. lateralis*, *P. parasitica*, *P. ramorum*, and *P. sojae*), *Phytopythium vexans*, and six *Pythium* species (*P. aphanidermatum*, *P. arrhenomanes*, *P. irregulare*, *P. iwayami*, *Pythium ultimum* var. *sporangiiferum*, and *P. ultimum* var. *ultimum*) were analyzed for possible bacterium-oomycete HGT events. To ensure a broad taxon sampling for the oomycetes as a whole, we downloaded all available oomycete genome data from public databases. The predicted proteomes of the *Peronosporales* species *Hyaloperonospora arabidopsidis* (71) and *Albugo laibachii* (72); the predicted proteomes of the *Saprolegniales* species *Saprolegnia parasitica* (26), *Saprolegnia diclina*, *Aphanomyces invadans*, and *Aphanomyces astaci* (Broad Institute); and the secretomes of the *Saprolegniales* species *Achyla hypogyna* and *Thraustotheca clavata* (27) were included in our local database. To cover taxon sampling of the stramenopiles, the predicted proteomes of the two diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (29, 73) and of the alga *Aureococcus anophagefferens* (74) were also included. In addition to our oocymete and stramenopile data, our database contained all available nonredundant prokaryotic protein data. To construct this portion and reduce redundancy, a representative genome from each prokaryotic species in the full NCBI GenBank database (75) was included. In total, just under 5 million protein sequences from 1,486 prokaryotic genomes were retained. More than 3 million sequences from 212 eukaryotic nuclear genomes, sampling a diverse range of animal, plant, and fungal lineages, were included (see Data Set S1 in the supplemental material).

TABLE 4 Identification of sequences with high bacterial homology corresponding to candidate events of HGT within oomycete genomes

Proteome	Initial size (no. of genes)	Size after Repbase filtering (no. of genes)	No. of OrthoMCL clusters (no. of sequences)	No. of OrthoMCL unclustered sequences	No. of intergenic bacterial hits
Phytophthora capsici	19,805	16,169	1,732 (8,879)	7,290	6
Phytophthora infestans	18,140	17,013	2,032 (9,459)	7,553	2
Phytophthora kernoviae	10,650	10,435	750 (3,244)	7,016	0
Phytophthora lateralis	11,635	10,539	880 (4,110)	6,337	14
Phytophthora parasitica	20,822	18,640	2,084 (10,153)	8,437	2
Phytophthora ramorum	15,743	13,403	1,639 (7,839)	5,564	5
Phytophthora sojae	26,584	22,210	2,418 (13,544)	8,666	2
Phytopythium vexans	11,958	11,634	1,097 (4,932)	6,702	7
Pythium aphanidermatum	12,312	12,002	1,144 (5,129)	6,873	11
Pythium arrhenomanes	13,805	13,224	1,221 (5,647)	7,577	18
Pythium irregulare	13,805	13,297	1,214 (5,888)	7,409	6
Pythium iwayami	14,875	14,279	1,303 (6,185)	8,094	6
Pythium ultimum var. *sporangiiferum*	14,096	13,915	917 (4,208)	9,707	13
Pythium ultimum var. *ultimum*	15,323	14,780	1,305 (6,661)	8,119	14

Identification of putative bacterium-oomycete HGT events. Our methods for identifying candidate bacterial HGT genes followed those of Richards et al. (23) in their analysis of fungal HGT genes in the oomycetes. Repetitive and transposable elements were identified and removed from each *Phytophthora* and *Phytopythium*/*Pythium* proteome by performing homology searches against Repbase (76) by the use of tBLASTn (77, 34) with an E value cutoff of 10^{-20} (Table 4). The remaining protein sequences in each oomycete proteome were then further filtered and clustered into groups of paralogs using OrthoMCL (33), with an E value cutoff of 10^{-20} and an inflation value of 1.5 (Table 4). Representative sequences from each group of paralogs, along with unclustered singleton sequences, were retrieved from their respective proteomes. These sequences were then queried against our local database using BLASTp with an E value cutoff of 10^{-20}.

Using bespoke python scripting, we identified 106 genes whose homology supported a bacterial transfer into an individual oomycete lineage (encoding proteins whose first hit outside their own genus was bacterial) and retrieved them for a second round of OrthoMCL clustering to remove redundancy in our datasets for each genus (Table 4). All retrieved protein sequences were clustered into groups of orthologs using OrthoMCL with an E value cutoff of 10^{-20} and an inflation value of 1.5 (Table 2). A total of 64 representative and singleton sequences from these datasets were then queried against our local database using BLASTp with an E value cutoff of 10^{-20} and an arbitrary limit for maximum hits per query sequence. The corresponding gene family for each candidate HGT gene was constructed from our BLASTp results.

Phylogenetic reconstruction of putative bacterium-oomycete HGT events. A total of 64 candidate HGT gene families were aligned using MUSCLE (78), and best-fit amino acid replacement models were selected for each alignment using ProtTest (38). Maximum-likelihood phylogenetic reconstruction for each alignment was carried out using PhyML (79) with 100 bootstrap replicates. Each phylogenetic tree was visualized and annotated with GenBank data using bespoke python scripting and iTOL (80). Additional phylogenetic analysis using consensus network methods was carried out using SplitsTree (57).

Analysis of bacterial contamination and taxon sampling in putative bacterium-oomycete HGT families. Seed genes and their directly adjacent gene were examined for their particular homology to determine whether candidate HGT genes were not simply the result of bacterial contamination of genomes along particular contigs or scaffolds. For each seed gene arising from *P. capsici*, the genomic location of that gene was identified by querying its corresponding protein sequence against the JGI *P. capsici* database (http://genome.jgi.doe.gov/PhycaF7) using tBLASTn with an E value cutoff of 10^{-4}. Homology data for each seed gene and their adjacent genes were provided by the JGI *P. capsici* genome browser (see Table S4 in the supplemental material). For each *Pythium* seed gene, the genomic location of the gene was identified by querying the corresponding protein sequence against the genomic scaffolds of the source species using tBLASTn with an E value cutoff of 10^{-4}, and then the seed gene's corresponding protein sequence and its two adjacent protein sequences were queried against the NCBI's nonredundant protein sequence database using BLASTp with an E value cutoff of 10^{-20} (see Table S4).

For studies of HGT in eukaryotes, particularly transfer between prokaryotes and eukaryotes, it is essential that genomic data cover as broad a range of taxa as possible to prevent as much as possible the introduction of bias into analysis and thus reduce the likelihood of obtaining false transfer events (36, 37). Comparison of the taxon sampling in our database with the NCBI data was performed by searching each seed gene's protein sequence against the NCBI nonredundant protein sequence database using BLASTp with an E value cutoff of 10^{-20}. The seed sequence and its homologs were aligned in MUSCLE, neighbor-joining trees were constructed in QuickTree (81) using 100 bootstrap replicates, and each tree was annotated with GenBank data using bespoke python scripting (not shown). Maximum-likelihood HGT phylogenies whose topology conflicted substantially with their corresponding neighbor-joining tree due to differences in taxon sampling were excluded from further analysis.

Characterization and functional annotation of putative bacterium-oomycete HGT families. For the remaining putative HGT families, bespoke python scripting was used to calculate the sequence length, GC content, and exon number of each oomycete gene present. The average sequence length, GC content, and exon number for each *Phytophthora*, *Phytopythium*, and *Pythium* genome were also calculated (data not shown). Multivariate codon usage analysis of each genome was carried out using GCUA (82) (see Fig. S2 in the supplemental material). To compare the properties of each putative HGT family with those of homologs in their potential bacterial donor, multivariate codon usage analysis of the genome of a representative potential donor as well as the relevant seed oomycete gene was also carried out using GCUA. Additionally, the sequence length and GC content of one or more bacterial sister genes were calculated using bespoke python scripting (see Table S2). Optimal local alignments of each seed protein sequence against a representative bacterial sister gene were generated using CLUSTAL Omega (83) (see Table S3). The putative function of each putative HGT family was annotated by performing initial Pfam homology searches of each seed protein sequence (84) (see Data Set S1) with an E value cutoff of 10^{-4} and BLAST homology searches against the NCBI's nonredundant protein database with an E value cutoff of 10^{-20}. To complement these initial annotations, each seed protein sequence was then analyzed in InterProScan (85). Signal peptide analysis and subcellular localization prediction analysis for each seed protein sequence were carried out using SignalP and TargetP, respectively (53, 54), with the default parameters.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://dx.doi.org/10.1128/mSphere.00195-16.

Figure S1, PDF file, 0.2 MB.
Figure S2, PDF file, 0.05 MB.
Figure S3, PDF file, 0.03 MB.
Figure S4, PDF file, 7.6 MB.
Figure S5, PDF file, 0.7 MB.
Table S1, DOCX file, 0.02 MB.
Table S2, DOCX file, 0.02 MB.
Table S3, DOCX file, 0.01 MB.
Table S4, DOCX file, 0.02 MB.
Data Set S1, XLSX file, 0.1 MB.

FUNDING INFORMATION

This work, including the efforts of Charley G. P. McCarthy, was funded by Irish Research Council (GOIPG/2015/2243).

REFERENCES

1. **Goldenfeld N, Woese C**. 2007. Biology's next revolution. Nature **445:** 369. http://dx.doi.org/10.1038/445369a.
2. **Dagan T, Artzy-Randrup Y, Martin W**. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A **105:**10039–10044. http://dx.doi.org/10.1073/pnas.0800679105.
3. **Keeling PJ, Palmer JD**. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet **9:**605–618. http://dx.doi.org/10.1038/nrg2386.
4. **Soucy SM, Huang J, Gogarten JP**. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet **16:**472–482. http://dx.doi.org/10.1038/nrg3962.
5. **Dunning Hotopp JC**. 2011. Horizontal gene transfer between bacteria and animals. Trends Genet **27:**157–163. http://dx.doi.org/10.1016/j.tig.2011.01.005.
6. **Hirt RP, Alsmark C, Embley TM**. 2015. Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. Curr Opin Microbiol **23:**155–162. http://dx.doi.org/10.1016/j.mib.2014.11.018.
7. **Alsmark C, Foster PG, Sicheritz-Ponten T, Nakjang S, Embley MT, Hirt RP**. 2013. Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. Genome Biol **14:**R19. http://dx.doi.org/10.1186/gb-2013-14-2-r19.
8. **Marcet-Houben M, Gabaldón T**. 2010. Acquisition of prokaryotic genes by fungal genomes. Trends Genet **26:**5–8. http://dx.doi.org/10.1016/j.tig.2009.11.007.
9. **Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, Disbennett K, Pfannkoch C, Sumin N, Sutton GG, Viswanathan LD, Walenz B, Goodstein DM, Hellsten U, Kawashima T, Prochnik SE, Putnam NH, Shu S, Blumberg B, Dana CE, Gee L, Kibler DF, Law L, Lindgens D, Martinez DE, Peng J, Wigge PA, Bertulat B, Guder C, Nakamura Y, Ozbek S, Watanabe H, Khalturin K, Hemmrich G, Franke A, Augustin R, Fraune S, Hayakawa E, Hayakawa S, Hirose M, Hwang JS, Ikeo K, Nishimiya-Fujisawa C, Ogura A, Takahashi T, Steinmetz PRH, et al.** 2010. The dynamic genome of Hydra. Nature **464:**592–596. http://dx.doi.org/10.1038/nature08830.
10. **Danchin EG, Rosso M-N, Vieira P, de Almeida-Engler J, Coutinho PM, Henrissat B, Abad P**. 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc Natl Acad Sci U S A **107:**17651–17656. http://dx.doi.org/10.1073/pnas.1008486107.
11. **Burki F**. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect J Biol **6:**a016147. http://dx.doi.org/10.1101/cshperspecta016147.
12. **Beakes GW, Glockling SL, Sekimoto S**. 2012. The evolutionary phylogeny of the oomycete "fungi". Protoplasma **249:**3–19. http://dx.doi.org/10.1007/s00709-011-0269-2.
13. **Burki F, Shalchian-Tabrizi K, Minge M, Skjæveland A, Nikolaev SI, Jakobsen KS, Pawlowski J**. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One **2**.
14. **Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ**. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A **106:**3859–3864. http://dx.doi.org/10.1073/pnas.0807880106.
15. **Gaston D, Roger AJ**. 2013. Functional divergence and convergent evolution in the plastid-targeted glyceraldehyde-3-phosphate dehydrogenases of diverse eukaryotic algae. PLoS One **8:**e70396. http://dx.doi.org/10.1371/journal.pone.0070396.
16. **Shalchian-Tabrizi K, Kauserud H, Massana R, Klaveness D, Jakobsen KS**. 2007. Analysis of environmental 18S ribosomal RNA sequences reveals unknown diversity of the Cosmopolitan phylum Telonemia. Protistologica **158:**173–180. http://dx.doi.org/10.1016/j.protis.2006.10.003.
17. **Jiang RH, Tyler BM**. 2012. Mechanisms and evolution of virulence in

oomycetes. Annu Rev Phytopathol **50:**295–318. http://dx.doi.org/10.1146/annurev-phyto-081211-172912.
18. **Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Lévesque CA, Buell CR**. 2013. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. PLoS One **8:**e75072. http://dx.doi.org/10.1371/journal.pone.0075072.
19. **Lévesque CA.de Cock AW**. 2004. Molecular phylogeny & taxonomy of the genus Pythium. Mycol Res **108**(Pt 12):1363–1383.
20. **De Cock AW, Lodhi AM, Rintoul TL, Bala K, Robideau GP, Abad ZG, Coffey MD, Shahzad S, Lévesque CA**. 2015. Phytopythium: molecular phylogeny and systematics. Persoonia **34:**25–39. http://dx.doi.org/10.3767/003158515X685382.
21. **Savory F, Leonard G, Richards TA**. 2015. The role of horizontal gene transfer in the evolution of the oomycetes. PLoS Pathog **11:**e1004805. http://dx.doi.org/10.1371/journal.ppat.1004805.
22. **Judelson HS**. 2012. Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. Eukaryot Cell **11:**1304–1312. http://dx.doi.org/10.1128/EC.00155-12.
23. **Richards TA, Soanes DM, Jones MD, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ**. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci U S A **108:**15258–15263. http://dx.doi.org/10.1073/pnas.1105100108.
24. **Richards TA, Dacks JB, Jenkinson JM, Thornton CR, Talbot NJ**. 2006. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. Curr Biol **16:**1857–1864. http://dx.doi.org/10.1016/j.cub.2006.07.052.
25. **Belbahri L, Calmin G, Mauch F, Andersson JO**. 2008. Evolution of the cutinase gene family: evidence for lateral gene transfer of a candidate Phytophthora virulence factor. Gene **408:**1–8.
26. **Jiang RH, de Bruijn I, Haas BJ, Belmonte R, Löbach L, Christie J, van den Ackerveken G, Bottin A, Bulone V, Díaz-Moreno SM, Dumas B, Fan L, Gaulin E, Govers F, Grenville-Briggs LJ, Horner NR, Levin JZ, Mammella M, Meijer HJ, Morris P, Nusbaum C, Oome S, Phillips AJ, van Rooyen D, Rzeszutek E, Saraiva M, Secombes CJ, Seidl MF, Snel B, Stassen JHM, Sykes S, Tripathy S, van den Berg H, Vega-Arreguin JC, Wawra S, Young SK, Zeng Q, Dieguez-Uribeondo J, Russ C, Tyler BM, van West P**. 2013. Distinctive expansion of potential virulence genes in the genome of the oomycete fish pathogen Saprolegnia parasitica. PLoS Genet **9:**e1003272. http://dx.doi.org/10.1371/journal.pgen.1003272.
27. **Misner I, Blouin N, Leonard G, Richards TA, Lane CE**. 2015. The secreted proteins of Achlya hypogyna and Thraustotheca clavata identify the ancestral oomycete secretome and reveal gene acquisitions by horizontal gene transfer. Genome Biol Evol **7:**120–135. http://dx.doi.org/10.1093/gbe/evu276.
28. **Whitaker JW, McConkey GA, Westhead DR**. 2009. The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. Genome Biol **10:**R36. http://dx.doi.org/10.1186/gb-2009-10-4-r36.
29. **Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret J-P, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kröger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jézéquel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, et al.** 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature **456:**239–244. http://dx.doi.org/10.1038/nature07410.
30. **Morris PF, Schlosser LR, Onasch KD, Wittenschlaeger T, Austin R, Provart N**. 2009. Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. PLoS One **4:**e6133. http://dx.doi.org/10.1371/journal.pone.0006133.
31. **Martens C, Vandepoele K, Van de Peer Y**. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. Proc Natl Acad Sci U S A **105:**3427–3432. http://dx.doi.org/10.1073/pnas.0712248105.
32. **Nosenko T, Bhattacharya D**. 2007. Horizontal gene transfer in chromalveolates. BMC Evol Biol **7:**173. http://dx.doi.org/10.1186/1471-2148-7-173.
33. **Li L, Stoeckert CJ, Roos DS**. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:**2178–2189. http://dx.doi.org/10.1101/gr.1224503.
34. **Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ**. 1997. Gapped BLAST and psi-blast: a new generation of protein database search programs. Nucleic Acids Res **25:**3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.
35. **Guindon S, Gascuel O**. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696–704. http://dx.doi.org/10.1080/10635150390235520.
36. **Gluck-Thaler E, Slot JC**. 2015. Dimensions of horizontal gene transfer in eukaryotic microbial pathogens. PLoS Pathog **11:**e1005156. http://dx.doi.org/10.1371/journal.ppat.1005156.
37. **Huang J**. 2013. Horizontal gene transfer in eukaryotes: the weak-link model. BioEssays **35:**868–875. http://dx.doi.org/10.1002/bies.201300007.
38. **Darriba D, Taboada GL, Doallo R, Posada D**. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics **27:**1164–1165. http://dx.doi.org/10.1093/bioinformatics/btr088.
39. **Koski LB, Morton RA, Golding GB**. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol **18:**404–412. http://dx.doi.org/10.1093/oxfordjournals.molbev.a003816.
40. **Yogev O, Yogev O, Singer E, Shaulian E, Goldberg M, Fox TD, Pines O**. 2010. Fumarase: a mitochondrial metabolic enzyme and a cytosolic/nuclear component of the DNA damage response. PLoS Biol **8:**e1000328. http://dx.doi.org/10.1371/journal.pbio.1000328.
41. **Jiang Y, Qian X, Shen J, Wang Y, Li X, Liu R, Xia Y, Chen Q, Peng G, Lin S-Y, Lu Z**. 2015. Local generation of fumarate promotes DNA repair through inhibition of histone H3 demethylation. Nat Cell Biol **17:**1158–1168. http://dx.doi.org/10.1038/ncb3209.
42. **Estévez M, Skarda J, Spencer J, Banaszak L, Weaver TM**. 2002. X-ray crystallographic and kinetic correlation of a clinically observed human fumarase mutation. Protein Sci **11:**1552–1557. http://dx.doi.org/10.1110/ps.0201502.
43. **Gerbod D, Edgcomb VP, Noël C, Vanácová S, Wintjens R, Tachezy J, Sogin ML, Viscogliosi E**. 2001. Phylogenetic relationships of class II fumarase genes from trichomonad species. Mol Biol Evol **18:**1574–1584. http://dx.doi.org/10.1093/oxfordjournals.molbev.a003944.
44. **Timmis JN, Ayliffe MA, Huang CY, Martin W**. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet **5:**123–135. http://dx.doi.org/10.1038/nrg1271.
45. **Pieretti I, Pesic A, Petras D, Royer M, Süssmuth RD, Cociancich S**. 2015. What makes Xanthomonas albilineans unique amongst xanthomonads? Front Plant Sci **6:**289. http://dx.doi.org/10.3389/fpls.2015.00289.
46. **Stammers DK, Ren J, Leslie K, Nichols CE, Lamb HK, Cocklin S, Dodds A, Hawkins AR**. 2001. The structure of the negative transcriptional regulator NmrA reveals a structural superfamily which includes the short-chain dehydrogenase/reductases. EMBO J **20:**6619–6626. http://dx.doi.org/10.1093/emboj/20.23.6619.
47. **Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A**. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res **13:**2129–2141. http://dx.doi.org/10.1101/gr.772403.
48. **Kim I-K, Yim H-S, Kim M-K, Kim D-W, Kim Y-M, Cha S-S, Kang S-O**. 2008. Crystal structure of a new type of NADPH-dependent quinone oxidoreductase (QOR2) from Escherichia coli. J Mol Biol **379:**372–384. http://dx.doi.org/10.1016/j.jmb.2008.04.003.
49. **Hoffman MT, Gunatilaka MK, Wijeratne K, Gunatilaka L, Arnold AE**. 2013. Endohyphal bacterium enhances production of indole-3-acetic acid by a foliar fungal endophyte. PLoS One **8:**e73132. http://dx.doi.org/10.1371/journal.pone.0073132.
50. **Hoffman MT, Arnold AE**. 2010. Diverse bacteria inhabit living hyphae of phylogenetically diverse fungal endophytes. Appl Environ Microbiol **76:**4063–4075. http://dx.doi.org/10.1128/AEM.02928-09.
51. **Strohl WR**. 2001. Biochemical engineering of natural product biosynthesis pathways. Metab Eng **3:**4–14. http://dx.doi.org/10.1006/mben.2000.0172.
52. **Sultana A, Kallio P, Jansson A, Wang J-S, Niemi J, Mäntsälä P, Schneider G**. 2004. Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation. EMBO J **23:**1911–1921. http://dx.doi.org/10.1038/sj.emboj.7600201.
53. **Petersen TN, Brunak S, von Heijne G, Nielsen H**. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods **8:**785–786. http://dx.doi.org/10.1038/nmeth.1701.

54. **Emanuelsson O, Nielsen H, Brunak S, von Heijne G**. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300:**1005–1016. http://dx.doi.org/10.1006/jmbi.2000.3903.
55. **Soanes DM, Richards TA, Talbot NJ**. 2007. Insights from sequencing fungal and oomycete genomes: what can we learn about plant disease and the evolution of pathogenicity? Plant Cell **19:**3318–3326. http://dx.doi.org/10.1105/tpc.107.056663.
56. **Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CM, Dorrance AE, Dou D, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee M-K, McDonald WH, Medina M, Meijer HJG, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JKC, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BWS, Terry A, Torto-Alalibo TA, Win J, Xu Z, Zhang H, et al.** 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science **313:**1261–1266. http://dx.doi.org/10.1126/science.1128796.
57. **Huson DH, Bryant D**. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267. http://dx.doi.org/10.1093/molbev/msj030.
58. **Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Romá-Mateo C, Theodosiou A, Mitchell AL**. 2012. The PRINTS database: a fine-grained protein sequence annotation and analysis resource-its status in 2012. Database (Oxford) 2012:bas019. http://dx.doi.org/10.1093/database/bas019.
59. **Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A**. 1992. The alpha/beta hydrolase fold. Protein Eng **5:**197–211. http://dx.doi.org/10.1093/protein/5.3.197.
60. **Van der Werf MJ, Overkamp KM, de Bont JA**. 1998. Limonene-1,2-epoxide hydrolase from Rhodococcus erythropolis DCL14 belongs to a novel class of epoxide hydrolases. J Bacteriol **180:**5052–5057.
61. **Van Loo B, Kingma J, Arand M, Wubbolts MG, Janssen DB**. 2006. Diversity and biocatalytic potential of epoxide hydrolases identified by genome analysis. Appl Environ Microbiol **72:**2905–2917. http://dx.doi.org/10.1128/AEM.72.4.2905-2917.2006.
62. **Janssen DB**. 2004. Evolving haloalkane dehalogenases. Curr Opin Chem Biol **8:**150–159. http://dx.doi.org/10.1016/j.cbpa.2004.02.012.
63. **Eevers N, Van Hamme JD, Bottos EM, Weyens N, Vangronsveld J**. 2015. Draft genome sequence of Methylobacterium radiotolerans, a DDE-degrading and plant growth-promoting strain isolated from Cucurbita pepo. Genome Announc **3:**e00488–15. http://dx.doi.org/10.1128/genomeA.00488-15.
64. **Thomas JE, Ou LT, All-Agely A**. 2008. DDE remediation and degradation. Rev Environ Contam Toxicol **194:**55–69. http://dx.doi.org/10.1007/978-0-387-74816-0_3.
65. **Moktali V, Park J, Fedorova-Abrams ND, Park B, Lee Y-H, Choi J, Kang S**. 2012. Systematic and searchable classification of cytochrome P450 proteins encoded by fungal and oomycete genomes. BMC Genomics **13:**525. http://dx.doi.org/10.1186/1471-2164-13-525.
66. **Sello MM, Jafta N, Nelson DR, Chen W, Yu J-H, Parvez M, Kgosiemang IK, Monyaki R, Raselemane SC, Qhanya LB, Mthakathi NT, Sitheni Mashele S, Syed K**. 2015. Diversity and evolution of cytochrome P450 monooxygenases in oomycetes. Sci Rep **5:**11572. http://dx.doi.org/10.1038/srep11572.
67. **Barabote RD, Thekkiniath J, Strauss RE, Vediyappan G, Fralick JA, San Francisco MJ**. 2011. Xenobiotic efflux in bacteria and fungi: a genomics update. Adv Enzymol Relat areas. Mol Biol **77:**237–306.
68. **Qutob D, Hraber PT, Sobral BW, Gijzen M**. 2000. Comparative analysis of expressed sequences in phytophthora sojae. Plant Physiol **123:**243–254. http://dx.doi.org/10.1104/pp.123.1.243.
69. **Jain R, Rivera MC, Lake JA**. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A **96:**3801–3806. http://dx.doi.org/10.1073/pnas.96.7.3801.
70. **Cohen O, Gophna U, Pupko T**. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol **28:**1481–1489. http://dx.doi.org/10.1093/molbev/msq333.
71. **Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, Kemen E, Thines M, Ah-Fong A, Anderson R, Badejoko W, Bittner-Eddy P, Boore JL, Chibucos MC, Coates M, Dehal P, Delehaunty K, Dong S, Downton P, Dumas B, Fabro G, Fronick C, Fuerstenberg SI, Fulton L, Gaulin E, Govers F, Hughes L, Humphray S, Jiang RHY, Judelson H, Kamoun S, Kyung K, Meijer H, Minx P, Morris P, Nelson J, Phuntumart V, Qutob D, Rehmany A, Rougon-Cardoso A, Ryden P, Torto-Alalibo T, Studholme D, Wang Y, Win J, Wood J, Clifton SW, Rogers J, Van den Ackerveken G, Jones JDG, et al.** 2010. Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. Science **330:**1549–1551. http://dx.doi.org/10.1126/science.1195203.
72. **Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, Robert-Seilaniantz A, Bailey K, Holub E, Studholme DJ, Maclean D, Jones JD**. 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of Arabidopsis thaliana. PLoS Biol **9:**e1001094. http://dx.doi.org/10.1371/journal.pbio.1001094.
73. **Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kröger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS**. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science **306:**79–86. http://dx.doi.org/10.1126/science.1101156.
74. **Gobler CJ, Berry DL, Dyhrman ST, Wilhelm SW, Salamov A, Lobanov AV, Zhang Y, Collier JL, Wurch LL, Kustka AB, Dill BD, Shah M, VerBerkmoes NC, Kuo A, Terry A, Pangilinan J, Lindquist EA, Lucas S, Paulsen IT, Hattenrath-Lehmann TK, Talmage SC, Walker EA, Koch F, Burson AM, Marcoval MA, Tang Y-Z, Lecleir GR, Coyne KJ, Berg GM, Bertrand EM, Saito MA, Gladyshev VN, Grigoriev IV**. 2011. Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. Proc Natl Acad Sci U S A **108:**4352–4357. http://dx.doi.org/10.1073/pnas.1016106108.
75. **Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW**. 2013. GenBank. Nucleic Acids Res **41:**D36–D42. http://dx.doi.org/10.1093/nar/gks1195.
76. **Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J**. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110:**462–467. http://dx.doi.org/10.1159/000084979.
77. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421. http://dx.doi.org/10.1186/1471-2105-10-421.
78. **Edgar RC**. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:**113. http://dx.doi.org/10.1186/1471-2105-5-113.
79. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O**. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59:**307–321. http://dx.doi.org/10.1093/sysbio/syq010.
80. **Letunic I, Bork P**. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res **44:**W242–W245. http://dx.doi.org/10.1093/nar/gkw290.
81. **Howe K, Bateman A, Durbin R**. 2002. QuickTree: building huge neighbour-joining trees of protein sequences. Bioinformatics **18:**1546–1547. http://dx.doi.org/10.1093/bioinformatics/18.11.1546.
82. **McInerney JO**. 1998. GCUA: general codon usage analysis. Bioinformatics **14:**372–373. http://dx.doi.org/10.1093/bioinformatics/14.4.372.
83. **Rice P, Longden I, Bleasby A**. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16:**276–277. http://dx.doi.org/10.1016/S0168-9525(00)02024-2.
84. **Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A**. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44:**D279–D285. http://dx.doi.org/10.1093/nar/gkv1344.
85. **Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S**. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics **30:**1236–1240. http://dx.doi.org/10.1093/bioinformatics/btu031.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE
Ecological and Evolutionary Science

Phylogenomic Reconstruction of the Oomycete Phylogeny Derived from 37 Genomes

Charley G. P. McCarthy, David A. Fitzpatrick

Department of Biology, Genome Evolution Laboratory, Maynooth University, Maynooth, Co. Kildare, Ireland

ABSTRACT The oomycetes are a class of microscopic, filamentous eukaryotes within the *Stramenopiles-Alveolata-Rhizaria* (SAR) supergroup which includes ecologically significant animal and plant pathogens, most infamously the causative agent of potato blight *Phytophthora infestans*. Single-gene and concatenated phylogenetic studies both of individual oomycete genera and of members of the larger class have resulted in conflicting conclusions concerning species phylogenies within the oomycetes, particularly for the large *Phytophthora* genus. Genome-scale phylogenetic studies have successfully resolved many eukaryotic relationships by using supertree methods, which combine large numbers of potentially disparate trees to determine evolutionary relationships that cannot be inferred from individual phylogenies alone. With a sufficient amount of genomic data now available, we have undertaken the first whole-genome phylogenetic analysis of the oomycetes using data from 37 oomycete species and 6 SAR species. In our analysis, we used established supertree methods to generate phylogenies from 8,355 homologous oomycete and SAR gene families and have complemented those analyses with both phylogenomic network and concatenated supermatrix analyses. Our results show that a genome-scale approach to oomycete phylogeny resolves oomycete classes and individual clades within the problematic *Phytophthora* genus. Support for the resolution of the inferred relationships between individual *Phytophthora* clades varies depending on the methodology used. Our analysis represents an important first step in large-scale phylogenomic analysis of the oomycetes.

IMPORTANCE The oomycetes are a class of eukaryotes and include ecologically significant animal and plant pathogens. Single-gene and multigene phylogenetic studies of individual oomycete genera and of members of the larger classes have resulted in conflicting conclusions concerning interspecies relationships among these species, particularly for the *Phytophthora* genus. The onset of next-generation sequencing techniques now means that a wealth of oomycete genomic data is available. For the first time, we have used genome-scale phylogenetic methods to resolve oomycete phylogenetic relationships. We used supertree methods to generate single-gene and multigene species phylogenies. Overall, our supertree analyses utilized phylogenetic data from 8,355 oomycete gene families. We have also complemented our analyses with superalignment phylogenies derived from 131 single-copy ubiquitous gene families. Our results show that a genome-scale approach to oomycete phylogeny resolves oomycete classes and clades. Our analysis represents an important first step in large-scale phylogenomic analysis of the oomycetes.

KEYWORDS oomycete, phylogeny, *Phytophthora*, species phylogeny, phylogenomics, supermatrix, supertrees

Received 24 February 2017 **Accepted** 24 March 2017 **Published** 12 April 2017

Citation McCarthy CGP, Fitzpatrick DA. 2017. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. mSphere 2:e00095-17. https://doi.org/10.1128/mSphere.00095-17.

Editor Aaron P. Mitchell, Carnegie Mellon University

Address correspondence to David A. Fitzpatrick, david.fitzpatrick@nuim.ie.

The oomycetes are a class of microscopic eukaryotes which include some of the most ecologically destructive marine and terrestrial eukaryotic species (1). Oomycete species display filamentous morphology and ecological roles very similar to those of fungi and were historically regarded as a basal fungal lineage (2). As morphological and molecular studies have improved since the latter half of the 20th century, the oomycetes have come to be understood as very distant relations of "true" fungi. They have independently evolved similar morphology and lifestyles through convergent evolution and limited interkingdom horizontal gene transfer (HGT) (2–5). Present phylogenomic studies place the oomycetes in the diverse stramenopiles lineage within the *Stramenopiles-Alveolata-Rhizaria* (SAR) eukaryotic supergroup (6–10) (Fig. 1). The stramenopiles were previously placed within *Chromista* (11) and then within the "chromalveolates" supergroup (*Chromista* plus *Alveolata*) on the basis of a hypothesized last common ancestor on the plastid lineage (12, 13). While early phylogenetic analyses supported the concept of a single origin for the "chromalveolate" plastid (14, 15), later plastome-wide and nuclear phylogenetic and HGT analyses have consistently failed to support a monophyletic chromalevolate grouping (16–21). In contrast, molecular evidence for the monophyly of the current SAR supergroup has been demonstrated in multiple phylogenetic analyses (18, 20, 22–26).

The oomycetes are thought to have diverged from diatoms between the Late Proterozoic and the mid-Paleozoic eras (~0.4 to 0.6 billion years ago [bya]) (27, 28) and have been found to have been present as early as the Devonian period (~400 million years ago [mya]) in the fossil record (29). Though many described species are phytopathogens, oomycete phytopathogenicity is thought to be a derived trait which has evolved independently in many lineages (30). Many species are as yet unsampled, and the class phylogeny of the oomycetes is still subject to revision; with current data, however, the oomycetes can be split into the earliest diverging clades and the later "crown" taxa (31–33) (Fig. 1). With the exception of some species infecting terrestrial nematodes (31), the earliest diverging oomycete clades are otherwise exclusively marine in habitat (1). The remaining "crown" oomycetes can be subdivided into the predominantly marine and freshwater "saprolegnian" branches and the predominantly terrestrial "peronosporalean" branches, which diverged in the Early Mesozoic era (1, 28, 34–36). The "saprolegnian" branches include the fish pathogen *Saprolegnia*, also known as "cotton mould" (37), and the animal- and plant-pathogenic *Aphanomyces* genus (34, 38). The "peronosporalean" branches include the best-characterized oomycete taxa, *Phytophthora* and *Pythium*, and the more basal *Albuginales* order (1, 35). The majority of "peronosporalean" oomycetes are phytopathogens, although *Pythium* includes species capable of infecting animals or acting as mycoparasitic biocontrol agents (39, 40) (Fig. 1).

Phytophthora is the largest genus (>120 described species) within the order *Peronosporales* and was divided into 10 phylogenetic clades on the basis of initial internal transcribed spacer (ITS) analysis and, later, combined nuclear and mitochondrial analyses (41, 42) (Fig. 2a). The largest clades (clades 1, 2, 7, and 8) are further divided into subclades, while the smallest clades (clades 5 and 10) contain fewer than five described species at present (43, 44). Initial ITS phylogeny data reported by Cooke et al. (41) suggested that *Phytophthora* was paraphyletic with respect to basal clades 9 and 10; however, later multigene and combined nuclear and mitochondrial studies have placed these clades within *Phytophthora* (42, 44, 45). Generally, species within *Phytophthora* clades do not share consistent morphological features or reproductive strategies, although clades 6 to 8 form a distinct branch of terrestrial species with predominantly nonpapillate sporangia within the genus tree (44). While many recent phylogenetic analyses have supported the current designation by Blair et al. (42) of 10 distinct phylogenetic clades within *Phytophthora*, many of the same analyses draw conflicting conclusions as to the relationships among these clades. In their analysis, Blair et al. (42) found strong support by maximum-likelihood, maximum-parsimony, and Bayesian methods for the 10 phylogenetic clades using data from seven highly conserved nuclear loci (including markers from 28S ribosomal DNA [rDNA], Hsp90, and β-tubulin)

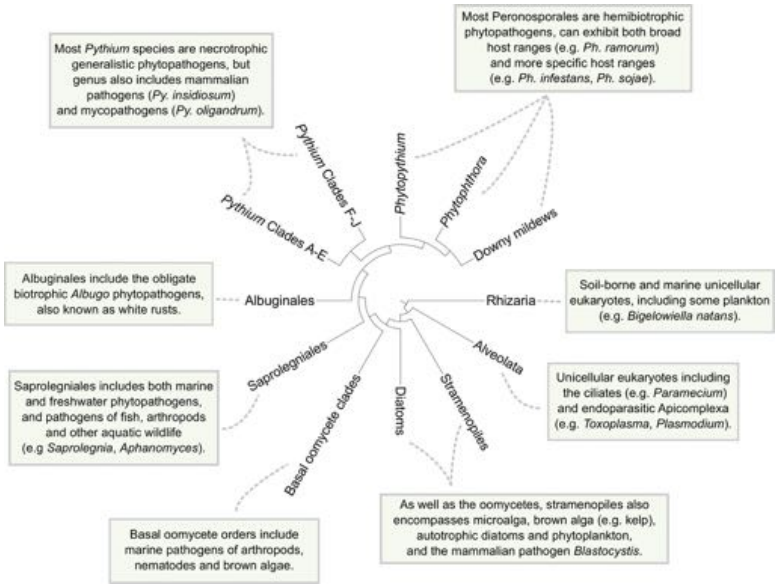


FIG 1 Consensus phylogeny of the oomycetes class within the greater SAR grouping, including information pertaining to various taxa. The cladogram was adapted from Judelson (10).

from 82 *Phytophthora* species (Fig. 2a). The relationship between the clades reported in Blair et al. (42) was mostly upheld in a follow-up analysis by Runge et al. (46) which included homologous data from an additional 39 *Phytophthora* species and other *Peronosporales* species. One noticeable difference was that their analysis placed clades 3, 6, and 7 as sister clades within a monophyletic clade with strong support by the minimum-evolution, maximum-likelihood, and Bayesian methods, while the clades were more distantly related in the analysis by Blair et al. (42) (Fig. 2a and b). The addition of four mitochondrial markers (*cox2*, *nad9*, *rps10*, and *secY*) in a later 11-locus analysis by Martin et al. (47), while topologically supporting the data from Blair et al. (42), displayed poor resolution for many interclade relationships (particularly for more extensively derived clades such as clades 1 to 5) within *Phytophthora* by the maximum-likelihood, maximum-parsimony, and Bayesian methods (Fig. 2c). A coalescent approach using a similar data set by the same authors showed improved Bayesian support among some *Phytophthora* clades (e.g., clades 1 to 5) but weaker support for other clades and a conflicting topology from the 11-locus analysis (47) (Fig. 2d).

Placement of other taxa within the *Peronosporales* order, namely, the "downy mildews," and the phylogeny of *Pythium* and the *Pythiales* order have also been difficult to resolve. The inclusion of two downy mildews species (*Hyaloperonospora arabidopsidis* and *Pseudoperonospora cubensis*) in an analysis conducted by Runge et al. placed the two species within *Phytophthora* clade 4 and sister to clade 1 species such as *Phytophthora infestans*, implying the existence of a paraphyletic *Phytophthora* genus (46) (Fig. 2b). However, a subsequent tree reconciliation analysis, inferred using a class phylogeny of 189 oomycete clusters of orthologous groups (COGs), placed *H. arabi-*

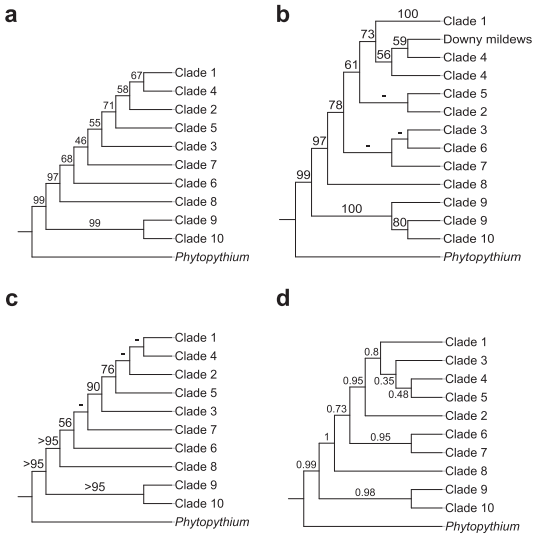


FIG 2 Congruence of the *Peronosporales* order among recent multilocus phylogenetic analyses. (a) Seven-locus maximum-likelihood (ML)/maximum-parsimony (MP)/Bayesian phylogeny of *Phytophthora* by Blair et al. (42). (b) Minimum-evolution (ME)/ML/Bayesian phylogeny of *Phytophthora* and downy mildews by Runge et al. (46). (c) Eleven-locus ML/MP/Bayesian phylogeny of *Phytophthora* by Martin et al. (47). (d) Six-locus coalescent phylogeny of *Phytophthora* by Martin et al. (47). Support values, where given, represent maximum-likelihood bootstrap support, except for panel d, where Bayesian posterior probabilities are given instead.

dopsidis as sister to members of the *Phytophthora* genus (48). Another downy mildew species, *Plasmopara halstedii*, was placed sister to *Phytophthora* clade 1 in similar phylogenetic analyses (36, 49). *Phytopythium*, a morphological intermediate between *Phytophthora* and *Pythium*, was reclassified from *Pythium* clade K to its own genus within the *Peronosporales* order based on a recent multigene phylogenetic analysis which placed the genus sister to *Phytophthora* (50). *Pythium* itself is divided into 10 clades, labeled A to J, which were initially circumscribed with its data and consistent with mitochondrial data (51). The main morphological difference between clades within *Pythium* is the development of the filamentous sporangium in species within clades A to C from the ancestral globose sporangium observed in the basal clades and *Phytopythium* (51, 52), with an intermediate contiguous sporangium developing in species within clade D (51) and an elongated sporagium in species within clade H (53). Otherwise, as in *Phytophthora*, phylogenetic clades generally do not correlate with distinct morphological characters in *Pythium* (51). A number of phylogenetic analyses suggest that *Pythium* is polyphyletic (36, 49, 52–55), and there has been recent suggestion that it be amended entirely into at least five new genera (53, 56).

Many of the aforementioned phylogenetic analyses of the oomycetes are based upon a small number of highly conserved nuclear and/or mitochondrial markers, either through consensus analysis or concatenated analysis. The selection of such markers, while usually robust, may unintentionally ignore other types of potential phylogenetic markers that might resolve conflicting analyses, such as lineages which include gene duplication events (20). One solution to the possible limitations of single-gene or

small-scale gene phylogenies is to assemble a consensus phylogeny for a given set of taxa using many sources of single-gene phylogenies through supertree analysis, which enables the inclusion of phylogenies with missing or duplicated taxa (57). Matrix representation using parsimony (MRP), in which character matrices are generated for each source phylogeny and merged into a single binary character matrix for maximum-parsimony alignment (58, 59), is one of the most commonly used supertree methods and has seen successful application in a number of eukaryotic phylogenomic studies (60–62). Other methods have been developed for inferring species phylogeny from paralogous gene phylogenies, the most successful of which has been gene tree parsimony (GTP) (63). GTP attempts to find the most parsimonious species tree from a set of source phylogenies with the lowest number of events required to explain incongruences (i.e., gene duplication events) between the source phylogenies and has seen application in large-scale phylogenetic analysis (64). Another method of large-scale phylogenetic analysis is the supermatrix approach of concatenating multiple character data sets for simultaneous analysis (65).

Since the publication of the genome sequences of *Phytophthora sojae* and *Phytophthora ramorum* in 2006 (66), the quantity of oomycete genomic data has steadily increased; currently, 37 oomycete species now have publicly available genomic data at the assembly level or higher (Table 1). With this in mind, we have conducted the first whole-genome phylogenetic analysis for the oomycetes as a class, using a variety of supertree and supermatrix approaches which have previously been used in fungal whole-genome phylogenetic analysis (60). In our analysis, we utilized protein data from 37 complete oomycete genomes and 6 complete SAR genomes (as outgroups). This represents all extant genomic data from the four "crown" oomycete orders and covers 8 of the 10 phylogenetic clades within *Phytophthora* and 7 of the 10 phylogenetic clades within *Pythium* (Table 1). Our whole-genome phylogenetic analysis of the oomycetes supports the four oomycete orders and the placement of *Phytopythium* within the *Peronosporales* and individual clades within *Phytophthora* and *Pythium*. The resolution of the *Peronosporales* as an order varied under different methods, probably due to missing data from clades 4 and 9 within *Phytophthora*. However, the overall order phylogenies are relatively congruent among our different species phylogenies. This analysis will provide a useful backbone to future genome phylogenies of the oomycetes utilizing more taxonomically extensive data sets.

RESULTS AND DISCUSSION

Identification of orthologous and paralogous oomycete and SAR gene families. For our supertree analyses, we constructed a data set containing 43 complete genomes, consisting of 37 from oomycete species and 6 outgroups from other species within the SAR supergroup (Materials and Methods; Table 1). Of these 37 oomycete genomes, 26 were from either *Phytophthora* species or *Pythium* species representing the majority of clades within both genera, and the remainder were sampled from all four of the "crown" orders (66–89). We downloaded proteomes for 23 oomycete species which were available from public databases, and we generated corresponding proteomes for the remaining 14 species from publicly available assembly data using bespoke oomycete reference templates with AUGUSTUS and GeneMark-ES (90, 91) (Table S1). In total, our final data set contained 702,132 protein sequences from 37 complete oomycete genomes and 6 complete SAR genomes (Table 1).

The initial step in determining the phylogeny of the 43 oomycete and SAR genomes in our data set through supertree methods was to identify groups of closely related orthologs or paralogs within our data set, which we termed gene families, and to use these groups to generate gene phylogenies to use as source data for our methods. To identify families of orthologous and paralogous genes in our data set, we set the following criteria:

(1) A single-copy gene family must contain no more than one orthologous gene per species and must be present in four or more species.

TABLE 1 Taxonomic and genomic information for the 43 oomycete and SAR species in this analysis[a]

Species name	Clade	Order	Class	Reference	Gene
Albugo candida	NA	*Albuginales*	*Oomycota*	Links et al. 2011 (73)	13310
Albugo labiachii	NA	*Albuginales*	*Oomycota*	Kemen et al. 2011 (74)	13804
Hyaloperonospora arabidopsidis	NA	*Peronosporales*	*Oomycota*	Baxter et al. 2010 (71)	14321
Phytophthora agathidicida	Clade 5	*Peronosporales*	*Oomycota*	Studholme et al. 2016 (70)	14110*
Phytophthora capsici	Clade 2	*Peronosporales*	*Oomycota*	Lamour et al. 2012 (72)	19805
Phytophthora cinnamomi	Clade 7	*Peronosporales*	*Oomycota*	Studholme et al. 2016 (70)	12942*
Phytophthora cryptogea	Clade 8	*Peronosporales*	*Oomycota*	Feau et al. 2016 (75)	11876*
Phytophthora fragariae	Clade 7	*Peronosporales*	*Oomycota*	Gao et al. 2015 (76)	13361*
Phytophthora infestans	Clade 1	*Peronosporales*	*Oomycota*	Haas et al. 2009 (69)	17797
Phytophthora kernoviae	Clade 10	*Peronosporales*	*Oomycota*	Sambles et al. 2015 (77)	10650
Phytophthora lateralis	Clade 8	*Peronosporales*	*Oomycota*	Quinn et al. 2013 (78)	11635
Phytophthora multivora	Clade 2	*Peronosporales*	*Oomycota*	Studholme et al. 2016 (70)	15006*
Phytophthora nicotianae	Clade 1	*Peronosporales*	*Oomycota*	Liu et al. 2016 (79)	10521
Phytophthora parasitica	Clade 1	*Peronosporales*	*Oomycota*	Broad Institute (INRA-310 v. 3)	27942
Phytophthora pinifolia	Clade 6	*Peronosporales*	*Oomycota*	Feau et al. 2016 (75)	19533*
Phytophthora pluvialis	Clade 3	*Peronosporales*	*Oomycota*	Studholme et al. 2016 (70)	18426*
Phytophthora pisi	Clade 7	*Peronosporales*	*Oomycota*	PRJEB6298	15495*
Phytophthora ramorum	Clade 8	*Peronosporales*	*Oomycota*	Tyler et al. 2006 (66)	15743
Phytophthora rubi	Clade 7	*Peronosporales*	*Oomycota*	PRJNA244739	15462*
Phytophthora sojae	Clade 7	*Peronosporales*	*Oomycota*	Tyler et al. 2006 (66)	26584
Phytophthora taxon Totara	Clade 3	*Peronosporales*	*Oomycota*	Studholme et al. 2016 (70)	16691*
Plasmopara halstedii	NA	*Peronosporales*	*Oomycota*	Sharma et al. 2015 (80)	15469
Plasmopara viticola	NA	*Peronosporales*	*Oomycota*	PRJNA329579	12048*
Phytopythium vexans	NA	*Peronosporales*	*Oomycota*	Adhikari et al. 2013 (67)	11958
Pilasporangium apinafurcum	NA	*Pythiales*	*Oomycota*	PRJDB3797	13184*
Pythium aphanidermatum	Clade A	*Pythiales*	*Oomycota*	Adhikari et al. 2013 (67)	12312
Pythium arrhenomanes	Clade B	*Pythiales*	*Oomycota*	Adhikari et al. 2013 (67)	13805
Pythium insidiosum	Clade C	*Pythiales*	*Oomycota*	Rujirawat et al. 2015 (81)	19290*
Pythium irregulare	Clade F	*Pythiales*	*Oomycota*	Adhikari et al. 2013 (67)	13805
Pythium iwayami	Clade G	*Pythiales*	*Oomycota*	Adhikari et al. 2013 (67)	14875
Pythium oligandrum	Clade D	*Pythiales*	*Oomycota*	Berger et al. 2016 (82)	14292*
Pythium ultimum var. *sporangiiferum*	Clade I	*Pythiales*	*Oomycota*	Adhikari et al. 2013 (67)	14096
Pythium ultimum var. *ultimum*	Clade I	*Pythiales*	*Oomycota*	Lévesque et al. 2010 (68)	15323
Aphanomyces astaci	NA	*Saprolegniales*	*Oomycota*	Broad Institute (APO3 v.2)	26259
Aphanomyces invadans	NA	*Saprolegniales*	*Oomycota*	Broad Institute (9901 v.2)	20816
Saprolegnia diclina	NA	*Saprolegniales*	*Oomycota*	PRJNA168273	18229
Saprolegnia parasitica	NA	*Saprolegniales*	*Oomycota*	Jiang et al. 2013 (83)	20121
Aureococcus anophagefferns	NA	*Pelagomonadales*	*Pelagophyceae*	Gobler et al. 2011 (84)	11501
Ectocarpus siliculosus	NA	*Ectocarpales*	*Phaeophyceae*	Cock et al. 2010 (87)	16269
Phaeodactylum tricornutum	NA	*Naviculales*	*Bacillariophyceae*	Bowler et al. 2008 (85)	10402
Thalassiosira psuedonana	NA	*Thalassiosirales*	*Coscinodiscophyceae*	Armbrust et al. 2004 (86)	11776
Paramecium tetraurelia	NA	*Peniculida*	*Oligohymenophorea*	Aury et al. 2006 (88)	39580
Bigelowiella natans	NA	*Chlorarachniophyceae*	*Cercozoa*	Curtis et al. 2012 (89)	21708

[a]Protein counts generated in this study from assembly data are highlighted with an asterisk (*). References are to the genome publications where possible and otherwise to the NCBI BioProject identifier or the Broad Institute strain identifier and assembly version. NA, not applicable.

(2) A multicopy gene family must contain at least four unique species, and two or more paralogs must be present in at least one of the species.

Using OrthoMCL (92), with an inflation value of 1.5 and a strict BLASTp cutoff value of 10^{-20} (93) and bespoke Python scripting, we identified over 56,000 homologous oomycete and SAR gene families in our data set. Of these, 2,853 families matched our criterion for single-copy families and 11,158 families matched our criterion for multicopy families. By aligning each of these gene families in MUSCLE (94) and sampling for highly conserved regions using Gblocks (95), both using the default parameters, and then carrying out permutation-tail possibility (PTP) tests for every remaining sampled alignment using PAUP* (96, 97), we were able to remove 576 single-copy gene families and 5,103 multicopy gene families with poor phylogenetic signal from our data. All

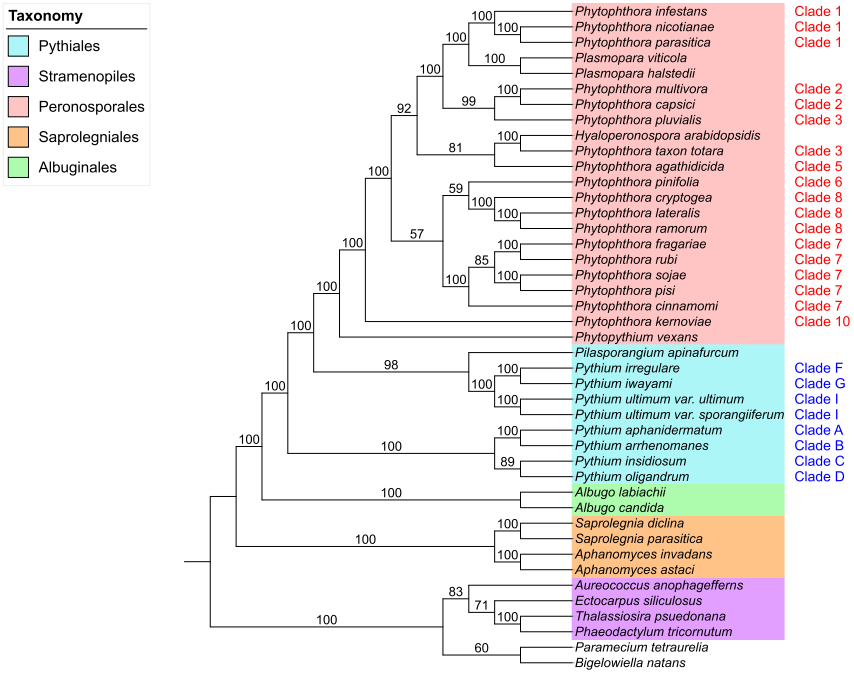


FIG 3 Matrix representation with parsimony (MRP) supertree of 37 oomycete species and 6 SAR species (2,280 source phylogenies). The supertree was generated in CLANN. The phylogeny is rooted at the SAR branch. *Phytophthora* clades as designated by Blair et al. (42) and *Pythium* clades as designated by de Cock et al. (50) are indicated in red and blue, respectively. No color, *P. tetraurelia* (*Alveolata*) and *B. natans* (*Rhizaria*).

remaining gene families had their evolutionary model estimated using ProtTest (98) (Table S2), and maximum-likelihood gene phylogenies were generated using PhyML with 100 bootstrap replicates (99). We generated phylogenetic reconstructions for 2,280 orthologous gene families (containing 35,622 genes) and 6,055 paralogous gene families (containing 174,282 genes). In total, from our 43-genome data set, we identified 8,335 individual gene phylogenies, containing 209,904 oomycete and SAR genes.

Supetree phylogenies fully resolve oomycete class and order phylogenies. All 2,280 orthologous single-copy gene phylogenies (35,622 genes in total) were used as input for CLANN (100), which implements a matrix representation using parsimony (MRP) method to determine consensus phylogeny for many source phylogenies with overlapping taxa or missing taxa. An MRP supertree phylogeny was generated in CLANN using a heuristic search with 100 bootstrap replicates. The supertree was visualized and annotated within the Interactive Tree of Life (iTOL) website (101) and rooted at the branch containing the SAR outgroups, *Paramecium tetraurelia* (*Alveolata*), *Bigelowiella natans* (*Rhizaria*), and four stramenopiles species (Fig. 3).

MRP supertree analysis of 2,280 orthologous single-copy oomycete gene phylogenies supported the four "crown" oomycete orders (*Saprolegniales*, *Albuginales*, *Pythiales*, and *Peronosporales*), with maximum bootstrap support (BP) (Fig. 3). The MRP

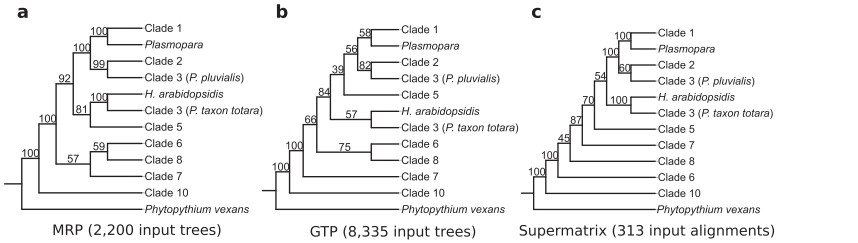


FIG 4 Congruence of the *Peronosporales* order data between our supertree and supermatrix methods. (a) MRP analysis. (b) GTP analysis. (c) Concatenated supermatrix analysis. For full phylogenies, refer to Fig. 3, 5, and 6, respectively.

supertree reflects the consensus phylogeny of the oomycetes (31–33) (Fig. 1). The *Saprolegniales* species represent the most basal "crown" order, and the *Albuginales* is a sister order to the *Pythiales* and *Peronosporales*. Within the *Pythiales* themselves, a highly supported split among *Pythium* clades A to D (100% BP) and clades F to I (100% BP) was observed, matching similar splits seen in small-scale analyses (51, 52) (Fig. 3). *Pilasporangium apinafurcum*, a *Pythiales* species, is placed sister to *Pythium* clades F to I (98% BP). *Phytopythium vexans* is placed at the base of the *Peronosporales* order (Fig. 3), supporting the recent reclassification of the *Phytopythium* genus from the *Pythiales* (50). Many individual *Phytophthora* clades within the *Peronosporales* are well supported. In addition, the "downy mildews" species in our data set (*Hyaloperonospora arabidopsidis* and two *Plasmopara* species) place as derived taxa within the *Peronosporales* order rather than as basal to *Phytophthora* (Fig. 3). The overall phylogeny of the *Peronosporales* in our MRP supertree is summarized in Fig. 4a and discussed in greater detail later in the text. As an additional analysis, a consensus supernetwork of the phylogenetic splits within the 2,280 single-copy gene phylogenies was generated in SplitsTree (102) (see Fig. S1 in the supplemental material). The network further highlights support for the four "crown" oomycete orders and the division of the *Pythiales* order as in the supertree phylogeny; it also recapitulates many of individual *Phytophthora* clades and intraorder relationships within the *Peronosporales* (Fig. 3 and 4a; Fig. S1).

Both the 2,280 single-copy phylogenies and the 6,055 multicopy phylogenies (209,904 genes in total) were used as input for DupTree (103), which uses a gene tree parsimony (GTP) method to determine consensus phylogeny for many source phylogenies that may include gene duplication events. The source data were bootstrapped with 100 replicates, and the resultant consensus GTP supertree was rooted at the branch containing *Paramecium tetraurelia*, *Bigelowiella natans*, and the other stramenopiles species (Fig. 5). As in the single-gene MRP supertree, all four individual crown oomycete orders and the oomycete class phylogeny are highly supported. The *Pythiales* order is once again split into highly supported sister branches containing clades A to D (100% BP) and clades F to I (100% BP) (Fig. 5). The *Peronosporales* order is highly supported again (100% BP), as is the placement of *Phytopythium vexans* at the base of this order (Fig. 5). As with the single-gene MRP supertree, the downy mildews (*P. viticola* and *P. halstedii*) are found as sister taxa to clade 1 *Phytophthora* species. However, it is worth pointing out that phylogenetic support for this grouping is weaker in the GTP supertree (58% BP) (Fig. 4b and 5) than in the MRP supertree, where support is very strong (100% BP) (Fig. 3). Overall, the phylogeny of the *Peronosporales* order in our GTP supertree displays weaker bootstrap support at some branches than in the single-gene MRP supertree. However, with the exception of the placement of clade 5, the overall taxonomic congruence between the two supertree approaches for the *Peronosporales* is high (Fig. 3, 4a and b, and 5).

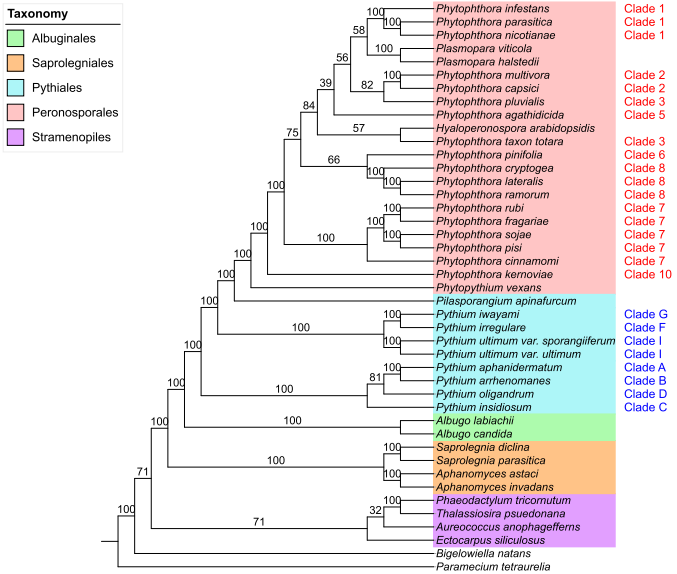


FIG 5 Gene tree parsimony (GTP) supertree of 37 oomycete species and 6 SAR species (8,335 source phylogenies). The supertree was generated in DupTree. The phylogeny is rooted at the SAR branch. *Phytophthora* clades as designated by Blair et al. (42) and *Pythium* clades as designated by de Cock et al. (50) are indicated in red and blue, respectively. No color, *P. tetraurelia* (*Alveolata*) and *B. natans* (*Rhizaria*).

The supermatrix approach based on ubiquitous *Peronosporales* gene phylogenies supports supertree phylogenies. As a complement to our supertree method phylogenies, we undertook a supermatrix approach to infer the oomycete species phylogeny using oomycete orthologs of known proteins corresponding to clusters of orthologous groups (COG) as phylogenetic markers (104). To identify oomycete COGs, we performed a reciprocal BLASTp analysis of all 458 *Saccharomyces cerevisiae* COGs against the 37 oomycete proteomes in our full data set (590,896 protein sequences in total) with an E value of 10^{-10}. Overall, 443 oomycete gene families that were reciprocal top hits to *S. cerevisiae* COGs were retrieved. Of the 443 COG families, 144 families contained an ortholog from all 37 oomycete species and were retained for analysis. A superalignment of 16,934 characters was generated by concatenating the 131 aligned families which retained alignment data after Gblocks sampling with FAScon CAT (105). The maximum-likelihood phylogeny of this superalignment was reconstructed in PhyML with 100 bootstrap replicates and an LG+I+G+F amino acid substitution model as selected by ProtTest, and the resultant consensus phylogeny was rooted at the *Saprolegniales* branch (Fig. S2). This initial supermatrix phylogeny supported the four "crown" orders similarly to our supertree phylogenies; however, poor resolution and inconsistent phylogeny were observed within the *Peronosporales*, particularly the placement of species from *Phytophthora* clades 7 and 8; for example, clade 7 species are not monophyletic (Fig. S2). To attempt to tease apart the data corresponding to the poor resolution of the *Peronosporales* in our maximum-likelihood phylogeny, a neighbor-joining network was generated for the COG superalignment in SplitsTree to visualize

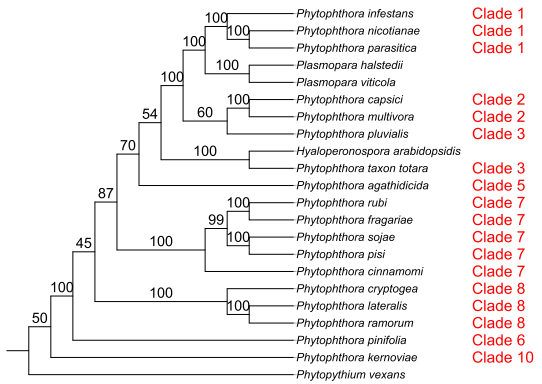


FIG 6 Maximum-likelihood (ML) supermatrix phylogeny of 22 *Peronosporales* species (313 ubiquitous *Pernosporales* gene families, 47,635 characters). The supermatrix phylogeny was generated in PhyML with a JTT+I+G+F amino acid substitution model. The cladogram is rooted at *Phytopythium vexans*. *Phytophthora* clades as designated by Blair et al. (2008) are shown in red.

the bifurcations within the superalignment (Fig. S3). As can be seen in the network, a significant amount of phylogenetic conflict is obvious and is represented as alternative splits among *Peronosporales* clades, a phenomenon that is consistent with poor bootstrap support and inconsistent topology (relative to supertrees) throughout the *Peronosporales* in this class-level supermatrix phylogeny (Fig. S2 and S3).

To extend our COG supermatrix phylogeny, we took the approach of generating a supermatrix from ubiquitous gene families within the 22 *Peronosporales* species in our data set. Using this approach, we hoped to extend the amount of available alignment data for species solely within *Peronosporales* to improve resolution of the order. We defined a ubiquitous *Peronosporales* gene family as containing exactly one ortholog from all 22 *Peronosporales* species in our data set. Using OrthoMCL, with a strict BLASTp E value of 10^{-20} and an inflation value of 1.5, we identified over 20,000 orthologous gene families in the 22 *Peronosporales* proteomes in our data set. From these families, we identified 352 ubiquitous gene families within *Peronosporales* using bespoke Python scripting; each family was then aligned in MUSCLE and sampled in Gblocks. After removing families which did not retain alignment data after Gblocks, we concatenated the remaining 313 gene families into a superalignment that was 47,365 amino acids in length. The maximum-likelihood phylogeny for this superalignment was generated with 100 bootstrap replicates and a JTT+I+G+F evolutionary model. The resultant consensus phylogeny was rooted at *Phytopythium vexans* (Fig. 6). While resolution of relationships among clades is still weak at some branches, the higher support seen on many other branches and the overall topology of the ubiquitous supermatrix phylogeny represent substantial improvements over the COG supermatrix. *Phytophthora* clades 1, 2, 7, and 8 are now all monophyletic, with 100% bootstrap support each. The genus is split between the basal lineages (*Phytopythium* and *Phytophthora* clades 6 to 10) and the more extensively derived *Phytophthora* clades (clades 1 to 5) and the downy mildews, which form a monophyletic group (70% BP) (Fig. 4c and 6), an inference that is also observed in our supertree species phylogenies and with the highest degree of congruence to the single-gene MRP supertree (Fig. 4a and b).

Resolution of the *Peronosporales* order in phylogenomic analysis. All three of our whole-genome species phylogenies strongly support the *Peronosporales* order

(Fig. 4) and display a high degree of congruence with one another. Each phylogeny also supports the recent reclassification of *Phytopythium* from the *Pythiales* to the *Peronosporales* as a basal taxon (50). All three phylogenies also show varying but strong bootstrap support (70 to 92% BP) for the divergence of *Phytophthora* clades 1 to 5 and the downy mildews (*Plasmopara* spp., *H. arabidopsidis*) from the remaining *Phytophthora* clades and *Phytopythium* at a single point (Fig. 4c). The relationships among these taxa across our phylogenies can be summarized as follows:

(1) The downy mildews species *Hyaloperonospora arabidopsidis* and *Phytophthora* taxon Totara (*Phytophthora* clade 3) are sister taxa, with maximum support in both MRP and supermatrix analysis (Fig. 4a and c). Therefore, *Phytophthora* clade 3 is not monophyletic in any of our species phylogenies (Fig. 4). *Phytophthora* taxon Totara has provisionally been assigned to clade 3 based on sequence similarity. Our species phylogenies suggest that it is not actually a clade 3 species.
(2) A close relationship between *Phytophthora* clades 1 and 2, the clade 3 species *Phytophthora pluvialis*, and the downy mildew species *Plasmopara viticola* and *Plasmopara halstedii* is observed in each phylogeny, with maximum support in both MRP and supermatrix analysis (Fig. 4a and c).

The placement of the clade 5 species *Phytophthora agathidcida* varies in each phylogeny, but it appears that the species is most closely related to *Phytophthora* taxon Totara and *H. arabidopsidis* within the *Peronosporales*, as is most apparent in the single-gene MRP supertree (81% BP) (Fig. 3 and 4a). As for the more basal clades, both the MRP and GTP phylogenies show support for the idea of clade 6 species *Phythophthora pinifolia* being sister to *Phytophthora* clade 8, with highest bootstrap support of 59% and 75%, respectively (Fig. 4a and b).

In our analysis, we set out to resolve relationships within the oomycetes where conflicts have arisen in different analyses, particularly in the *Peronosporales* order (Fig. 2). With respect to the divergence of *Phytophthora* clades 1 to 5 and the downy mildews from the remaining basal taxa in the *Peronosporales* (i.e., *Phytophthora* clades 6 to 10 and *Phytopythium*), our results are congruent with the small-scale analyses performed by Blair et al. and Martin et al. (42, 47) (Fig. 2a, c, and d), with closest topological similarity to the latter authors' 6-locus coalescence method phylogeny (Fig. 2d), despite a lack of data from *H. arabidopsidis* and *Plasmopara* species in both analyses and the inclusion of *H. arabidopsidis* data in the analysis carried out by Runge et al. (46) (Fig. 2b). Our own analysis lacks data from any species in *Phytophthora* clade 4, which is still unsampled in terms of genome sequencing. In the analysis by Runge et al., *H. arabidopsidis* branches within paraphyletic *Phytophthora* clade 4; were there a representative species from clade 4 available, a greater degree of resolution for the relationships among *Phytophthora* clades 3 to 5 and *Hyaloperonospora* might be observed. However, it is not clear whether the placement of *H. arabidopsidis* relative to *Phytophthora* clade 1 would then recapitulate that described by Runge et al. (46). Similarly, with regard to the basal taxa, our result are relatively congruent with the linearized relationships seen in previous analyses (Fig. 2), although the close relationship of clade 6 species *Phytophthora pinifolia* to *Phytophthora* clade 7 seen in our two supertree methods is not reflected in any of the multilocus phylogenies (Fig. 4a and b). The resolution of the relationships among *Phytophthora* clades 6, 7, and 8 varies both in support and sister group relationships among our analyses (Fig. 4); however, similar variation can be observed between the highlighted multilocus phylogenies (42, 46, 47) (Fig. 2). The lack of available genomic data from *Phytophthora* clade 9 also prevents any conclusions regards its placement in a whole-genome phylogeny; however, we would expect that it would branch as a sister to clade 10 species such as *Phytophthora kernoviae*, as the relationship between clades 9 and 10 has been highly supported in multilocus analyses (42, 46, 47).

The use of supertree and phylogenomic methods in oomycete systematics. Our analysis is the first large-scale genome phylogeny of the oomycetes as a class, using all extant genomic data from 37 oomycete species. Our analysis has recapitulated the four crown orders of the oomycetes and many relationships within the two largest-sampled orders, the *Pythiales* and the *Peronosporales*. During our analysis, we were conscious of potential characteristics of oomycete genomes that could obfuscate phylogenomic analysis. The role of HGT and its impact on the quality of our analyses were considered; it has been shown that supertree and supermatrix analyses are thought to be susceptible to misleading signal in data sets where a large degree of HGT has occurred, particularly in MRP analysis (106). While HGT from other microbial eukaryotes, fungi, and prokaryotes has been identified within oomycete genomes, the majority of these events are thought to be ancestral or to have not occurred in proportions large enough to impact our results (4, 5, 107). Other factors, such as fast-evolving regions of genomes or ancestral gene loss or duplication events within the oomycetes, are not likely to have affected our analysis, given our genome-wide scale of data acquisition and our strict filtering of gene families with poor phylogenetic signal (10, 48, 96). Intraspecific hybridization within the *Phytophthora* genus has been increasingly reported in the literature and usually occurs in nature among *Phytophthora* species within the same phylogenetic clade (108). A number of hybrid species or hybridization events have been described in *Phytophthora* clades 6 to 8 (108–110); however, none of these species are present in our data set. Also, where hybridization has occurred, it has been between closely related species and, in the case of *Phytophthora* species, those from the same phylogenetic clade. Taking this into consideration, hybridization should affect intraclade relationships to a greater degree than interclade relationships.

Compared with fungi, particularly in light of the ongoing 1,000 fungal genomes project (http://1000.fungalgenomes.org), there is a relative dearth of genomic data for both the earliest diverging lineages and the "crown" taxa within the oomycetes. With the greater sampling of genomic sequencing of the oomycetes likely to occur in the future, it is our view that subsequent genome phylogenies of the oomycetes will match the success of other eukaryotic genome phylogenies at resolving individual problematic clade and species relationships (60, 62). We suspect that, with a broader sampling of all *Phytophthora* clades and downy mildew species, we would see better resolution of the *Peronosporales* within any subsequent oomycete genome phylogenies. Similar approaches with other oomycete taxa, such as *Pythium*, may disentangle some of the phylogenetic conflicts seen in recent analyses (49, 53). Similarly, sequencing of more *Saprolegniales* species or basal oomycete species and their inclusion in similar analyses will potentially help uncover further aspects of oomycete evolution, including the evolution of phytopathogenicity. Such analyses, for which ours is a first step, would also provide the benefit of establishing a robust phylogeny for a eukaryotic group with such devastating ecological impact and would hopefully encourage further genomics and phylogenomics research into the oomycetes.

Conclusions. Using 37 oomycete genomes and 6 SAR genomes, we have carried out the first whole-genome phylogenetic analysis of the oomycetes as a class. The different methods that we used in our analysis support the four "crown" oomycete orders and support many individual phylogenetic clades within genera. Our analysis also generally supports the placement of *Phytopythium* within the *Peronosporales*, the placement of the downy mildews within the *Phytophthora* genus, and the topology of clades within the *Pythiales* order. However, resolution of the *Peronosporales* as an order remains weak at some branches, possibly due to a lack of genomic data for some phylogenetic clades within *Phytophthora*. As the amount of genomic data available for the oomycetes increases, future genome phylogenies of the class should resolve these branches, as well as those within currently unsampled basal lineages or undersampled taxa such as *Saprolegnia*. Our analysis represents an important backbone for oomycete phylogenetics upon which future analyses can be based.

MATERIALS AND METHODS

Data set assembly. The predicted proteomes for 29 SAR species (23 oomycete species, 4 other stramenopile species, the alveolate species *Paramecium tetraurelia*, and the rhizarian species *Bigelowiella natans*) were obtained from public databases (Table 1). Predicted proteomes for a further 14 oomycete species (10 *Phythophthora* species, 2 *Pythium* species, *Plasmopara viticola*, and *Pilasporangium apinafurcum*) were generated from publically available assembly data using AUGUSTUS (90). Templates for *ab initio* protein prediction with AUGUSTUS were generated from assembly and expressed sequence tag (EST) data from a number of reference oomycete species (*Phytophthora sojae, Phytophthora capsici*, *Pythium ultimum* var. *ultimum*, and *Plasmopara halstedii*) (Table S1). *Ph. capsici* was used as a reference for *Phytophthora* species from clades 1 to 5, while *Ph. sojae* was used as a reference for *Phytophthora* species from clades 6 to 10. *Py. ultimum* var. *ultimum* was used as a reference for two *Pythium* species and *Pi. apinafurcum*. *P. halstedii* was used as a reference for *P. viticola*. GeneMark-ES (91) was used in conjuction with AUGUSTUS for protein prediction for *Pi. apinafurcum*. The taxonomy, assembly, and prediction statistics for each of the 14 assemblies included in this study are summarized in Table S1. Our final data set contained 702,132 protein sequences from 37 oomycete genomes and 6 SAR genomes (66–89) (Table 1; Table S1).

Identification and reconstruction of gene phylogenies in oomycete and SAR genomes. All 702,132 protein sequences in our data set were filtered and clustered into 56,638 orthologous gene families using OrthoMCL (92), with a BLASTp E value cutoff of 10^{-20} (93) and an inflation value of 1.5. Using bespoke Python scripting, we identified and retrieved two types of gene family containing 200 sequences or fewer from the 56,638 families within our data set as follows:

(1) A total of 2,853 single-copy gene families (single-copy orthologs present in ≥4 species.
(2) A total of 11,158 multicopy gene families (≥1 paralog[s] present in ≥4 species).

Each of these gene families was retrieved and aligned in MUSCLE (94), and highly conserved regions of these alignments were sampled using Gblocks (95) with the default parameters. A total of 266 single-copy gene families and a total of 4,928 multicopy gene families did not retain alignment data after Gblocks sampling and were discarded. Permutation-tail probability (PTP) tests (96) were carried out for every remaining sampled gene family in PAUP* (97), using 100 replicates, to determine whether a given sampled gene family had phylogenetic signal. Those sampled gene families whose PTP test result had a P value of ≤0.05 were considered to have signal and were retained. A total of 2,280 single-copy sampled gene families (containing 35,622 genes in total) and a total of 6,055 multicopy sampled gene families (containing 174,282 genes in total) ultimately satisfied our filtering process. Best-fit amino acid replacement models were selected for every remaining sampled gene family using ProtTest (Table S2), and maximum-likelihood phylogenetic reconstruction was carried out using PhyML with 100 bootstrap replicates.

Supertree analyses of single-copy and paralogous gene phylogenies. Maximum-parsimony supertree analysis of 2,280 single-copy gene phylogenies (containing 35,622 genes in total) was carried out using CLANN, by performing a subtree prune and regraft (SPR) heuristic search with 100 bootstrap replicates (100). This phylogeny was visualized and annotated as a cladogram using the Interactive Tree of Life (iTOL) website (101) (Fig. 3). As an additional analysis, a consensus supernetwork of phylogenetic multifurcations within the 2,280 individual gene phylogenies was generated in SplitsTree (102) (see Fig. S1 in the supplemental material). Gene tree parsimony (GTP) supertree analyses of all 8,335 gene phylogenies (containing 209,904 genes in total) was carried out using DupTree (103) and a rooted SPR heuristic search of 100 bootstrapped replicates of each phylogeny. A consensus phylogeny was generated from all individual replicates and was visualized and annotated as a cladogram using iTOL (Fig. 5).

Identification and supermatrix analysis of ubiquitous oomycete gene phylogenies. A reciprocal BLASTp search was carried out with an E value cutoff of 10^{-10} among all 37 oomycetes proteomes in our data set (590,896 protein sequences in total) and 458 core orthologous genes (COGs) in *Saccharomyces cerevisiae* from the CEGMA data set (93, 104). A total of 443 oomycete gene families representing oomycete top hits to *S. cerevisiae* COGs were retrieved, among which 144 families contained an ortholog from all 37 oomycete species in our data set. Each of these 144 families was aligned in MUSCLE and was sampled for highly conserved regions using Gblocks with the default parameters. After 13 families which failed to retain alignment data after Gblocks sampling were removed, the remaining 131 sampled alignments (containing 4,847 genes in total) were concatenated into a superalignment of 16,934 aligned positions. This superalignment was bootstrapped 100 times using Seqboot, and maximum-likelihood phylogenetic trees were generated for each individual replicate using PhyML, with an LG+I+G+F amino acid substitution model as selected by ProtTest. A consensus tree was generated from these replicate trees using Consense, and the consensus tree was visualized and annotated as a cladogram using iTOL (Fig. S2). A neighbor-joining network of phylogenetic splits in the original superalignment was generated in SplitsTree (Fig. S3).

Identification and supermatrix analysis of ubiquitous *Peronosporales* gene phylogenies. A total of 347,375 protein sequences from the 22 *Peronosporales* proteomes in our data set were filtered and clustered into 22,803 orthologous gene families using OrthoMCL, with a BLASTp E value cutoff of 10^{-20} and an inflation value of 1.5. Using bespoke Python scripting, we identified 352 ubiquitous *Peronosporales* gene families, which we defined as any family which had exactly one representative ortholog from all 22 *Peronosporales* species in our data set. Each of these families was aligned in MUSCLE and sampled for highly conserved regions using Gblocks with the default parameters. After 39 gene families which did not retain alignment data after sampling were removed, the remaining 313 sampled alignments (containing 6,886 genes in total) were concatenated into a single superalignment of 47,365 aligned positions. This superalignment was bootstrapped 100 times using Seqboot, and maximum-likelihood phylogenetic trees were generated for each individual replicate using PhyML with a JTT+I+G+F amino acid substitution model, as selected by ProtTest. A consensus tree was generated from these replicate trees using Consense, and the consensus tree was visualized and annotated as a cladogram using iTOL (Fig. 6).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSphere.00095-17.

FIG S1, PDF file, 0.1 MB.
FIG S2, PDF file, 0.1 MB.
FIG S3, PDF file, 0.02 MB.
TABLE S1, DOCX file, 0.02 MB.
TABLE S2, DOCX file, 0.01 MB.

ACKNOWLEDGMENTS

We acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support. We also thank Joyce Reilly for generating corroborating results through her undergraduate research project.

C.G.P.M. is funded by Irish Research Council grant number GOIPG/2015/2242.

REFERENCES

1. Beakes GW, Glockling SL, Sekimoto S. 2012. The evolutionary phylogeny of the oomycete "fungi." Protoplasma 249:3–19. https://doi.org/10.1007/s00709-011-0269-2.
2. Lévesque CA. 2011. Fifty years of oomycetes—from consolidation to evolutionary and genomic exploration. Fungal Divers 50:35–46. https://doi.org/10.1007/s13225-011-0128-7.
3. Richards TA, Dacks JB, Jenkinson JM, Thornton CR, Talbot NJ. 2006. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. Curr Biol 16:1857–1864. https://doi.org/10.1016/j.cub.2006.07.052.
4. Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci U S A 108:15258–15263. https://doi.org/10.1073/pnas.1105100108.
5. Savory F, Leonard G, Richards TA. 2015. The role of horizontal gene transfer in the evolution of the oomycetes. PLoS Pathog 11:e1004805. https://doi.org/10.1371/journal.ppat.1004805.
6. Burki F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect Biol 6:a016147. https://doi.org/10.1101/cshperspecta.016147.
7. Tsui CKM, Marshall W, Yokoyama R, Honda D, Lippmeier JC, Craven KD, Peterson PD, Berbee ML. 2009. Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding. Mol Phylogenet Evol 50:129–140. https://doi.org/10.1016/j.ympev.2008.09.027.
8. Cavalier-Smith T, Chao EEY. 2006. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J Mol Evol 62:388–420. https://doi.org/10.1007/s00239-004-0353-8.
9. Riisberg I, Orr RJS, Kluge R, Shalchian-Tabrizi K, Bowers HA, Patil V, Edvardsen B, Jakobsen KS. 2009. Seven Gene phylogeny of heterokonts. Protist 160:191–204. https://doi.org/10.1016/j.protis.2008.11.004.
10. Judelson HS. 2012. Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. Eukaryot Cell 11:1304–1312. https://doi.org/10.1128/EC.00155-12.
11. Cavalier-Smith T. 1981. Eukaryote kingdoms: seven or nine? Biosystems 14:461–481. https://doi.org/10.1016/0303-2647(81)90050-2.
12. Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol 46:347–366. https://doi.org/10.1111/j.1550-7408.1999.tb04614.x.
13. Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. J Eukaryot Microbiol 56:1–8. https://doi.org/10.1111/j.1550-7408.2008.00371.x.
14. Bachvaroff TR, Sanchez Puerta MV, Delwiche CF. 2005. Chlorophyll c-containing plastid relationships based on analyses of a multigene data set with all four chromalveolate lineages. Mol Biol Evol 22:1772–1782. https://doi.org/10.1093/molbev/msi172.
15. Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. Proc Natl Acad Sci U S A 99:15507–15512. https://doi.org/10.1073/pnas.242379899.
16. Janouskovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci U S A 107:10949–10954. https://doi.org/10.1073/pnas.1003335107.
17. Rice DW, Palmer JD. 2006. An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. BMC Biol 4:31. https://doi.org/10.1186/1741-7007-4-31.
18. Gaston D, Roger AJ. 2013. Functional divergence and convergent evolution in the plastid-targeted glyceraldehyde-3-phosphate dehydrogenases of diverse eukaryotic algae. PLoS One 8:e70396. https://doi.org/10.1371/journal.pone.0070396.
19. Harper JT, Waanders E, Keeling PJ. 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. Int J Syst Evol Microbiol 55:487–496. https://doi.org/10.1099/ijs.0.63216-0.
20. Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rümmele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of Rhizaria with chromalveolates. Mol Biol Evol 24:1702–1713. https://doi.org/10.1093/molbev/msm089.
21. Keeling PJ. 2001. Foraminifera and Cercozoa are related in actin phylogeny: two orphans find a home? Mol Biol Evol 18:1551–1557. https://doi.org/10.1093/oxfordjournals.molbev.a003941.
22. Baldauf SL. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977. https://doi.org/10.1126/science.290.5493.972.
23. Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One 2:e790. https://doi.org/10.1371/journal.pone.0000790.
24. Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A 106:3859–3864. https://doi.org/10.1073/pnas.0807880106.
25. Shalchian-Tabrizi K, Kauserud H, Massana R, Klaveness D, Jakobsen KS. 2007. Analysis of environmental 18S ribosomal RNA sequences reveals unknown diversity of the Cosmopolitan phylum Telonemia. Protist 158:173–180. https://doi.org/10.1016/j.protis.2006.10.003.
26. Moreira D, von der Heyden S, Bass D, López-García P, Chao E, Cavalier-Smith T. 2007. Global eukaryote phylogeny: combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria,

Retaria and Excavata. Mol Phylogenet Evol 44:255–266. https://doi.org/10.1016/j.ympev.2006.11.001.
27. Dick MW. 2001. Straminipilous fungi. Kluwer Academic Publishers, Amsterdam, The Netherlands.
28. Matari NH, Blair JE. 2014. A multilocus timescale for oomycete evolution estimated under three distinct molecular clock models. BMC Evol Biol 14:101. https://doi.org/10.1186/1471-2148-14-101.
29. Taylor TN, Krings M, Kerp H. 2006. Hassiella monospora gen. et sp. nov., a microfungus from the 400 million year old Rhynie chert. Mycol Res 110:628–632. https://doi.org/10.1016/j.mycres.2006.02.009.
30. Thines M, Kamoun S. 2010. Oomycete–plant coevolution: recent advances and future prospects. Curr Opin Plant Biol 13:427–433. https://doi.org/10.1016/j.pbi.2010.04.001.
31. Hakariya M, Hirose D, Tokumasu S. 2007. A molecular phylogeny of Haptoglossa species, terrestrial Peronosporomycetes (oomycetes) endoparasitic on nematodes. Mycoscience 48:169–175. https://doi.org/10.1007/S10267-007-0355-7.
32. Beakes D, Honda D, Thines M, Beakes GW, Honda D, Thines M. 2014. Systematics of the Straminipila: Labyrinthulomycota, Hyphochytriomycota, and Oomycota, p 39–97. *In* The mycota. Springer, Berlin, Germany.
33. Sekimoto S, Yokoo K, Kawamura Y, Honda D. 2008. Taxonomy, molecular phylogeny, and ultrastructural morphology of Olpidiopsis porphyrae sp. nov. (oomycetes, straminipiles), a unicellular obligate endoparasite of Bangia and Porphyra spp. (Bangiales, Rhodophyta). Mycol Res 112:361–374. https://doi.org/10.1016/j.mycres.2007.11.002.
34. Jiang RHY, Tyler BM. 2012. Mechanisms and evolution of virulence in oomycetes. Annu Rev Phytopathol 50:295–318. https://doi.org/10.1146/annurev-phyto-081211-172912.
35. Thines M. 2014. Phylogeny and evolution of plant pathogenic oomycetes-a global overview. Eur J Plant Pathol 138:431–447. https://doi.org/10.1007/s10658-013-0366-5.
36. Riethmüller A, Voglmayr H, Göker M, Weiß M, Oberwinkler F. 2002. Phylogenetic relationships of the downy mildews (Peronosporales) and related groups based on nuclear large subunit ribosomal DNA sequences. Mycologia 94:834–849. https://doi.org/10.2307/3761698.
37. Hulvey JP, Padgett DE, Bailey JC. 2007. Species boundaries within Saprolegnia (Saprolegniales, Oomycota) based on morphological and DNA sequence data. Mycologia 99:421–429. https://doi.org/10.3852/mycologia.99.3.421.
38. Kamoun S. 2003. Molecular genetics of pathogenic oomycetes. Eukaryot Cell 2:191–199. https://doi.org/10.1128/EC.2.2.191-199.2003.
39. Gaastra W, Lipman LJA, De Cock AWAM, Exel TK, Pegge RBG, Scheurwater J, Vilela R, Mendoza L. 2010. Pythium insidiosum: an overview. Vet Microbiol 146:1–16. https://doi.org/10.1016/j.vetmic.2010.07.019.
40. Benhamou N, le Floch G, Vallance J, Gerbore J, Grizard D, Rey P. 2012. Pythium oligandrum: an example of opportunistic success. Microbiology 158:2679–2694. https://doi.org/10.1099/mic.0.061457-0.
41. Cooke DE, Drenth A, Duncan JM, Wagels G, Brasier CM. 2000. A molecular phylogeny of Phytophthora and related oomycetes. Fungal Genet Biol 30:17–32. https://doi.org/10.1006/fgbi.2000.1202.
42. Blair JE, Coffey MD, Park SY, Geiser DM, Kang S. 2008. A multi-locus phylogeny for Phytophthora utilizing markers derived from complete genome sequences. Fungal Genet Biol 45:266–277. https://doi.org/10.1016/j.fgb.2007.10.010.
43. Weir BS, Paderes EP, Anand N, Uchida JY, Pennycook SR, Bellgard SE, Beever RE. 2015. A taxonomic revision of Phytophthora clade 5 including two new species, Phytophthora agathidicida and P. Cocois. Phytotaxa 205:21–38. https://doi.org/10.11646/phytotaxa.205.1.2.
44. Kroon LPNM, Bakker FT, Van Den Bosch GBM, Bonants PJM, Flier WG. 2004. Phylogenetic analysis of Phytophthora species based on mitochondrial and nuclear DNA sequences. Fungal Genet Biol 41:766–782. https://doi.org/10.1016/j.fgb.2004.03.007.
45. Martin FN, Tooley PW. 2003. Phylogenetic relationships of Phytophthora ramorum, P. nemorosa, and P. pseudosyringae, three species recovered from areas in California with sudden oak death. Mycol Res 107:1379–1391. https://doi.org/10.1017/S0953756203008785.
46. Runge F, Telle S, Ploch S, Savory E, Day B, Sharma R, Thines M. 2011. The inclusion of downy mildews in a multi-locus-dataset and its reanalysis reveals a high degree of paraphily in *Phytophthora*. IMA Fungus 2:163–171. https://doi.org/10.5598/imafungus.2011.02.02.07.
47. Martin FN, Blair JE, Coffey MD. 2014. A combined mitochondrial and nuclear multilocus phylogeny of the genus Phytophthora. Fungal Genet Biol 66:19–32. https://doi.org/10.1016/j.fgb.2014.02.006.
48. Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2012. Reconstruction of oomycete genome evolution identifies differences in evolutionary trajectories leading to present-day large gene families. Genome Biol Evol 4:199–211. https://doi.org/10.1093/gbe/evs003.
49. Robideau GP, Rodrigue N, André Lévesque C. 2014. Codon-based phylogenetics introduces novel flagellar gene markers to oomycete systematics. Mol Phylogenet Evol 79:279–291. https://doi.org/10.1016/j.ympev.2014.04.009.
50. de Cock AWAM, Lodhi AM, Rintoul TL, Bala K, Robideau GP, Abad ZG, Coffey MD, Shahzad S, Lévesque CA. 2015. Phytopythium: molecular phylogeny and systematics. Persoonia Mol Phylogeny Evol Fungi 34:25–39. https://doi.org/10.3767/003158515X685382.
51. Lévesque CA, de Cock AW. 2004. Molecular phylogeny and taxonomy of the genus Pythium. Mycol Res 108:1363–1383. https://doi.org/10.1017/s0953756204001431.
52. Villa NO, Kageyama K, Asano T, Suga H. 2006. Phylogenetic relationships of pythium and Phytophthora species based on ITS rDNA, cytochrome oxidase II and beta-tubulin gene sequences. Mycologia 98:410–422. https://doi.org/10.3852/mycologia.98.3.410.
53. Uzuhashi S, Tojo M, Kakishima M. 2010. Phylogeny of the genus Pythium and description of new genera. Mycoscience 51:337–365. https://doi.org/10.1007/S10267-010-0046-7.
54. Hyde KD, Nilsson RH, Alias SA, Ariyawansa HA, Blair JE, Cai L, de Cock AWAM, Dissanayake AJ, Glockling SL, Goonasekara ID, Gorczak M, Hahn M, Jayawardena RS, van Kan JAL, Laurence MH, Lévesque CA, Li X, Liu JK, Maharachchikumbura SSN, Manamgoda DS, Martin FN, McKenzie EHC, McTaggart AR, Mortimer PE, Nair PVR, Pawłowska J, Rintoul TL, Shivas RG, Spies CFJ, Summerell BA, Taylor PWJ, Terhem RB, Udayanga D, Vaghefi N, Walther G, Wilk M, Wrzosek M, Xu JC, Yan JY, Zhou N. 2014. One stop shop: backbones trees for important phytopathogenic genera I (2014). Fungal Divers 67 21–125. https://doi.org/10.1007/s13225-014-0298-1.
55. Rahman AK, Sugitani N, Hatsu M, Takamizawa K. 2003. A role of xylanase, alpha-L-arabinofuranosidase, and xylosidase in xylan degradation. Can J Microbiol 49:58–64. https://doi.org/10.1139/w02-114.
56. Huang JH, Chen CY, Lin YS, Ann PJ, Huang HC, Chung WH. 2013. Six new species of Pythiogeton in Taiwan, with an account of the molecular phylogeny of this genus. Mycoscience 54:130–147. https://doi.org/10.1016/j.myc.2012.09.007.
57. Bininda-Emonds ORP. 2004. The evolution of supertrees. Trends Ecol Evol 19:315–322. https://doi.org/10.1016/j.tree.2004.03.015.
58. Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10. https://doi.org/10.2307/1222480.
59. Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1:53–58. https://doi.org/10.1016/1055-7903(92)90035-F.
60. Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol 6:99. https://doi.org/10.1186/1471-2148-6-99.
61. Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol 24:1752–1760. https://doi.org/10.1093/molbev/msm095.
62. Beck RMD, Bininda-Emonds ORP, Cardillo M, Liu FG, Purvis A. 2006. A higher-level MRP supertree of placental mammals. BMC Evol Biol 6:93. https://doi.org/10.1186/1471-2148-6-93.
63. Cotton JA, Page RDM. 2003. Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. Mol Phylogenet Evol 29:298–308. https://doi.org/10.1016/S1055-7903(03)00109-X.
64. Casewell NR, Wagstaff SC, Harrison RA, Wüster W. 2011. Gene tree parsimony of multilocus snake venom protein families reveals species tree conflict as a result of multiple parallel gene loss. Mol Biol Evol 28:1157–1172. https://doi.org/10.1093/molbev/msq302.
65. de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol Evol 22:34–41. https://doi.org/10.1016/j.tree.2006.10.002.
66. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CM, Dorrance AE, Dou D, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee MK, McDonald WH, Medina M, Meijer HJ, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JK, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BW, Terry A, Torto-Alalibo TA, Win J, Xu Z, Zhang H, et al. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313:1261–1266. https://doi.org/10.1126/science.1128796.
67. Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Lévesque CA, Buell CR. 2013. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. PLoS One 8:e75072. https://doi.org/10.1371/journal.pone.0075072.
68. Lévesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, Huitema E, Raffaele S, Robideau GP, Thines M, Win J, Zerillo MM, Beakes GW, Boore JL, Busam D, Dumas B, Ferriera S, Fuerstenberg SI, Gachon CMM, Gaulin E, Govers F, Grenville-Briggs L, Horner N, Hostetler J, Jiang RHY, Johnson J, Krajaejun T, Lin H, Meijer HJG, Moore B, Morris P, Phuntmart V, Puiu D, Shetty J, Stajich JE, Tripathy S, Wawra S, van West P, Whitty BR, Coutinho PM, Henrissat B, Martin F, Thomas PD, Tyler BM, De Vries RP, Kamoun S, Yandell M, Tisserat N, Buell CR. 2010. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biol 11:R73. https://doi.org/10.1186/gb-2010-11-7-r73.
69. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T, Bozkurt TO, Ah-Fong AMV, Alvarado L, Anderson VL, Armstrong MR, Avrova A, Baxter L, Beynon J, Boevink PC, Bollmann SR, Bos JIB, Bulone V, Cai G, Cakir C, Carrington JC, Chawner M, Conti L, Costanzo S, Ewan R, Fahlgren N, Fischbach MA, Fugelstad J, Gilroy EM, Gnerre S, Green PJ, Grenville-Briggs LJ, Griffith J, Grünwald NJ, Horn K, Horner NR, Hu C-H, Huitema E, Jeong D-H, Jones AME, Jones JDG, Jones RW, Karlsson EK, Kunjeti SG, Lamour K, Liu Z, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 461:393–398. https://doi.org/10.1038/nature08358.
70. Studholme DJ, McDougal RL, Sambles C, Hansen E, Hardy G, Grant M, Ganley RJ, Williams NM. 2016. Genome sequences of six Phytophthora species associated with forests in New Zealand. Genomics Data 7:54–56. https://doi.org/10.1016/j.gdata.2015.11.015.
71. Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, Kemen E, Thines M, Ah-Fong A, Anderson R, Badejoko W, Bittner-Eddy P, Boore JL, Chibucos MC, Coates M, Dehal P, Delehaunty K, Dong S, Downton P, Dumas B, Fabro G, Fronick C, Fuerstenberg SI, Fulton L, Gaulin E, Govers F, Hughes L, Humphray S, Jiang RHY, Judelson H, Kamoun S, Kyung K, Meijer H, Minx P, Morris P, Nelson J, Phuntumart V, Qutob D, Rehmany A, Rougon-Cardoso A, Ryden P, Torto-Alalibo T, Studholme D, Wang Y, Win J, Wood J, Clifton SW, Rogers J, Van den Ackerveken G, Jones JDG, et al. 2010. Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. Science 330:1549–1551. https://doi.org/10.1126/science.1195203.
72. Lamour KH, Stam R, Jupe J, Huitema E. 2012. The oomycete broad-host-range pathogen Phytophthora capsici. Mol Plant Pathol 13:329–337. https://doi.org/10.1111/j.1364-3703.2011.00754.x.
73. Links MG, Holub E, Jiang RHY, Sharpe AG, Hegedus D, Beynon E, Sillito D, Clarke WE, Uzuhashi S, Borhan MH. 2011. De novo sequence assembly of Albugo candida reveals a small genome relative to other biotrophic oomycetes. BMC Genomics 12:503. https://doi.org/10.1186/1471-2164-12-503.
74. Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, Robert-Seilaniantz A, Bailey K, Holub E, Studholme DJ, Maclean D, Jones JDG. 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of Arabidopsis thaliana. PLoS Biol 9:e1001094. https://doi.org/10.1371/journal.pbio.1001094.
75. Feau N, Taylor G, Dale AL, Dhillon B, Bilodeau GJ, Birol I, Jones SJM, Hamelin RC. 2016. Genome sequences of six Phytophthora species threatening forest ecosystems. Genomics Data 10:85–88. https://doi.org/10.1016/j.gdata.2016.09.013.
76. Gao R, Cheng Y, Wang Y, Wang Y, Guo L, Zhang G. 2015. Genome sequence of Phytophthora fragariae var. fragariae, a quarantine plant-pathogenic fungus. Genome Announc 3:e00034-15. https://doi.org/10.1128/genomeA.00034-15.
77. Sambles C, Schlenzig A, O'Neill P, Grant M, Studholme DJ. 2015. Draft genome sequences of Phytophthora kernoviae and Phytophthora ramorum lineage EU2 from Scotland. Genomics Data 6:193–194. https://doi.org/10.1016/j.gdata.2015.09.010.
78. Quinn L, O'Neill PA, Harrison J, Paskiewicz KH, McCracken AR, Cooke LR, Grant MR, Studholme DJ. 2013. Genome-wide sequencing of Phytophthora lateralis reveals genetic variation among isolates from Lawson cypress (Chamaecyparis lawsoniana) in Northern Ireland. FEMS Microbiol Lett 344:179–185. https://doi.org/10.1111/1574-6968.12179.
79. Liu H, Ma X, Yu H, Fang D, Li Y, Wang X, Wang W, Dong Y, Xiao B. 2016. Genomes and virulence difference between two physiological races of Phytophthora nicotianae. GigaScience 5:3. https://doi.org/10.1186/s13742-016-0108-7.
80. Sharma R, Xia X, Cano LM, Evangelisti E, Kemen E, Judelson H, Oome S, Sambles C, van den Hoogen DJ, Kitner M, Klein J, Meijer HJG, Spring O, Win J, Zipper R, Bode HB, Govers F, Kamoun S, Schornack S, Studholme DJ, Van den Ackerveken G, Thines M. 2015. Genome analyses of the sunflower pathogen Plasmopara halstedii provide insights into effector evolution in downy mildews and Phytophthora. BMC Genomics 16:741. https://doi.org/10.1186/s12864-015-1904-7.
81. Rujirawat T, Patumcharoenpol P, Lohnoo T, Yingyong W, Lerksuthirat T, Tangphatsornruang S, Suriyaphol P, Grenville-Briggs LJ, Garg G, Kittichotirat W, Krajaejun T. 2015. Draft genome sequence of the pathogenic oomycete Pythium insidiosum strain Pi-S, isolated from a patient with pythiosis. Genome Announc 3:e00574-15. https://doi.org/10.1128/genomeA.00574-15.
82. Berger H, Yacoub A, Gerbore J, Grizard D, Rey P, Sessitsch A, Compant S. 2016. Draft genome sequence of biocontrol agent Pythium oligandrum strain Po37, an oomycota. Genome Announc 4:e00215-16. https://doi.org/10.1128/genomeA.00215-16.
83. Jiang RHY, de Bruijn I, Haas BJ, Belmonte R, Löbach L, Christie J, van den Ackerveken G, Bottin A, Bulone V, Díaz-Moreno SM, Dumas B, Fan L, Gaulin E, Govers F, Grenville-Briggs LJ, Horner NR, Levin JZ, Mammella M, Meijer HJG, Morris P, Nusbaum C, Oome S, Phillips AJ, van Rooyen D, Rzeszutek E, Saraiva M, Secombes CJ, Seidl MF, Snel B, Stassen JHM, Sykes S, Tripathy S, van den Berg H, Vega-Arreguin JC, Wawra S, Young SK, Zeng Q, Dieguez-Uribeondo J, Russ C, Tyler BM, van West P. 2013. Distinctive expansion of potential virulence genes in the genome of the oomycete fish pathogen Saprolegnia parasitica. PLoS Genet 9:e1003272. https://doi.org/10.1371/journal.pgen.1003272.
84. Gobler CJ, Berry DL, Dyhrman ST, Wilhelm SW, Salamov A, Lobanov AV, Zhang Y, Collier JL, Wurch LL, Kustka AB, Dill BD, Shah M, VerBerkmoes NC, Kuo A, Terry A, Pangilinan J, Lindquist EA, Lucas S, Paulsen IT, Hattenrath-Lehmann TK, Talmage SC, Walker EA, Koch F, Burson AM, Marcoval MA, Tang YZ, Lecleir GR, Coyne KJ, Berg GM, Bertrand EM, Saito MA, Gladyshev VN, Grigoriev IV. 2011. Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. Proc Natl Acad Sci U S A 108:4352–4357. https://doi.org/10.1073/pnas.1016106108.
85. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kröger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jézéquel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, et al. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456:239–244. https://doi.org/10.1038/nature07410.
86. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kröger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306:79–86. https://doi.org/10.1126/science.1101156.
87. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury JM, Badger JH, Beszteri B, Billiau K, Bonnet E, Bothwell JH, Bowler C, Boyen C, Brownlee C, Carrano CJ, Charrier B, Cho GY, Coelho SM, Collén J, Corre E, Da Silva C, Delage L, Delaroque N, Dittami SM, Doulbeau S, Elias M, Farnham G, Gachon CMM, Gschloessl B, Heesch S, Jabbari K, Jubin C, Kawai H, Kimura K, Kloareg B, Küpper FC, Lang D, Wincker P. 2010. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. Nature 465:617–621. https://doi.org/10.1038/nature09016.
88. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Câmara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouël A, Lepère G, Malinsky

S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Bétermier M, Weissenbach J, Scarpelli C, Schächter V, Sperling L, Meyer E, Cohen J, Wincker P. 2006. Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. Nature 444:171–178. https://doi.org/10.1038/nature05230.

89. Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, Hopkins JF, Kuo A, Rensing SA, Schmutz J, Symeonidi A, Elias M, Eveleigh RJM, Herman EK, Klute MJ, Nakayama T, Oborník M, Reyes-Prieto A, Armbrust EV, Aves SJ, Beiko RG, Coutinho P, Dacks JB, Durnford DG, Fast NM, Green BR, Grisdale CJ, Hempel F, Henrissat B, Höppner MP, Ishida K-I, Kim E, Kořený L, Kroth PG, Liu Y, Malik S-B, Maier UG, McRose D, Mock T, Neilson JAD, Onodera NT, Poole AM, Pritham EJ, Richards TA, Rocap G, Roy SW, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65. https://doi.org/10.1038/nature11681.
90. Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32:W309–W312. https://doi.org/10.1093/nar/gkh379.
91. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18:1979–1990. https://doi.org/10.1101/gr.081612.108.
92. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503.
93. Ramsay L, Macaulay M, Degli Ivanissevich S, MacLean K, Cardle L, Fuller J, Edwards KJ, Tuvesson S, Morgante M, Massari A, Maestri E, Marmiroli N, Sjakste T, Ganal M, Powell W, Waugh R. 2000. A simple sequence repeat-based linkage map of barley. Genetics 156:1997–2005. https://doi.org/10.1093/nar/25.17.3389.
94. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
95. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.
96. Faith DP, Cranston PS. 1991. Could a cladogram this short have arisen by chance alone?: on permutation tests for cladistic structure. Cladistics 7:1–28. https://doi.org/10.1111/j.1096-0031.1991.tb00020.x.
97. Swofford LD. 2002. PAUP*: phylogenetic analysis using parsimony (* and other methods) version 4.0 beta. Sinauer Associates, Sunderland, MA.
98. Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165. https://doi.org/10.1093/bioinformatics/btr088.
99. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. https://doi.org/10.1093/sysbio/syq010.
100. Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21:390–392. https://doi.org/10.1093/bioinformatics/bti020.
101. Letunic I, Bork P. 2007. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128. https://doi.org/10.1093/bioinformatics/btl529.
102. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267. https://doi.org/10.1093/molbev/msj030.
103. Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24:1540–1541. https://doi.org/10.1093/bioinformatics/btn230.
104. Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071.
105. Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. Mol Phylogenet Evol 56:1115–1118. https://doi.org/10.1016/j.ympev.2010.04.024.
106. Lapierre P, Lasek-Nesselquist E, Gogarten JP. 2014. The impact of HGT on phylogenomic reconstruction methods. Brief Bioinform 15:79–90. https://doi.org/10.1093/bib/bbs050.
107. McCarthy CGP, Fitzpatrick DA. 2016. Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages. mSphere 1:e00195-16. https://doi.org/10.1128/mSphere.00195-16.
108. Burgess TI. 2015. Molecular characterization of natural hybrids formed between five related indigenous clade 6 Phytophthora species. PLoS One 10:e0134225. https://doi.org/10.1371/journal.pone.0134225.
109. Bertier L, Leus L, D'hondt L, de Cock AW, Höfte M. 2013. Host adaptation and speciation through hybridization and polyploidy in Phytophthora. PLoS One 8:e85385. https://doi.org/10.1371/journal.pone.0085385.
110. Husson C, Aguayo J, Revellin C, Frey P, Ioos R, Marçais B. 2015. Evidence for homoploid speciation in Phytophthora alni supports taxonomic reclassification in this species complex. Fungal Genet Biol 77:12–21. https://doi.org/10.1016/j.fgb.2015.02.013.

CHAPTER SIX

Multiple Approaches to Phylogenomic Reconstruction of the Fungal Kingdom

Charley G.P. McCarthy, David A. Fitzpatrick[1]
Maynooth University, Maynooth, County Kildare, Ireland
[1]Corresponding author: e-mail address: david.fitzpatrick@nuim.ie

Contents

1. Introduction 212
 1.1 The Phylogeny of the Fungal Kingdom 212
 1.2 *Saccharomyces cerevisiae* and the Origin of Modern Fungal Genomics 213
 1.3 Fungal Genomics and Phylogenomics Beyond the Yeast Genome 214
 1.4 The 1000 Fungal Genomes Project 215
2. Phylogenomic Reconstructions of the Fungal Kingdom 216
 2.1 Supermatrix Phylogenomic Analysis of Fungi 225
 2.2 Parsimony Supertree Phylogenomic Analysis of Fungi 232
 2.3 Bayesian Supertree Phylogenomic Analysis of Fungi 240
 2.4 Phylogenomics of Fungi Based on Gene Content 244
 2.5 Alignment-Free Phylogenomic Analysis of Fungi 247
3. A Genome-Scale Phylogeny of 84 Fungal Species From Seven Phylogenomic Methods 252
 3.1 Higher-Level Genome Phylogeny of the Fungal Kingdom 252
 3.2 Multiple Phylogenomic Methods Show Moderate Support for the Modern Designations of Mucoromycota and Zoopagomycota 254
 3.3 Pezizomycotina as a Benchmark for Phylogenomic Methodologies 255
 3.4 The Use of Phylogenomics Methods in Fungal Systematics 257
4. Concluding Remarks 259
Acknowledgments 260
References 260

Abstract

Fungi are possibly the most diverse eukaryotic kingdom, with over a million member species and an evolutionary history dating back a billion years. Fungi have been at the forefront of eukaryotic genomics, and owing to initiatives like the 1000 Fungal Genomes Project the amount of fungal genomic data has increased considerably over the last 5 years, enabling large-scale comparative genomics of species across the kingdom. In this chapter, we first review fungal evolution and the history of fungal genomics. We then review in detail seven phylogenomic methods and reconstruct the phylogeny of 84 fungal species from 8 phyla using each method. Six methods have seen extensive use in previous fungal studies, while a Bayesian supertree method is novel to fungal phylogenomics. We find that both established and novel phylogenomic methods can accurately reconstruct the fungal kingdom. Finally, we discuss the accuracy and suitability of each phylogenomic method utilized.

Advances in Genetics, Volume 100
ISSN 0065-2660
https://doi.org/10.1016/bs.adgen.2017.09.006

Author's personal copy

1. INTRODUCTION

1.1 The Phylogeny of the Fungal Kingdom

The fungi are one of the six kingdoms of life sensu Cavalier-Smith, sister to the animal kingdom, and are thought to span approximately 1.5 million species found across a broad range of ecosystems (Baldauf & Palmer, 1993; Berbee & Taylor, 1992; Cavalier-Smith, 1998; Hawksworth, 2001; Nikoh, Hayase, Iwabe, Kuma, & Miyata, 1994). While the overall fossil record of the fungi is poor due to their simple morphology, fungal fossils have been identified dating back to the Ordovician period approximately 400 million years ago (Redecker, 2000) and molecular clock analyses suggest that the fungi originated in the Precambrian eon approximately 0.76–1.06 billion years ago (Berbee & Taylor, 2010). Classic studies into fungal evolution were based on the comparison of morphological or biochemical characteristics; however, the broad range of diversity within the fungal kingdom had limited the efficacy of some of these studies (Berbee & Taylor, 1992; Heath, 1980; Léjohn, 1974; Taylor, 1978). Since the development of phylogenetic approaches within systematics and the incorporation of molecular data into phylogenetic analyses, our understanding of the evolution of fungi has improved substantially (Guarro, Gené, & Stchigel, 1999).

Initial phylogenetic analyses of fungal species had revealed that there were four distinct phyla within the fungal kingdom: the early-diverging Chytridiomycota and Zygomycota, and the Ascomycota and Basidiomycota. The Chytridiomycota grouping was later subject to revision (James et al., 2006), and in their comprehensive classification of the fungal kingdom in 2007 Hibbet et al. formally abandoned the phylum Zygomycota (Hibbett et al., 2007). Instead, Hibbet et al. treated zygomycete species as four *incertae sedis* subphyla (Entomophthoromycotina, Kickellomycotina, Mucoromycotina, and Zoopagomycotina) and subsequently described one subkingdom (the Dikarya) and seven phyla namely Chytridiomycota, Neocallimastigomycota, Blastocladiomycota, Microsporidia, Glomeromycota, Ascomycota, and

Basidiomycota (Hibbett et al., 2007). More recent phylogenetic classification of the zygomycetes has led to the circumscription of the Mucoromycota and Zoopagomycota phyla (Spatafora et al., 2016). Furthermore, recent phylogenetic analyses have shown that *Rozella* species occupy a deep branching position in the fungal kingdom (James et al., 2006; Jones, Forn, et al., 2011), the clade containing these species are now termed the Cryptomycota phylum (Jones, Forn, et al., 2011; Jones, Richards, Hawksworth, & Bass, 2011).

1.2 *Saccharomyces cerevisiae* and the Origin of Modern Fungal Genomics

In terms of genomic data, fungi are by far the highest sampled eukaryotic kingdom, with assembly data available for over 1000 fungal species on the NCBI's GenBank facility as of May 2017. Many of these species also have multiple strains sequenced (the most extreme example being *S. cerevisiae*, which has over 400 strain assemblies available on GenBank). This reflects both the ubiquity of fungi in many areas of biological and medical study and the relative simplicity of sequencing fungal genomes with modern sequencing technology. Fungi have been the exemplar group in eukaryote genetics and genomics, from the first determination of a nucleic acid sequence taken from *S. cerevisiae* by Holley and company in the late 1960s to the sequencing of the first eukaryotic genome in the mid-1990s (Goffeau et al., 1996; Holley et al., 1965). The genome of *S. cerevisiae* was sequenced through a massive international collaboration that grew to involve approximately 600 scientists in 94 laboratories and sequencing centers from across 19 countries between 1989 and 1996 (Engel et al., 2014; Goffeau et al., 1996; Goffeau & Vassarotti, 1991). Throughout the early 1990s, each of the standard 16 nuclear chromosomes of *S. cerevisiae*, sourced from the common laboratory strain 288C and its isogenic derivative strains AB972 and FY1679, was individually sequenced and published by participating researchers (Engel et al., 2014 briefly summarize each of these sequencing projects) with the initial publication of chromosome III involving 35 European laboratories on its own (Oliver et al., 1992). The complete genome sequence of *S. cerevisiae* 288C was finally published in 1996, with 5885 putative protein-coding genes and 275 transfer RNA genes identified across the genome's ~12 million base pairs (Goffeau et al., 1996).

In the intervening years the *S. cerevisiae* 288C reference genome has been constantly updated and refined as individual genes or chromosomes have been reanalyzed or even resequenced, and all of these revisions have been recorded and maintained by the Saccharomyces Genome Database (Fisk et al., 2006). It is worth noting, however, that such was the attention paid

to the original sequencing project by its contributors that the most recent major update of the *S. cerevisiae* 288C reference genome, a full resequencing of the derivative AB972 strain using far less labor-intensive modern sequencing and annotation techniques, made only minor alterations to the original genome annotation overall (Engel et al., 2014). Much of our understanding regarding the processes of genome evolution in eukaryotes since 1996 has also been derived from the study of the *S. cerevisiae* 288C genome, including the confirmation that the *S. cerevisiae* genome had undergone a whole-genome duplication (WGD) event (Kellis, Birren, & Lander, 2004; Wolfe & Shields, 1997), the effect of interspecific hybridization on genome complexity (De Barros Lopes, Bellon, Shirley, & Ganter, 2002), evidence that interdomain horizontal gene transfer (HGT) from prokaryotes into eukaryotes has occurred (Hall & Dietrich, 2007), to the ongoing development of an entirely synthetic genome through the Sc2.0 project (Annaluru et al., 2014).

1.3 Fungal Genomics and Phylogenomics Beyond the Yeast Genome

As more model organisms from other eukaryotic kingdoms had their genomes sequenced, *S. cerevisiae* 288C provided a useful comparison as the reference fungal genome, even for more complex eukaryotes like *Drosophila melanogaster*. However, the later sequencing of other model fungal species *Schizosaccharomyces pombe* and *Neurospora crassa* showed the limits of relying solely on *S. cerevisiae* as a reference for the entire fungal kingdom, particularly the latter; *N. crassa* was found to have a far larger genome than either *S. cerevisiae* or *S. pombe* and over 57% of genes predicted in *N. crassa* had no homolog in either of the other two sequenced fungal genomes (Galagan et al., 2003; Galagan, Henn, Ma, Cuomo, & Birren, 2005; Wood et al., 2002). Borne out of a lull in fungal genomic advances and the increasing sophistication of sequencing technology, the Fungal Genome Initiative (FGI) was set up by a number of research organizations in the early 2000s, under the aegis of the Broad Institute (Cuomo & Birren, 2010). Collaborators within the FGI were tasked with the sequencing and annotating the genomes of over 40 species from across the fungal kingdom, with a broad scope of species selected for analysis, medically significant human fungal pathogens like *Candida albicans* and *Aspergillus fumigatus*, commercially important species such as *Penicillium chrysogenum* and *Sclerotinia sclerotiorum*, as well as basal fungal species such as *Phycomyces blakesleeanus* (Cuomo & Birren, 2010). Between 2004 and 2012, in approximately the same amount

of time it had taken to sequence each individual chromosome of *S. cerevisiae* 288C in the 1990s, over 100 fungal genomes were sequenced and made publicly available on facilities like GenBank and the Joint Genome Institute (JGI)'s Genome Portal website (Benson et al., 2013; Grigoriev, Nordberg, et al., 2011).

The steady increase in genomic data available for fungi from the first decade of this century on, while still sampled mainly from the Ascomycota and Basidiomycota phyla, allowed for a greater range of fungal genomic analyses to be conducted. This included phylogenomic analyses of the fungal kingdom using a variety of different methods (which we will discuss in detail in the following section) and comparative investigations such as analysis of the evolution of pathogenicity in genera like *Candida* or *Aspergillus* (Butler et al., 2009; Galagan, Calvo, Cuomo, et al., 2005; Jackson et al., 2009), the extent of inter-/intrakingdom HGT both to and from fungal genomes (Fitzpatrick, Logue, & Butler, 2008; Marcet-Houben & Gabaldón, 2010; Richards et al., 2011; Szöllősi, Davín, Tannier, Daubin, & Boussau, 2015), identification of clusters of secondary metabolites (Keller, Turner, & Bennett, 2005; Khaldi et al., 2010), and syntenic relationships across *Saccharomyces* and *Candida* (Byrne & Wolfe, 2005; Fitzpatrick, O'Gaora, Byrne, & Butler, 2010). The wealth of genomic data available for some fungal orders or classes has allowed for easier automation of the sequencing and annotation of novel-related species, through the development of reference transcriptomic or proteomic data for gene prediction software such as AUGUSTUS or quality assessment software for genome assembly such as BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015; Stanke, Steinkamp, Waack, & Morgenstern, 2004).

1.4 The 1000 Fungal Genomes Project

The recent deluge of genomic data available for the fungal kingdom comes as a result of the 1000 Fungal Genomes Project, an initiative headed by the JGI. The project (which can be found at http://genome.jgi.doe.gov/pages/fungi-1000-projects.jsf) aims to provide genomic sequence data from at least one species from every circumscribed fungal family, either from projects headed by the JGI, projects which have been incorporated into the MycoCosm database or through community-led nomination and provision of sequencing material. The project has an inbuilt preference for sequencing projects arising from families with no sequenced species to date, or only one other reference genome at the time of nomination. Assembly and

annotation data are then hosted at the JGI's MycoCosm facility as well as other publically available databases (Grigoriev et al., 2014). This community-wide effort has led to a staggering increase in the number of fungal genomes available within the last 5 years; Grigoriev et al. (2014) quoted the number of genomes present in MycoCosm at over 250 at the end of 2013; as of May 2017 there are 772 fungal genomes available to download from the facility, with another 500 species nominated for sequencing. The project has seen a large increase particularly in the amount of data available from fungal phyla outside of the Dikarya, with 58 genomes currently available from the zygomycetes, the Chytridiomycota, Neocallimastigomycota, and Blastocladiomycota. There are many other fungal families with species yet to be nominated for sequencing, including many families from the Pezizomycotina subphylum within Ascomycota and the Chytridiomycota phylum. It is hoped that the wealth of fungal genomic data arising from the 1000 Fungal Genomes Project will help, among countless other scenarios, to fuel the search for novel biosynthetic products and to better understand the ecological effects of different families within the fungal kingdom (Grigoriev, Cullen, et al., 2011). The initiative will also enable the large-scale comparative analysis of hundreds of fungal species from across the fungal kingdom, including kingdom-level phylogenomic reconstructions.

2. PHYLOGENOMIC RECONSTRUCTIONS OF THE FUNGAL KINGDOM

Phylogenetic inference arising from molecular data has, in the past, predominately relied on single genes or small numbers of highly conserved genes or nuclear markers. While usually these markers make for robust individual phylogenies, potential conflicts can occur between individual phylogenies depending on the marker(s) used. The selection of such markers may also overlook other gene families which may be phylogenetically informative, such as gene duplication events or HGT events (Bininda-Emonds, 2004). With the advent of genome sequencing and the increasing sophistication of bioinformatics software and techniques, it has become common practice to reconstruct the evolutionary relationships of species by utilizing large amounts of phylogenetically informative genomic data. Such data can include ubiquitous or conserved genes, individual orthologous and paralogous gene phylogenies, shared genomic content, or compositional signatures of genomes (Fig. 1). Methods of phylogenomic analysis, in other words phylogenetic reconstruction of species using genome-scale data, have

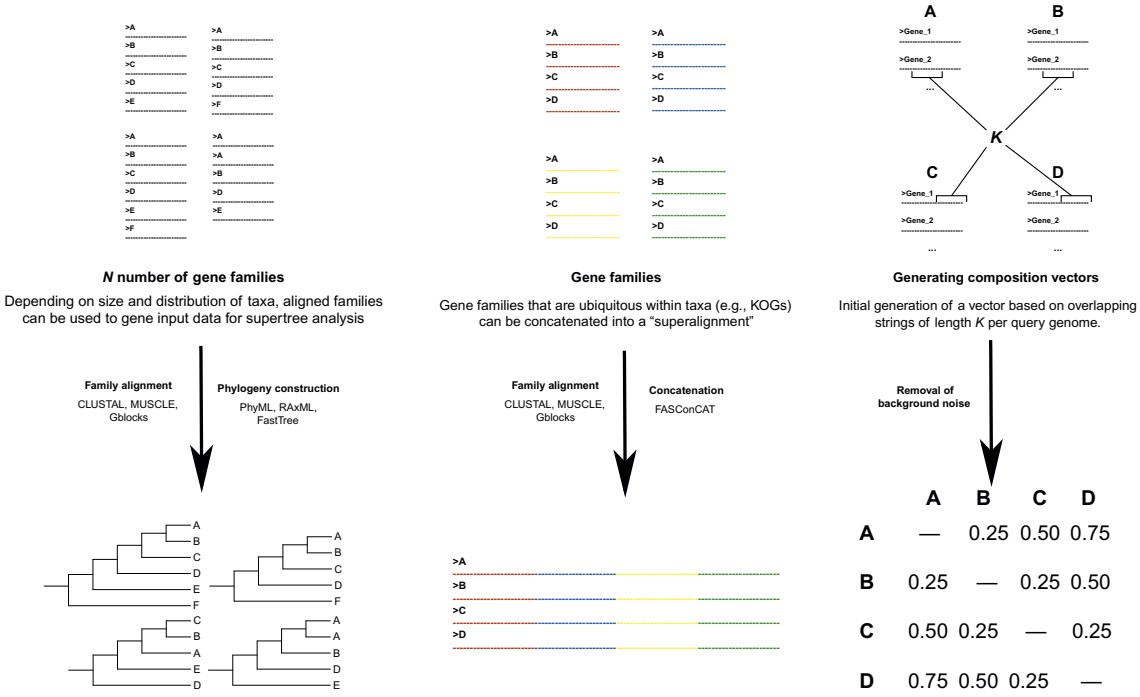


Fig. 1 See figure legend on next page

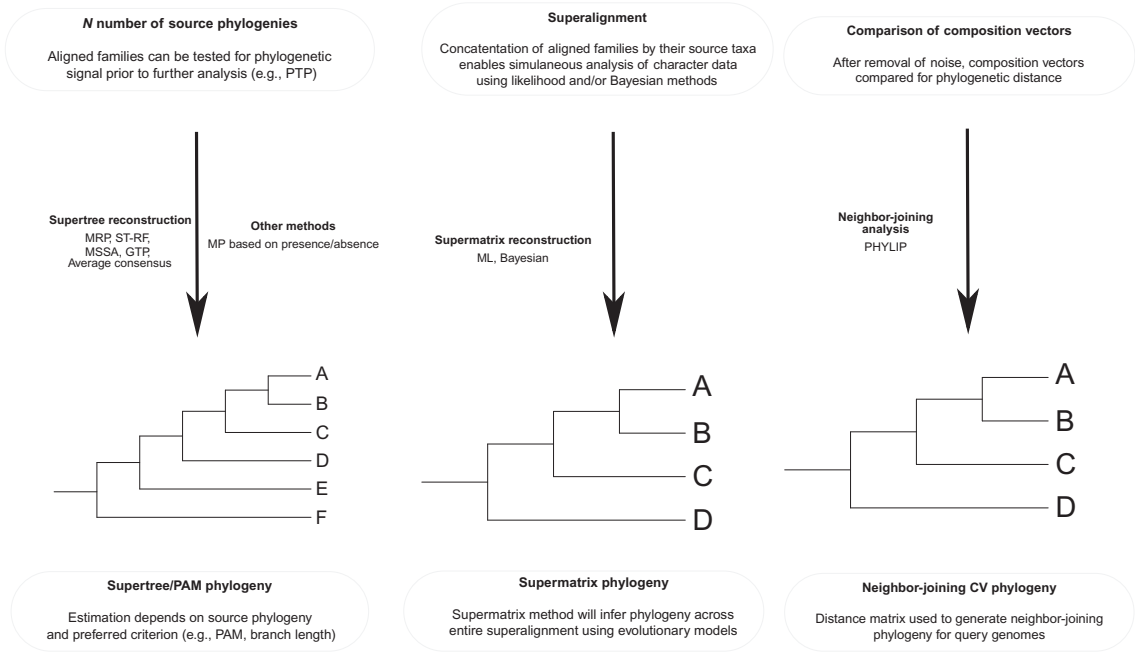


Fig. 1 Illustrative comparison of common phylogenomic methods. *Left*: supertree and presence–absence methods, *middle*: supermatrix methods, and *right*: composition vector methods.

all been developed for each of these types of potential phylogenetic marker and each comes with their advantages and disadvantages. Many phylogenomic analyses of the fungal kingdom have been carried out using these methods.

In this section, we review in turn each established approach to phylogenomic reconstruction from molecular data present in the literature and review each approach's application in previous fungal phylogenomic analyses. To demonstrate both the application and accuracy of all of these approaches to reconstructing phylogeny from genome-scale data, we have conducted our own phylogenomic analyses of the fungal kingdom using each method (Fig. 2). We have carried out such analyses to take advantage of both the greater coverage of the fungal kingdom arising from the 1000 Fungal Genomes Project and the advances in phylogenetic methodologies in the years following many of the analyses that we review below. In total, 84 fungal genomes from across 8 phyla (Table 1) were selected for our

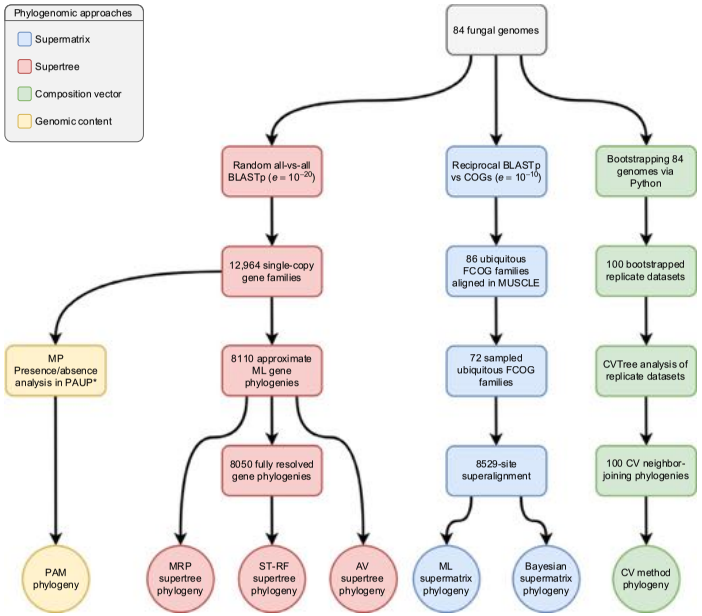


Fig. 2 Summary of the methodology of all 7 phylogenomic analyses of 84 fungal species carried out in this review.

Table 1 List of Species Used in Phylogenomic Analysis

Species	Phylum	Subphylum	Class	MycoCosm ID
Bipolaris maydis	Ascomycota	Pezizomycotina	Dothideomycetes	CocheC4_1
Cenococcum geophilum	Ascomycota	Pezizomycotina	Dothideomycetes	Cenge3
Hysterium pulicare	Ascomycota	Pezizomycotina	Dothideomycetes	Hyspu1_1
Zymoseptoria tritici	Ascomycota	Pezizomycotina	Dothideomycetes	Mycgr3
Aspergillus niger	Ascomycota	Pezizomycotina	Eurotiomycetes	Aspni7
Coccidioides immitis	Ascomycota	Pezizomycotina	Eurotiomycetes	Cocim1
Endocarpon pusillum	Ascomycota	Pezizomycotina	Eurotiomycetes	EndpusZ1
Exophiala dermatitidis	Ascomycota	Pezizomycotina	Eurotiomycetes	Exode1
Phaeomoniella chlamydospora	Ascomycota	Pezizomycotina	Eurotiomycetes	Phach1
Blumeria graminis	Ascomycota	Pezizomycotina	Leotiomycetes	Blugr1
Botrytis cinerea	Ascomycota	Pezizomycotina	Leotiomycetes	Botci1
Arthrobotrys oligospora	Ascomycota	Pezizomycotina	Orbiliomycetes	Artol1
Dactylellina haptotyla	Ascomycota	Pezizomycotina	Orbiliomycetes	Monha1
Pyronema omphalodes	Ascomycota	Pezizomycotina	Pezizomycetes	Pyrco1
Tuber melanosporum	Ascomycota	Pezizomycotina	Pezizomycetes	Tubme1
Coniochaeta ligniaria	Ascomycota	Pezizomycotina	Sordariomycetes	Conli1
Hypoxylon sp. *EC38*	Ascomycota	Pezizomycotina	Sordariomycetes	HypEC38_3

Magnaporthe grisea	Ascomycota	Pezizomycotina	Sordariomycetes	Maggr1
Neurospora crassa	Ascomycota	Pezizomycotina	Sordariomycetes	Neucr_trp3_1
Ophiostoma piceae	Ascomycota	Pezizomycotina	Sordariomycetes	Ophpic1
Phaeoacremonium minimum	Ascomycota	Pezizomycotina	Sordariomycetes	Phaal1
Xylona heveae	Ascomycota	Pezizomycotina	Xylonomycetes	Xylhe1
Candida albicans	Ascomycota	Saccharomycotina	Saccharomycetes	Canalb1
Lipomyces starkeyi	Ascomycota	Saccharomycotina	Saccharomycetes	Lipst1_1
Ogataea polymorpha	Ascomycota	Saccharomycotina	Saccharomycetes	Hanpo2
Saccharomyces cerevisiae	Ascomycota	Saccharomycotina	Saccharomycetes	SacceM3707_1
Saitoella complicata	Ascomycota	Taphrinomycotina	N/A	Saico1
Pneumocystis jirovecii	Ascomycota	Taphrinomycotina	Pneumocystidomycetes	Pneji1
Schizosaccharomyces cryophilus	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schcy1
Schizosaccharomyces japonicus	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schja1
Schizosaccharomyces octosporus	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schoc1
Schizosaccharomyces pombe	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schpo1
Protomyces lactucaedebilis	Ascomycota	Taphrinomycotina	Taphrinomycetes	Prola1
Taphrina deformans	Ascomycota	Taphrinomycotina	Taphrinomycetes	Tapde1_1
Agaricus bisporus	Basidiomycota	Agaricomycotina	Agaricomycetes	Agabi_varbur_1

Continued

Table 1 List of Species Used in Phylogenomic Analysis—cont'd

Species	Phylum	Subphylum	Class	MycoCosm ID
Auricularia subglabra	Basidiomycota	Agaricomycotina	Agaricomycetes	Aurde3_1
Botryobasidium botryosum	Basidiomycota	Agaricomycotina	Agaricomycetes	Botbo1
Fibulorhizoctonia	Basidiomycota	Agaricomycotina	Agaricomycetes	Fibsp1
Gloeophyllum trabeum	Basidiomycota	Agaricomycotina	Agaricomycetes	Glotr1_1
Heterobasidion annosum	Basidiomycota	Agaricomycotina	Agaricomycetes	Hetan2
Jaapia argillacea	Basidiomycota	Agaricomycotina	Agaricomycetes	Jaaar1
Punctularia strigosozonata	Basidiomycota	Agaricomycotina	Agaricomycetes	Punst1
Serendipita indica	Basidiomycota	Agaricomycotina	Agaricomycetes	Pirin1
Serpula lacrymans	Basidiomycota	Agaricomycotina	Agaricomycetes	SerlaS7_3_2
Sistotremastrum suecicum	Basidiomycota	Agaricomycotina	Agaricomycetes	Sissu1
Sphaerobolus stellatus	Basidiomycota	Agaricomycotina	Agaricomycetes	Sphst1
Wolfiporia cocos	Basidiomycota	Agaricomycotina	Agaricomycetes	Wolco1
Calocera cornea	Basidiomycota	Agaricomycotina	Dacrymycetes	Calco1
Dacryopinax primogenitus	Basidiomycota	Agaricomycotina	Dacrymycetes	Dacsp1
Basidioascus undulatus	Basidiomycota	Agaricomycotina	Geminibasidiomycetes	Basun1
Cryptococcus neoformans	Basidiomycota	Agaricomycotina	Tremellomycetes	Cryne_JEC21_1
Cutaneotrichosporon oleaginosus	Basidiomycota	Agaricomycotina	Tremellomycetes	Triol1

Wallemia sebi	Basidiomycota	Agaricomycotina	Wallemiomycetes	Walse1
Leucosporidium creatinivorum	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Leucr1
Microbotryum lychnidis-dioicae	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Micld1
Rhodotorula graminis	Basidiomycota	Pucciniomycotina	Microbotryomycetes	Rhoba1_1
Mixia osmundae	Basidiomycota	Pucciniomycotina	Mixiomycetes	Mixos1
Puccinia graminis	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucgr2
Tilletiaria anomala	Basidiomycota	Ustilaginomycotina	Exobasidiomycetes	Tilan2
Malassezia sympodialis	Basidiomycota	Ustilaginomycotina	Malasseziomycetes	Malsy1_1
Sporisorium reilianum	Basidiomycota	Ustilaginomycotina	Ustilaginomycetes	Spore1
Ustilago maydis	Basidiomycota	Ustilaginomycotina	Ustilaginomycetes	Ustma1
Allomyces macrogynus	Blastocladiomycota	N/A	Blastocladiomycetes	GCA_000151295.1
Catenaria anguillulae	Blastocladiomycota	N/A	Blastocladiomycetes	Catan2
Batrachochytrium dendrobatidis	Chytridiomycota	N/A	Chytridiomycetes	GCA_000149865.1
Rhizoclosmatium globosum	Chytridiomycota	N/A	Chytridiomycetes	Rhihy1
Spizellomyces punctatus	Chytridiomycota	N/A	Chytridiomycetes	Spipu1
Gonapodya prolifera	Chytridiomycota	N/A	Monoblepharidomycetes	Ganpr1
Rozella allomycis	Cryptomycota	N/A	N/A	Rozal1_1
Rhizophagus irregularis	Mucoromycota	Glomeromycotina	Glomeromycetes	Gloin1

Continued

Table 1 List of Species Used in Phylogenomic Analysis—cont'd

Species	Phylum	Subphylum	Class	MycoCosm ID
Mortierella elongate	Mucoromycota	Mortierellomycotina	N/A	Morel2
Phycomyces blakesleeanus	Mucoromycota	Mucoromycotina	N/A	Phybl2
Rhizopus oryzae	Mucoromycota	Mucoromycotina	N/A	Rhior3
Umbelopsis ramanniana	Mucoromycota	Mucoromycotina	N/A	Umbra1
Anaeromyces robustus	Neocallimastigomycota	N/A	Neocallimastigomycetes	Anasp1
Neocallimastix californiae	Neocallimastigomycota	N/A	Neocallimastigomycetes	Neosp1
Orpinomyces sp. *C1A*	Neocallimastigomycota	N/A	Neocallimastigomycetes	Orpsp1_1
Piromyces finnis	Neocallimastigomycota	N/A	Neocallimastigomycetes	Pirfi3
Basidiobolus meristosporus	Zoopagomycota	Entomophthoromycotina	Basidiobolomycetes	Basme2finSC
Conidiobolus thromboides	Zoopagomycota	Entomophthoromycotina	Entomophthoromycetes	Conth1
Coemansia reversa	Zoopagomycota	Kickxellomycotina	N/A	Coere1
Linderina pennispora	Zoopagomycota	Kickxellomycotina	N/A	Linpe1
Martensiomyces pterosporus	Zoopagomycota	Kickxellomycotina	N/A	Marpt1
Ramicandelaber brevisporus	Zoopagomycota	Kickxellomycotina	N/A	Rambr1

Genome data from MycoCosm (http://genome.jgi.doe.gov/programs/fungi/index.jsf) has previously been published and MycoCosm ID is given in final column. GENBANK accessions given for *Allomyces macrogynus* and *Batrachochytrium dendrobatidis*.

large-scale phylogenomic reconstructions of the fungal kingdom. Where possible, we included at least one published representative genome from each order covered by the 1000 Fungal Genomes Project in our dataset. All genomic data were ultimately obtained from the JGI's MycoCosm facility (Grigoriev et al., 2014). Our analyses include the first phylogenomic reconstruction of fungi carried out using a Bayesian supertree approach, and the first large-scale gene content approach to fungal phylogenomics that has been conducted in at least a decade. We discuss, in brief, the methodology and the results of each reconstruction and their accuracy (or otherwise) in reconstructing the phylogeny of both basal fungal lineages and the Dikarya. In Section 3, we discuss the overall phylogeny of the fungal kingdom arising from our analyses and compare with previous literature.

2.1 Supermatrix Phylogenomic Analysis of Fungi

The two best-established alignment-based approaches to reconstructing phylogeny on a genomic scale are the "supertree" method, in which a consensus phylogeny is derived from many individual gene phylogenies (discussed in Section 2.2), and the "supermatrix" method which we discuss here. Supermatrix method phylogeny is the simultaneous analysis of a phylogenetic matrix, also referred to as a "superalignment," constructed from all available character data from a given set of taxa. Generally supermatrices are constructed from concatenating highly conserved markers (e.g., rRNA genes, mitochondrial markers) for small-scale multigene phylogenies, and from homologs of conserved orthologous genes (known as COGs, or sometimes KOGs in eukaryotes) for genome-scale phylogenies (de Queiroz & Gatesy, 2007; Koonin et al., 2004). Supermatrix approaches can also incorporate statistically powerful maximum-likelihood and Bayesian methods of phylogenomic analysis. Described in simple terms, given an alignment of sequences and a suitable evolutionary model, maximum-likelihood phylogenetic analysis examines all possible trees by their possible parameters (e.g., topology, site support, branch length) and returns the most likely phylogenetic tree for the alignment (Page & Holmes, 1998). Similarly, Bayesian analysis incorporates phylogenetic likelihoods to calculate the posterior probability of a phylogeny, which is the probability of that phylogeny given the alignment data (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001).

One advantage of a supermatrix approach to phylogenomic analysis over a supertree approach is the retention of character evidence in analysis in the former approach; most supertree methods can be considered estimations

Author's personal copy

using individual trees based on summarized character data, at least two steps removed from any actual sequence data, whereas a supermatrix approach entails direct analysis of combined character data (Creevey & McInerney, 2009; de Queiroz & Gatesy, 2007). Supermatrix methods also have the potential to resolve deep branches and reveal so-called hidden supports within phylogenies that supertree methods may overlook (de Queiroz & Gatesy, 2007). However, supermatrix analysis requires ubiquitous sequences from all taxa being investigated, which restricts the available pool of character data and may overlook miss important phylogenetic information from phylogenies with gene deletion, gene duplication, or horizontal gene transfer events that supertree methods can utilize (Creevey & McInerney, 2009). Compositional biases may also have an effect on supermatrix methods, though phylogenetic models have been developed which can ameliorate errors that these biases may induce during analysis (Lartillot, Brinkmann, & Philippe, 2007; Lartillot & Philippe, 2004). In practice, many phylogenomic analyses utilize both supertree and supermatrix methods in tandem to reconstruct phylogeny in a "total evidence" approach (Kluge, 1989) and will often comment on the taxonomic congruence (or otherwise) of the resulting phylogenies.

2.1.1 Fungal Phylogenomics Using the Supermatrix Approach

Supermatrix analysis has been widely used in fungal phylogenomics. One of the initial comparisons of individual gene phylogenies with genome-scale species phylogenies used a maximum-parsimony analysis among other methods to reconstruct the phylogeny of seven *Saccharomyces* species and *C. albicans*; the authors showed that incongruence among individual gene phylogenies could be resolved with high support using a concatenated alignment (Rokas, Williams, King, & Carroll, 2003). Initial genome-based phylogenies of Ascomycota using 17 genomes and both supertree and supermatrix methods resolved both Pezizomycotina and Saccharomycotina, as well as placing *S. pombe* as an early-diverging branch within Ascomycota (Robbertse, Reeves, Schoch, & Spatafora, 2006). Robbertse et al. (2006) generated a superalignment of 195,664 amino acid characters in length derived from 781 gene families, which produced identical topologies under both neighbor-joining and maximum-likelihood criteria. The first large-scale phylogenomic analysis of fungi used a 67,101-character superalignment derived from 531 eukaryotic COGs found in 21 fungal genomes, all of which were sampled from Ascomycota and Basidiomycota (Kuramae, Robert, Snel, Weiß, & Boekhout, 2006). A more extensive phylogenomic analysis from the same year produced 2 highly congruent genome phylogenies from 42 fungal

genomes using 2 methods: a matrix representation with parsimony (MRP) supertree derived from 4805 single-copy gene families (which we discuss in greater detail in Section 2.2.1), and a 38,000-character superalignment derived from 153 ubiquitous gene families (Fitzpatrick, Logue, Stajich, & Butler, 2006).

Most of the relationships resolved in Fitzpatrick et al. (2006) were further supported by a 31,123-character superalignment from 69 proteins conserved in up to 60 fungal genomes generated by Marcet-Houben, Marceddu, and Gabaldón (2009), although they found a large degree of topological conflict within a 21-species Saccharomycotina clade (Marcet-Houben & Gabaldón, 2009; Marcet-Houben et al., 2009). A later follow-up analysis to Fitzpatrick et al. (2006) by Medina, Jones, and Fitzpatrick (2011) reconstructed the phylogeny of 103 fungal species by performing Bayesian analysis on a 12,267-site superalignment derived from 87 gene families with a phyletic range of over half of their dataset, in addition to supertree analysis (Medina et al., 2011). A recent phylogenomic analysis of 46 fungal genomes, including 25 zygomycetes species, reconstructed the phylogeny of the early-diverging fungal lineages using a 60,383-character superalignment (Spatafora et al., 2016). Another recent phylogenomic analysis used a 28,807-site superalignment derived from 136 gene families from 40 eukaryotic genomes to investigate the evolution of sourcing carbon from algal and plant pectin in early-diverging fungi (Chang et al., 2015). Finally, a comparison of the dynamics of genome evolution between 28 Dikarya species and cyanobacteria used a supermatrix phylogeny of 24,514 amino acid characters from 529 fungal gene families with large phyletic range as a scaffold to infer rates of intrakingdom HGT within Dikarya that were near similar to those within cyanobacteria (Szöllősi et al., 2015).

To extend the analyses described above, we carried out supermatrix analysis using maximum-likelihood and Bayesian methods on a superalignment constructed from orthologous genes conserved throughout 84 species from 8 phyla within the fungal kingdom.

2.1.2 Phylogenomic Reconstruction of 84 Fungal Species From 72 Ubiquitous Gene Families Using Maximum-Likelihood and Bayesian Supermatrix Analysis

A reciprocal BLASTp search was carried out between all protein sequences from our 84-genome dataset and 458 core orthologous genes (COGs) from *S. cerevisiae* obtained from the CEGMA dataset, with an *e*-value cutoff of 10^{-10} (Camacho et al., 2009; Parra, Bradnam, & Korf, 2007), from which

456 COG families were retrieved (2 *S. cerevisiae* COGs did not return any homologs). From these, 86 ubiquitous fungal COG families, i.e., families containing a homolog from all 84 input species, were identified. Each ubiquitous fungal COG family was aligned in MUSCLE, and conserved regions of each alignment were sampled in Gblocks using the default parameters (Castresana, 2000; Edgar, 2004). Fourteen alignments did not retain any character data after Gblocks filtering and were removed from further analysis. The remaining 72 sampled alignments were concatenated into a superalignment of 8529 aligned positions using the Perl program FASconCat (Kück & Meusemann, 2010). This superalignment was bootstrapped 100 times using Seqboot (Felsenstein, 1989), and maximum-likelihood phylogenetic trees were generated for each individual replicate using PhyML with an LG + I + G amino acid substitution model as selected by ProtTest (Darriba, Taboada, Doallo, & Posada, 2011; Guindon et al., 2010). A consensus phylogeny was generated from all 100 individual replicate phylogenies using CLANN (Creevey & McInerney, 2005). Markov Chain Monte Carlo (MCMC) Bayesian phylogenetic inference was carried out on the same superalignment using PhyloBayes MPI with the default CAT + GTR amino acid substitution model, running 2 chains for 1000,000 iterations and sampling every 100 iterations (Lartillot & Philippe, 2004; Lartillot, Rodrigue, Stubbs, & Richer, 2013). Both chains were judged to have converged after 100,000 iterations and a consensus Bayesian phylogeny was generated with a burn-in of 1000 trees. Both supermatrix phylogenies were visualized using the Interactive Tree of Life (iTOL) website and annotated according to the NCBI's taxonomy database (Federhen, 2012; Letunic & Bork, 2016). Both supermatrix phylogenies were rooted at *Rozella allomycis*, which is the most basal species in evolutionary terms in our dataset (Jones, Forn, et al., 2011) and is the root for all the phylogenies we present hereafter (Figs. 3 and 4).

2.1.3 Supermatrix Analyses of 84 Fungal Species Accurately Reconstructs the Fungal Kingdom

We reconstructed the phylogeny of the fungal kingdom by generating a superalignment of 72 concatenated ubiquitous gene families and performing ML analysis using PhyML and Bayesian analysis using a parallelized version of PhyloBayes. Both ML and Bayesian analysis reconstruct the phylogeny of our fungal dataset with a high degree of accuracy relative to other kingdom phylogenies in the literature and in most cases recover the eight fungal phyla in our dataset (Figs. 3 and 4). Here, we discuss the results of both our analyses with regard to the basal fungal lineages, and the two Dikarya

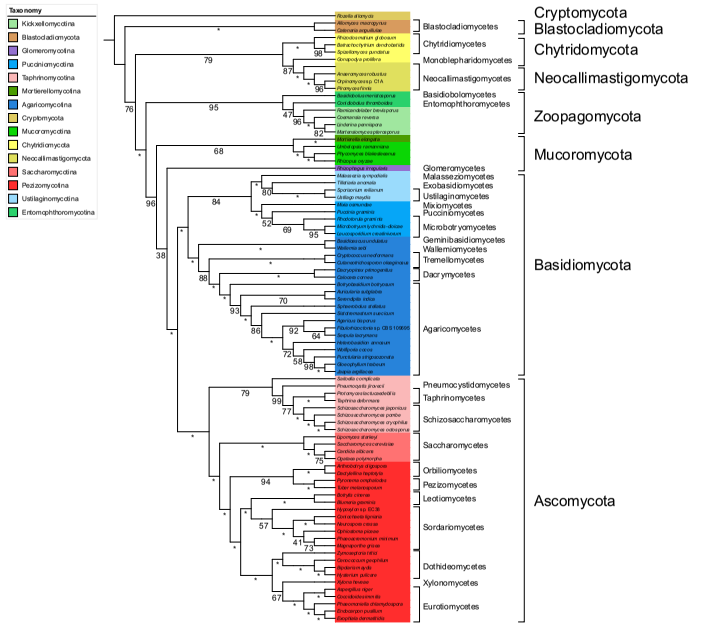


Fig. 3 ML phylogeny of 84 fungal species from a 8529-character superalignment derived from 72 ubiquitous fungal COG families sampled in Gblocks using PhyML with a LG+I+G model. Bootstrap supports shown on branches. Maximum bootstrap support designated with an *asterisk* (*).

phyla. Further in this chapter, we use these supermatrix analyses as the point of comparison for our other phylogenomic methods.

2.1.3.1 Basal Fungi

In our ML supermatrix phylogeny, Blastocladiomycota emerge as the earliest-diverging fungi with maximum bootstrap support (henceforth abbreviated to BP) after rooting at *R. allomycis* (Fig. 3). Chytridiomycota and Neocallimastigomycota are placed as sister clades with 79% BP, surprisingly the Chytridiomycota species *Gonapodya prolifera* branches as sister to Neocallimastigomycota (87% BP). The Chytridiomycetes class is monophyletic with maximum bootstrap support, as is the Neocallimastigomycetes class (Fig. 3). The former zygomycetes phylum Zoopagomycota is strongly supported as a monophyletic clade with 95% BP (Fig. 3). The other former

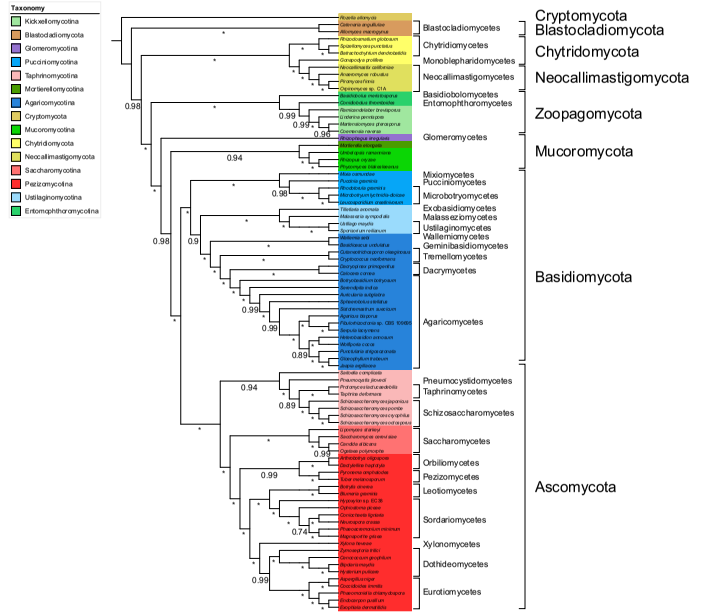


Fig. 4 Bayesian phylogeny of 84 fungal species from an 8529-character superalignment derived from 72 ubiquitous fungal COG families sampled in Gblocks using PhyloBayes MPI with a CAT+GTR model. Posterior probabilities shown on branches with a burn-in of 1000 trees. Maximum posterior probability support designated with an *asterisk* (*).

zygomycetes phylum Mucoromycota is paraphyletic and split between a clade containing the Mucoromycotina and Mortierellomycotina species *Mortierella elongata* that has 68% BP, and the Glomeromycotina species *Rhizophagus irregularis* branching basal to Dikarya with lower support (38% BP). The placement of Mucoromycota as the closest phyla to Dikarya has near-maximum support (96% BP) which matches other analysis (Spatafora et al., 2016).

The Bayesian supermatrix phylogeny is in near-total agreement with the ML phylogeny in resolving the relationships of the basal fungi in our dataset (Fig. 4). The relationship between Chytridiomycota and Neocallimastigomycota in the Bayesian phylogeny mirrors that seen in the ML phylogeny, with all branches receiving maximum support as monophyletic with a Bayesian posterior probability (henceforth abbreviated to PP) equal to 1 (Fig. 4). The

Zoopagomycota are monophyletic with full support, with a topology matching the ML phylogeny with strong branch support throughout (Fig. 4). There is also a close association between the three Mucoromycota subphyla: Glomeromycota branches earlier in the Bayesian phylogeny than in the ML phylogeny, which receives maximum support in the Bayesian phylogeny, and the sister relationship between Mucoromycotina and *M. elongata* receives strong support (0.94 PP) in the Bayesian phylogeny (Fig. 4). Both the ML and Bayesian place the Mucoromycota as the basal phylum that is most closely related to Dikarya (Fig. 4).

2.1.3.2 Basidiomycota

In the ML phylogeny, the three subphyla within Basidiomycota are fully resolved with maximum BP, with 84% BP for the placement of Ustilagomycotina and Pucciniomycotina as sister clades (Fig. 3). *Basidioascus undulatus* and *Wallemia sebi* branch at the base of Agaricomycotina with maximum BP, while the other classes with the subphyla are all fully supported. There is also high support (88% BP) for the placement of Tremellomycetes as sister to Dacrymycetes and Agaricomycetes (Fig. 3). The Tremellomycetes, including *Cryptococcus neoformans*, are monophyletic. The Dacrymycetes are also monophyletic with maximum BP. The forest saprophyte *Botryobasidium botryosum* is placed at the base of the Agaricomycetes, which has some strong intraclade resolution with weaker branch supports toward the tips of the clade (Fig. 3). *Malassezia sympodialis*, a commensal fungi of humans and animals, is placed at the base of the Ustilagomycotina. The Exobasidiomycetes species *Tilletiaria anomala* branches between *M. sympodialis* and the Ustilagomycetes. The Pucciniomycotina are monophyletic with full support (Fig. 3). The most highly represented Pucciniomycotina class, the Microbotryomycetes, are monophyletic with 69% BP (Fig. 3).

The Bayesian phylogeny reflects the ML phylogeny in its resolution of the Basidiomycota as monophyletic with full support (Fig. 4). The phylogeny places Pucciniomycotina at the base of the phylum with maximum support. Resolution of branches within Pucciniomycotina is substantially improved under Bayesian phylogeny (Fig. 4). There is high support (0.9 PP) for a sister relationship between Ustilagomycotina and Agaricomycotina (Fig. 4). The Exobasidiomycetes species *T. anomala* now branches at the base of the Ustilagomycotina, which is resolved with maximum PP. There is maximum support for the placement of *M. sympodialis* as sister to the Ustilagomycetes, which are monophyletic (Fig. 4). As in the ML phylogeny, *B. undulatus* and *W. sebi* branch at the base of Agaricomycotina with maximum support, while

the other classes with the subphyla all have maximum support and have similar topology under Bayesian analysis. There is a large improvement in the support of branches in the Agaricomycotina in the Bayesian phylogeny relative to the ML phylogeny (Fig. 4).

2.1.3.3 Ascomycota

Both the ML and Bayesian supermatrix phylogenies display near-identical topologies for the Ascomycota, and Bayesian analysis shows stronger support for some branches toward the tips of the phylogeny than the ML phylogeny does (Figs. 3 and 4). The three subphyla within Ascomycota are fully resolved, with maximum BP support for Saccharomycotina and Pezizomycotina and 79% BP for the monophyly of Taphrinomycotina in the ML phylogeny (contrast with 0.94 PP for the monophyly of Taphrinomycotina in the Bayesian phylogeny; Figs. 3 and 4). The placement of Taphrinomycotina as an ancestral clade within Ascomycota is fully supported, and within Taphrinomycotina, there is high support (77% BP/0.89 PP) for a sister relationship between Schizosaccharomycetes and Taphrinomycetes. Six of the seven classes within Pezizomycotina in our dataset with two or more representatives (i.e., all bar Xylonomycetes) are monophyletic, most of which receive maximum BP and/or PP support. Many of the relationships between classes are also well supported in both phylogenies, with lower support (67% BP) for a sister relationship between the Xylonomycetes species *Xylona heveae* and the Eurotiomycetes class in the ML phylogeny; in the Bayesian phylogeny *X. heveae* branches sister to a clade containing Dothideomycetes and Eurotiomycetes with maximum PP support (Figs. 3 and 4). The Dothideomycetes are monophyletic in both phylogenies and branch into two clades with high support under both ML and Bayesian reconstruction (Figs. 3 and 4). The Orbiliomycetes and Pezizomycetes are placed as the most basal Pezizomycotina classes, with strong support (94% BP/0.99 BP) for a sister relationship (Figs. 3 and 4). The Leotiomycetes and Sordariomycetes are also placed as a sister clades with maximum support in both phylogenies. The major difference in the resolution of the Sordariomycetes between the supermatrix phylogenies is the stronger branch supports within the order under Bayesian analysis (Figs. 3 and 4).

2.2 Parsimony Supertree Phylogenomic Analysis of Fungi

The most common supertree methods for reconstructing genome phylogenies are grounded in parsimony methods, in which changes to character states (i.e., evolutionary events such as presence of a given taxon in a tree or even a tree branch) are calculated and phylogeny is reconstructed using

as little state changes as possible. The first supertree construction method to see widespread use in large-scale phylogenetic and phylogenomic analysis was the MRP method. MRP, which was developed independently by Baum (1992) and Ragan (1992), enables the use of source phylogenies with overlapping or missing taxa in generating a consensus phylogeny (Baum, 1992; Ragan, 1992). The method generates a matrix (referred to as a Baum–Ragan matrix) where each column represents one internal branch in each given source phylogeny such that the number of columns within the matrix is equal to the number of internal branches across all source phylogenies, and assigns a score of 1 to taxa from a given source phylogeny *P* which are present in the clade defined by internal branch *A*, 0 to taxa present in *P* but not within the clade defined by *A*, and ? to taxa that are not present in *P* (Creevey & McInerney, 2009). The Baum–Ragan matrix is then subject to parsimony analysis, with equal weighting given to each source phylogeny, and reconstructs the supertree phylogeny with the minimum of evolutionary changes required which includes all taxa represented across all source phylogenies. Similar parsimony methods, most notably gene tree parsimony (Slowinski & Page, 1999), extend MRP to include source phylogenies containing duplicated taxa; however, we do not cover such methods in this subsection. Parsimony-based supertree methods like MRP are generally quite accurate in reconstructing phylogeny for large datasets, although some issues have been observed (which we discuss in Section 2.3).

2.2.1 Matrix Representation With Parsimony Analysis in Fungal Phylogenomics

Many phylogenomic analyses of fungi have used parsimony methods. The first large-scale phylogenomic analysis of fungi to use MRP in supertree reconstruction was by Fitzpatrick et al. (2006), who carried out a phylogenomic reconstruction of fungi using 42 genomes from Dikarya and the zygomycete *Rhizopus oryzae* using both supertree and supermatrix methods (Fitzpatrick et al., 2006). Using a random BLASTp approach to identify homologous gene families, where randomly selected query sequences are sequentially searched against a full database and then both query sequences and homologs (if any) are sequentially removed from the database, Fitzpatrick et al. (2006) utilized 4805 single-copy gene phylogenies for MRP supertree reconstruction using the software package CLANN (Creevey & McInerney, 2005, 2009). The MRP phylogeny resolved the Pezizomycotina and Saccharomycotina subphyla within Ascomycota and inferred the Sordariomycetes and the Leotiomycetes as sister classes within Pezizomycotina. The MRP phylogeny also resolved two major clades

within the Saccharomycotina: a monophyletic clade of species that translate the codon CTG as serine instead of leucine (the "CTG clade"), and a grouping of species that have undergone whole genome duplication (the "WGD clade") and their closest relatives. The authors compared the MRP phylogeny with a maximum-likelihood supermatrix phylogeny reconstructed using 38,000 characters from 153 gene families (as detailed in the previous subsection); both were highly congruent with conflict only in the placement of the sole Dothideomycetes species represented, *Stagonospora nodourum*. The authors also complemented their MRP phylogeny with two other supertree methods implemented in CLANN: a most similar supertree analysis (MSSA) method phylogeny which was identical to the MRP supertree (Creevey et al., 2004) and an average consensus (AV) method phylogeny based on branch lengths (Lapointe & Cucumel, 1997), which the authors believed to suffer from long-branch attraction in the erroneous placement of some species within the WGD clade in Saccharomycotina (Fitzpatrick et al., 2006). A follow-up analysis to Fitzpatrick et al. (2006) by Medina et al. (2011) using 103 genomes was extended to include multicopy gene families using the gene tree parsimony (GTP) method and successfully resolved the major groupings within the fungal kingdom (Medina et al., 2011). Using both a random BLASTp and a Markov Clustering Algorithm (MCL)-based approach with varying inflation values to identify orthologous gene families, the authors used as many as 30,012 single and paralogous gene phylogenies as input for supertree reconstruction.

As a follow-up to the supertree reconstructions of the fungal kingdom by Fitzpatrick et al. (2006) and Medina et al. (2011), we ran supertree analysis for 84 fungal species using MRP and AV methods and source phylogenies identified via a random BLASTp approach described later.

2.2.2 Phylogenomic Reconstruction of 84 Fungal Species From 8110 Source Phylogenies Using MRP and AV Supertree Methods

Following Fitzpatrick et al. (2006), families of homologous protein sequences within our 84-genome dataset were identified using BLASTp with an *e*-value cutoff of 10^{-20} by randomly selecting a query sequences from our database, finding all homologous sequences via BLASTp (Camacho et al., 2009), and removing the entire family from the database before reformatting and repeating. 12,964 single-copy gene families, which contained no more than one homolog from 4 or more taxa, were identified. Each single-copy gene family was aligned in MUSCLE, and conserved regions of each alignment were sampled using Gblocks with the default parameters (Castresana, 2000; Edgar, 2004). Sampled alignments were tested for phylogenetic signal using

the PTP test as implemented in PAUP* with 100 replicates (Faith & Cranston, 1991; Swofford, 2002). 8110 sampled alignments which retained character data after Gblocks filtering and passed the PTP test were retained for phylogenomic reconstruction. 8110 approximately maximum-likelihood gene phylogenies were generated with FastTree, using the default JTT+CAT protein evolutionary model (Price, Dehal, & Arkin, 2010). All 8110 single-copy gene phylogenies were used to generate a matrix representation with parsimony (MRP) supertree using CLANN, with 100 bootstrap replicates (Creevey & McInerney, 2005). To complement the MRP supertree, an average consensus (AV) supertree was generated from the same input dataset in CLANN, with 100 bootstrap replicates. Both supertrees were visualized in iTOL and annotated according to the NCBI's taxonomy database. Both supertrees were rooted at *R. allomycis* (Figs. 5 and 6).

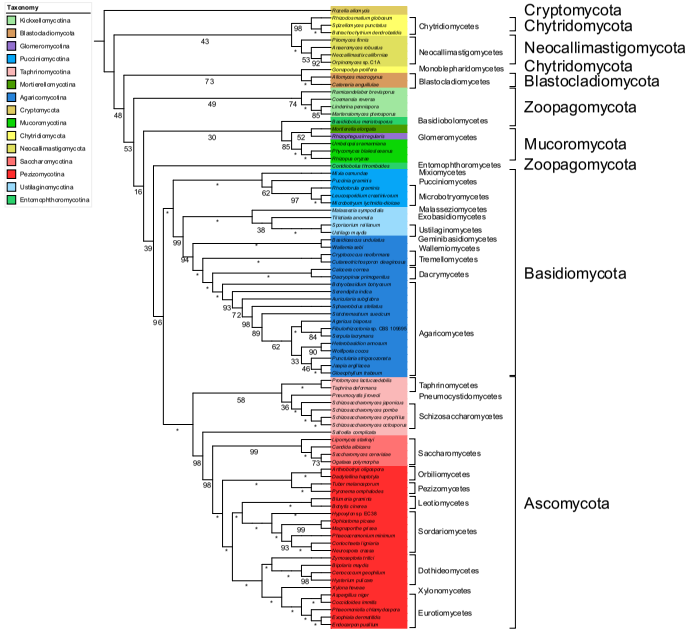


Fig. 5 Matrix representation with parsimony (MRP) phylogeny of 84 fungal species derived from 8110 source phylogenies. Bootstrap supports shown on branches. Maximum bootstrap support designated with an *asterisk* (*).

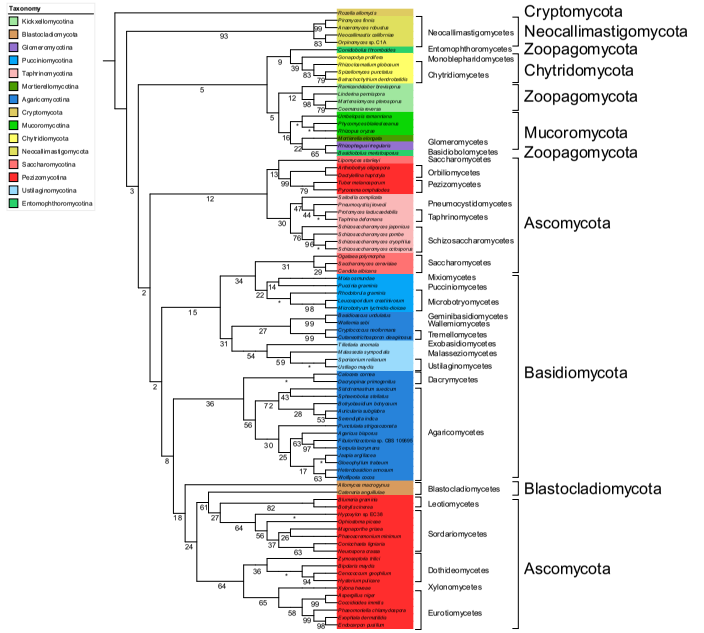


Fig. 6 Average consensus (AV) phylogeny of 84 fungal species derived from 8110 source phylogenies. Bootstrap supports shown on branches. Maximum bootstrap support designated with an *asterisk* (*).

2.2.3 MRP Phylogenomic Analysis of 84 Fungal Species Is Highly Congruent With Supermatrix Phylogenomic Analyses

We reconstructed the overall phylogeny of 8110 single-copy source phylogenies from our 84-genome dataset using an MRP supertree method analysis as implemented in CLANN (Fig. 5). MRP supertree reconstruction of the fungal kingdom recovers the majority of the eight fungal phyla in our dataset and is effective in resolving the Dikarya. However, there is poorer resolution of some of the basal phyla due to smaller taxon sampling perhaps having a negative influence on the distribution of basal taxa within our source phylogenies (we return to this in Section 3). Overall our MRP analysis is highly congruent with our supermatrix phylogenies detailed earlier, with some variation in the placement and resolution in some branches. We discuss the results of our MRP analysis for the basal fungal lineages and both Dikarya

phyla and note some of the congruences and incongruences where noteworthy with our supermatrix phylogenies (Figs. 3–5).

2.2.3.1 Basal Fungi

After rooting at *R. allomycis*, the Neocallimastigomycota and Chytridiomycota (bar *G. prolifera*) emerge as the earliest-diverging fungal lineages. *G. prolifera* branches basal to the Blastocladiomycota with 73% BP (Fig. 5). This arrangement of the Neocallimastigomycota, Chytridiomycota, and Blastocladiomycota has poor support in general (43% BP for a sister relationship between Neocallimastigomycotina and 4 Chytridiomycota species); however with the exception of the aforementioned placement of *G. prolifera* the individual phyla receive maximum or near-maximum support as monophyletic (Fig. 5). Zoopagomycota is paraphyletic in our MRP phylogeny; a monophyletic Kicxellomycotina clade receives 74% BP support (Fig. 5), while as in the supermatrix phylogenies (Figs. 3 and 4) Entomophthoromycotina is paraphyletic. In our MRP analysis, *Basidiobolus meristosporus* branches at the base of Mucoromycota and *Conidiobolus thromboides* branches at the base of Dikarya, but those relationships are poorly supported (30% and 39% BP, respectively; Fig. 5). The Glomeromycotina species *R. irregularis* branches sister to the Mortierellomycota representative *M. elongata* with weak support (52% BP), but Murocomycota (the placement of Glomeromycotina, Mortierellomycota, and Mucoromycotina) receives higher support (85% BP). The monophyly of Mucoromycotina is also fully supported (Fig. 5). Overall many of the associations between basal phyla we observed in our supermatrix phylogenies are present in our MRP analysis as well; however, the overall placement of the basal fungal lineages varies between supermatrix and MRP analyses, such as the placement of Blastocladiomycota as a later-diverging clade than either Chytridiomycota or Neocallimastigomycota under MRP supertree analysis (Figs. 3–5).

2.2.3.2 Basidiomycota

The Basidiomycota are recovered with maximum support in our MRP phylogeny (Fig. 5). The Pucciniomycotina emerge as the most basal subphylum with maximum support, with *Mixia osmundae* branching at the base of the subphylum and *Puccinia graminis* placed as sister to the Microbotryomycetes (who are monophyletic with 97% BP). This reflects the topology of Pucciniomycotina seen in our supermatrix phylogenies (Figs. 3–5). The Ustilagomycotina and Agaricomycotina branch as sister subphyla with 99% BP and both are monophyletic; the former is fully supported

at the branch level and the latter has 94% BP. *M. sympodialis* is placed at the base of Ustilagomycotina, reflecting the resolution of the Ustilagomycotina under ML supermatrix analysis (Figs. 3 and 5). In the Agaricomycotina, *W. sebi* and *B. undulatus* branch at the base of the subphylum with maximum support. The three larger classes from Agaricomycotina in our dataset (Agaricomycetes, Dacrymycetes, and Tremellomycetes) are all monophyletic and are recovered with maximum support (Fig. 5). The MRP phylogeny of the Basidiomycota is highly congruent overall with the supermatrix phylogenies, with comparable branch support (Figs. 3–5).

2.2.3.3 Ascomycota

Our MRP phylogeny supports the Ascomycota as a monophyletic group with maximum BP (Fig. 5). There is greater support along many deeper branches in the Ascomycota in our MRP phylogeny than in our ML supermatrix phylogeny and support is comparable with our Bayesian phylogeny; we ascribe this to a larger abundance of smaller source phylogenies containing closely related Ascomycotina species in our dataset (Figs. 3–5). Taphrinomycotina emerges as the earliest-diverging lineage but is paraphyletic; *Saitoella complicata* branches as an intermediate between Taphrinomycotina and a Saccharomycotina–Pezizomycotina clade with 98% BP, while the remaining members are monophyletic with weak support (58% BP). *Pneumocystis jirovecii* is placed as a sister taxon to Schizosaccharomycetes in our MRP analysis with weak support (36% BP); in the supermatrix phylogenies it was sister to Taphrinomycetes. The Taphrinomycetes and Schizosaccharomycetes themselves are monophyletic with maximum BP (Fig. 5). The Saccharomycotina are monophyletic with 99% BP (Fig. 5). The six larger classes (i.e., all bar Xylonomycetes) in our dataset from Pezizomycotina are all supported as monophyletic and receive maximum BP, with Pezizomycetes and Orbiliomycetes branching as the basal sister clades (Fig. 5). The MRP phylogeny mirrors Bayesian supermatrix reconstruction in placing a single origin for three classes (Xylonomycetes, Eurotiomycetes, and Dothideomycetes) with maximum support (Figs. 4 and 5). As in both supermatrix phylogenies, Dothideomycetes are split into two clades with high or maximum support. In the Sordariomycetes, MRP analysis reflects the ML supermatrix phylogeny in placing *Hypoxylon* sp. EC58 at the base of the class (Figs. 3 and 5). The MRP phylogeny of the Ascomycota is highly congruent with both of our supermatrix phylogenies with comparable branch supports, which is aided by the broad range of genomic data available for the phylum (Figs. 3–5).

2.2.4 Average Consensus Phylogenomic Reconstruction of 84 Fungal Species Is Affected by Long-Branch Attraction Artifacts

To complement our MRP phylogeny, we generated an average consensus (AV) method supertree phylogeny (Fig. 6) using the same set of input phylogenies as implemented in CLANN following Fitzpatrick et al. (2006). AV phylogeny infers phylogeny based on the branch lengths of source phylogenies, by computing the average value of the path-length matrices associated with said source phylogenies, and then using a least-squares method to find the source matrix closest to this average value (Lapointe & Cucumel, 1997). The tree that is associated with this source matrix is the average consensus phylogeny for the total set of source phylogenies, and the method is thought to work best with a set of source phylogenies of similar size (Lapointe & Cucumel, 1997). Our AV phylogeny was rooted at *R. allomycis* (Fig. 6). Given the results we obtained from our AV phylogeny, we believe that the method is susceptible to long-branch attraction (Felsenstein, 1978), as reported by Fitzpatrick et al. (2006). Long-branch attraction occurs when two very divergent taxa or clades with long branch lengths (i.e., many character changes occurring over time) are inferred as each other's closest relative due to convergent evolution of a given character (e.g., amino acid substitution), and is a common problem in parsimony and distance-based methods (Bergsten, 2005; Felsenstein, 1978). In the AV phylogeny, we recovered the two Blastocladiomycota species in our dataset within a large paraphyletic Pezizomycotina clade (Fig. 6). Additionally, the Ascomycota are paraphyletic: one clade containing two Pezizomycotina classes (Pezizomycetes and Orbiliomycetes), the Taphrinomycotina and the Saccharomycotina species *Lipomyces starkeyi* places at the base of Dikarya, while three Saccharomycotina species (including *S. cerevisiae*) appear as a sister clade to Pucciniomycotina (Fig. 6). The Agaricomycotina are also paraphyletic; Tremellomycetes and two basal Basidiomycota species (*B. undulatus* and *W. sebi*) appear closer to Ustilagomycota (Fig. 6). Many of the supports throughout the tree are extremely poor (almost all of the incongruences we highlighted all have <40% BP), which seems to be another effect of long-branch attraction (Fig. 6). Due to the breadth of fungal taxa, we have sampled for our multiple analyses, and the timescale of the evolution of the fungal kingdom being approximately 1 billion years old, it is unsurprising that a method using branch lengths to infer a close relationship between actually distantly related species that both have long branches, a classic example of the "Felsenstein Zone" (Bergsten, 2005; Huelsenbeck & Hillis, 1993). Ultimately, our AV phylogeny (Fig. 6) seems to confirm one of the concerns of Fitzpatrick et al. (2006) in a much more stark fashion that the AV method is not appropriate

for large-scale phylogenomic reconstructions containing taxa sampled from across many phyla without prior predictive analysis of the potential for long branch attraction in such datasets (Su & Townsend, 2015).

2.3 Bayesian Supertree Phylogenomic Analysis of Fungi

While parsimony-based supertree reconstructions are generally reliable, concerns have been raised in the past as to some of the underlying methodology of MRP reconstruction and the effects that factors like input tree sizes (Pisani & Wilkinson, 2002; Wilkinson, Thorley, Pisani, Lapointe, & McInerney, 2004). There has long been the desire for a supertree method that infers phylogeny from source trees with more statistical rigor like Bayesian and maximum-likelihood inference methods. While Bayesian and ML analyses are the standard for supermatrix reconstruction, such methods have been difficult to implement in the past for supertree analysis due to computational limitations, most of which is down to the necessity of tree searching for the best supertree (i.e., calculating likelihoods for all possible supertrees given a set of source phylogenies).

It is only in recent years that phylogenomic inference based on ML and Bayesian methods has been implemented for supertree analysis; one such model for supertree likelihood estimation was first described by Steel and Rodrigo (2008) and then refined the following year (Bryant & Steel, 2009; Steel & Rodrigo, 2008). The Steel and Rodrigo method of likelihood estimation (henceforth referred to as ST-RF) is based on modeling the incongruences between input gene phylogenies and a corresponding unknown or provided supertree phylogeny. Two recent implementations of ST-RF ML analysis have been reported: the first a heuristic method of estimating approximate ML supertrees based on subtree pruning and regrafting implemented in the Python software L.U.St. by Akanni, Creevey, Wilkinson, and Pisani (2014), and the second a heuristic Bayesian MCMC criterion by Akanni, Wilkinson, Creevey, Foster, and Pisani (2015) implemented in the Python software package p4 (Akanni et al., 2014, 2015; Foster, 2004). Akanni et al. (2015) tested the Bayesian MCMC implementation on both a large kingdom-wide metazoan dataset and a smaller Carnivora dataset, notably the analysis produced a Bayesian supertree in full agreement with both the literature on metazoan relationships and a previous MRP supertree analysis on the same dataset (Holton & Pisani, 2010).

No parametric supertree reconstruction has been carried out for the fungal kingdom to date, and with that in mind we reconstructed the phylogeny

of our 84-genome dataset with the MCMC Bayesian criterion developed by Akanni et al. (2015) using a slightly amended gene phylogeny dataset from our MRP and AV supertree phylogenies.

2.3.1 Heuristic MCMC Bayesian Supertree Reconstruction of 84 Fungal Genomes From 8050 Source Phylogenies

MCMC Bayesian supertree analysis was carried out on the single-copy phylogeny dataset using the ST-RF model as implemented in p4 (Akanni et al., 2015; Foster, 2004; Steel & Rodrigo, 2008). As ST-RF analysis is currently only implemented in p4 for fully bifurcating phylogenies, 60 phylogenies were removed from the total single-copy phylogeny dataset, for an input dataset of 8050 gene phylogenies. Two separate MCMC analyses with 4 chains each were ran for 30,000 generations with $\beta = 1$, sampling every 20 generations. The analyses converged after 30,000 generations, and a consensus phylogeny based on posterior probability of splits was generated from 150 supertrees sampled after convergence following Akanni et al. (2015). This consensus phylogeny was visualized in iTOL and annotated according to the NCBI's taxonomy database, and rooted at *R. allomycis* (Fig. 7).

2.3.2 Supertree Reconstruction With a Heuristic MCMC Bayesian Method Highly Congruent With MRP and Supermatrix Phylogenies

Using 8050 of the 8110 individual gene phylogenies which we identified in our MRP supertree analysis, we have reconstructed the first parametric supertree of the fungal kingdom (Fig. 7). We selected the ST-RF MCMC Bayesian supertree reconstruction method implemented in p4 for reconstruction over the heuristic method implemented in L.U.St. due to tractability issues regarding large datasets in the latter method (Akanni et al., 2014, 2015). Two ST-RF analyses were carried out for 30,000 generations, and the analyses were adjudged to have converged after 20,000 generations. To construct a phylogeny from our MCMC analysis, we sampled 150 trees generated after convergence and built a consensus tree in p4, where branch support values are the estimated posterior probabilities of a given split (i.e., bipartition) within a phylogeny (Fig. 7). Our ST-RF MCMC analysis is highly congruent with both our MRP supertree phylogeny and supermatrix phylogenies and supports the monophyly of the majority of the eight fungal phyla in our dataset (Fig. 7). Below, we detail the resolution of the basal and Dikarya lineages under ST-RF analysis.

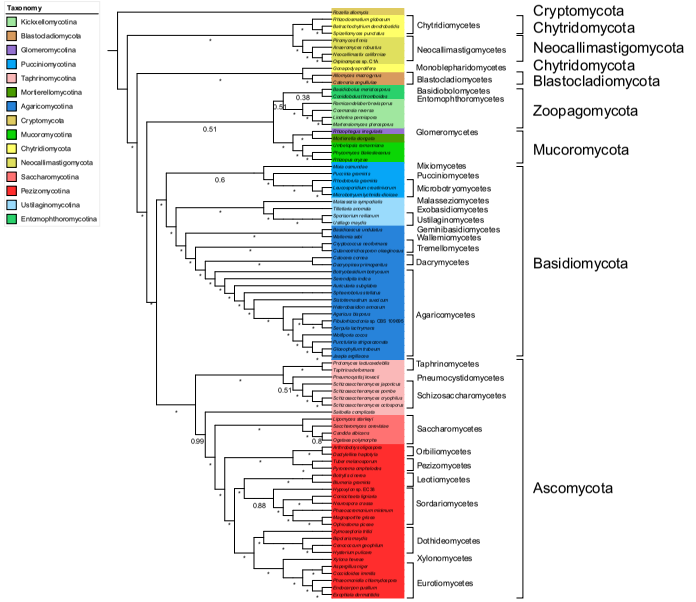


Fig. 7 MCMC Bayesian supertree phylogeny of 84 fungal species derived from 8050 fully bifurcating source phylogenies. Phylogeny generated in p4 using ST-RF model of maximum-likelihood supertree estimation running for 30,000 generations with $\beta = 1$. Posterior probabilities of bipartition(s) within 150 trees sampled after convergence shown on branches. Maximum posterior probability support designated with an *asterisk* (*).

2.3.2.1 Basal Fungi

After rooting at *R. allomycis*, the Neocallimastigomycota and Chytridiomycota (except *G. prolifera*) form a sister group relationship with maximum PP (Fig. 7). The Blastocladiomycota emerge after this branch, and the Chytridiomycota species *G. prolifera* branches as sister to the phylum with maximum PP (Fig. 7). There is weak support (0.51 PP) for a monophyletic clade containing both former zygomycetes phyla Zoopagomycota and Mucoromycota as sister clades (Fig. 7). Notably, unlike MRP and supermatrix analysis, ST-RF phylogeny places the Entomophthoromycotina as monophyletic but with very weak support (0.38 PP). There is also weak support for the placement the Entomophthoromycotina as basal within Zoopagomycota. Kickxellomycotina are monophyletic with maximum

support. The monophyly of Mucoromycota is fully supported, with *R. irregularis* (Glomeromycotina) and *M. elongata* (Mortierellomycotina) branching as sister taxa (Fig. 7).

2.3.2.2 Basidiomycota

The Basidiomycota are supported as a monophyletic group with maximum PP (Fig. 7). There is weak support for the monophyly of Pucciniomycotina (0.6 PP); however, the deeper branches within the subphyla are all fully supported and their topology reflects both the MRP supertree and ML supermatrix phylogenies discussed earlier (Figs. 3, 5, and 7). There is full support for a sister relationship between Ustilaginomycotina and Agaricomycotina, and both these subphyla are fully supported. In Ustilaginomycotina, *M. sympodialis* is the basal species with maximum support (Fig. 7), as in our supermatrix and MRP supertree phylogenies. The topology of the Agaricomycotina is nearly identical on the class level to both the MRP and supermatrix phylogenies, with *B. undulatus* and *W. sebi* branching as basal species, the Tremellomycetes forming a monophyletic intermediate clade, and a fully supported sister relationship between the Dacrymycetes and the Agaricomycetes (Fig. 7).

2.3.2.3 Ascomycota

The monophyly of the Ascomycota is supported with maximum PP, as is the monophyly of two of the three subphyla in Ascomycota (Fig. 7). Taphrinomycotina is paraphyletic as in the MRP phylogeny, with *S. complicata* branching sister to Saccharomycota with near-maximum support (0.99 PP) and the remaining Taphrinomycotina species are placed as a monophyletic clade with maximum PP (Figs. 5 and 7). The Taphrinomycetes branch at the base of the Taphrinomycotina clade, and there is weak support (0.51 PP) for the placement of *P. jirovecii* as sister to the Schizosaccharomycotina (Fig. 7). The Saccharomycotina are fully supported as monophyletic (1.0 PP) with *L. starkeyi* placed at the base of the subphyla. The monophyly of the Pezizomycotina is also fully supported and there is maximum support for the monophyly of the six larger represented classes within the subphylum (Fig. 7). Additionally, the relationships between the individual classes within Pezizomycotina are identical to the topology seen in both the MRP supertree phylogeny and the ML supermatrix phylogeny (Figs. 3, 5, and 7). The Orbiliomycetes and Pezizomycetes branch as the earliest-diverging clades within Pezizomycotina with maximum PP, the Sordariomycetes and Leotiomycetes are sister classes with maximum PP and a monophyletic

Author's personal copy

Dothideiomycetes–Xylonomycetes–Eurotiomycetes clade receives maximum PP (Fig. 7).

2.4 Phylogenomics of Fungi Based on Gene Content

A common alternative to phylogenomic reconstruction using gene phylogenies is to take a "gene content" approach in which evolutionary relationships between species are derived from shared genomic content, such as the presence or absence of conserved orthologous genes (COGs) or the overall proportion of shared genes between two species, working under the assumption that species that share more of their genome are closely related (Snel, Bork, & Huynen, 1999; Snel, Huynen, & Dutilh, 2005). In the case of presence–absence analyses, a matrix can be constructed for the species under investigation, which can then have their phylogeny reconstructed via parsimony methods. Analyses based on proportions of shared genes can entail the construction of distance matrices for all input species, with values equal to the inverse ratio of shared genes (i.e., if two species share 75% of their genes, their distance is 0.25), which is then used to construct a neighbor-joining phylogeny. The advantages of such approaches are the relative tractability of parsimony or distance-based gene content methods, and their potential to use more information from genomes rather than the sourcing of data from smaller sets of gene families required by supertree or supermatrix approaches (Creevey & McInerney, 2009). However, the gene content approach is by its very nature a "broad strokes" approach and can ignore potentially important phylogenetic information from individual gene phylogenies such as HGT events, and assumes the same evolutionary history for missing orthologs or genomic content among species (Page & Holmes, 1998).

2.4.1 Gene Content Approaches to Phylogenomics in Fungi

Gene content approaches to phylogenomic reconstruction have seen application in a number of phylogenomics studies, although its greatest use predated many of the now common supertree and supermatrix methods. One of the earliest phylogenomic studies used a distance-based approach based on shared gene content to reconstruct the phylogeny of 13 unicellular species, including *S. cerevisiae* (Snel et al., 1999). Another study used a weighted distance matrix approach to reconstruct the phylogeny of 23 prokaryote and eukaryote species, including *S. cerevisiae* and partial genomic data from *S. pombe* (Tekaia, Lazcano, & Dujon, 1999). The most extensive gene content-based phylogenomic reconstruction of fungi was an analysis of 21 fungal genomes and 4 other eukaryote genomes in 2006 (Kuramae et al., 2006). In their

analysis, the authors generated a presence–absence matrix (PAM) of 4852 COGs in fungal genomes as a complement to a supermatrix phylogeny using 531 concatenated proteins which was reconstructed using four different methods (MP, ML, neighbor-joining, and Bayesian inference). The authors reconstructed the phylogeny of all 25 genomes using this PAM and found that the PAM phylogeny differ most in the placement of *S. pombe* within Saccharomycetes as opposed to its basal position in Ascomycetes as seen in their supermatrix reconstructions (Kuramae et al., 2006).

To test the accuracy of inferring the phylogeny of a large genomic dataset using simple parsimony methods based on shared genomic content, we carried out a simple parsimony-based PAM phylogenomic reconstruction of 84 fungal species based on the presence of orthologs from single-copy gene families.

2.4.2 Phylogenomic Reconstruction of 84 Fungal Species Based on COG PAM

A simple PAM was generated for 84 fungal genomes based on their representation across 12,964 single-copy gene families identified via the random BLASTp approach detailed in Section 2.2. Parsimony analysis of this matrix was carried out using PAUP* with 100 bootstrap replicates. The resultant consensus phylogeny generated by PAUP* was visualized using iTOL and annotated according to the NCBI's taxonomy database. The phylogeny was rooted at *R. allomycis* (Fig. 8).

2.4.3 COG PAM Approach Displays Erroneous Placement of Branches Within Dikarya

We generated a simple PAM phylogeny for the 84 fungal genomes in our dataset by checking for the presence or absence of all 84 species across the 12,964 single-copy phylogenies we generated during our supertree analyses via random BLASTp searches and using the PAM as input for parsimony analysis (Fig. 8). The simple PAM phylogeny shows some level of congruence with the other phylogenomic analyses described here along certain branches (Fig. 8). The monophyly of Neocallimastigomycota, Chytridiomycota, and Blastocladiomycota all displays maximum or near-maximum BP, and there is 72% BP for a sister relationship between Chytridiomycota and Neocallimastigomycota (Fig. 8). The Zoopagomycota and Mucoromycota are placed in one monophyletic clade with 82% BP, with the two Entomophthoromycotina species in our dataset branching as closely related to the Mucoromycota (Fig. 8). However, some glaring conflicts with the

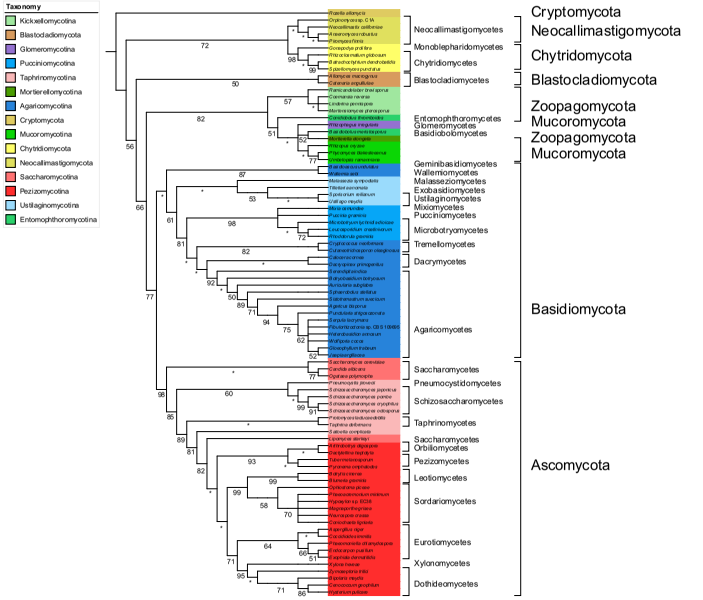


Fig. 8 Maximum parsimony (MP) phylogeny of 84 fungal species based on the presence of homologs from 12,964 single-copy gene families identified via random BLASTp searches. Bootstrap supports shown on branches. Maximum bootstrap support designated with an *asterisk* (*).

other phylogenomic methods we carried out can be observed within the Dikarya lineage. Most notably, the Agaricomycotina and Saccharomycotina are both paraphyletic in our single-copy PAM approach; for the former, *W. sebi* and *B. undulatus* branch at the base of the Basidiomycota adjacent to Ustilagomycotina, while in the latter three of the four Saccharomycotina (excluding *L. starkeyi*) species branch in our dataset at the base of the Ascomycota, implying that Taphrinomycotina diverged later than Saccharomycotina (Fig. 8). There is uncertain placement of clades within the Basidiomycota subphyla in particular. In the Ascomycota, the Taphrinomycotina are paraphyletic and *S. complicata* branches adjacent to *L. starkeyi*. The monophyly of all six larger Pezizomycotina classes are supported, many with relatively high or even maximum BP; however, there is poorer resolution of many relationships within these classes with the clearest examples

being the Sordariomycetes and Eurotiomycetes (Fig. 8). In short, our PAM phylogeny is able to retrieve relationships with some level of accuracy within the fungal kingdom, but the method lacks the ability to resolve some of the more divergent relationships within fungi to the degree that some of our supermatrix or supertree phylogenies have illustrated.

2.5 Alignment-Free Phylogenomic Analysis of Fungi

Another alternative to the alignment-based methods of phylogenomic reconstruction we have detailed earlier is the use of a string-based comparison of genomes to infer phylogeny, based on the assumption that under such comparisons each species should have a characteristic genomic signature that can act as a phylogenetic marker (Delsuc, Brinkmann, & Philippe, 2005). Some analyses have thus used signatures such as distribution of protein folds or frequency of oligonucleotides from genetic and genomic data to infer phylogeny (Campbell, Mrázek, & Karlin, 1999; Lin & Gerstein, 2000; Pride, Meinersmann, Wassenaar, & Blaser, 2003). The most widely used alignment-free phylogenomic method, the composition vector (CV) approach, was first implemented by Qi, Luo, and Hao (2004) and by Qi, Wang, and Hao (2004), who used the approach to reconstruct the phylogeny of 87 prokaryote species from 11 bacterial and 2 archaeal phyla (Qi, Wang, et al., 2004). In their analysis, the authors detail the CV method for reconstructing phylogeny using genome-scale data, which we recount as follows:

1. Given a nucleic acid or amino acid sequence of length L in a genome, count the appearances of overlapping strings (i.e., oligonucleotides or oligopeptides) of a length K and construct a frequency vector of length 4^K for nucleic acid sequences and 20^K for amino acid sequences.
2. Subtract background noise, to account for random mutation at the molecular level, from each frequency vector to generate an overall composition vector for a given genome.
3. Calculate a distance matrix for the set of composition vectors corresponding to the set of input genomes.
4. Generate a neighbor-joining phylogeny from the distance matrix using software such as Neighbor or PAUP*.

The main advantages of the composition vector approach over traditional alignment-based methods of inferring phylogeny are the removal of artificial selection of phylogenetic markers from the process of reconstruction (the only variable in the method is K, the length of overlapping oligopeptides), and the relative speed with which the approach can infer phylogeny for large

datasets over alignment-based supertree or supermatrix methods. Hence, it may be useful for quick phylogenomic identification of newly sequenced genomes against published data and as an independent verification step of previous alignment-based phylogenetic or phylogenomic analysis (Wang, Xu, Gao, & Hao, 2009). On that point however, interpreting the accuracy or otherwise of CV phylogenomic reconstructions is generally dependent on prior knowledge of the phylogeny of given taxa derived from alignment-based phylogenetic or phylogenomic analyses. An approach to inferring phylogeny based on nucleotide or amino acid composition may also be susceptible to compositional biases, and there has not been to the best of our knowledge a rigorous analysis of the potential effect these may have on accuracy of phylogenomic inference, as there have been for the supertree or supermatrix methods referred to earlier.

2.5.1 Composition Vector Method Phylogenomics of Fungi

Many of the phylogenomic analyses using the CV method have analyzed large prokaryotic datasets or broad global datasets sampled from many phyla or kingdoms across the three domains of life, whose phylogenies were recovered with quality comparative to alignment-based phylogenomic analyses. The most extensive application of the composition vector approach in fungal phylogenomics was an 85-genome analysis by Wang et al. (2009) using a CV implementation in the software program CVTree (Qi, Luo, et al., 2004; Wang et al., 2009). For their analysis, Wang et al. (2009) reconstructed the phylogeny of the fungal kingdom using 81 genomes from 4 fungal phyla (Basidiomycota, Ascomycota, Chytridiomycota, and Mucoromycota) as well as the microsporidian *Encephalitozoon cuniculi* and 3 eukaryotic outgroup taxa. The authors described the resolution of both the Basidiomycota and Ascomycota in detail in their analysis; the three subphyla within Basidiomycota were recovered but with poor bootstrap support due to issues with taxon sampling (only 12 Basiomycota species had genomic data at the time of the analysis), while the main focus of the authors analysis was on the resolution of 65 Ascomycota species. Within the Ascomycota, the Taphrinomycota (represented by three *Schizosaccharomyces* species) were fully resolved and in the Saccharomycotina the two clades described by Fitzpatrick et al. (2006), the CTG clade and the WGD clade, were also recovered. CV reconstruction recovered four classes within Pezizomycotina; the Dothideomycetes and Eurotiomycetes were placed as sister taxa with maximum support, as were the Sordariomycetes and Leotiomycetes.

To complement our phylogenomic analyses based on source gene phylogenies or identification of shared orthologs, we carried out alignment-free analysis of 84 fungal species using the composition vector method as implemented in CVTree.

2.5.2 Phylogenomic Reconstruction of 84 Fungal Species Using the CV Approach

Composition vector analysis was carried out on 84 genomes using CVTree with $K=5$ (Qi, Luo, et al., 2004). We selected $K=5$ as the best compromise of both computational requirements and resolution power. As the CV method does not generate bootstrapped phylogenies, we generated 100 bootstrap replicates of our 84-genome representative dataset using bespoke Python scripting and ran composition vector analysis on each replicate dataset (Zuo, Xu, Yu, & Hao, 2010). 100 replicate neighbor-joining phylogenies were calculated from their corresponding CVTree output distance matrices using Neighbor (Felsenstein, 1989). The majority-rule consensus phylogeny for all 100 composition vector replicate trees was generated using Consense (Felsenstein, 1989) and was visualized in iTOL, and annotated according to the NCBI's taxonomy database. The phylogeny was rooted at *R. allomycis* (Fig. 9).

2.5.3 Composition Vector Phylogenomic Reconstruction of 84 Fungal Species Is Congruent With Alignment-Based Methods

We carried out composition vector method phylogenomic reconstruction of our 84-genome dataset to complement the alignment-based and genomic content methods we detailed earlier (Fig. 9). Our composition vector analysis displays adequate levels of taxonomic congruence with our supermatrix and supertree analyses detailed in previous sections, supporting all the monophyly of each major fungal phyla and many of the subphyla within (Fig. 9). There are however some variations in topology and support between the basal lineages and within the Pezizomycotina subphylum in our CV phylogeny compared to our supermatrix and supertree phylogenies.

2.5.3.1 Basal Fungi

After rooting at *R. allomycis*, the Neocallimastigomycota emerge as the earliest-diverging fungal lineage (Fig. 9). The monophyly of Neocallimastigomycetes is also fully supported. Monophyletic Blastocladiomycota and Chytridiomycota clades branch as sister phyla with 62% BP. The monophyly of

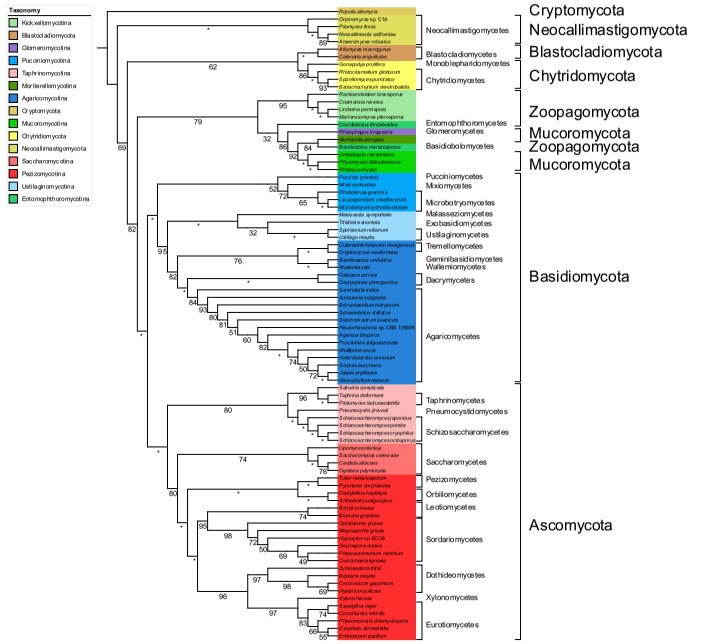


Fig. 9 Composition vector (CV) method phylogeny of 84 fungal species generated from 100 bootstrapped replicates of an 84-genome dataset. Bootstrap supports shown on branches. Maximum bootstrap support designated with an *asterisk* (*).

Blastocladiomycota receives maximum support, and notably unlike our MRP and supermatrix phylogenies *G. prolifera* branches within the Chytridiomycota with 86% BP (Figs. 3–5 and 9). In contrast to both supermatrix phylogenies and the MRP and ST-RF phylogenies, and like the AV and PAM phylogenies the two zygomycetes fungal phyla (Mucoromycota, Zoopagomycota) are placed within one monophyletic clade with 79% BP (Figs. 3–9). Kickxellomycotina are monophyletic with 95% BP and branch at the base of this Zoopagomycota–Mucoromycota clade. Resolution of the relationship between the rest of the former zygomycetes subphyla is harder to ascertain and has weaker support; the two Entomophthoromycotina species branch distant from each other with *B. meristosporus* branching within Mucoromycota adjacent to Mortierellomycotina and *C. thromboides* branching beside the Glomeromycotina species *R. irregularis*, similar to what is

seen under PAM phylogenomic analysis (Figs. 8–9). Like the MRP phylogeny (Fig. 5), *R. irregularis* is within a paraphyletic Mucoromycota clade instead of at the base of the Dikarya as seen in the supermatrix phylogenies (Figs. 3, 4, and 9).

2.5.3.2 Basidiomycota

Pucciniomycotina is placed as the earliest-diverging subphylum within Basidiomycota with 52% BP, and the Ustilagomycotina and Agaricomycotina subphyla are sister clades with 95% BP (Fig. 9). The most-represented class within the Pucciniomycotina, the Microbotryomycetes, is monophyletic with 65% BP (Fig. 9), while unlike the rest of our phylogenies discussed earlier *P. graminis* is placed as the most basal species within Pucciniomycotina. Within the Ustilaginomycotina, *M. sympodialis* are placed as the basal lineage sister to the Exobasidiomycetes representative *T. anomala* similar to its position under ML supermatrix reconstruction and MRP reconstruction (Figs. 3, 5, and 9). The Agaricomycetes are monophyletic with 84% BP, with varying support for relationships within the class but a topology identical to both supermatrix phylogenies and MRP phylogeny with the exception of the placement of Tremellomycetes within a monophyletic ancestral branch adjacent to *B. undulatus* and *W. sebi* (Figs. 3–5 and 9).

2.5.3.3 Ascomycota

Within the Ascomycota, all three subphyla are resolved as monophyletic clades (Fig. 9). Taphrinomycotina is placed as the most basal subphylum within Ascomycota with maximum support, while the Pezizomycotina and Saccharomycotina are sister subphyla with 80% BP (Fig. 9). The Taphrinomycotina are monophyletic with 80% BP, and CV phylogeny displays maximum support for a sister relationship between *P. jirovecii* and the Schizosaccharomycetes and near-maximum (96% BP) support for a similar relationship between *S. complicata* and the two Taphrinomycetes representatives in our dataset (Fig. 9). The Saccharomycotina are monophyletic with 74% support (Fig. 9). All six larger classes from the Pezizomycotina represented in our dataset are resolved as monophyletic. The Orbiliomycetes and Pezizomycetes are placed as both sister subphyla and the earliest-diverging Pezizomycotina clades, both with maximum BP. The Leotiomycetes and Sordariomycetes are also sister clades with 95% BP. As our MRP phylogeny, the Eurotiomycetes are placed as sister to the Xylonomycetes species *X. heveae* with 97% BP (Figs. 5 and 9).

3. A GENOME-SCALE PHYLOGENY OF 84 FUNGAL SPECIES FROM SEVEN PHYLOGENOMIC METHODS

There is a large degree of congruence in the resolution of the fungal kingdom in most of the phylogenomic analyses we described in Section 2, which speaks to the quality of the genomic data we obtained from MycoCosm and the relative accuracy of the majority of the phylogenomic methods we utilized. In constructing a dataset for our analyses, we selected one representative from as many fungal orders as had been sequenced to date; this was to generate a phylogeny that was representative on the order level (though we do not focus on order phylogeny in this review) and to avoid overrepresentation of highly sampled taxa such as Eurotiomycetes or Saccharomycotina. Many of the best-known phylogenetic relationships within the fungal kingdom were recovered in our analyses, such as the monophyly of Dikarya as a whole (Hibbett et al., 2007). However, our analyses also supports more recent studies that have attempted to resolve outstanding branches of the fungal tree of life (Spatafora et al., 2016). In this section, we briefly describe the main trends seen across our seven phylogenomic reconstructions of the fungal kingdom and their congruence with previous studies and comment on the reconstructions of both the well-studied and highly represented Pezizomycotina subphylum and some of the newly circumscribed basal phyla. Finally, we discuss the suitability of the phylogenomic methods we have described and applied in this review for future fungal systematics studies.

3.1 Higher-Level Genome Phylogeny of the Fungal Kingdom

Despite variations in the resolution of some branches, there is a trend across the majority of phylogenies conducted of support or partial support for the eight phyla described in our dataset. Fig. 10 shows the congruence on the phylum level within the fungal kingdom in five of our seven phylogenetic reconstructions. We will refer to Fig. 10 and the subfigures (Figs. 10A–D) in Fig. 10 when comparing the different reconstructions on the phylum level and to the corresponding full phylogenies themselves for comparisons at lower levels here and elsewhere (average consensus and gene content phylogenies are omitted from Fig. 10 on the basis of erroneous placement of taxa). Beginning with the Cryptomycota species *R. allomycis*, the next-earliest-diverging clade within the fungal kingdom is the Blastocladiomycota under both supermatrix analyses followed by Neocallimastigomycota and Chytridiomycota

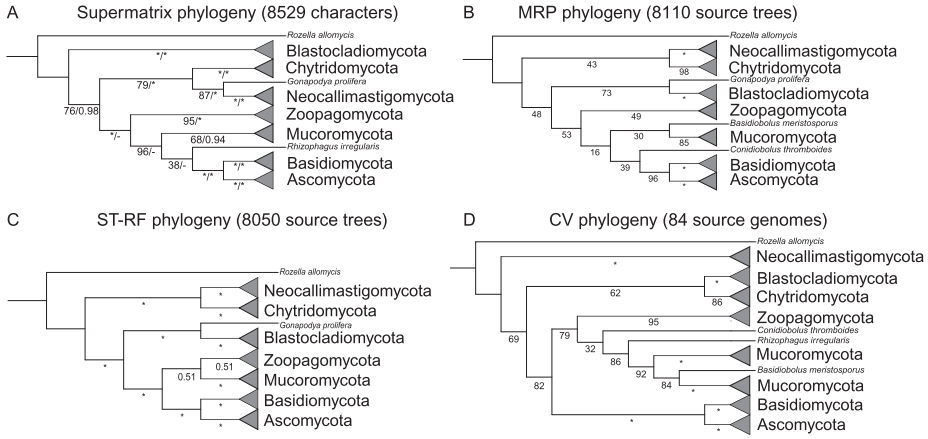
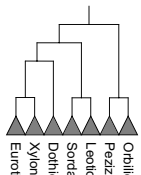


Fig. 10 Congruence of eight fungal phyla under five phylogenomic reconstructions. All clades bar Cryptomycota (represented *Rozella allomycis*) collapsed by phylum, paraphyletic species displayed as individual leaves. *Gonapodya prolifera* = Chytridiomycota, *Rhizophagus irregularis* = Mucoromycota, all other species except *R. allomycis* = Zoopagomycota. Refer to Figs. 3, 4, 5, 7, and 9, respectively, for original phylogenies. (A) ML and Bayesian supermatrix phylogenies. Branch supports given as ML bootstrap supports and, where topology is identical, Bayesian posterior probabilities. Maximum bootstrap or posterior probability support designated with an *asterisk* (*). (B) MRP supertree phylogeny. Branch supports given as bootstrap supports. Maximum bootstrap support designated with an *asterisk* (*). (C). MCMC Bayesian supertree phylogeny using ST-RF ML method. Branch supports given as posterior probabilities of bipartition(s). Maximum posterior probability support designated with an *asterisk* (*). (D) CV phylogeny. Branch supports given as bootstrap supports. Maximum bootstrap support designated with an *asterisk* (*).



	Supermatrix		Supertree			PAM	CV
	ML	PB	MRP	AV	ST-RF		
Orbiliomycetes	✓	✓	×	×	✓	✓	✓
Pezizomycetes	✓	✓	×	×	✓	✓	✓
Leotiomycetes	✓	✓	×	×	✓	✓	✓
Sordariomycetes	✓	✓	×	×	✓	✓	✓
Dothideomycetes	✓	Xy	×	×	✓	Eu	✓
Xylonomycetes	✓	Do	×	×	✓	✓	✓
Eurotiomycetes	✓	✓	×	✓	✓	Do	✓

Fig. 11 Congruence of Pezizomycotina under seven phylogenomic methods. Placement of classes identical to topology on the left (see text) indicated with a *tick*, varying placement of classes indicated by the first two letters of a class. Average consensus (AV) phylogeny produced paraphyletic Pezizomycotina and so entire column labeled with *crosses*. Refer to text for discussion of topology of Pezizomycotina under AV phylogeny. Refer to Figs. 3–9 for original phylogenies.

(Fig. 10A). Other analyses place Neocallimastigomycota and Chytridiomycota (except *G. prolifera*) as closest to *R. allomyis* (Fig. 10B–D).

We describe the resolution of the former zygomycetes in greater detail later, but in the five phylogenies in Fig. 10 all support at least a sister relationship between the two zygomycetes phyla Zoopagomycota and Mucoromycota. The placement of the Glomeromycotina species *R. irregularis* varies, but Mucoromycota is generally placed as sister to the Dikarya (Fig. 10). The Basidiomycota are fully supported as monophyletic in each of the five phylogenies represented in Fig. 10, and all bar ML supermatrix reconstruction is in exact agreement with the two most extensive fungal genome phylogenies containing all three Basidiomycota subphyla (Medina et al., 2011; Wang et al., 2009). The Ascomycota are also fully supported as monophyletic in each of the five phylogenies represented in Fig. 10, with the only major variation being the placement of *S. complicata* within (or paraphyletic to) Taphrinomycotina (Fig. 10). The Saccharomycotina are monophyletic in all five phylogenies (Fig. 10). We discuss the class-level phylogeny within Pezizomycotina in greater detail in Section 3.3 and Fig. 11, but to briefly summarize here we see strong-to-maximum support for all six of the larger classes that were present in our dataset, and support for the two unofficial "Sordariomyceta" and "Dothideomyceta" groupings within Pezizomycotina (Schoch et al., 2009).

3.2 Multiple Phylogenomic Methods Show Moderate Support for the Modern Designations of Mucoromycota and Zoopagomycota

There is moderate support for the recent designation of the zygomycetes phyla Zoopagomycota and Mucoromycota by Spatafora et al. (2016) across

most of our phylogenomic methods (Fig. 10). Previously the species within these two phyla were classified within Zygomycota, a phylum-level classification that had dated back to the 1950s until it was formally disputed by Hibbett et al. (2007). Six *incertae sedis* zygomycetes subphyla were later circumscribed (Hoffmann, Voigt, & Kirk, 2011), and subsequent phylogenetic analyses informally classified the zygomycetes subphyla into two groups, which were later established as Mucoromycota and Zoopagomycota (Chang et al., 2015; Spatafora et al., 2016).

Our phylogenomic analyses included 11 species from the 2 zygomycetes phyla, with the best resolution found in the ST-RF phylogeny where Zoopagomycota and Mucoromycota are placed as sister phyla with 0.51 PP and branch sister to Dikarya (Fig. 10C). Notably, our ST-RF phylogeny is the only phylogeny that resolves Entomophthoromycotina as a monophyletic clade (Fig. 7), albeit with extremely weak posterior probability support (0.38 PP). Within Zoopagomycota in our ST-RF phylogeny, Entomophthoromycotina branch as the basal clade with 0.51 PP, sister to Kickxellomycotina (Fig. 7). Our ST-RF phylogeny also places *R. irregularis* (Glomeromycotina) adjacent to *M. elongata* (Mortierellomycotina) within the Mucoromycota (Fig. 7). Within Mucoromycota, Mortiellomycotina and Mucoromycotina are supported as sister subphyla throughout the majority of our phylogenies (e.g., Bayesian supermatrix analysis, Fig. 4), with high to maximum support. Both of these phylum-level topologies are in agreement with Spatafora et al. (2016), though their phylogeny does not support a distinctive monophyletic branch containing both Zoopagomycota and Mucoromycota (Fig. 10C). The majority of our remaining phylogenomic analysis all shows some degree of support for both Zoopagomycota and Mucoromycota in relative agreement with Spatafora et al. (2016); however, in each of these phylogenies there is some conflict in either subphylum-level topology or lower BP/PP support due to issues of taxon sampling or low gene tree coverage in our dataset (of our 8110 source phylogenies for MRP analysis over 3500 contain 7 taxa or less; Fig. 10). With greater sampling of species from these lineages, we hope to see more consistent support of both the Zoopagomycota and Mucoromycota in future genome phylogenies using these methods, in line with what appears to be moderate-to-strong support for the new classification in our analyses based on total evidence (Kluge, 1989).

3.3 Pezizomycotina as a Benchmark for Phylogenomic Methodologies

The Pezizomycotina are by far the most sampled subphylum within the fungal kingdom in terms of genome sequencing (375 Pezizomycotina species

Author's personal copy

have genomic data available from MycoCosm as of May 2017). Reflecting this, 22 Pezizomycotina species representing 7 classes are present in our 84-genome dataset (>25% of our final dataset). As a well-represented clade within our dataset at both the subphylum and individual class level, we are able to see how multiple phylogenomic analyses conducted in a total evidence approach (Kluge, 1989) are able to resolve a single clade of closely related classes containing some important ecological and pathogenic fungi. In every phylogenomic reconstruction, we attempted bar average consensus (AV) phylogeny, Pezizomycotina were monophyletic with maximum bootstrap or posterior probability branch support, and every class within Pezizomycotina is monophyletic with high or maximum BP or PP support (Figs. 3–5 and 7–9). There is a consistent trend within each of these phylogenies in the resolution of relationships between Pezizomycetes classes:

1. The Orbiliomycetes and Pezizomycetes always branch as the basal classes within Pezizomycotina and are always sister taxa (Figs. 3–5 and 7–9).
2. The relationship between Sordariomycetes and Leotiomycetes (within "Sordariomyceta" sensu Schoch et al., 2009) is always present and is fully supported in each phylogeny (Figs. 3–5 and 7–9).
3. The relationship between Dothideomycetes, Xylonomycetes, and Eurotiomycetes (within "Dothideomyceta" sensu Schoch et al., 2009) is always present and is fully supported in each phylogeny (Figs. 3–5 and 7–9).

Fig. 11 displays on the left the topology of the Pezizomycotina classes supported under ML supermatrix reconstruction, MRP supertree reconstruction, and ST-RF supertree reconstruction (Figs. 3, 5, and 7) and indicates the congruence (or otherwise) of Pezizomycotina under every phylogenomic analysis we attempted (Figs. 3–9). All methods bar AV are highly congruent in their resolution of the Pezizomycotina subphylum, with placement of the Xylonomycetes class the most notable variation. Even within the highly aberrant AV phylogeny, sister relationships such as those between Orbiliomycetes and Pezizomycetes or the association of classes within Sordariomyceta or Dothideomyceta can still be observed, though with lower resolution and support (Fig. 6). There is a high degree of congruence between our genome phylogenies of Pezizomycotina (Fig. 11) and the most extensive molecular phylogenies of Pezizomycotina that we could find in the literature derived from either small concatenated sets or whole genomes (Medina et al., 2011; Spatafora et al., 2006; Wang et al., 2009). The relative consistency of our analyses both with each other and with previous literature suggests that the resolution of Pezizomycotina could be considered a good benchmark for the accuracy of novel or existing

phylogenomic methods (e.g., ST-RF analysis) when incorporated into a total evidence analysis, as the subphylum is large and diverse (the 10th edition of Ainsworth & Bisby's Dictionary of the Fungi estimates close to 70,000 Pezizomycetes species) but also densely sampled in genomic terms and containing a number of genomes of reference quality (Kirk, Cannon, Minter, & Stalpers, 2008).

3.4 The Use of Phylogenomics Methods in Fungal Systematics

Phylogenomic analyses with larger datasets across a wider spectrum of taxa are becoming more and more computationally tractable as methods of identifying potential phylogenetic markers on a genome-wide scale (e.g., identification and reconstruction of orthologous gene phylogenies in supertree analysis) and genome-scale reconstruction improve. In as much as the majority of our multiple analyses strongly support the major phyla of the fungal kingdom, we can also treat our analyses as measures of the accuracy of each of these phylogenomic methods in the reconstruction of large datasets. Supermatrix, MRP and ST-RF supertree, and CV method reconstructions all appear to arrive at relatively congruent results and may be useful for approximating a total evidence style approach for phylogenomic analyses of fungi. Simplified parsimony methods like our PAM phylogeny or branch length-based methods like our average consensus phylogeny may be useful for the reconstruction of smaller but well-represented datasets (for example, our PAM phylogeny does reconstruct the Pezizomycotina with support and topology close to supertree and supermatrix phylogenies) but for phylum or kingdom-wide analyses issues such as long-branch attraction begin to emerge (Bergsten, 2005). Long-branch attraction is thought to be an issue with MRP reconstruction as well, and while it is likely a factor in the weaker supports in some of the ancestral branches in our MRP phylogeny (for example, the weak supports in some of the internal branches grouping the basal phyla together), the MRP phylogeny seems to have been relatively immune to the topological effects of long-branch attraction that are very apparent in our branch length-dependent average consensus method phylogeny (Pisani & Wilkinson, 2002).

For our supertree analyses, we identified groups of orthologous proteins using a sequential random BLASTp approach as implemented by Fitzpatrick et al. (2006), where a random sequence from a given database is searched against that entire database, and then the sequence and its homologs (if any) are removed and the database reformatted (Fitzpatrick et al., 2006). Overall, this

ad hoc approach to identifying orthology within our dataset seems to have been sufficient as a first step to generating source gene phylogenies; however, it may have had an impact downstream on resolution of internal branches within our MRP analysis. It is possible that a random BLASTp approach is too conservative, in that the orthologous families it identifies are missing members or that two "separate" orthologous families may in fact be one large orthologous family. Other established methods of identifying orthologous families, such as the OrthoMCL pipeline, have been used in phylogenomic analyses and can be tuned for granularity (i.e., orthologous cluster size) which may produce broader source phylogenies (Li, Stoeckert, & Roos, 2003). However, the large SQL-dependent computational overhead required for the current implementation of OrthoMCL was not considered suitable for an analysis of this scale.

Most of the phylogenomic methods we attempted are relatively tractable even for a dataset as large as ours. Depending on computational resources and available data, some of the methods we have discussed may be more appropriate for future fungal phylogenomic analyses than others. The most common techniques like MRP analysis and both ML and Bayesian supermatrix analysis were both tractable and produced phylogenies with largely congruent topologies and supports on most branches (although we should note that we utilized the parallelized version of PhyloBayes for our Bayesian analysis). The heuristic MCMC Bayesian supertree reconstruction we attempted using the ST-RF model as implemented in p4 was also relatively tractable despite not being parallelized, and Akanni et al. (2015) note that the method is far more efficient than the approximate ML reconstruction implemented in L.U.St. (Akanni et al., 2015). However, ST-RF analysis using either p4 or L.U.St. is currently only able to use fully resolved input phylogenies. While in our case this meant only 60 single-copy phylogenies (<1% of our total dataset) had to be removed before carrying out analysis, this may cause issues for more polytomous datasets. Bayesian and ML supertree reconstruction is certainly a promising development for phylogenomics, and hopefully methods like ST-RF should see more widespread use in future phylogenomic analysis as they mature.

Phylogenomic reconstruction using average consensus as implemented in CLANN was extremely inefficient time-wise and returned a severely erroneous phylogeny, so while it is certainly desirable for branch lengths to be incorporated in supertree reconstruction, a branch length-based method like AV is not appropriate for this kind of large-scale analysis. While PAM method reconstruction was straightforward to carry out, as we state earlier there were issues with erroneous placement of taxa and as such we

do not recommend the method for large-scale datasets. Finally, composition vector method analysis produced a phylogeny relatively congruent to our alignment-based methods at $K=5$. Other CV method analyses have recommended K-values between 5 and 7 for most datasets (Zuo, Li, & Hao, 2014), however with the size of our dataset and the increase in computational resources required for generating distance matrices for eukaryotic genomes at $K>5$ in CVTree we felt that $K=5$ was the best compromise between accuracy and computational tractability. We would recommend however as in Section 2.5 that CV analysis should be used in conjunction with alignment-based methods for eukaryotic datasets, as interpretation of CV analysis requires a priori knowledge of the phylogeny of a given dataset.

4. CONCLUDING REMARKS

Fungi make up one of the major eukaryotic kingdoms, with an estimated 1.5 million member species inhabiting a diverse variety of ecological niches and an evolutionary history dating back over a billion years. It is imperative that evolutionary relationships within the fungal kingdom are well understood by analysis of as much quality phylogenetic data as is available with the most accurate methodologies possible. In this chapter, we discussed the evolutionary diversity of the fungal kingdom and the important role that fungi have had in the area of genomic and phylogenomics. We have reviewed previous phylogenomic analyses of the fungal kingdom over the last decade, and using seven phylogenomic methods, we have reconstructed the phylogeny of 84 fungal species across 8 fungal phyla. We found that established supermatrix and supertree methods produced relatively congruent phylogenies that were in large agreement with the literature. We also conducted the first analysis of the fungal kingdom using a heuristic MCMC Bayesian approach to supertree reconstruction previously used in Metazoa and found that this novel supertree approach resolves the fungal kingdom with a high degree of accuracy. The majority of our analyses overall show moderate-to-strong support of the newly assigned zygomycete phyla Mucoromycota and Zoopagomycota and strongly support the monophyly of Dikarya, while within the highly sampled Pezizomycotina subphylum there is a large amount of congruence between different phylogenomic methods as to the resolution of class relationships within the subphylum. We also conclude that supermatrix and supertree analyses remain the exemplar methods of phylogenomic reconstruction for fungi, based on their accuracy and

Author's personal copy

computational tractability. We believe through both our discussion of the ecological diversity of the fungal kingdom and the history of its study on the genomic level we have demonstrated the need for a robust fungal tree of life with a broad representation, and that through our multiple phylogenomic analysis we have generated an important backbone for future comparative genomic analysis of fungi, particularly with the constantly increasing amount of quality genomic data arising from the 1000 Fungal Genomes Project and its certain use in future studies.

ACKNOWLEDGMENTS

We wish to acknowledge the JGI and all individual contributors to the 1000 Fungal Genomes Project for both the sheer scope of their undertaking and the quantity and quality of genomic data that they have made publically available and that we were able to use in this chapter. We also wish to acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support. C.G.P.M. is funded by an Irish Research Council Government of Ireland Postgraduate Scholarship (Grant No. GOIPG/2015/2242).

REFERENCES

Akanni, W. A., Creevey, C. J., Wilkinson, M., & Pisani, D. (2014). L.U.St: A tool for approximated maximum likelihood supertree reconstruction. *BMC Bioinformatics*, *15*(1), 183. https://doi.org/10.1186/1471-2105-15-183.

Akanni, W. A., Wilkinson, M., Creevey, C. J., Foster, P. G., & Pisani, D. (2015). Implementing and testing Bayesian and maximum-likelihood supertree methods in phylogenetics. *Royal Society Open Science*, *2*(8), 140436. https://doi.org/10.1098/rsos.140436.

Annaluru, N., Muller, H., Mitchell, L. A., Ramalingam, S., Stracquadanio, G., Richardson, S. M., et al. (2014). Total synthesis of a functional designer eukaryotic chromosome. *Science*, *344*(6179), 55–58. https://doi.org/10.1126/science.1249252.

Baldauf, S. L., & Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(24), 11558–11562. https://doi.org/10.1073/pnas.90.24.11558.

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, *41*(1), 3–10. https://doi.org/10.2307/1222480.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), D36–42. https://doi.org/10.1093/nar/gks1195.

Berbee, M. L., & Taylor, J. W. (1992). Detecting morphological convergence in true fungi, using 18S ribosomal RNA gene sequence data. *Biosystems*, *28*(1–3), 117–125.

Berbee, M. L., & Taylor, J. W. (2010). Dating the molecular clock in fungi—How close are we? *Fungal Biology Reviews*, *24*(1–2), 1–16. https://doi.org/10.1016/j.fbr.2010.03.001.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x.

Bininda-Emonds, O. R. P. (2004). The evolution of supertrees. *Trends in Ecology and Evolution*, *19*(6), 315–322. https://doi.org/10.1016/j.tree.2004.03.015.

Bryant, D., & Steel, M. (2009). Computing the distribution of a tree metric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *6*(3), 420–426. https://doi.org/10.1109/TCBB.2009.32.

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. C. M. A. S., Sakthikumar, S., Munro, C. A., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature*, *459*(7247), 657–662. https://doi.org/10.1038/nature08064.

Byrne, K. P., & Wolfe, K. H. (2005). The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, *15*(10), 1456–1461. https://doi.org/10.1101/gr.3672305.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421.

Campbell, A., Mrázek, J., & Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(16), 9184–9189. https://doi.org/10.1073/pnas.96.16.9184.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Cavalier-Smith, T. (1998). A revised six-kingdom system of life. *Biological Reviews of the Cambridge Philosophical Society*, *73*(3), 203–266.

Chang, Y., Wang, S., Sekimoto, S., Aerts, A. L., Choi, C., Clum, A., et al. (2015). Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants. *Genome Biology and Evolution*, *7*(6), 1590–1601. https://doi.org/10.1093/gbe/evv090.

Creevey, C. J., Fitzpatrick, D. A., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., et al. (2004). Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings Biological Sciences/The Royal Society*, *271*(1557), 2551–2558. https://doi.org/10.1098/rspb.2004.2864.

Creevey, C. J., & McInerney, J. O. (2005). Clann: Investigating phylogenetic information through supertree analyses. *Bioinformatics*, *21*(3), 390–392. https://doi.org/10.1093/bioinformatics/bti020.

Creevey, C. J., & McInerney, J. O. (2009). Trees from trees: Construction of phylogenetic supertrees using CLANN. *Methods in Molecular Biology*, *537*, 139–161. https://doi.org/10.1007/978-1-59745-251-9_7.

Cuomo, C. A., & Birren, B. W. (2010). The fungal genome initiative and lessons learned from genome sequencing. *Methods in Enzymology*, *470*(C), 833–855. https://doi.org/10.1016/S0076-6879(10)70034-3.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics*, *27*(8), 1164–1165. https://doi.org/10.1093/bioinformatics/btr088.

De Barros Lopes, M., Bellon, J. R., Shirley, N. J., & Ganter, P. F. (2002). Evidence for multiple interspecific hybridization in Saccharomyces sensu stricto species. *FEMS Yeast Research*, *1*(4), 323–331. https://doi.org/10.1016/S1567-1356(01)00051-4.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, *6*(5), 361–375. https://doi.org/10.1038/nrg1603.

de Queiroz, A., & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology and Evolution*, *22*(1), 34–41. https://doi.org/10.1016/j.tree.2006.10.002.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., et al. (2014). The reference genome sequence of Saccharomyces cerevisiae: Then and now. *G3 (Bethesda)*, *4*(3), 389–398. https://doi.org/10.1534/g3.113.008995.

Faith, D. P., & Cranston, P. S. (1991). Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics*, *7*(1), 1–28. https://doi.org/10.1111/j.1096-0031.1991.tb00020.x.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Research*, *40*(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, *27*(4), 401. https://doi.org/10.2307/2412923.

Felsenstein, J. (1989). PHYLIP—Phylogeny inference package—v3.2. *Cladistics*, *5*(2), 164–166. https://doi.org/10.1111/j.1096-0031.1989.tb00562.x.

Fisk, D. G., Ball, C. A., Dolinski, K., Engel, S. R., Hong, E. L., Issel-Tarver, L., et al. (2006). Saccharomyces cerevisiae S288C genome annotation: A working hypothesis. *Yeast*, *23*(12), 857–865. https://doi.org/10.1002/yea.1400.

Fitzpatrick, D. A., Logue, M. E., & Butler, G. (2008). Evidence of recent interkingdom horizontal gene transfer between bacteria and Candida parapsilosis. *BMC Evolutionary Biology*, *8*(1), 181. https://doi.org/10.1186/1471-2148-8-181.

Fitzpatrick, D. A., Logue, M. E., Stajich, J. E., & Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, *6*(1), 99. https://doi.org/10.1186/1471-2148-6-99.

Fitzpatrick, D. A., O'Gaora, P., Byrne, K. P., & Butler, G. (2010). Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. *BMC Genomics*, *11*(1), 290. https://doi.org/10.1186/1471-2164-11-290.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic Biology*, *53*(3), 485–495. https://doi.org/10.1080/10635150490445779.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., et al. (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature*, *422*(6934), 859–868. https://doi.org/10.1038/nature01554.

Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L.-J., Wortman, J. R., Batzoglou, S., et al. (2005). Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature*, *438*(7071), 1105–1115. https://doi.org/10.1038/nature04341.

Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A., & Birren, B. (2005). Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Research*, *15*(12), 1620–1631. https://doi.org/10.1101/gr.3767105.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science*, *274*(5287), 546–567. https://doi.org/10.1126/science.274.5287.546.

Goffeau, A., & Vassarotti, A. (1991). The European project for sequencing the yeast genome. *Research in Microbiology*, *142*(7–8), 901–903. https://doi.org/10.1016/0923-2508(91)90071-H.

Grigoriev, I. V., Cullen, D., Goodwin, S. B., Hibbett, D., Jeffries, T. W., Kubicek, C. P., et al. (2011). Fueling the future with fungal genomics. *Mycology*, *2*(3), 192–209. https://doi.org/10.1080/21501203.2011.584577.

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., et al. (2014). MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Research*, *42*(D1), D699–704. https://doi.org/10.1093/nar/gkt1183.

Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., et al. (2011). The genome portal of the department of energy joint genome institute. *Nucleic Acids Research*, *40*(D1), 1–7. https://doi.org/10.1093/nar/gkr947.

Guarro, J., Gené, J., & Stchigel, A. M. (1999). Developments in fungal taxonomy. *Clinical Microbiology Reviews*, *12*(3), 454–500. https://doi.org/0893-8512/99/$04.00?0.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010.

Hall, C., & Dietrich, F. S. (2007). The reacquisition of biotin prototrophy in Saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*, *177*(4), 2293–2307. https://doi.org/10.1534/genetics.107.074963.

Hawksworth, D. L. (2001). The magnitude of fungal diversity: The 1.5 million species estimate revisited. *Mycological Research*, *105*(12), 1422–1432. https://doi.org/10.1017/S0953756201004725.

Heath, I. B. (1980). Variant mitoses in lower eukaryotes: Indicators of the evolution of mitosis? *International Review of Cytology*, *64*(C), 1–80. https://doi.org/10.1016/S0074-7696(08)60235-1.

Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., et al. (2007). A higher-level phylogenetic classification of the fungi. *Mycological Research*, *111*(5), 509–547. https://doi.org/10.1016/j.mycres.2007.03.004.

Hoffmann, K., Voigt, K., & Kirk, P. M. (2011). Mortierellomycotina subphyl. nov., based on multi-gene genealogies. *Mycotaxon*, *115*(1), 353–363. https://doi.org/10.5248/115.353.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., et al. (1965). Structure of a ribonucleic acid. *Science (New York, N.Y.)*, *147*(3664), 1462–1465. https://doi.org/10.1126/science.147.3664.1462.

Holton, T. A., & Pisani, D. (2010). Deep genomic-scale analyses of the metazoa reject coelomata: Evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biology and Evolution*, *2*(1), 310–324. https://doi.org/10.1093/gbe/evq016.

Huelsenbeck, J. P., & Hillis, D. M. (1993). Success of phylogenetic methods in the four taxon case. *Systematic Biology*, *42*(3), 247–264. https://doi.org/10.1093/sysbio/42.3.247.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, *294*(5550), 2310–2314. https://doi.org/10.1126/science.1065889.

Jackson, A. P., Gamble, J. A., Yeomans, T., Moran, G. P., Saunders, D., Harris, D., et al. (2009). Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans. *Genome Research*, *19*(12), 2231–2244. https://doi.org/10.1101/gr.097501.109.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., et al. (2006). Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature*, *443*(7113), 818–822. https://doi.org/10.1038/nature05110.

Jones, M. D. M., Forn, I., Gadelha, C., Egan, M. J., Bass, D., Massana, R., et al. (2011). Discovery of novel intermediate forms redefines the fungal tree of life. *Nature*, *474*(7350), 200–203. https://doi.org/10.1038/nature09984.

Jones, M. D. M., Richards, T. A., Hawksworth, D. L., & Bass, D. (2011). Validation and justification of the phylum name Cryptomycota phyl. nov. *IMA Fungus*, *2*(2), 173–175. https://doi.org/10.5598/imafungus.2011.02.02.08.

Keller, N. P., Turner, G., & Bennett, J. W. (2005). Fungal secondary metabolism—From biochemistry to genomics. *Nature Reviews Microbiology*, *3*(12), 937–947. https://doi.org/10.1038/nrmicro1286.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, *428*(VN-(6983)), 617–624. https://doi.org/10.1038/nature02424.

Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H., et al. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*, *47*(9), 736–741. https://doi.org/10.1016/j.fgb.2010.06.003.

Author's personal copy

Kirk, P. M., Cannon, P. F., Minter, D. W., & Stalpers, J. A. (2008). *Ainsworth & Bisby's dictionary of the fungi* (10th ed.). Wallingford, UK: CABI.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among epicrates (Boidae, serpentes). *Systematic Biology*, *38*(1), 7–25. https://doi.org/10.1093/sysbio/38.1.7.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, *5*(2), R7. https://doi.org/10.1186/gb-2004-5-2-r7.

Kück, P., & Meusemann, K. (2010). FASconCAT: Convenient handling of data matrices. *Molecular Phylogenetics and Evolution*, *56*(3), 1115–1118. https://doi.org/10.1016/j.ympev.2010.04.024.

Kuramae, E. E., Robert, V., Snel, B., Weiß, M., & Boekhout, T. (2006). Phylogenomics reveal a robust fungal tree of life. *FEMS Yeast Research*, *6*(8), 1213–1220. https://doi.org/10.1111/j.1567-1364.2006.00119.x.

Lapointe, F.-J., & Cucumel, G. (1997). The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, *46*(2), 306–312. https://doi.org/10.1093/sysbio/46.2.306.

Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, *7*(Suppl. 1), S4. https://doi.org/10.1186/1471-2148-7-S1-S4.

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112.

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, *62*(4), 611–615. https://doi.org/10.1093/sysbio/syt022.

Léjohn, H. B. (1974). *Biochemical parameters of fungal phylogenetics*. *In T. Dobzhansky, M. K. - Hecht, & W. C. Steere (Eds.), Evolutionary biology* (pp. 79–125). Boston, MA: Springer. *https://doi.org/10.1007/978-1-4615-6944-2_3*.

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, *44*(W1), W242–W245. https://doi.org/10.1093/nar/gkw290.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503.

Lin, J., & Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. *Genome Research*, *10*(6), 808–818. https://doi.org/10.1101/gr.10.6.808.

Marcet-Houben, M., & Gabaldón, T. (2009). The tree versus the forest: The fungal tree of life and the topological diversity within the yeast phylome. *PLoS One*, *4*(2). e4357 https://doi.org/10.1371/journal.pone.0004357.

Marcet-Houben, M., & Gabaldón, T. (2010). Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics: TIG*, *26*(1), 5–8. https://doi.org/10.1016/j.tig.2009.11.007.

Marcet-Houben, M., Marceddu, G., & Gabaldón, T. (2009). Phylogenomics of the oxidative phosphorylation in fungi reveals extensive gene duplication followed by functional divergence. *BMC Evolutionary Biology*, *9*(1), 295. https://doi.org/10.1186/1471-2148-9-295.

Medina, E. M., Jones, G. W., & Fitzpatrick, D. A. (2011). Reconstructing the fungal tree of life using phylogenomics and a preliminary investigation of the distribution of yeast

prion-like proteins in the fungal kingdom. *Journal of Molecular Evolution*, *73*(3–4), 116–133. https://doi.org/10.1007/s00239-011-9461-4.

Nikoh, N., Hayase, N., Iwabe, N., Kuma, K., & Miyata, T. (1994). Phylogenetic relationship of the kingdoms Animalia, Plantae, and Fungi, inferred from 23 different protein species. *Molecular Biology and Evolution*, *11*(5), 762–768. https://doi.org/0737-4038/94f I lOS-0005$02.

Oliver, S. G., Van Der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., et al. (1992). The complete DNA sequence of yeast chromosome III. *Nature*, *357*(6373), 38–46. https://doi.org/10.1038/357038a0.

Page, R. D. M., & Holmes, E. C. (1998). *Molecular evolution: A phylogenetic approach*. Oxford, UK: Blackwell Science.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071.

Pisani, D., & Wilkinson, M. (2002). Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology*, *51*(1), 151–155. https://doi.org/10.1080/106351502753475925.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS One*, *5*(3). e9490 https://doi.org/10.1371/journal.pone.0009490.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M., & Blaser, M. J. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*, *13*(2), 145–156. https://doi.org/10.1101/gr.335003.

Qi, J., Luo, H., & Hao, B. (2004). CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, *32*(Suppl. 2), W45–7. https://doi.org/10.1093/nar/gkh362.

Qi, J., Wang, B., & Hao, B. I. (2004). Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Journal of Molecular Evolution*, *58*(1), 1–11. https://doi.org/10.1007/s00239-003-2493-7.

Ragan, M. A. (1992). Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, *1*(1), 53–58. https://doi.org/10.1016/1055-7903(92)90035-F.

Redecker, D. (2000). Glomalean fungi from the ordovician. *Science*, *289*(5486), 1920–1921. https://doi.org/10.1126/science.289.5486.1920.

Richards, T. A., Soanes, D. M., Jones, M. D. M., Vasieva, O., Leonard, G., Paszkiewicz, K., et al. (2011). Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15258–15263. https://doi.org/10.1073/pnas.1105100108.

Robbertse, B., Reeves, J. B., Schoch, C. L., & Spatafora, J. W. (2006). A phylogenomic analysis of the Ascomycota. *Fungal Genetics and Biology*, *43*(10), 715–725. https://doi.org/10.1016/j.fgb.2006.05.001.

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*(6960), 798–804. https://doi.org/10.1038/nature02053.

Schoch, C. L., Sung, G. H., López-Giráldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., et al. (2009). The ascomycota tree of life: A phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic Biology*, *58*(2), 224–239. https://doi.org/10.1093/sysbio/syp020.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Author's personal copy

Slowinski, J. B., & Page, R. D. (1999). How should species phylogenies be inferred from sequence data? *Systematic Biology*, *48*(4), 814–825. https://doi.org/10.1080/106351599260030.

Snel, B., Bork, P., & Huynen, M. a. (1999). Genome phylogeny based on gene content. *Nature Genetics*, *21*(1), 108–110. https://doi.org/10.1038/5052.

Snel, B., Huynen, M. A., & Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annual Review of Microbiology*, *59*(1), 191–209. https://doi.org/10.1146/annurev.micro.59.030804.121233.

Spatafora, J. W., Chang, Y., Benny, G. L., Lazarus, K., Smith, M. E., Berbee, M. L., et al. (2016). A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia*, *108*(5), 1028–1046. https://doi.org/10.3852/16-042.

Spatafora, J., Sung, G., Johnson, D., Hesse, C., O'Rourke, B., Serdani, M., et al. (2006). A five-gene phylogeny of Pezizomycotina. *Mycologia*, *98*(6), 1018–1028. https://doi.org/10.3852/mycologia.98.6.1018.

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*(Suppl. 2) W309-12https://doi.org/10.1093/nar/gkh379.

Steel, M., & Rodrigo, A. (2008). Maximum likelihood supertrees. *Systematic Biology*, *57*(2), 243–250. https://doi.org/10.1080/10635150802033014.

Su, Z., & Townsend, J. P. (2015). Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: Analytical predictions of long-branch effects. *BMC Evolutionary Biology*, *15*(86), 86. https://doi.org/10.1186/s12862-015-0364-7.

Swofford, L. D. (2002). *PAUP*: Phylogenetic analysis using parsimony (* and other methods). Version 4.0 beta*. Sunderland, MA: Sinauer.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., & Boussau, B. (2015). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1678), 20140335. https://doi.org/10.1098/rstb.2014.0335.

Taylor, F. J. R. (1978). Problems in the development of an explicit hypothetical phylogeny of the lower eukaryotes. *Biosystems*, *10*(1–2), 67–89. https://doi.org/10.1016/0303-2647(78)90031-X.

Tekaia, F., Lazcano, A., & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Research*, *9*(6), 550–557. https://doi.org/10.1101/gr.9.6.550.

Wang, H., Xu, Z., Gao, L., & Hao, B. (2009). A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evolutionary Biology*, *9*(1), 195. https://doi.org/10.1186/1471-2148-9-195.

Wilkinson, M., Thorley, J. L., Pisani, D. E., Lapointe, F.-J., & McInerney, J. O. (2004). *Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (Ed.), Vol. 3. Phylogenetic supertrees: Combining information to reveal the Tree of Life* (pp. 227–246). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-1-4020-2330-9_11.

Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*(6634), 708–713. https://doi.org/10.1038/42711.

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., et al. (2002). The genome sequence of Schizosaccharomyces pombe. *Nature*, *415*(6874), 871–880. https://doi.org/10.1038/nature724.

Zuo, G., Li, Q., & Hao, B. (2014). On K-peptide length in composition vector phylogeny of prokaryotes. *Computational Biology and Chemistry*, *53*(Part A), 166–173. https://doi.org/10.1016/j.compbiolchem.2014.08.021.

Zuo, G., Xu, Z., Yu, H., & Hao, B. (2010). Jackknife and bootstrap tests of the composition vector trees. *Genomics, Proteomics and Bioinformatics*, *8*(4), 262–267. https://doi.org/10.1016/S1672-0229(10)60028-9.

MICROBIAL GENOMICS

RESEARCH ARTICLE

McCarthy and Fitzpatrick, *Microbial Genomics* 2019;5

DOI 10.1099/mgen.0.000243

Pan-genome analyses of model fungal species

Charley G. P. McCarthy[1,2] and David A. Fitzpatrick[1,2,*]

Abstract

The concept of the species 'pan-genome', the union of 'core' conserved genes and all 'accessory' non-conserved genes across all strains of a species, was first proposed in prokaryotes to account for intraspecific variability. Species pan-genomes have been extensively studied in prokaryotes, but evidence of species pan-genomes has also been demonstrated in eukaryotes such as plants and fungi. Using a previously published methodology based on sequence homology and conserved microsynteny, in addition to bespoke pipelines, we have investigated the pan-genomes of four model fungal species: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. Between 80 and 90 % of gene models per strain in each of these species are core genes that are highly conserved across all strains of that species, many of which are involved in housekeeping and conserved survival processes. In many of these species, the remaining 'accessory' gene models are clustered within subterminal regions and may be involved in pathogenesis and antimicrobial resistance. Analysis of the ancestry of species core and accessory genomes suggests that fungal pan-genomes evolve by strain-level innovations such as gene duplication as opposed to wide-scale horizontal gene transfer. Our findings lend further supporting evidence to the existence of species pan-genomes in eukaryote taxa.

DATA SUMMARY

All the genomic sequence data has been previously uploaded to the National Center for Biotechnology Information (NCBI) GenBank, and links to relevant articles or NCBI BioProject pages are included in Table S1 (available with the online version of this article). Gene model prediction and post-processing pan-genome analysis pipelines are available from https://github.com/chmccarthy/pangenome-pipelines.

INTRODUCTION

Many fields of eukaryote functional and comparative genomics rely on the use of curated reference genomes intended to be broadly representative of a given species. Regardless of their quality, reference genomes do not and cannot contain all genetic information for a species due to genetic and genomic variation between individuals within a species [1]. To account for such variation, it has become increasingly common to refer to species with multiple genomes sequenced in terms of their 'pan-genome', which is defined as the union of all genes observed across all isolates/strains of a species (Fig. 1). The pan-genome of a species is then usually subdivided into two components. (i) The 'core' genome, containing genes conserved across all observed genomes from a species. These genes are usually, but not always, essential for the viability of an individual organism [2]. (ii) The 'accessory' or 'dispensable' genome, containing genes specific to sets of isolate genomes or individual isolate genomes within a species. These genes could influence phenotypic differences between isolates; for example, in bacteria, antibiotic-resistant and antibiotic-susceptible isolates of the same species may have different accessory genomes [2].

A species' pan-genome can evolve as a consequence of lifestyle: sympatric species may have large pan-genomes (and thus a large degree of intraspecific variation), while environmentally isolated or highly specialized species have smaller pan-genomes [2–5]. The existence of a species pan-genome in prokaryotes was first demonstrated across eight pathogenic strains of *Streptococcus agalactiae* in 2005 [6], and was quickly confirmed by similar analysis of exemplar bacteria and archaea, including *Haemophilus influenzae*, *Escherichia coli* and *Sulfolobus islandicus* [7–11]. Over 40 prokaryote species had their pan-genomes described in the literature by 2013 [2]. Many tools for pan-genome analysis have been published in recent years, which utilize methods such as whole-genome alignment, read mapping, clustering algorithms or de Bruijn graph construction [12–16].

Although the concept of the species pan-genome is well-established in comparative prokaryote genomics, it has only recently been extended to comparative intraspecific studies of eukaryotes. This is despite repeated observation of intraspecific genomic content variation in eukaryotes dating back to the first intraspecific comparative analyses of *Saccharomyces cerevisiae* genomes in the mid-2000s [17–20]. The relative dearth of eukaryotic pan-genome analysis in the literature is due in part to the relative difficulty of sequencing and analysing large eukaryotic genome datasets relative to prokaryotes [21]. Additionally, while horizontal gene transfer (HGT) is thought to be the driving influence in prokaryotic gene family and pan-genome evolution, HGT occurs in far lower rates in eukaryotes and is more difficult to detect [22–26]. Despite these challenges, there have been a number of recent studies of intraspecific variation within diverse eukaryote taxa that show strong evidence for the existence of a eukaryotic pan-genome in some form. For example, comparative analysis of nine diverse cultivars of *Brassica oleracea* found that ~19 % of all genes analysed were part of the *B. oleracea* accessory genome, with ~2 % of these being cultivar-specific [27]. A similar comparison of seven geographically diverse wild soybean (*Glycine soja*) strains found approximately the same 80 : 20 proportion of core to accessory gene content within the wild soybean pan-genome, while larger accessory genome sizes have been reported in wheat, maize, grasses and *Medicago* [28–32]. Individual strains of the coccolitophore *Emiliania huxleyi* have an accessory complement of up to 30 % of their total gene content, which varies with geographical location [33]. In fungi, a number of studies of the *Saccharomyces cerevisiae* pan-genome, including a recent large-scale analysis of genome evolution across 1011 strains, have shown evidence for an accessory genome of varying size, as well as large variation in subterminal regions across multiple *Saccharomyces cerevisiae* strains [13, 34–36], and recent analysis of the *Zymoseptoria tritici* pan-genome found that up to 40 % of genes in the total *Z. tritici* pan-genome were either lineage or strain-specific [37].

The methods of pan-genome evolution within eukaryotes in the absence of rampant HGT appears to vary among species, and can include genome rearrangement events or more discrete adaptive evolution processes. In plants, accessory genomes may evolve as a result of varying levels of ploidy, heterozygosity and whole-genome duplication within species, as well as adaptive changes and the evolution of phenotypic differences, such as in *B. oleracea* [27]. Adaptive evolution has also influenced the evolution of the *Emiliania huxleyi* pan-genome, with strains containing varying amounts of nutrient acquisition and metabolism as a result of niche specialization [33]. High levels of functionally redundant accessory genome content can be observed within the *Z. tritici* species pan-genome, which is thought to arise from the species' own genome defence mechanisms inducing polymorphisms as opposed to gene duplication events [37]. Peter *et al.* [36] observed a large proportion of accessory genes within *Saccharomyces cerevisiae* appear to have arisen via introgression from closely related *Saccharomyces* species, with a smaller number originating from HGT events with other yeasts [36].

In this study, we have adapted a method of prokaryotic pan-genome analysis that identifies putative pan-genomic structure within species by accounting for conserved genomic neighbourhoods (CGNs) between strain genomes and applied it to eukaryote analysis [38] (Fig. S1). We have used this method in tandem with bespoke pre- and post-processing pipelines that analyse the extent of gene duplication within species pan-genomes (available from https://github.com/chmccarthy/pangenome-pipelines) to construct and characterize the pan-genomes of four exemplar fungal species: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. All four species are model organisms in eukaryotic genomics and play important roles in human health and lifestyles; *Saccharomyces cerevisiae* is used

IMPACT STATEMENT

Recent prokaryotic genomic studies of multiple individuals from the same species has uncovered large differences in the gene content between individuals. It has become increasingly common to refer to species with multiple genomes sequenced in terms of their 'pan-genome'. The pan-genome is the union of 'core' conserved genes and all 'accessory' non-conserved genes across all strains of a species. Species pan-genomes have been analysed in many prokaryotic species, but have been recently demonstrated in eukaryotes such as plants and fungi as well. Here, we have investigated the pan-genomes of four model fungal species namely, *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. Each species is a model organism for fungal evolutionary biology, genomics and comparative genomics. Our results show that between 80 and 90 % of gene models per strain are core genes that are highly conserved, many of which are involved in housekeeping and conserved survival processes. The remaining accessory gene models are clustered within subterminal regions, and may be involved in pathogenesis and antimicrobial resistance. Analysis of the ancestry of species core and accessory genomes suggests that fungal pan-genomes evolve by strain-level innovations such as gene duplication as opposed to wide-scale horizontal gene transfer. Our findings lend further supporting evidence to the existence of species pan-genomes in eukaryote taxa.

Received 5 September 2018; Accepted 23 November 2018; Published 4 February 2019

Author affiliations: [1]Genome Evolution Laboratory, Department of Biology, Maynooth University, Maynooth, Co. Kildare, Ireland; [2]Human Health Research Institute, Maynooth University, Maynooth, Co. Kildare, Ireland.

***Correspondence:** David A. Fitzpatrick, david.fitzpatrick@mu.ie

Keywords: fungal pangenomes; comparative genomics; yeast; *Aspergillus*; *Cryptococcus*.

Abbreviations: 100GS, 100-genomes strains; BGC, biosynthetic gene cluster; CGN, conserved genomic neighbourhood; DP, dispensable pathway; GO, gene ontology; HGT, horizontal gene transfer; NCBI, National Center for Biotechnology Information; SGD, *Saccharomyces* Genome Database; UPR, unfolded protein response.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and seven supplementary figures are available with the online version of this article.

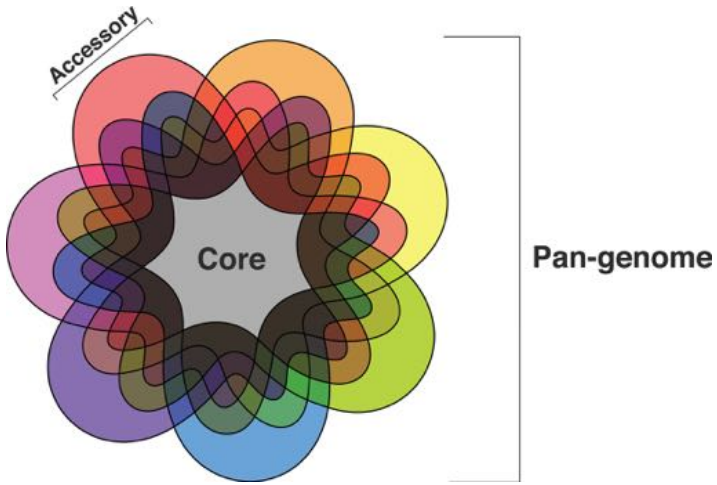


Fig. 1. Seven-set Venn diagram representing a hypothetical species pan-genome. Each set represents genes/gene models conserved across strains of a given species. The core species genome (grey) is defined as the set of all genes/gene models conserved across all strains of a species, while the accessory genome consists of all genes/gene models not universally conserved within a species.

extensively in biotechnology, *Candida albicans* is an opportunistic invasive pathogen and the second-most common cause of fungal infection, *Cryptococcus neoformans* var. *grubii* is an intracellular pathogen that causes meningitis in immunocompromised hosts, and *Aspergillus fumigatus* is an opportunistic respiratory pathogen [39–43]. We have found strong evidence for pan-genomic structure within all four fungal species. In line with previous analyses of other eukaryotes, we found that approximately 80–90 % of fungal species' pan-genomes are composed of core genes, while the remainder is composed of strain or lineage-specific accessory genes. Analysis of the origin of fungal pan-genomes suggests that fungal accessory genomes are enriched for genes of eukaryotic origin and arise via eukaryotic innovations such as gene duplication as opposed to large-scale HGT. Functionally, fungal core genomes are enriched for both housekeeping processes and essential survival processes in pathogenic species, whereas many fungal accessory gene models are found within clusters in the terminal and subterminal regions of genomes and are enriched for processes that may be implicated in fungal pathogenicity or antimicrobial resistance. Our findings complement the increasing amount of studies showing evidence for pan-genomic structure in eukaryote species.

METHODS

Dataset assembly

For each of the four fungal species chosen, we obtained strain genome assemblies from the National Center for Biotechnology Information's (NCBI's) GenBank facility (Table S1). Strains were selected based on geographical and environmental diversity where possible (Table S1). The predicted protein set from each species' reference genome was also obtained from GenBank. For each strain genome in each species dataset, translated gene model and gene model location prediction was performed using a bespoke prediction pipeline consisting of three parts (Fig. S2).

(i) Reference proteins were queried against individual strain genomes using Exonerate with a heuristic protein2genome search model [44]. Translated gene model top hits whose sequence length was ≥50 % of the query reference protein's sequence length were considered homologues and included in the strain gene model set. The genomic locations of these gene models were included in the strain genomic locations dataset.

(ii) *Ab initio* hidden Markov model-dependent gene model prediction was carried out using GeneMark-ES, with self-training and a fungal-specific branch point site prediction model enabled [45]. Predicted gene models whose genomic locations did not overlap with any gene models previously predicted via the first step were included in the strain gene model set. The genomic locations of these gene model were also included in the strain genomic locations dataset.

(iii) Finally, position weight matrix-dependent gene model prediction was carried out for all remaining non-coding regions of the genome using TransDecoder [46]. For *Saccharomyces cerevisiae* and *Candida albicans* strain genomes, these gene models were additionally screened against a dataset of known 'dubious' pseudogenes in each species taken from their respective public repositories using BLASTP with an *E* value cut-off of 10^{-4} [47, 48]. Predicted gene models whose top BLASTP hit against a known dubious pseudogene had a sequence coverage of ≥70 % were removed from further processing. All remaining predicted gene models with a length of ≥200 aa and a coding potential score of 100 or greater as assigned by TransDecoder were included in the final strain gene model set. Their corresponding genomic locations were also included in the strain genomic locations dataset.

Thus, for each strain genome in a species dataset, a gene model set and corresponding genomic location set was constructed using two initial independent prediction methods; a search for gene models orthologous to the reference protein set and an *ab initio* prediction approach, followed by a 'last resort' approach for predicting gene models in genomic regions for which gene models had not been previously called. We used this approach to ensure consistency in gene models calls between strains and to reduce the potential of poor heterogenous gene model calling within each species dataset, which would in turn reduce the number of false positives/negatives in our analysis. The completeness of each set of predicted gene models was assessed using BUSCO with the appropriate BUSCO dataset for each species [49] (Table S1). For each species dataset, all strain genome gene model sets were combined and an all-vs-all BLASTP search was carried out for all predicted gene models using an *E* value cut-off of 10^{-4}. The results of the BLASTP search were used as input for PanOCT along with the combined genomic location data for each strain genome in a species dataset [38]. Further information for each species dataset is detailed below.

Saccharomyces cerevisiae

Genomic data for 100 *Saccharomyces cerevisiae* strains were obtained from the NCBI's GenBank facility. Of these 100 genomes, 99 had previously been included in the geographically and phenotypically diverse '100-genomes strains' (100GS) resource for *Saccharomyces cerevisiae* [50]. For our analysis, we excluded the 100GS European vineyard strain M22 as its lower assembly quality prevented us from carrying out *ab initio* gene model prediction using GeneMark-ES [45, 50]. In its place, we included the European commercial winemaking strain Lalvin EC118 [51]. The protein set for the reference *Saccharomyces cerevisiae* strain S288C was also obtained from GenBank [40]. Construction of the *Saccharomyces cerevisiae* pan-genome dataset was performed as detailed above, with potentially dubious gene model predictions for each strain genome checked against a dataset of 689 known dubious *Saccharomyces cerevisiae* gene models obtained from the *Saccharomyces* Genome Database (SGD) [17]. The completeness of each strain's gene model dataset was assessed using 1711 *Saccharomyces cerevisiae* BUSCOs from the Saccharomycetales dataset; on average ~1677 BUSCOS (~98 %) were retrieved as complete gene models in each strain (Table S1). In total, 575 940 gene models and corresponding unique genomic locations were predicted for 100 *Saccharomyces cerevisiae* genomes (Table S1).

Candida albicans

Genomic data for 34 *Candida albicans* strains were obtained from the NCBI's GenBank facility, encompassing predominantly clinical or presumed-clinical strains isolated from North America, Europe and the Middle East (Table S1). The protein set for the reference *Candida albicans* strain SC5314 was also obtained from GenBank [41]. Construction of the *Candida albicans* pan-genome dataset was performed as detailed above, with potentially dubious gene model predictions for each genome checked against a dataset of 152 known dubious gene models from *Candida albicans* SC5314 obtained from the *Candida* Genome Database [48]. The completeness of each strain's gene model dataset was assessed using 1711 *Saccharomyces cerevisiae* BUSCOs from the Saccharomycetales dataset; on average ~1642 BUSCOs (~96 %) were retrieved as complete gene models in each strain (Table S1). In total, 203786 gene models and their corresponding unique genomic locations were predicted for 34 *Candida albicans* genomes (Table S1).

Cryptococcus neoformans var. *grubii*

Genomic data for 25 *Cryptococcus neoformans* var. *grubii* strains were obtained from the NCBI's GenBank facility, encompassing both clinical and wild-type strains sampled from North America and Southern African regions (Table S1). The protein set for the reference *Cryptococcus neoformans* var. *grubii* strain H99 was also obtained from GenBank [42]. Construction of the *Cryptococcus neoformans* var. *grubii* pan-genome dataset was performed as detailed above, with the exception that a check for known dubious gene models was not carried out as no such data were available for *Cryptococcus neoformans* var. *grubii*. The completeness of each strain's gene model dataset was assessed using the 1335 BUSCOs from the Basidiomycota dataset; on average ~987 BUSCOs (~74 %) were retrieved as complete gene models in each strain (Table S1). In total, 170241 gene models and their corresponding genomic locations were predicted for 25 *Cryptococcus neoformans* var. *grubii* genomes (Table S1).

Aspergillus fumigatus

Genomic data for 12 *Aspergillus fumigatus* strains were obtained from the NCBI's GenBank facility, including both clinical and wild-type strains isolated from the Northern and Southern hemispheres, and the International Space Station (Table S1). The protein set for the reference *Aspergillus fumigatus* strain AF293 was also obtained from GenBank [43]. Construction of the *Aspergillus fumigatus* pan-genome dataset was performed as detailed above, with the exception that a check for known dubious gene models was not carried out as no such data was available for *Aspergillus fumigatus*. The completeness of each strain's gene model dataset was assessed using 4046 *Aspergillus nidulans* BUSCOs from the Eurotiomycetes dataset; on average ~3410 BUSCOs (~84 %) were retrieved as complete gene models in each strain (Table S1). In total, 116230 putative proteins and their corresponding unique genomic locations were predicted for 12 *Aspergillus fumigatus* genomes (Table S1).

Pan-genome analysis of fungal species

Analysis of the pan-genomes of the four fungal species in our study was performed using the Perl software PanOCT [38]. PanOCT is a graph-based method that uses both BLAST score ratio [52] and CGN [53] approaches to establish clusters of syntenically conserved orthologues across multiple genomes for species pan-genome analysis (Fig. S1). The use of genomic context in addition to sequence similarity in PanOCT allowed us to distinguish between multiple homologous sequences within any genome analysed (i.e. paralogues) [38]. We used CGN (window size=5, the default value) as our criterion for defining conserved gene evolution between strains of fungal species. In the sections below, we refer to gene models containing an orthologue from all strains present in a species dataset as core gene models (and thus part of the core genome) and those missing an orthologue from one or more strains as accessory clusters (and thus part of the accessory genome). After removing invalid or low-quality BLASTP hits in each species dataset (Table S1), the initial core and accessory genomes for each species dataset were constructed using PanOCT with the default parameters.

To assess the influence of duplication and microsynteny loss on fungal pan-genomes, we processed the results of the PanOCT analysis using a multi-step Python/R post-processing pipeline. This first step of this pipeline was an iterative search for independent syntenic clusters with the potential to be merged based on reciprocal sequence similarity. Starting with accessory clusters of size $n-1$ (where n is the number of strains in a dataset), parallelized all-vs-all BLASTP searches of all remaining gene models from accessory clusters (e=10^{-4}) were performed, and this output was parsed to identify instances where two accessory clusters with no overlapping strain representation could be merged into one cluster based on the following criteria. (i) Each member gene model in a 'query' cluster of size m had a reciprocal BLASTP strain top hit with a sufficient number of member gene models in a 'subject' cluster of size $n-m$ or smaller. (ii) The size of the resulting 'merged' cluster was $\leq n$.

This approach attempted to account for loss-of-synteny events such as rearrangements or other artefacts arising from different genome sequencing and assembly methods. Merged accessory clusters that now had an orthologous gene model from each strain in a dataset (i.e. whose size=n) were recategorized as core clusters, although for this study such recategorizations were a rare occurrence.

The second step of our post-processing pipeline assessed the influence of gene duplication on fungal pan-genome evolution by analysing the proportion of accessory gene models that were potentially paralogous to the core genome. Gene models from accessory clusters were assessed for sequence similarity to core gene models from the initial all-vs-all BLASTP search used as input for PanOCT. If accessory gene models were sufficiently similar to every gene model from a given core cluster (E value cut-off of $1e^{-4}$), then that accessory cluster was classified as being a paralogous cluster or a cluster of duplicated core gene models. This approach attempted to account for duplication events followed by subsequent gene loss, rearrangement in strains or strain-/lineage-specific expansions of gene families. Using a sequence-based approach of pan-genome analysis, as opposed to genome alignment or other methods, also facilitated the downstream application of systematic functional analysis of species pan-genomes; e.g. gene ontology (GO)-slim enrichment, which is detailed below. We visualized the distribution of syntenic orthologues within fungal accessory genomes using the UpSet technique, an alternative to Venn or Euler diagrams, which visualizes intersections of sets and their occurrences using a matrix representation [54]. This technique, implemented in the R package UpSetR, allowed us to see the number of shared syntenic orthologues (intersections) across different strains (sets) within a species dataset [55, 56]. Singleton gene models from each reference strain genome were functionally characterized by searching against their corresponding reference protein set using BLASTP (e=10^{-4}).

Phylogenomic reconstruction of intraspecific phylogenies

Phylogenomic reconstruction of intraspecific lineages was carried out for all four fungal species using a supermatrix approach. For each fungal pan-genome dataset, all core orthologue clusters whose smallest gene model was at least 90 % the length of the longest gene model were retrieved from the dataset. Each cluster was aligned in MUSCLE with the default parameters, and for each cluster alignment phylogenetically informative character sites were extracted using PAUP* [57, 58]. Sampled alignments retaining character data were concatenated into a superalignment using FASConCAT [59].

In total, (i) 4311 *Saccharomyces cerevisiae* core clusters (431 100 gene models) passed the minimum sequence length criterion and retained alignment data after sampling,

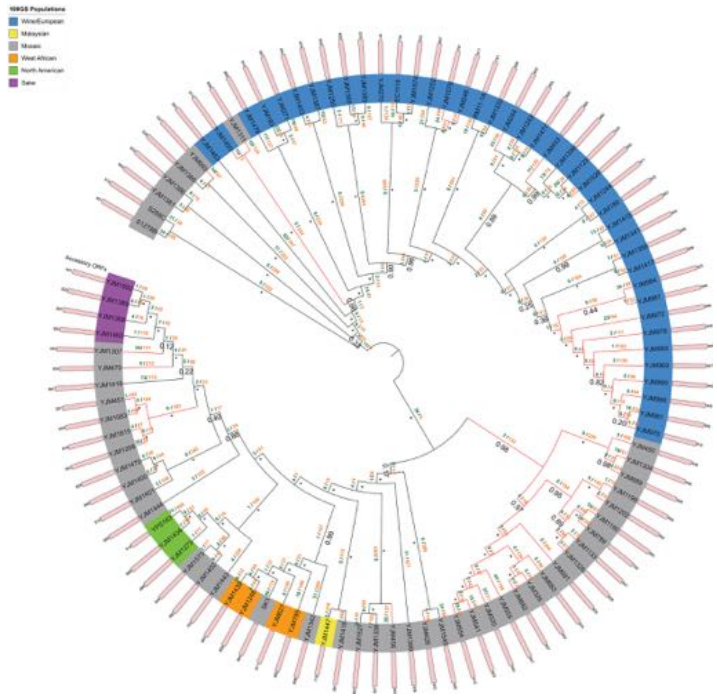


Fig. 2. Approximate maximum-likelihood supermatrix phylogeny of the *Saccharomyces cerevisiae* pan-genome dataset based on 4311 core orthologue clusters. *Saccharomyces cerevisiae* populations are as assigned by Strope *et al.*, clinical strains are indicated by red branches. Numbers below branches refer to Shimodaira–Hasegawa local supports, maximum supports are indicated by asterisks. Dollo parsimony analysis of gene model gain/loss events is annotated above branches in green and orange, respectively.

and were concatenated into a 100 genome superalignment containing 54 860 aa sites.

(ii) 4327 *Candida albicans* core clusters (68 904 gene models) retained alignment data after sampling, and were concatenated into a 34 genome superalignment containing 31 999 aa sites.

(iii) 4512 *Cryptococcus neoformans* var. *grubii* core clusters (112 800 gene models) retained alignment data after sampling, and were concatenated into a 25 genome superalignment containing 47 811 aa sites.

(iv) 5 724 *Aspergillus fumigatus* core clusters (68 904 gene models) retained alignment data after sampling for phylogenetically informative residues, and were concatenated into a 12 genome superalignment containing 20 760 aa sites.

Approximate maximum-likelihood phylogenomic reconstruction was performed for each superalignment using FastTree with the default JTT+CAT evolutionary model and Shimodaira–Hasegawa local supports [60]. All phylogenomic trees were rooted at the midpoint and annotated using the iTOL website [61] (Figs 2–5). A binary matrix was generated for the presence/absence of all orthologue clusters across all strains within each species accessory genome. Each species matrix was mapped onto the corresponding intraspecific supermatrix phylogeny and Dollo parsimony

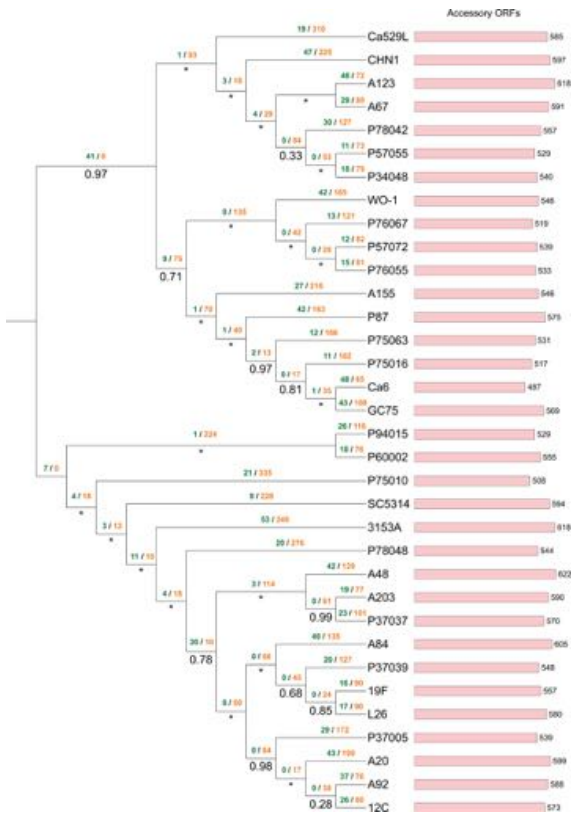


Fig. 3. Approximate maximum-likelihood supermatrix phylogeny of the *Candida albicans* pan-genome dataset based on 4327 core orthologue clusters. Numbers below branches refer to Shimodaira–Hasegawa local supports, maximum supports are indicated by asterisks. Dollo parsimony analysis of gene model gain/loss events is annotated above branches in green and orange, respectively.

analysis was performed on each matrix using Count (Figs 2–5) [62, 63]. Orthologue gain and loss events were manually annotated onto each intraspecific phylogeny.

Functional annotation and GO enrichment analysis of fungal species pan-genomes

Pfam, InterPro and GO annotation for all four fungal datasets was carried out using InterProScan [64–67]. The total numbers of proteins with at least one annotation per database from the original putative protein sets per species are given in Table 1. Enrichment analysis of GO terms was carried out for the core and accessory complements of each species' pan-genome by mapping all GO terms per species to their species GO-slim counterparts (or to the general GO-slim term basket for *Cryptococcus neoformans* var. *grubii*) and performing a Fischer's exact test analysis with

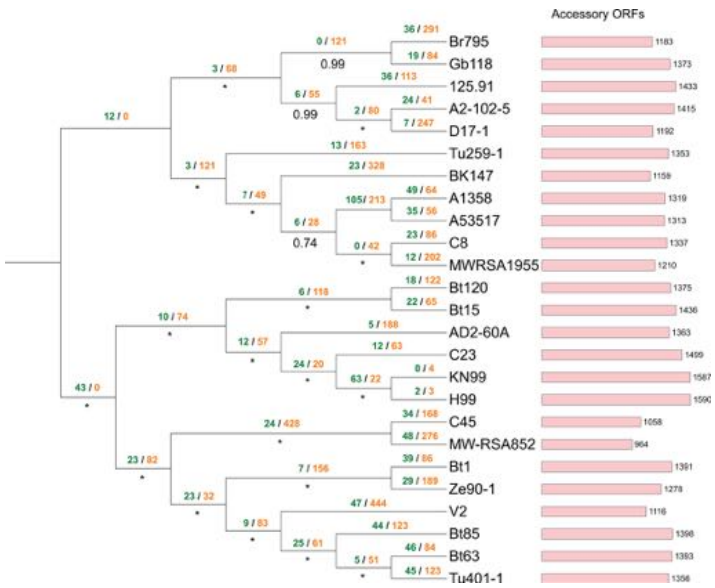


Fig. 4. Approximate maximum-likelihood supermatrix phylogeny of the *Cryptococcus neoformans* var. *grubii* pan-genome dataset based on 4512 core orthologue clusters. Numbers below branches refer to Shimodaira–Hasegawa local supports, maximum supports are indicated by asterisks. Dollo parsimony analysis of gene model gain/loss events is annotated above branches in green and orange, respectively.

parent term propagation and false discovery rate correction (P<0.05) for all complements using the Python package GOAtools (Table S2) [67–69]. False discovery rate correction was applied for all Fischer's exact tests in GOAtools using a P value distribution generated from 500 resampled P values.

Putative ancestral history of fungal core and accessory genomes

The putative evolutionary history of fungal core and accessory genomes was analysed by querying all gene models per species against a >5 million protein dataset sampled from 1109 bacterial and 488 archaeal genomes obtained from UniProt, using BLASTP with an E value cut-off of 10^{-20} [70]. Gene models were filtered by their ancestral history into three classifications using the following criteria. (i) Gene models whose hits were exclusively from bacterial or archaeal sequences were classified as 'bacterial' or 'archaeal' in origin, respectively. (ii) Gene models whose hits contained both bacterial and archaeal sequences were classified as 'undefined prokaryote' in origin. (iii) Gene models that did not hit any protein sequence in the dataset were classified as 'eukaryotic' in origin (Table S3). Pearson's χ^2 tests were carried out to determine the significance of prokaryote and eukaryote origin frequencies within the complements of each species pan-genome [68] (Table S3).

Extent of HGT in fungal accessory genomes

The extent of HGT in each fungal accessory genome was assessed by randomly selecting representative gene models from each accessory cluster and searching these using BLASTP with an E value cut-off of $1e^{-20}$ against a dataset representative of fully sequenced prokaryotic and eukaryotic species. This dataset was composed of over 8 million protein sequences from 1698 genomes sampled from all three domains of life that had been used in previous interdomain HGT analysis [71], as well as all predicted gene models per species dataset. Putative interdomain HGT events were

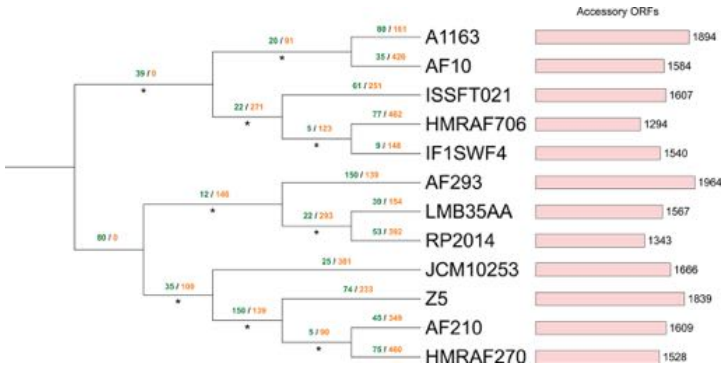


Fig. 5. Approximate maximum-likelihood supermatrix phylogeny of the *Aspergillus fumigatus* pan-genome dataset based on 5724 core orthologue clusters. Numbers below branches refer to Shimodaira–Hasegawa local supports, maximum supports are indicated by asterisks. Dollo parsimony analysis of gene model gain/loss events is annotated above branches in green and orange, respectively.

identified by locating gene models whose first top hit outside either the sequence's source species or genus was prokaryotic in origin. Putative HGT events identified by either filter are given per species in Table S3. Putative intrakingdom fungal HGT events were identified by filtering the same BLASTP output for gene models whose first top hit outside the sequence's source species was fungal in origin but not from the same genus (Table S3).

Chromosomal location of core and accessory gene models in species reference genomes

Pearson's χ^2 tests were carried out for the global frequencies of core and accessory gene models along the subterminal regions of chromosomes, which we defined as approximately the first and last 10 % of each chromosome, in each reference genome (Table S4). Pearson's χ^2 tests were also carried out for the frequencies of core and accessory gene models per chromosome for each reference genome (Table S4) [68]. The chromosomal locations of core and accessory gene models along each reference genome were visualized using the Ruby software PhenoGram [72].

Distribution of knockout viability phenotypes in *Saccharomyces cerevisiae* S288C

All available knockout phenotype data for *Saccharomyces cerevisiae* S288C were obtained from the SGD [73]. A reciprocal BLASTP search was carried out between all 5815 *Saccharomyces cerevisiae* S288C gene models from our *Saccharomyces cerevisiae* pan-genome dataset and the reference protein set for *Saccharomyces cerevisiae* S288C with an *E* value cut-off of 10^{-20} to match predicted proteins to orthologues from the reference protein set. Knockout phenotype viability data, if available, was then inferred for each of our *Saccharomyces cerevisiae* S288C gene models that had a reciprocal reference orthologue. Pearson's χ^2 tests were carried out for the frequencies of knockout phenotype viability in both the core and accessory genomes of *Saccharomyces cerevisiae* S288C (Table S5).

Distribution of 'dispensable pathway' (DP) genes in the *Saccharomyces cerevisiae* pan-genome

Data for 14 DP gene clusters containing 41 genes found in *Saccharomyces cerevisiae* was taken from a previously

Table 1. Number of gene models in our four fungal pan-genome datasets with at least one annotation term per annotation type

Percentage of annotated gene models relative to pan-genome datasets shown in parentheses.

Species	Pfam	InterPro	GO
Saccharomyces cerevisiae	468 511 (81 %)	455 582 (79 %)	312 161 (54 %)
Candida albicans	161 235 (79 %)	155 271 (76 %)	105 694 (52 %)
Cryptococcus neoformans	111 305 (65 %)	106 655 (63 %)	72 243 (42 %)
Aspergillus fumigatus	83 239 (71 %)	79 231 (68 %)	54 457 (46 %)

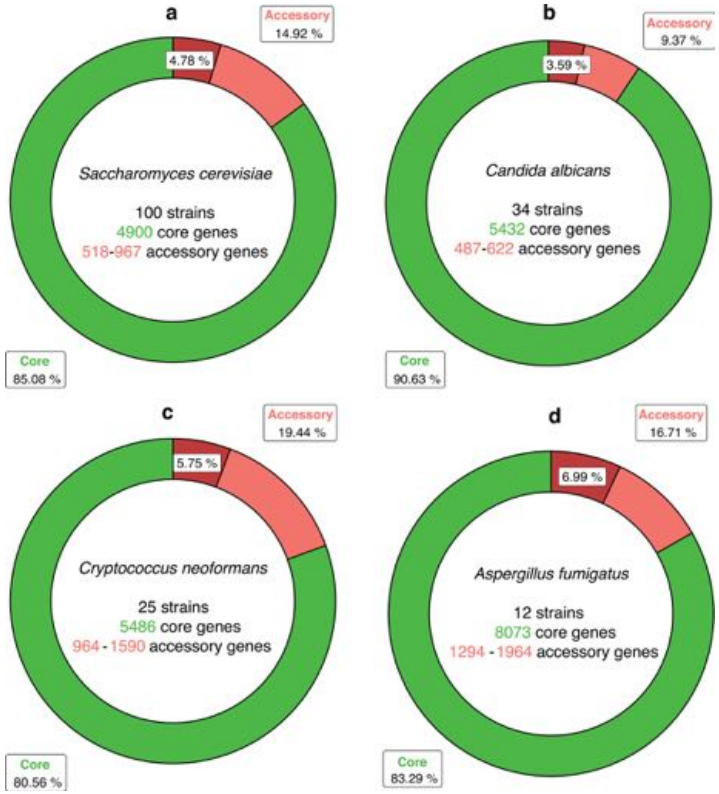


Fig. 6. Pan-genomes of four fungal species. (a) *Saccharomyces cerevisiae*, (b) *Candida albicans*, (c) *Cryptococcus neoformans* var. *grubii*, (d) *Aspergillus fumigatus*. The ring charts represent the total number of gene models in pan-genome complements expressed as a proportion of total pan-genome size. Sections in dark-red represent duplicated core gene models in the accessory genome.

published analysis of biotin reacquisition in yeast species [74]. A total of 38 DP genes were extracted from the *Saccharomyces cerevisiae* S288C reference protein set, encompassing 13 of the 14 DP clusters. A reciprocal BLASTP search was performed between these genes and all 5815 *Saccharomyces cerevisiae* S288C gene models from the *Saccharomyces cerevisiae* pan-genome dataset with an *E* value cut-off of 10^{-20} to identify DP genes in our predicted gene model set. All 38 DP genes had a unique reciprocal match with a predicted gene model in *Saccharomyces cerevisiae* S288C. A binary matrix was generated for the presence/absence of syntenic orthologues of DP genes from *Saccharomyces cerevisiae* S288C in the *Saccharomyces cerevisiae* pan-genome dataset (Table S5).

Distribution of biosynthetic gene clusters (BGCs) in the *Aspergillus fumigatus* pan-genome

Data for 33 known BGCs encompassing 307 genes in *Aspergillus fumigatus* Af293 were obtained from a previous analysis of secondary metabolism in *Aspergillus fumigatus* [75]. *Aspergillus fumigatus* Af293 gene models from the *Aspergillus fumigatus* pan-genome dataset were matched to their homologues from the reference gene data set using a reciprocal BLASTP search with an *E* value cut-off of 10^{-20}. A binary matrix was constructed for the presence/absence of syntenic orthologues of the 307 putative BGC genes from *Aspergillus fumigatus* Af293 within the *Aspergillus fumigatus* pan-genome dataset (Table S5).

RESULTS

Analysis of the *Saccharomyces cerevisiae* pan-genome

Overall, 575 940 gene models were predicted across all 100 *Saccharomyces cerevisiae* strains with a mean of 5759 gene models predicted per strain (Tables 2 and S1). These 575 940 gene models were distributed across 7750 unique syntenic orthologue clusters (Table 2). The core *Saccharomyces cerevisiae* genome contained 4900 gene models, which were conserved across 100 *Saccharomyces cerevisiae* strains (490 000 gene models in total, 85 % of the total species pan-genome). For individual strain genomes, this corresponded to between 83 and 90 % of their total predicted gene model content (Fig. 6a, Table S1). The remaining 85 940 predicted gene models were accessory gene models, distributed across 2850 clusters, with strain accessory genome sizes ranging from 518 to 967 gene models per *Saccharomyces cerevisiae* strain (mean size =~859 gene models). Further analysis of the *Saccharomyces cerevisiae* species accessory genome identified that ~32 % of accessory gene models (776 clusters, 4.77 % of the total species pan-genome) were duplicates of core gene models conserved across one or more strains. This corresponded to a mean of 275 gene models per *Saccharomyces cerevisiae* strain, and 27 511 gene models in total (Tables 2 and S1). Overall, 455 syntenic clusters (encompassing 45 045 accessory gene models) were missing a syntenic orthologue in only one other strain and 1416 accessory gene models were singletons. Analysis of the distribution of orthologues within the *Saccharomyces cerevisiae* accessory genome using the R package UpSetR showed that the most frequent sets are singleton gene models or syntenic clusters missing a syntenic orthologue in one strain, with YPS163 having the most singleton genes (74 in total) (Fig. S3). Other strains (e.g. YJM1477) lacked singleton gene models altogether (Fig. 2). There were 13 756 gene models (from 1935 syntenic clusters) that did not have a syntenic orthologue in *Saccharomyces cerevisiae* S288C. Of these non-reference gene models, 1385 were singleton gene models found only in one strain. The widest-distributed non-reference gene model was present in 93 strains and there was no accessory gene model solely missing from *Saccharomyces cerevisiae* S288C. YPS163 had the smallest accessory genome of the 100 yeast strains (518 gene models) and YJM271 had the largest (967 gene models) (Fig. 2).

Phylogenomic reconstruction of all 100 *Saccharomyces cerevisiae* strains resolved two major groups; a clade containing strains and mosaics derived from Malaysian, West African, North American and sake populations, and a clade containing strains and mosaics derived from wine/European populations (Fig. 2). Each of the non-mosaic populations as assigned by Strope *et al.* [50] present in the dataset (except the singleton Malaysian strain YJM1447) resolved to a monophyletic geographical group [50]; the placement of the mosaic laboratory strain SK-1 in a West African clade is consistent with its West African origin [76], and the clinical mosaic strain YJM1311 is of predominantly wine/European ancestry; hence, its placement at the base of the wine/European clade [50] (Fig. 2). Many of the remaining mosaic strains branched close to non-mosaic clades that shared their dominant population fraction as determined by Strope *et al.* [50]; for example, many of the clinical mosaic strains placed adjacent to the sake clade had predominantly sake population ancestry [50] (Fig. 2). Three strains (YJM248, YJM1252, YJM1078) identified by Strope *et al.* [50] as having a higher relative proportion of introgressed genes than other *Saccharomyces cerevisiae* strains (potentially arising from recent hybridization with *Saccharomyces paradoxus*) formed a monophyletic branch within the previously described wine/European clade [50].

Analysis of the *Candida albicans* pan-genome

A total of 203 786 gene models were predicted across all 34 *Candida albicans* strain genomes, with a mean of 5993 gene models predicted per strain, distributed across 7325 unique syntenic orthologue clusters (Tables 2 and S1). The core *Candida albicans* genome contained 5432 gene models that were conserved across 34 *Candida albicans* strains (184 688 in total, 90 % of the total species pan-genome). This corresponded to between 89 and 91 % of the total predicted gene models for each strain genome (Fig. 6b, Table S1). The remaining 19 098 predicted gene models were accessory gene models, distributed across 1893 clusters, with strain accessory genome sizes ranging from 487 to 622 gene models per *Candida albicans* strain (mean size =~561 gene models) (Tables 2 and S1). Further analysis of the *Candida albicans* species accessory genome identified that ~38 % of accessory gene models (1013 clusters, ~3.59 % of the total species pan-genome) were duplicates of core gene models conserved across one or more strains. This corresponded to a mean of 215 gene models per *Candida albicans* strain, and 7312 gene models in total (Tables 2 and S1). Of the 19 098 *Candida albicans* accessory gene models identified, 3624 accessory gene models (from 268 syntenic clusters) were missing a syntenic orthologue in only one other strain, while 928 gene models were singletons. UpSet analysis of the distribution of orthologues within the *Candida albicans* accessory genome showed that 1056 gene models (32 syntenic clusters) from 33 *Candida albicans* strains were missing an orthologue in *Candida albicans* WO-1 and *Candida albicans* 3153A had 53 putative gene models with no orthologue in any other strain (Fig. S4). SC5314 had the smallest number of singleton gene models (nine in total). *Candida albicans* A48 had the largest accessory genome (622 gene models) and *Candida albicans* Ca6 had the smallest (487 gene models) (Fig. 3). Phylogenomic reconstruction of all 34 *Candida albicans* strains resolved two main groups when rooted at the midpoint; one containing the exemplar *MTL*-homozygous strain WO-1 and a ladderized group containing the reference strain SC5314 (Fig. 3).

Analysis of the *Cryptococcus neoformans* var. *grubii* pan-genome

A total of 170 241 gene models were predicted across all 25 *Cryptococcus neoformans* var. *grubii* strain genomes, with a mean of 6809 gene models predicted per strain, distributed across 8193 unique syntenic orthologue clusters (Tables 2 and S1). The core *Cryptococcus neoformans* var. *grubii* genome contained 5486 gene models that were conserved across 25 *Cryptococcus neoformans* var. *grubii* strains (137 150 in total, 80 % of the total species pan-genome). This corresponded to between 76 and 85 % of the total predicted gene models for each strain genome (Fig. 6c, Table S1). The remaining 33 091 predicted gene models were accessory gene models distributed across 2698 clusters, with strain accessory genome sizes ranging from 964 to 1654 gene models per *Cryptococcus neoformans* var. *grubii* strain (mean size =~1334 gene models) (Table S1). Detailed analysis of the *Cryptococcus neoformans* var. *grubii* species accessory genome identified that ~29 % of accessory gene models (776 clusters, ~5.8 % of the total species pan-genome) were duplicates of core gene models conserved across one or more strains. This corresponded to a mean of ~391 gene models per *Cryptococcus neoformans* var. *grubii* strain, and 9794 gene models in total (Tables 2 and S1). Overall 674 *Cryptococcus neoformans* var. *grubii* clusters (encompassing 16 032 accessory gene models) were missing a syntenic orthologue in only one other strain and 668 accessory gene models were singletons. UpSet analysis of the distribution of orthologues within the *Cryptococcus neoformans* var. *grubii* accessory genome showed that 3600 gene models (150 syntenic clusters) from 24 *Cryptococcus neoformans* var. *grubii* strains were missing an orthologue in *Cryptococcus neoformans* var. *grubii* MWRSA852, whereas the *Cryptococcus neoformans* var. *grubii* A1358 genome had 49 putative gene models with no orthologue in any other strain (Fig. S5). KN99 had no singleton gene models, but it should be noted that that strain is an isogenic derivative of the reference H99 strain. *Cryptococcus neoformans* var. *grubii* H99 itself had the largest accessory genome (1590 gene models) and *Cryptococcus neoformans* var. *grubii* MW-RSA852 had the smallest (964 gene models) (Fig. 4). The most frequent sets found in the accessory genome include both singleton genes and clusters missing orthologues from one or two strains. Phylogenomic reconstruction of all 25 strains using a 47 811-site amino acid supermatrix derived from the core *Cryptococcus neoformans* var. *grubii* genome resolved two monophyletic groups when rooted at the midpoint (Fig. 4).

Analysis of the *Aspergillus fumigatus* pan-genome

A total of 116 311 gene models were predicted across all 12 *Aspergillus fumigatus* strain genomes, distributed across 11 075 unique syntenic orthologue clusters, with a mean of 9692 gene models predicted per strain. The core *Aspergillus fumigatus* genome contained 8073 core gene models that were conserved across 12 *Aspergillus fumigatus* strains (96 876 in total, 83 % of the total species pan-genome). This corresponded to between 80 and 86 % of the total predicted gene models for each strain genome (Fig. 6d, Table S1). The remaining 19 435 predicted gene models were accessory gene models distributed across 3002 clusters, with strain accessory genome sizes ranging from 1294 to 1964 gene models per *Aspergillus fumigatus* strain (mean size =~1619 gene models) (Table S1). Detailed analysis of the *Aspergillus fumigatus* species accessory genome identified that ~41 % of accessory gene models (1170 clusters, ~6.9 % of the total species pan-genome) were duplicates of core gene models that were conserved across one or more strains. This corresponded to a mean of 677 gene models per *Aspergillus fumigatus* strain, and 8127 gene models in total. Overall, 7953 gene models (from 958 syntenic clusters) were missing a syntenic orthologue in only one other strain, whereas 723 gene models were singletons.

UpSet analysis of the orthologue distribution in the *Aspergillus fumigatus* accessory genome found that 2167 gene models (197 syntenic clusters) from 11 *Aspergillus fumigatus* strains were missing an orthologue in *Aspergillus fumigatus* IFISWF4 and the reference *Aspergillus fumigatus* Af293 genome has 150 putative gene models with no

Table 2. Pan-genomes of four fungal species: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*

Duplicated core gene models (GMs) and clusters in the accessory genomes are given in parentheses.

		Core genome		Accessory genome		Pan-genome	
Species	Strains	GMs	Clusters	GMs	Clusters	GMs	Clusters
Saccharomyces cerevisiae	100	490 000	4900	85 940 (27 511)	2850 (776)	575 940	7750
Candida albicans	34	184 688	5432	19 098 (7312)	1893 (1013)	203786	7325
Cryptococcus neoformans	25	137 150	5486	33 091 (9974)	2698 (776)	170241	8193
Aspergillus fumigatus	12	96 876	8073	19 435 (8127)	3002 (1170)	116 311	11 075

orthologue in any other strain (Fig. S6). The latter may be due to a lower degree of strain sampling within the *Aspergillus fumigatus* dataset or the reference genome having a higher-quality assembly than other strains of *Aspergillus fumigatus*. The Z5 strain has the smallest number of singleton gene models (nine in total). *Aspergillus fumigatus* Af293 has the largest accessory genome (1964 gene models) and *Aspergillus fumigatus* HMRAF706 has the smallest (1294 gene models) (Fig. 5). Phylogenomic reconstruction of all 12 strains using a 20 760-site amino acid supermatrix derived from the core *Aspergillus fumigatus* genome resolved two monophyletic groups when rooted at the midpoint, one containing both International Space Station strains and *Aspergillus fumigatus* Af10, and one containing all three environmental strains as well as *Aspergillus fumigatus* Af293 and Af210 (Fig. 5). The placement of the two International Space Station strains as well as the aforementioned individual clinical strains is in relative agreement with the most extensive intraspecific *Aspergillus fumigatus* phylogeny published [77].

GO enrichment in fungal core and accessory genomes

Analysis of the distribution of GO terms in fungal core genomes shows that many housekeeping biological processes, such as translation, nucleic acid metabolism and oligopeptide metabolism, are significantly over-represented in each species (P<0.05) (Table S2). Furthermore, molecular function terms for enzymatic and nucleic acid binding activity are also significantly over-represented (Table S2). In fungal accessory genomes, terms relating to transport and localization of proteins, carbohydrate metabolism, as well as protein modification and carboxyl acid metabolism, are significantly over-represented in many species (Table S2). Terms relating to housekeeping processes are significantly under-represented in fungal accessory genomes compared to core genomes. There are no common or synonymous cellular component or molecular function terms that are significantly under-represented across all four fungal accessory genomes in our analysis. However, terms relating to the functions of intracellular membrane-bound organelles are significantly over-represented in the accessory genomes of both *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus* (Table S2).

Many broad and granular housekeeping terms relating to nucleic acid and protein biological processes are significantly over-represented within the core genome of *Saccharomyces cerevisiae* (Table S2). In addition to transport processes, genes potentially involved in vitamin metabolism and protein dephosphorylation are significantly over-represented within the core genome of *Saccharomyces cerevisiae*. Similar terms are also significantly over-represented within the core genome of *Candida albicans* (Table S2). The *Cryptococcus neoformans* var. *grubii* core genome is significantly over-represented in some unique terms involved in regulation of homeostasis and biological quality, functional pathways such as the unfolded protein response (UPR) pathway, as well as signal transduction (Table S2). There are fewer terms that are significantly over-represented within the *Cryptococcus neoformans* var. *grubii* accessory genome than in the other fungal accessory genomes in this study. Those terms that are significantly over-represented in the *Cryptococcus neoformans* var. *grubii* accessory genome are also found elsewhere, e.g. transport. The core *Aspergillus fumigatus* genome is significantly over-represented in terms related to small molecule biosynthesis and other biosynthetic processes (Table S2). Within the *Aspergillus fumigatus* core genome, terms relating to vesicle-mediated transport and carboxylic acid metabolism are significantly over-represented, these terms are also significantly over-represented in the *Saccharomyces cerevisiae* core genome.

Ancestral origin of fungal core and accessory genomes

The ancestral origin of fungal core and accessory genomes was inferred via BLASTP searches ($1e^{-20}$) of fungal gene models against >5 million prokaryotic sequences from >1500 bacterial and archaeal genomes. Gene models that had hits with prokaryotic sequences exclusively were classified as having originated within the prokaryotes (broken down further by prokaryotic kingdom in Table S3), and gene models that lacked a BLASTP hit against the prokaryotic database were classified as having originated within the eukaryotes. Using these criteria, for each fungal pan-genome dataset between 69 and 77 % of all gene models were inferred as eukaryotic in origin. Similar proportions of gene models inferred as having originated within eukaryotes were also observed in fungal core genomes. Higher proportions of gene models with a putative origin within eukaryotes was observed in fungal accessory genomes (74–81 % of all accessory gene models in each species). Statistical analysis of the ancestral history of each fungal species pan-genome found that each fungal accessory genome was significantly enriched for genes of eukaryotic origin and each fungal core genome was significantly enriched for genes of prokaryotic origin (P<0.05) (Table S3).

Interdomain and intrakingdom HGT into fungal accessory genomes

Systematic screening for interdomain HGT events in each fungal accessory genome revealed small numbers of putative HGT events from prokaryote sources per species, ranging from a single event in the *Candida albicans* accessory genome to 11 events in the *Aspergillus fumigatus* accessory genome (Table S3). The distribution of these putative HGT genes in fungal accessory genomes varies from strain-unique singleton genes (particularly in *Saccharomyces cerevisiae*) to more widely distributed genes (as seen in *Cryptococcus neoformans* and *Aspergillus fumigatus*) (Table S3). The majority of potential prokaryote donors are soil-dwelling bacteria, such as *Clostridium pasteurianum* (a donor to the *Aspergillus fumigatus* accessory genome) and *Acinetobacter pittii* (a donor to the *Saccharomyces cerevisiae* accessory genome). We then applied a similar screen for recent HGT from other fungal species, which suggested up to 8 % of fungal accessory genomes may have arisen via intrakingdom HGT. The largest extent of such intradomain HGT appeared to have occurred into the accessory genomes of *Cryptococcus neoformans* and *Aspergillus fumigatus* (420 and 391 potential events, respectively) (Table S3). In each accessory genome, putative HGT-derived gene models appear to have been transferred mainly from closely related species or species that share similar niches. For example, *Aspergillus fumigatus* is a potential donor of three *Candida albicans* accessory gene models (Table S3). However, further comprehensive investigations are required to confidently confirm that these HGT events are bona fide.

Chromosomal location of core and accessory genomes in fungal reference genomes

Between 17 and 21 % of all predicted gene models for each fungal reference strain lie in the subterminal regions of that strain's genome. Approximately 15 % of all core gene models in both *Saccharomyces cerevisiae* S288C and *Cryptococcus neoformans* var. *grubii* H99 are found in their subterminal regions, whereas this proportion is higher in *Candida albicans* SC5314 and *Aspergillus fumigatus* Af293 (~21 and ~18 % of all core gene models, respectively). *Candida albicans* SC5314 has a lower proportion of accessory gene models (115 of 594 gene models, ~19 % of its total accessory genome) found in subterminal regions than the other three fungal species, where that proportion is ~28–33 % of their total accessory genomes. There is a statistically significant bias (P<0.05) towards accessory gene models in the subterminal regions of *Saccharomyces cerevisiae* S288c, *Cryptococcus neoformans* var. *grubii* H99 and *Aspergillus fumigatus* Af293, with a corresponding bias (P<0.05) towards core gene models in the non-subterminal regions of each genome (Table S4). In contrast, there is no significant pattern in the distribution of accessory gene models in *Candida albicans* SC5314, and instead its subterminal regions are significantly enriched for core gene models (P<0.05) (Table S4). Statistical analysis of core and accessory gene model enrichment per chromosome in each reference genome found that at least one chromosome was significantly enriched for core gene models and another chromosome was significantly enriched for accessory gene models per genome (P<0.05) (Table S4). The number of chromosomes per genome that were significantly biased towards either core or accessory gene models ranged from two in *Candida albicans* SC5314 (chromosomes 2 and 7) to six in *Saccharomyces cerevisiae* S288C (chromosomes I–III, VI, VIII and XIII) (Table S4). Visualizing chromosomal plots showed that clustering of accessory genes mostly occurred in subterminal regions of fungal genomes (Fig. S7a–d). There are some exceptions: some chromosomes in *Saccharomyces cerevisiae* S288c, *Cryptococcus neoformans* var. *grubii* H99 and *Aspergillus fumigatus* Af293 had at least one larger accessory gene cluster closer to the chromosomal midpoint (Fig. S7a, c–d). In contrast, there appeared to be no major clustering of accessory genes in any chromosome in *Candida albicans* SC5314 (Fig. S7b).

Knockout viability of core and accessory genes in *Saccharomyces cerevisiae* S288C

A total of 5343 predicted *Saccharomyces cerevisiae* S288C gene models from the species pan-genome dataset, encompassing 4730 core gene models and 613 accessory gene models, were assigned their reference homologue's corresponding knockout viability phenotype. The remaining 472 predicted gene models from *Saccharomyces cerevisiae* S288C did not have a knockout viability phenotype assigned to them, either due to the lack of a unique reciprocal BLASTP hit or a lack of viability data for the reference homologue (Table S5). Those *Saccharomyces cerevisiae* S288C gene models that had knockout phenotype data were predominantly knockout-viable;~79 % of annotated core gene models and ~88 % of annotated accessory gene models had a reciprocal reference homologue with a viable knockout phenotype (Table S5). There was no significant bias in the distribution of knockout viability within the core *Saccharomyces cerevisiae* S288C genome, i.e. the core genome was enriched for neither knockout-viable or knockout-inviable gene models (of those which had knockout phenotype data available) (Table S5). The *Saccharomyces cerevisiae* S288C accessory genome, however, was over-represented for knockout-viable gene models (P<0.05) (Table S5).

DP gene clusters in the *Saccharomyces cerevisiae* pan-genome

All 38 reference DP genes had a unique reciprocal homologue within the set of predicted *Saccharomyces cerevisiae* S288C gene models taken from our pan-genome dataset (Table S5). One of the 13 reference DP clusters was syntenically conserved within all strains in the *Saccharomyces cerevisiae* pan-genome dataset; a three-member *GAL* cluster involved in galactose utilization. Some clusters are widely conserved within the dataset, but are missing a member gene in a small number of strains; these include a three-member *BIO* cluster that mediates biotin uptake, a *SNO1-SNZ1* vitamin B6 metabolism cluster and a large six-member *DAL-DCG* cluster that enables utilization of allantoin as a nitrogen source (Table S5). Other clusters had more patchy distribution within the species pan-genome, most notably a three-member *ARR* gene cluster that confers arsenic resistance was missing a member gene (*ARR3*) in 49 out of 100 strains (Table S5). Some clusters, such as a four-member *FIT/FRE* iron uptake cluster, are completely missing in a small number of strains (Table S5).

BGCs in the *Aspergillus fumigatus* pan-genome

A total of 307 known biosynthetic genes from 33 BGCs in *Aspergillus fumigatus* Af293 had a unique reciprocal homologue within the set of predicted *Aspergillus fumigatus* Af293 gene models from the *Aspergillus fumigatus* pan-genome [75]. A total of 240 of the 307 known biosynthetic genes were core genes found in all 12 *Aspergillus fumigatus* strains, none of which were unique to *Aspergillus fumigatus* Af293 alone (Table S5). There were 14 *Aspergillus fumigatus* BGCs that were completely conserved (i.e. all genes within

that cluster are core genes), which included known mycotoxin-producing BGCs such as fumagillin and gliotoxin clusters (Table S5). Other BGCs were found to have one or two genes missing, potentially due to synteny loss or pseudogenization. Some BGCs showed far more variable distribution within the *Aspergillus fumigatus* pan-genome; for example, a polyketide synthase (PKS) cluster was wholly conserved in four strains (Af293, Z5, HMRAF270 and JCM10253) and absent or translocated in the other eight, and a fusarielin-like cluster was completely absent from A1163 and only partially present in some strains but was wholly conserved in others (Table S5).

DISCUSSION

Applying genomic context in eukaryotic pan-genome analysis

To investigate pan-genomic structure within four fungal species, we adapted a method previously used in bacterial pan-genome analysis and implemented in PanOCT (Pan-genome Ortholog Clustering Tool) [38]. Our rationale for using this method to construct species pan-genomes was that it allowed us to investigate intraspecific variability on a gene-to-gene level, as opposed to defining core and accessory genomes based on families of related gene models (e.g. a core gene family may be present in all strains of a species, but the number of genes belonging to that family will usually vary between strains). This allowed us to see which genes and biological functions were relatively conserved in their distribution, and which had varying expansion and distribution in fungal species. A similar approach was used in a previous analysis of genome variation in *Saccharomyces* species, but was limited to assessing syntenic conservation of reference homologues using immediately adjacent genes [34]. To ensure consistency between strain genomes in each of our datasets we constructed a custom gene model prediction pipeline that used three different predictive methods to generate a unique set of predicted gene models and their genomic locations (i.e. no isoforms) per strain genome (Fig. S2) [44–46]. As our definition of what constitutes a core or accessory gene model is quite stringent compared to other pan-genome analyses, we also developed a post-processing pipeline that attempted to account for loss of microsynteny between fungal strain genomes and to also examine the extent of duplication of core genome content within fungal accessory genomes.

Pan-genomes of four model fungi

We chose to investigate the potential pan-genomic structure of four model fungal species: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. In addition to their impact on human lifestyle, each species chosen is a model organism for fungal evolutionary biology, genomics and comparative genomics. *Saccharomyces cerevisiae* was the first eukaryote to have its genome sequenced, and the other three species each had their genome sequenced during the initial wave of fungal genomics research in the early to mid-2000s [39–43, 78]. Our selection covers fungal species with different genomic characteristics; *Saccharomyces cerevisiae* has undergone ancestral whole-genome duplication and *Candida albicans* has an alternative genetic code [79, 80], whereas *Cryptococcus neoformans* and *Aspergillus fumigatus* are more intron-dense than either *Saccharomyces cerevisiae* or *Candida albicans* and extensive alternative splicing occurs in *Cryptococcus* species [81, 82]. Our selection also covers fungal species with different evolutionary histories. *Saccharomyces cerevisiae*, *Candida albicans* and *Aspergillus fumigatus* are members of the fungal phylum Ascomycota; the former two are closely related members of the subphylum Saccharomycotina, which includes many typical commensal and pathogenic yeasts that reproduce by budding, while *Aspergillus fumigatus* is a member of the large subphylum Pezizomycotina of filamentous fungi [78]. *Cryptococcus neoformans* var. *grubii* superficially resembles many yeast species and also replicates by budding, but is a member of the phylum Basidiomycota and is more closely related to multicellular fungi within the subphylum Agaricomycotina than other yeast species [78]. Genome assemblies available on GenBank for each species at the time of writing ranged from 12 for *Aspergillus fumigatus* to >400 for *Saccharomyces cerevisiae* [36].

Our species pan-genome for *Saccharomyces cerevisiae* was constructed using genomic data from 100 strains, 99 of which were previously included in the 100GS resource (Table S1) [50]. The resource includes 7 *Saccharomyces cerevisiae* genomes sequenced prior to 2015 and 93 *Saccharomyces cerevisiae* genomes sequenced *de novo* by the 100GS authors, taken from diverse genotypic and phenotypic backgrounds (populations referred to henceforth are as assigned by the 100GS authors after Liti *et al.* [50, 83]). The resource covers strains from laboratory, biotech, clinical and wild populations, which makes it an excellent dataset for carrying out *Saccharomyces cerevisiae* population genomics and pan-genomics studies of this kind. In their analysis, the 100GS authors screened *Saccharomyces cerevisiae* strains for aneuploidy, introgressed genes, phenotypically relevant single-nucleotide polymorphisms and non-reference genomic content [50]. The 100GS authors also assessed levels of resistance to environmental stresses such as sulphite and copper resistance, as well as fungicides such as ketoconazole [50].

A more recent study of 1011 *Saccharomyces cerevisiae* genomes included an analysis of the pan-genome of *Saccharomyces cerevisiae* in which the authors of that study detected non-reference genomic content by aligning strain genomes to the S288C genome using BLASTN, and extracting and annotating unique non-reference genes using an integrative multi-method procedure [36, 84]. Notably, despite a tenfold difference in the number of input strains, and different methods of identifying core and accessory genome content, both their study (4940 core genes) and our own (4900 core gene models) predict a similar-sized core *Saccharomyces cerevisiae* genome [36]. The 1011 genome study predicted an almost identical accessory genome to our analysis also; they identified 2856 accessory genes with varying distribution across 1011 genomes [36], whereas we identified an accessory genome of 2850 genes for our pan-genome dataset. The 1011 genome study also observed a number of evolutionary and functional trends within the *Saccharomyces cerevisiae* accessory genome; accessory genes were clustered within the subterminal regions of *Saccharomyces cerevisiae* genomes and some accessory genes may have originated via HGT from divergent yeast species or other fungi [36]. We observe similar trends in our analysis of the *Saccharomyces cerevisiae* accessory genome.

For the remaining three species, we constructed species pan-genome datasets based on strain genome assemblies that were available from GenBank at the start of our analyses. For each of these datasets, we attempted to sample strain genomes with as many diverse characteristics (e.g. geographical location, phenotype) as was possible with the genome assembly data available. Although there are a smaller number of strains sampled for these species pan-genomes, the sizes of these species' core and accessory genomes are in line with our analysis of *Saccharomyces cerevisiae*, as well as larger analyses of species pan-genomes in fungi and other taxa. The *Candida albicans* species pan-genome dataset was constructed using data from 34 strains, predominantly clinical in origin, including both homozygous and heterozygous *MTL* mating-type strains (Table S1) [85]. A substantial amount of genome assembly data available for *Candida albicans* comes from strains isolated in hospitals; of the 34 strains in our dataset, 14 strains were clinical isolates from the USA alone (Table S1). A number of other strains were isolated from European and Middle East sources, but for 13 strains no information was available on the isolate source for the genome from GenBank. Perhaps as a consequence of a lower degree of environmental diversity due to sampling primarily clinical strains, the *Candida albicans* pan-genome has the smallest proportion of accessory gene content of the four species analysed in this study (~9% of the entire species pan-genome). The *Candida albicans* pan-genome also has the lowest degree of variation in accessory genome size between individual strains of the four species analysed (Figs 3 and 6b). The UpSet distribution of the *Candida albicans* accessory genome illustrates this lower degree of variability within the *Candida albicans* pan-genome, as the most frequent sets are either singleton clusters or clusters that are missing an orthologue from one strain (Fig. S4). Despite this caveat, however, the *Candida albicans* pan-genome otherwise exhibits many of the same functional and evolutionary trends seen in the other three species we have investigated (as detailed below). With a broader sampling of strains found outside of a clinical context, a more accurate picture of the size of the *Candida albicans* accessory genome will be attained.

In contrast to *Candida albicans*, both our *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus* pan-genome datasets were constructed using a diverse array of strain genomes taken from both clinical and wild environments. The *Cryptococcus neoformans* var. *grubii* pan-genome dataset was constructed using clinical strain genomes isolated predominantly from human immunodeficiency virus positive patients from the USA and Botswana and wild-type strains sampled from Southern Africa sources (Table S1). *Cryptococcus neoformans* var. *grubii* has the largest proportion of accessory genes of the four species analysed (~20% of the entire species pan-genome). As *Cryptococcus neoformans* is an intracellular pathogen in humans, it has to adapt to extreme variations in environmental stresses in order to survive. This is thought to lead to the high level of genomic rearrangement and instability seen in *Cryptococcus neoformans* [86]. It is possible that this in turn creates more novel genetic content, which may explain the higher level of accessory genome content seen in *Cryptococcus neoformans* var. *grubii*. Genomic instability as a result of pathogenic lifestyle fuelling pan-genome evolution has previously been observed in the wheat pathogen *Z. tritici* [37]. The *Aspergillus fumigatus* pan-genome dataset was constructed using 12 strain genomes sampled from clinical environments in the UK, USA and Canada, wild-type samples taken from China and from South American forest floors, and 2 strains isolated from surfaces within the International Space Station [77] (Table S1). Approximately 15% of the *Aspergillus fumigatus* pan-genome is made up of accessory gene content, which is predominantly clustered in the subterminal regions of chromosomes (discussed below). There is a greater degree of variation in the accessory genome sizes of individual *Aspergillus fumigatus* strains than in the other species analysed, we believe that this is primarily an artefact of the smaller number of genomes in our dataset (at the time of writing our *Aspergillus fumigatus* dataset included almost all strain genomes available as assembly data on GenBank).

Broad trends across fungal pan-genomes

Fungal core and accessory genomes enriched for potential infection and survival processes

Between 65 and 81% of gene models per species pan-genome had at least one Pfam domain, while the proportion of gene models with GO data was between 42 and 54% per species (Table 1). This variation is primarily down to a lack of human annotation for some species, and for *Cryptococcus neoformans var. grubii* in particular the lack of a dedicated GO-slim dataset. This can be seen in our statistical analyses of the distribution of GO terms in individual species pan-genomes; *Saccharomyces cerevisiae* currently has a far more detailed array of ontological terms than *Aspergillus fumigatus* for example (Table S2). In spite of gaps in ontological data for some of our species of interest, there are a number of patterns we can observe across multiple species in our GO analyses of fungal core and accessory genomes, as well as unique patterns of enrichment in some species. Many housekeeping terms such as translation, nucleic acid metabolism and oligopeptide metabolism are statistically over-represented in each fungal core genome we have analysed ($P<0.05$) (Table S2). There is an over-representation of similar cellular component terms in each of the three 'yeast' core genomes (i.e. all excluding *Aspergillus fumigatus*) (Table S2). This may reflect the morphological distinctions

between these three species and *Aspergillus fumigatus*; however, the lack of dedicated annotation data for *Cryptococcus neoformans* var. *grubii* makes a definitive observation difficult. Terms relating to transport, localization and CAZY processes are statistically over-represented in fungal accessory genomes (Table S2). In part this is to be expected, as many fungi have varying numbers of copies of genes involved in CAZY and transport processes [87]. Terms relating to housekeeping processes are statistically under-represented in fungal accessory genomes, which may be due to potential gene dosage effects. The similar patterns of statistical over-representation for terms relating to intracellular membrane-bound organelles in the accessory genomes of *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus* may reflect infection or in-host survival processes for both pathogenic species (Table S2). Both the *Candida albicans* core and accessory species genome share similarly over-represented terms to their *Saccharomyces cerevisiae* counterparts, a reflection of the two species' relatively close evolutionary relationship (Table S2).

Many of the terms that are over-represented in the *Cryptococcus neoformans* var. *grubii* core genome may reflect the species' lifestyle as an intracellular pathogen (Table S2). Such terms include regulation of homeostasis and biological quality (e.g. cell mass), which are vital for *Cryptococcus neoformans* var. *grubii* to survive the plethora of environmental stresses it encounters in the host. Similarly, UPR is an over-represented molecular function in the *Cryptococcus neoformans* var. *grubii* core genome; the UPR pathway is known to influence thermoregulation in *Cryptococcus neoformans* var. *grubii* particularly during the initial infection period [88]. Another over-represented term in the *Cryptococcus neoformans* var. *grubii* core genome is signal transduction; many signal transduction pathways in *Cryptococcus neoformans* var. *grubii* play an important role in cell differentiation as well as pathogenicity (Table S2) [89]. The core *Aspergillus fumigatus* genome is enriched for small molecule biosynthesis and other biosynthetic processes, which concurs with previous comparative studies of *Aspergillus* species [90, 91] (Table S2). This also appears to agree with our findings of BGC conservation within the *Aspergillus fumigatus* species pan-genome (Table S5). Both transport and localization processes are over-represented within the *Aspergillus fumigatus* accessory genome, which may have an indirect role in the infection processes of *Aspergillus fumigatus*. *Aspergillus fumigatus* strain pathogenesis may, therefore, be influenced by accessory genome evolution, particularly within subterminal regions [92].

The fungal core genome is more ancient in origin than the fungal accessory genome

Our statistical analysis of the ancestral history of each fungal species pan-genome found that gene models of eukaryotic origin are statistically over-represented within fungal accessory genomes, while gene models of prokaryotic origin are statistically over-represented in fungal core genomes (P<0.05) (Table S3). In other words, genes of prokaryotic origin appear to be more likely to be syntenically conserved and universally retained within these fungal species (Table S3). This appears consistent with the observation that prokaryote-derived genes in *Saccharomyces cerevisiae* are essential for survival [70]. However, it appears that the accessory genome contains more genes that arose at some point during the evolution of eukaryotes and that may be more likely to be variably retained or lost within strains of fungal species (Table S3). This would concur with our analysis of the gains and losses of syntenic orthologues in fungal accessory genomes, which are largely mediated at the strain level.

HGT may only play a limited role in fungal pan-genome evolution

Given the extent of HGT in prokaryotes and its role in generating novel genetic content and in the evolution of prokaryotic gene families, it is likely that HGT plays a significant role in prokaryote pan-genome evolution. HGT in eukaryotes is known to be far less frequent than in prokaryotes however, so its impact on eukaryotic pan-genome evolution may be limited. We examined the extent of HGT into fungal accessory genomes from two potential sources of novel genetic content: prokaryote species and other species within the fungal kingdom. A screen for interdomain HGT events in each fungal accessory genome following previous methodology [71, 93] revealed low numbers of putative HGT events from prokaryote sources into fungal accessory genomes per species (Table S3). Gene transfer between prokaryotes and eukaryotes is a subject of some controversy, with different studies suggesting that interdomain HGT is alternately non-existent or a rare but real occurrence [25, 26, 94]. Regardless, from our analysis it appears that interdomain HGT is not an influencing factor on accessory genome evolution (and hence, pan-genome evolution) within fungi. We then applied a similar screen for HGT from other fungal species into fungal accessory genomes, and found that up to 8 % of fungal accessory genomes may be derived from intrakingdom HGT. There are caveats to consider when interpreting this finding however; although some of these events may be genuine incidences of HGT, it is equally plausible that these genes have undergone pseudogenization or have otherwise lost synteny in one or more strains/lineages. That the majority of potential donor species are close relatives in each analysis we performed may in part suggest this; for example, 96 of the 102 putative HGT events into the *Saccharomyces cerevisiae* accessory genome have a potential donor from the species in the same phylum (Saccharomycotina) and 379 of the 392 putative HGT events into the *Aspergillus fumigatus* accessory genome suggest transfer from other species in the subphylum Pezizomycotina (132 from *Penicillium* species alone) (Table S3). Although there appears to be greater evidence for intrakingdom HGT having a role to play in fungal accessory genome evolution than interdomain HGT, it is our opinion that a dedicated analysis of intrakingdom HGT in fungal accessory genomes using robust phylogenetic methods is required to test the true role of intrakingdom HGT in fungal pan-genome evolution.

Eukaryotic processes such as gene duplication may influence fungal pan-genome evolution

Between 29 and 41 % of fungal accessory genomes contain gene models which appear to be duplicates of core gene models that have undergone subsequent loss in some strains, possibly by pseudogenization, microsynteny loss or expansion in other strains (Tables 2 and S1). *Cryptococcus neoformans* var. *grubii* has the smallest proportion of these duplicated core gene models (and consequently, the highest proportion of accessory gene models that have potentially arisen via other processes) and *Aspergillus fumigatus* has the largest (Table S1). This accounts for between 3 and 7 % of the total size of fungal pan-genomes, with the smallest proportion in *Candida albicans* and the largest in *Aspergillus fumigatus* (Fig. 6, Table S1). These results appear to indicate that gene duplication, which is the driving factor of gene family expansion in eukaryotes, does play an important role in the evolution of fungal accessory genomes (and pan-genomes as a whole) [95, 96]. The larger proportion of duplicated core genes in *Aspergillus fumigatus* appears to reflect the greater extent of gene duplication and paralogue diversity within that species relative to *Cryptococcus neoformans* var. *grubii* and *Saccharomyces cerevisiae* [97]. Preliminary annotation of these gene models shows that many have putative or known functions in transport and outer membrane processes, which are processes that are often mediated by expanded gene families in fungi.

Mapping the presence or absence of syntenic orthologues within fungal accessory genomes finds that for each species the majority of syntenic orthologue loss events, through chromosomal rearrangement or gene loss, or the gain of new genes, has occurred within strains as opposed to more ancestral branches (Figs 2–5). We searched each set of singleton gene models from each reference genome against the reference protein set to assess the putative function(s) of some of these strain-unique genes. Many singleton gene models are homologous to membrane proteins, DNA/RNA-binding or transposition-related genes (e.g. *gag/pol* retrotransposons in *Saccharomyces cerevisiae*, DDE1 transposases in *Aspergillus fumigatus*), which are usually independently expanded or redistributed within individual fungal genomes [83, 98]. Between 30 and 60 % of singleton gene models within each species pan-genome dataset had at least one Pfam domain, a lower proportion than that seen in each species dataset (65–81 %) as a whole, which may be another artefact of gaps in human annotation (Table S2). Closely related strains of many species also appear to have similar accessory genome sizes (e.g. many clades within the *Saccharomyces cerevisiae* 100GS dataset, the reduced sizes of both *Cryptococcus neoformans* var. *grubii* C45 and MW-RSA852 relative to most other strains) (Figs 2–4). There is greater variation in the sizes of strain accessory genomes in *Aspergillus fumigatus*; however, this may be an artefact of taxon sampling (Fig. 5). *Saccharomyces cerevisiae* S288C itself had 31 singleton gene models not found in any other *Saccharomyces cerevisiae* strain. By comparison, the 100GS authors located 108 genes present in $\geq$1 strain but not in S288C and 28 genes unique to S288C [50]. In total, these analyses suggest that fungal pan-genomes evolve by innovations originating within fungi on the strain level, such as gene duplication or rearrangement, as opposed to being influenced by factors such as HGT from prokaryotic sources or larger species-level events.

Subterminal regions of fungal genomes may be harbours of accessory genome content

Analysis of the global distribution of core and accessory gene models shows that there is a statistically significant bias towards accessory gene models in the subterminal regions within three of the four reference genomes in our study and a statistically significant bias towards core gene models outside these subterminal regions in the same genomes (P<0.05) (Fig. S7a, c, d, Table S4). The sole exception is *Candida albicans* SC5314, wherein there is a statistically significant bias for core gene models within subterminal regions (P<0.05) (Fig. S7b, Table S4). The subterminal regions of chromosomes are usually areas of genomic instability in eukaryotes, so it is unsurprising that we observe greater breakdown of synteny in these regions [99]. Terminal and subterminal regions of chromosomes (i.e. telomeres and subtelomeric regions) are also known hotspots of recombination in fungi, which can lead to the evolution of novel genetic content, and in some fungi such recombinatory hotspots are potentially enriched for secreted proteins [100]. All fungal reference genomes possess at least one chromosome that is enriched for accessory gene models; these chromosomes may have undergone recombination or translocation events that lead to the breakdown of synteny or the eventual evolution of novel genes (Table S4). Such translocation events are known to have occurred within some strains of *Saccharomyces cerevisiae* and *Aspergillus fumigatus* in particular [86, 99, 101, 102]. In some reference genomes such as *Aspergillus fumigatus* Af293 large clusters of accessory genome content can be observed outside the subterminal regions, which may reflect instances of strain- or lineage-specific genomic rearrangement events (Fig. S7). Such rearrangements are linked to environmental adaptation and reproductive isolation in *Saccharomyces cerevisiae* genomes [103]. In *Cryptococcus neoformans* var. *grubii*, the greater degree of accessory genome content found outside subterminal regions may be a reflection of the role that genomic rearrangement plays in shaping the genomes of individual strains within the host [86].

Fungal core and accessory genomes encompass various biological pathways and phenotypes

Due to its position as arguably the most complete fungal model organism, there is a wealth of manually annotated functional data available for *Saccharomyces cerevisiae* that is lacking for other species. One such collection is the

systematic mutation set available from the SGD, which includes amongst other datasets a systematically constructed genome-wide set of deletion phenotypes for many different strains of *Saccharomyces cerevisiae* [17, 73]. Using reciprocal BLASTP searches against the reference protein set as well as data from the systematic mutation set, we inferred the knockout viability of the core and accessory genomes of *Saccharomyces cerevisiae* S288C. We found that the core *Saccharomyces cerevisiae* S288C genome is not significantly over-represented for either knockout-viable or knockout-inviable genes (Table S5). This may reflect the fact less than 20 % of the genes encoded in the *Saccharomyces cerevisiae* S288C genome are thought to be essential for growth and, thus, likely knockout-inviable [104]. It is worth observing, however, that 962 of the 1031 predicted gene models with an inviable knockout phenotype are within the core *Saccharomyces cerevisiae* genome (Table S5). In contrast, there is a significant proportion of gene models within the *Saccharomyces cerevisiae* S288C accessory genome that are associated with a viable knockout phenotype ($P<0.05$), which appears to reinforce the more variable nature of species accessory genomes relative to core genomes (Table S5).

Unlike filamentous fungi such as *Aspergillus* species, many yeasts lack BGCs. Somewhat analogous to BGCs in *Saccharomyces cerevisiae* are small DP gene clusters of functionally related genes, which have been lost in other *Saccharomyces* and related species but were later regained in *Saccharomyces cerevisiae* via HGT or neofunctionalization [74]. Hall and Dietrich [74] previously described 14 such clusters, encompassing 38 reference and another 3 non-reference genes, which are involved in many different metabolic processes [74]. Our analysis of the distribution of 38 reference DP genes within the *Saccharomyces cerevisiae* pan-genome found one DP cluster that appears to be completely conserved in the pan-genome; a cluster on chromosome II containing three *GAL* genes that mediates the degradation of galactose to galactose-1-phosphate within the glycolysis pathway [105] (Table S5). Other clusters were highly conserved across almost all strains but not universally conserved in our dataset, i.e. a small number of strains. Such highly conserved clusters include two clusters involved in the metabolism of B vitamins; a three gene *BIO* biotin uptake cluster on chromosome XIV and a *SNO1-SNZ1* vitamin B6 metabolism cluster on chromosome XIII (Table S5) [74]. Another highly conserved six gene *DAL-DCG* cluster found on chromosome IX, the largest DP cluster, allows *Saccharomyces cerevisiae* to use allantoin as its sole nitrogen source through a pathway in which allantoin is converted to urea, which is then converted into ammonium by *DUR1-2* [106]. A *SAM4-SAM3* cluster that enables the usage of *S*-adenosylmethionine as a sulphur source has one of the two member genes missing in four strains (and is entirely absent in YJM969) (Table S5).

It is possible that some strains may simply be missing a syntenic orthologue of one or more genes in a cluster due to pseudogenization or synteny loss due to chromosomal rearrangement. Other DP clusters have more patchy distribution within the *Saccharomyces cerevisiae* species pan-genome, particularly those within subterminal regions in *Saccharomyces cerevisiae* S288C, which may indicate a greater breakdown of synteny or gene loss within these clusters. For some clusters, this may be due to functional redundancy; for example, three DP clusters are involved in vitamin B1 and B6 metabolism, the aforementioned *SNO1-SNZ1* cluster is conserved across almost all 100 strains whereas the other two clusters have patchier distribution or are totally missing in some strains (e.g. in the Indonesian strain YJM1244, two clusters are completely conserved but the other is absent) (Table S5). Other potential causes for this varying distribution of DP clusters may include environmental adaptations. One DP cluster that confers arsenic resistance is prevalent in many wine/European strains, but has much patchier conservation in non-European strains or strains with Malaysian or West African ancestry (such as SK1). One member gene of this cluster, *ARR3*, is absent in 49 out of the 100 strains in our dataset, including many mosaic strains with wine/European and Malaysian ancestry. Increased arsenic resistance has been observed in strains of European ancestry, likely as a result of anthropogenic influence on soil composition, which may explain the *ARR* cluster's absence in some non-European strains [34, 76]. Additionally, the *ARR* cluster is located in the subterminal regions of chromosome XVI in *Saccharomyces cerevisiae* S228C; this suggests gene loss or chromosomal rearrangements amongst other events may be responsible for the absence of *ARR3* in the *ARR* cluster of many strains [34, 107].

Within the aspergilli and other fungi, functionally related genes involved in secondary metabolism pathways are often arranged into BGCs within the subterminal regions of chromosomes. These BGCs are involved in a range of infection and survival processes in the aspergilli, and subterminal regions themselves are believed to mediate the infection process of *Aspergillus fumigatus* in the human host [91, 92, 99, 108]. Our analysis of known BGCs in the *Aspergillus fumigatus* pan-genome found 14 BGCs that were completely conserved, a number of which are involved in the production of mycotoxins. Other BGCs have one or two syntenic orthologues that are missing in other strains, in these cases the majority of these genes may play more indirect roles in cluster function and, therefore, be less likely to be conserved within clusters, while some are only partially present or completely absent in some strains but are highly conserved in others (Table S5). An analysis of variation of *Aspergillus fumigatus* BGCs using short-read data by Lind *et al.* [109] found similar patterns of BGC variation to our gene-level functional analysis [109]. Lind *et al.* [109] observed some trends that explain the variation in BGCs within *Aspergillus fumigatus* in both their analysis and ours; for example, a fusarielin-like cluster we identified as missing from A1163 and partially present in other strains has gained pseudogenizing mutations in some strains but not others, whereas variation in other accessory BGCs is due to factors such as transposable elements (as is the case in a 27 member PKS cluster) or lineage-specific gene acquisition/loss events [109]. This suggests that some BGCs are invariably conserved due to the importance of their function (such as gliotoxins), while others may be lost in particular strains due to environmental adaptations or other factors.

Other remarks

Compared to the increasing amount of software designed to construct and characterize bacterial and archaeal pan-genomes, little dedicated pan-genome software exists for eukaryote taxa. Our overall method of analysis, bespoke gene model prediction followed by pan-genome construction using PanOCT as the anchor method, is ad hoc but may point towards a sufficiently optimized syntenic method of pan-genome construction for eukaryotes in the future. On this point, it is worth noting that PanOCT's current implementation has an exponential memory usage curve per genome added, which makes analysis of prokaryotic or eukaryotic datasets of this scale difficult without dedicated high-performance computational facilities [38]. The relative lack of GO information for some fungal species (e.g. *Cryptococcus neoformans* var. *grubii*, which currently lacks a dedicated GO-slim dataset) may have affected our functional characterization of fungal pan-genomes. We attempted to ameliorate this lack of data by using other sources of genomic information (e.g. knockout data from SGD for *Saccharomyces cerevisiae*), though their efficacy is ultimately dependent on human annotation. One caveat of large-scale pan-genome analysis of this kind may be the usage of genomes assembled via a reference-based approach as opposed to *de novo* approaches, which may then lead to an underestimation of accessory genome sizes within species pan-genomes due to underestimation of sequence diversity or inheritance of assembly artefacts from the reference genome [110]. The majority of genomes used for each species dataset were assembled using *de novo* approaches, for example, the 100GS dataset is predominantly *de novo* sequenced strains, so the potential effects of overreliance on reference-based assembly data may have been reduced in our study [50].

The size of a species pan-genome and its complements are ultimately dependent on the amount and the geographical or phenotypical variety of genomic data sampled. Methodological differences notwithstanding, our 100 strain analysis of the *Saccharomyces cerevisiae* pan-genome and the 1011 strain analysis by Peter *et al.* [36] predict similar-sized pan-genomes [36]. In contrast, our reconstruction of the *Candida albicans* pan-genome likely underestimates the true size of the *Candida albicans* accessory genome due to a lack of non-clinical genomic data. The greater variation of accessory genome sizes between individual strains of *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus* may be an artefact of there being fewer strain genomes available for both species, which would in turn affect the sizes of those species' pan-genomes. There have been attempts to estimate the 'true' size of bacterial pan-genomes from existing data using different mathematical models, which vary from inferring almost infinite pan-genomes that increase in size with each strain added to stricter models that infer a more finite structure for most bacterial species [5, 6, 111]. Future analysis of fungal species pan-genomes should attempt to quantify their true size using similar methods.

Conclusions

Evidence for the existence of pan-genomic structure has been demonstrated in eukaryotic taxa using a variety of methodologies. Using computational methods based on sequence similarity and conserved synteny between strains, we have constructed and characterized species pan-genomes for four model fungi: *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans* var. *grubii* and *Aspergillus fumigatus*. Defining core genomes as containing gene models syntenically conserved throughout species and accessory genomes as containing gene models of varying syntenic conservation and distribution throughout species, we find strong evidence for pan-genomic structure within fungi. Between 80 and 90 % of all potential gene models in fungal species are core gene models, with the remainder being accessory gene models that are strain-specific or specific to individual groups of strains. Fungal core genomes are enriched for genes of ancient origin and facilitate many essential metabolic, regulatory and survival processes in both commensal and pathogenic species. Fungal accessory genomes are enriched for genes of more recent origin, appear to evolve and vary in size by processes like gene duplication and gain/loss events within strains, and are enriched for genes involved in molecule transport and carbohydrate metabolism amongst other functions. Our analysis supports the growing amount of evidence for pan-genomic structure in eukaryotes.

Funding information
C.G.P.M. is funded by an Irish Research Council Government of Ireland Postgraduate Scholarship (grant no. GOIPG/2015/2242).

Acknowledgements
The authors would like to acknowledge the original contributors to all sequencing data used in this analysis for making their data publicly available. The authors would also like to acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data bibliography
1. McCarthy CGP, Fitzpatrick DA. Pipelines for eukaryotic pangenome analysis. Github: https://github.com/chmccarthy/pangenome-pipelines (2018).

References
1. **Parfrey LW, Lahr DJG, Katz LA.** The dynamic nature of eukaryotic genomes. *Mol Biol Evol* 2008;25:787–794.
2. **Rouli L, Merhej V, Fournier PE, Raoult D.** The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* 2015;7:72–85.
3. **Diene SM, Merhej V, Henry M, El Filali A, Roux V *et al.*** The rhizome of the multidrug-resistant *Enterobacter aerogenes* genome

reveals how new "killer bugs" are created because of a sympatric lifestyle. *Mol Biol Evol* 2013;30:369–383.
4. **Lefébure T, Pavinski Bitar PD, Suzuki H, Stanhope MJ.** Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biol Evol* 2010;2:646–655.
5. **Snipen L, Almøy T, Ussery DW.** Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* 2009; 10:385.
6. **Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D *et al.*** Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci USA* 2005;102:13950–13955.
7. **Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J *et al.*** Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 2007;8: R103.
8. **Young JPW, Crossman LC, Johnston AWB, Thomson NR, Ghazoui ZF *et al.*** The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol* 2006;7:R34.
9. **Boissy R, Ahmed A, Janto B, Earl J, Hall BG *et al.*** Comparative supragenomic analyses among the pathogens *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. *BMC Genomics* 2011;12:187.
10. **Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF *et al.*** The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 2008;190:6881–6893.
11. **Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ.** Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc Natl Acad Sci USA* 2009;106:8605–8610.
12. **Marcus S, Lee H, Schatz MC.** SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* 2014; 30:3476–3483.
13. **Song G, Dickins BJA, Demeter J, Engel S, Dunn B *et al.*** AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of *Saccharomyces cerevisiae*. *PLoS One* 2015;10:e0120671.
14. **Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.*** Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
15. **Jandrasits C, Dabrowski PW, Fuchs S, Renard BY.** seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. *BMC Genomics* 2018;19:47.
16. **Chaudhari NM, Gupta VK, Dutta C.** BPGA - an ultra-fast pan-genome analysis pipeline. *Sci Rep* 2016;6:24373.
17. **Engel SR, Cherry JM.** The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database. *Database* 2013;2013:bat012.
18. **Wei W, McCusker JH, Hyman RW, Jones T, Ning Y *et al.*** Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci USA* 2007;104:12825–12830.
19. **Gu Z, David L, Petrov D, Jones T, Davis RW *et al.*** Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 2005;102:1092–1097.
20. **Ronald J, Tang H, Brem RB.** Genomewide evolutionary rates in laboratory and wild yeast. *Genetics* 2006;174:541–544.
21. **Golicz AA, Batley J, Edwards D.** Towards plant pangenomics. *Plant Biotechnol J* 2016;14:1099–1105.
22. **Lapierre P, Gogarten JP.** Estimating the size of the bacterial pan-genome. *Trends in Genetics* 2009;25:107–110.
23. **McInerney JO, McNally A, O'Connell MJ.** Why prokaryotes have pangenomes. *Nat Microbiol* 2017;2:17040.
24. **Keeling PJ, Palmer JD.** Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* 2008;9:605–618.
25. **Ku C, Martin WF.** A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. *BMC Biol* 2016;14:89.
26. **Martin WF.** Too much eukaryote LGT. *Bioessays* 2017;39: 1700115.
27. **Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H *et al.*** The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun* 2016;7:13390.
28. **Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H *et al.*** The pangenome of hexaploid bread wheat. *Plant J* 2017;90: 1007–1013.
29. **Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G *et al.*** Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 2014;26:121–135.
30. **Zhou P, Silverstein KAT, Ramaraj T, Guhlin J, Denny R *et al.*** Exploring structural variation and gene family architecture with de novo assemblies of 15 Medicago genomes. *BMC Genomics* 2017;18:261.
31. **Gordon SP, Contreras-Moreira B, Woods DP, des Marais DL, Burgess D *et al.*** Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat Commun* 2017;8:2184.
32. **Li YH, Zhou G, Ma J, Jiang W, Jin LG *et al.*** De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* 2014;32:1045–1052.
33. **Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC *et al.*** Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* 2013;499:209–213.
34. **Bergström A, Simpson JT, Salinas F, Barré B, Parts L *et al.*** A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol* 2014;31:872–888.
35. **Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G.** Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res* 2012;22:908–924.
36. **Peter J, de Chiara M, Friedrich A, Yue JX, Pflieger D *et al.*** Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 2018;556:339–344.
37. **Plissonneau C, Hartmann FE, Croll D.** Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol* 2018;16:5.
38. **Fouts DE, Brinkac L, Beck E, Inman J, Sutton G.** PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res* 2012;40:e172–e172.
39. **Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ *et al.*** Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25:1422–1423.
40. **Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B *et al.*** Life with 6000 genes. *Science* 1996;274:546–567.
41. **Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S *et al.*** The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci USA* 2004;101:7329–7334.
42. **Janbon G, Ormerod KL, Paulet D, Byrnes EJ, Yadav V *et al.*** Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. *grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genet* 2014;10: e1004261.
43. **Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS *et al.*** Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* 2005;438:1151–1156.
44. **Slater G, Birney E.** Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 2005;6:31.

45. **Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M.** Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 2008;18:1979–1990.
46. **Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD *et al.*** De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;8:1494–1512.
47. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al.*** BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
48. **Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M *et al.*** The *Candida* genome database (CGD): incorporation of assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res* 2017;45:D592–D596.
49. **Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM.** BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–3212.
50. **Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA *et al.*** The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res* 2015;25:762–774.
51. **Novo M, Bigey F, Beyne E, Galeote V, Gavory F *et al.*** Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci USA* 2009;106:16333–16338.
52. **Rasko DA, Myers GSA, Ravel J.** Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* 2005;6:2.
53. **Deniélou YP, Sagot MF, Boyer F, Viari A.** Bacterial syntenies: an exact approach with gene quorum. *BMC Bioinformatics* 2011;12: 193.
54. **Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H.** UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 2014;20:1983–1992.
55. **Conway JR, Lex A, Gehlenborg N.** UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;33:2938–2940.
56. **R Core Team.** *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013.
57. **Edgar RC.** MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5: 113.
58. **Swofford LD.** *PAUP*:Phylogenetic Analysis Using Parsimony (* and Other Methods)*. Sunderland, MA, USA: Sinauer Associates; 2002.
59. **Kück P, Meusemann K.** FASconCAT: convenient handling of data matrices. *Mol Phylogenet Evol* 2010;56:1115–1118.
60. **Price MN, Dehal PS, Arkin AP.** FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5: e9490.
61. **Letunic I, Bork P.** Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;44:W242–W245.
62. **Farris JS.** Phylogenetic analysis under Dollo's law. *Syst Biol* 1977;26:77–88.
63. **Csűös M.** Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 2010;26:1910–1912.
64. **Jones P, Binns D, Chang HY, Fraser M, Li W *et al.*** InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236–1240.
65. **Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J *et al.*** The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;44:D279–D285.
66. **Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK *et al.*** InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 2012;40:D306–D312.
67. **Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC *et al.*** Expansion of the gene ontology knowledgebase and resources:the gene ontology consortium. *Nucleic Acids Res* 2017;45:D331–D338.
68. **Agresti A.** *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc; 2002.
69. **Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A *et al.*** GOATOOLS: a Python library for Gene Ontology analyses. *Sci Rep* 2018;8:10872.
70. **Cotton JA, McInerney JO.** Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc Natl Acad Sci USA* 2010; 107:17252–17255.
71. **McCarthy CGP, Fitzpatrick DA.** Systematic search for evidence of interdomain horizontal gene transfer from prokaryotes to oomycete lineages. *mSphere* 2016;1:e00195-16.
72. **Wolfe D, Dudek S, Ritchie MD, Pendergrass SA.** Visualizing genomic information across chromosomes with PhenoGram. *BioData Min* 2013;6:18.
73. **Giaever G, Nislow C.** The yeast deletion collection: a decade of functional genomics. *Genetics* 2014;197:451–465.
74. **Hall C, Dietrich FS.** The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* 2007;177:2293–2307.
75. **Lind AL, Lim FY, Soukup AA, Keller NP, Rokas A.** An LaeA- and BrlA-dependent cellular network governs tissue-specific secondary metabolism in the human pathogen *Aspergillus fumigatus*. *mSphere* 2018;3:e00050–18.
76. **Warringer J, Zörgö E, Cubillos FA, Zia A, Gjuvsland A *et al.*** Trait variation in yeast is defined by population history. *PLoS Genet* 2011;7:e1002111.
77. **Knox BP, Blachowicz A, Palmer JM, Romsdahl J, Huttenlocher A *et al.*** Characterization of *Aspergillus fumigatus* isolates from air and surfaces of the International Space Station. *mSphere* 2016;1:e00227-16.
78. **McCarthy CGP, Fitzpatrick DA.** Multiple approaches to phylogenomic reconstruction of the fungal kingdom. In: *Advances in Genetics*; 2017. pp. 211–266.
79. **Wolfe KH.** Origin of the yeast whole-genome duplication. *PLoS Biol* 2015;13:e1002221.
80. **Santos MAS, Tuite MF.** The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans*. *Nucleic Acids Res* 1995; 23:1481–1486.
81. **Stajich JE, Dietrich FS, Roy SW.** Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol* 2007;8:R223.
82. **Gonzalez-Hilarion S, Paulet D, Lee KT, Hon CC, Lechat P *et al.*** Intron retention-dependent gene regulation in *Cryptococcus neoformans*. *Sci Rep* 2016;6:32252.
83. **Liti G, Carter DM, Moses AM, Warringer J, Parts L *et al.*** Population genomics of domestic and wild yeasts. *Nature* 2009;458: 337–341.
84. **Yue JX, Li J, Aigrain L, Hallin J, Persson K *et al.*** Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* 2017;49:913–924.
85. **Lockhart SR, Pujol C, Daniels KJ, Miller MG, Johnson AD *et al.*** In *Candida albicans*, white-opaque switchers are homozygous for mating type. *Genetics* 2002;162:737–745.

86. **Fraser JA, Huang JC, Pukkila-Worley R, Alspaugh JA, Mitchell TG *et al.*** Chromosomal translocation and segmental duplication in *Cryptococcus neoformans*. *Eukaryot Cell* 2005;4:401–406.
87. **Wisecaver JH, Slot JC, Rokas A.** The evolution of fungal metabolic pathways. *PLoS Genet* 2014;10:e1004816.
88. **Cheon SA, Jung KW, Bahn YS, Kang HA.** The unfolded protein response (UPR) pathway in *Cryptococcus*. *Virulence* 2014;5:341–350.
89. **Lengeler KB, Davidson RC, D'Souza C, Harashima T, Shen WC *et al.*** Signal transduction cascades regulating fungal development and virulence. *Microbiol Mol Biol Rev* 2000;64:746–785.
90. **Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC *et al.*** SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol* 2010;47:736–741.
91. **Andersen MR, Nielsen JB, Klitgaard A, Petersen LM, Zachariasen M *et al.*** Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc Natl Acad Sci USA* 2013;110:E99–E107.
92. **McDonagh A, Fedorova ND, Crabtree J, Yu Y, Kim S *et al.*** Subtelomere directed gene expression during initiation of invasive aspergillosis. *PLoS Pathog* 2008;4:e1000154.
93. **Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G *et al.*** Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc Natl Acad Sci USA* 2011;108:15258–15263.
94. **Marcet-Houben M, Gabaldón T.** Acquisition of prokaryotic genes by fungal genomes. *Trends Genet* 2010;26:5–8.
95. **Lynch M, Conery JS.** The evolutionary fate and consequences of duplicate genes. *Science* 2000;290:1151–1155.
96. **Treangen TJ, Rocha EP.** Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 2011;7:e1001284.
97. **Yang E, Hulse AM, Cai JJ.** Evolutionary analysis of sequence divergence and diversity of duplicate genes in *Aspergillus fumigatus*. *Evol Bioinform Online* 2012;8:623–644.
98. **Perez-Nadales E, Nogueira MF, Baldin C, Castanheira S, El Ghalid M *et al.*** Fungal model systems and the elucidation of pathogenicity determinants. *Fungal Genet Biol* 2014;70:42–67.
99. **Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P *et al.*** Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet* 2008;4:e1000046.
100. **Croll D, Lendenmann MH, Stewart E, McDonald BA.** The impact of recombination hotspots on genome evolution of a fungal plant pathogen. *Genetics* 2015;201:1213–1228.
101. **Schmidt KH, Viebranz E, Doerfler L, Lester C, Rubenstein A.** Formation of complex and unstable chromosomal translocations in yeast. *PLoS One* 2010;5:e12007.
102. **Colson I, Delneri D, Oliver SG.** Effects of reciprocal chromosomal translocations on the fitness of *Saccharomyces cerevisiae*. *EMBO Rep* 2004;5:392–398.
103. **Hou J, Friedrich A, de Montigny J, Schacherer J.** Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in *Saccharomyces cerevisiae*. *Curr Biol* 2014;24:1153–1159.
104. **Giaever G, Chu AM, Ni L, Connelly C, Riles L *et al.*** Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002;418:387–391.
105. **Slot JC, Rokas A.** Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc Natl Acad Sci USA* 2010;107:10136–10141.
106. **Naseeb S, Delneri D.** Impact of chromosomal inversions on the yeast DAL cluster. *PLoS One* 2012;7:e42022.
107. **Maciaszczyk E, Wysocki R, Golik P, Lazowska J, Ulaszewski S.** Arsenical resistance genes in *Saccharomyces douglasii* and other yeast species undergo rapid evolution involving genomic rearrangements and duplications. *FEMS Yeast Res* 2004;4:821–832.
108. **Keller NP, Turner G, Bennett JW.** Fungal secondary metabolism — from biochemistry to genomics. *Nat Rev Microbiol* 2005;3:937–947.
109. **Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM *et al.*** Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol* 2017;15:e2003583.
110. **Ekblom R, Wolf JB.** A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl* 2014;7:1026–1042.
111. **Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J *et al.*** Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 2007;8:R103.

genes

MDPI

Article

Pangloss: A Tool for Pan-Genome Analysis of Microbial Eukaryotes

Charley G. P. McCarthy [1,2,*] and David A. Fitzpatrick [1,2]

[1] Genome Evolution Laboratory, Department of Biology, Maynooth University, W23 F2K8 Maynooth, Ireland
[2] Human Health Research Institute, Maynooth University, W23 F2K8 Maynooth, Ireland
* Correspondence: Charley.McCarthy@mu.ie

Received: 7 June 2019; Accepted: 8 July 2019; Published: 10 July 2019

check for updates

Abstract: Although the pan-genome concept originated in prokaryote genomics, an increasing number of eukaryote species pan-genomes have also been analysed. However, there is a relative lack of software intended for eukaryote pan-genome analysis compared to that available for prokaryotes. In a previous study, we analysed the pan-genomes of four model fungi with a computational pipeline that constructed pan-genomes using the synteny-dependent Pan-genome Ortholog Clustering Tool (PanOCT) approach. Here, we present a modified and improved version of that pipeline which we have called Pangloss. Pangloss can perform gene prediction for a set of genomes from a given species that the user provides, constructs and optionally refines a species pan-genome from that set using PanOCT, and can perform various functional characterisation and visualisation analyses of species pan-genome data. To demonstrate Pangloss's capabilities, we constructed and analysed a species pan-genome for the oleaginous yeast *Yarrowia lipolytica* and also reconstructed a previously-published species pan-genome for the opportunistic respiratory pathogen *Aspergillus fumigatus*. Pangloss is implemented in Python, Perl and R and is freely available under an open source GPLv3 licence via GitHub.

Keywords: pangenomes; bioinformatics; microbial eukaryotes; fungi

1. Introduction

Species pan-genomes have been extensively studied in prokaryotes, where pan-genome evolution is primarily driven by rampant horizontal gene transfer (HGT) [1–4]. Pan-genome evolution in prokaryotes can also vary substantially as a result of lifestyle and environmental factors; opportunistic pathogens such as *Pseudomonas aeruginosa* have large "open" pan-genomes with large proportions of accessory genes, whereas obligate intracellular parasites such as *Chlamydia* species have smaller "closed" pan-genomes with larger proportions of conserved core genes and a smaller pool of novel genetic content [5–7]. Studies of pan-genome evolution within eukaryotes has not been as extensive as that of prokaryotes to date, as eukaryote genomes are generally more difficult to sequence and assemble in large numbers relative to prokaryote genomes. However, consistent evidence for pan-genomic structure within eukaryotes has been demonstrated in plants, fungi and plankton [8–12]. Unlike prokaryote pan-genomes, eukaryote pan-genomes evolve via a variety of processes besides HGT, these include variations in ploidy and heterozygosity within plants [8], and cases of introgression, gene duplication and repeat-induced point mutation in fungi and plankton [9–12].

The majority of software and pipelines available for pan-genome analysis are explicitly or implicitly intended for prokaryote datasets. For example, the commonly-cited pipeline Roary is intended for use with genomic location data generated by the prokaryote genome annotation software Prokka [13,14]. A number of other methodologies such as seq-seq-pan or SplitMEM use genome alignment or de Bruijn graph-based approaches for pan-genome construction, which are

usually computationally impracticable for eukaryote analysis [15,16]. Other common pan-genome methodologies, such as the Large Scale BLAST Score Ratio (LS-BSR) approach or the Markov Cluster Algorithm (MCL)/MultiParanoid-dependent Pan-genome Analysis Pipeline (PGAP), may have potential application in eukaryote pan-genome analysis but as of writing no such application has occurred [17–20]. Of the eukaryote pan-genome analyses in the literature, some construct pan-genomes by mapping and aligning sequence reads using pipelines such as the Eukaryotic Pan-genome Analysis Toolkit (EUPAN) [8,12,21], or have constructed and characterised eukaryote pan-genomes using bespoke BLAST-dependent or clustering algorithm-dependent sequence clustering approaches [9,10,12]. In a previous article, we constructed and analysed the species pan-genomes of four model fungi including *Saccharomyces cerevisiae*, using the synteny-based Pan-genome Ortholog Clustering Tool (PanOCT, https://sourceforge.net/projects/panoct/) method in addition to our own prediction and analysis pipelines [11,22]. PanOCT was initially developed for prokaryote pan-genome analysis, and constructs a pan-genome from a given dataset by clustering homologous sequences from different input genomes together into clusters of syntenic orthologs based on a measurement of local syntenic conservation between these sequences, referred to as a conserved gene neighbourhood (CGN) score, and BLAST score ratio (BSR) assessment of sequence similarity [22,23]. Crucially, this synteny-based approach allows PanOCT to distinguish between paralogous sequences within the same genome when assessing orthologous sequences between genomes [11].

Here, we present a refined and improved version of our PanOCT-based pan-genome analysis pipeline which we have called Pangloss. Pangloss incorporates reference-based and ab initio gene model prediction methods, and synteny-based pan-genome construction using PanOCT with an optional refinement based on reciprocal sequence similarity between clusters of syntenic orthologs. Pangloss can also perform a number of downstream characterisation analyses of eukaryote pan-genomes, including Gene Ontology (GO-slim) term enrichment in core and accessory genomes, selection analyses in core and accessory genomes and visualisation of pan-genomic data. To demonstrate the pipeline's capabilities we have constructed and analysed a species pan-genome for the oleaginous yeast *Yarrowia lipolytica* using Pangloss [24]. *Y. lipolytica* is one of the earliest-diverging yeasts and has seen various applications as a non-conventional yeast model for protein secretion, regulation of dimorphism and lipid accumulation, and is a potential alternative source for biofuels and other oleochemicals [25–31]. We have also reconstructed the species pan-genome of the opportunistic respiratory pathogen *Aspergillus fumigatus* from a previous study as a control [11]. Pangloss is implemented in Python, Perl and R, and is freely available under an open source GPLv3 licence from http://github.com/chmccarthy/Pangloss.

2. Materials and Methods

2.1. Implementation

Pangloss is predominantly written in Python with some R and Perl components, and is compatible with macOS and Linux operating systems. Pangloss performs a series of gene prediction, gene annotation and functional analyses to characterise the pan-genomes of microbial eukaryotes. These analyses can be enabled by the user by invoking their corresponding flags on the command line, and many of the parameters of these analyses are controlled by Pangloss using a configuration file. The various dependencies for eukaryote pan-genome analysis using Pangloss are given in Table 1 along with versions tested and the workflow of Pangloss is given in Figure 1, both are described in greater detail below [32–45]. A user manual as well as further installation instructions and download locations for all dependencies of Pangloss are available from http://github.com/chmccarthy/Pangloss/.

Table 1. List of various dependencies for Pangloss, versions tested in parentheses. PanOCT included with Pangloss. See http://github.com/chmccarthy/Pangloss/ for download location and detailed installation instructions for each dependency.

Dependencies	Function
Python (2.7.10) *. BioPython (1.7.3) [32]	Base environment for Pangloss.
Exonerate (2.4) [33], GeneMark-ES (4.3.8) [38], TransDecoder (5.5) [39]	Gene model prediction.
BLAST+ (2.9.0) [40]	All-vs.-all sequence similarity search, dubious gene similarity search.
BUSCO (3.1) [41]	Gene model set completeness analysis.
PanOCT (3.2) [22]	Pan-genome construction.
MUSCLE (3.8.31) [42], PAML (4.8) [43]	Selection analysis of core/accessory cluster alignment using yn00.
InterProScan (5.34)† [44], GOATools (0.8.12) [45]	Functional classification and functional enrichment analysis of pan-genome.
R (3.6), ggplot (3.2) [34], ggrepel (0.8.1), UpSetR (1.4) [35], Bioconductor (3.9) [36], KaryoploteR (1.10.3) [37]	Visualisation of pan-genome size and distributions across genomes.

* Required for all analyses. † InterProScan is only available for Linux distributions.

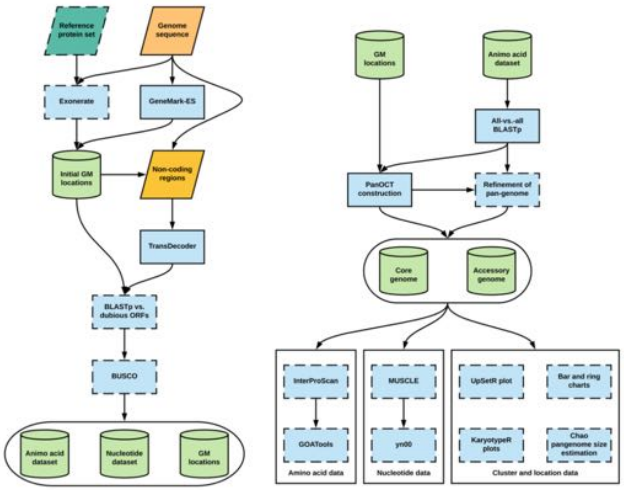


Figure 1. Workflow of Pangloss. Optional analyses represented with dotted borders. Refer to implementation for further information. GM: Gene model.

2.1.1. Gene Model Prediction and Annotation

By default, Pangloss performs its own gene model prediction to generate nucleotide and protein sequence data for all gene models from each genome in a dataset (Figure 1). Pangloss also generates a set of PanOCT-compatible gene model location data for each genome. Gene model prediction can be skipped by including the argument --no_pred at the command-line if such data has already been generated, or the user can solely run gene model prediction with no downstream analysis by including the argument --pred_only at the command-line. For each genome in a dataset, up to three methods of prediction are used:

1. All predicted protein sequences from a user-provided reference genome are queried against each genome using Exonerate (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate), with a heuristic protein2genome search model [33]. Translated gene model top-hits with an alignment score of ≥90% of the maximum possible alignment score as determined by Exonerate are retained as potential gene models. This search step is parallelized through Python's multiprocessing library and can be optionally disabled by the user by including the argument --no_exonerate at the command-line.
2. Ab initio hidden Markov model (HMM)-dependent gene model prediction is performed using GeneMark-ES (http://exon.gatech.edu/GeneMark/) with self-training enabled [38]. If the species of interest is fungal, the user can enable a fungal-specific branch point site prediction model in the configuration file. If the user has also predicted gene models via step 1, those gene models whose locations do not overlap with gene models predicted via GeneMark-ES are incorporated into the latter dataset.
3. All remaining non-coding regions of the genome are extracted and subjected to position weight matrix (PWM)-dependent gene model prediction using TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki) [39]. Any remaining predicted gene models with a length of ≥200 amino acids are included in the final gene model dataset.

There are a number of optional steps after that the user can take to assess the quality of gene model prediction within a dataset (Figure 1). The user can filter gene model sets for potential pseudogenes by querying a set of known dubious genes (either user-curated or from an appropriate resource such as the *Saccharomyces* Genome Database) against each gene model set using BLASTp (enabled via the --qc command-line argument) [46,47]. Any gene models whose top BLASTp hit against a dubious gene has sequence coverage of ≥70% are removed from further analysis. The completeness of each gene model set can also be assessed using BUSCO (https://gitlab.com/ezlab/busco) (enabled via the --busco command-line argument), with the appropriate dataset assigned by the user [41].

2.1.2. BLASTp and PanOCT Analysis

By default, all predicted gene models within a dataset are combined and an all-vs.-all BLASTp search is performed within Pangloss with a user-defined e-value cut-off (default = 10^{-4}) (Figure 1). However, if the user prefers to perform the all-vs.-all BLASTp step on their own high-performance computational environment they can skip the search via the --no_blast command-line argument. The BLASTp search data, along with all gene models and gene model location datasets combined, are used as input for PanOCT. For a pan-genome dataset of syntenic ortholog clusters as constructed by Pangloss, clusters that contain an ortholog from all input genomes are classified as "core" clusters (containing "core" gene models) and clusters missing an ortholog from ≥1 input genomes are classified as "accessory" clusters (containing "accessory" gene models) [11]. Pangloss also generates nucleotide and amino acid datasets for every core and accessory cluster for further downstream analyses.

2.1.3. Refinement of Pan-Genome Construction Based on Reciprocal Sequence Similarity

After construction of the initial pan-genome, the user has the option of refining the pan-genome with Pangloss via the --refine command-line argument (Figure 1). This method attempts to refine the PanOCT-derived microsyntenic pan-genome by accounting for microsynteny loss due to genome assembly artefacts or genomic rearrangements. In this method, Pangloss first extracts all accessory clusters from the accessory genome and parses the previously-generated all-vs.-all BLASTp data used for PanOCT. For each accessory cluster $\boldsymbol{A}$, Pangloss extracts the BLASTp data for each ortholog in $\boldsymbol{A}$

and generates a list of BLASTp top-hits to each strain genome not represented in ***A*** with ≥30% sequence identity. If this list matches another accessory cluster ***B*** in the accessory genome, Pangloss will then check if each ortholog in ***B*** has a reciprocal strain top-hit to each ortholog in ***A***. If ***A*** and ***B*** satisfy this criterion they are merged into a new cluster ***AB***, and ***A*** and ***B*** themselves are subsequently removed from the accessory genome. If this new cluster ***AB*** has an ortholog from every input strain genome in the dataset it is then reclassified as a core cluster [11].

2.1.4. Functional Annotation and Characterisation of Pan-Genome Components

There are optional arguments in Pangloss through which the user can characterise pan-genomes once they are constructed (Figure 1). If InterProScan (https://www.ebi.ac.uk/interpro/download.html) is installed, the user can select to have the entire pan-genome dataset annotated with Pfam, InterPro and gene ontology (GO) information via the --ips command-line argument [44]. Additionally, if GOAtools (https://github.com/tanghaibao/goatools) is installed, the output from InterProScan can be used to perform GO-enrichment analysis of the core and accessory components of the pan-genome via the --go command-line argument, using Fischer's exact test (FET) with parent term propagation and false discovery rate correction ($p < 0.05$) using a p-value distribution generated from 500 resampled p-values [45,48].

2.1.5. Selection Analysis of Pan-Genome Using yn00

The user can perform selection analysis on core and accessory gene model clusters using yn00 from the PAML (http://abacus.gene.ucl.ac.uk/software/paml.html#download) package of phylogenetic software (enabled via the --yn00 command-line argument) (Figure 1) [43]. For each cluster in a pangenome dataset, an amino acid alignment is performed using MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/) with the default parameters. A corresponding nucleotide alignment is then generated by Pangloss by transferring gaps in the amino acid alignment into the nucleotide data for the same cluster. yn00 selection analysis is handled by Biopython's Bio.Phylo.PAML module (https://biopython.org/) and is run with the default parameters (universal genetic code, equal weighting of pathways between codons and estimated codon frequencies). From each cluster alignment, Pangloss will report where available the estimated transition/transversion rate ratio of the cluster (κ) and the number of pairwise alignments within the cluster that show evidence of positive selection according to Yang and Nielsen's method where the d_N/d_S ratio (ω) is ≥ 1, if $\omega \neq \infty$ [49].

2.1.6. Visualisation of Pan-Genome Data

A number of optional methods of visualising pan-genome data are incorporated into Pangloss (Figure 1). A simple ring chart of the proportion of core and accessory gene models in a pangenome dataset is generated in R using the --size command-line argument. The same flag also generates a bar chart for the distribution of syntenic cluster sizes within a pangenome dataset and estimates the true size of the pan-genome using the Chao lower bound method in R, as previously implemented in the prokaryote pan-genome analysis package micropan [50,51]. The Chao lower bound method estimates the size of a population given a set of occurrence data for that population from singleton and doubleton occurrences [50]. In the case of pan-genomic data we can estimate the true number of syntenic clusters within a pan-genome ($\hat{N}$) given the observed number of clusters (N) from the numbers of 1-member and 2-member clusters in the pan-genome (y_1 and y_2, respectively), as given by the equation [50]:

$$\hat{N} = N + \frac{y_1^2}{2y_2}$$

The Chao lower bound method is a conservative method of estimating true pan-genome size, but it is worth noting that this estimation may be skewed in cases of overabundance of singleton data (e.g., singleton genes arising from highly fragmented genomes) [51,52]. The distribution of syntenic orthologous gene models within the species accessory genome can be visualised using the R package UpSetR via the --upset command-line argument [35]. This generates an ortholog distribution plot based on the UpSet technique of visualising intersections of sets and their occurrences within a dataset using matrix representation, allowing for more input sets than similar Venn-based or Euler-based methods [53]. Finally, karyotype plots of the genomic locations of core and accessory gene models along each chromosome/contig within a genome, coloured by either pan-genome component or by syntenic cluster size, can be generated for each genome in a dataset using the Bioconductor package KaryoploteR (https://bioconductor.org/packages/release/bioc/html/karyoploteR.html) via the --karyo command-line argument [36,37].

2.2. Dataset Assembly

2.2.1. *Yarrowia lipolytica*

Nuclear genome assembly data for seven *Yarrowia lipolytica* strains was obtained from GenBank. Each strain genome was selected based on geographic and environmental distribution, information on which is found in Table S1 [24,54–56]. Gene model and gene model location prediction was carried out for all *Y. lipolytica* strain genomes using Pangloss (Figure 1). GeneMark-ES gene model prediction was performed with a fungal branching point model and TransDecoder gene model prediction was performed with an amino acid sequence length cut-off of ≥200 aa. All predicted gene model sets were filtered against a set of 936 known pseudogenes or dubious open reading frames (ORFs) from *Saccharomyces cerevisiae* and *Candida albicans* obtained from the *Saccharomyces* and *Candida* Genome Database websites respectively, with a BLASTp e-value cut-off of 10^{-4} [47,57]. Gene models with sequence coverage of ≥70% to a pseudogene/dubious ORF were removed from the dataset (Table S1). BUSCO analysis for each strain gene model set was performed using the Saccharomycetales dataset (Table S1). In total, 45,533 gene models were predicted across our entire *Y. lipolytica* pan-genome dataset, with an average of 6504 gene models per strain and BUSCO completeness per gene model set ranging from approximately 83–89% (87.9% average) (Table S1).

2.2.2. *Aspergillus fumigatus*

Nuclear genome assembly data for 12 *Aspergillus fumigatus* strains was obtained from GenBank. Each strain genome was previously used to construct an initial *A. fumigatus* species pan-genome using a similar approach to that implemented in Pangloss, and strains were selected based on geographic and environmental distribution including both clinical and wild-type strains [11] (Table S1). Gene model and gene model location prediction was carried out for all *A. fumigatus* genomes using Pangloss (Figure 1). GeneMark-ES gene model prediction was performed with a fungal branching point model and TransDecoder gene model prediction was performed with an amino acid sequence length cut-off of ≥200 aa. No filtering for pseudogenes or dubious ORFs was performed for the *A. fumigatus* dataset as no such data is available. BUSCO analysis for each strain gene model set was performed using the Eurotiomycetes dataset (Table S1). In total, 113,414 gene models were predicted across our entire *A. fumigatus* pan-genome dataset, with an average of 9451 gene models per strain and BUSCO completeness per gene model set ranging from approximately 93–97% (96% average) (Table S1).

2.3. Pangenome Analysis

2.3.1. *Yarrowia lipolytica*

An all-vs.-all BLASTp search for the entire *Y. lipolytica* dataset was performed within Pangloss with an e-value cut-off of 10^{-4}. PanOCT analysis for the *Y. lipolytica* dataset was performed within Pangloss using the default parameters for PanOCT (CGN window = 5, sequence identity cut-off ≥35%). Pan-genome refinement was carried out within Pangloss (Table S1). Pfam, InterPro and gene ontology annotation of the dataset was performed using InterProScan with the default parameters [44,58–60].

GO-slim enrichment analysis was carried out for both the core and accessory *Y. lipolytica* genomes using GOATools. GO terms were mapped to the general GO-slim term basket and a Fischer's exact test (FET) analysis with parent term propagation and false discovery rate (FDR) correction ($p < 0.05$) with a *p*-value distribution generated from 500 resampled *p*-values [45,48,60]. yn00 analysis of the *Y. lipolytica* pan-genome dataset was performed within Pangloss with the default parameters [43,49]. All plots were generated within Pangloss using its various R components as detailed above (Figures 1–5).

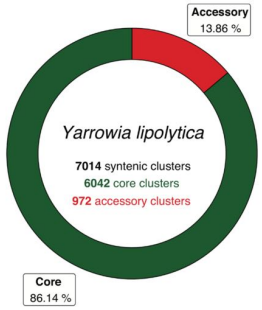


Figure 2. Pan-genome of *Yarrowia lipolytica* represented as a ring chart of proportions of core and accessory ortholog clusters within the total dataset. Modified from original figure generated by Pangloss. Core proportions coloured in green, accessory proportions coloured in red.

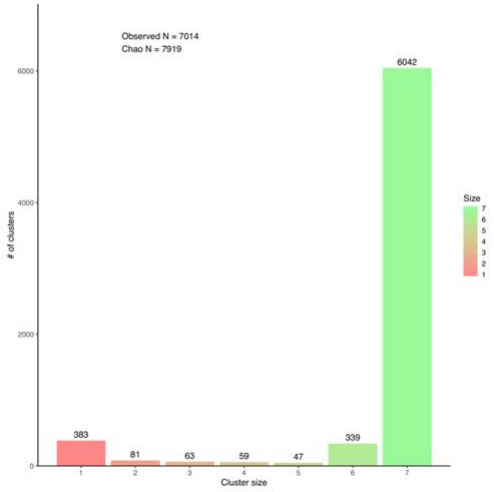


Figure 3. Bar chart representing the distribution of syntenic cluster sizes within *Yarrowia lipolytica* pan-genome and Chao's lower bound estimation of true pan-genome size. Figure generated by Pangloss.

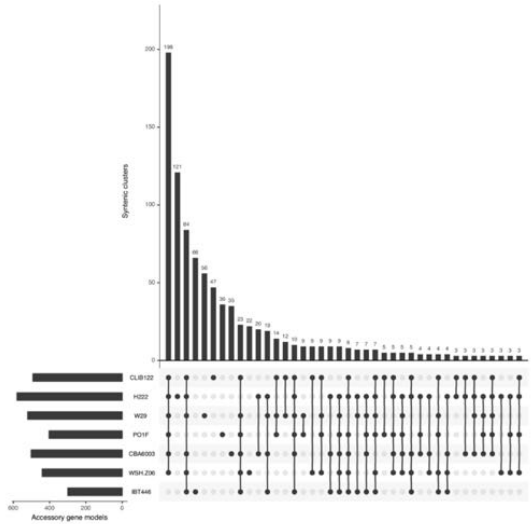


Figure 4. UpSet plot of the distribution of syntenic orthologs within the *Yarrowia lipolytica* accessory genome, ranked by syntenic cluster frequency. UpSet plots represent intersections between sets within data as a matrix and the number of occurrences of those intersections as a bar chart. In our case, the set intersection matrix represents clusters which contain a syntenic ortholog from 1–6 strains in our dataset and the number of their occurrences is given by the bar chart. Numbers of singleton clusters range from 22 in WSH-Z06 to 121 in H222. Figure generated by Pangloss.

2.3.2. *Aspergillus fumigatus*

An all-vs.-all BLASTp search for the entire *A. fumigatus* dataset was performed within Pangloss with an e-value cut-off of 10^{-4}. PanOCT analysis for the *A. fumigatus* dataset was performed within Pangloss using the default parameters for PanOCT (CGN window = 5, sequence identity cut-off ≥35%). Pan-genome refinement was carried out within Pangloss (Table S1).

3. Results

3.1. Analysis of the Yarrowia lipolytica Pan-Genome

A *Y. lipolytica* species pan-genome was constructed with Pangloss via PanOCT using publicly-available assembly data from seven strains, including the reference CLIB122 strain and a number of other industrially-relevant strains [24,54–56] (Table S1). Strain genomes ranged in size from 19.7–21.3 Mb, and the majority had been assembled to near-scaffold quality (Table S1). A total of 45,533 valid *Y. lipolytica* gene models were predicted by Pangloss after filtering for known pseudogenes from model yeasts, for an average of ~6505 gene models per strain genome (Table S1). Pangloss constructed a refined species pan-genome for *Y. lipolytica* containing 6042 core syntenic clusters (42,294 gene models in total) and 972 accessory syntenic clusters (3239 gene models in total) (Figure 2, Table 2 and Table S1). This gives a core:accessory proportion split of approximately 92:8 in terms of gene models and 87:13 in terms of unique syntenic clusters (Figure 2, Table S1). These core:accessory proportions were similar

to our previous analyses of other yeasts such as *Saccharomyces cerevisiae* (85:15) and *Candida albicans* (91:9) [11]. Accessory genome size in individual *Y. lipolytica* strains varied from 303 gene models in IBT446 to 583 gene models in H222 (Table S1). Using Chao's lower bound method, the size of the *Y. lipolytica* pan-genome was estimated to contain 7970 syntenic clusters (Figure 3). 341 syntenic clusters were missing an ortholog in one strain, with 202 clusters missing an ortholog from IBT446 only, and 390 syntenic clusters consisted of a singleton gene model (Figures 3 and 4). The number of singleton gene models in individual strains varied from 23 gene models in WSH-Z06 and CBA6003 to 121 gene models in H222 (Figure 4). Karyotype plots were generated for each *Y. lipolytica* strain in our dataset and display varying amounts of accessory gene models distributed across the six chromosomes of *Y. lipolytica* (e.g., CLIB122 in Figure 5a,b). This is similar to our previous observation of accessory genome distribution within the *Candida albicans* pan-genome, which may have arisen due to a lack of non-clinical strain genomes for that species [11]. A large accessory region in chromosome D in CLIB122 (NC_006070.1, Figure 5a,b) appears to be the result of a gapped region in the same chromosome in PO1f, presumably arising from sequencing artefacts (Figure 5a,b).

Table 2. Pan-genomes of *Yarrowia lipolytica* and *Aspergillus fumigatus*. Refer to Table S1 for further information including strain assembly statistics, BUSCO completeness and links to relevant literature.

Species	Strains	Core Genome		Accessory Genome		Pan-Genome	
		Gene Models	Clusters	Gene Models	Clusters	Gene Models	Clusters
Yarrowia lipolytica	7	42,294	6042	3239	972	45,533	7014
Aspergillus fumigatus	12	92,016	7668	21,398	3727	113,414	11,395

3.2. Characterisation of the Yarrowia lipolytica Pan-Genome

Selection analysis was performed for all non-singleton clusters in the *Y. lipolytica* core and accessory genome using yn00, which estimates synonymous and non-synonymous rates of substitution within a gene family using pairwise comparisons [43]. Of the 6042 core clusters in the *Y. lipolytica* pan-genome dataset, 453 clusters had at least one pairwise alignment which had $\omega \geq 1$ (7% of all core clusters), whereas for the 582 non-singleton accessory clusters only 52 clusters had at least one pairwise alignment with $\omega \geq 1$ (9% of all non-singleton accessory clusters). It is possible that the low levels of positive selection (i.e., clusters with ≥1 pairwise alignment with $\omega \geq 1$) within the accessory genome reflects the potential lack of evolutionary distance between the strains in our *Y. lipolytica* dataset. The *Y. lipolytica* pangenome dataset was annotated with Pfam, InterPro and gene ontology data using InterProScan [44,58–60]. Approximately 77% of the total dataset (35,139 gene models) contained at least one Pfam domain. GO-slim enrichment analysis was performed for both core and accessory genomes using GOATools with the default parameters as implemented in Pangloss (Table S2). Unlike our previous analysis of term enrichment in fungal pan-genomes, transport processes appear to be enriched within the core *Y. lipolytica* genome and processes relating to the production of organic and aromatic compounds are enriched within the accessory *Y. lipolytica* genome (Table S2) [11]. The former may be due to the array of the lipid transport systems that *Y. lipolytica* uses to live in environments rich in hydrophobic substrates [61]. Similarly, genes whose functions are related to intracellular organelle function are enriched in the *Y. lipolytica* core genome—this may encompass the accumulation of lipids and fatty acids within organelles and lipid body formation within the *Y. lipolytica* cell (Table S2) [62].

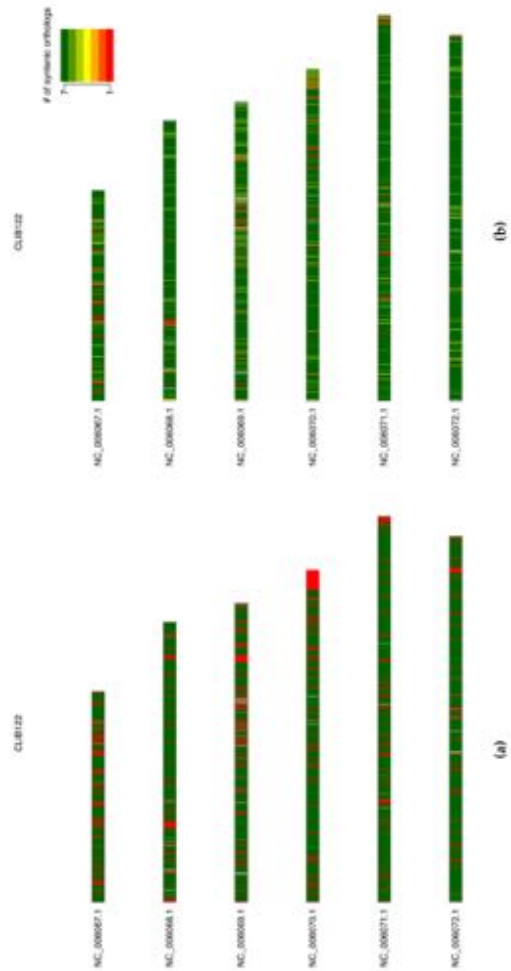


Figure 5. Karyotype plots of core and accessory gene model locations across the six chromosomes of *Yarrowia lipolytica* strain CLIB122. Left: **(a)** Gene model locations coloured by source pan-genome component (core: green, accessory: red). Right: **(b)** Gene model locations coloured by the size of their source syntenic cluster. Non-coding regions coloured in grey. Both figures generated by Pangloss.

3.3. Reanalysis of the Aspergillus fumigatus Pan-Genome

As a way of assessing the quality of Pangloss's pan-genome construction we also reconstructed a species pan-genome for *Aspergillus fumigatus*, the opportunistic agent of invasive aspergillosis, using a previously-analysed dataset containing both clinical and wild-type strains [11,63] (Table 2, Table S1). A total of 113,414 valid *A. fumigatus* gene models were predicted by Pangloss with an average of ~9451 gene models per strain genome (Table 2, Table S1). Pangloss constructed a refined species pan-genome for *A. fumigatus* containing 7668 core syntenic clusters (92,016 gene models in total) and 1783 accessory syntenic clusters (21,398 gene models in total) (Table 2, Table S1). This gives a core:accessory proportion split of approximately 81:19 in terms of gene models and 67:33 in terms of unique syntenic clusters (Table 2, Table S1). These core:accessory proportions are relatively in line with our previous study of the same *A. fumigatus* pan-genome dataset, which found core:accessory proportion splits of 83:17 in terms of gene models and 73:27 in terms of unique syntenic clusters [11]. Variation between the two *A. fumigatus* pan-genome analyses is a result of performing gene prediction using Exonerate in our initial analysis but not in this subsequent reanalysis [11].

4. Discussion

As pan-genome analysis of eukaryotes becomes more commonplace, ideally the amount of software to construct and characterise eukaryote pan-genome should begin to match that which is already available for prokaryotes. Our software pipeline Pangloss applies a sequence similarity and synteny-based approach from prokaryote pan-genome analysis, implemented in the previously-published Perl software PanOCT, to eukaryote pan-genome analysis and allows the user to perform their own gene prediction and downstream characterisation and visualisation of pan-genome data from one self-contained script [11,22]. Although our pipeline has been designed for eukaryote pan-genome analysis, as PanOCT is a prokaryote method in origin, Pangloss should also support prokaryote datasets—albeit with some modifications to gene model prediction strategies by the user. Unlike other common gene clustering approaches, such as MCL, PanOCT incorporates local synteny via assessing the CGN between potential orthologs as a criterion to clustering in addition to sequence similarity [19,22]. This makes PanOCT distinct from most clustering approaches in that it can distinguish orthologs from paralogs (i.e., if one assumes that "true" orthologs are more likely to be located in relatively-similar regions of their respective genomes they then should in turn be more likely to cluster together when syntenic conservation is taken into consideration). This is of particular relevance to eukaryote pan-genomes, as gene duplication plays a substantial role in eukaryote gene family and genome evolution [11,64]. Although this approach is more stringent than clustering gene families based on approaches like MCL or BLAST searches alone, it is potentially more reflective of evolution on a gene-level basis within strains of the same species.

There are ways in which our approach can be improved upon in future methodologies, both in terms of prediction and analytic strategies. For example, Pangloss has an optional Exonerate-based gene model prediction strategy which searches input genomes for translated homologs of reference sequences [33]. This is an exhaustive approach that may pick up potential gene models missed by GeneMark-ES and/or TransDecoder, but it is also time-inefficient. To search all 6472 reference protein sequences from *Y. lipolytica* CLIB222 against a single *Y. lipolytica* genome takes, on average, four hours on three threads on a server running Ubuntu 18.04.2 LTS (approximately nine sequences per minute per thread), whereas both GeneMark-ES gene model prediction with fungal point branching and subsequent ORF prediction in non-coding regions with TransDecoder performed on the same genome with the same number of threads typically takes ~30–35 min. It is for this reason primarily that we have made the Exonerate-based strategy optional for any gene prediction that is performed by Pangloss. Furthermore, PanOCT's memory usage increases exponentially per strain added, notwithstanding the potentially complex distribution of gene models between strains themselves [11,22]. Constructing a species pan-genome using PanOCT from a small and relatively well-conserved dataset, such as that for our *Y. lipolytica* or *A. fumigatus* studies, should be achievable on most standard hardware. For larger datasets,

such as our previous pan-genome analysis of 100 *Saccharomyces cerevisiae* genomes; however, it may be preferable to perform such analysis on a high-performance computational environment or otherwise an alternative synteny-based method of pan-genome construction may be more appropriate [11]. Finally, we would encourage users to interrogate and visualise the results of analysis using Pangloss and adjust the input parameters where appropriate for their data. In our case, the parameters which were chosen for use in Pangloss for this analysis (e.g., BLAST e-value cut-off, CGN window size) are largely based on those from our previous analysis of fungal pan-genomes or other studies using PanOCT [11,22]. Depending on the size of a pan-genome dataset or the species of interest, different cut-offs may be more suitable (e.g., for species with longer average gene lengths a lower sequence identity cut-off for PanOCT clustering than the default (>35%) may be more appropriate). Many of these parameters can be adjusted in the configuration file provided with Pangloss.

5. Conclusions

Pan-genome analysis of eukaryotes has become more common, but many of the available software for pan-genome analysis are intended for use with prokaryote data. We have developed Pangloss, a pipeline that allows users to generate input data and construct species pan-genomes for microbial eukaryotes using the synteny-dependent PanOCT method and various downstream characterisation analyses. To demonstrate the capabilities of our pipeline we constructed a species pan-genome for *Yarrowia lipolytica*, an oleaginous yeast with potential biotechnological applications, and performed various functional and data visualisation analyses using Pangloss. The *Y. lipolytica* pangenome is similar in terms of core and accessory genome proportions to previously analysed fungal pan-genomes but is unique in that biological processes such as transport are statistically-enriched in the core genome. We also used Pangloss to reconstruct a species pan-genome for the respiratory pathogen *Aspergillus fumigatus* using a previously-analysed dataset and found that Pangloss generated a similar pan-genomic structure for *A. fumigatus* to that of our previous analysis. Building on our previous work on fungal pan-genomes, this study not only provides further evidence for pan-genomic structure within eukaryote species but also presents a methodological pipeline for future eukaryote pan-genome analysis.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/7/521/s1. Table S1, Information for *Yarrowia lipolytica* and *Aspergillus fumigatus* pan-genome datasets. Core gene models labelled in green, accessory gene models labelled in red. References and strain information taken from cited articles where available, otherwise from GenBank or similar resources with relevant links included. Table S2. GO-slim enrichment analysis for the *Yarrowia lipolytica* pan-genome dataset. Fischer's exact test with FDR correction ($p < 0.05$) carried out using GOATools within Pangloss. All terms present in the table are either significantly over- or under-represented in either the *Y. lipolytica* core or accessory genome. Significantly over-represented terms labelled green, significantly under-represented terms labelled red.

Author Contributions: C.G.P.M: Conceptualization, methodology, software, formal analysis, investigation, data curation, Writing—Original draft, and Writing—Review and editing. D.A.F: Conceptualization, methodology, investigation, Writing—Review and editing, supervision, and project administration.

Funding: CGPM is funded by an Irish Research Council Government of Ireland Postgraduate Scholarship (Grant No. GOIPG/2015/2242).

Acknowledgments: The authors would like to acknowledge the original contributors to all sequencing data used in this analysis for making their data publicly available. The authors would also like to acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Donati, C.; Medini, D.; Ward, N.L.; Angiuoli, S.V.; Crabtree, J.; Jones, A.L.; Durkin, A.S.; et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13950–13955. [CrossRef] [PubMed]

2. Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **2005**, *15*, 589–594. [CrossRef] [PubMed]
3. Rouli, L.; Merhej, V.; Fournier, P.E.; Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **2015**, *7*, 72–85. [CrossRef] [PubMed]
4. Vernikos, G.; Medini, D.; Riley, D.R.; Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **2015**, *23*, 148–154. [CrossRef] [PubMed]
5. Mosquera-Rendón, J.; Rada-Bravo, A.M.; Cárdenas-Brito, S.; Corredor, M.; Restrepo-Pineda, E.; Benítez-Páez, A. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. *Bmc Genom.* **2016**, *17*, 45. [CrossRef] [PubMed]
6. Lefébure, T.; Bitar, P.D.P.; Suzuki, H.; Stanhope, M.J. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biol. Evol.* **2010**, *2*, 646–655. [CrossRef] [PubMed]
7. Sigalova, O.; Chaplin, A.V.; Bochkareva, O.O.; Shelyakin, P.V.; Filaretov, V.A.; Akkuratov, E.; Burskaya, V.; Gelfand, M.S. *Chlamydia* pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction. *bioRxiv* **2018**, 506121. [CrossRef]
8. Golicz, A.A.; Bayer, P.E.; Barker, G.C.; Edger, P.P.; Kim, H.R.; Martinez, P.A.; Chan, C.K.K.; Severn-Ellis, A.; McCombie, W.R.; Parkin, I.A.P.; et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **2016**, *7*, 13390. [CrossRef] [PubMed]
9. Plissonneau, C.; Hartmann, F.E.; Croll, D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *Bmc Biol.* **2018**, *16*, 5. [CrossRef]
10. Peter, J.; De Chiara, M.; Friedrich, A.; Yue, J.-X.; Pflieger, D.; Bergstrom, A.; Sigwalt, A.; Barré, B.; Freel, K.; Llored, A.; et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **2018**, *556*, 339–344. [CrossRef]
11. McCarthy, C.G.P.; Fitzpatrick, D.A. Pan-genome analyses of model fungal species. *Microb. Genom.* **2019**, *5*, 1–23. [CrossRef] [PubMed]
12. Read, B.A.; Kegel, J.; Klute, M.J.; Kuo, A.; Lefebvre, S.C.; Maumus, F.; Mayer, C.; Miller, J.; Monier, A.; Salamov, A.; et al. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **2013**, *499*, 209–213. [CrossRef] [PubMed]
13. Page, A.J.; Cummins, C.A.; Hunt, M.; Wong, V.K.; Reuter, S.; Holden, M.T.G.; Fookes, M.; Falush, D.; Keane, J.A.; Parkhill, J. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **2015**, *31*, 3691–3693. [CrossRef] [PubMed]
14. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef] [PubMed]
15. Jandrasits, C.; Dabrowski, P.W.; Fuchs, S.; Renard, B.Y. seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. *Bmc Genom.* **2018**, *19*, 47. [CrossRef] [PubMed]
16. Marcus, S.; Lee, H.; Schatz, M.C. SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* **2014**, *30*, 3476–3483. [CrossRef]
17. Sahl, J.W.; Caporaso, J.G.; Rasko, D.A.; Keim, P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *PeerJ* **2014**, *2*, e332. [CrossRef] [PubMed]
18. Enright, A.J.; Van Dongen, S.; Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584. [CrossRef]
19. Alexeyenko, A.; Tamas, I.; Liu, G.; Sonnhammer, E.L.L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **2006**, *22*, e9–e15. [CrossRef]
20. Zhao, Y.; Wu, J.; Yang, J.; Sun, S.; Xiao, J.; Yu, J. PGAP: Pan-genomes analysis pipeline. *Bioinformatics* **2012**, *28*, 416–418. [CrossRef]
21. Hu, Z.; Sun, C.; Lu, K.C.; Chu, X.; Zhao, Y.; Lu, J.; Shi, J.; Wei, C. EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics* **2017**, *33*, 2408–2409. [CrossRef] [PubMed]
22. Fouts, D.E.; Brinkac, L.; Beck, E.; Inman, J.; Sutton, G. PanOCT: Automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res.* **2012**, *40*, e172. [CrossRef] [PubMed]
23. Rasko, D.A.; Myers, G.S.A.; Ravel, J. Visualization of comparative genomic analyses by BLAST score ratio. *Bmc Bioinform.* **2005**, *6*, 2. [CrossRef] [PubMed]
24. Dujon, B.; Sherman, D.; Fischer, G.; Durrens, P.; Casaregela, S.; Lafontaine, I.; De Montigny, J.; Marck, C.; Neuvéglise, C.; Talla, E.; et al. Genome evolution in yeasts. *Nature* **2004**, *430*, 35–44. [CrossRef] [PubMed]

25. Shen, X.-X.; Zhou, X.; Kominek, J.; Kurtzman, C.P.; Hittinger, C.T.; Rokas, A. Reconstructing the Backbone of the *Saccharomycotina* Yeast Phylogeny Using Genome-Scale Data. *G3 Genes Genomes Genet.* **2016**, *6*, 3927–3939. [CrossRef] [PubMed]
26. O'Brien, C.E.; McCarthy, C.G.P.; Walshe, A.E.; Shaw, D.R.; Sumski, D.A.; Krassowski, T.; Fitzpatrick, D.A.; Butler, G. Genome analysis of the yeast *Diutina catenulata*, a member of the *Debaryomycetaceae/Metschnikowiaceae* (CTG-Ser) clade. *PLoS ONE* **2018**, *13*, e0198957. [CrossRef]
27. Nicaud, J.M. *Yarrowia lipolytica*. *Yeast* **2012**, *29*, 409–418. [CrossRef]
28. Adrio, J.L. Oleaginous yeasts: Promising platforms for the production of oleochemicals and biofuels. *Biotechnol. Bioeng.* **2017**, *114*, 1915–1920. [CrossRef]
29. Friedlander, J.; Tsakraklides, V.; Kamineni, A.; Greenhagen, E.H.; Consiglio, A.L.; MacEwen, K.; Crabtree, D.V.; Afshar, J.; Nugent, R.L.; Hamilton, M.A.; et al. Engineering of a high lipid producing *Yarrowia lipolytica* strain. *Biotechnol. Biofuels* **2016**, *9*, 77. [CrossRef]
30. Qiao, K.; Wasylenko, T.M.; Zhou, K.; Xu, P.; Stephanopoulos, G. Lipid production in *Yarrowia lipolytica* is maximized by engineering cytosolic redox metabolism. *Nat. Biotechnol.* **2017**, *35*, 173–177. [CrossRef]
31. Zeng, W.; Fang, F.; Liu, S.; Du, G.; Chen, J.; Zhou, J. Comparative genomics analysis of a series of *Yarrowia lipolytica WSH-Z06* mutants with varied capacity for α-ketoglutarate production. *J. Biotechnol.* **2016**, *239*, 76–82. [CrossRef] [PubMed]
32. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef] [PubMed]
33. Slater, G.S.C.; Birney, E. Automated generation of heuristics for biological sequence comparison. *Bmc Bioinform.* **2005**, *6*, 31. [CrossRef] [PubMed]
34. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **2011**, *3*, 180–185. [CrossRef]
35. Conway, J.R.; Lex, A.; Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **2017**, *33*, 2938–2940. [CrossRef]
36. Obenchain, V.; Bravo, H.C.; Huber, W.; Lawrence, M.; Carlson, M.; MacDonald, J.; Carey, V.J.; Irizarry, R.A.; Love, M.I.; Hahne, F.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **2015**, *12*, 115–121. [CrossRef]
37. Gel, B.; Serra, E. KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **2017**, *33*, 3088–3090. [CrossRef]
38. Ter-Hovhannisyan, V.; Lomsadze, A.; Chernoff, Y.O.; Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **2008**, *18*, 1979–1990. [CrossRef]
39. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. [CrossRef]
40. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *Bmc Bioinform.* **2009**, *10*, 421. [CrossRef]
41. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef] [PubMed]
42. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef] [PubMed]
43. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef] [PubMed]
44. Jones, P.; Binns, D.; Chang, H.-Y.Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef] [PubMed]
45. Klopfenstein, D.V.; Zhang, L.; Pedersen, B.S.; Ramírez, F.; Vesztrocy, A.W.; Naldi, A.; Mungall, C.J.; Yunes, J.M.; Botvinnik, O.; Weigel, M.; et al. GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **2018**, *8*, 10872. [CrossRef] [PubMed]
46. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]

47. Engel, S.R.; Cherry, J.M. The new modern era of yeast genomics: Community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database. *Database* **2013**, *2013*, bat012. [CrossRef] [PubMed]
48. Agresti, A. *Categorical Data Analysis*; Wiley Series in Probability and Statistics; John Wiley and Sons Inc.: Hoboken, NJ, USA, 2002; ISBN 0471360937.
49. Yang, Z.; Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **2000**, *17*, 32–43. [CrossRef] [PubMed]
50. Chao, A. Non-parametric estimation of the classes in a population. *Scand. J. Stat.* **1984**, *11*, 265–270. [CrossRef]
51. Snipen, L.; Liland, K.H. micropan: An R-package for microbial pan-genomics. *Bmc Bioinform.* **2015**, *16*, 1–8. [CrossRef] [PubMed]
52. Böhning, D.; Kaskasamkul, P.; van der Heijden, P.G.M. A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika* **2019**, *82*, 361–384. [CrossRef]
53. Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuillemot, R.; Pfister, H. UpSet: Visualization of intersecting sets. *Ieee Trans. Vis. Comput. Graph.* **2014**, *20*, 1983–1992. [CrossRef] [PubMed]
54. Liu, L.; Alper, H.S. Draft Genome Sequence of the Oleaginous Yeast *Yarrowia lipolytica* PO1f, a Commonly Used Metabolic Engineering Host. *Genome Announc.* **2014**, 2. [CrossRef] [PubMed]
55. Magnan, C.; Yu, J.; Chang, I.; Jahn, E.; Kanomata, Y.; Wu, J.; Zeller, M.; Oakes, M.; Baldi, P.; Sandmeyer, S. Sequence assembly of *Yarrowia lipolytica* strain W29/CLIB89 shows transposable element diversity. *PLoS ONE* **2016**, *11*, e0162363. [CrossRef] [PubMed]
56. Devillers, H.; Neuvéglise, C. Genome Sequence of the Oleaginous Yeast *Yarrowia lipolytica* H222. *Microbiol. Resour. Announc.* **2019**, 8. [CrossRef] [PubMed]
57. Skrzypek, M.S.; Binkley, J.; Binkley, G.; Miyasato, S.R.; Simison, M.; Sherlock, G. The *Candida* Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.* **2017**, *45*, D592–D596. [CrossRef]
58. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2015**, *44*, D279–D285. [CrossRef]
59. Hunter, S.; Jones, P.; Mitchell, A.; Apweiler, R.; Attwood, T.K.; Bateman, A.; Bernard, T.; Binns, D.; Bork, P.; Burge, S.; et al. InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res.* **2012**, *40*, D306–D312. [CrossRef]
60. Carbon, S.; Dietze, H.; Lewis, S.E.; Mungall, C.J.; Munoz-Torres, M.C.; Basu, S.; Chisholm, R.L.; Dodson, R.J.; Fey, P.; Thomas, P.D.; et al. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Res.* **2017**, *45*, D331–D338. [CrossRef]
61. Thevenieau, F.; Beopoulos, A.; Desfougeres, T.; Sabirova, J.; Albertin, K.; Zinjarde, S.; Nicaud, J.-M. Uptake and Assimilation of Hydrophobic Substrates by the Oleaginous Yeast *Yarrowia lipolytica*. In *Handbook of Hydrocarbon and Lipid Microbiology*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1513–1527.
62. Mlíčková, K.; Roux, E.; Athenstaedt, K.; D'Andrea, S.; Daum, G.; Chardot, T.; Nicaud, J.M. Lipid accumulation, lipid body formation, and acyl coenzyme A oxidases of the yeast *Yarrowia lipolytica*. *Appl. Env. Microbiol.* **2004**, *70*, 3918–3924. [CrossRef]
63. Nierman, W.C.; Pain, A.; Anderson, M.J.; Wortman, J.R.; Kim, H.S.; Arroyo, J.; Berriman, M.; Abe, K.; Archer, D.B.; Bermejo, C.; et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **2005**, *438*, 1151–1156. [CrossRef] [PubMed]
64. Friedman, R.; Hughes, A.L. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **2001**, *11*, 373–381. [CrossRef] [PubMed]