Studies on the Modular Evolution of Genes

A thesis submitted to the National University of Ireland for the degree of
Doctor of Philosophy

Presented by:

Robert James Leigh B.Sc.(Hons.)
Genome Evolution Laboratory
Department of Biology
Maynooth University
Co. Kildare
Ireland

February 2019

Supervisors:

Dr. David A. Fitzpatrick B.Sc.(Hons.), Ph.D., Pg.Dip.H.E.
Maynooth University, Maynooth, Co. Kildare, Ireland

Professor James O. McInerney B.Sc.(Hons.), Ph.D., D.Sc., F.L.S.
University of Nottingham, Nottingham, United Kingdom

Head of Department:

Professor Paul N. Moynagh B.A.(mod.), Ph.D., M.R.I.A.

Table of Contents

Acknowledgements

Firstly and foremostly, I'd like to express my sincerest gratitude to my supervisors Dr. David Fitzpatrick and Professor James McInerney. Without their help, guidance, support and most undying patience, none of this work would have been possible.

I would like to thank the Maynooth University John and Pat Hume Scholarship for funding my research over the past four years. I would like to thank the technical and administitive staff at the MU Department of Biology, specifically Michelle, Terry, Frances, Gemma, and Trish for their support and guidance throughout the past 4 years.

This work likely would not have been possible without the support and friendship of my lab partners Charley, Eoin, Jamie, Sera, Ricardo, Matt, David, Martin and Fiona. I couldn't imagine a better group of people to slog through 4 years with.

I would like to thank my friends for putting up with my ever inflating head and frequent caffeine fuelled existential crises during the years, especially Sean, Caoimhe, Alex, Kevin, Jason, Becky, Jonathan, Shane, and Nate.

Lastly, but certainly not least, I would like to thank my parents, Patricia and Paul, and aunts Claire and Carol for their support and guidance throughout not only for this chapter of my life, but throughout my entire life

This thesis is dedicated to the memory of

Michelle (Shelly) Behan

and

Rafeek Khan

Two beloved friends who passed during the course of my studies.

Declaration

This thesis has not been submitted in whole, or in part, to this or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:_____

Robert J. Leigh

Abbreviations

-c	Mutual overlap
-x	Minimum composite family size
-y	Minimum component family size
α	Critical alpha (significance threshold)
AG	Alternation of generations
AMR	Antimicrobial resistance
B	Bonferroni corrected
bf	*bona-fide*
BLAST	Basic local alignment search tool
BLASTP	Protein BLAST
BUSCO	Benchmarked universal single copy orthologs
C	Connectivity score; combinatoric
c	Comparisons
Ca^{2+}	Calcium cation
CDF	Cumulative distribution function
cGMP	Cyclic guanosine monophosphate
CV	Coefficient of variation
CycPE	Cyclical polyploidisation events
DDC	Duplication, degeneration, and complementation
DFS	Depth first search
DNA	Deoxyribonucleic acid
DR	Distance ratio
E	Edges; Expectation value

e-value	Expectation value
EAC	Escape from adaptive conflict
Φ	Cumulative distribution function
f_b	Birth rate (frequency)
f_d	Decay rate (frequency)
*fdf*BLAST	Find differential fusions BLAST
FLH	Full length homolog
G	Graph
HGT	Horizontal gene transfer
HSP	High scoring pairs
IAD	Innovation, amplification, divergence
JGI	Joint Genome Institute
K^+	Potassium cation
KOG	Eukaryote orthologous group
KS	Kolmogorov-Smirnoff
LG	Le and Gascuel
MDR	Mutation during redundancy
n	number
NC	Nested composite
nfSSN	Non-familial SSN
NR	Non-remodelled
NRPS	Non-ribosomal peptide synthases
ns	Non-selelected
ORF	Open reading frame
outfmt	Output format

P	Probability value
PE	Polyploidisation events
PFAM	Protein families
pH	Potential of hydrogen
pident	Percentage identity
PKS	Polyketide synthases
PPI	Protein-protein interaction
Q	Q-function
qend	Query end position
qlen	Query sequence length
qseqid	Query sequence ID
qstart	Query start position
RC	Remodelling categories
RhGC	Rhodopsin-guanyl cyclase
RNA	Ribonucleic acid
SC	Strict composite
send	Subject end position
siRNA	Small interfering ribonucleic acid
slen	Subject sequence length
SN	Strict component
sseqid	Subject sequence ID
SSN	Sequence similarity network
sstart	Subject start position
Sup	Supremum
T_b	Traits birthed

T_d	Traits decayed
TE	Transposable elements
TNT	Tree analysis with New Technology
UTC	Ulvophyceae, Trebouxiophyceae, and Chlorophyceae
V	Vertices
Z	Z-score

Index of figures

Chapter I

Chapter II

Chapter III

Chapter IV

Index of tables

Chapter II

Chapter III

Chapter IV

Publications

Leigh R.J., Moran R.J., Pathmanathan J.S., Bapteste E., O'Connell M.J., Fitzpatrick D.A. & McInerney J.O. (in prep.) "Gene remodelling is rampant in fungi and plants".

Leigh R.J., Moore M., & McInerney J.O. (in prep.) "Reconstructing the ancestral metabolic pathways of LECA".

Moran R.J., Leigh R.J., Fitzpatrick D.A., McInerney J.O., & O'Connell M.J. (in prep.) "Gene remodelling in metazoa"

Conferences attended

Molecular and Genome Evolution Symposium (2015), Manchester, UK

Virtual Institute of Bioinformatics & Evolution (2015), Dublin, Ireland

Molecular and Genome Evolution Symposium (2016), Manchester, UK

Virtual Institute of Bioinformatics & Evolution (2016), Dublin, Ireland

Young Systematists' Forum (2016), London, United Kingdom

13th European Conference on Fungal Genetics (2016), Paris, France

Irish Fungal Society & British Mycological Society (2016), Dublin Ireland

Society for Molecular Biology & Evolution (2017), Austin, TX, USA

Abstract

Gene evolution is primarily studied through the observations of comparative cumulative point mutations between homologs. Genes also evolve through "remodelling", the process of repurposing and reorganising genes and gene fragments into novel sequences. Gene remodelling is a relatively underappreciated evolutionary concept. Remodelling events circumscribe the development of novel sequences *via* fusion or fission events and through the shuffling of exons or domains. To date, all studies into remodelling have focussed on specific remodelling events, for example gene fusions in cancer samples, or have used small datasets (<15 species). As such, a comparative remodelling analyses between two taxonomic Kingdoms has yet to be completed. In 2018, CompositeSearch was developed to overcome the computational bottlenecks associated with mining all possible combinations that may attribute to remodelling events. We used CompositeSearch to investigate the comparative extent of remodelling within large fungal (107 species) and plant (50 species) datasets. We observed approximately 50% of fungal genes and 61% of plant genes to have a history of remodelling despite robust controls against Type I errors. We observed the rate of remodelled family birth and decay to be clocklike in both datasets, and that remodelled genes were considerably more homoplastic than non-remodelled genes. Functional overrepresentation analysis concluded that remodelled genes were associated with rapidly evolving systems, such as secondary metabolism, and with phenotypic novelty, such as flowering in angiosperms.

Remodelling events have been associated with the development of antimicrobial resistance (AMR). As CompositeSearch does not discern between a fusion event and any other remodelling event, we developed CompositeBLAST to detect novel AMR fusion events. CompositeBLAST was considerably faster and more sensitive than previously published fusion detection tools. Using this software, we detected previously unreported mupirocin and vancomycin resistance genes as being derived from remodelling events.

Chapter I:

Introduction

1.1 Gene evolution

1.1.1. Introduction to modular evolution

Molecular evolution refers to the accruement of changes in biological sequences (such as DNA, RNA, or translated amino acid sequences) over successive generations. Molecular evolution is a relatively young field in biology, which utilises principles and practices from genetics, biochemistry, computational biology, and bioinformatics. Gene evolution, a subdiscipline of molecular evolution, focusses on changes in protein-coding genes over generations and is achieved through comparative genetics and comparative genomics. Themes in gene evolution explore, for example, the evolutionary origins of new genes within genomes (Haggerty *et al*., 2012), the evolution of novel functions (Hughes, 2005), and phenotype evolution (Zhang, 2018).

The emergence of novel genes in a genome, whether through duplication, horizontal gene transfer (HGT) or successive point mutations can promote significant functional innovation or phenotypic changes (Chandrasekaran and Betrán, 2008; Haggerty *et al*., 2012). A common example of this can be observed in the development of antimicrobial resistance in bacterial populations *via* plasmid mediated HGT (Lerminiaux & Cameron, 2019). Gene deletion may also confer considerable phenotypic changes (Bailey *et al*., 2019), and is often observed during the transition to parasitic life cycles (Sun *et al*., 2018). For example, successive losses of secondary metabolite production, cellulose degrading, and hemicellulose degrading genes have been implicated in the transition towards obligate biotrophy in the fungal phytopathogen *Blumera graminis* (Spanu *et al.*, 2010).

With the exception of HGT, genetic novelty in genomes typically arises due to "**genetic redundancy**", where a particular biological function is carried out by two or more genes (Force

et al., 1999), usually resulting from a duplication event (discussed further in Section 1.1.3.). In such instances, due to a lack in selection pressures, one duplicate is free to acquire mutations and evolve down its own pathway without affecting organismal fitness (Kleinjan *et al.*, 2008; Xia *et al*., 2016). Of course, there are exceptions to this, such as duplications resulting in gene toxicity or deleterious mutations in paralogs (Birchler and Veitia, 2012). As deleterious mutants are unlikely to persist in the genome, this thesis is exclusively concerned with "successful" mutants (divergent sequences that persist in the genome record). The potential fates of duplicated genes are discussed in *subsection 1.1.2.1.*

The cynosure of this thesis is the study of "modular evolution" in protein-coding genes, and, as such, the terms "gene" and "protein" are used interchangeably. Specifically, we investigate the genesis of novel gene families through "**gene remodelling**", a broad term describing the process of novel gene generation from existing gene sequences through processes such as domain shuffling (Kawashima *et al*., 2009) within a gene, or gene fusion and gene fission between independent genes (Leonard and Richards, 2012).

Typically, gene evolution is considered to be "tree-like", examined through the lens of vertical inheritance, and involves the study of the rates, distribution, and effect of nucleotide substitutions over time or between phenotypes (Haggerty *et al.*, 2012). Comparatively, gene remodelling examines the "network-like" aspects of gene evolution (Jachiet *et al*., 2014; Pathmanathan *et al*., 2018), where instead of having a single point of origin, a gene family can be formed from multiple, non-homologous progenitors (Haggerty *et al*., 2012) (Figure 1.1.1.). Remodelled gene families are the central focus of this thesis.

1.1.1.1. Selfish gene theory

The selfish gene theory (Dawkins, 1989) argues that genes compete for survival within genomes, tending to promote phenotypic selection that aids their own propagation. Briefly,

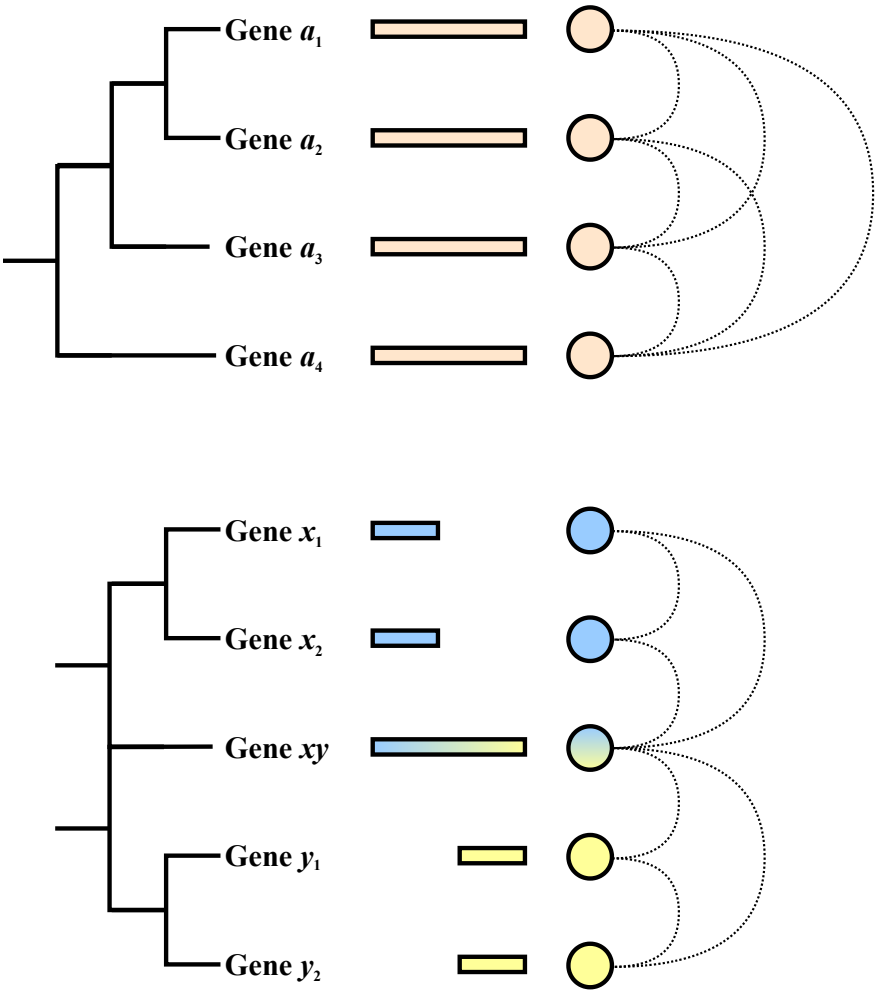

Figure 1.1.1. "Tree-like" *vs* "network-like" gene evolution

The top phylogeny depicts tree-like evolution as the number of roots (n_{roots}) does not exceed 1 and does not possess merged nodes. The bottom phylogeny depicts network-like evolution as $n_{roots} > 1$ and due to the possession of merged nodes. The bars beside each gene name represent a BLAST alignment (Altschul *et al*., 1997) between each gene, and colour annotations represent membership within a gene family. Circles beside the bars represent network vertices, where each vertex is a gene and each connection (edge) represents homology. All genes in the top phylogeny are connected to each other. Gene families *x* and *y* in the bottom phylogeny do not share homology with each other, however both share homology with family *xy*.

adaptations that are beneficial for the propagation for a specific allele will be selected against genes in direct competition, resulting in a biased allele distribution. As traits are inherited almost exclusively *via* genetic propagation, evolutionary processes, such as selection, are better understood through the observation of genes as opposed to exclusively relying on phenotypic data such as fossils.

1.1.1.2. Gene duplication

The duplication of genes is a major driving force underlying expansions in the genetic repertoire, and subsequently lays the foundation for the generation of genetic novelty (Wagner, 1998; Garsmeur *et al.*, 2014). Duplicates arising within an organism are known as **paralogs**, whereas genes arising from a common ancestor prior to a speciation event are known as **orthologs**. Duplication mediated evolution has been under consistent, intense scrutiny since Ohno first described the phenomenon in detail (Ohno, 1970). Paralogs are usually indistinguishable from each other immediately after a duplication event, and are thus functionally redundant (Force *et al.*, 1999). Redundancy reduces selective pressure on at least one paralog, which may then accumulate point mutations resulting in innovation or decay (Rice and Palmer, 2006; Wang and Paterson, 2011).

1.1.1.2.1. Potential fates of duplicated genes

The potential fates of duplicated genes are commonly debated (*eg.* Wagner, 1998; Rastogi and Liberles, 2005; Nardmann and Werr, 2013). In multicellular organisms, unless a duplication event occurs in germ cells, paralogs are present in just one cell, which must compete with neighbouring cells to ensure paralogous survival. Comparatively, duplicates

arising in single celled organisms or within the germ cells of a multicellular organism must compete with other individuals within a population that do not possess this paralogous pair. If a paralog is deleterious to normal function, thus reducing host fitness, it is likely to be removed from the gene pool prior to reproduction or within a few generations of the initial duplication event (Veitia, 2005). As mentioned above, silent mutations may not be selected against and may persist within a population. Comparatively, if a paralog is advantageous, it is likely to be selected for and passed on to successive generations (Force *et al.*, 1999).

Gene dosage refers to the sum of translated sequences by a specific gene in a given cell over time (Qian *et al.*, 2008). If gene concentration is suboptimal, a duplication event could hypothetically increase fitness. Therefore, unless inhibited, increases in gene copy number enables increased protein production. If amplified protein production is advantageous, little variation, if any, accumulates within the paralogous sequence while maintenance remains advantageous (Lynch and Conery, 2000). The slow accumulation of point mutations in a positively selected paralog, or point mutations accrued once paralogous maintenance is no longer advantageous results in altered functionality (Hughes *et al.*, 2014). Comparatively, a paralog may be deleterious to an organism due to interference with tightly coordinated pathways. Prominent fitness declination arising from paralogs is exemplified in human neuropathies where paralogs of α-synuclein have been associated with early onset Parkinson's disease (Rice and McLysaght, 2017). Parkinson's disease progresses through the accumulation of Lewy bodies in nerve tissues (Jellinger and Korczyn, 2018). Lewy bodies are composed of granular tissue and α-synuclein variants. As α-synuclein is involved in embryonic nerve development, germline mutations would likely render a host non-viable. A duplication event allows for α-synuclein variants to arise thus contributing to Lewy body formation (Rice and McLysaght, 2017).

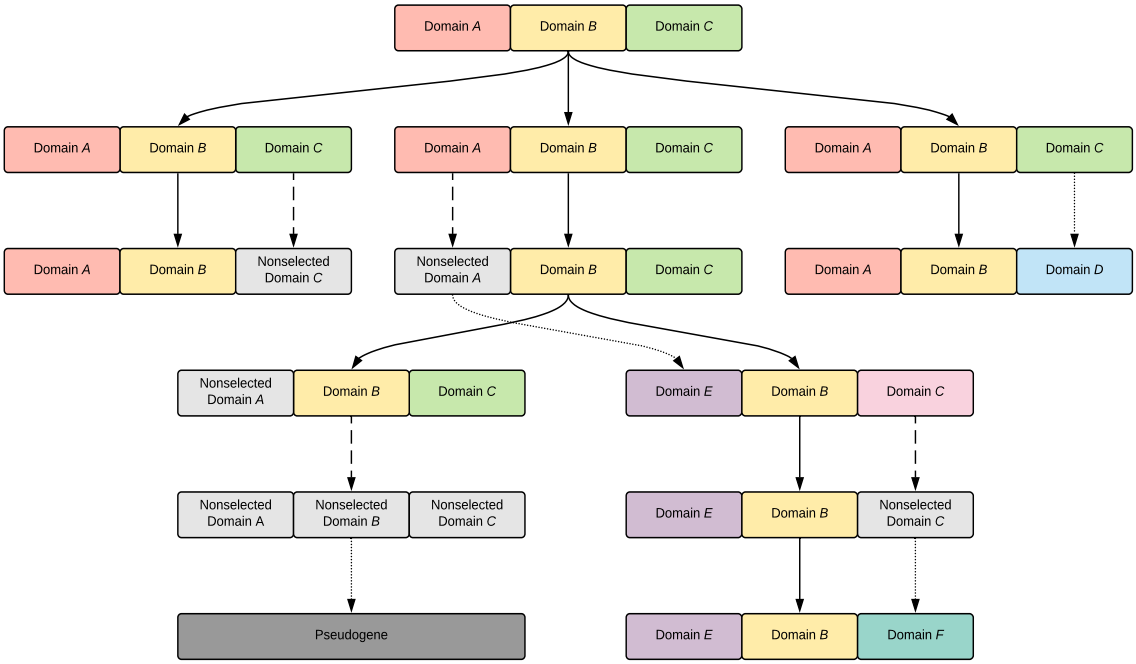

Figure 1.1.2. Paralog fates

Each box represents a conserved gene domain and each domain triplet represents a 3-domain gene (*eg*. ABC and EBF). Each descending level represents a duplication event. A solid line between genes represents vertical inheritance between a parent and paralog, a broken line (dashed) represents loss of selection (ns), and a broken line (dotted) represents a change in function. Loss of selection in domains A and C following the second duplication event (resulting in $A_{ns}BC$ and ABC_{ns}) represents a **subfunctionalisation event** (functional partitioning), and the change in function (ABC→ABD) represents a **neofunctionalization event** (evolution of functional novelty). The loss of selection (and subsequent loss of function) exhibited during the third and fourth duplication events ($A_{ns}BC \rightarrow A_{ns}B_{ns}C_{ns}$) represents a **pseudogenisation event** (complete loss of function). The development of functional novelty in a subfunctionalised gene ($EBC \rightarrow EBC_{ns} \rightarrow EBF$) as exhibited during the third and fourth duplication events represents a **neosubfunctionalisation event** (the evolution of functional novelty in functionally partitioned paralogs).

In cases where paralogs are retained, but not functionally conserved, one of three scenarios arise (Figure 1.1.2.):

(*i.*): **Pseudogenisation** (Levasseur and Pontarotti, 2011) where a paralog accumulates successive point mutations rendering it functionally deficient resulting in its transition to a non-coding gene;

(*ii.*): **Neofunctionalisation** (Xia *et al.*, 2016), where one paralog retains its original function and the other, being under less selective pressure evolves a new function through the accumulation of beneficial mutations or through the convergent evolution of a conserved domain, resulting in a "false fusion" event (a gene that appears to be a fusion of two ORFs but did not arise from such) such as an epaktologous event (Nagy and Patthy, 2011). Fusion events are a prominent subset of remodelling events (Leonard and Richards, 2012) and are discussed in detail in *subsection 1.1.1.2.1.*; or

(*iii.*): **Subfunctionalisation** (Kleinjan *et al.*, 2008), where paralogs of multifunctional genes undergo functional partitioning resulting in differential expression. Ancestral, pre-subfunctionalised genes often contain more than one catalytic domain (Bashton and Chothia, 2007), resulting in selective pressure being applied to only one portion of each paralog.

As such, successive deletion events are free to take place within non-selected regions, resulting in a "one-sided fission" event if this further results in a partial gene deletion (Des Marais and Rausher, 2008). Fission events, like fusion events, are discussed in their own

section later in this chapter (*subsection 1.1.2.1.*). It is also possible for neofunctionalisation to occur in subfunctionalised genes, ultimately cumulating to **neosubfunctionalisation** (Rastogi and Liberles, 2005). Neofunctionalisation or pseudoegenisation are hypothesised to be the most likely fates of non-conserved paralogs (Wagner, 1998; Lynch and Conery, 2000). Neosubfunctionalised genes lend credence to the theory that subfunctionalised genes are intermediates in protein evolution and are rarely retained. Due to the considerable variation in neosubfunctionalised genes from their original paralog, it may be possible that these genes may only share "hidden homology", and, as such, may be undetectable using sequence similarity searches (Janeček, 2008).

1.1.1.2.1.1. Paralog neofunctionalisation

Neofunctionalisation refers to the evolution of functional novelty in one paralogous gene while the other paralog retains its individual function as a consequence of selective pressures. Neofunctionalisation is hypothesised to follow one of two evolutionary models (Figure 1.1.3.):

(*i.*): <u>The "**mutation during redundancy**" (MDR) model</u>

The MDR model (He and Zhang, 2005) strictly states that differential selective pressures are exerted on each paralog following a duplication event. Specifically, one paralog is under selective pressure to retain its original function and the other is under much less pressure and is free to accumulate mutations. MDR mediated neofunctionalisation is hypothesised to be rare due to the fact that deleterious or silent mutations are much more

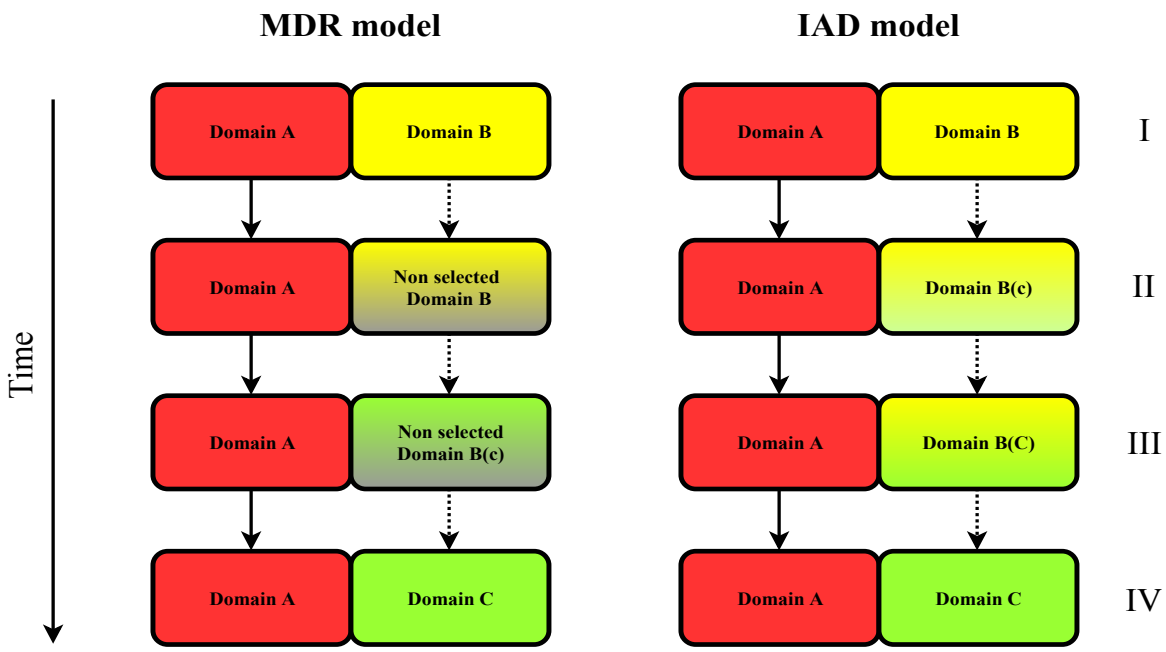

Figure 1.1.3. Comparison of MDR and IAD neofunctionalization models over time

Each box represents a domain and each domain pair represents a paralog undergoing neofunctionalisation (Gene AB→AC). Each level represents a timepoint (annotated by Roman numerals I-IV). A solid line between domains represents functional conservation and a broken (dashed) line represents functional divergence. A colour transition to green represents beneficial mutations leading to functional novelty and a transition to grey represents loss of selection (and subsequent loss of function). For illustrative purposes, Domain A is functionally retained at each time point in each model further emphasising the divergence of Domain B→C. For the MDR model, Domain B undergoes loss of selection and begins to acquire beneficial mutations resulting in functional novelty (which is selected for), eventually becoming the dominant function (Gene AC). For the IAD model, enzyme promiscuity is observed (indicated by the green colour transition without first undergoing a loss of selection), where the secondary function is selected for, eventually becoming the dominant function (Gene AC).

frequent than beneficial mutations and that these accumulations would have to be large scale and localised (Rastogi and Liberles, 2005), which is statistically unlikely to occur compared to the accumulation of deleterious mutations thus leading to pseudogenization (Loewe and Hill, 2010); and

(*ii.*): The "**innovation, amplification, and divergence**" (IAD) model

Conversely, the IAD model (Andersson *et al.*, 2015) postulates that novel functions arise alongside original function through enzyme promiscuity, the phenomenon where enzymes catalyse a secondary reaction in addition to their main catalytic reaction. As the rate of beneficial promiscuity increases, the secondary function is further amplified through successive duplication events (Weng, 2014). IAD mediated evolution is most commonly associated with plasmids where it has been observed experimentally by Andersson and company (2015). Promiscuity tends to polarise after a number of duplication events, thus leading to distinct functions between paralogs (Weng, 2014). Briefly, the key difference between the MDR and IAD models is that MDR mediated neofunctionalisation emerges through non-selection of a paralog which allows for the progressive accumulation of beneficial mutations over time, whereas IAD mediated neofunctionalisation emerges through catalytic domain promiscuity, where futher duplicates are selected for one promiscuous function resulting in selective divergence over time.

1.1.1.2.1.2. Paralog subfunctionalisation

Subfunctionalisation is the evolution of functional partition within multifunctional paralogs (Rastogi and Liberles, 2005). As with neofunctionalisation, subfunctionalisation is also hypothesised to arise from one of two models:

(*i*): The "**duplication, degeneration, and complementation**" (DDC) model

The DDC model states non-selected paralogs may accumulate mutations reducing their capacity to perform a particular function (Figure 1.1.4.). This lack of functionality is complemented by the other paralog, resulting in a specialist partitioned function for the degenerated paralog. As the multifunctional paralog is no longer required to carry out the function of the degenerated paralog, it also evolves towards subfunctionalisation due to a lack of selection for the secondary function (Force *et al.*, 1999; Hellsten *et al.*, 2007). In this model subfunctionalised paralogs functionally complement the pre-duplication multifunctional gene but may now freely evolve their own expression patterns and undergo further duplications and modifications without disrupting its complement (assuming these events would not be deleterious due to gene dosage toxicity).

(*ii.*): The "**escape from adaptive conflict**" (EAC) model

In contrast to the DDC model, the EAC model initiates with evolution towards simultaneous functional promiscuity prior to a duplication event (Des Marais and Rausher, 2008). In scenarios where functional optimisation is unlikely to be achieved for both functions in a single gene (due to, for example,

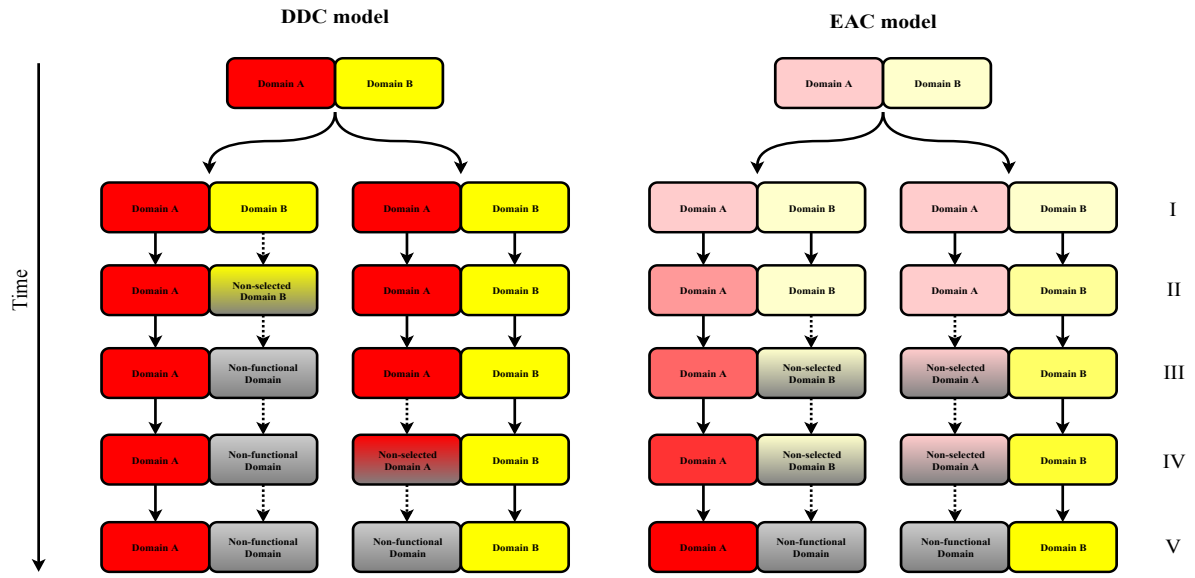

Figure 1.1.4. A comparison of DDC and EAC subfunctionalisaton models over time

Each box represents a domain and each domain pair represents a paralog undergoing subfunctionalisation (Gene AB→A and C). Each level represents a timepoint (annotated by Roman numerals I-V). A solid line between domains represents functional conservation and a broken (dashed) line represents functional divergence. A colour transition to grey represents loss of selection and function. Transitions from dullness to vibrancy (in hue) represents a selective pressure increase. In the DDC model, Domain A is conserved in one paralog and Domain B in the other. In this scenario, one paralog underwent loss of selection relatively early post-duplication, and was thus only expressed to perform one function (as indicated by the transition to grey in Domain B). Consequently, the second paralog was selected for its alternative function (Domain B) and displayed lack of selection for Domain A (as indicated by the transition to grey) resulting in two distinct, subfunctionalised paralogs. In the EAC model, one domain in each paralog evolves towards an optimized function (as represented by increases in vibrancy), a loss of selection is observed in the other domain (as indicated by the transition to grey) resulting in two distinct paralogs with complementary domains, and subsequent optimised complementary functions.

substrate competition or structural hindrances), a duplication event followed by a subfunctionalisation event may be advantageous, ultimately resulting in the same outcome as the DDC model. An example of subfunctionalisation can be observed in RNAse1 (*RNS1*) of the primate genus *Pygathrix* (Zhang, 2003). *Pygathrix* spp. have adapted to predominant folivory which is atypical of the capabilities of Old-World primate metabolism. In most primates, *RNS1* is a bifunctional enzyme, assisting in viral defence and in nutritionally derived nucleic acid degradation, and optimally functions at neutral pH (Lomax *et al.*, 2017). As *Pygathrix* have evolved folivory in their recent evolutionary history, they lack a specialised genetic arsenal for leaf degradation, relying on a low stomach pH to degrade cellulose (Zhang, 2003; Liu and Wang, 2016). As *RNS1* displays severely hampered functionality at low pH levels, a duplicated *RNS1* gene was positively selected for increased functionality in acidic environments. Both paralogs evolved to optimise a function for viral defence at neutral pH or for the metabolism of nucleic acids at a lower pH cumulating to an EAC mediated subfunctionalisation event (Liu and Wang, 2016).

1.1.1.2.2. Duplication mechanisms

As gene duplication leads to functional redundancy, which may subsequently lead to functional, genotypic, and phenotypic novelty, understanding the underlying genetic factors leading to duplication is of importance, especially those leading to large-scale genetic redundancy.

Gene duplication events are commonly ascribed to transposable elements (Marburger *et al.*, 2018), slipped strand mispairing (Levinson and Gutman, 1987), ectopic recombination

(Petrov *et al.*, 2003), polyploidy (Clark and Donoghue, 2018), and aneuploidy (Leitch and Leitch, 2008). Variances in ploidy are common in fungi (Albertin and Marullo, 2012) and plants (Tank *et al.*, 2015), as such these were deemed to be excellent candidates to investigate for modular genetic evolutionary patterns due to their propensity to introduce vast amounts of redundant genetic material over short time frames. These studies are the topics of Chapters III and IV.

Whole genome duplication (WGD) or polyploidization is a massive evolutionary driving force. Evidence of WGD has been observed throughout eukaryotes (Macqueen and Johnston, 2014; Ren *et al.*, 2018), and especially in Viridiplantae (Garsmeur *et al.*, 2014), where their influence is particularly prevalent in grass genomes (Thiel *et al.*, 2009). Due to the considerable genetic redundancy, subsequent subfunctionalisation, and neofunctionalisation arising post-WGD, it is implicit in many niche adaption and speciation events (Fawcett *et al.*, 2009).

Polyploidy refers to one of two genomic states, **autopolyploid** (where extra chromosomes are intergenomic in origin) or **allopolyploid** (where extra chromosomes are intragenomic in origin; Figure 1.1.5.). Allopolyploids arise from hybridization events between closely related species, where copies of both parental genomes are maintained. A prominent example of this is observed within the allohexaploid genome of *Triticum aestivum* (common wheat) where three distinct ancestral genomes are maintained (Blanc and Wolfe, 2004). The evolution toward wheat allohexaploidy is believed to have arisen through two distinct hybridization events, where the origin of each ancestral genome has been identified (Petersen *et al.*, 2006). In the first instance, a hybridization event occurred between two diploid species, *Triticum uratu* (einkhorn wheat) and *Aegilops speltoides*, resulting in the emergence of *Triticum turgidum* (durum wheat), an extant crop species harvested for pasta production

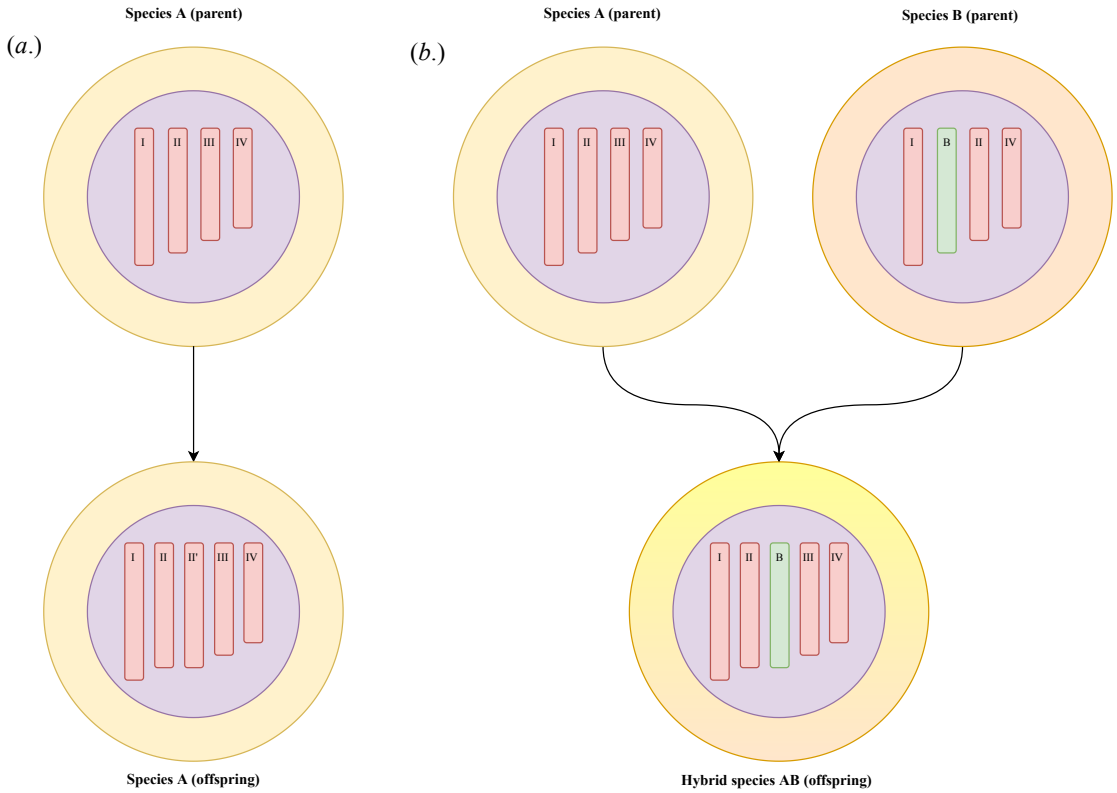

Figure 1.1.5. Comparison of auto- and allopolyploidy

Each orange or yellow circle represents a cell and each purple circle represents its nucleus. Within each nucleus are a series of bars, representing chromosomes. Chromosomes shared by species A and B are red and chromosomes unique to species B are green. Duplicated chromosomes are annotated with a prime (*eg.* II'). Scenario (*a.*) depicts autoploidy, where an extra copy of a chromosome is inherited during duplication, in this case, chromosome II is duplicated (II→{II, II'}). Such scenarios may arise from replication errors such as chromosomal segregation errors (Potapova and Gorbsky, 2017). Scenario (*b.*) represents alloploidy, the inheritance of new chromosomal permutations *via* sexual reproduction between closely related (yet distinct) species. Comparatively, chromosomal copy number increases arising from sexual reproduction between two organisms from the same species is known as polysomy (not shown).

(Kerby and Kuspira, 1987). A second hybridization event between the allotetraploid *T. turgidum* and diploid *Aegilops tauschii* (Tausch's goatgrass) resulted in the emergence of *T. aestivum* (Gill *et al.*, 1991; Petersen *et al.*, 2006).

Comparatively, autopolyploidy is often associated with **polysomy**, the addition of at least one extra chromosomal duplicate from a parent of the same species (Soltis *et al.,* 2014). Incorrect gamete chromosome reduction results in a preserved diploid state. Specifically, autopolyploidy arises through diploid gamete fertilization, resulting in autotriploidy, arising from one diploid gamete and one haploid gamete, or autotetraploids, arising from two diploid gametes (Parisod *et al.*, 2010).

1.1.2. Mechanisms of gene evolution

1.1.2.1. Point mutations

Point mutations (single base pair substitutions, insertions, or deletions) are the most common mechanism of gene evolution, usually arising from erroneous DNA replication (Frömmel and Holzhütter, 1985). Such errors may be spontaneous or may be the result of an environmental interference such as ultraviolet waves or the presence of reactive oxygen species (Fitch, 1967; Livnat, 2013). A point mutation results in one of five outcomes depending on their location within a codon (Thuriaux *et al*, 1982; Friedlander *et al.*, 1983; Hervás-Aguilar *et al.*, 2007; Barrick *et al.*, 2009):

(*i.*): **Silent mutations** do not phenotypically alter an organism as they do not change the codon encoded amino acid, however, subsequent mutations could be deleterious;

(*ii.*): **Nonsense mutations** result in the generation of a stop codon within a gene resulting in a truncated gene product, an X-linked **nonsense mutation** in the human dystrophin gene (*DMD*) results in the onset of Duchenne muscular dystrophy (Gardner *et al.*, 1995);

(*iii.*): **Conservative missense mutations** result in an amino acid substitution to an amino acid with similar chemical properties. In such instances, function may be conserved unless the mutation occurs within a highly conserved gene;

(*iv.*): **Non-conservative missense mutations** result in an amino acid substitution that drastically alters the chemical properties of their site. These mutations are often deleterious, for example, a non-conservative missense mutation at position 20 (A→T) of the β-globin gene (*HBB*) results in the onset of sickle cell anemia (Li *et al.*, 2011); and finally,

(*v.*): **Frameshift mutations** which arise through a nucleobase insertion or deletion event, resulting in codon triplet disruption and the translation of a divergent protein sequence, as such frameshift mutations can be highly deleterious (Zhang *et al.*, 2018). An example of such a mutation is observed in a subset of cystic fibrosis patients where a deletion of a thymine residue at position 1213 of the transmembrane conductance regulator gene (*MRP7*), presents as a misfolded protein product (Lannuzzi *et al.*, 1991).

While point mutations can be highly deleterious, they can also result in advantageous structural or functional novelty (Zakharova *et al.*, 1999). The accumulation of point mutations drives selection towards a neofunctionalised or subfunctionalised state. A single missense mutation can also be highly beneficial. An example of this can be observed in the evolution of antimicrobial resistance. A missense mutation (Asp→Asn) at position 87 of a gyrase gene *gyrA* in clinically relevant gammaproteobacteria leads to fluoroquinolone resistance (Willmott and Maxwell, 1993).

1.1.2.2. Gene remodelling

Considering the sheer extent of genetic variation observed throughout life, it is highly improbable that all variances can be attributed to successive accumulations of point mutations within gene lineages (Haggerty *et al*., 2012; Jachiet *et al.*, 2014 Pathmanathan *et al.,* 2018). It is likely that the process of **gene remodelling** played a significant role during the evolution of complexity. Gene remodelling refers to the rearrangement and modification of existing genes forming new genes from processes such as gene fusion and fission (Leonard and Richards, 2012), domain, and exon shuffling (van Rijk and Bloemendal, 2003; Morgante *et al.*, 2005), and *de novo* generation (McLysaght and Guerzoni, 2015). As gene remodelling is the central theme of this thesis, it is imperative to define specific terms used in remodelling description (Figures 1.1.6.-7.):

(*i*.): A gene that shares significant similarity across a portion of its sequence length with two (or more) unrelated gene sequences is known as a "**composite"** (as it is composed of other, distinct sequences).

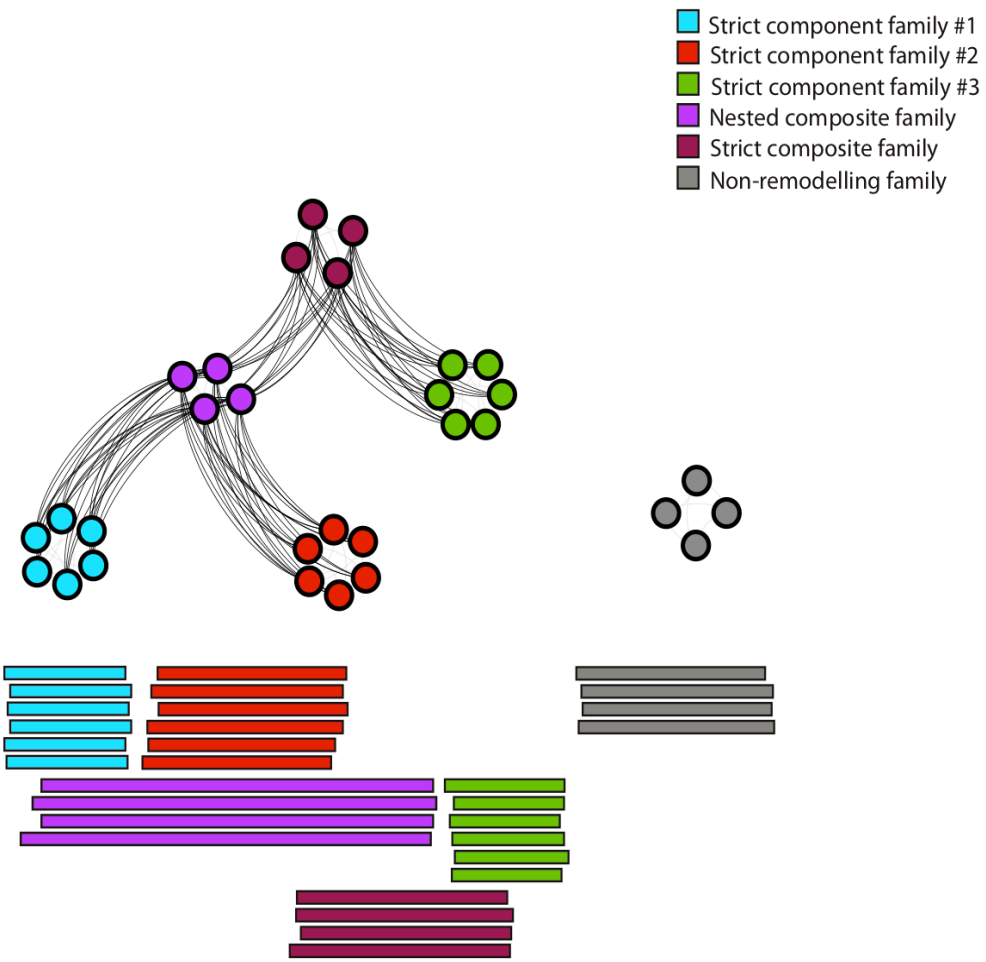

Figure 1.1.6. Remodelling categories (network format)

Each vertex represents a gene and each cluster of genes (annotated with one of six colours) represents a gene family. Edges represent a "remodelled relationship" (homology *via* remodelling) between genes in different families. Bars at the bottom of the figure represent BLAST alignments between each gene with respect to every other gene. Alignments from genes of the red and blue **strict component** families constitute significant distinct portions of genes of the purple **nested composite** family (and *vice versa*). Similarly, genes of the purple and green families display considerable homology with distinct portions of the mahogany **strict composite genes**. Grey genes are **non-remodelled**.

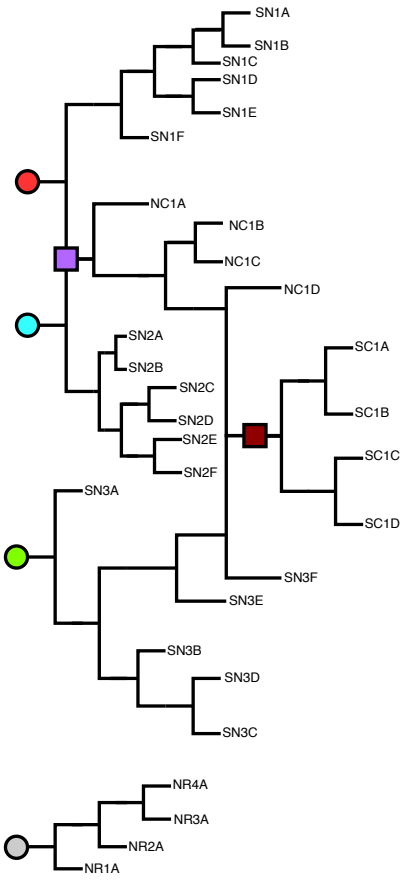

Figure 1.1.7. Remodelling categories (phylogeny format)

This diagram contains two phylogenies, the upper phylogeny represents one of many different topologies that could be inferred from the remodelled gene network in Figure 1.2.6. In this phylogeny, each circle represents the birth of a strict component family and all genes emerging after this event are included within the family. Boxes represent gene fusion events (fusion families) and all genes emerging afterwards are members of that family. Colour annotations are consistent with Figure 1.2.6. (blue, red, and green represent struct components; purple represents a nested composite; and mahogany represents a strict composite). The lower phylogeny represents a non-remodelled gene phylogeny and is included for completeness.

(*ii*.): The terms **"fusion"** and "**fission"** describe a composite gene with a clear, directional evolutionary history (either two genes merged together to form one gene or one gene split to form two genes). The term composite is bidirectional and is used to avoid incorrectly referring to a pre-fission gene as a fusion or *vice versa* (Figure 1.1.8.).

(*iii*.): A "**component"** gene is a sequence detected as constituting a significant portion of a composite gene.

(*iv*.): A "**remodelling event**" is the scenario leading to the formation of composite or component genes.

(*v*.): A gene that is identified as a composite in one remodelling event and as a component in another is known as a **"nested composite"** (**NC**).

(*vi*.): A composite gene that is not further remodelled (is not detected as a component gene within another remodelling event) is a "**strict composite**" (**SC**).

(*vii*.): A component gene that is not further remodelled is a "**strict component**" (**SN**).

(*viii*.): A gene that is not detected as having a history of remodelling is "**non-remodelled**" (**NR**); and

(*ix*.): NC, SC, SN, and NR categories are "**remodelling categories**" (**RC**).

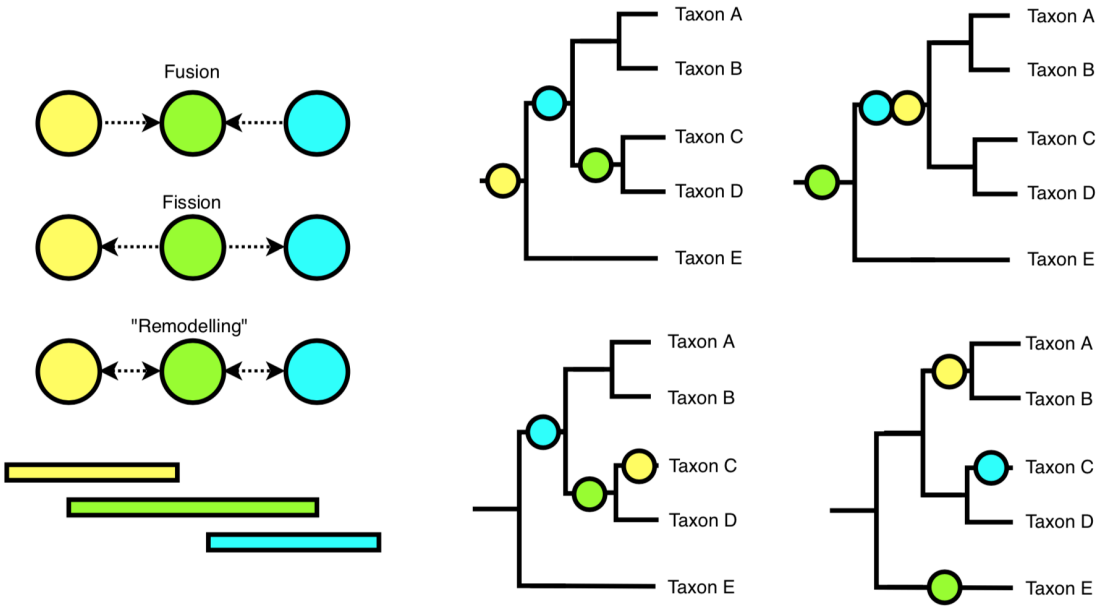

Figure 1.1.8. Comparison of remodelling event types

The networks on the left of the diagram represent three types of remodelling event, where yellow and blue circles represent component genes and the green circle represents a composite. The bars below the network represent sequence alignments between the three genes. Annotation of remodelling events as “fusion” or “fission” requires information on component emergence (directed graphs (discussed in section 1.2.)) whereas annotation as “remodelled” does not (undirected graph). The upper left phylogeny depicts a fusion event as both components emerged prior to the formation of the composite. The upper right phylogeny depicts a fission event as both components emerged after the composite gene. The lower trees do not provide enough evidence to support annotation as “fusion” or “fission”. As annotation as “remodelling” does not require graph directionality, all four events can be accurately annotated as remodelled.

Gene remodelling refers to the process of gene evolution beyond just cumulative mutations. In addition to point mutations, remodelling mediated evolution *via* fusion, fission, and shuffling of domains, exons, and fragments gives rise to vast complexity in emerging composite and component genes (Leonard and Richards, 2012; Pathmanathan *et al.*, 2018). Composite and component genes therefore may be homologous to a wide array of unrelated gene families (Haggerty *et al.*, 2012). Remodelling analyses have been completed on small datasets due to the restrictive computing power and time restraints disentangling such extensive similarity until the advent of CompositeSearch (Pathmanathan *et al.,* 2018).

1.1.2.2.1. Gene fusion and fission

Gene fusions possess a defined evolutionary direction where two previously separate open reading frames (ORFs) merge to form a new composite ORF (Leonard and Richards, 2012). Fusion events arise through local (gene mediated) or global (chromosome mediated) mechanisms. The most simple fusion mechanism is the degeneration of a stop codon, resulting in a readthrough of two ORFs as one. Interestingly, gene fusion events often arise through the same "local" mechanisms as duplication, including *via* transposable elements (TE; Bennetzen, 2005). TEs often erroneously insert captured gene sequences into new genomic positions during their own replication processes. If a sequence is inserted into the promoter region or if the resultant protein is non-functional, both the duplicate and target region would be functionally negated (Kapitonov and Jurka, 2001). If a TE inserts a gene into a non-disruptive region, an effective duplication event would have taken place (Lisch, 2013). However, if a captured sequence is inserted into the coding region of a sequence without disrupting transcription machinery or the promotor region, and assuming the stop codon of the donor sequence or target region does not prematurely terminate the readthrough, a fused gene would

be generated (Bennetzen, 2005; Figure 1.1.9.). While TEs are implicit in the evolution of complexity, a misplaced insertion may result in considerable loss of fitness, especially if a signalling pathway is disrupted (Koivunen *et al.*, 2008; Tamura *et al.*, 2016). With the general exception of plants which suffer less consequence (Feschotte, 2002; Nystedt *et al.*, 2013), eukaryotes implement a number of TE mediated misplacement safeguards such as siRNA silencing of TE transcripts (Poetsch *et al.*, 2018). TEs have been identified as the aetiology for haemophilia by insertion mediated truncation of the Factor VIII blood clotting gene, *hemA* (Kazazian *et al.*, 1988). Due to the positional restraints and safeguards against TEs, it is likely that these are relatively rare facilitators of fused gene formation.

Comparatively, chromosomal events are more likely contribute more fused genes to a species' repertoire, especially in lineages that are prone to polyploidisation such as fungi and plants (Nakamura *et al.*, 2007; Leonard and Richards, 2012). After a polyploidization event, massive chromosomal rearrangements and redundant gene compaction take place (Clark and Donoghue, 2018). This environment promotes the three chromosomal events that promote fusion events (Leonard and Richards, 2012) (Figure 1.1.10.):

(*i.*): **Translocation** of a chromosomal segment on to another chromosome

(*ii.*): **Inversion** of a segment within a chromosome; and

(*iii.*): **interstitial deletion**, the deletion of a chromosomal segment resulting in the merger of two previously non-adjacent chromosomal regions.

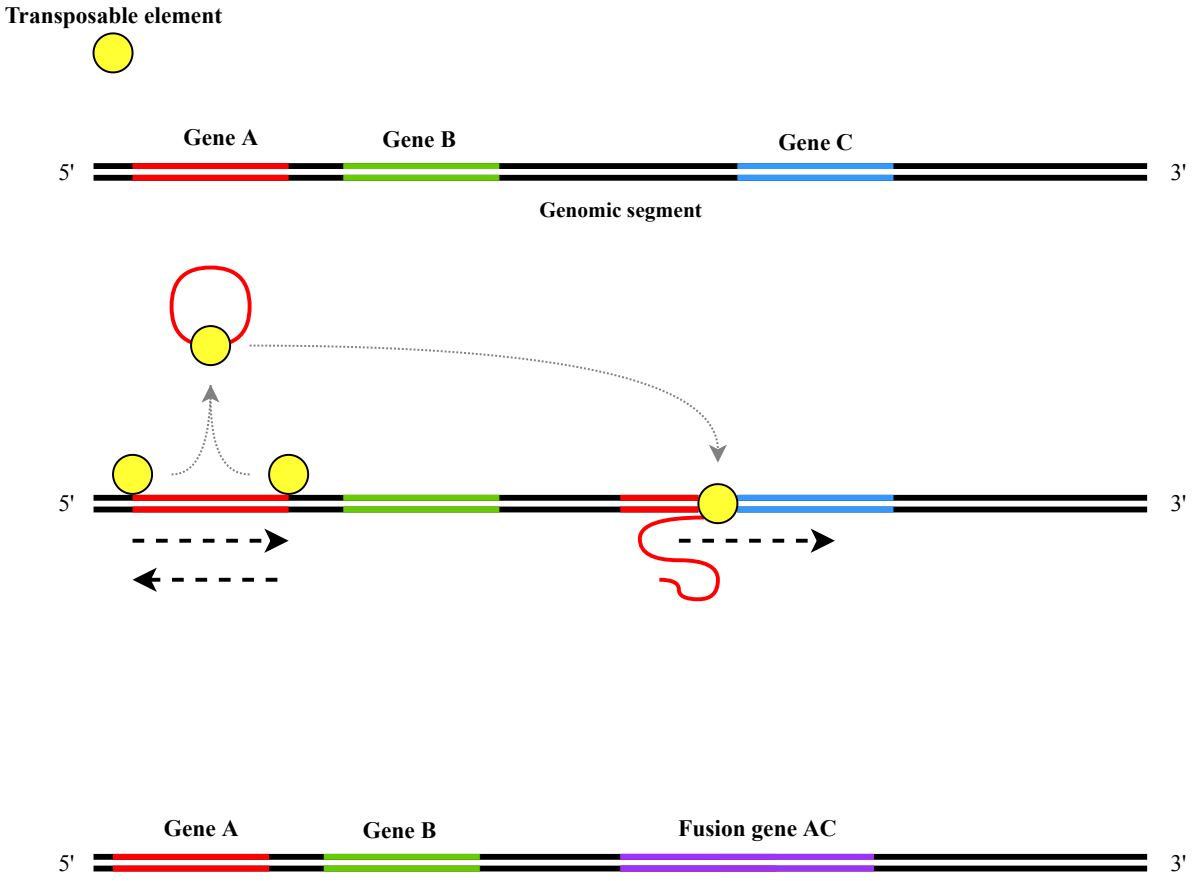

Figure 1.1.9. Transposable element mediated gene fusion

In this scenario, a transposable element (yellow circle) such as a transposon, binds and translocates a copy of Gene A and inserts the duplicated sequence before the N-terminus of Gene C. A resultant missense mutation leading to the degradation of a stop codon allows the readthrough of Genes A and C as a single fused ORF. Black dashed arrows represent TE movement with respect to a captured gene sequence. Grey dotted arrows represent TE movement throughout a genome.

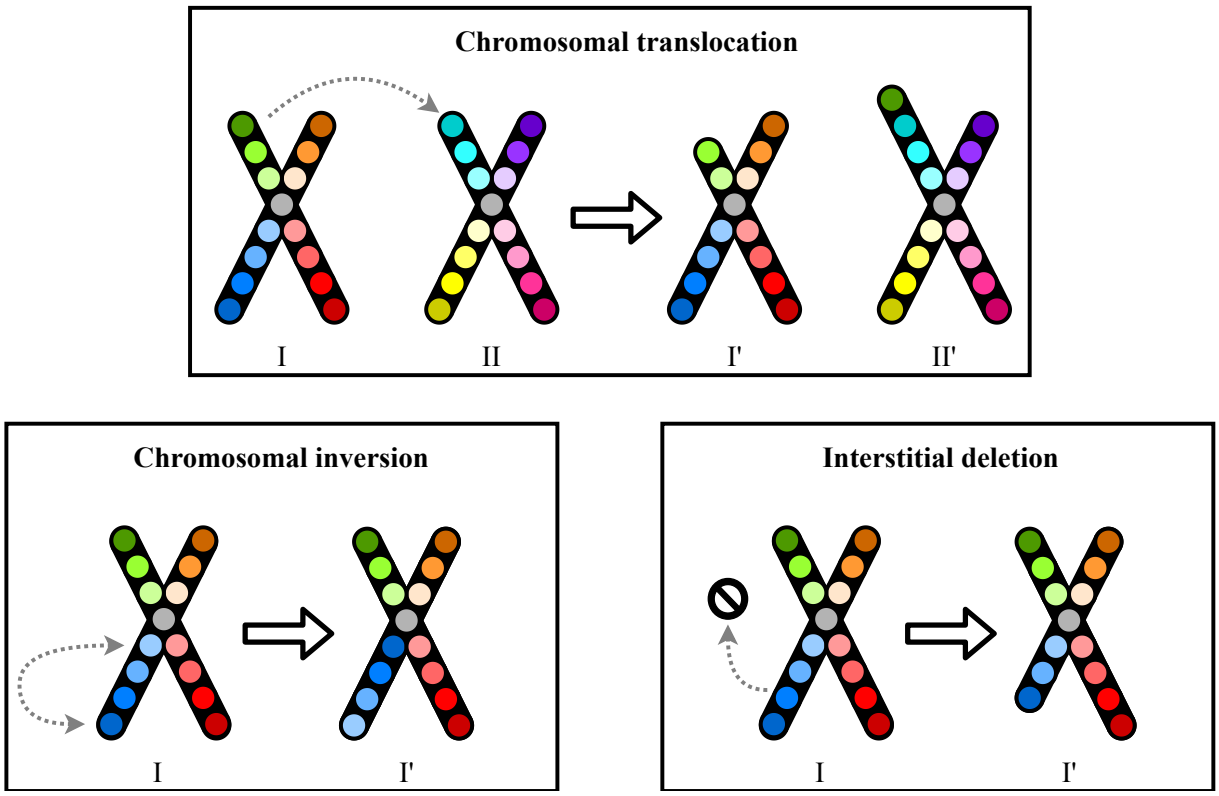

Figure 1.1.10. Chromosomal rearrangements that promote gene remodelling

Chromosomal segments are represented by different coloured circles, movement of segments are represented by grey arrows, and the transition of one chromosomal state to another is represented by a black arrow. Wild type chromosomes (I and II) are drawn to the left of the black arrow, and the resultant aberration (I' and II') are drawn to the right. During **translocation**, the q-arm of Chromosome I is translocated to the end of a q-arm in Chromosome II, resulting in severe truncation and elongation of Chromosomes I' and II' respectively. During **inversion**, a p-arm is inverted in Chromosome I, so what was once the telomere (dark red circle) is now adjacent to the centromere (grey circle), displacing all genes at the extremities of the p-arm. Finally, during interstitial deletion, two interstitial segments are deleted from the p-arm in Chromosome I, resulting in the fusion of telomeric segments to the centromeric segments. All of these events promote gene remodelling by physically disrupting genomic architecture and the splintering of DNA sequences.

Gene fusion events have been identified as the aetiology of many cancers due to the disruption of signalling pathways (Nam *et al.*, 2007). An example of this can be observed in the the oncogenic *BCR-ABL1* fusion which arises through a translocation event (Hochhaus *et al.*, 2011) between chromosomes 9 and 22 (t(9;22)(q34;q11)). Chromosomes displaying this event are known as "Philadelphia chromosomes" and are prevalent in many disease phenotypes, especially leukaemia (Kang *et al.*, 2016). Unfused *ABL1* is a tyrosine kinase involved in mitotic progression and in stress response (Quentmeier *et al.*, 2005). The function of wild-type *BCR* is yet to be fully elucidated. While gene fusions have been attributed to the emergence of disease phenotypes, evolutionary history is replete with hallmarks of advantageous fusion events (Enright and Ouzounis, 2001; Richards *et al.*, 2006). Fusion events have also been implicit in the rapid development of antimicrobial resistance in clinical settings (Williams *et al.*, 2005; Coleman *et al.*, 2015). The functional prowess of fusion (composite) genes are discussed in detail in Chapters II and III, and the extent of fused AMR genes are discussed in Chapter IV. A prominent example of a gene fusion event leading to a highly advantageous adaption can be observed in rhodopsin-guanyl cyclase (RhGC) gene within the Blastocladiomycota (Avelar *et al.*, 2014). In their unfused states, rhodopsin functions as a photoreceptor and guanyl-cyclase as a G-protein coupled receptor. In *Blastocladia emersonii*, RhGC is localised in the zoospore flagellar eyespot. In RhGC, light-activated rhodopsin initiates cGMP synthesis. Plasma membrane hyperpolarisation arises from the opening of K^+-selective channels by cGMP resulting in the opening of voltage activated Ca^{2+} channels (Avelar *et al.,* 2014). Increased Ca^{2+} results in increased flagellar beating, resulting in phototactic response. RhGC is essential for phototaxis is *B. emersonii* zoospores.

Comparative to fusions, fissions refer to the splitting of an ancestral ORF into two or more ORFs (Leonard and Richards, 2012). As mentioned in *subsection 1.1.1.2.1.2.*, a one-sided fission may arise through a nonsense mutation along a sequence, resulting in a highly

truncated yet still viable sequence (Leonard and Richards, 2012). Gene fission events are believed to occur much less frequently than fusions due to additional formation requirements (Durrens *et al.,* 2008). In addition to the generation of a stop codon, a second start codon preceded by a promoter region must be appropriated within a gene for a fission to occur (Leonard and Richards, 2012).

1.2. Graph theory

As mentioned previously, evolution *via* gene remodelling is "network-like" as opposed to "tree-like". A "tree" structure refers a subset of graph structures that displays a single root (if any) and is completely devoid of merged nodes. Comparatively, "network-like" refers to a subset of graph structures that has merged nodes and more than one root (Haggerty *et al.*, 2012). In discrete mathematics, graph theory focusses on modelling pairwise relationships between two or more objects. A graph ($G = V,E$) is composed of **nodes** (**vertices**) and the relationships between them **(edges)** and may be **directed**, with asymmetric, orientated edges between vertices, or **undirected**, with symmetric, unorientated edges between vertices. This thesis relies exclusively on undirected graphs (as previous gene remodelling studies have done (Jachiet *et al.*, 2014; Pathmanathan *et al*., 2018)), where each vertex is a gene and edges between vertices are formed through sequence similarity (homology), resulting in a **sequence similarity network** (SSN). A small set of graph theory terminologies are used throughout this this thesis (Figure 1.2.1.):

(*i.*): A **connected component** (not to be confused with "component" genes as discussed in *subsection 1.1.2.8.*) refers to a group of vertices that are connected (*via* edges) to each other but not to any group;

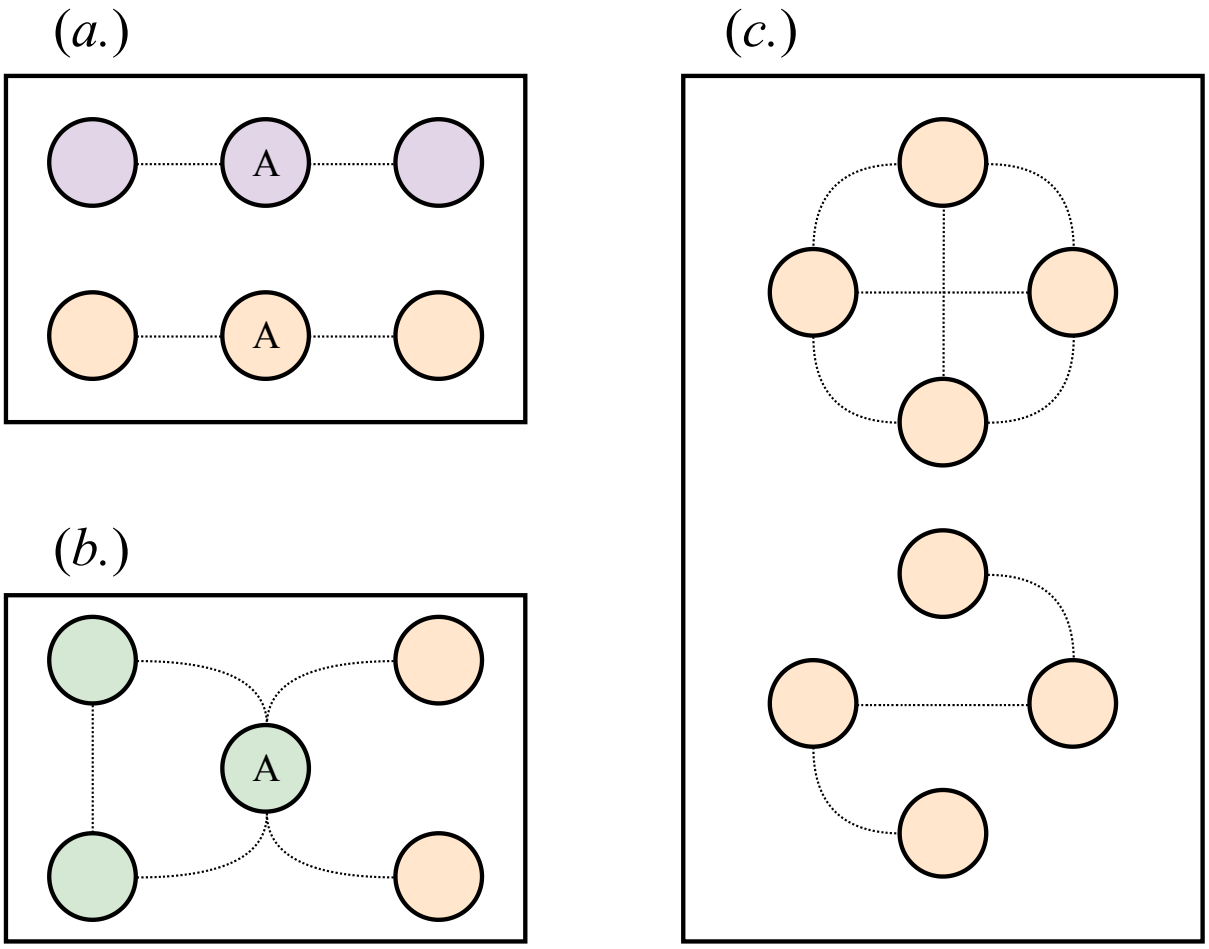

Figure 1.2.1. Graph theory nomenclature

Graph (*a*.) depicts two **connected components**, distinct vertex groups (purple and orange) that are unconnected to each other. Graph (*b.*) illustrates a **subgraph** (annotated by green) within a connected component (non-included vertices are annotated with orange). The upper connected component in Graph (*c*.) depicts a **clique** (where all vertices are connected to all other vertices) and the lower connected component depicts a **quasi-clique** (where all vertices are not connected to all other vertices). For clarity, connected components in Graphs (*a*.) and (*b*.) are also quasi cliques. A **cluster** (or **family**) refers to any clique or quasi-clique in a graph. **Articulation points** (annotated as "A") in Graphs (*a*.) and (*b*.) are vertices that connect two otherwise disconnected clusters.

(*ii.*): A **subgraph** refers to specific nodes and their associated edges within a connected component;

(*iii.*): A **clique** refers to a connected component where all vertices share an edge with all other vertices;

(*iv.*): A **quasi-clique** refers to a connected component where not all vertices share an edge with all other vertices.

(*v.*): A **cluster** (or **family**) refers to both cliques and quasi-cliques

(*vi.*): An **articulation point** is a vertex (or cluster or vertices) that connect two otherwise unconnected clusters.

In their studies, Jachiet *et al*. (2012) and Pathmanathan *et al*., (2018) primarily focussed on viral datasets, however, other network based remodelling studies have investigated the role of remodelling in pathway evolution (Richards *et al*., 2006; Ocaña-Pallarès *et al*., 2019) and as evidence for eukaryogenesis *via* bacterial-archaeal chimerism (Alvarez-Ponce *et al*., 2013).

Homologous relationships (edges) between genes (vertices) in an SSN are easily established using tools such as BLASTP (Altschul *et al*., 1997) or DIAMOND (Buchfink *et al*., 2015), or through the possession of conserved protein domains such as PFAM domains (Finn *et al*., 2008) and InterPro domains (Apweiler *et al*., 2001) as assigned by tools such as InterProScan (Jones *et al*., 2014). All SSNs in this thesis are constructed using BLASTP and the specific criteria for cutting edges are discussed in Section 1.3.2.

1.3. Tools used for remodelled gene detection in large datasets

1.3.1. *fdf*BLAST (Leonard and Richards, 2012)

Prior to the release of CompositeSearch, remodelled gene detection required considerable computational resources required for large scale combinatorial comparisons (Pathmanathan *et al*., 2018). Due to this bottleneck, *in silico* studies on gene remodelling tended to focus on fusion or fission events pertaining to single genes (*eg.* Alvelar *et al*., 2014), on specific biochemical pathways (*eg.* Richards *et al*., 2006), or between two organisms (*eg.* Nakamura *et al*., 2007). In 2012, Leonard and Richards released *fdf*BLAST, the first tool for for high-throughput gene remodelling events. In particular, *fdf*BLAST detects **differential** gene fusions (instances where a fused gene is observed in one genome and both components are observed in another genome). For the purposes of this section, the terms "fusions" and "composites" are used interchangeably. Leonard and Richards applied this technology to nine fungal genomes, where a total of 63 gene fusions were identified.

*fdf*BLAST has five main steps (Figure 1.3.1.):

(*i.*) Comparative reciprocal BLAST

Each genome in a multi-genome dataset is queried against each other genome using BLASTall (Altschul *et al*., 1997) to search for the presence of single copy full length homologs (1:1 relationship), the presence of genes with multiple homologs in another genome (1:2 relationship), and the presence of multiple homologs in a genome that hit a single gene in another genome (2:1 relationship). This step ensures that a bidirectional hit is observed between the

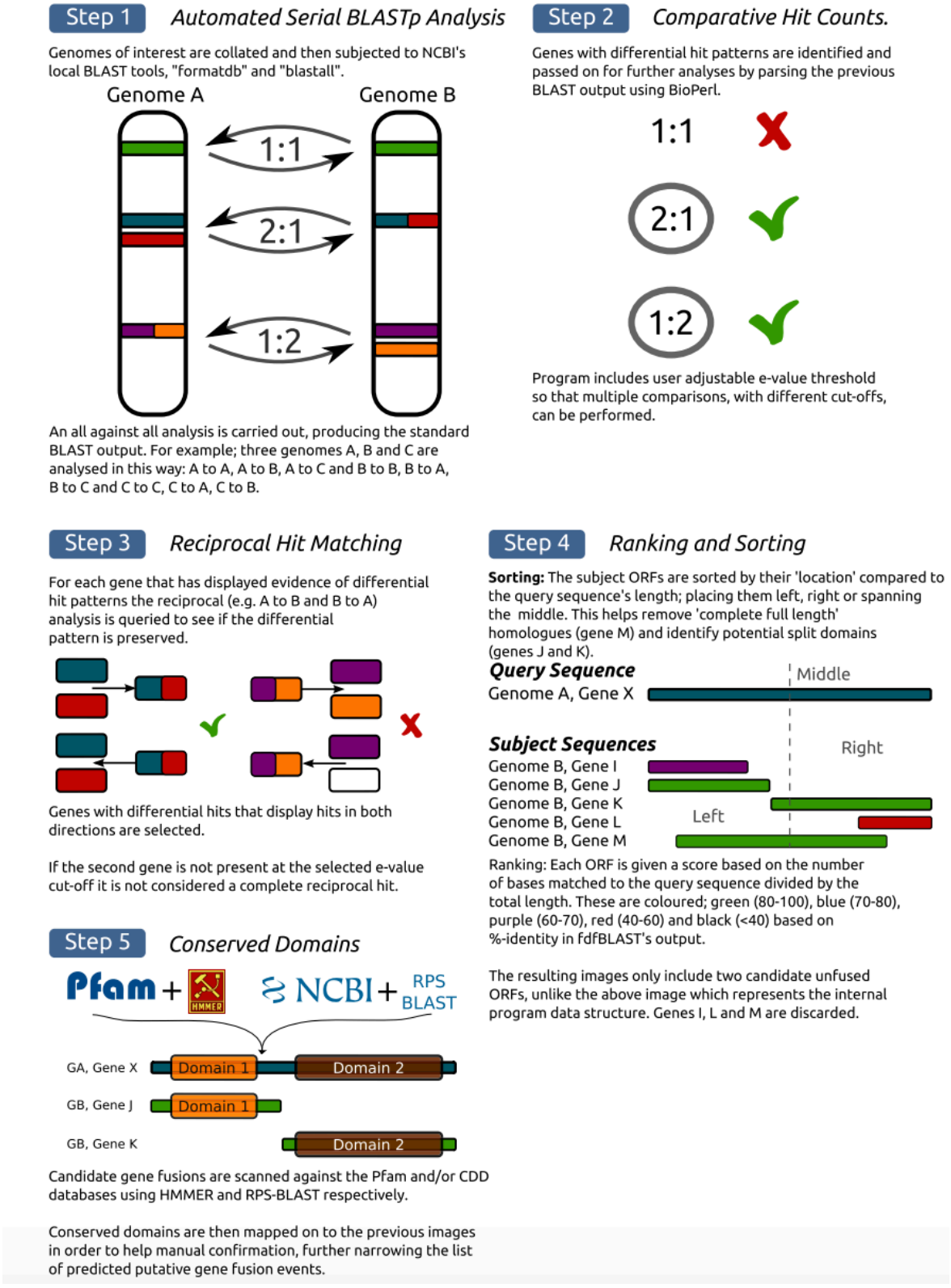

Figure 1.3.1. *fdf*BLAST pipeline

An illustration of the *fdf*BLAST pipeline used by Leonard and Richards (2012).

fusion and each component. If a potential fusion is not found to have bidirectional hits, it is discarded from further analyses.

(*ii*.) Comparative hit counts

All instances of 1:2 and 2:1 relationships are extracted while all 1:1 relationships are excluded from further analyses. 2:1 and 1:2 relationships are selected to detect instances where one gene aligns to two different homologs at different coordinates along its sequence (fused gene sequence). Such instances may then be parsed for differential fusion detection. 1:1 relationships cannot satisfy this requirement, as such they are discarded.

(*iii*.) Reciprocal hit matching

Alignments between each gene in each potential fusion (1:2 and 2:1) relationship are assessed to ensure that hits are bidirectional between both genomes. This is a quality control procedure to ensure confidence in homology.

(*iv*.) Ranking and sorting

A "terminal alignment category" (C-terminal or N-terminal) is assigned to each component based on its positional alignment along the fusion sequence. If the component sequence terminates within the first 50% of the fusion it is annotated (sorted) as "C-terminal terminating" and if it terminates within the final 50% of the gene, it is annotated as "N-terminal terminating". This process

is replicated for the initiation (start) of each component alignment against the fused gene, and each component is annotated as “C-terminal initiating” or “N-terminal initiating”. Instances where genes initiate and terminate within the same terminal are assigned to that terminal (C-terminal or N-terminal aligning) and genes that initiate within the C-terminus but terminate within the N-terminus are annotated as “middle aligning”. If a middle aligning component has a user defined (default = 90%) length similar to the fusion, it is considered to be a full length homolog and is discarded.

Retained components are then ranked based on two separate methodologies:

(*a*.): Percentage identity sequence similarity

The percentage of total identical amino acids (“pident”) mapped to the fusion sequence (where higher percentages are ranked higher); and

(*b*.): Geographical distance ratios between components

The geographical distance ratio (DR) between each N-terminal and each C-terminal component alignment (query sequences) along the fused gene is calculated using the formula:

$$DR = \frac{qend_N}{qstart_C}$$

where:

DR: Distance ratio

$qend_N$: Alignment termination position of the query N-terminal component (qend) against its potential fusion (subject sequence)

$qstart_C$: Alignment initiation position of the query C-terminal component (qstart) against its potential fusion (subject sequence)

For each potential fusion, the DR between all potential C- and N-termini are binned between 0.1 and 1 at increments of 0.1. Distance ratios are inversely correlated with the distances between alignments. For example, a distance ratio of 1 would be observed when two component alignments are adjacent, whereas a distance ratio of 0.1 would be observed when alignments are relatively far apart in relation to the fusion sequence (Figure 1.3.2.).

(*v*.) Conserved domain detection

HMMscan (Johnson *et al.*, 2010) is used to assign PFAM domains (Bateman, 2002) to each component and fusion. A fusion must share domains with each of its components to be considered valid and fusions not possessing conserved PFAM domains are discarded. Graphical outputs are constructed for each fusion/fission event, which may then be manually curated.

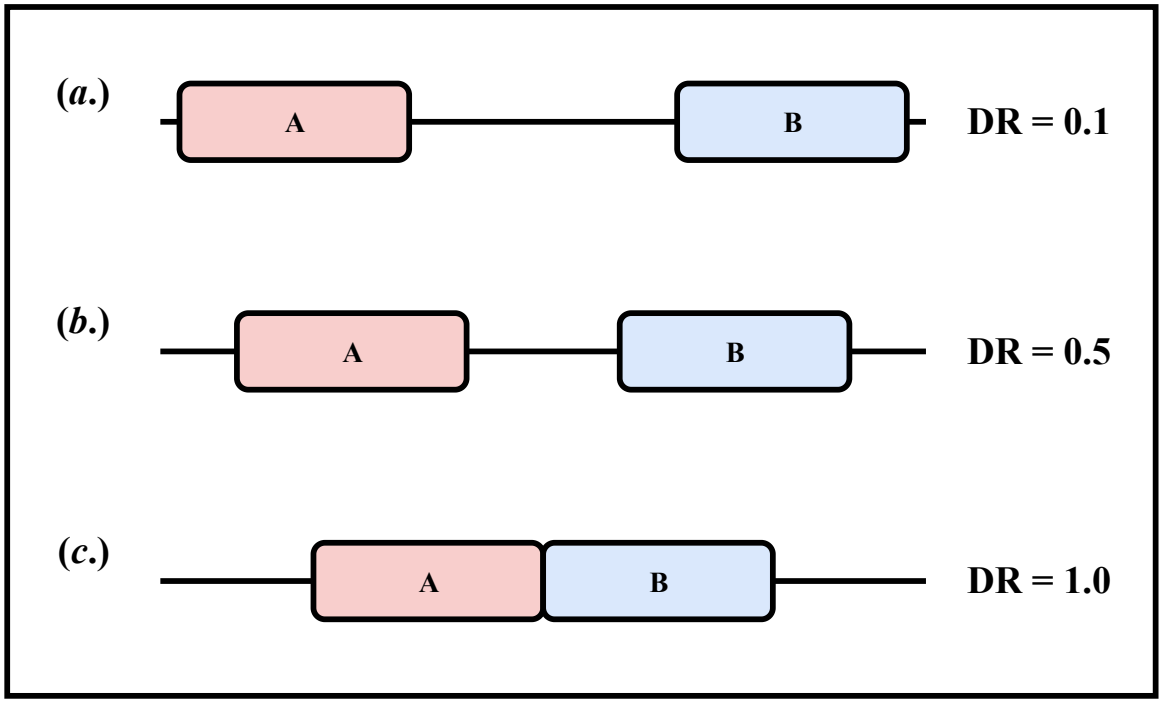

Figure 1.3.2. *fdf*BLAST distance ratios

Each scenario (*a*.-*c*.) represents the distance ratio (DR) calculated for two components (A and B) against a composite gene (solid black line) where DR = [0.1, 0.2, …, 1.0]. The DR is inversely correlated to the distance between domains (Leonard and Richards, 2012)

1.3.2. CompositeSearch (Pathmanathan *et al*., 2018)

To our knowledge, CompositeSearch is the only program for high throughput remodelled gene detection since *fdf*BLAST (Leonard and Richards, 2012) and its predecessor, fusedTriplets (Jachiet *et al*., 2014). CompositeSearch is a high throughput C++ program for detecting gene remodelling events in SSNs constructed using BLASTP (Altschul, 1997). CompositeSearch detects families acting as articulation points (composite families) between at least two otherwise unrelated families (component families) and differs from *fdf*BLAST by searching for all instances of gene remodelling, not just differentially distributed gene fusion events, thus CompositeSearch is a **sensitive** detection tool, whereas *fdf*BLAST is a **selective** detection tool. A CompositeSearch analysis identifies gene remodelling in five steps:

(*i*.): <u>SSN construction</u>

CompositeSearch requires an SSN constructed using BLASTP (tabular format (-outfmt 6)) as input. Low complexity sequences and repetitive sequences can be detected by BLAST as being highly similar; as these potentially non-homologous and relatively common gene architectures could cause a potentially high degree of false positive results, they are excluded from further analyses using BLAST incorporated "SEG" (-seg yes) and soft masking (-soft_masking true) filters. This step is taken as a precaution as the implementation of compositional statistics in the BLAST+ algorithm already aims to reduce influence from such alignments (Pathmanathan *et al.*, 2018). As CompositeSearch requires alignment information in a defined order, the executable BLAST command was applied:

blastp -query <input file> -db <database file> -out <output file> -seg yes -soft_masking true -outfmt “6 qseqid sseqid evalue pident bitscore qstart qend qlen sstart send slen”

where:

-query:	Dataset of query sequences (FASTA format)
-db:	Database of subject sequences (FASTA format)
-out:	Output file
-seg	SEG filter
-soft_masking:	Soft masking filter
-outfmt:	Output file format (6; tabular)
qseqid:	Query sequence ID
sseqid:	Subject sequence ID
evalue:	Expect value (*E*)
pident:	Percentage identity
bitscore:	Bitscore
qstart:	Start of query alignment (to subject)
qend:	End of query alignment (to subject)
qlen:	Length of query sequence
sstart:	Start of subject alignment (to query)
send:	End of subject alignment (to query)
slen:	Length of subject sequence

BLAST flags between quotation marks (6, qseqid - slen) refer to the sequence columns that appear in in the -outfmt6 (tabular) output file.

(*ii*.) "cleanBlastp" executable

The C++ program "cleanBlastp" is provided with CompositeSearch which converts each gene ID (qseqids and sseqids) to a unique integer which removes self-hits, removes redundant hits, and in cases where multiple high scoring pairs (HSPs) are observed, selects the best hit (based on the lowest *e*-value). Three output files are produced by cleanBlastp:

(*a.*): A "**cleanBlast**" file with retained BLASTP statistics (in tabular format) for CompositeSearch network analyses (explained in step (*iii*.).)

(*b*.): A "**cleanBlast.genes**" file containing a list of all converted gene IDs which is used during CompositeSearch analysis.

(*c*.): A "**cleanBlast.dico**" file containing gene IDs and their mapped integer IDs. This file is not used by CompositeSearch but is provided to allow the user to identify genes during downstream analyses.

The "cleanBlast" and "cleanBlast.genes" files contain data from the BLASTP analysis (described in step (*i*.)) which has been optimised for processing by CompositeSearch. This "cleaning" method has been implemented to avoid simultaneous memory access issues and to allow CompositeSearch to be parallelized for rapid analyses (Pathmanathan *et al.*, 2018)

(*iii*.) CompositeSearch executable

CompositeSearch, is used to identify instances of gene remodelling between homologous gene families. The executable is invoked using the command:

./compositesearch -i cleanBlast file -n cleanBlast.genes -m composites
-c 80 -p 30 -e 1e-05 -l 20 -x 2 -y 2

where:

-i:	Input file (cleanBlast file produced in step (*ii*.))
-n:	Input gene file (cleanBlast.genes file)
-m:	Mode, only "composites" was used throughout this thesis
-c:	Mutual overlap filter (default = 80%)
-p:	Minimum percentage identity filter (default = 30%)
-e:	Minimum BLASTP *e*-value filter (default = $E \leq 10$)
-l:	Maximum overlap allowed between component family alignments against a given composite (default = 20 amino acids)

-x Minimum composite family size filter (default = 1; the reasoning behind why this was increased to 2 during all analyses in this thesis is discussed in section 2.2.2.)

-y: Minimum component family size filter (default = 1; the reasoning behind why this was increased to 2 during all analyses in this thesis is discussed in section 2.2.2.)

-t: Number of threads used

As each flag is used during different steps of a CompositeSearch analysis, they are discussed as they arise. The "cleanBlast" and "cleanBlast.genes" files (generated in step (*ii*.)) are processed by CompositeSearch to assign genes to clusters (families) and then to determine the presence of gene remodelling events *via* the identification of clusters acting as articulation points between at least two other clusters.

Genes are clustered into families using a three step method:

(*a.*): Edge detection

An edge is established between two gene (vertex) alignments if the pident ≥ the user defined minimum pident (-p) and if the *e*-value ≤ the user defined maximum *e*-value.

(*b.*): Vertex clustering:

A Depth First Search (DFS) algorithm is used to cluster vertices (with intervertex edges as defined in step (*i*.)) into connected

components (CC) if the mutual coverage score ≥ the user defined minimum (-c).

(*c.*): Family assignment:

As overextended BLAST alignments may incorrectly introduce genes to a family (Mills and Pearson, 2013), a mutual coverage score (S_{mc}) is computed for each potential family (as defined by Pathmanethan *et al*., 2018), and instances where $S_{mc} < 1$ were subjected to Louvian community detection (Blondel *et al*., 2008). The application of Louvian community detection is to prevent potential composites and their components being assigned to the same families. Genes remaining in clusters after DFS and Louvian community detection are considered to be gene families. A connectivity score for each family (C_f) is calculated using the formula:

$$C_f = \frac{(2 \times n_E)}{\left(n_V \times (n_V - 1)\right)}$$

where:

n_E: The sum of edges in a given family; and

n_V: The sum of vertices in a given family

The family detection step produces three output files, "family.edges" (an edge list between vertices in each family), "family.nodes" (a list of vertices (integer gene IDs) assigned to each

family), and “family.info” (a tab delimited file containing family attributes).

Remodelling events between gene families are determined using a two-step method:

(*a*.): Non-familial homology extraction

Instances of non-familial homology (edges between vertices that are not assigned to the same family) are extracted from the “cleanBlast” file. Instances with genes assigned to families with sizes < -x (minimum composite family size) or -y (minimum component family size) are discarded resulting in an non-familial SSN (nfSSN).

(*b*.): Composite detection

Each gene in each family with a size ≥ -x are considered to be potential composites. Each gene in each family with a size ≥ -y are considered to be potential components. For each potential composite, all hits are extracted from the nfSSN and compared. A composite is considered to be a *bona fide* composite if genes from two distinct component gene families map to two distinct

regions of the composite with an overlap ≥ -l (the maximum permitted overlap). If a family contains a single *bona fide* composite it is considered to be a "composite family" and all potential component genes associated with a *bona fide* composite are considered to be component genes, and their families annotated as such (Figure 1.3.3.). Gene families can be both composites and components, such instances are discussed in *subsection 1.1.2.2*.

This step results in the production of two files:

(*i*.): A "**compositesinfo**" file containing mean alignment information for each composite against its associated component gene families; and

(*ii*.): A "**compositefamiliesinfo**" file containing composite family attributes that may be used for downstream processing.

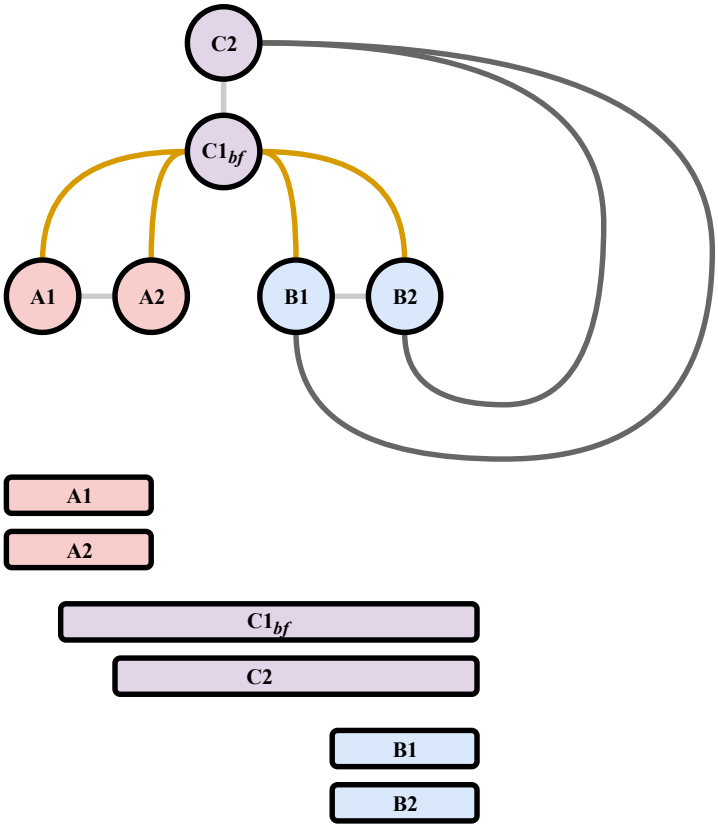

Figure 1.3.3. A *bona fide* composite

The graph illustrates relationships between three gene families (A, B, and C). Families A and B (annotated as red and blue) are component families and family C (annotated as purple) is a composite family. Intrafamilial relationships are annotated as light grey, interfamilial relationships used to infer a remodelling event are annotated as gold, and interfamilial relationships not used to infer a remodelling relationship are annotated as dark grey. Bars below the graph represent BLAST sequence similarity alignments between each sequence. An instance of remodelling was ascertained in this graph due to gene C2 acting as an articulation point (*bona fide* composite) between families A and B. Gene C1 is a truncated homolog of C2 and is assigned to the same family (and therefore a composite), however, as a significant sequence similarity relationship was only detected between C1 and members of family B (but not between family A) it could not serve as an articulation point and is not a *bona fide* composite.

1.4. Statistical tests

This thesis relies heavily on a set of distinct probabilistic statistical methods for the inference of significance which are discussed here prior to their usage in Chapters III-V.

1.4.1. Data comparison

1.4.1.1. Population pairwise comparisons

Differences between two populations are often ascertained by comparing means using a *t*-test (Student, 1908; Welch, 1947) assuming the data follows a Gaussian distribution. However, as data populations in this thesis was anticipated (and later observed) to be highly skewed (*eg.* as observed in gene family sizes (Demuth and Hahn, 2009)), a non-parametric Mann-Whitney *U* test (Mann and Whitney, 1947) was used to compare if two populations were stochastically different. Both population medians are stated to be equal under the null hypothesis (H_O:η_a=η_b;H_A:η_a≠η_b) in a Mann-Whitney *U* test, and the *U* statistic is calculated using the formula:

$$U_a = \sum R_{a,b} - \frac{n_a(n_a+1)}{2};\ U_b = \sum R_{a,b} - \frac{n_b(n_b+1)}{2};\ U = \min(U_a, U_b)$$

where:

a:	Population series *a*
b:	Population series *b*
R:	Rank
n:	Population size

The *P*-value associated with the *U* statistic is ascertained from the normal distribution the corrected Z-statistic ($Z_c(\mu,\sigma^2)$) using the formulae:

$$P = f(Z_c(\mu,\sigma^2)) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(Z_c)^2}{2}}; Z_c = \frac{U - \left(\frac{n_a n_b}{2}\right)}{\sqrt{\frac{n_a n_b}{12}\left((n_a + n_b + 1)\sum_{i=1}^{k} \frac{t_i^3 - t_i}{(n_a + n_b)(n_a + n_b - 1)}\right)}}$$

where:

$f(Z_c(\mu,\sigma^2))$:	Function of the corrected Z-score
e:	Euler's number (~2.71828)
π:	Archimedes' constant (~3.14159)
n:	Population size
i:	Specific rank
k:	Number of distinct ranks
t_i:	Number of subjects (t) sharing a specific rank (i)

1.4.1.2. Comparison of proportions

A Fisher's exact test (Fisher, 1922) is a statistical test used to determine if there are non-random associations, such as similar proportions, between two categorical variables. In general, a Fisher's exact test is a probabilistic test where both proportions (π) are stated to be equal under the null hypothesis ($H_O:\pi_1=\pi_2;H_A:\pi_1\neq\pi_2$) and is calculated using the formula:

$$P = \frac{((X_a+X_b)!\times(Y_a+Y_b)!\times(X_a+Y_b)!\times(X_a+Y_b)!)}{((X_a+X_b+Y_a+Y_b)!\times X_a!\times X_b!\times Y_a!\times Y_b!)}$$

where:

X_a: Subset *a* of set {X} ($a \subset X$)

X_b: Remaining groups (*b*) of set {X} ($b = a^c$)

Y_a: Subgroup *a* of set {Y} ($a \subset X$)

Y_b: Remaining groups (*b*) of set {Y} ($b = a^c$)

1.4.1.3. Comparison of data points to data series

In this thesis, a **data series** is defined as "a set of discrete, real numbers (data points; $n \geq 30$) from a given experimental dataset". Each data point may be compared to its series using a cumulative distribution function (Φ) or one dimensional Gaussian Q-function (Q) (Nikolić, Perić and Marković, 2017). Both functions are tail distribution probability functions of the **standard normal distribution**, and, as Q is derived from Φ ($Q = 1-\Phi$), both are inversely correlated with each other. Under the null hypothesis (for Φ), a given data point (x) is stated to be greater than or equal to a random point on a Gaussian distribution $\Phi(X)$ when sampled from the data series ($H_O:\Phi(x)\geq\Phi(X);H_A:\Phi(x)>\Phi(X)$) and for Q, x is stated to be less than or equal to $\Phi(X)$ ($H_O:Q(x)\leq 1-\Phi(X);H_A:Q(x)>1-\Phi(X)$). In this thesis, we were concerned with "bursts" (data points that were statistically likely to be distributed on the right tail of a Gaussian distribution ($\Phi(x)\geq 0.95;Q(x)\leq 0.05$)), so Q was calculated for each data point during such comparisons. Q was calculated using 3 steps (Figure 1.4.1.):

(*i*.): <u>Determination of normality</u>

As Q requires a sereisto follow a standard gaussian distribution, the data series was subjected to a Kolmogorov-Smirnoff (KS) test ($\alpha = 0.05$)

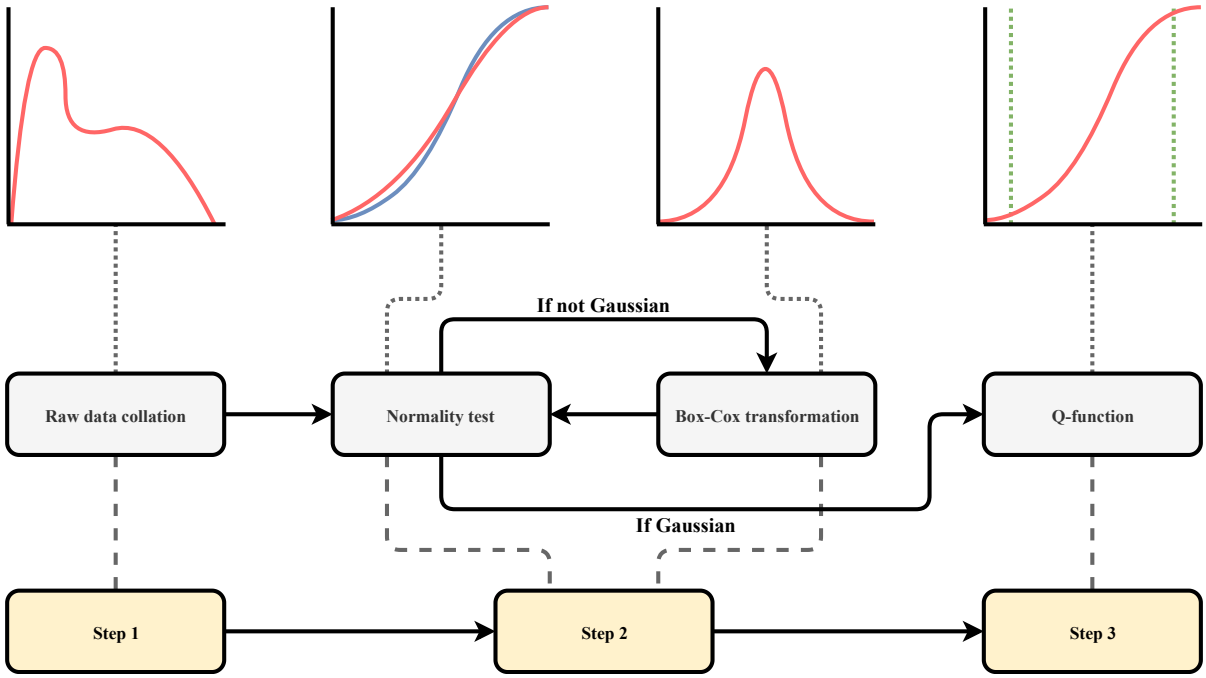

Figure 1.4.1. Q-function pipeline

Each yellow box represents one of three steps in the Q-function pipeline described in *subsection 1.4.2.3.* Each step is connected to its associated process *via* a grey broken line (dashed). Each process is connected to its associated graph *via* a grey broken line (dotted). Graphs were drawn for illustrative purposes and are rough approximations of statistical distribution. The graph associated with step 1 depicts a histogram of raw data (x) connected by a spline. The left graph associated with step 2 (normality test) depicts $g(x)$ (red) and $\widehat{F}(x)$ (blue). The right graph associated with step 2 (Box-Cox transformation) depicts x subsequent to transformation ($\lambda(x)$). Finally, the graph associated with step 3 depicts $g(x)$ with $\Phi(x)$ and $Q(x)$ significance limit boundaries (0.05, 0.95) depicted as green broken lines (dotted).

(Kolmogorov, 1933; Smirnoff, 1948), where a series was considered to follow a Gaussian distribution if $P > \alpha$ was observed. Instances where $P \leq \alpha$ were transformed to a Gaussian distribution using a Box-Cox transformation (Box and Cox, 1985) and normality was redetermined using a KS test ($\alpha = 0.05$).

A KS statistic (T) is calculated using the formula:

$$T = \sup_x \left| g(x) - \widehat{F}(x); g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{Z^2}{2}\right)}; Z = \frac{x - \mu}{\sigma}; \widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq x\}} \right.$$

where:

x:	Series variable
sup:	Supremum
g:	Gaussian function
Z:	Z-score (standard score)
$\widehat{F}$:	Empirical distribution function
$I_{\{x_i \leq x\}}$:	Indicator function for events $(i) \leq x$
n:	Count of elements in a series

A Lilliefors correction (Lilliefors, 1967) is employed to determine that T is sampled from a Gaussian distribution using the formulae:

$$T_{n,\alpha} = \frac{\alpha_{KS}(n)}{f(n)}; f(n) = \frac{\alpha_{KS}(n) + n}{\sqrt{n}} - 0.1$$

where:

n: Sample size

α: Critical alpha (0.05)

$\alpha_{KS}(n)$: α for n from KS α table (Smirnoff, 1948)

As $n \leq 50$ is observed for every invocation of the KS test, $\alpha_{KS}(n)$ was estimated to be 0.895 for each test. Instances where $T > T_{n,\alpha}$ ($P > 0.05$) were considered to be sampled from a Gaussian distribution.

A Box-Cox transformation is performed using the formula:

$$(x+1)_{\lambda} = \frac{(x+1)^{\lambda} - 1}{\lambda}; \lambda = (-5, -4.99, -4.98, \ldots, 5)$$

where:

x: Series variable

λ: Box-Cox power parameter

The mean and standard deviation is computed for each transformed series and the λ yielding the lowest standard deviation is selected as the most appropriate transformation.

(*ii.*): Data standardisation

Each transformed series was standardized by transforming each data point to its *Z*-score (standard score) using the formula described in step (*i*).

(*iii*.): Calculation of Φ and *Q*

The error function (E) was used to calculate Φ for each Z-score in the series using the formulae:

$$\Phi(Z) = \frac{1}{2}\left[1+\mathrm{E}\left(\frac{Z}{\sqrt{2}}\right)\right];\ \mathrm{E}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y} dy$$

where:

x:	Independent variable x; integral upper-bound
y:	Independent variable y
e:	Euler's number (~2.71828)
π:	Archimedes' constant (~3.14159)
dy:	Differential of y

Q was then calculated by subtracting Φ from 1.

Every primary invocation of the KS test in this thesis reported each series as non-Gaussian ($P < 0.05$) and every Box-Cox transformation invoked to rectify this was successful ($P > 0.05$).

1.4.2.4. Correlations between two data series

A multitude of data series can be extrapolated from object characteristics in a given set, for example, series of (*i*.) genome sizes, (*ii*.) the sum of genes in a given genomes, and (*iii*.) the sum of chromosomes can be constructed from a set of given genomes. In this thesis, diverse characteristics are compared across genomic sets ($n \geq 50$). As such characteristics may compare a Gaussian series to a non-Gaussian series, correlations were derived using non-parametric Spearman's ρ correlation matrices (Spearman, 1904). A correlation matrix is a collection ($n > 2$) of pairwise correlation tests (*eg.* Spearman's ρ test) where each matrix cell represents one test. For a Spearman's ρ test each variable is ranked in each series (rank variables), where non-distinct variables are each assigned to the same rank, and ρ is calculated using the formula:

$$\rho = \frac{\sigma_{x,y}}{\sigma_x \times \sigma_y}; \sigma_{x,y} = \frac{\sum\left(\left(x_n\text{-}\mu_x\right) \times \left(y_n\text{-}\mu_y\right)\right)}{|c|_R}$$

where:

$\sigma_{x,y}$:	Covariance between series x and y
$\|c\|_R$:	The count of ranks
x_n:	Data point n is series x

1.4.3. Control for Type I errors

A Bonferroni-Dunn correction (Bonferroni adjustment) is applied to every statistical test in this thesis to control for Type I errors (Bonferroni, 1936; Dunn, 1959; 1961). A Bonferroni correction is an adjustment of the critical α, the maximum permitted *P*-value allowed for the inference of statistical significance. The critical α is adjusted using the formula:

$$\alpha_B = \frac{\alpha}{|c|}$$

where:

α: Critical α (0.05)

$|c|$: The count of all comparisons made during a study

α_B: The Bonferroni adjusted critical α

A Bonferroni correction may also be applied to the *P*-value, where a *P*-value can be adjusted to a maximum of 1, using the formula:

$$P_B = P \times |c|$$

where:

P_B: Bonferroni adjusted *P*-value

Significance is determined in probabilistic statistical tests when $P \leq \alpha$ (or $P < \alpha$) when Type I errors are not controlled. By using a Bonferroni correction, a result is not determined to be significant unless $P \leq \alpha_B$ or $P_B \leq \alpha$, resulting in more robust statistical results. The critical α is adjusted in every invocation of the Bonferroni correction throughout this thesis, with the exception of results produced by "find_enrichment.py" from GOATOOLS (Klopfenstein *et al.*, 2018), which adjusts the *P*-value (*subsections 2.2.3.4; 3.2.5.1.*).

1.5. Aims of this thesis

Efforts to understand the effect, scope, and breadth of modular gene evolution and gene remodelling have been carried out on small datasets due to the computational limitations

imposed by parsing such extensive combinatoric calculations. We aim to uncover the extent of remodelling throughout the evolutionary history of the genome. From there, we aim to investigate and functional or phylogenomic distribution biases using robust bioinformatic and biostatistical methodologies. We believe gene remodelling is an underrepresented and underappreciated evolutionary process and aim to highlight its extent and effects.

Chapter II:

Bioinformatic and Biostatistical Analyses of Gene Remodelling in Fungi

2.1. Introduction

2.1.1. Introduction to mycology

Fungi, with more than five million estimated species, constitute one of the most diverse and speciose eukaryote kingdoms (Blackwell, 2011). Fungi are excellent candidates for assessing trends in eukaryote evolution as they have relatively small genomes, display simple, yet specialised cellular morphologies, exhibit short life cycles, and are amenable to genetic manipulation (Taylor *et al.*, 1993; Botstein *et al*., 1997; Pazouki and Panda, 2000; Calvo *et al.*, 2002; Berger *et al.*, 2005; Alby and Bennett, 2010; Karathia *et al.*, 2011; Leducq, 2014; Mohanta and Bae, 2015). The value of fungi to experimental biology is exemplified by *Saccharomyces cerevisiae* (baker's yeast) being the first eukaryote organism to have its entire genome sequenced (Goffeau *et al.*, 1996); by the wealth of highly curated genomic and proteomic resource databases available for model fungi (Skrzypek and Hirschman, 2011; Stajich *et al.*, 2012; McDowall *et al.*, 2015); and databases dedicated to fungal chromosomal architecture and synteny (Byrne and Wolfe, 2005; Fitzpatrick *et al.*, 2010). Fungi occupy a vast range of ecological niches from mutualistic internal and external mycobiomes (Brundrett, 2002; Seed, 2014; Hager and Ghannoum, 2017) to deadly plant and animal pathogens (Sanglard, 2002; Rosenblum *et al.*, 2008; Brefort *et al.*, 2009; Ma *et al.*, 2010) and have been found occupying niches in hostile environments such as deep sea hydrothermal vents (Le Calvez *et al.*, 2009; Burgaud *et al.*, 2014).

The fungal kingdom is highly diverse, and are known to virtually all known habitable ecosystems (Barron, 2003; Le Calvez *et al.*, 2009; Seed, 2014; Mohanta and Bae, 2015). Fungi are reported to be the earliest eukaryote lineage to have engaged in terrestrialisation (Liu *et al.*, 2006), subsequently altering edaphic geochemistry and enabling terrestrialisation in early plant

lineages (Bidartondo *et al.*, 2011; Quirk *et al.*, 2015). Due to their incredibly vast morphological and biochemical characteristics (Pazouki and Panda, 2000; Calvo *et al.*, 2002) and the low probability of fungal macrostructure fossilisation (Redecker *et al.*, 2000), phylogenetic relationships between fungal lineages are most accurately reconstructed using genomic data (McCarthy and Fitzpatrick, 2017). Fungi are hypothesised to have diverged from the metazoa approximately 833-1891 Ma (Betts *et al.*, 2018). At present, there are eight major fungal phyla (Spatafora *et al.*, 2017) (Figure 2.1.1.). The Cryptomycota and Microsporidia form a clade of deep branching zoosporic true fungi (Bass *et al.*, 2018). The placement of these lineages amongst the fungi has been debated due to the lack of chitinous cell walls and a non-trophic life-cycle, existing as obligate parasites with resultantly reduced gene sets (Capella-Gutiérrez *et al.*, 2012; James *et al.*, 2013). The Blastocladiomycota and Chytridiomycota both constitute aquatic phyla and maintain motile flagella in their spores (James *et al.*, 2006). Both Chytridiomycota and Blastocladiomycota are observed to maintain cholesterol as the fatty acid constituent of cell walls as opposed to ergosterol (Liu *et al.*, 2006). The loss of the flagellum and transition from cell wall cholesterol to ergosterol has been attributed to mycoterrestrial evolution, eventually leading to the emergence of the Mucuromycota and later the Glomeromycota (Spatafora *et al.*, 2017). The Dikarya are the sister clade to Glomeromycota and circumscribes the Ascomycota and Basidiomycota phyla. Dikarya are distinguished from other phyla by the maintenance of two distinct nuclei during their life cycles (Spatafora *et al.*, 2017). Basidiomycota maintain two nuclei during their entire life cycles while Ascomycota maintain a second nucleus during reproduction (Gladfelter and Berman, 2009).

2.1.2. Secondary metabolism

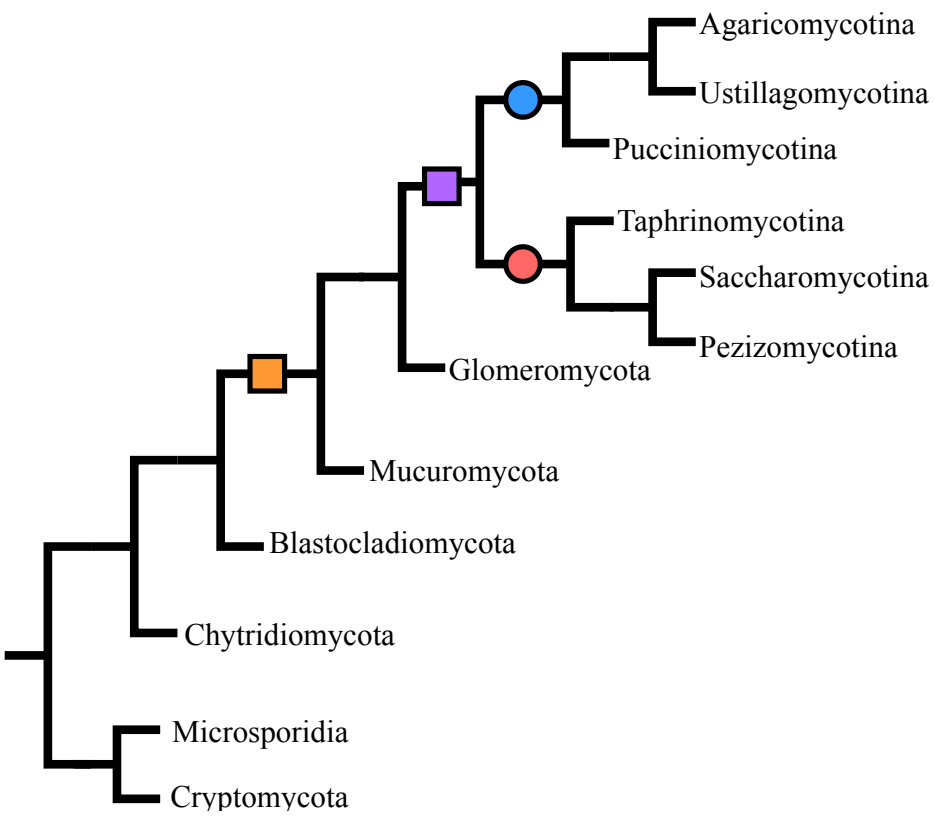

Figure 2.1.1. Representative phylogeny of the major fungal phyla

This phylogeny was adapted from Spatafora *et al.* (2017). Each leaf node represents a phylum or subphylum. Red and blue circles represent the Ascomycota and Basidiomycota phyla respectively. The purple box represents the emergence of the Dikarya. The orange box represents the transition from cholesterol to ergosterol during mycoterrestrialisation and subsequent loss of flagella (Liu *et al.,* 2006).

Secondary metabolites are small, bioactive compounds whose production, while not essential for cell viability, may confer considerable fitness in niche environments (Keller *et al.,* 2005). Like plants and bacteria, fungi have the capacity to produce a wide arsenal of secondary metabolites, primarily polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS) (Calvo *et al.*, 2002; Keller *et al.*, 2005; Perrin *et al.*, 2007; Liu *et al.*, 2015). The evolution of many secondary metabolic pathways, and the regulatory machinery for these pathways, has co-evolved with fungal sexual development, likely as a mechanism to protect spores (Calvo *et al.*, 2002). Secondary metabolism is particularly prominent in Pezizomycotina (Arvas *et al.*, 2007; Lah *et al.*, 2011). Fungal secondary metabolites also present as potent virulence factors for infection of plant and animals hosts, enabling the evolution of many devastating pathogens (Idnurm and Howlett, 2001; Perrin *et al.*, 2007; Ma *et al.*, 2010; Calvo and Cary, 2015). Such metabolites are of considerable economic value, for example, the revolutionary antibiotic, Penicillin, an NRPS, was first isolated from *Penicillium notatum* (Pezizomycotina) in 1929 (Fleming, 1929). This discovery profoundly altered the scope of modern medicine by introducing widespread antibiotic use (Gaynes, 2017).

2.1.3. Fungal chromosome dynamics

The relationship between chromosomal aberration and gene remodelling events has previously been established (Mitelman *et al.*, 2004; Leonard and Richards, 2012; Kloosterman and Hochstenbach, 2014). Large scale chromosomal rearrangements occur frequently throughout the course of evolution (Chang *et al.*, 2013). Such rearrangements have been observed in cells that have adapted to selective conditions during experimental evolution analyses (Gordon *et al.*, 2009; Naseeb *et al.*, 2017). Chromosomal rearrangements in *S. cerevisiae* have been implicated in expanding its nitrogen assimilation repertoire in nitrogen

poor environments (Hellborg *et al*., 2008). This adaption arose *via* rearrangement of the DAL gene cluster to a tightly packed genomic unit, much like a bacterial operon. These structural rearrangements allow for rapid co-expression of genes during nitrogen starvation (Wong and Wolfe, 2005). Experimental reversions of these rearrangements (to mimic those in *Naumovia castelii*) severely reduced the ability of *S. cerevisiae* to survive under nitrogen starvation (Naseeb and Delneri, 2012).

Wild type *S. cerevisiae* sampled from Evolution Canyon, Israel were observed to be significantly more copper resistant due to inversions and segmental duplications in chromosomes XII and XIII (Chang *et al.*, 2013). When samples of this strain were subjected to fluctuating copper concentrations a high frequency of chromosomal reversion to the model configuration was observed, which remained fixed in populations not exposed to high copper concentrations (Chang *et al.*, 2013). These results suggest that yeast chromosomes are malleable and chromosomal rearrangements may serve as a mechanism for rapid adaption. Such chromosomal rearrangements also promote gene remodelling events (Leonard and Richards, 2012).

Fungi do not maintain specialised tissues for most, if any, of their life cycles (Mulder *et al.*, 2007). A lack of pressure to maintain such structures may allow these species to undergo such chromosomal anomalies that may be otherwise deleterious in organisms maintaining multicellular anatomies (Mitelman *et al.*, 2004; Raudsepp and Chowdhary, 2016; Potapova and Gorbsky, 2017).

Despite many species possessing intron rich genomes, Fungi have relatively low rates of alternative splicing when compared to other eukaryotes (McGuire *et al.*, 2008; Grutzmann *et al.*, 2014), with many cases of duplication and subfunctionalisation being associated with intron containing genes (Rastogi and Liberles, 2005; Hickman and Rusche, 2010; Marshall *et al.*, 2013). Due to the propensity for fungal paralogs to undergo subfunctionalization it is

reasonable to assume that these generally follow a "subneofunctionalization" model, whereby subfunctionalized paralogs slowly transition toward novel functionality (He and Zhang, 2005).

One prominent methodology of researching remodelling event evolution, especially *via* gene fusion and fission, is the use of network models over traditional tree models (Haggerty *et al.*, 2014). As a fusion gene typically shares ancestry with two unrelated genes, a typical tree model is inappropriate (as discussed in *subsection 1.1.2.2.1.*) instead an *N*-rooted fusion graph or network is used (Haggerty *et al.*, 2014; Coleman *et al.*, 2015). These networks are typically either protein-protein interaction networks (PPI) or, as is the case throughout this thesis, sequence similarity networks (SSN). In these networks, a gene or protein is displayed as a vertex, and its interaction or detected homology to another gene is displayed as an edge connecting the two vertices (as discussed in section 1.3.). This method is used as it provides greater capacity to visualise composite nodes, HGT, and other combinatorial events on a gene and genome level (Haggerty *et al.*, 2014). The use of these methods has allowed for in-depth exploration into alternative evolutionary mechanisms such as gene remodelling (Jachiet *et al.*, 2013; Pathmanathan *et al.*, 2018).

2.2: Methodology

2.2.1: Benchmarking of CompositeSearch

While CompositeSearch was found to be more sensitive than its predecessors, MosaicFinder and fusedTriplets (Jachiet *et al.,* 2013; Pathmanathan *et al.,* 2018) when detecting remodelled genes in viruses, there are no reports on its use for eukaryotic remodelling event detection. A total of 63 composite genes were reportedly identified in 9 fungal genomes (Table 2.2.1.) using *fdf*BLAST (Leonard and Richards, 2012). We benchmarked

CompositeSearch against *fdf*BLAST on the 9 genomes used by Leonard and Richards (2012). We performed a reciprocal BLAST sequence similarity search on the concatenated genome dataset using an *e*-value stringency (cut-off) score (*E*) of $E \leq 1e^{-10}$ to mirror the *E* used for *fdf*BLAST, we then processed the BLAST output file through cleanBLASTp and CompositeSearch using default parameters. Composite genes identified by both *fdf*BLAST and CompositeSearch are reported in Table 2.3.1.

2.2.2: Development of a composite family quality control procedure

False positive matches (Type I errors) may occur in sequence similarity searches due to "poor gene calls" from poorly assembled genomes (Richards, 2018) As CompositeSearch identifies gene remodelling events based on partially aligning genes, such reports would potentially result in false reports of gene remodelling events. We constructed a pipeline to test the effect of poor genome annotation in datasets used for composite detection (Figure 2.2.1.). This approach consisted of controlling composite family size (-*x*) and component family size (-*y*). The pipeline consisted of 7 steps:

(*i*.): A subset of genes (5%; n = 4,451) were randomly selected and extracted from the dataset. "Pseudo-fissions" were induced in each subset sequence by artificially splitting at its midpoint to form two separate sequences.

(*ii*.): A sequence similarity network (SSN) was constructed using BLAST ($E \leq 1e^{-05}$) using the parameters required by CompositeSearch as described in subsection 1.3.2.

Table 2.2.1. Fungal genomes used for composite detection by Leonard and Richards, 2012

These 9 reference genomes are sampled from across the fungal Tree of Life and constitute a small yet highly diverse dataset. Originally, this dataset also included two Microsporidia, however, as no composites were detected in these taxa by Leonard and Richards (2012), they were excluded from their published analysis.

Species	GenBank ID	n_{genes}	Taxonomy
Allomyces macrogynus	28583	17,600	Blastocladiomycota
Batrachochytrium dendrobatidis JEL423	109871	8,732	Chytridiomycota
Coprinopsis cinerea (strain FGSC 9003)	5346	13,394	Basidiomycota; Agaricomycotina
Ustilago maydis 521	5270	6,522	Basidiomycota; Ustillagomycotina
Schizosaccharomyces pombe 972h	4896	5,010	Ascomycota; Taphrinomycotina
Saccharomyces cerevisiae S288c	4932	5,885	Ascomycota; Saccharomycotina
Neurospora crassa OR74A	5141	9,908	Ascomycota; Pezizomycotina
Mucor circinelloides f. lusitanicus	29924	10,930	Mucoromycota
Rhizopus oryzae RA 99880	64495	17,459	Mucoromycota

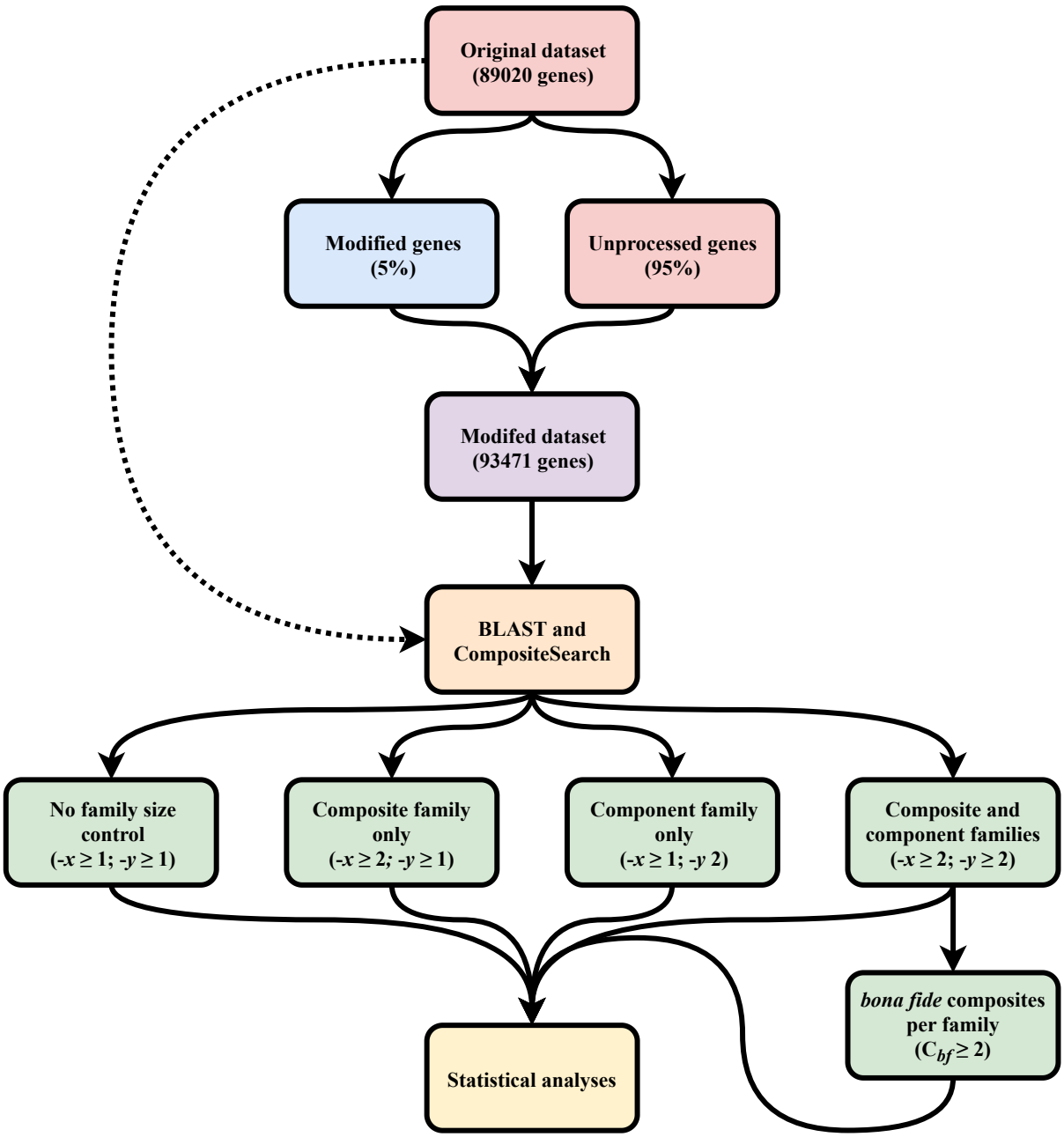

Figure 2.2.1. Effect of low quality genomes in composite detection analyses

This graphic illustrates the pipeline described in subsection 2.2.2. Each level depicts a step in the pipeline for each iteration. Solid black lines indicate the directional succession of each step. The broken line depicts the generation of the control dataset which does not introduce "pseudo-fissions" into the dataset.

(*iii*.): A concise SSN was constructed by parsing the SSN through "cleanBlastp" (Pathmanathan *et al*., 2018). This removed self-hits, redundant hits, and selected the longest HSP in cases where multiple HSPs were observed between hits.

(*iv*.): A series of four CompositeSearch analyses were performed on the concise SSN:

(*a*.): $-x \geq 1$; $-y \geq 1$ (default settings)

(*b*.): $-x \geq 2$; $-y \geq 1$

(*c*.): $-x \geq 1$; $-y \geq 2$

(*d*.): $-x \geq 2$; $-y \geq 2$

A fifth analysis ("strict filter") was included by extracting composite families where the sum of *bona fide* composites ($-x_{bf}$) ≥ 2 from the "compositefamiliesinfo" output file produced during analysis (*d*.; $-x \geq 2$; $-y \geq 2$). As described in subsection 1.3.2., a *bona fide* composite is identified as having non-overlapping homology to members of two distinct component families.

(*v*.): Data for each of the five analyses was recorded for downstream comparisons.

(*vi*.): A series of pseudoreplicates was constructed by performing steps 1-5 100 times, and a control dataset (a dataset without pseudofissions) was constructed by performing steps 2-5 on the original dataset.

(*vii*.): Control and experimental datasets were compared:

(*a*.): We assumed 5% of identified composite families from the control dataset were false positives (FP_C).

(*b*.): Experimental false positives (FP_E) were computed using the formula:

$$FP_E = (CF_E - CF_C) + FP_C$$

where:

C: Control dataset (*n*)

E: Experimental dataset (*n*)

FP: False positives (*n*)

CF: Sum of detected composite families (*n*)

(*c*.) A one-tailed Fisher's exact test ($H_O:\pi(x)\leq\pi(X);H_O:\pi(x)>\pi(X)$) was performed between control and experimental false positives for each iteration (Table 2.3.2., Figure 2.3.1.).

As the purpose of this was to identify a filter where control and experimental false positives were statistically similar, a result with an insignificant *P*-value ($P \geq 0.05$) was considered to be successful.

2.2.3. Development of a composite gene analysis pipeline

We developed a modular pipeline for remodelled gene detection and analysis (Figure 2.2.2.). This pipeline consisted of five modules which are described below (*subsections 2.2.3.1-2.2.3.5.*). This pipeline was an attempt to identify:

(*a*.) Gene remodelling trends across the dataset

(*b*.) Functional trends in remodelled gene categories

(*c*.) Gene remodelling trends across phylogenies

2.2.3.1. Database construction and quality control

A dataset of 107 fungal genomes (Table 2.2.2.) was obtained from Leonard and Richards (2012) and taxonomic lineages were obtained from the sources provided by the authors. The quality of these genomes were assessed using BUSCO *v*3.0.2 (Simão *et al*., 2015; Waterhouse *et al*., 2017) with the fungal single copy orthologous dataset (fungi_odb9) from OrthoDB (Zdobnov *et al*., 2017; Kriventsev*a et al*., 2018). Genome size (Mbp) and GC% were obtained from source for each taxon and genome density (n_{genes}/Mbp) was obtained by dividing the number of genes in a given genome by its respective genome size (Table 2.3.3.). Descriptive statistics (mean, median, standard deviation, quartiles, and co-efficient of variation (CV)) were calculated for collated genomic statistics and BUSCO completion (Table 2.3.4.). The dataset consisted of 1,150,995 canonical protein sequences sampled from across five major fungal phyla (Ascomycota (71 genomes), Basidiomycota (30 genomes), Blastocladiomycota (1 genome), Chytridiomycota (2 genomes), and Mucoromycota (3 genomes)). This heavily biased towards the Dikarya, and specifically to the Ascomycota, due to the availability of the data at the time of Leonard and Richards' initial publication.

2.2.3.2. CompositeSearch analysis, quality control, and annotation

A sequence similarity network (SSN) was constructed using BLAST ($E \leq 1e^{-05}$) using the parameters required by CompositeSearch as discussed in subsection 1.3.2., resulting in

Table 2.2.2. Dataset of fungal genomes

Each of the 107 fungal species used in this chapter is provided with its GenBank ID (if applicable), the number of genes observed in its genome (n_{genes}), and its taxonomic lineage. Taxonomic clades are ranked left to right in increasing order of specificity (phylum, subphylum, class, order, and family). A total of 101 species (Ascomycota and Basidiomycota) belong within a subkingdom "Dikarya", and all belong within a subphylum. The 6 genomes that do not belong to a subphylum are annotated as "N/A".

Binomial classification	GenBank ID	n_{genes}	Taxonomic lineage
Dothistroma septosporum	--	12580	Ascomycota; Pezizomycotina; Dothideomycete; Capnodiales; Mycosphaerellaceae
Mycosphaerella fijiensis CIRAD86	83344	10313	Ascomycota; Pezizomycotina; Dothideomycete; Capnodiales; Mycosphaerellaceae
Mycosphaerella graminicola IPO323	54734	10933	Ascomycota; Pezizomycotina; Dothideomycete; Capnodiales; Mycosphaerellaceae
Septoria musiva	--	10233	Ascomycota; Pezizomycotina; Dothideomycete; Capnodiales; Mycosphaerellaceae
Septoria populicola	--	9739	Ascomycota; Pezizomycotina; Dothideomycete; Capnodiales; Mycosphaerellaceae
Baudoinia compniacensis	--	10153	Ascomycota; Pezizomycotina; Dothideomycete; Capnodiales; Teratosphaeriaceae
Hysterium pulicare	--	12352	Ascomycota; Pezizomycotina; Dothideomycete; Hysteriales; Hysteriaceae
Rhytidhysteron rufulum	--	12117	Ascomycota; Pezizomycotina; Dothideomycete; Hysteriales; Hysteriaceae
Alternaria brassicicola ATCC 96836	29001	10688	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae
Cochliobolus heterostrophus	5016	9633	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae
Cochliobolus sativus	--	12250	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae
Leptosphaeria maculans	--	12469	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae
Pyrenophora teres	--	11799	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae
Pyrenophora triticirepentis strain Pt1CBFP	45151	12169	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae

Binomial classification	GenBank ID	n_{genes}	Taxonomic lineage
Setosphaeria turcica	--	11702	Ascomycota; Pezizomycotina; Dothideomycete; Pleospoales; Pleosporaceae
Aspergillus aculeatus	--	10828	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus carbonarius	40993	11624	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus clavatus	5057	9120	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus flavus	5059	12587	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus fumigatus Af293	5085	9887	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus nidulans FGSCA4	41734	10560	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus oryzae RIB40	5062	12063	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Aspergillus terreus NIH 2624	33178	10406	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
*Neosartorya fischeri (*NRRL 181)	36630	10403	Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Aspergillaceae
Blastomyces dermatitidis	5039	9522	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Ajellomycetaceae
Histoplasma capsulatum (strain NAm1)	5037	9251	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Ajellomycetaceae
Paracoccidioides brasiliensis Pb01	121759	9136	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Ajellomycetaceae
Microsporum canis CBS 113480	63405	8765	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Arthrodermataceae
Microsporum gypseum CBS 118893	489714	8876	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Arthrodermataceae
Trichophyton equinum CBS127.97	63418	8560	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Arthrodermataceae
Coccidioides immitis RS	5501	10654	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Onygenaceae
Coccidioides posadasii str. Silveira	199306	10124	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Onygenaceae
Uncinocarpus reesii	33188	7798	Ascomycota; Pezizomycotina; Eurotiomycetes; Onygenales; Onygenaceae
Botryotinia cinerea (strain B05.10)	40559	16448	Ascomycota; Pezizomycotina; Leotiomycetes; Helotiales; Sclerotiniaceae
Sclerotinia sclerotiorum (strain ATCC 18683)	5180	14522	Ascomycota; Pezizomycotina; Leotiomycetes; Helotiales; Sclerotiniaceae
Cryphonectria parasitica	5116	11184	Ascomycota; Pezizomycotina; Sordariomycete; Diaporthales; Cryphonectriaceae
Acremonium alcalophilum	--	9521	Ascomycota; Pezizomycotina; Sordariomycete; Glomerellales; Plectosphaerellaceae
Verticillium alboatrum VaMs.102	526221	10220	Ascomycota; Pezizomycotina; Sordariomycete; Glomerellales; Plectosphaerellaceae
Verticillium dahliae VdLs.17	27337	10535	Ascomycota; Pezizomycotina; Sordariomycete; Glomerellales; Plectosphaerellaceae
Trichoderma atroviride IMI 202040	63577	11100	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Hypocreaceae

Binomial classification	GenBank ID	n_{genes}	Taxonomic lineage
Trichoderma reesei QM6a	51453	9143	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Hypocreaceae
Trichoderma virens Gv298	29875	11643	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Hypocreaceae
Fusarium graminearum species complex	5518	13321	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Nectriaceae
Fusarium oxysporum f. sp. lycopersici	5507	17608	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Nectriaceae
Fusarium verticillioides	117187	14195	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Nectriaceae
Nectria haematococca mpVI	70790	15707	Ascomycota; Pezizomycotina; Sordariomycete; Hypocreales; Nectriaceae
Magnaporthe grisea 7015	148305	11109	Ascomycota; Pezizomycotina; Sordariomycete; Magnaporthales; Magnaporthaceae
Chaetomium globosum CBS 148.51	38033	11124	Ascomycota; Pezizomycotina; Sordariomycete; Sordariales; Chaetomiaceae
Sporotrichum thermophile	--	9166	Ascomycota; Pezizomycotina; Sordariomycete; Sordariales; Chaetomiaceae
Thielavia terrestris	35720	9815	Ascomycota; Pezizomycotina; Sordariomycete; Sordariales; Chaetomiaceae
Podospora anserina DSM 980	5145	10601	Ascomycota; Pezizomycotina; Sordariomycete; Sordariales; Lasiosphaeriaceae
Neurospora crassa OR74A	5141	9908	Ascomycota; Pezizomycotina; Sordariomycete; Sordariales; Sordariaceae
Neurospora tetrasperma FGSC 2508	40127	10640	Ascomycota; Pezizomycotina; Sordariomycete; Sordariales; Sordariaceae
Wickerhamomyces anomalus	--	6423	Ascomycota; Saccharomycotina; Saccharomycetes; Phaffomycetaceae; Wickerhamomyces
Candida albicans SC5314	5476	6205	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Debaryomycetaceae
Candida caseinolytica	--	4657	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Debaryomycetaceae
Candida glabrata CBS 138	5478	5202	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Debaryomycetaceae
Candida tenuis	--	5533	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Debaryomycetaceae
Debaryomyces hansenii CBS767	4959	6272	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Debaryomycetaceae
Spathaspora passalidarum	--	5983	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Debaryomycetaceae
Yarrowia lipolytica CLIB122	4952	6448	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Dipodascaceae
Lipomyces starkeyi	--	8192	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Lipomycetaceae
Hansenula polymorpha	--	5177	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Pichaiaceae
Pichia membranifaciens	--	5546	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Pichaiaceae
Pichia stipitis CBS 6054	4924	5807	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Pichaiaceae
Ashbya gossypii ATCC 10895	33169	4717	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae

Binomial classification	**GenBank ID**	n_{genes}	**Taxonomic lineage**
Saccharomyces cerevisiae S288c	4932	5885	Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae
Schizosaccharomyces cryophilus oy26	653667	5057	Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae
Schizosaccharomyces japonicus yFS27	4897	4814	Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae
Schizosaccharomyces octosporus yFS286	4899	4925	Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae
Schizosaccharomyces pombe 972h	4896	5010	Ascomycota; Taphrinomycotina; Schizosaccharomycetes; Schizosaccharomycetales; Schizosaccharomycetaceae
Agaricus bisporus var. *burnettii* JB137-S8	597362	11289	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Agaricaceae
Schizophyllum commune H48	5334	13181	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Agaricaceae
Pleurotus ostreatus	5322	11603	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Pleurotaceae
Fomitopsis pinicola	--	14724	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Polyporaceae
Trametes versicolor	--	14296	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Polyporaceae
Wolfiporia cocos	--	12746	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Polyporaceae
Coprinopsis cinerea (strain FGSC 9003)	5346	13394	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Psathyrellaceae
Laccaria bicolor (strain S238NH82)	29883	19036	Basidiomycota; Agaricomycotina; Agaricomycete; Agaricales; Tricholomataceae
Auricularia delicata	--	23577	Basidiomycota; Agaricomycotina; Agaricomycete; Auriculariales; Auriculariaceae
Coniophora putinea	--	13761	Basidiomycota; Agaricomycotina; Agaricomycete; Boletales; Coniophorineae
Serpula lacrymans S7.3	85982	14495	Basidiomycota; Agaricomycotina; Agaricomycete; Boletales; Coniophorineae
Phlebia brevispora	--	16170	Basidiomycota; Agaricomycotina; Agaricomycete; Corticales; Corticiaceae
Punctularia strigosozonata	--	11538	Basidiomycota; Agaricomycotina; Agaricomycete; Corticiales; Punctulariaceae
Gloeophyllum trabeum	--	11846	Basidiomycota; Agaricomycotina; Agaricomycete; Gloeophyllales; Gloeophyllaceae
Fomitiporia mediterranea	--	11333	Basidiomycota; Agaricomycotina; Agaricomycete; Hymenochaetales; Hymenochaetaceae
Ganoderma sp.	--	12910	Basidiomycota; Agaricomycotina; Agaricomycete; Polyporales; Ganodermataceae
Bjerkandera adusta	--	15473	Basidiomycota; Agaricomycotina; Agaricomycete; Polyporales; Meruliaceae
Ceriporiopsis subvermispora	--	12125	Basidiomycota; Agaricomycotina; Agaricomycete; Polyporales; Phanerochaetaceae
Phanerochaete chrysosporium RP78	5306	10048	Basidiomycota; Agaricomycotina; Agaricomycete; Polyporales; Phanerochaetaceae
Phlebiopsis gigantea	--	11891	Basidiomycota; Agaricomycotina; Agaricomycete; Polyporales; Phanerochaetaceae
Dichomitus squalens	--	12290	Basidiomycota; Agaricomycotina; Agaricomycete; Polyporales; Polyporaceae

Binomial classification	**GenBank ID**	n_{genes}	**Taxonomic lineage**
Heterobasidion annosum	13563	12299	Basidiomycota; Agaricomycotina; Agaricomycete; Russulales; Bondarzewiaceae
Dacryopinax sp.	--	10242	Basidiomycota; Agaricomycotina; Dacrymycete; Dacrymycetales; Dacrymycetaceae
Cryptococcus neoformans var. *grubii* H99	5207	6967	Basidiomycota; Agaricomycotina; Tremellomycete; Tremellales; Tremellaceae
Tremella mesenterica	5217	8313	Basidiomycota; Agaricomycotina; Tremellomycete; Tremellales; Tremellaceae
Rhodotorula graminis	--	7283	Basidiomycota; Pucciniomycotina; Microbotryomycetes; Sporidiobolales; Sporidiobolales
Sporobolomyces roseus IAM 13481	40563	5536	Basidiomycota; Pucciniomycotina; Microbotryomycetes; Sporidiobolales; Sporidiobolales
Melampsora laricis-populina	203908	16831	Basidiomycota; Pucciniomycotina; Pucciniomycetes; Pucciniales; Melampsoraceae
Puccinia graminis f. sp. tritici	5297	20566	Basidiomycota; Pucciniomycotina; Pucciniomycetes; Pucciniales; Pucciniaceae
Ustilago maydis 521	5270	6522	Basidiomycota; Ustilaginomycotina; Ustilaginomycetes; Ustilaginales; Ustilaginaceae
Allomyces macrogynus	28583	17600	Blastocladiomycota; N/A; Blastocladiomycetes; Blastocladiales; Blastocladiaceae
Batrachochytrium dendrobatidis JEL423	109871	8732	Chytridiomycota; N/A; Chytridiomycetes; Rhizophydiales; Rhizophydiales
Spizellomyces punctatus	109760	8804	Chytridiomycota; N/A; Chytridiomycetes; Spizellomycetales; Spizellomycetaceae
Mucor circinelloides f. lusitanicus	29924	10930	Mucoromycota; Mucuromycotina; Mucoromycetes; Mucorales; Mucoraceae
Phycomyces blakesleeanus	4837	16528	Mucoromycota; Mucuromycotina; Mucoromycetes; Mucorales; Phycomycetaceae
Rhizopus oryzae RA 99880	64495	17459	Mucoromycota; Mucuromycotina; Mucoromycetes; Mucorales; Rhizopodaceae

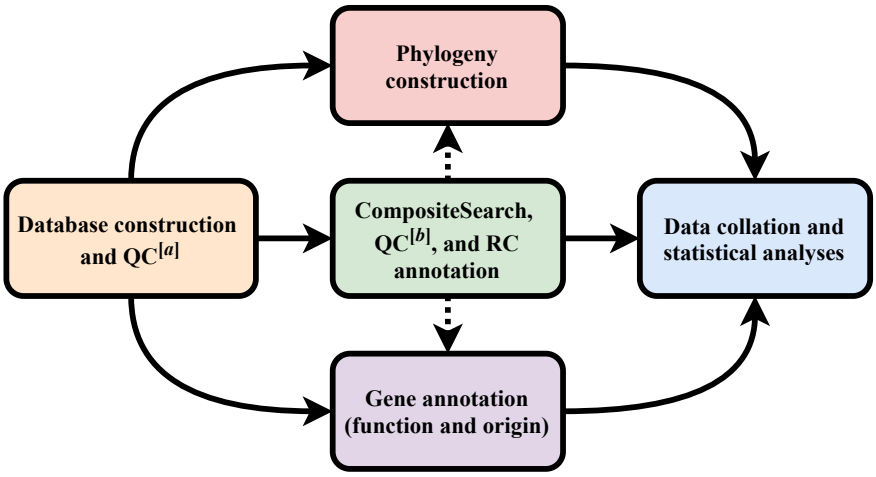

Figure 2.2.2. Remodelled gene analysis pipeline

The orange box represents the initial genome database construction and quality control using BUSCO ($QC^{[a]}$) with an appropriate BUSCO set from OrthoDB (*subsection 2.2.3.1.*). The red box represents a phylogeny construction by aligning groups of highly distributed orthologs (using MUSCLE) and merging them to construct a superalignment which is used to build a phylogeny using PhyML with a model as decided by ProtTest (*subsection 2.2.3.4*.). The purple box represents assigning PFAM and GO terms to each gene using InterProScan, and assignin putative "Domains-of-origin" using the method described by Cotton and McInerney, 2010. The green box represents the identification of remodelled genes using BLASTP and ComposteSearch. The second round of quality control ($QC^{[b]}$) refers to the family-size control strategy discussed in *subsection 2.2.3.2.*). Solid arrows represent the "flow" of data and dashed arrows refer to data mergers, where analyses rely on the output of two datasers. Finally, the blue box refers to the collation and analysis of data (section 2.2.3.).

$1.325e^{12}$ pairwise sequence comparisons. The SSN was parsed through "cleanBlastp" to produce a concise SSN. The functionality of "cleanBlastp" is discussed in subsection 1.3.2. CompositeSearch was performed on the concise SSN using default parameters with composite family size control ($-x \geq 2$) and component family size control ($-y \geq 2$) parameters. The strict filter protocol (as described in section 2.2.2.) was applied to reduce potential Type I errors. Genes were annotated with a remodelling category (RC) as per *subsection 1.1.2.2*.

2.2.3.3. Trends in gene family sizes

Descriptive statistics were computed for family sizes in each RC (Table 2.3.5., Figure 2.3.2.). As considerable variation was observed (CV ≤ 294.6%), datasets were compared using a two-tailed Mann-Whitney *U* test (H_O:$\eta_1=\eta_2$;H_A:$\eta_1\neq\eta_2$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$) and a $P \leq \alpha_B$ was considered statistically significant (Table 2.3.6).

2.2.3.4. Phylogenetic and character state reconstruction

2.2.3.4.1. Phylogenetic reconstruction

A phylogeny of the organisms in this analysis was constructed by obtaining all yeast "euKaryote Orthologous Groups" (KOGs) (Tatusov *et al.*, 2003) from JGI (https://www.genome.jgi.doe.gov) and searching them against our dataset using BLASTP ($E \leq 1e^{-20}$). KOG sequences were retained for phylogenetic construction if they were present in 95% (~102 species), resulting in 277 KOG sequences that were sampled for reconstruction. The reciprocal best BLAST hits was identified for each of the 277 KOGs in each species (if

present) and was extracted as the representative sequence for that species resulting in the creation of 277 KOG families. Each KOG family was aligned using MUSCLE *v*6 with default parameters (Edgar, 2004). Poorly aligned positions and other non-informative positions (such as positions with no substitutions or only one substitution) were removed from each alignment (resulting in removal of the entire gene alignment in 28 cases) by using Gblocks *v*0.91*b* with default parameters (Castresana, 2000). A total of 249 KOG families were retained for phylogenetic reconstruction. A superalignment was constructed from the 249 alignments using FASconCAT *v*1.0 with default parameters (Kuck and Meusemann, 2010). Bootstrap support for internal branches were assessed in PhyML *v*3.0 (Guindon *et al.*, 2010) using 100 bootstrap samples and the LG+I+G (Le and Gascuel, 2008) model, as selected by ProtTest *v*3.0 (Darriba *et al.*, 2011). A majority-rule consensus tree was then constructed using PAUP* (Swafford, 2002) and visualised using iTOL *v*3 (Letunic and Bork, 2016) (Figure 2.3.3.).

2.2.3.4.2. Character state reconstruction

A presence-absence matrix was constructed for each CompositeSearch defined sequence family for each species. Each branch was assigned a node ID using the "naked -; tplot;" functions with "Tree analysis with New Technology" (TNT) *v*1.5 (Goloboff *et al.*, 2008). The TNT "-apo" function was used to plot character states from a presence-absence matrix of gene families to their inferred branches on an independently constructed phylogeny (Figure 2.3.4.), this function also determines whether a character state (trait; T) was gained (birthed; T_b) and lost (decayed; T_d) for each branch on the phylogeny by comparing families with its preceding branch. As TNT does not assign any character states to the root node or any node immediately succeeding it, we appended two "pseudo-outgroups" to the phylogenetic

Newick file and to the presence-absence matrix (Figure 2.3.4.) For the presence-absence matrix, each pseudo-outgroup was assigned a "0" (absence) for each character state.

2.2.3.4.3. Comparison of homoplastic proportions

Any character state that was observed to have been gained more than once by TNT ($T_{b(f>1)}$) across the phylogeny was considered homoplastic. We calculated the "homoplastic proportion" (HP) for each RC (HP_{RC}) using the formula:

$$HP_{RC} = \frac{\Sigma(T_{b(FS>1)})}{\Sigma(T_b)}$$

A two-tailed Fisher's exact test ($H_O:\pi(x)=\pi(X);H_A:\pi(x)\neq\pi(X)$) was used to compare HP rates between each RC. A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$) and a $P \leq \alpha_B$ was considered statistically significant (Table 2.3.7).

2.2.3.4.4. Construction of rate series

The sum of RC trait births (T_b) and decays (T_d) were calculated for each branch. We calculated the T_b/Meganuum (Ma; f_b) and T_d/Ma (f_d) using the formula:

$$f_x = \frac{T_x}{\kappa \times \tau};\ \tau = \frac{DT}{LP}$$

where:

f: Frequency (rate)

T_x: The sum of birthed traits (T_b) **or** decayed traits (T_d) at a given branch

κ: Branch length

τ: "Time constant"

DT: Divergence time of eukaryotes

LP: Longest root-to-leaf path in our phylogeny

Fungal molecular clock studies are fraught with difficulties due to the lack of reliable fossil records for accurate node calibration (Berbee and Taylor, 2009; Prieto and Wedin, 2013). A highly calibrated molecular clock analysis of species across the Tree of Life estimated Dikarya to have reportedly between 392.1-1823 Ma (Betts *et al.,* 2018). The farthest child (longest path; LP) from the root of Dikarya (Node 118 in our phylogeny (Figure 2.3.4)) was reported to be *Pichia membranifaciens* (κ = 0.61630503). We calculated τ using 392.1 and 1823 Ma as DT. This indicated that κ = 1 represents 636.21 to 2957.95 Ma. As 2957.97 Ma is considerably greater than the divergence time for Eukaryotes (Betts *et al*., 2018), we determined τ = 636.21 to be most appropriate. When adjusted for the LP from the root (*P. membranifaciens*; LP = 0.88999) the divergence time for the MCRA was determined to be 566.22 Ma. This timepoint was also deemed suitable as established Chytridiomycota and Blastocladiomycota fossils (our earliest diverging lineages) have dated to approximately the Devonian-Carboniferous Periods (400-300 Ma; Krings *et al*., 2009*a*; Krings *et al.,* 2009*b*) and the suspected Glomeromycota or Mucuromycota *Prototaxites loganii* has been dated to the Middle Devonian (386 Ma; Retallack and Landing, 2014). We calculated f_x for each branch using the formula described in *subsection 2.2.3.3.2.2.* (Table 2.3.5.). The root node was excluded from further analyses as it did not possess a κ.

2.2.3.4.4.1. Comparison of rates between and within nodes and tips

For each RC, three sets of f_b and f_d descriptive statistics were calculated (Table 2.3.8.) for each of:

(*a*.) The entire phylogeny

(*b*.) The subset of internal nodes within the phylogeny (branches)

(*c*.) The subset of tip nodes within the phylogeny (leaves)

For each RC, to determine whether considerable differences exist between tips ($\mu f_{x(t)}$) and nodes ($\mu f_{x(n)}$) Mann-Whitney U tests ($H_O:\eta_1=\eta_2;H_A:\eta_1\neq\eta_2$) were used to compare:

(*a*.) $\mu f_{b(n)}$ and $\mu f_{b(t)}$

(*b*.) $\mu f_{d(n)}$ and $\mu f_{d(t)}$

A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 2$; $\alpha_B = 0.025$) was applied to each set and instances where $P \leq \alpha_B$ were considered statistically significant.

To determine whether considerable rate differences exist between RCs within each set, Mann-Whitney U tests ($H_O:\eta_1=\eta_2;H_A:\eta_1\neq\eta_2$) were used to compare:

(*a*.) Each $\mu f_{b(n)}$ to each other $\mu f_{b(n)}$

(*b*.) Each $\mu f_{b(t)}$ to each other $\mu f_{b(t)}$

(*c*.) Each $\mu f_{d(n)}$ to each other $\mu f_{d(n)}$

(*d*.) Each $\mu f_{d(t)}$ to each other $\mu f_{d(t)}$

Again, a Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$) was applied to each set and a $P \leq \alpha_B$) was considered statistically significant (Table 2.3.9.)

2.2.3.4.4.2.. Investigation for evolutionary bursts

An evolutionary burst is a significant increase in f_b or f_d compared to the background rate. To determine the presence of any evolutionary bursts, rate series were transformed to a Gaussian distribution using a Box-Cox transformation and bursts were determined by using a Q-function (H_O:$Q(x)>1-\Phi(x)$;H_A:$Q(x)\leq 1-\Phi(x)$) as described in *subsection 1.4.1.3.* This procedure was implemented for:

(a.) f_b and f_d for all tips and internal nodes ($\{f_{x(y)}\}$; y)

(b.) $f_{b(n)}$ and $f_{d(n)}$ only ($\{f_{x(n)}\}$; x)

A Bonferroni correction was applied to each set ($\alpha = 0.05$; $|c_{(y)}| = 212$; $|c_{(x)}| = 105$; $\alpha_{B(y)} = 2.6e^{-04}$; $\alpha_{B(x)} = 4.8e^{-04}$) and instances where $P_{(y)} \leq \alpha_{B(y)}$ or $P_{(x)} \leq \alpha_{B(x)}$ were considered statistically significant (Tables 2.3.9.-10..)

2.2.3.5. Gene annotation (function and origin)

2.2.3.5.1 Functional annotation of each RC

Each sequence in the dataset was assigned a PFAM domain (Bateman *et al.,* 2004) and gene ontology (GO) functional terms (Ashburner *et al*., 2000) using InterProScan *v*5 (Jones *et al*., 2014) using the "PfamA" model ($E \leq 1e^{-03}$) and "--goterms" flag. GO terms for each

sequences were grouped into a "GO-slim" subset (broad category) terms using the "map_to_slim.py" script from GOATOOLS (Klopfenstein *et al.*, 2018) using the curated generic ontology map (produced by https://www.geneontology.org). Each sequence was mapped back to its CompositeSearch defined family and RC annotation. For each RC, overrepresentated GO terms were identified by comparing to the background of all other genes in the dataset using the "find_enrichment.py" script from GOATOOLS. The "find_enrichment.py" script uses a applies a Bonferroni correction to *P*-values to report statistical significance ($P_B = P \times |c|$; $P_{B(\max)} < 1$) and instances where $P_B \leq 0.05$ were considered significantly overrepresented (Tables 2.3.12.-19.).

2.2.3.5.2. Functional enrichments during the divergence of Pezizomycotina

Considerable phenotypic innovations are hypothesised to have emerged during the divergence of Pezizomycotina from Saccharomycotina, such as a wide expansion in secondary metabolism (Wisecaver *et al*., 2014) and a transition from a predominantly asexual life cycle to being predominantly sexual (Ojeda-Lopéz *et al.,* 2018). The transition from asexuality to sexuality has led to a bias towards the maintenance of a multicellular teleomorph during at least some of the life cycle (in most species (Wynns, 2015)) as opposed to only maintaining a yeast-like unicellular anamorph (Nagy *et al*., 2018). Pezizomycotina reportedly underwent considerable genomic expansions and genomic rearrangements during this divergence, leading to larger, more complex genomes and more rapid evolutionary rates when compared to other Ascomycota subphyla (Kelkar and Ochman, 2012). Due to the correlations between genome rearrangement and gene remodelling (Leonard and Richards, 2012), and between gene remodelling and phenotype evolution (Richards *et al.,* 2006*;* Alvelar *et al*., 2014) it was hypothesised that gene remodelling may be correlated with some of the innovations observed

during this transition. Genes for each RC were extracted from families assigned to Node 115 (the root of Pezizomycotina), resulting in four sets. Again, overrepresentation was established for each RC set using "find_enrichment.py" using all genes from all sampled Pezizomycotina as a background and the curated Aspergillus GO-slim ontology map (Table 2.3.20.). In all cases, a Bonferroni correction was applied to the *P*-value and instances where $P_B \leq 0.05$ were considered significantly overrepresented.

2.2.3.5.3. Gene origin annotation

A method for assigning genes to a potential "Domain-of-origin" (Bacteria, Archaea, or Eukaryote; DO) has previously been described (Cotton & McInerney, 2010). Each gene in the dataset was searched ($E \leq 1e^{-06}$) against a large bacterial and archaeal dataset obtained from McCarthy & Fitzpatrick (2016). If a gene unambiguously returns hits in only one DO (bacteria or archaea), it was considered to originate in that DO. If a gene was reported to have hits in both DOs, it was annotated as "Undefined Prokaryote". This annotation is necessary to prevent DO mis-annotation due to the possibility of HGT between bacteria and archaea. If a hit was not reported for a gene it was considered to be of eukaryote origin. DOs were assigned to each family by majority rule of their associated genes (Table 2.3.21.**)**. For each RC, DOs were compared to all other DOs using a two-tailed Fisher's exact test ($H_O: \pi(x) = \pi(X); H_A: \pi(x) \neq \pi(X)$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$) and instances where $P \leq \alpha_B$ were considered statistically significant.

2.2.3.6. Trends between gene remodelling and genome characteristics

As discussed in *subsection 2.1.4.*, fungal genomes have been observed to undergo frequent rounds of WGD and chromosomal rearrangements, thus providing an environment that promotes gene remodelling events. "Retained remodelling events" (or "internal remodelling events") were established for each genome using the protocol discussed in *subsection 2.2.3.2* (Figure 2.2.3.). Genome remodelling category proportions (RCPs) were calculated for genes obtained from the "globally remodelled" dataset (GRCP) and the "retained remodelling" subset (RRCP) by dividing the number of genes in each RC assigned to a given genome by the number of genes in the genome (Tables 2.3.23.-24; Figure 2.3.5.).

2.2.3.6.2. Correlations between genomic characteristics and genomic remodelled proportions

We collected data for four genomic characteristics for each of our sampled taxa for comparative analyses:

(*a*.) Genome size (Mbp)

(*b*.) Genome density (n_{genes}/Mbp)

(*c*.) Genome guanine-cytosine content (GC %); and

(*d*.) BUSCO genome completeness (C%)

Genome size and GC% for each fungal genome assembly was obtained from their respective repositories (Table 2.3.3.). Genome density was calculated by dividing the sum of genes in a given genome by its genome size. BUSCO completeness (%) was obtained from section *2.2.2.* and RCPs were obtained from *subsection 2.2.3.5.1.* (Table 2.3.21.). Correlations were established using a Spearman's ρ correlation test (H_O:$X_1 \propto X_2$;H_A:$X_1 \not\propto X_2$) between each

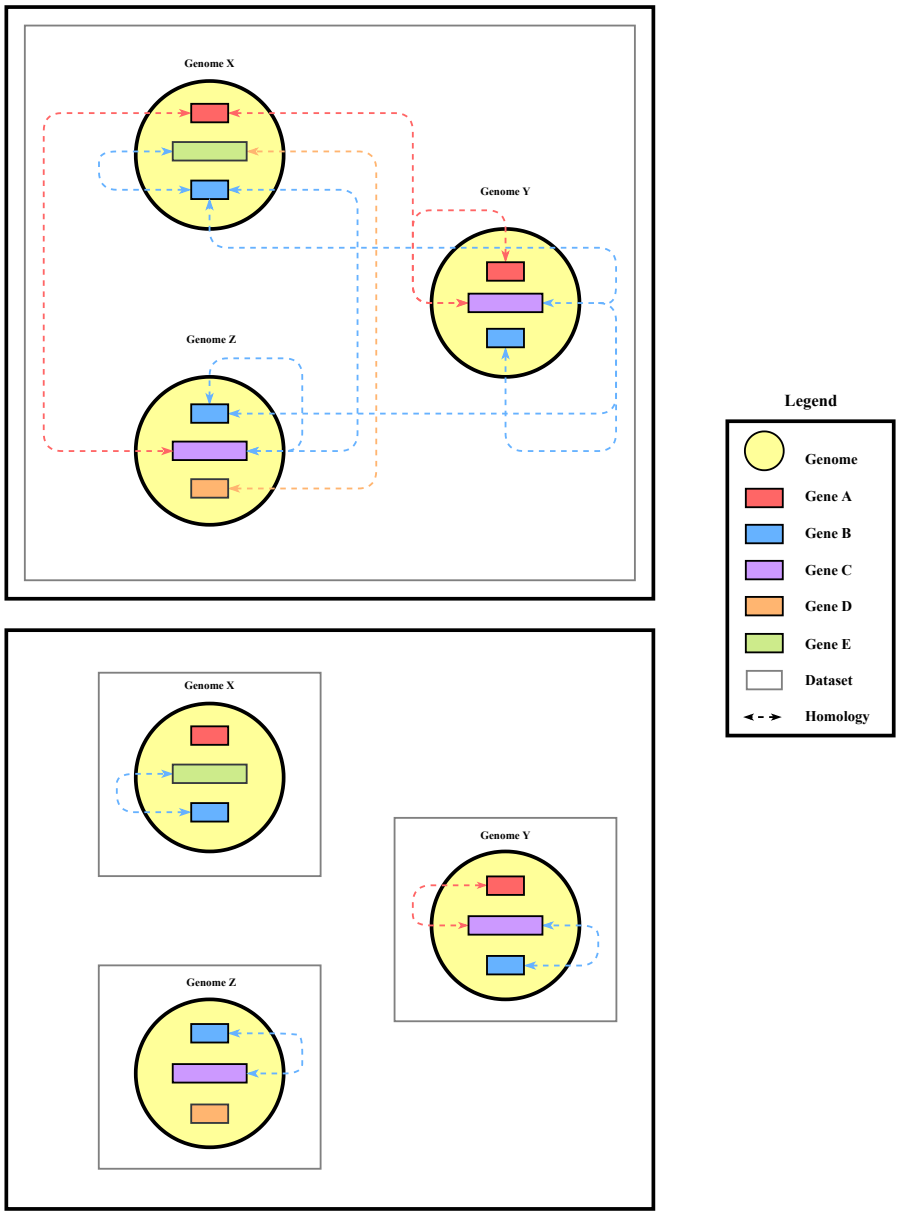

Figure 2.2.3. Global remodelling *vs* internal (retained remodelling)

The top image displays "global remodelling" where remodelling events are detected within multigenome datasets. The bottom image displays "internal remodelling", a subset of global remodelling where both component families and the composite family are required to be observed within the same genome. Higher rates of internal remodelling are expected for genomes that have underwent WGD events due to subsequent chromosomal restructuring post-WGD. Only Gene C in Genome Y is observed to be "internally remodelled".

RCP and each of the four genomic characteristics (Figure 2.3.6.) A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 5$; $\alpha_B = 0.01$) to each set and a $P \leq \alpha_B$ was considered statistically significant. This procedure was performed on both the globally remodelled and internally remodelled datasets.

2.3. Results

2.3.1: Detection of composite genes identified by both *fdf*BLAST and CompositeSearch

All composite genes were searched against the 54 fusions (of a reported 63) provided by Leonard and Richards (2012) using BLASTp ($E \leq 1e^{-50}$) and a mutual coverage of 90%. The reciprocal best hits were selected based on lowest *e*-value score. This analysis returned hits for 52 of the 58 queried fusion sequences provided by Leonard and Richard (SI; 2012) (Table 2.3.1.) thus illustrating the effectiveness of CompositeSearch. Fusions unidentified by CompositeSearch did not possess two non-overlapping component families.

2.3.2: Effect of controlling remodelled family sizes for type I error reduction

Investigations into whether controlling composite and component family sizes would minimize Type I errors in composite detection were performed. It was found that controlling either composites or components lead to a considerable reduction in errors. Restricting component families to contain a minimum of 2 constituent genes and restricting composite families to contain a minimum of 2 *bona fide* composite genes ($-x_{(bf)} \geq 2$; $-y \geq 2$) lead to a statistically insignificant difference between the control and experimental datasets (Table 2.3.2.; Figure 2.3.1.). These results indicate that these measures are successful in curtailing

Table 2.3.1. Composite genes identified by both *fdf*BLAST and CompositeSearch

Each composite gene identified by both packages was assigned its "Fusion ID" (Leonard and Richards, 2012), its composite family ID, and its remodelling category. A total of 52 genes were identified, of which 42 (80.77%) were observed to be nested composites.

Fusion ID	Composite family	Remodelling category
Fusion_46_XP_013958031	F59067	Strict composite
Fusion_10_XP_011394556	F26749	Nested composite
Fusion_11_XP_011387847	F42983	Nested composite
Fusion_12_XP_001837408	F24039	Nested composite
Fusion_13_XP_011392186	F49622	Nested composite
Fusion_14_XP_011391823	F153	Strict composite
Fusion_15_XP_011386537	F25597	Nested composite
Fusion_16_XP_002475223	F11272	Nested composite
Fusion_17_XP_013025771	F10645	Nested composite
Fusion_18_XP_013024184	F10952	Nested composite
Fusion_19_NP_593279	F39079	Nested composite
Fusion_1_XP_011390436	F62013	Strict composite
Fusion_20_NP_593238	F32521	Nested composite
Fusion_21_NP_594836	F45115	Strict composite
Fusion_22_NP_595325	F82042	Strict composite
Fusion_23_XP_013021078	F567	Nested composite
Fusion_24_NP_587822	F36060	Nested composite
Fusion_25_NP_588353	F26239	Nested composite
Fusion_26_NP_588314	F30380	Nested composite
Fusion_27_XP_001729752	F46089	Strict composite
Fusion_29_XP_001729201	F21421	Nested composite
Fusion_2_NP_009575	F8546	Nested composite
Fusion_30_NP_594305	F55554	Nested composite
Fusion_31_Ept02722	F13970	Nested composite
Fusion_33_XP_003711695	F22075	Nested composite

Fusion ID	Composite family	Remodelling category
Fusion_34_XP_001402280	F44699	Nested composite
Fusion_35_XP_964702	F26746	Nested composite
Fusion_36_XP_002470642	F33988	Nested composite
Fusion_37_XP_003655292	F21030	Nested composite
Fusion_39_XP_013939227	F41438	Nested composite
Fusion_3_NP_009652	F21777	Nested composite
Fusion_40_XP_001273409	F6256	Nested composite
Fusion_41_Elq37550	F26451	Nested composite
Fusion_43_XP_003303532	F22310	Nested composite
Fusion_44_Eha22344	F41895	Nested composite
Fusion_45_XP_006453987	F30712	Nested composite
Fusion_47_XP_001932096	F6215	Nested composite
Fusion_48_XP_003296895	F70420	Nested composite
Fusion_4_NP_014575	F50877	Nested composite
Fusion_50_50574	F39066	Nested composite
Fusion_51_XP_012052038	F39078	Nested composite
Fusion_53_XP_001903306	F3907	Strict composite
Fusion_55_XP_008030890	F43617	Strict composite
Fusion_57_Aaw40768	F45813	Nested composite
Fusion_58_XP_001220095	F61041	Strict composite
Fusion_59_XP_008024409	F3806	Nested composite
Fusion_60_Oal07102	F42519	Nested composite
Fusion_62_XP_006682148	F80820	Nested composite
Fusion_6_Kne57841	F45915	Strict composite
Fusion_7_Kne72089	F590	Nested composite
Fusion_8_Xp_011388676	F21108	Nested composite
Fusion_9_Gaq46133	F48168	Nested composite

Table 2.3.2. Effect of poor genome quality on composite detection analyses

Each of the five detection stringency criteria (composites ≥ *x*; components ≥ *y*) are displayed for each of the 100 experimental iterations. Within each criteria, for each iteration, five values are presented, (*i*.) the sum of composites detected in the control (C), and (*ii*.) experimental (E) datasets with (*iii*., *iv*.) the sums of type I errors associated with C and E, and (*v*.) the *P*-value derived from each iteration. Insignificant results ($P > 0.05$; $n = 100$ (100%)) were only observed when *bona fide* composite and component family sizes were controlled, illustrating the effectiveness of this implementation in controlling type I errors.

	Composites ≥ 1, Components ≥ 1					Composites ≥ 2, Components ≥ 1					Composites ≥ 1, Components ≥ 2					Composites ≥ 2, Components ≥ 2					Composites (*bona fide*) ≥ 2, Components ≥ 2				
	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*
1	8113	9098	405	1390	<0.0001	2683	3777	134	1228	<0.0001	2818	3966	140	1288	<0.0001	1105	1137	55	87	0.0058	658	669	32	43	0.132
2	8683	9881	434	1632	<0.0001	2985	4400	149	1564	<0.0001	2953	4187	55	1289	<0.0001	1181	1211	59	89	0.0104	506	513	25	32	0.222
3	8883	10018	444	1579	<0.0001	2826	3878	141	1193	<0.0001	2922	3980	56	1114	<0.0001	1126	1149	56	79	0.0333	497	500	24	27	0.396
4	8771	9897	438	1564	<0.0001	2839	3999	141	1301	<0.0001	2928	4330	57	1459	<0.0001	1073	1098	53	78	0.0211	510	519	25	34	0.158
5	8775	9933	438	1596	<0.0001	2774	3643	138	1007	<0.0001	2762	3817	58	1113	<0.0001	1132	1159	56	83	0.0163	458	466	22	30	0.175
6	8317	9330	415	1428	<0.0001	2703	3900	135	1332	<0.0001	2806	3960	59	1213	<0.0001	1129	1158	56	85	0.0111	559	564	27	32	0.309
7	8877	10011	443	1577	<0.0001	2616	3453	130	967	<0.0001	2741	3920	60	1239	<0.0001	1197	1233	59	95	0.0031	530	536	26	32	0.264
8	8400	9555	420	1575	<0.0001	2574	3825	128	1379	<0.0001	2851	3912	61	1122	<0.0001	1119	1153	55	89	0.0038	655	663	32	40	0.213
9	8339	9418	416	1495	<0.0001	2966	3885	148	1067	<0.0001	2973	4304	62	1393	<0.0001	1066	1091	53	78	0.0211	606	617	30	41	0.126
10	8130	9208	406	1484	<0.0001	2537	3485	126	1074	<0.0001	2812	3870	63	1121	<0.0001	1208	1239	60	91	0.009	497	505	24	32	0.184
11	8721	9912	436	1627	<0.0001	2684	3738	134	1188	<0.0001	2881	3893	64	1076	<0.0001	1168	1203	58	93	0.0037	641	651	32	42	0.156
12	8585	9765	429	1609	<0.0001	2543	3329	127	913	<0.0001	2968	3948	65	1045	<0.0001	1143	1176	57	90	0.0053	671	676	33	38	0.325
13	8591	9779	429	1617	<0.0001	2946	3951	147	1152	<0.0001	2835	4134	66	1365	<0.0001	1077	1107	53	83	0.008	688	695	34	41	0.252
14	8571	9704	428	1561	<0.0001	2767	4034	138	1405	<0.0001	2890	4039	67	1216	<0.0001	1092	1116	54	78	0.0262	701	706	35	40	0.329
15	8293	9412	414	1533	<0.0001	2814	3801	140	1127	<0.0001	2825	4104	68	1347	<0.0001	1077	1105	53	81	0.0119	548	552	27	31	0.354
16	8007	9061	400	1454	<0.0001	2527	3757	126	1356	<0.0001	2839	3946	69	1176	<0.0001	1185	1209	59	83	0.0307	499	508	24	33	0.153

	Composites ≥ 1, Components ≥ 1					Composites ≥ 2, Components ≥ 1					Composites ≥ 1, Components ≥ 2					Composites ≥ 2, Components ≥ 2					Composites (*bona fide*) ≥ 2, Components ≥ 2				
	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*
17	8652	9820	432	1600	<0.0001	2820	4062	141	1383	<0.0001	2757	4080	70	1393	<0.0001	1154	1190	57	93	0.0028	530	534	26	30	0.351
18	8196	9223	409	1436	<0.0001	2830	3714	141	1025	<0.0001	2945	3865	71	991	<0.0001	1166	1192	58	84	0.021	486	488	24	26	0.448
19	8652	9766	432	1546	<0.0001	2562	3742	128	1308	<0.0001	2984	4181	72	1269	<0.0001	1170	1197	58	85	0.0176	691	699	34	42	0.22
20	8874	9985	443	1554	<0.0001	2957	4132	147	1322	<0.0001	2773	3735	73	1035	<0.0001	1145	1169	57	81	0.0289	522	525	26	29	0.399
21	8536	9694	426	1584	<0.0001	2785	4030	139	1384	<0.0001	2886	4188	74	1376	<0.0001	1189	1212	59	82	0.0363	568	573	28	33	0.312
22	8027	9072	401	1446	<0.0001	2746	3594	137	985	<0.0001	2986	4019	75	1108	<0.0001	1207	1242	60	95	0.0041	637	643	31	37	0.28
23	8708	9846	435	1573	<0.0001	2682	3602	134	1054	<0.0001	2814	3903	76	1165	<0.0001	1129	1160	56	87	0.0075	526	534	26	34	0.192
24	8636	9844	431	1639	<0.0001	2966	4333	148	1515	<0.0001	2863	4219	77	1433	<0.0001	1114	1147	55	88	0.0047	612	619	30	37	0.24
25	8831	10024	441	1634	<0.0001	2747	4029	137	1419	<0.0001	2900	4183	78	1361	<0.0001	1095	1117	54	76	0.0372	572	577	28	33	0.312
26	8777	9954	438	1615	<0.0001	2557	3789	127	1359	<0.0001	2868	3950	79	1161	<0.0001	1139	1165	56	82	0.0196	609	619	30	40	0.15
27	8635	9699	431	1495	<0.0001	2675	3689	133	1147	<0.0001	2982	4363	80	1461	<0.0001	1094	1124	54	84	0.0084	548	555	27	34	0.23
28	8738	9834	436	1532	<0.0001	2805	3692	140	1027	<0.0001	2848	4206	81	1439	<0.0001	1111	1143	55	87	0.0058	487	495	24	32	0.184
29	8866	10053	443	1630	<0.0001	2795	3726	139	1070	<0.0001	2872	3847	82	1057	<0.0001	1070	1102	53	85	0.0053	555	566	27	38	0.116
30	8621	9786	431	1596	<0.0001	2999	4388	149	1538	<0.0001	2784	3725	83	1024	<0.0001	1064	1093	53	82	0.0098	496	501	24	29	0.299
31	8199	9228	409	1438	<0.0001	2907	4288	145	1526	<0.0001	2863	4242	84	1463	<0.0001	1137	1169	56	88	0.0061	596	599	29	32	0.404
32	8925	10002	446	1523	<0.0001	2681	3639	134	1092	<0.0001	2888	4191	85	1388	<0.0001	1115	1148	55	88	0.0047	476	478	23	25	0.447
33	8758	9891	437	1570	<0.0001	2838	3699	141	1002	<0.0001	2798	4028	86	1316	<0.0001	1126	1154	56	84	0.0135	515	518	25	28	0.397
34	8203	9203	410	1410	<0.0001	2790	4039	139	1388	<0.0001	2889	4059	87	1257	<0.0001	1091	1119	54	82	0.0124	447	449	22	24	0.446
35	8922	10112	446	1636	<0.0001	2818	3696	140	1018	<0.0001	2787	4155	88	1456	<0.0001	1065	1098	53	86	0.0043	472	478	23	29	0.253
36	8770	9893	438	1561	<0.0001	2532	3639	126	1233	<0.0001	2924	4299	89	1464	<0.0001	1177	1204	58	85	0.0175	519	523	25	29	0.348
37	8935	10010	446	1521	<0.0001	2633	3849	131	1347	<0.0001	2773	3769	90	1086	<0.0001	1092	1120	54	82	0.0124	598	603	29	34	0.314
38	8370	9434	418	1482	<0.0001	2568	3460	128	1020	<0.0001	2761	4027	91	1357	<0.0001	1183	1215	59	91	0.0071	600	606	30	36	0.277
39	8217	9219	410	1412	<0.0001	2660	3791	133	1264	<0.0001	2898	4045	92	1239	<0.0001	1059	1090	52	83	0.0061	471	473	23	25	0.447
40	8822	9949	441	1568	<0.0001	2838	4160	141	1463	<0.0001	2867	3831	93	1057	<0.0001	1139	1163	56	80	0.028	565	570	28	33	0.312
41	8713	9910	435	1632	<0.0001	2504	3498	125	1119	<0.0001	2789	3885	94	1190	<0.0001	1063	1094	53	84	0.0065	518	524	25	31	0.26
42	8560	9710	428	1578	<0.0001	2801	4162	140	1501	<0.0001	2817	4109	95	1387	<0.0001	1087	1116	54	83	0.0102	517	526	25	34	0.158
43	8486	9617	424	1555	<0.0001	2739	3996	136	1393	<0.0001	2965	4312	96	1443	<0.0001	1136	1163	56	83	0.0163	527	530	26	29	0.399
44	8358	9452	417	1511	<0.0001	2969	3906	148	1085	<0.0001	2761	3722	97	1058	<0.0001	1062	1094	53	85	0.0053	572	579	28	35	0.234
45	8747	9899	437	1589	<0.0001	2641	3551	132	1042	<0.0001	2883	4189	98	1404	<0.0001	1198	1225	59	86	0.0182	577	582	28	33	0.312
46	8639	9845	431	1637	<0.0001	2974	3866	148	1040	<0.0001	2851	3932	99	1180	<0.0001	1154	1178	57	81	0.0289	691	697	34	40	0.288
47	8272	9290	413	1431	<0.0001	2790	4093	139	1442	<0.0001	2760	4127	100	1467	<0.0001	1088	1118	54	84	0.0084	571	580	28	37	0.169
48	8617	9761	430	1574	<0.0001	2863	3836	143	1116	<0.0001	2851	3825	101	1075	<0.0001	1130	1161	56	87	0.0075	490	499	24	33	0.154
49	8380	9538	419	1577	<0.0001	2686	3684	134	1132	<0.0001	2810	3984	102	1276	<0.0001	1051	1074	52	75	0.0293	620	623	31	34	0.407
50	8901	10037	445	1581	<0.0001	2514	3374	125	985	<0.0001	2839	3887	103	1151	<0.0001	1155	1188	57	90	0.0052	504	513	25	34	0.158
51	8846	9992	442	1588	<0.0001	2889	4283	144	1538	<0.0001	2967	4245	104	1382	<0.0001	1152	1175	57	80	0.0343	502	505	25	28	0.398

	Composites ≥ 1, Components ≥ 1					Composites ≥ 2, Components ≥ 1					Composites ≥ 1, Components ≥ 2					Composites ≥ 2, Components ≥ 2					Composites (*bona fide*) ≥ 2, Components ≥ 2				
	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*	C	E	Type I Errors (C)	Type I Errors (E)	*P*
52	8909	9999	445	1535	<0.0001	2712	3963	135	1386	<0.0001	2923	4063	105	1245	<0.0001	1126	1160	56	90	0.0041	578	587	28	37	0.169
53	8852	9914	442	1504	<0.0001	2586	3841	129	1384	<0.0001	2962	3957	106	1101	<0.0001	1069	1096	53	80	0.0145	480	489	24	33	0.154
54	8309	9455	415	1561	<0.0001	2501	3727	125	1351	<0.0001	2775	3688	107	1020	<0.0001	1080	1105	54	79	0.0219	528	534	26	32	0.264
55	8856	9945	442	1531	<0.0001	2547	3583	127	1163	<0.0001	2826	3972	108	1254	<0.0001	1172	1203	58	89	0.0082	509	514	25	30	0.303
56	8342	9494	417	1569	<0.0001	2602	3640	130	1168	<0.0001	2800	3930	109	1239	<0.0001	1074	1106	53	85	0.0053	516	524	25	33	0.188
57	8642	9808	432	1598	<0.0001	2694	3966	134	1406	<0.0001	2897	4202	110	1415	<0.0001	1109	1140	55	86	0.0072	509	519	25	35	0.131
58	8074	9134	403	1463	<0.0001	2656	3579	132	1055	<0.0001	2829	4165	111	1447	<0.0001	1157	1188	57	88	0.0079	460	465	23	28	0.296
59	8450	9485	422	1457	<0.0001	2998	4342	149	1493	<0.0001	2882	4158	112	1388	<0.0001	1075	1108	53	86	0.0042	599	608	29	38	0.173
60	8493	9649	424	1580	<0.0001	2882	4223	144	1485	<0.0001	2890	3915	113	1138	<0.0001	1132	1163	56	87	0.0075	599	602	29	32	0.404
61	8825	9955	441	1571	<0.0001	2676	3893	133	1350	<0.0001	2794	3878	114	1198	<0.0001	1202	1239	60	97	0.0027	536	541	26	31	0.306
62	8389	9518	419	1548	<0.0001	2541	3348	127	934	<0.0001	2776	3984	115	1323	<0.0001	1182	1205	59	82	0.0363	470	472	23	25	0.447
63	8338	9357	416	1435	<0.0001	2967	3907	148	1088	<0.0001	2808	3743	116	1051	<0.0001	1207	1238	60	91	0.009	496	500	24	28	0.346
64	8061	9165	403	1507	<0.0001	2723	3586	136	999	<0.0001	2944	3911	117	1084	<0.0001	1174	1199	58	83	0.0251	507	515	25	33	0.188
65	8625	9762	431	1568	<0.0001	2934	3935	146	1147	<0.0001	2968	4251	118	1401	<0.0001	1089	1114	54	79	0.0219	619	627	30	38	0.207
66	8260	9262	413	1415	<0.0001	2615	3727	130	1242	<0.0001	2890	4071	119	1300	<0.0001	1152	1180	57	85	0.014	619	624	30	35	0.317
67	8127	9169	406	1448	<0.0001	2856	4089	142	1375	<0.0001	2842	4165	120	1443	<0.0001	1080	1102	54	76	0.0372	503	508	25	30	0.303
68	8493	9528	424	1459	<0.0001	2971	4341	148	1518	<0.0001	2987	4335	121	1469	<0.0001	1200	1234	60	94	0.005	566	573	28	35	0.234
69	8656	9850	432	1626	<0.0001	2775	3707	138	1070	<0.0001	2919	4120	122	1323	<0.0001	1182	1216	59	93	0.0047	549	558	27	36	0.166
70	8534	9720	426	1612	<0.0001	2766	4015	138	1387	<0.0001	2955	4241	123	1409	<0.0001	1163	1191	58	86	0.0146	647	657	32	42	0.156
71	8141	9263	407	1529	<0.0001	2873	4123	143	1393	<0.0001	2762	3729	124	1091	<0.0001	1177	1204	58	85	0.0175	500	507	25	32	0.223
72	8755	9854	437	1536	<0.0001	2738	4097	136	1495	<0.0001	2793	3818	125	1150	<0.0001	1104	1132	55	83	0.013	672	676	33	37	0.366
73	8346	9355	417	1426	<0.0001	2600	3462	130	992	<0.0001	2922	3896	126	1100	<0.0001	1077	1104	53	80	0.0144	576	580	28	32	0.356
74	8211	9269	410	1468	<0.0001	2894	3982	144	1232	<0.0001	2923	4198	127	1402	<0.0001	1201	1226	60	85	0.0267	553	559	27	33	0.268
75	8765	9901	438	1574	<0.0001	2506	3368	125	987	<0.0001	2899	4132	128	1361	<0.0001	1082	1105	54	77	0.0313	604	607	30	33	0.406
76	8702	9804	435	1537	<0.0001	2762	3891	138	1267	<0.0001	2951	4251	129	1429	<0.0001	1174	1205	58	89	0.0082	506	509	25	28	0.398
77	8165	9159	408	1402	<0.0001	2931	4024	146	1239	<0.0001	2789	3774	130	1115	<0.0001	1120	1143	56	79	0.0333	617	628	30	41	0.126
78	8317	9384	415	1482	<0.0001	2733	3575	136	978	<0.0001	2921	3856	131	1066	<0.0001	1166	1198	58	90	0.0068	621	627	31	37	0.28
79	8826	9915	441	1530	<0.0001	2610	3888	130	1408	<0.0001	2793	3795	132	1134	<0.0001	1141	1174	57	90	0.0053	532	538	26	32	0.264
80	8419	9576	420	1577	<0.0001	2806	4137	140	1471	<0.0001	2906	3837	133	1064	<0.0001	1186	1211	59	84	0.0259	636	646	31	41	0.153
81	8954	10190	447	1683	<0.0001	2794	3816	139	1161	<0.0001	2840	3899	134	1193	<0.0001	1169	1192	58	81	0.0353	575	583	28	36	0.2
82	8377	9409	418	1450	<0.0001	2683	3582	134	1033	<0.0001	2895	4070	135	1310	<0.0001	1191	1217	59	85	0.0217	649	658	32	41	0.183
83	8488	9621	424	1557	<0.0001	2592	3506	129	1043	<0.0001	2955	4074	136	1255	<0.0001	1109	1141	55	87	0.0058	544	549	27	32	0.309
84	8339	9489	416	1566	<0.0001	2776	4010	138	1372	<0.0001	2897	4286	137	1526	<0.0001	1116	1143	55	82	0.0157	600	609	30	39	0.177
85	8861	10018	443	1600	<0.0001	2802	3771	140	1109	<0.0001	2881	3816	138	1073	<0.0001	1059	1085	52	78	0.0168	472	475	23	26	0.394
86	8539	9646	426	1533	<0.0001	2872	3891	143	1162	<0.0001	2777	3688	139	1050	<0.0001	1199	1237	59	97	0.002	610	617	30	37	0.24

	Composites ≥ 1, Components ≥ 1					Composites ≥ 2, Components ≥ 1					Composites ≥ 1, Components ≥ 2					Composites ≥ 2, Components ≥ 2					Composites (*bona fide*) ≥ 2, Components ≥ 2				
	C	**E**	**Type I Errors (C)**	**Type I Errors (E)**	***P***	**C**	**E**	**Type I Errors (C)**	**Type I Errors (E)**	***P***	**C**	**E**	**Type I Errors (C)**	**Type I Errors (E)**	***P***	**C**	**E**	**Type I Errors (C)**	**Type I Errors (E)**	***P***	**C**	**E**	**Type I Errors (C)**	**Type I Errors (E)**	***P***
87	8062	9066	403	1407	<0.0001	2982	4250	149	1417	<0.0001	2765	3644	140	1019	<0.0001	1199	1235	59	95	0.0031	534	540	26	32	0.264
88	8876	9982	443	1549	<0.0001	2778	3803	138	1163	<0.0001	2823	4017	141	1335	<0.0001	1072	1097	53	78	0.0211	476	482	23	29	0.253
89	8915	9984	445	1514	<0.0001	2758	3773	137	1152	<0.0001	2747	3763	142	1158	<0.0001	1203	1239	60	96	0.0033	595	603	29	37	0.203
90	8467	9520	423	1476	<0.0001	2933	3953	146	1166	<0.0001	2799	4125	143	1469	<0.0001	1189	1216	59	86	0.0182	508	513	25	30	0.303
91	8773	9983	438	1648	<0.0001	2532	3347	126	941	<0.0001	2770	3837	144	1211	<0.0001	1069	1097	53	81	0.0119	621	626	31	36	0.32
92	8211	9236	410	1435	<0.0001	2626	3728	131	1233	<0.0001	2749	3595	145	991	<0.0001	1072	1096	53	77	0.0253	531	537	26	32	0.264
93	8724	9867	436	1579	<0.0001	2946	4394	147	1595	<0.0001	2749	3958	146	1355	<0.0001	1180	1217	59	96	0.0025	610	615	30	35	0.317
94	8701	9883	435	1617	<0.0001	2779	3894	138	1253	<0.0001	2853	3745	147	1039	<0.0001	1114	1143	55	84	0.0107	695	699	34	38	0.368
95	8840	9928	442	1530	<0.0001	2707	3914	135	1342	<0.0001	2925	4056	148	1279	<0.0001	1061	1093	53	85	0.0053	512	517	25	30	0.303
96	8676	9862	433	1619	<0.0001	2705	3610	135	1040	<0.0001	2803	4057	149	1403	<0.0001	1191	1215	59	83	0.0307	452	459	22	29	0.21
97	8037	9040	401	1404	<0.0001	2831	3922	141	1232	<0.0001	2901	4063	150	1312	<0.0001	1197	1226	59	88	0.0126	617	625	30	38	0.207
98	8690	9900	434	1644	<0.0001	2765	3925	138	1298	<0.0001	2810	4019	151	1360	<0.0001	1150	1173	57	80	0.0343	535	543	26	34	0.192
99	8502	9589	425	1512	<0.0001	2779	4103	138	1462	<0.0001	2770	3663	152	1045	<0.0001	1056	1085	52	81	0.0093	619	624	30	35	0.317
100	8914	10160	445	1691	<0.0001	2780	4082	139	1441	<0.0001	2832	4120	153	1441	<0.0001	1209	1241	60	92	0.0074	617	629	30	42	0.105

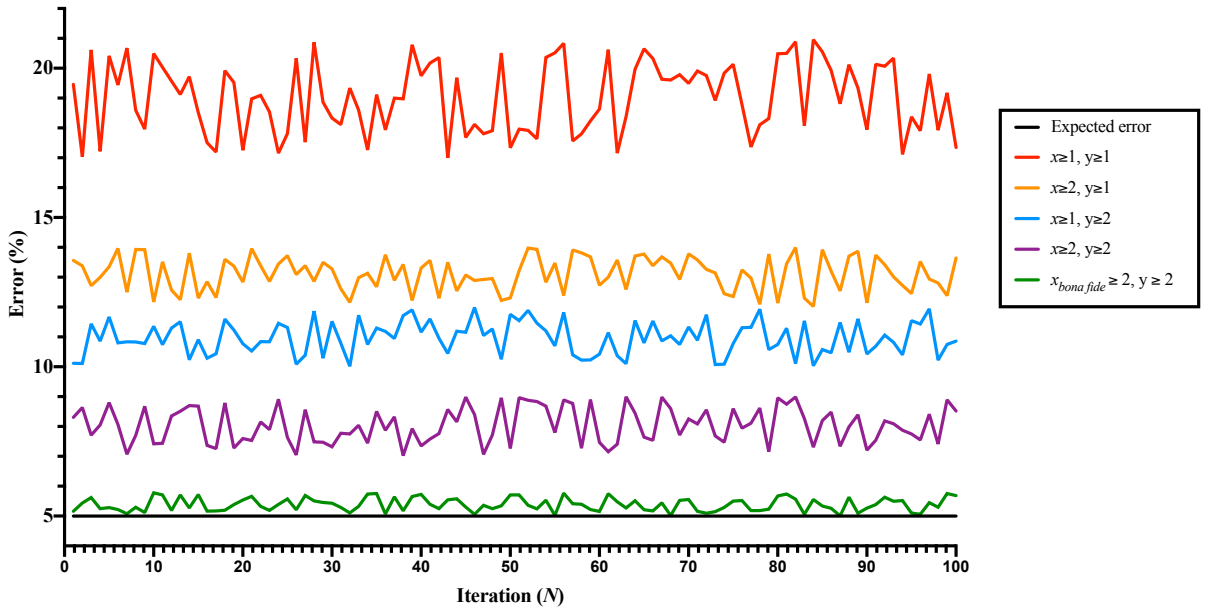

Figure 2.3.1: Effect of poor genome quality on composite detection analyses

Each error rate (C and E) for the five stringency criteria are plotted (as annotated in the legend) as per Table 2.3.2. This figure serves as a visual aid to further demonstrate the effectiveness of controlling the sum of *bona fide* composite genes in a composite family alongside controlling component family sizes in comparison to other stringency criteria.

Type I errors resulting from potential cases poor genome annotation and was therefore used in every CompositeSearch analysis throughout this thesis.

2.3.3. Fungal genome dataset quality control and genomic statistics

Genome completeness and genomic statistics were collated (Table 2.3.3). A mean completeness score (C) of 95.6±4.0% (C ∈ {77.2, …, 100}) was observed across the 107 species (Table 2.2.4). A large standard deviation was observed for duplicated BUSCO genes (2.47±8.01) due to a massive amount of duplicates in *Allomyces macrogynus* (72.1%). As *A. macrogynus* is autotetraploid (Emerson and Wilson, 1954; Albertin and Marullo, 2012), such high levels of duplication in expected single copy orthologs is not surprising.

2.3.4. Gene remodelling is rampant in fungi

In total, of the 1,150,918 protein coding sequences in our dataset, 111,768 (9.71%) were excluded due to low complexity as detected by BLAST or due to being a singleton, resulting in a sample of 1,039,150 genes within 81,476 non-singleton families (Table 2.3.5; Figure 2.3.2). Remodelled genes accounted for approximately 73.89% of all sampled genes and 50.39% of all families, with the remaining 26.11% of genes and 49.69% of families being non-remodelled. Nested composites had the highest representation of all remodelled genes, accounting for 33.97% of all sampled genes (and 21.25% of all families) and strict components were best represented amongst remodelled families, accounting for 26.51% of all families (and 18.72% of all genes). These results illustrate that approximately 60.13% of all fungal genes, 68.04% of all sampled genes, and 49.69% of families have a history in remodelling.

Table 2.3.3. Completeness and characteristics for 107 fungal genomes

Each taxon is annotated with its genome size (GS), GC content, number of genes, genome density, and genome completeness. Genome completeness is given as a percentage of expected orthologs where C is "completeness". Completeness is the cumulation of singleton (S) and duplicated (D) orthologs. Fragmented (F) and missing (M) orthologs detract from C. Taxa are arranged based on their taxonomic clades.

		Genomic statistics				BUSCO completeness (%) (fungi_odb *v.*09; *n* = 230)				
Binomial classification	**Taxonomic clade**	**GS (Mbp)**	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Dothistroma septosporum	Ascomycota; Pezizomycotina	30.68	52.4	12580	410.039	96.9	96.9	0	2.4	0.7
Mycosphaerella fijiensis CIRAD86	Ascomycota; Pezizomycotina	73.73	45.2	10313	139.875	93.5	92.8	0.7	2.1	4.4
Mycosphaerella graminicola IPO323	Ascomycota; Pezizomycotina	39.68	51.2	10933	275.529	97.9	97.9	0	0.7	1.4
Septoria musiva	Ascomycota; Pezizomycotina	28.92	51.1	10233	353.838	98.9	98.6	0.3	1	0.1
Septoria populicola	Ascomycota; Pezizomycotina	32.11	50.3	9739	303.301	99	99	0	1	0
Baudoinia compniacensis	Ascomycota; Pezizomycotina	21.88	54.8	10513	480.484	95.9	95.9	0	3.4	0.7
Hysterium pulicare	Ascomycota; Pezizomycotina	38.25	48.8	12352	322.928	94.8	93.8	1	4.5	0.7
Rhytidhysteron rufulum	Ascomycota; Pezizomycotina	39.86	47.9	12117	303.989	94.8	94.1	0.7	3.1	2.1
Alternaria brassicicola ATCC 96836	Ascomycota; Pezizomycotina	31.04	50.8	10688	344.33	85.9	84.5	1.4	10	4.1
Cochliobolus heterostrophus	Ascomycota; Pezizomycotina	32.09	50.7	9633	300.187	96.2	96.2	0	1	2.8
Cochliobolus sativus	Ascomycota; Pezizomycotina	33.21	49.8	12250	368.865	98.6	98.3	0.3	0.3	1.1
Leptosphaeria maculans	Ascomycota; Pezizomycotina	45.12	45.3	12469	276.352	94.4	94.1	0.3	4.8	0.8
Pyrenophora teres	Ascomycota; Pezizomycotina	54.1	45.3	11799	218.096	98.2	97.2	1	1	0.8
Pyrenophora triticirepentis strain Pt1CBFP	Ascomycota; Pezizomycotina	37.36	50.9	12169	325.723	96.2	95.9	0.3	2.4	1.4

		Genomic statistics				BUSCO completeness (%) (fungi_odb *v.*09; *n* = 230)				
Binomial classification	**Taxonomic clade**	**GS (Mbp)**	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Setosphaeria turcica	Ascomycota; Pezizomycotina	43.01	51.2	11702	272.076	96.2	95.5	0.7	2.8	1
Aspergillus aculeatus	Ascomycota; Pezizomycotina	35.19	50.9	10828	307.701	97.9	97.2	0.7	0.7	1.4
Aspergillus carbonarius	Ascomycota; Pezizomycotina	34.12	51.7	11624	340.68	93.1	87.2	5.9	1	5.9
Aspergillus clavatus	Ascomycota; Pezizomycotina	27.85	45.2	9120	327.469	98.6	98.6	0	1	0.4
Aspergillus flavus	Ascomycota; Pezizomycotina	39.91	48.4	12587	315.385	93.8	92.8	1	4.5	1.7
Aspergillus fumigatus Af293	Ascomycota; Pezizomycotina	28.81	49.8	9887	343.179	96.5	96.2	0.3	2.4	1.1
Aspergillus nidulans FGSCA4	Ascomycota; Pezizomycotina	30.07	50.3	10560	351.181	98.9	98.6	0.3	1	0.1
Aspergillus oryzae RIB40	Ascomycota; Pezizomycotina	36.58	48.3	12063	329.77	89.6	88.6	1	4.1	6.3
Aspergillus terreus NIH 2624	Ascomycota; Pezizomycotina	29.23	52.9	10406	356.004	90	89	1	6.2	3.8
Neosartorya fischeri (NRRL 181)	Ascomycota; Pezizomycotina	31.77	49.5	10403	327.447	98.2	97.9	0.3	1.4	0.4
Blastomyces dermatitidis	Ascomycota; Pezizomycotina	66.27	37.1	9522	143.685	95.8	95.5	0.3	3.8	0.4
Histoplasma capsulatum (strain NAm1)	Ascomycota; Pezizomycotina	33.03	46.3	9251	280.079	86.6	86.6	0	10.7	2.7
Paracoccidioides brasiliensis Pb01	Ascomycota; Pezizomycotina	32.93	42.9	9136	277.437	94.5	94.5	0	4.1	1.4
Microsporum canis CBS 113480	Ascomycota; Pezizomycotina	23.26	47.5	8765	376.827	97.9	97.2	0.7	1.7	0.4
Microsporum gypseum CBS 118893	Ascomycota; Pezizomycotina	23.27	48.4	8876	381.435	96.6	95.9	0.7	2.8	0.6
Trichophyton equinum CBS127.97	Ascomycota; Pezizomycotina	24.16	47.3	8560	354.305	93.7	93.4	0.3	4.5	1.8
Coccidioides immitis RS	Ascomycota; Pezizomycotina	29.02	46	10654	367.126	97.9	95.5	2.4	1.7	0.4
Coccidioides posadasii str. Silveira	Ascomycota; Pezizomycotina	27.58	46.6	10124	367.078	91.7	91.4	0.3	5.5	2.8
Uncinocarpus reesii	Ascomycota; Pezizomycotina	22.35	48.6	7798	348.904	85.9	85.9	0	10	4.1
Botryotinia cinerea (strain B05.10)	Ascomycota; Pezizomycotina	42.63	42	16448	385.832	89.3	88.6	0.7	8.6	2.1
Sclerotinia sclerotiorum ATCC 18683	Ascomycota; Pezizomycotina	38.91	48.6	14522	373.22	94.8	94.5	0.3	4.5	0.7
Cryphonectria parasitica	Ascomycota; Pezizomycotina	49.6	50.8	11184	225.484	97.9	97.9	0	1.7	0.4
Acremonium alcalophilum	Ascomycota; Pezizomycotina	54.42	46.4	9521	174.954	99.7	99.7	0	0.3	0

		Genomic statistics				BUSCO completeness (%) (fungi_odb *v.*09; *n* = 230)				
Binomial classification	**Taxonomic clade**	**GS (Mbp)**	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Verticillium alboatrum VaMs.102	Ascomycota; Pezizomycotina	36.46	56.5	10220	280.307	77.2	76.9	0.3	19.7	3.1
Verticillium dahliae VdLs.17	Ascomycota; Pezizomycotina	33.9	55.6	10535	310.767	93.4	93.4	0	5.5	1.1
Trichoderma atroviride IMI 202040	Ascomycota; Pezizomycotina	36.14	49.7	11100	307.139	98.2	97.9	0.3	0.3	1.5
Trichoderma reesei QM6a	Ascomycota; Pezizomycotina	32.68	53.6	9143	279.774	99.3	99.3	0	0.3	0.4
Trichoderma virens Gv298	Ascomycota; Pezizomycotina	39.02	49.2	11643	298.385	96.9	96.9	0	1.4	1.7
Fusarium graminearum species complex	Ascomycota; Pezizomycotina	36.46	48.3	13321	365.359	99	99	0	0.7	0.3
Fusarium oxysporum f. sp. lycopersici	Ascomycota; Pezizomycotina	61.39	48.4	17608	286.822	94.4	94.1	0.3	3.8	1.8
Fusarium verticillioides	Ascomycota; Pezizomycotina	41.84	48.7	14195	339.269	95.8	95.5	0.3	3.1	1.1
Nectria haematococca mpVI	Ascomycota; Pezizomycotina	51.29	50.8	15707	306.239	98.2	97.2	1	0.3	1.5
Magnaporthe grisea 7015	Ascomycota; Pezizomycotina	44.56	47.8	11109	249.304	87.9	87.9	0	5.5	6.6
Chaetomium globosum CBS 148.51	Ascomycota; Pezizomycotina	34.34	55.6	11124	323.937	85.9	85.9	0	10.3	3.8
Sporotrichum thermophile	Ascomycota; Pezizomycotina	38.74	51.4	8806	227.31	94.1	94.1	0	3.1	2.8
Thielavia terrestris	Ascomycota; Pezizomycotina	36.91	54.7	9815	265.917	98.3	97.6	0.7	0.3	1.4
Podospora anserina DSM 980	Ascomycota; Pezizomycotina	34.72	52.2	10601	305.328	98.3	97.6	0.7	1.7	0
Neurospora crassa OR74A	Ascomycota; Pezizomycotina	41.1	48.2	9908	241.071	99.3	96.9	2.4	0.7	0
Neurospora tetrasperma FGSC 2508	Ascomycota; Pezizomycotina	39.15	49.4	10640	271.775	97.5	97.2	0.3	1	1.5
Wickerhamomyces anomalus	Ascomycota; Saccharomycotina	14.15	35	6423	453.922	95.2	94.5	0.7	3.1	1.7
Candida albicans SC5314	Ascomycota; Saccharomycotina	14.28	33.5	6205	434.524	99.6	99.3	0.3	0.3	0.1
Candida caseinolytica	Ascomycota; Saccharomycotina	9.18	45.4	4657	507.298	92	91.7	0.3	2.8	5.2
Candida glabrata CBS 138	Ascomycota; Saccharomycotina	12.47	38.6	5202	417.161	99.3	96.9	2.4	0.7	0
Candida tenuis	Ascomycota; Saccharomycotina	10.75	43	5533	514.698	95.5	95.5	0	0.7	3.8
Debaryomyces hansenii CBS767	Ascomycota; Saccharomycotina	12.18	36.3	6272	514.943	100	99.7	0.3	0	0
Spathaspora passalidarum	Ascomycota; Saccharomycotina	13.18	37.5	5983	453.945	96.9	96.9	0	3.1	0

		Genomic statistics				BUSCO completeness (%) (fungi_odb *v.*09; *n* = 230)				
Binomial classification	**Taxonomic clade**	**GS (Mbp)**	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Yarrowia lipolytica CLIB122	Ascomycota; Saccharomycotina	20.55	49	6448	313.771	98.6	98.6	0	1	0.4
Lipomyces starkeyi	Ascomycota; Saccharomycotina	21.27	47	8192	385.143	95.5	93.8	1.7	3.4	1.1
Hansenula polymorpha	Ascomycota; Saccharomycotina	8.9	47.9	5177	581.685	96.2	95.9	0.3	2.8	1
Pichia membranifaciens	Ascomycota; Saccharomycotina	11.43	45.1	5546	485.214	98.3	98.3	0	1	0.7
Pichia stipitis CBS 6054	Ascomycota; Saccharomycotina	15.44	41.2	5807	376.101	97.2	96.9	0.3	1	1.8
Ashbya gossypii ATCC 10895	Ascomycota; Saccharomycotina	9.14	51.7	4717	516.083	98.6	98.6	0	1.4	0
Saccharomyces cerevisiae S288c	Ascomycota; Saccharomycotina	12.16	38.2	5885	483.964	99.6	91.7	7.9	0.3	0.1
Schizosaccharomyces cryophilus oy26	Ascomycota; Taphrinomycotina	11.55	37.7	5057	437.835	99.3	91.7	7.6	0.7	0
Schizosaccharomyces japonicus yFS275	Ascomycota; Taphrinomycotina	11.73	44.1	4814	410.401	98.6	91.4	7.2	1.4	0
Schizosaccharomyces octosporus yFS286	Ascomycota; Taphrinomycotina	11.63	37.9	4925	423.474	99.6	91.7	7.9	0.3	0.1
Schizosaccharomyces pombe 972h	Ascomycota; Taphrinomycotina	12.59	36	5010	397.935	100	91.7	8.3	0	0
Agaricus bisporus var. *burnettii* JB137-S8	Basidiomycota; Agaricomycotina	32.61	46.7	11289	346.182	94.8	94.1	0.7	1.7	3.5
Schizophyllum commune H48	Basidiomycota; Agaricomycotina	38.48	57.4	13181	342.542	97.6	95.9	1.7	1.4	1
Pleurotus ostreatus	Basidiomycota; Agaricomycotina	34.36	50.76	11603	337.689	99.4	96.6	2.8	0.3	0.3
Fomitopsis pinicola	Basidiomycota; Agaricomycotina	41.61	55.4	14724	353.857	98.3	96.6	1.7	1.7	0
Trametes versicolor	Basidiomycota; Agaricomycotina	44.79	57.4	14296	319.178	96.2	95.5	0.7	0.7	3.1
Wolfiporia cocos	Basidiomycota; Agaricomycotina	50.48	52	12746	252.496	99	97.6	1.4	0.3	0.7
Coprinopsis cinerea (strain FGSC 9003)	Basidiomycota; Agaricomycotina	36.19	51.6	13394	370.102	96.9	95.9	1	2.8	0.3
Laccaria bicolor (strain S238NH82)	Basidiomycota; Agaricomycotina	58.68	47	19036	324.404	95.2	93.1	2.1	1.7	3.1
Auricularia delicata	Basidiomycota; Agaricomycotina	43.2	57.1	23577	545.764	97.2	93.4	3.8	1	1.8
Coniophora putinea	Basidiomycota; Agaricomycotina	42.97	52.3	13761	320.247	95.5	93.8	1.7	1	3.5
Serpula lacrymans S7.3	Basidiomycota; Agaricomycotina	42.79	45.3	14495	338.747	93.4	93.1	0.3	3.1	3.5
Phlebia brevispora	Basidiomycota; Agaricomycotina	46.4	52	16170	348.491	97.3	94.5	2.8	2.1	0.6

		Genomic statistics				BUSCO completeness (%) (fungi_odb *v.*09; *n* = 230)				
Binomial classification	**Taxonomic clade**	**GS (Mbp)**	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Punctularia strigosozonata	Basidiomycota; Agaricomycotina	34.17	54.9	11538	337.665	94.5	93.8	0.7	1.4	4.1
Gloeophyllum trabeum	Basidiomycota; Agaricomycotina	37.18	52.9	11846	318.612	97.6	96.6	1	2.1	0.3
Fomitiporia mediterranea	Basidiomycota; Agaricomycotina	63.35	41.7	11333	178.895	98.3	96.9	1.4	0.7	1
Ganoderma sp.	Basidiomycota; Agaricomycotina	43.29	56.1	12910	298.221	99	98.3	0.7	0.7	0.3
Bjerkandera adusta	Basidiomycota; Agaricomycotina	40.2	54.8	15473	384.9	98.3	97.6	0.7	0.7	1
Ceriporiopsis subvermispora	Basidiomycota; Agaricomycotina	37.87	53.8	12125	320.174	88.9	88.6	0.3	1.7	9.4
Phanerochaete chrysosporium RP78	Basidiomycota; Agaricomycotina	29.84	57	10048	336.729	94.5	93.8	0.7	1.4	4.1
Phlebiopsis gigantea	Basidiomycota; Agaricomycotina	27.9	54.7	11891	426.201	98.2	97.2	1	0.7	1.1
Dichomitus squalens	Basidiomycota; Agaricomycotina	42.75	55	12290	287.485	94.4	93.4	1	1	4.6
Heterobasidion annosum	Basidiomycota; Agaricomycotina	27.98	52.9	12299	439.564	95.8	95.5	0.3	3.4	0.8
Dacryopinax sp.	Basidiomycota; Agaricomycotina	29.5	52	10242	347.186	95.9	94.5	1.4	1	3.1
Cryptococcus neoformans var. *grubii* H99	Basidiomycota; Agaricomycotina	18.92	48.2	6967	368.235	98.6	97.9	0.7	1	0.4
Tremella mesenterica	Basidiomycota; Agaricomycotina	28.64	46.8	8313	290.258	97.2	96.9	0.3	2.4	0.4
Rhodotorula graminis	Basidiomycota; Pucciniomycotina	20.78	67.8	7283	350.481	96.2	95.9	0.3	2.8	1
Sporobolomyces roseus IAM 13481	Basidiomycota; Pucciniomycotina	20.8	53.8	5536	266.154	87.2	86.9	0.3	7.9	4.9
Melampsora laricis-populina	Basidiomycota; Pucciniomycotina	97.8	41	16831	172.096	95.9	93.8	2.1	3.4	0.7
Puccinia graminis f. sp. tritici	Basidiomycota; Pucciniomycotina	81.6	43.3	20566	252.034	88.9	80.3	8.6	7.9	3.2
Ustilago maydis 521	Basidiomycota; Ustilaginomycotina	19.66	54	6522	331.74	92.1	91.4	0.7	4.5	3.4
Allomyces macrogynus	Blastocladiomycota	57.06	60.5	17600	308.447	96.6	24.5	72.1	3.1	0.3
Batrachochytrium dendrobatidis JEL423	Chytridiomycota	23.9	39.4	8732	365.356	97.2	94.8	2.4	1.7	1.1
Spizellomyces punctatus	Chytridiomycota	23.91	47.6	8804	368.214	95.2	92.4	2.8	2.4	2.4
Mucor circinelloides f. lusitanicus	Mucoromycota	36.57	42.2	10930	298.879	96.2	73.8	22.4	2.8	1
Phycomyces blakesleeanus	Mucoromycota	53.37	35.8	16528	309.687	94.5	78.3	16.2	2.4	3.1

		Genomic statistics				BUSCO completeness (%) (fungi_odb *v.*09; *n* = 230)				
Binomial classification	**Taxonomic clade**	**GS (Mbp)**	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Rhizopus oryzae RA 99880	Mucoromycota	39.06	35.4	17459	446.979	83.4	52.4	31	12.1	4.5

Table 2.3.4. Descriptive statistics for 107 fungal genomes

Descriptive statistics calculated for each category in Table 2.3.3. Column IDs are annotated as per Table 2.3.3.

	Genome statistics				BUSCO completeness (%)				
	GS (Mbp)	**GC (%)**	**Genes (*n*)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Minimum	8.9	33.5	4657	140	77.2	24.5	0	0	0
$\eta_{0.25}$	23.9	45.3	8765	298	94.5	92.8	0.3	1	0.4
η ($\eta_{0.50}$)	34.2	48.8	10640	338	96.5	95.5	0.7	1.7	1.1
$\eta_{0.75}$	41.6	52.2	12352	376	98.3	97.2	1.4	3.4	2.8
Maximum	97.8	67.8	23577	582	100	99.7	72.1	19.7	9.4
μ	34.2±15.5	48.4±6.11	10757±3632	341.0±81.7	95.6±4.0	93.10±9.14	2.47±8.01	2.72±3.0	1.71±1.76
CV	45.3%	12.6%	33.8%	23.9%	4.18%	9.81%	324%	110%	103%

Table 2.3.5. Extent of remodelled genes and families in fungi

The number (*n*) of genes in the entire dataset ($Genes_{(a)}$), the sum of genes in the sampled dataset ($Genes_{(s)}$; $Genes_{(a)}$ – Excluded), and the number of gene families attributed to each RC are presented with their associated proportion (%) within their respective populations. The "Excluded" category is only observed in $Genes_{(a)}$ as genes in this category were not used for sampling by CompositeSearch, and thus were excluded from $Genes_{(s)}$ and Families respectively

	n			**%**		
RC	**$Genes_{(a)}$**	**$Genes_{(s)}$**	**Families**	**$Genes_{(a)}$**	**$Genes_{(s)}$**	**Families**
NC	353039	353039	17317	30.67	41.58	21.25
SC	30110	30110	2067	2.62	3.55	2.54
SN	194513	194513	21603	16.90	22.91	26.51
NR	271309	271309	40489	23.57	31.96	49.69
Excluded	301947	N/A	N/A	26.24	N/A	N/A

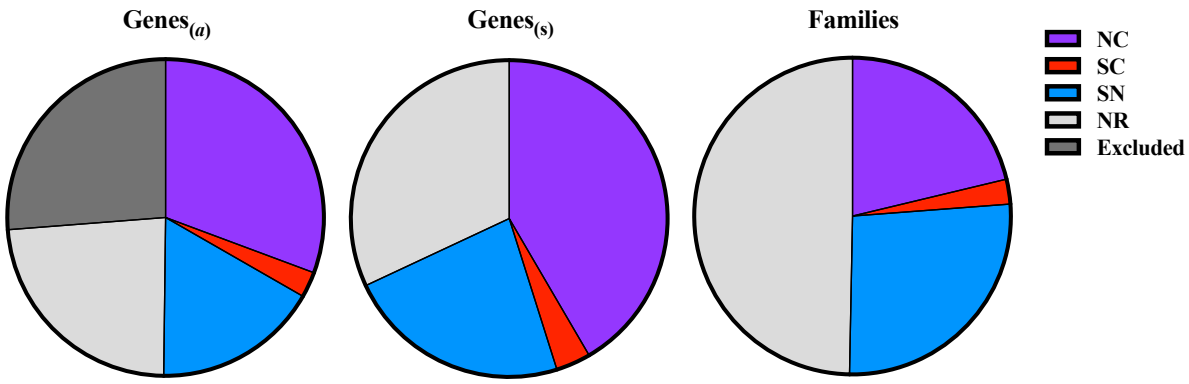

Figure 2.3.2. Extent of remodelled genes and families in fungi

Each pie chart represents one of three datasets (Genes$_{(a)}$, Genes$_{(s)}$, and Families) from Table 2.3.2. Again, the "Excluded" category is only observed in Genes$_{(a)}$ as genes in this category were not used for sampling by CompositeSearch, and thus were excluded from Genes$_{(s)}$ and Families respectively

2.3.5. Variance in gene family sizes

Gene families were observed to display incredible variation for each RC (C.V. = 145.8% - 294.6%) with considerable bias observed towards smaller families ($\eta_{0.25} \leq 3$; $\eta_{0.50} \leq 5$; $\eta_{0.75} \leq 18$ (Table 2.3.6.). Nested composites were observed to be most variant (C.V. = 294.6%) and to have the largest range ($2 \geq n \leq 1933$) compared to other categories ($2 \geq n \leq 320$). On average, nested composites were reported to form larger families ($n = 20.39 \pm 60.07$) than other RC categories ($n_{SC} = 14.57 \pm 21.25$; $n_{SN} = 9.0 \pm 16.81$; $n_{NR} = 6.7 \pm 12.15$) (Figure 2.3.4.). Family sizes for each RC were observed to be significantly different ($P \leq \alpha_B \leq 8.33e^{-03}$) from each other ($P \leq 2.79e^{-05}$) with the exception of NC *vs.* NR ($P = 0.11$) using a two-tailed Mann-Whitney *U* test (Table 2.3.7.).

2.3.6. Comparison of evolutionary rates

2.3.6.1. Phylogenetic annotation

A phylogeny was constructed using 277 highly distributed (present in $n \geq 102$ (~95%) species) KOG gene family alignments (Figure 2.3.3.). Internal nodes were annotated as per the "-apo" function in TNT (Figure 2.3.4.). All major clades are in their correct placements as Chyrididomycota was basal to Blastocladiomycota (thus constituting the "zoosporic true fungi" (ZTF)), the ZTF were basal to the Mucoromycota (thus forming the "monokaryotes"), and the monokaryotes were basal to the Dikarya (Spatafora *et al*., 2016).

Table 2.3.6. Descriptive statistics for fungal gene family sizes

Statistics describing family size distribution characteristics were tabulated for each RC. Each RC is assigned a mean (μ), median (η), quartiles ($\eta_{0.25, 0.50, 0.75}$), minima, maxima, and CV.

	NC	SC	SN	NR
n	17317	2067	21603	40489
Minimum	2.00	2.00	2.00	2.00
$\eta_{0.25}$	3.00	2.00	2.00	2.00
η ($\eta_{0.50}$)	5.00	5.00	3.00	3.00
$\eta_{0.75}$	18.0	17.0	7.00	5.00
Maximum	1933	178	320	225
μ	20.4±60.1	14.6±21.2	9.00±16.8	6.70±12.1
CV	295%	146%	187%	181%

Table 2.3.7. Comparison of fungal family size distributions

Samples (*a*.) and (*b*.) refer to the RCs being tested. The Mann-Whitney U statistic (U) is provided alongside a *P*-value for each comparison. All comparisons were considered statistically significant ($P \leq \alpha_B \leq 8.33e^{-03}$) except for NC *vs.* NR ($P = 0.11$).

Sample (*a*.)	**Sample (*b*.)**	**U**	***P***
NC	SC	$7.19e^{08}$	$6.48e^{-38}$
NC	SN	$5.41e^{08}$	$6.19e^{-89}$
NC	NR	$4.41e^{08}$	0.11
SC	SN	$4.28e^{08}$	$5.00e^{-15}$
SC	NR	$3.56e^{08}$	$2.79e^{-05}$
SN	NR	$2.48e^{08}$	$7.02e^{-31}$

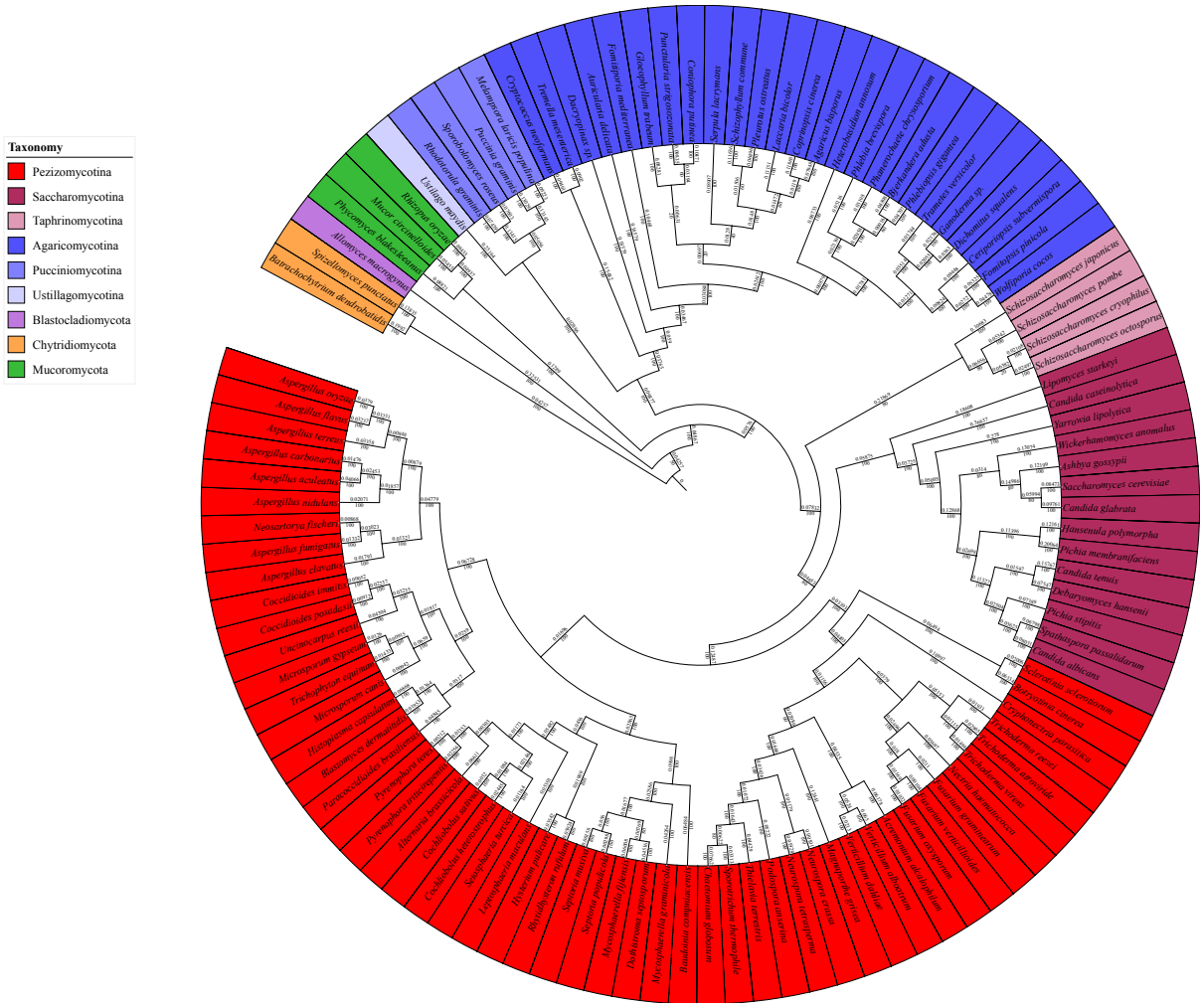

Figure 2.3.3: Representative fungal phylogeny

This fungal phylogeny constructed from a maximum likelihood superalignment 249 highly distributed KOG genes ($n_{species} \geq 102$). We were content with the placement of all higher taxonomic clades in this tree with the exception of one contentious species, *Ustiago maydis*. In this tree *U. maydis* represents the Ustillagomycotina subphylum (smut fungi), which should be sister to the Agaricomycotina and not the Pucciniomycotina as displayed here (Spatafora *et al.*, 2017). This could just be due to the complement of genes in this particular species, and perhaps a larger Ustilagomycotina sample size would better resolve their phylogenetic placement.

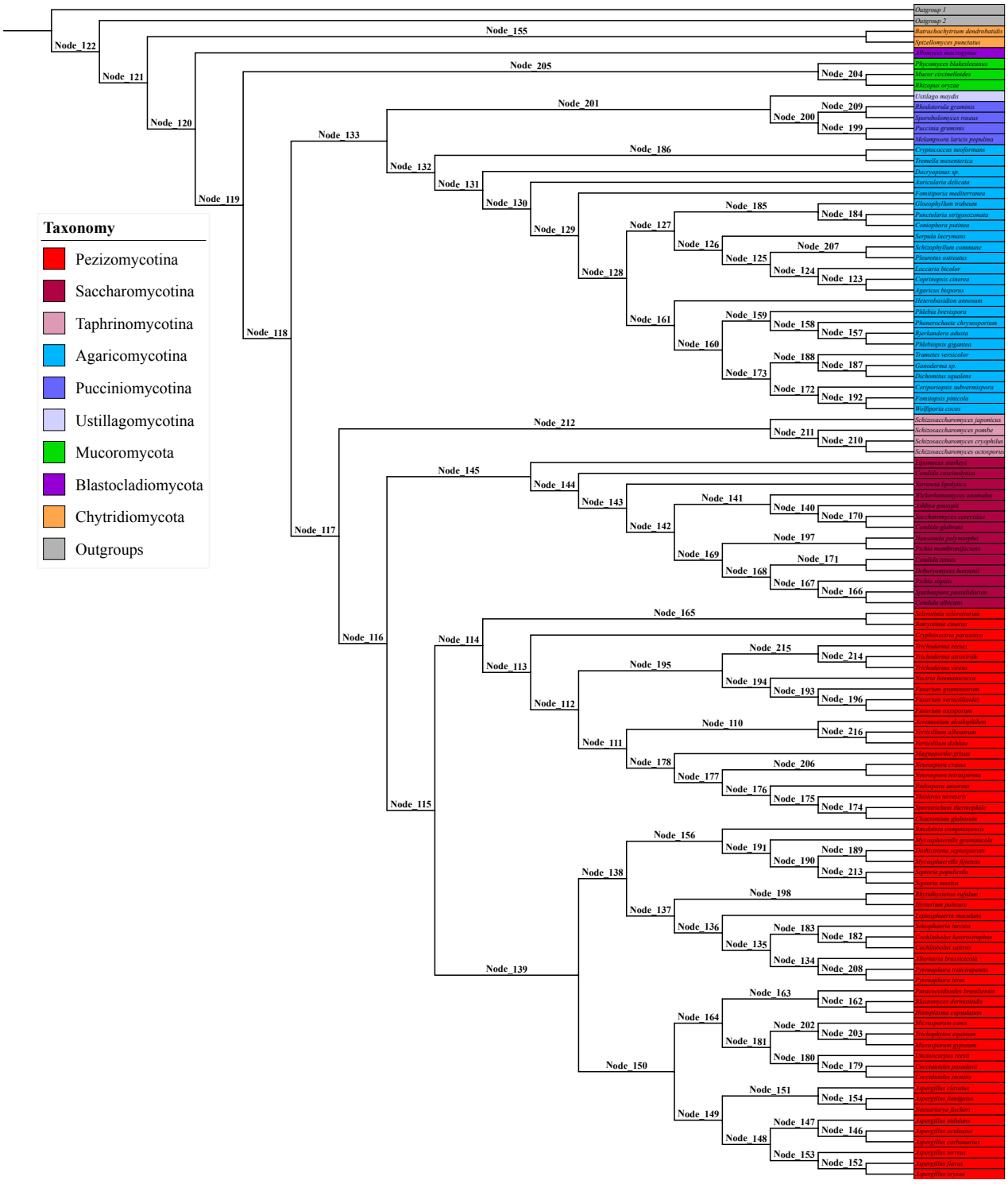

Figure 2.3.4: Representative fungal phylogeny annotated with internal node annotations

An illustration of the representative fungal phylogeny (Figure 2.3.3) with the two pseudo-outgroups prepended to the original root node and annotated using TNT (as per *subsection 2.2.3.4.2.*)

2.3.4.2. Remodelled genes are more homoplastic than non-remodelled genes

When sampled from across the entire phylogeny (all data), significant differences ($P \leq \alpha \leq 8.33e^{-03}$) were observed between all comparisons ($P \leq 7.10e^{-03}$) except for between NC and RC ($P = 0.497$) (Table 2.3.7.). Remodelled genes were considerably more homoplastic than non-remodelled genes ($HP_{NR} = 0.305$), where composite genes were observed to be most homoplastic ($HP_{NC} = 0.487$; $HP_{SC} = 0.464$; $HP_{SN} = 0.433$).

When sampled from exclusively from internal nodes, again, all comparisons were observed to be significantly different ($P \leq 3.40e^{-14}$) except for between NC and RC ($P = 0.671$). Again, remodelled genes were observed to be more homoplastic ($HP_{NC} = 0.120$; $HP_{SC} = 0.123$; $HP_{SN} = 0.073$) in comparison to non-remodelled families ($HP_{NR} = 0.058$).

These results suggest that remodelled genes, especially strict and nested composites, are much more likely to disobey Dollo's Law of Irreversibility (Dollo, 1893; Gould, 1970) and evolve through multiple, independent events, which would be expected of fused genes (Leonard and Richards, 2012). These results also highlight the extent of homoplasy in tips (leaf nodes) in comparison to internal nodes from the increases observed due to their inclusion.

2.3.6.3. Dynamic evolutionary rates observed within and between remodelling categories

When sampled from across the phylogeny, birth rates displayed considerable variation ($0.00846 \leq f_b \leq 416.4$; $221.1 \leq CV \leq 276.1$) within each RC (Table 2.3.8.). Significant differences ($P \leq \alpha \leq 8.33e^{-03}$) were observed between each RC ($P \leq 1.72e^{-05}$) except for between NC and SN ($P = 0.952$) (Table 2.3.9). Even greater variance rates were observed for decay rates within each RC ($0.00846 \leq f_b \leq 322$; $315.8 \leq CV \leq 377.8$), however significant

Table 2.3.8. Homoplastic proportion comparisons between fungal remodelling categories

Pairwise comparisons between each HP_{RC} ($RC_{(a)}$ and $RC_{(b)}$) sampled from across (*i*.) the entire phylogeny and (*ii*.) exclusively from internal nodes are displayed. The sum of homoplastic families (n_H) and all families (n_{all}) used to calculate the HPs are shown. Significant *P*-values ($P \leq \alpha_B \leq 8.33e^{-03}$) are emboldened.

			$RC_{(a)}$			$RC_{(b)}$			
Dataset	**$RC_{(a)}$**	**$RC_{(b)}$**	**n_H**	**n_{all}**	**HP**	**n_H**	**n_{all}**	**HP**	***P***
All data	NC	SC	8431	17317	0.487	978	2107	0.464	$4.97e^{-02}$
	NC	SN	8431	17317	0.487	9418	21727	0.433	$\mathbf{7.50e^{-26}}$
	NC	NR	8431	17317	0.487	12350	40489	0.305	**0**
	SC	SN	978	2107	0.464	9418	21727	0.433	$\mathbf{7.10e^{-03}}$
	SC	NR	978	2107	0.464	12350	40489	0.305	$\mathbf{4.31e^{-50}}$
	SN	NR	9418	21727	0.433	12350	40489	0.305	$\mathbf{2.73e^{-222}}$
Exclusively internal nodes	NC	SC	2080	17317	0.120	260	2107	0.123	0.671
	NC	SN	2080	17317	0.120	1593	21727	0.073	$\mathbf{2.42e^{-55}}$
	NC	NR	2080	17317	0.120	2334	40489	0.058	$\mathbf{1.21e^{-137}}$
	SC	SN	260	2107	0.123	1593	21727	0.073	$\mathbf{1.61e^{-14}}$
	SC	NR	260	2107	0.123	2334	40489	0.058	$\mathbf{4.24e^{-28}}$
	SN	NR	1593	21727	0.073	2334	40489	0.058	$\mathbf{3.40e^{-14}}$

Table 2.3.9: Descriptive statistics for observed evolutionary rates in fung

Descriptive statistics for evolutionary birth (f_b) and decay (f_d) rates were calculated. The considerable variance observed in across the entire phylogeny (CV ≤ 378.8%) is due to the considerable differences between leaf and branch node subsets (Table 2.3.9). In particular, considerable variance is observed in the subset of leaf nodes (CV ≤ 309.4%), which is due to using a small κ during transformation.

		RC	μ	σ	SE	CV(%)	Min.	Max.	$\eta_{0.25}$	$\eta_{0.50}$	$\eta_{0.75}$
Entire phylogeny (*n* = 212)	f_b	NC	14.59	40.27	2.766	276.1	0.00846	416.4	2.722	6.464	12.13
		SC	1.859	4.83	0.3317	259.8	0.00846	42.56	0.391	0.8433	1.617
		SN	14.7	38.28	2.629	260.5	0.00846	328.1	3.001	6.259	11.72
		NR	21.59	48.6	3.338	225.1	0.00846	405.2	4.93	10.18	19.44
	f_d	NC	6.992	25.48	1.75	364.4	0.004832	322	0.6173	0.6173	5.216
		SC	0.955	3.016	0.2071	315.8	0.004832	37.79	0.08478	0.3211	0.833
		SN	5.11	18.22	1.251	356.5	0.004832	226.8	0.4969	1.6	3.667
		NR	6.602	25.01	1.717	378.8	0.004832	314.5	0.5778	2.244	4.643
Exclusively internal nodes (*n* = 105)	f_b	NC	6.852	6.071	0.5925	88.61	0.02664	30.11	2.288	4.787	10.93
		SC	0.9429	0.8233	0.08035	87.32	0.02443	3.285	0.3006	0.6578	1.315
		SN	6.784	7.742	0.7556	114.1	0.02664	60.62	2.369	4.789	8.424
		NR	13.36	15.9	1.552	119.1	0.02664	95.57	3.855	7.982	15.52
	f_d	NC	2.242	3.245	0.3167	144.8	0.01015	23.5	0.518	0.518	2.891
		SC	0.394	0.6931	0.06764	175.9	0.01015	5.306	0.06323	0.1765	0.3956
		SN	1.62	2.092	0.2041	129.1	0.01015	13.14	0.3583	0.9843	2.064
		NR	1.938	2.47	0.2411	127.5	0.01015	14.91	0.44	0.9515	2.697
Exclusively leaf nodes (*n* = 107)	f_b	NC	22.18	55.45	5.36	250.0	0.00846	416.4	4.325	7.49	15.17
		SC	2.758	6.642	0.6421	240.8	0.00846	42.56	0.5263	0.9193	1.971
		SN	22.46	52.31	5.057	232.9	0.00846	328.1	4.371	8.919	15.49
		NR	29.66	65.73	6.354	221.6	0.00846	405.2	5.873	11.83	21.18
	f_d	NC	11.65	35.18	3.401	301.9	0.004832	322	1.159	1.159	7.901
		SC	1.506	4.125	0.3988	274.0	0.004832	37.79	0.168	0.5102	1.107
		SN	8.535	25.15	2.431	294.7	0.004832	226.8	0.941	2.749	6.096
		NR	11.18	34.58	3.343	309.4	0.004832	314.5	1.095	3.66	8.107

Table 2.3.10. Comparison of family birth and decay rates in fungi

Pairwise comparisons between each HP_{RC} ($RC_{(a)}$ and $RC_{(b)}$) sampled from across (*i*.) the entire phylogeny and (*ii*.) exclusively from internal nodes are displayed. The U statistic and *P*-values are provided for each comparison. Statistically significant comparisons ($P \leq \alpha \leq 8.33e^{-03}$) are emboldened.

		$RC_{(a)}$	$RC_{(b)}$	U	*P*
All data	f_b	NC	SC	39883	**$2.56e^{-43}$**
		NC	SN	22395.5	0.952
		NC	NR	17060	**$1.79e^{-05}$**
		SC	SN	4977	**$1.01e^{-43}$**
		SC	NR	3663	**$2.93e^{-50}$**
		SN	NR	17048	**$1.72e^{-05}$**
	f_d	NC	SC	35982	**$9.36e^{-27}$**
		NC	SN	24967	0.048
		NC	NR	22974	0.691
		SC	SN	10376	**$9.06e^{-22}$**
		SC	NR	9469	**$6.62e^{-25}$**
		SN	NR	20515	0.121
Exclusively internal nodes (branches)	f_b	NC	SC	9983.5	**$3.20e^{-24}$**
		NC	SN	5698.5	0.674
		NC	NR	4061.5	**$9.86e^{-04}$**
		SC	SN	1098.5	**$1.19e^{-23}$**
		SC	NR	780.5	**$6.17e^{-27}$**
		SN	NR	3870.5	**$1.93e^{-04}$**
	f_d	NC	SC	8990	**$2.85e^{-15}$**
		NC	SN	6220.5	0.108
		NC	NR	5823	0.481
		SC	SN	2399	**$1.55e^{-12}$**
		SC	NR	2254	**$1.37e^{-13}$**
		SN	NR	5113.5	0.365
Exclusively leaf nodes	f_b	NC	SC	10031.5	**$1.93e^{-21}$**
		NC	SN	5473	0.579
		NC	NR	4444.5	**$4.73e^{-03}$**
		SC	SN	1302.5	**$1.63e^{-22}$**
		SC	NR	994.5	**$1.59e^{-25}$**
		SN	NR	4592.5	0.0125
	f_d	NC	SC	9305	**$2.69e^{-15}$**
		NC	SN	6434.5	0.117
		NC	NR	5758	0.942
		SC	SN	2447	**$4.64e^{-13}$**
		SC	NR	2113	**$1.55e^{-15}$**
		SN	NR	5035.5	0.128

rate differences were only observed between SC and each other RC ($P \leq 9.06e^{-22}$) and not between any other comparison ($P > 0.048$).

Comparatively, when sampled from the subset of exclusively internal nodes, considerably less variance was observed for both birth rates ($0.00846 \leq f_b \leq 322$; $87.37 \leq CV \leq 119.1$) and decay rates ($0.02443 \leq f_b \leq 95.57$; $127.5 \leq CV \leq 175.9$) respectively for each RC. Again, only significant differences were observed between SC and each other RC ($P \leq 1.55e^{-12}$) and not between any other comparison ($P > 0.108$).

Finally, when sampled from the subset of exclusively leaf nodes, considerable variation in birth rates ($0.00846 \leq f_b \leq 416.4$; $221.6 \leq CV \leq 250$) and decay rates ($0.004832 \leq f_b \leq 322$; $274 \leq CV \leq 309.4$) respectively for each RC. Significant differences in birth rates were observed between each SC comparison ($P \leq 1.93e^{-21}$) and between NC and SN ($P = 4.73e^{-03}$). Significant decay rate differences ($P \leq 4.64e^{-13}$) were only observed between SC comparisons.

Significant differences ($P \leq \alpha \leq 0.0125$) were observed between internal exclusive nodes and leaf exclusive nodes for each specific RC (*eg.* NC *vs* NC) for both birth rates ($P \leq 9.27e^{-03}$) and decay rates ($P \leq 1.51e^{-06}$) (Table 2.3.10).

These results suggest that despite the considerable variance observed within RCs, with the exception of SCs, RCs evolve at a relatively similar rate. The significant evolutionary rate increases observed between internal nodes and leaf nodes in each RC is likely due to genomic innovation during speciation (Gogarten and Townsend, 2005). Synapomorphic families observed at internal nodes were evolutionarily "successful", meaning they were retained post speciation events. Such differences were expected as genome sequencing provides a "snapshot in time" (Klimke *et al*., 2011) of a given genome, with no guarantee that an innovation will persist. The evolutionary rate increase is likely influenced by homoplasy and epaktologous events which would be consistant with the HP differences observed in Table 2.3.7.

Table 2.3.11. Comparison of evolutionary rates between internal and leaf nodes

Each test samples gene families of the same RC and compares those sampled from leaves to those sampled from internal nodes. U is the Mann-Whitney U statistic. Every result was considered statistically significant ($P \leq \alpha_B \leq 0.0125$).

	RC	*U*	*P*
f_b	NC	4170	$1.19e^{-03}$
	SC	4455	$9.27e^{-03}$
	SN	3703	$1.82e^{-05}$
	NR	4438	$8.29e^{-03}$
f_d	NC	3268	$1.44e^{-07}$
	SC	3469	$1.51e^{-06}$
	SN	3158	$3.66e^{-08}$
	NR	2959	$2.65e^{-09}$

2.3.6.3. Evolution via *gene remodelling is clocklike in Fungi*

When sampling from across the phylogeny, significant evolutionary bursts ($P \leq \alpha_B \leq 2.35e^{-04}$) were only observed during two speciation events (Table 2.3.11). Significant bursts of nested composite births ($f_b = 416.44$/Ma; $P = 5.7e^{-05}$), strict component births ($f_b = 307.012$/Ma; $P = 2.2e^{-04}$), and non-remodelled family births ($f_b = 404.28$/Ma; $P = 2.2e^{-04}$) were observed during the divergence of *Coccidioides immitis* from its MRCA with *Coccidoides posadasii*. Significant bursts of strict component births ($f_b = 328.06$/Ma; $P = 1.7e^{-04}$), non-remodelled family births ($f_b = 405.16$/Ma; $P = 2.0e^{-04}$), nested composite decay ($f_d = 322.01$/Ma; $P = 9.5e^{-05}$), strict component decay ($f_d = 226.77$/Ma; $P = 1.0e^{-04}$), and non-remodelled family decay ($f_d = 314.45$/Ma; $P = 8.44e^{-05}$) were observed during the divergence of *Fusarium verticillioides* from its MRCA with *Fusarium oxysporum*. These results are influenced by short branch lengths associated with *C. immitis* ($\kappa = 0.0005$; Table 2.3.11) and *F. verticillioides* ($\kappa = 0.001$).

When sampling from the subset of internal branches, a significant burst ($P \leq 4.76e^{-04}$) was only observed for strict component births ($f_b = 60.62$/Ma; $P = 3.0e^{-04}$) in "Node_196", branch within genus *Fusarium*, representing the divergence of the MRCA of *F. verticillioides* and *F. oxysporum* (*Fov*-MCRA) from *F. graminearum* (Table 2.3.12). In contrast to the branches representing *C. immitis* and *F. verticilliodes*, the branch length for "Node_196" ($\kappa = 0.01169$) was within the average branch lengths for internal branches (0.04169±0.03859), however, the sum of strict component families assigned to this branch ($n = 450$) was considerably greater ($n > \mu + (3\times\sigma)$) than the mean assignment of strict component families per internal branch (109.4±89.987).

The burst observed at "Node_196" corresponds to genomic expansions during the divergence of *Fov*-MCRA from *F. graminearum* (Ma *et al.*, 2010). These expansions are

Table 2.3.12. Evolutionary bursts across the fungal phylogeny

Each site (branch or leaf node) is annotated with its associated branch length (κ), character state changes (T_b and T_d), birth and decay rates (f_b and f_d) and their associated *P*-values derived from a *Q*-function ($P(f_b)$ and $P(f_d)$). λ values used for Box-Cox transformation are given beneath Q-function RC identifiers. Significant evolutionary bursts ($P \leq \alpha_B \leq 2.36e^{-04}$) are emboldened and underlined in the $P(f_b)$ and $P(f_d)$ columns. Significant bursts were observed at two sites, the leaf nodes representing speciation events for *Coccoides immitis* and *Fusarium verticilliodes*, respectively. Significant *P*-values are emboldened.

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Acremonium alcalophilum	0.061776	238	29	192	321	161	23	85	108	6.081	0.763	4.911	8.193	4.122	0.611	2.188	2.773	0.502	0.501	0.568	0.559	0.323	0.3	0.401	0.398
Agaricus bisporus	0.079481	326	33	336	527	157	25	135	180	6.467	0.672	6.664	10.442	3.125	0.514	2.69	3.579	0.484	0.541	0.478	0.487	0.387	0.341	0.352	0.339
Allomyces macrogynus	0.325309	487	69	855	1934	0	0	0	0	2.358	0.338	4.136	9.349	0.005	0.005	0.005	0.005	0.755	0.737	0.617	0.52	0.997	0.997	0.997	0.997
Alternaria brassicicola	0.06611	243	31	296	303	448	57	274	414	5.801	0.761	7.061	7.228	10.675	1.379	6.538	9.867	0.516	0.502	0.461	0.594	0.14	0.147	0.169	0.142
Ashbya gossypii	0.121993	34	4	50	57	35	8	32	67	0.451	0.064	0.657	0.747	0.464	0.116	0.425	0.876	0.96	0.97	0.937	0.951	0.787	0.716	0.76	0.66
Aspergillus aculeatus	0.040663	294	50	235	401	159	22	126	146	11.403	1.971	9.123	15.539	6.185	0.889	4.909	5.682	0.315	0.226	0.384	0.368	0.237	0.222	0.221	0.24
Aspergillus carbonarius	0.014757	307	35	289	414	236	32	158	161	32.807	3.835	30.89	44.204	25.244	3.515	16.936	17.256	0.086	0.098	0.097	0.111	0.046	0.05	0.052	0.073

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Aspergillus clavatus	0.017915	167	26	156	288	74	11	32	46	14.74	2.369	13.775	25.356	6.58	1.053	2.895	4.124	0.245	0.184	0.267	0.231	0.224	0.191	0.335	0.307
Aspergillus flavus	0.032367	275	34	203	357	140	9	92	98	13.403	1.7	9.907	17.385	6.847	0.486	4.516	4.808	0.271	0.264	0.359	0.335	0.217	0.354	0.238	0.274
Aspergillus fumigatus	0.013323	197	24	175	221	142	22	158	166	23.359	2.949	20.763	26.19	16.87	2.713	18.758	19.701	0.142	0.141	0.169	0.222	0.082	0.069	0.045	0.06
Aspergillus nidulans	0.020712	274	30	253	371	91	11	71	86	20.87	2.353	19.276	28.231	6.982	0.911	5.464	6.602	0.164	0.186	0.185	0.204	0.213	0.217	0.201	0.21
Aspergillus oryzae	0.037896	240	25	411	279	248	29	190	253	9.996	1.078	17.088	11.613	10.328	1.244	7.922	10.535	0.353	0.393	0.212	0.455	0.146	0.163	0.138	0.132
Aspergillus terreus	0.033578	353	56	330	416	186	37	157	199	16.571	2.668	15.494	19.52	8.754	1.779	7.396	9.362	0.216	0.16	0.236	0.301	0.172	0.113	0.149	0.15
Auricularia delicata	0.165791	789	109	983	1569	0	0	0	0	7.49	1.043	9.329	14.885	0.009	0.009	0.009	0.009	0.439	0.404	0.377	0.381	0.995	0.989	0.994	0.994
Batrachochytrium dendrobatidis	0.199203	271	30	359	460	0	0	0	0	2.146	0.245	2.841	3.638	0.008	0.008	0.008	0.008	0.775	0.811	0.714	0.761	0.996	0.992	0.995	0.995
Baudoinia compniacensis	0.064841	277	27	316	490	112	13	68	94	6.739	0.679	7.684	11.902	2.739	0.339	1.673	2.303	0.471	0.538	0.435	0.448	0.418	0.445	0.466	0.442
Bjerkandera adusta	0.048984	362	51	437	780	70	12	62	107	11.648	1.669	14.055	25.061	2.278	0.417	2.022	3.466	0.309	0.268	0.262	0.233	0.462	0.392	0.42	0.346
Blastomyces dermatitidis	0.01953	237	30	157	262	51	7	38	46	19.155	2.495	12.716	21.167	4.185	0.644	3.139	3.783	0.183	0.173	0.288	0.279	0.32	0.289	0.316	0.326
Botryotinis cinerea	0.063341	355	43	412	463	185	31	118	120	8.834	1.092	10.249	11.514	4.616	0.794	2.953	3.003	0.389	0.39	0.349	0.458	0.298	0.244	0.33	0.38
Candida albicans	0.080313	79	9	93	163	49	3	55	80	1.566	0.196	1.84	3.21	0.979	0.078	1.096	1.585	0.834	0.854	0.806	0.786	0.652	0.798	0.566	0.53
Candida caseinolytica	0.366372	133	17	172	194	71	6	61	82	0.575	0.077	0.742	0.837	0.309	0.03	0.266	0.356	0.946	0.96	0.927	0.945	0.843	0.931	0.831	0.814
Candida glabrata	0.097608	30	6	35	77	40	4	26	46	0.499	0.113	0.58	1.256	0.66	0.081	0.435	0.757	0.955	0.929	0.946	0.914	0.728	0.792	0.756	0.689
Candida tenuis	0.157671	97	5	112	192	74	11	56	63	0.977	0.06	1.126	1.924	0.748	0.12	0.568	0.638	0.899	0.974	0.882	0.867	0.705	0.709	0.707	0.721
Ceriporiopsis subvermispora	0.094478	264	25	333	479	149	19	137	220	4.409	0.433	5.557	7.986	2.496	0.333	2.296	3.677	0.596	0.672	0.532	0.566	0.44	0.45	0.389	0.333

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Chaetomium globosum	0.076624	360	21	249	347	233	43	224	312	7.405	0.451	5.128	7.139	4.8	0.903	4.615	6.421	0.443	0.66	0.556	0.598	0.289	0.219	0.233	0.216
Coccidioides immitis	0.000517	136	13	100	132	45	3	31	48	416.441	42.556	307.012	404.283	139.827	12.159	97.271	148.946	**$5.7e^{-05}$**	$4.30e^{-04}$	**$2.23e^{-04}$**	**$2.18e^{-04}$**	0.001	0.008	0.001	0.001
Coccidioides posadasii	0.009125	102	18	161	143	172	23	95	117	17.743	3.273	27.906	24.805	29.801	4.134	16.537	20.326	0.2	0.122	0.113	0.236	0.036	0.04	0.054	0.058
Cochliobolus heterostrophus	0.024454	256	27	233	247	281	35	163	362	16.519	1.8	15.041	15.941	18.126	2.314	10.541	23.333	0.217	0.249	0.244	0.36	0.074	0.084	0.099	0.047
Cochliobolus sativus	0.003197	242	32	240	340	56	8	46	44	119.461	16.223	118.477	167.638	28.022	4.424	23.106	22.122	0.005	0.006	0.006	0.006	0.039	0.037	0.032	0.051
Coniophora putinea	0.104709	405	47	456	701	150	17	153	208	6.095	0.721	6.86	10.538	2.267	0.27	2.312	3.137	0.501	0.519	0.469	0.485	0.463	0.504	0.387	0.369
Coprinopsis cinerea	0.118645	374	47	336	631	186	25	139	188	4.968	0.636	4.465	8.373	2.477	0.344	1.855	2.504	0.562	0.558	0.596	0.552	0.442	0.441	0.441	0.422
Cryphonectria parasitica	0.109973	499	63	476	713	59	12	45	59	7.146	0.915	6.818	10.205	0.858	0.186	0.657	0.858	0.453	0.444	0.471	0.494	0.679	0.602	0.678	0.664
Cryptococcus neoformans	0.094612	227	31	189	337	18	1	17	17	3.788	0.532	3.157	5.615	0.316	0.033	0.299	0.299	0.638	0.612	0.688	0.662	0.84	0.921	0.815	0.837
Dacryopinax sp.	0.18579	0	0	0	0	665	56	372	423	0.008	0.008	0.008	0.008	5.634	0.482	3.156	3.587	>0.999	>0.999	>0.999	>0.999	0.256	0.356	0.315	0.338
Debaryomyces hansenii	0.075474	120	18	158	281	24	3	35	39	2.52	0.396	3.311	5.873	0.521	0.083	0.75	0.833	0.74	0.697	0.676	0.65	0.769	0.786	0.651	0.67
Dichomitus squalens	0.036296	223	26	313	383	98	24	111	171	9.7	1.169	13.598	16.629	4.287	1.083	4.85	7.449	0.362	0.369	0.27	0.348	0.314	0.186	0.224	0.188
Dothistroma septosporum	0.045165	257	35	301	510	111	13	78	130	8.979	1.253	10.51	17.784	3.898	0.487	2.749	4.559	0.384	0.349	0.342	0.328	0.336	0.353	0.347	0.285
Fomitiporia mediterranea	0.104677	338	50	378	615	30	4	29	48	5.09	0.766	5.691	9.25	0.465	0.075	0.45	0.736	0.555	0.5	0.525	0.523	0.786	0.806	0.75	0.694
Fomitopsis pinicola	0.053252	433	52	568	775	96	7	97	123	12.81	1.564	16.795	22.905	2.863	0.236	2.893	3.66	0.283	0.286	0.217	0.257	0.407	0.539	0.335	0.334
Fusarium graminearum	0.021103	178	12	152	248	94	18	63	77	13.333	0.968	11.396	18.547	7.076	1.415	4.767	5.81	0.272	0.427	0.319	0.316	0.211	0.143	0.227	0.235
Fusarium oxysporum	0.010319	330	22	411	568	136	9	83	136	50.419	3.503	62.757	86.672	20.868	1.523	12.795	20.868	0.04	0.111	0.027	0.032	0.061	0.133	0.077	0.056

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Fusarium verticillioides	0.00104	180	17	216	267	212	24	149	207	273.634	27.212	328.058	405.16	322.011	37.795	226.768	314.452	$3.29e^{-04}$	$2.00e^{-03}$	**$1.71e^{-04}$**	**$2.15e^{-04}$**	**$9.50e^{-05}$**	$8.94e^{-04}$	**$1.12e^{-04}$**	**$8.44e^{-05}$**
Ganoderma sp.	0.021757	209	32	288	393	98	9	89	121	15.171	2.384	20.878	28.464	7.152	0.722	6.502	8.814	0.238	0.183	0.167	0.202	0.208	0.264	0.17	0.16
Gloeophyllum trabeum	0.082815	332	52	512	709	63	10	51	67	6.32	1.006	9.737	13.476	1.215	0.209	0.987	1.291	0.491	0.415	0.364	0.41	0.606	0.572	0.59	0.576
Hansenula polymorpha	0.121609	159	13	145	167	103	12	72	123	2.068	0.181	1.887	2.171	1.344	0.168	0.944	1.603	0.782	0.867	0.802	0.851	0.584	0.627	0.6	0.527
Heterobasidion annosum	0.087328	354	46	542	671	196	20	123	175	6.39	0.846	9.773	12.095	3.546	0.378	2.232	3.168	0.487	0.469	0.363	0.443	0.357	0.417	0.396	0.367
Histoplasma capsulatum	0.068685	188	22	206	237	294	39	184	226	4.325	0.526	4.737	5.446	6.751	0.915	4.234	5.195	0.602	0.615	0.579	0.669	0.219	0.216	0.251	0.258
Hysterium pulicare	0.035422	404	52	306	514	164	21	122	157	17.971	2.352	13.623	22.852	7.322	0.976	5.458	7.011	0.197	0.186	0.27	0.258	0.204	0.204	0.201	0.199
Laccaria bicolor	0.111506	434	60	708	969	169	13	104	141	6.132	0.86	9.994	13.673	2.396	0.197	1.48	2.002	0.5	0.464	0.357	0.406	0.45	0.586	0.495	0.475
Leptosphaeria maculans	0.053006	258	30	251	398	196	18	118	179	7.68	0.919	7.473	11.832	5.842	0.563	3.529	5.338	0.432	0.443	0.443	0.45	0.248	0.319	0.29	0.252
Lipomyces starkeyi	0.186085	368	58	388	472	45	5	29	31	3.117	0.498	3.286	3.995	0.389	0.051	0.253	0.27	0.689	0.631	0.678	0.742	0.813	0.87	0.837	0.849
Magnaporthe grisea	0.124413	335	39	345	485	308	29	192	225	4.245	0.505	4.371	6.14	3.904	0.379	2.438	2.855	0.607	0.627	0.602	0.638	0.335	0.416	0.375	0.391
Melampsora laricis-populina	0.092229	488	70	603	967	67	4	35	44	8.334	1.21	10.294	16.497	1.159	0.085	0.614	0.767	0.407	0.359	0.348	0.35	0.617	0.782	0.692	0.686
Microsporum canis	0.009418	147	16	114	180	52	5	36	49	24.701	2.837	19.193	30.209	8.846	1.001	6.175	8.345	0.131	0.148	0.186	0.188	0.171	0.2	0.179	0.169
Microsporus gypseum	0.012602	145	15	113	165	74	11	62	64	18.21	1.996	14.219	20.705	9.355	1.497	7.858	8.107	0.194	0.223	0.258	0.285	0.161	0.135	0.139	0.173
Mucor circinelloides	0.054553	160	27	190	287	40	4	17	37	4.639	0.807	5.503	8.298	1.181	0.144	0.519	1.095	0.582	0.484	0.535	0.555	0.612	0.665	0.724	0.613
Mycosphaerella fijiensis	0.060081	303	39	354	434	301	40	232	340	7.953	1.046	9.287	11.38	7.901	1.073	6.096	8.921	0.421	0.403	0.378	0.462	0.19	0.187	0.181	0.158
Mycosphaerella graminicola	0.04264	270	37	315	471	165	20	100	148	9.99	1.401	11.649	17.399	6.119	0.774	3.723	5.493	0.353	0.316	0.313	0.334	0.239	0.249	0.278	0.246

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Nectria haematococca	0.03697	477	57	392	603	70	11	72	88	20.322	2.466	16.709	25.679	3.019	0.51	3.104	3.784	0.17	0.175	0.218	0.227	0.395	0.342	0.319	0.326
Neosartorya fischeri	0.008676	242	25	227	356	61	10	51	59	44.026	4.711	41.308	64.68	11.233	1.993	9.421	10.871	0.052	0.072	0.06	0.058	0.133	0.099	0.114	0.128
Neurospora crassa	0.00101	107	21	110	148	41	6	24	30	168.109	34.245	172.779	231.929	65.376	10.896	38.914	48.254	0.002	0.001	0.002	0.002	0.008	0.009	0.012	0.013
Neurospora tetrasperma	0.019285	90	11	93	118	108	23	84	106	7.417	0.978	7.662	9.699	8.884	1.956	6.928	8.721	0.442	0.424	0.436	0.509	0.17	0.101	0.159	0.161
Node112	0.045345	82	10	51	74	14	1	6	13	2.877	0.381	1.802	2.6	0.52	0.069	0.243	0.485	0.709	0.706	0.81	0.823	0.769	0.82	0.843	0.768
Node113	0.020864	173	31	124	167	10	1	6	6	13.108	2.411	9.417	12.656	0.829	0.151	0.527	0.527	0.276	0.18	0.374	0.429	0.685	0.654	0.721	0.754
Node114	0.011563	150	23	101	181	3	0	6	6	20.525	3.262	13.865	24.739	0.544	0.136	0.952	0.952	0.168	0.123	0.265	0.237	0.762	0.679	0.598	0.643
Node115	0.044031	417	51	250	273	48	9	41	79	14.922	1.856	8.96	9.781	1.749	0.357	1.499	2.856	0.242	0.241	0.389	0.507	0.524	0.431	0.492	0.391
Node116	0.033931	74	7	52	56	30	2	24	40	3.474	0.371	2.455	2.64	1.436	0.139	1.158	1.899	0.661	0.714	0.748	0.821	0.569	0.674	0.553	0.487
Node117	0.12487	64	7	44	58	40	2	29	39	0.818	0.101	0.566	0.743	0.516	0.038	0.378	0.504	0.918	0.94	0.947	0.952	0.77	0.908	0.779	0.762
Node118	0.04653	21	0	9	9	29	8	24	43	0.743	0.034	0.338	0.338	1.013	0.304	0.845	1.486	0.926	0.991	0.973	0.981	0.645	0.473	0.625	0.544
Node119	0.078317	66	2	45	31	0	0	0	0	1.345	0.06	0.923	0.642	0.02	0.02	0.02	0.02	0.857	0.973	0.906	0.959	0.988	0.961	0.986	0.987
Node120	0.057596	35	2	23	23	0	0	0	0	0.982	0.082	0.655	0.655	0.027	0.027	0.027	0.027	0.898	0.956	0.937	0.958	0.983	0.939	0.981	0.982
Node121	0.048674	400	43	246	321	0	0	0	0	12.949	1.421	7.976	10.398	0.032	0.032	0.032	0.032	0.28	0.312	0.424	0.489	0.98	0.924	0.977	0.978
Node122	0.042568	0	0	0	0	0	0	0	0	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.999	0.989	0.999	0.999	0.977	0.91	0.973	0.975
Node123	0.011903	92	18	70	117	61	12	52	46	12.281	2.509	9.376	15.582	8.187	1.717	6.999	6.206	0.294	0.172	0.376	0.367	0.184	0.117	0.157	0.222
Node124	0.014721	19	5	29	45	16	1	10	13	2.135	0.641	3.203	4.912	1.815	0.214	1.174	1.495	0.776	0.556	0.685	0.695	0.515	0.566	0.55	0.543

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Node125	0.014804	30	2	28	40	22	2	14	24	3.291	0.319	3.079	4.353	2.442	0.319	1.593	2.654	0.675	0.752	0.695	0.723	0.445	0.461	0.477	0.408
Node126	0.0128	35	3	38	54	21	4	21	27	4.421	0.491	4.789	6.754	2.701	0.614	2.701	3.438	0.596	0.636	0.576	0.613	0.421	0.299	0.351	0.348
Node127	0.006965	29	1	22	34	9	0	5	8	6.77	0.451	5.19	7.898	2.257	0.226	1.354	2.031	0.47	0.66	0.552	0.569	0.464	0.551	0.516	0.472
Node128	0.026124	42	4	41	58	5	0	2	1	2.587	0.301	2.527	3.55	0.361	0.06	0.181	0.12	0.734	0.766	0.741	0.766	0.823	0.844	0.877	0.924
Node129	0.030842	149	27	117	177	0	0	0	0	7.644	1.427	6.014	9.071	0.051	0.051	0.051	0.051	0.433	0.311	0.509	0.529	0.968	0.869	0.963	0.966
Node130	0.034674	406	46	258	340	0	0	0	0	18.449	2.131	11.741	15.458	0.045	0.045	0.045	0.045	0.191	0.208	0.311	0.369	0.972	0.885	0.967	0.97
Node131	0.058997	0	0	0	0	20	1	14	18	0.027	0.027	0.027	0.027	0.559	0.053	0.4	0.506	0.999	0.994	0.999	0.999	0.757	0.863	0.77	0.761
Node132	0.037652	0	0	0	0	4	0	4	6	0.042	0.042	0.042	0.042	0.209	0.042	0.209	0.292	0.999	0.986	0.999	0.998	0.885	0.896	0.861	0.84
Node133	0.098772	79	8	44	55	29	4	20	30	1.273	0.143	0.716	0.891	0.477	0.08	0.334	0.493	0.865	0.901	0.93	0.941	0.782	0.795	0.798	0.765
Node134	0.003028	20	3	25	81	9	2	11	18	10.902	2.077	13.497	42.569	5.191	1.557	6.23	9.864	0.328	0.214	0.272	0.118	0.273	0.13	0.177	0.142
Node135	0.011734	74	7	78	200	22	1	14	10	10.047	1.072	10.582	26.925	3.081	0.268	2.009	1.474	0.351	0.396	0.34	0.215	0.39	0.506	0.421	0.546
Node136	0.054824	110	14	127	283	29	0	14	17	3.182	0.43	3.67	8.142	0.86	0.029	0.43	0.516	0.684	0.673	0.65	0.561	0.678	0.935	0.758	0.758
Node137	0.045604	172	19	111	175	15	2	17	15	5.963	0.689	3.86	6.066	0.551	0.103	0.62	0.551	0.508	0.533	0.636	0.642	0.759	0.741	0.69	0.746
Node138	0.028646	133	23	96	144	2	0	6	2	7.353	1.317	5.322	7.956	0.165	0.055	0.384	0.165	0.445	0.334	0.545	0.567	0.906	0.858	0.777	0.9
Node139	0.016957	102	15	64	74	5	0	5	5	9.547	1.483	6.025	6.952	0.556	0.093	0.556	0.556	0.366	0.3	0.508	0.605	0.758	0.764	0.711	0.745
Node140	0.149864	97	7	149	279	66	8	49	54	1.028	0.084	1.573	2.937	0.703	0.094	0.524	0.577	0.893	0.954	0.834	0.802	0.717	0.761	0.722	0.739
Node141	0.031404	28	5	52	59	14	0	7	14	1.452	0.3	2.653	3.003	0.751	0.05	0.4	0.751	0.846	0.766	0.73	0.798	0.704	0.872	0.77	0.69

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Node142	0.128676	75	1	42	90	57	5	29	41	0.928	0.024	0.525	1.112	0.708	0.073	0.366	0.513	0.905	0.995	0.952	0.925	0.715	0.81	0.784	0.759
Node143	0.050053	77	15	45	96	34	9	22	17	2.449	0.502	1.445	3.046	1.099	0.314	0.722	0.565	0.746	0.629	0.847	0.796	0.628	0.465	0.658	0.742
Node144	0.037245	47	6	46	50	26	4	16	16	2.026	0.295	1.983	2.152	1.139	0.211	0.717	0.717	0.786	0.77	0.792	0.852	0.62	0.569	0.66	0.699
Node145	0.058749	59	14	56	73	7	1	3	9	1.605	0.401	1.525	1.98	0.214	0.054	0.107	0.268	0.829	0.693	0.839	0.864	0.883	0.862	0.923	0.851
Node146	0.024533	190	25	121	261	94	3	60	45	12.237	1.666	7.816	16.786	6.087	0.256	3.908	2.947	0.295	0.269	0.43	0.345	0.24	0.518	0.268	0.384
Node147	0.018571	39	5	27	43	8	0	5	2	3.385	0.508	2.37	3.724	0.762	0.085	0.508	0.254	0.668	0.626	0.755	0.757	0.702	0.783	0.728	0.857
Node148	0.00879	62	6	33	46	6	0	5	4	11.265	1.252	6.08	8.404	1.252	0.179	1.073	0.894	0.318	0.349	0.506	0.551	0.6	0.611	0.571	0.655
Node149	0.047793	223	19	160	266	10	1	11	11	7.367	0.658	5.295	8.781	0.362	0.066	0.395	0.395	0.444	0.548	0.546	0.539	0.822	0.829	0.772	0.8
Node150	0.067278	247	35	193	255	7	2	5	10	5.794	0.841	4.532	5.981	0.187	0.07	0.14	0.257	0.517	0.471	0.592	0.645	0.895	0.818	0.901	0.855
Node151	0.013209	98	10	91	213	11	2	11	11	11.78	1.309	10.947	25.464	1.428	0.357	1.428	1.428	0.306	0.336	0.33	0.229	0.57	0.431	0.504	0.554
Node152	0.032306	234	26	294	604	109	18	63	95	11.434	1.314	14.353	29.436	5.352	0.924	3.114	4.671	0.314	0.335	0.256	0.194	0.266	0.214	0.318	0.28
Node153	0.008481	61	4	45	75	27	4	29	34	11.491	0.927	8.525	14.085	5.189	0.927	5.56	6.487	0.313	0.44	0.404	0.397	0.273	0.214	0.198	0.214
Node154	0.030228	150	22	151	296	31	5	16	34	7.852	1.196	7.904	15.443	1.664	0.312	0.884	1.82	0.425	0.363	0.427	0.37	0.535	0.466	0.615	0.498
Node155	0.042568	174	24	148	314	0	0	0	0	6.462	0.923	5.502	11.631	0.037	0.037	0.037	0.037	0.484	0.442	0.535	0.455	0.977	0.91	0.973	0.975
Node156	0.099798	149	29	136	193	38	4	21	21	2.362	0.472	2.158	3.055	0.614	0.079	0.346	0.346	0.754	0.647	0.775	0.795	0.741	0.797	0.793	0.818
Node157	0.009157	109	13	113	196	23	5	19	31	18.882	2.403	19.568	33.815	4.12	1.03	3.433	5.493	0.186	0.181	0.181	0.163	0.323	0.195	0.296	0.246
Node158	0.029582	39	11	48	74	25	2	18	23	2.125	0.638	2.604	3.985	1.381	0.159	1.01	1.275	0.777	0.557	0.735	0.742	0.578	0.64	0.585	0.579

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Node159	0.021386	64	11	60	137	17	3	15	26	4.777	0.882	4.483	10.142	1.323	0.294	1.176	1.984	0.573	0.456	0.595	0.496	0.587	0.482	0.55	0.477
Node160	0.017813	72	9	69	115	21	1	23	25	6.442	0.882	6.177	10.236	1.941	0.176	2.118	2.294	0.485	0.456	0.501	0.493	0.499	0.615	0.408	0.443
Node161	0.008789	34	6	27	56	13	1	9	17	6.26	1.252	5.008	10.194	2.504	0.358	1.788	3.219	0.493	0.349	0.563	0.495	0.439	0.431	0.449	0.363
Node162	0.013638	94	12	100	217	30	4	28	24	10.949	1.498	11.64	25.124	3.573	0.576	3.342	2.881	0.326	0.297	0.313	0.233	0.356	0.314	0.302	0.389
Node163	0.031702	73	16	80	160	32	3	19	12	3.669	0.843	4.016	7.982	1.636	0.198	0.992	0.645	0.647	0.47	0.625	0.566	0.539	0.585	0.589	0.719
Node164	0.023896	70	7	45	86	69	5	20	21	4.67	0.526	3.026	5.723	4.604	0.395	1.381	1.447	0.58	0.616	0.699	0.657	0.298	0.406	0.512	0.551
Node165	0.064938	307	27	335	904	23	2	24	26	7.455	0.678	8.133	21.905	0.581	0.073	0.605	0.654	0.441	0.538	0.418	0.269	0.75	0.812	0.694	0.716
Node166	0.036247	46	5	68	158	18	0	9	16	2.038	0.26	2.992	6.895	0.824	0.043	0.434	0.737	0.785	0.798	0.702	0.607	0.687	0.891	0.756	0.694
Node167	0.025037	14	5	22	50	1	0	5	2	0.942	0.377	1.444	3.202	0.126	0.063	0.377	0.188	0.903	0.71	0.848	0.786	0.926	0.837	0.78	0.888
Node168	0.113266	105	3	140	324	16	3	4	17	1.471	0.056	1.957	4.51	0.236	0.056	0.069	0.25	0.844	0.977	0.795	0.715	0.873	0.857	0.949	0.859
Node169	0.026975	10	0	17	24	2	0	5	11	0.641	0.058	1.049	1.457	0.175	0.058	0.35	0.699	0.938	0.975	0.891	0.9	0.901	0.849	0.792	0.704
Node170	0.059942	58	3	83	258	30	5	18	29	1.547	0.105	2.203	6.792	0.813	0.157	0.498	0.787	0.835	0.936	0.771	0.611	0.689	0.644	0.732	0.681
Node171	0.015968	26	5	49	112	6	1	9	6	2.658	0.591	4.922	11.123	0.689	0.197	0.984	0.689	0.728	0.581	0.568	0.468	0.72	0.587	0.591	0.707
Node172	0.008238	34	5	26	60	24	4	9	32	6.678	1.145	5.152	11.639	4.77	0.954	1.908	6.296	0.474	0.376	0.555	0.455	0.291	0.208	0.434	0.219
Node173	0.012714	52	6	40	55	16	2	12	9	6.552	0.865	5.069	6.923	2.102	0.371	1.607	1.236	0.48	0.462	0.559	0.606	0.481	0.422	0.475	0.586
Node174	0.00622	101	12	115	154	92	20	51	58	25.774	3.285	29.312	39.167	23.5	5.306	13.14	14.909	0.124	0.122	0.105	0.133	0.052	0.028	0.075	0.088
Node175	0.016433	118	11	86	159	78	7	61	39	11.383	1.148	8.322	15.304	7.556	0.765	5.93	3.826	0.315	0.375	0.411	0.372	0.198	0.252	0.186	0.324

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Node176	0.010719	75	14	53	130	27	3	19	21	11.144	2.199	7.918	19.209	4.106	0.587	2.933	3.226	0.321	0.2	0.426	0.306	0.324	0.31	0.332	0.363
Node177	0.036259	81	8	63	101	48	2	24	45	3.555	0.39	2.774	4.422	2.124	0.13	1.084	1.994	0.655	0.7	0.72	0.719	0.478	0.69	0.569	0.476
Node178	0.016658	37	8	47	75	22	10	23	25	3.586	0.849	4.529	7.171	2.17	1.038	2.265	2.453	0.653	0.468	0.592	0.596	0.473	0.193	0.392	0.427
Node179	0.022372	186	27	292	1023	55	10	42	44	13.138	1.967	20.586	71.944	3.934	0.773	3.021	3.162	0.276	0.226	0.17	0.047	0.334	0.249	0.325	0.367
Node180	0.032349	77	8	107	221	38	7	20	33	3.79	0.437	5.248	10.787	1.895	0.389	1.02	1.652	0.638	0.669	0.549	0.478	0.505	0.41	0.583	0.52
Node181	0.018174	44	6	23	57	28	1	17	12	3.892	0.605	2.076	5.016	2.508	0.173	1.557	1.124	0.631	0.573	0.783	0.69	0.439	0.62	0.483	0.607
Node182	0.010843	135	18	114	305	41	5	29	43	19.714	2.754	16.67	44.357	6.088	0.87	4.349	6.378	0.176	0.153	0.218	0.111	0.24	0.226	0.246	0.217
Node183	0.02146	71	11	47	107	23	3	11	10	5.274	0.879	3.516	7.91	1.758	0.293	0.879	0.806	0.544	0.457	0.661	0.569	0.523	0.483	0.616	0.676
Node184	0.031942	77	7	83	103	29	7	37	44	3.838	0.394	4.134	5.118	1.476	0.394	1.87	2.214	0.635	0.698	0.617	0.685	0.563	0.406	0.438	0.451
Node185	0.006311	26	9	22	28	8	1	6	11	6.725	2.491	5.729	7.223	2.242	0.498	1.744	2.989	0.472	0.173	0.523	0.594	0.465	0.348	0.455	0.381
Node186	0.154817	249	26	278	490	0	0	0	0	2.538	0.274	2.833	4.985	0.01	0.01	0.01	0.01	0.738	0.787	0.715	0.691	0.994	0.987	0.994	0.994
Node187	0.02055	132	21	167	362	41	6	28	38	10.173	1.683	12.85	27.765	3.212	0.535	2.218	2.983	0.348	0.266	0.286	0.208	0.38	0.331	0.397	0.381
Node188	0.035135	106	12	98	208	13	2	10	10	4.787	0.582	4.429	9.35	0.626	0.134	0.492	0.492	0.573	0.585	0.598	0.52	0.737	0.682	0.734	0.765
Node189	0.007692	80	7	74	128	40	4	31	35	16.552	1.635	15.326	26.361	8.378	1.022	6.539	7.357	0.217	0.274	0.239	0.221	0.18	0.196	0.169	0.191
Node190	0.016766	46	8	47	75	9	1	16	12	4.406	0.844	4.5	7.125	0.937	0.187	1.594	1.219	0.597	0.47	0.594	0.598	0.661	0.6	0.477	0.589
Node191	0.025855	108	5	97	158	22	1	22	15	6.626	0.365	5.958	9.666	1.398	0.122	1.398	0.973	0.476	0.718	0.512	0.51	0.575	0.705	0.509	0.638
Node192	0.0272	118	20	117	210	67	5	42	54	6.877	1.214	6.819	12.193	3.93	0.347	2.485	3.178	0.465	0.358	0.471	0.441	0.334	0.439	0.37	0.366

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Node193	0.017997	126	11	161	307	40	13	29	30	11.092	1.048	14.149	26.9	3.581	1.223	2.62	2.707	0.323	0.402	0.26	0.216	0.355	0.165	0.358	0.404
Node194	0.026958	262	30	199	432	32	1	22	24	15.335	1.808	11.661	25.247	1.924	0.117	1.341	1.458	0.235	0.247	0.312	0.232	0.502	0.715	0.519	0.549
Node195	0.037903	110	15	107	156	23	4	12	13	4.603	0.664	4.479	6.511	0.995	0.207	0.539	0.581	0.584	0.545	0.595	0.623	0.649	0.573	0.717	0.738
Node196	0.011693	223	17	450	710	130	27	68	82	30.11	2.42	60.623	95.573	17.609	3.764	9.275	11.157	0.099	0.179	0.029	0.026	0.077	0.046	0.116	0.124
Node197	0.113963	60	6	93	162	47	4	35	55	0.841	0.097	1.296	2.248	0.662	0.069	0.497	0.772	0.915	0.943	0.863	0.846	0.728	0.821	0.732	0.685
Node198	0.019885	210	34	200	381	39	7	36	33	16.678	2.767	15.888	30.195	3.162	0.632	2.925	2.688	0.215	0.152	0.23	0.188	0.384	0.293	0.332	0.405
Node199	0.131449	165	12	197	342	44	0	17	25	1.985	0.155	2.368	4.101	0.538	0.012	0.215	0.311	0.79	0.89	0.756	0.736	0.763	0.983	0.857	0.832
Node200	0.050559	27	7	15	20	5	0	7	5	0.87	0.249	0.497	0.653	0.187	0.031	0.249	0.187	0.911	0.808	0.955	0.958	0.895	0.928	0.84	0.889
Node201	0.02886	44	4	16	26	9	1	4	9	2.451	0.272	0.926	1.471	0.545	0.109	0.272	0.545	0.746	0.788	0.905	0.899	0.761	0.73	0.828	0.749
Node202	0.065903	151	24	167	371	39	6	23	29	3.625	0.596	4.007	8.872	0.954	0.167	0.572	0.716	0.65	0.578	0.626	0.536	0.657	0.629	0.705	0.699
Node203	0.00915	73	13	83	182	42	9	38	24	12.712	2.405	14.43	31.436	7.387	1.718	6.7	4.295	0.285	0.181	0.255	0.179	0.202	0.117	0.165	0.298
Node204	0.049373	159	18	207	360	7	1	10	6	5.094	0.605	6.622	11.493	0.255	0.064	0.35	0.223	0.555	0.573	0.48	0.458	0.865	0.835	0.791	0.871
Node205	0.128902	225	21	215	467	7	2	1	3	2.756	0.268	2.634	5.707	0.098	0.037	0.024	0.049	0.719	0.792	0.732	0.658	0.941	0.911	0.983	0.967
Node206	0.051792	277	32	339	1398	79	10	46	81	8.437	1.002	10.319	42.458	2.428	0.334	1.426	2.489	0.403	0.416	0.347	0.118	0.446	0.449	0.504	0.424
Node207	0.015061	63	9	77	97	34	4	32	41	6.679	1.044	8.14	10.227	3.653	0.522	3.444	4.383	0.474	0.404	0.418	0.494	0.351	0.337	0.295	0.294
Node208	0.013527	142	24	192	656	34	3	30	57	16.617	2.905	22.427	76.344	4.067	0.465	3.602	6.74	0.216	0.143	0.153	0.042	0.326	0.365	0.285	0.206
Node209	0.134154	107	22	161	205	21	1	14	21	1.265	0.269	1.898	2.414	0.258	0.023	0.176	0.258	0.866	0.791	0.8	0.835	0.863	0.951	0.88	0.855

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Node210	0.063819	47	9	77	196	10	2	8	8	1.182	0.246	1.921	4.852	0.271	0.074	0.222	0.222	0.876	0.81	0.798	0.698	0.858	0.809	0.854	0.871
Node211	0.066565	114	4	169	356	12	0	6	5	2.716	0.118	4.014	8.43	0.307	0.024	0.165	0.142	0.723	0.924	0.625	0.55	0.843	0.95	0.886	0.912
Node212	0.218688	307	31	342	509	67	4	27	20	2.214	0.23	2.465	3.666	0.489	0.036	0.201	0.151	0.768	0.824	0.747	0.76	0.779	0.913	0.865	0.907
Node213	0.056002	148	18	221	873	75	11	45	44	4.182	0.533	6.231	24.531	2.133	0.337	1.291	1.263	0.611	0.611	0.498	0.239	0.477	0.446	0.528	0.581
Node214	0.011153	168	17	183	278	33	5	19	27	23.817	2.537	25.931	39.319	4.792	0.846	2.819	3.946	0.138	0.17	0.126	0.132	0.29	0.231	0.341	0.317
Node215	0.051531	160	18	179	370	55	12	36	55	4.911	0.58	5.49	11.316	1.708	0.397	1.129	1.708	0.565	0.587	0.536	0.463	0.529	0.405	0.559	0.512
Node216	0.052398	305	29	303	750	144	16	75	85	9.179	0.9	9.119	22.528	4.35	0.51	2.28	2.58	0.378	0.449	0.384	0.262	0.311	0.342	0.391	0.415
Paracoccidioides brasiliensis	0.048447	161	21	140	189	158	28	105	120	5.256	0.714	4.575	6.164	5.159	0.941	3.439	3.926	0.545	0.522	0.589	0.637	0.274	0.211	0.296	0.318
Phanerochaete chrysosporium	0.031977	259	29	307	374	347	38	243	355	12.78	1.475	15.14	18.433	17.106	1.917	11.994	17.499	0.283	0.302	0.242	0.318	0.08	0.104	0.084	0.071
Phlebia brevispora	0.051163	481	55	694	851	108	13	55	75	14.808	1.72	21.351	26.175	3.349	0.43	1.72	2.335	0.244	0.26	0.163	0.222	0.371	0.384	0.459	0.439
Phlebiopsis gigantea	0.047068	229	24	352	471	111	16	99	148	7.681	0.835	11.788	15.762	3.74	0.568	3.339	4.976	0.432	0.473	0.309	0.363	0.345	0.317	0.302	0.267
Phycomyces blakesleeanus	0.088207	378	34	399	422	28	0	13	12	6.754	0.624	7.128	7.538	0.517	0.018	0.249	0.232	0.47	0.564	0.458	0.582	0.77	0.967	0.839	0.867
Pichia membranifaciens	0.200642	86	16	105	170	101	11	65	105	0.682	0.133	0.83	1.34	0.799	0.094	0.517	0.83	0.934	0.91	0.916	0.908	0.693	0.762	0.725	0.67
Pichia stipitis	0.073887	79	10	113	198	19	4	28	49	1.702	0.234	2.425	4.233	0.425	0.106	0.617	1.064	0.819	0.82	0.75	0.729	0.8	0.735	0.691	0.619
Pleurotus ostreatus	0.064965	396	47	412	714	136	14	95	150	9.605	1.161	9.992	17.299	3.315	0.363	2.323	3.653	0.365	0.371	0.357	0.336	0.373	0.427	0.386	0.334
Podospora anserina	0.052217	339	29	260	466	78	15	64	135	10.235	0.903	7.857	14.057	2.378	0.482	1.957	4.094	0.346	0.448	0.428	0.398	0.452	0.356	0.428	0.309
Puccinia graminis	0.130292	443	55	465	819	95	15	77	94	5.356	0.676	5.622	9.892	1.158	0.193	0.941	1.146	0.54	0.539	0.529	0.503	0.617	0.592	0.601	0.603

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Punctularia strigosozonata	0.098351	376	39	430	681	171	20	128	169	6.025	0.639	6.888	10.899	2.749	0.336	2.062	2.717	0.505	0.556	0.468	0.474	0.417	0.447	0.415	0.403
Pyrenophora teres	0.002117	294	37	283	322	82	10	95	96	218.992	28.209	210.826	239.778	61.615	8.166	71.265	72.008	0.001	0.001	0.001	0.002	0.009	0.015	0.003	0.006
Pyrenophora triticirepentis	0.022557	265	33	265	347	150	16	102	135	18.535	2.369	18.535	24.249	10.522	1.185	7.177	9.477	0.19	0.184	0.194	0.242	0.143	0.171	0.153	0.148
Rhizopus oryzae	0.094333	357	34	504	412	92	16	83	92	5.965	0.583	8.415	6.882	1.55	0.283	1.4	1.55	0.508	0.584	0.408	0.608	0.552	0.492	0.508	0.535
Rhodotorula graminis	0.074386	271	35	304	472	33	4	36	33	5.747	0.761	6.445	9.995	0.718	0.106	0.782	0.718	0.519	0.502	0.488	0.5	0.713	0.737	0.642	0.699
Rhytidhysteron rufulum	0.05024	409	40	345	492	166	28	125	144	12.827	1.283	10.825	15.424	5.225	0.907	3.942	4.537	0.282	0.342	0.334	0.37	0.271	0.218	0.266	0.286
Saccharomyces cerevisiae	0.084726	68	8	74	107	28	3	24	37	1.28	0.167	1.391	2.004	0.538	0.074	0.464	0.705	0.865	0.879	0.853	0.862	0.763	0.808	0.744	0.702
Schizophyllum commune	0.11636	399	54	497	752	189	19	148	198	5.403	0.743	6.727	10.172	2.567	0.27	2.013	2.688	0.537	0.51	0.475	0.495	0.433	0.504	0.421	0.405
Schizosaccharomyces cryophilus	0.021077	29	4	46	61	17	4	24	33	2.237	0.373	3.505	4.624	1.342	0.373	1.864	2.536	0.766	0.712	0.662	0.709	0.584	0.42	0.439	0.419
Schizosaccharomyces japonicus	0.106829	69	12	74	111	28	3	16	20	1.03	0.191	1.103	1.648	0.427	0.059	0.25	0.309	0.893	0.858	0.885	0.886	0.799	0.848	0.839	0.833
Schizosaccharomyces octosporus	0.024974	31	9	25	51	18	4	17	14	2.014	0.629	1.636	3.273	1.196	0.315	1.133	0.944	0.788	0.561	0.827	0.782	0.61	0.464	0.558	0.644
Schizosaccharomyces pombe	0.053416	43	6	38	69	14	0	10	16	1.295	0.206	1.148	2.06	0.441	0.029	0.324	0.5	0.863	0.845	0.88	0.858	0.794	0.933	0.803	0.763
Sclerotinia sclerotiorum	0.020058	384	40	336	395	117	9	66	48	30.169	3.213	26.408	31.031	9.247	0.784	5.25	3.84	0.098	0.125	0.122	0.182	0.163	0.247	0.208	0.323
Septoria musiva	0.001576	133	19	171	194	72	6	47	53	133.606	19.941	171.495	194.427	72.786	6.979	47.859	53.841	0.004	0.004	0.002	0.004	0.007	0.019	0.008	0.011
Septoria populicola	0.002978	123	15	138	185	96	19	78	110	65.45	8.445	73.367	98.175	51.199	10.556	41.698	58.588	0.023	0.026	0.019	0.024	0.014	0.01	0.011	0.009
Serpula lacrymans	0.089065	493	38	626	655	127	25	121	164	8.718	0.688	11.065	11.577	2.259	0.459	2.153	2.912	0.393	0.534	0.327	0.456	0.464	0.368	0.405	0.387
Setosphaeria turcica	0.011641	251	30	239	410	96	16	72	82	34.027	4.186	32.406	55.496	13.098	2.295	9.857	11.207	0.081	0.086	0.09	0.077	0.112	0.084	0.108	0.123

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Spathaspora passalidarum	0.067977	73	8	124	158	88	13	99	117	1.711	0.208	2.89	3.677	2.058	0.324	2.312	2.728	0.818	0.843	0.71	0.759	0.486	0.457	0.387	0.402
Spizellomyces punctatus	0.138351	444	43	353	515	0	0	0	0	5.056	0.5	4.022	5.862	0.011	0.011	0.011	0.011	0.557	0.631	0.625	0.651	0.993	0.984	0.993	0.993
Sporobolomyces roseus	0.076031	152	17	205	245	123	14	84	109	3.163	0.372	4.259	5.086	2.563	0.31	1.757	2.274	0.686	0.713	0.609	0.686	0.434	0.468	0.453	0.445
Sporotrichum thermophile	0.033132	188	26	187	293	148	14	95	157	8.966	1.281	8.919	13.948	7.069	0.712	4.554	7.496	0.385	0.342	0.39	0.4	0.211	0.267	0.236	0.187
Thielavia terrestris	0.044292	201	33	218	293	101	16	64	77	7.168	1.207	7.772	10.433	3.62	0.603	2.307	2.768	0.452	0.36	0.432	0.488	0.353	0.303	0.388	0.398
Trametes versicolor	0.037437	280	34	408	559	70	12	51	77	11.798	1.47	17.172	23.512	2.981	0.546	2.183	3.275	0.305	0.303	0.211	0.25	0.398	0.326	0.401	0.359
Tremella mesenterica	0.090196	220	31	234	331	37	6	17	25	3.851	0.558	4.095	5.786	0.662	0.122	0.314	0.453	0.634	0.598	0.62	0.654	0.727	0.704	0.808	0.779
Trichoderma atroviride	0.029649	249	19	187	285	103	14	88	130	13.254	1.06	9.967	15.162	5.514	0.795	4.718	6.945	0.273	0.399	0.358	0.375	0.26	0.244	0.229	0.201
Trichoderma reesei	0.019329	109	10	105	168	68	11	53	57	8.945	0.895	8.62	13.743	5.611	0.976	4.391	4.716	0.386	0.451	0.401	0.405	0.256	0.204	0.244	0.278
Trichoderma virens	0.014963	228	27	224	314	62	11	67	78	24.055	2.941	23.635	33.089	6.618	1.261	7.143	8.298	0.136	0.141	0.143	0.167	0.223	0.161	0.154	0.17
Trichophyton equinum	0.01433	153	22	144	149	211	15	121	140	16.892	2.523	15.905	16.453	23.254	1.755	13.382	15.466	0.212	0.171	0.23	0.351	0.052	0.114	0.073	0.084
Uncinocarpus reesii	0.043036	203	31	214	282	247	42	199	263	7.451	1.169	7.852	10.336	9.058	1.57	7.305	9.642	0.441	0.369	0.428	0.491	0.166	0.129	0.151	0.145
Ustilago maydis	0.251644	244	41	234	301	84	15	91	89	1.53	0.262	1.468	1.886	0.531	0.1	0.575	0.562	0.837	0.797	0.845	0.87	0.765	0.749	0.705	0.743
Verticillium alboatrum	0.065004	250	25	259	225	360	51	220	282	6.069	0.629	6.287	5.465	8.729	1.257	5.344	6.843	0.503	0.561	0.495	0.668	0.173	0.161	0.205	0.204
Verticillium dahliae	0.021303	251	24	214	286	82	14	49	59	18.593	1.845	15.863	21.175	6.124	1.107	3.689	4.427	0.189	0.243	0.23	0.278	0.239	0.182	0.28	0.292
Wickerhamomyces anomalus	0.13074	227	21	267	358	44	8	44	43	2.741	0.264	3.222	4.316	0.541	0.108	0.541	0.529	0.721	0.795	0.683	0.725	0.762	0.732	0.716	0.754
Wolfiporia cocos	0.044784	314	38	389	527	114	19	85	113	11.056	1.369	13.688	18.531	4.036	0.702	3.018	4.001	0.324	0.323	0.268	0.316	0.328	0.27	0.325	0.314

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.09**	**SC λ = 0.03**	**SN λ = 0.09**	**NR λ = 0.13**	**NC λ = 0.08**	**SC λ = -0.02**	**SN λ = 0.07**	**NR λ = 0.07**
Yarrowia lipolytica	0.277999	187	27	172	269	58	8	42	37	1.063	0.158	0.978	1.527	0.334	0.051	0.243	0.215	0.889	0.887	0.899	0.895	0.833	0.87	0.843	0.875

Table 2.3.13. Investigation for evolutionary bursts across internal nodes of the fungal phylogeny

Each site (branch or leaf node) is annotated with its associated branch length (κ), character state changes (T_b and T_d), birth and decay rates (f_b and f_d) and their associated *P*-values derived from a *Q*-function ($P(f_b)$ and $P(f_d)$). λ values used for Box-Cox transformations are given beneath the RC identifiers in the *Q*-function columns. Significant evolutionary bursts ($P \leq \alpha_B \leq 2.36e^{-04}$) are emboldened and underlined in the $P(f_b)$ and $P(f_d)$ columns. A significant burst was observed at one site, "Node 196", which represents the divergence of *Fusarium verticillioides* and *F. oxysorum* from *F. graminearum*.

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.33**	**SC λ = 0.30**	**SN λ = 0.26**	**NR λ = 0.24**	**NC λ = 0.17**	**SC λ = -0.05**	**SN λ = 0.17**	**NR λ = 0.18**
Node112	0.045345	82	10	51	74	14	1	6	13	2.877	0.381	1.802	2.6	0.52	0.069	0.243	0.485	0.7	0.7	0.787	0.802	0.709	0.763	0.803	0.696
Node113	0.020864	173	31	124	167	10	1	6	6	13.108	2.411	9.417	12.656	0.829	0.151	0.527	0.527	0.137	0.071	0.239	0.363	0.597	0.534	0.642	0.677
Node114	0.011563	150	23	101	181	3	0	6	6	20.525	3.262	13.865	24.739	0.544	0.136	0.952	0.952	0.04	0.028	0.122	0.154	0.699	0.567	0.477	0.524
Node115	0.044031	417	51	250	273	48	9	41	79	14.922	1.856	8.96	9.781	1.749	0.357	1.499	2.856	0.102	0.132	0.257	0.451	0.384	0.28	0.34	0.211
Node116	0.033931	74	7	52	56	30	2	24	40	3.474	0.371	2.455	2.64	1.436	0.139	1.158	1.899	0.642	0.708	0.713	0.8	0.442	0.56	0.418	0.321
Node117	0.12487	64	7	44	58	40	2	29	39	0.818	0.101	0.566	0.743	0.516	0.038	0.378	0.504	0.915	0.921	0.936	0.936	0.711	0.89	0.72	0.688
Node118	0.04653	21	0	9	9	29	8	24	43	0.743	0.034	0.338	0.338	1.013	0.304	0.845	1.486	0.923	0.973	0.963	0.968	0.543	0.322	0.512	0.394
Node119	0.078317	66	2	45	31	0	0	0	0	1.345	0.06	0.923	0.642	0.02	0.02	0.02	0.02	0.86	0.953	0.893	0.944	0.982	0.962	0.98	0.98
Node120	0.057596	35	2	23	23	0	0	0	0	0.982	0.082	0.655	0.655	0.027	0.027	0.027	0.027	0.898	0.936	0.926	0.943	0.977	0.934	0.973	0.974
Node121	0.048674	400	43	246	321	0	0	0	0	12.949	1.421	7.976	10.398	0.032	0.032	0.032	0.032	0.141	0.217	0.299	0.43	0.973	0.913	0.968	0.97
Node122	0.042568	0	0	0	0	0	0	0	0	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.993	0.97	0.995	0.994	0.97	0.893	0.964	0.966

		T_b				T_d				f_b				f_d				$P(f_b)$				P(f_d)			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.33**	**SC λ = 0.30**	**SN λ = 0.26**	**NR λ = 0.24**	**NC λ = 0.17**	**SC λ = -0.05**	**SN λ = 0.17**	**NR λ = 0.18**
Node123	0.011903	92	18	70	117	61	12	52	46	12.281	2.509	9.376	15.582	8.187	1.717	6.999	6.206	0.157	0.063	0.241	0.292	0.048	0.042	0.032	0.064
Node124	0.014721	19	5	29	45	16	1	10	13	2.135	0.641	3.203	4.912	1.815	0.214	1.174	1.495	0.776	0.527	0.635	0.663	0.373	0.425	0.414	0.392
Node125	0.014804	30	2	28	40	22	2	14	24	3.291	0.319	3.079	4.353	2.442	0.319	1.593	2.654	0.659	0.747	0.648	0.694	0.287	0.31	0.322	0.229
Node126	0.0128	35	3	38	54	21	4	21	27	4.421	0.491	4.789	6.754	2.701	0.614	2.701	3.438	0.557	0.622	0.495	0.572	0.26	0.161	0.18	0.167
Node127	0.006965	29	1	22	34	9	0	5	8	6.77	0.451	5.19	7.898	2.257	0.226	1.354	2.031	0.386	0.649	0.464	0.522	0.31	0.408	0.371	0.302
Node128	0.026124	42	4	41	58	5	0	2	1	2.587	0.301	2.527	3.55	0.361	0.06	0.181	0.12	0.729	0.761	0.705	0.742	0.78	0.797	0.847	0.903
Node129	0.030842	149	27	117	177	0	0	0	0	7.644	1.427	6.014	9.071	0.051	0.051	0.051	0.051	0.336	0.215	0.407	0.477	0.96	0.834	0.952	0.955
Node130	0.034674	406	46	258	340	0	0	0	0	18.449	2.131	11.741	15.458	0.045	0.045	0.045	0.045	0.057	0.097	0.167	0.295	0.964	0.858	0.957	0.959
Node131	0.058997	0	0	0	0	20	1	14	18	0.027	0.027	0.027	0.027	0.559	0.053	0.4	0.506	0.994	0.978	0.996	0.995	0.693	0.825	0.707	0.687
Node132	0.037652	0	0	0	0	4	0	4	6	0.042	0.042	0.042	0.042	0.209	0.042	0.209	0.292	0.992	0.967	0.994	0.993	0.86	0.873	0.827	0.794
Node133	0.098772	79	8	44	55	29	4	20	30	1.273	0.143	0.716	0.891	0.477	0.08	0.334	0.493	0.867	0.887	0.918	0.925	0.727	0.726	0.745	0.693
Node134	0.003028	20	3	25	81	9	2	11	18	10.902	2.077	13.497	42.569	5.191	1.557	6.23	9.864	0.197	0.103	0.128	0.05	0.111	0.048	0.042	0.023
Node135	0.011734	74	7	78	200	22	1	14	10	10.047	1.072	10.582	26.925	3.081	0.268	2.009	1.474	0.227	0.323	0.2	0.134	0.225	0.358	0.256	0.396
Node136	0.054824	110	14	127	283	29	0	14	17	3.182	0.43	3.67	8.142	0.86	0.029	0.43	0.516	0.67	0.664	0.59	0.512	0.588	0.928	0.691	0.683
Node137	0.045604	172	19	111	175	15	2	17	15	5.963	0.689	3.86	6.066	0.551	0.103	0.62	0.551	0.439	0.499	0.573	0.604	0.696	0.651	0.599	0.667
Node138	0.028646	133	23	96	144	2	0	6	2	7.353	1.317	5.322	7.956	0.165	0.055	0.384	0.165	0.352	0.244	0.455	0.52	0.886	0.818	0.716	0.872
Node139	0.016957	102	15	64	74	5	0	5	5	9.547	1.483	6.025	6.952	0.556	0.093	0.556	0.556	0.246	0.202	0.407	0.563	0.694	0.683	0.628	0.665
Node140	0.149864	97	7	149	279	66	8	49	54	1.028	0.084	1.573	2.937	0.703	0.094	0.524	0.577	0.893	0.935	0.814	0.78	0.639	0.678	0.643	0.656
Node141	0.031404	28	5	52	59	14	0	7	14	1.452	0.3	2.653	3.003	0.751	0.05	0.4	0.751	0.848	0.761	0.692	0.776	0.623	0.838	0.707	0.59
Node142	0.128676	75	1	42	90	57	5	29	41	0.928	0.024	0.525	1.112	0.708	0.073	0.366	0.513	0.904	0.98	0.941	0.908	0.637	0.748	0.726	0.684
Node143	0.050053	77	15	45	96	34	9	22	17	2.449	0.502	1.445	3.046	1.099	0.314	0.722	0.565	0.743	0.614	0.83	0.773	0.52	0.313	0.557	0.661
Node144	0.037245	47	6	46	50	26	4	16	16	2.026	0.295	1.983	2.152	1.139	0.211	0.717	0.717	0.787	0.765	0.766	0.833	0.51	0.429	0.559	0.602
Node145	0.058749	59	14	56	73	7	1	3	9	1.605	0.401	1.525	1.98	0.214	0.054	0.107	0.268	0.832	0.685	0.82	0.845	0.857	0.824	0.904	0.809
Node146	0.024533	190	25	121	261	94	3	60	45	12.237	1.666	7.816	16.786	6.087	0.256	3.908	2.947	0.159	0.164	0.307	0.268	0.085	0.371	0.104	0.203
Node147	0.018571	39	5	27	43	8	0	5	2	3.385	0.508	2.37	3.724	0.762	0.085	0.508	0.254	0.65	0.611	0.723	0.731	0.619	0.709	0.651	0.817

		T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.33**	**SC λ = 0.30**	**SN λ = 0.26**	**NR λ = 0.24**	**NC λ = 0.17**	**SC λ = -0.05**	**SN λ = 0.17**	**NR λ = 0.18**
Node148	0.00879	62	6	33	46	6	0	5	4	11.265	1.252	6.08	8.404	1.252	0.179	1.073	0.894	0.186	0.263	0.403	0.502	0.483	0.48	0.441	0.542
Node149	0.047793	223	19	160	266	10	1	11	11	7.367	0.658	5.295	8.781	0.362	0.066	0.395	0.395	0.351	0.517	0.457	0.487	0.779	0.776	0.71	0.74
Node150	0.067278	247	35	193	255	7	2	5	10	5.794	0.841	4.532	5.981	0.187	0.07	0.14	0.257	0.451	0.421	0.515	0.609	0.872	0.76	0.878	0.815
Node151	0.013209	98	10	91	213	11	2	11	11	11.78	1.309	10.947	25.464	1.428	0.357	1.428	1.428	0.171	0.247	0.189	0.147	0.444	0.28	0.355	0.405
Node152	0.032306	234	26	294	604	109	18	63	95	11.434	1.314	14.353	29.436	5.352	0.924	3.114	4.671	0.181	0.245	0.113	0.113	0.106	0.099	0.148	0.106
Node153	0.008481	61	4	45	75	27	4	29	34	11.491	0.927	8.525	14.085	5.189	0.927	5.56	6.487	0.179	0.382	0.275	0.326	0.111	0.099	0.054	0.059
Node154	0.030228	150	22	151	296	31	5	16	34	7.852	1.196	7.904	15.443	1.664	0.312	0.884	1.82	0.325	0.28	0.303	0.295	0.399	0.315	0.499	0.334
Node155	0.042568	174	24	148	314	0	0	0	0	6.462	0.923	5.502	11.631	0.037	0.037	0.037	0.037	0.406	0.383	0.442	0.392	0.97	0.893	0.964	0.966
Node156	0.099798	149	29	136	193	38	4	21	21	2.362	0.472	2.158	3.055	0.614	0.079	0.346	0.346	0.752	0.635	0.746	0.773	0.672	0.729	0.738	0.764
Node157	0.009157	109	13	113	196	23	5	19	31	18.882	2.403	19.568	33.815	4.12	1.03	3.433	5.493	0.053	0.071	0.053	0.086	0.156	0.086	0.128	0.08
Node158	0.029582	39	11	48	74	25	2	18	23	2.125	0.638	2.604	3.985	1.381	0.159	1.01	1.275	0.777	0.529	0.697	0.716	0.454	0.516	0.459	0.439
Node159	0.021386	64	11	60	137	17	3	15	26	4.777	0.882	4.483	10.142	1.323	0.294	1.176	1.984	0.528	0.402	0.519	0.439	0.466	0.332	0.413	0.309
Node160	0.017813	72	9	69	115	21	1	23	25	6.442	0.882	6.177	10.236	1.941	0.176	2.118	2.294	0.407	0.402	0.397	0.435	0.353	0.484	0.242	0.268
Node161	0.008789	34	6	27	56	13	1	9	17	6.26	1.252	5.008	10.194	2.504	0.358	1.788	3.219	0.419	0.263	0.478	0.437	0.281	0.279	0.288	0.182
Node162	0.013638	94	12	100	217	30	4	28	24	10.949	1.498	11.64	25.124	3.573	0.576	3.342	2.881	0.196	0.198	0.17	0.15	0.189	0.173	0.133	0.209
Node163	0.031702	73	16	80	160	32	3	19	12	3.669	0.843	4.016	7.982	1.636	0.198	0.992	0.645	0.624	0.42	0.559	0.519	0.404	0.448	0.465	0.629
Node164	0.023896	70	7	45	86	69	5	20	21	4.67	0.526	3.026	5.723	4.604	0.395	1.381	1.447	0.537	0.598	0.653	0.621	0.133	0.255	0.365	0.401
Node165	0.064938	307	27	335	904	23	2	24	26	7.455	0.678	8.133	21.905	0.581	0.073	0.605	0.654	0.346	0.506	0.292	0.187	0.684	0.751	0.606	0.626
Node166	0.036247	46	5	68	158	18	0	9	16	2.038	0.26	2.992	6.895	0.824	0.043	0.434	0.737	0.786	0.793	0.657	0.566	0.599	0.866	0.689	0.594
Node167	0.025037	14	5	22	50	1	0	5	2	0.942	0.377	1.444	3.202	0.126	0.063	0.377	0.188	0.902	0.703	0.83	0.763	0.91	0.787	0.72	0.857
Node168	0.113266	105	3	140	324	16	3	4	17	1.471	0.056	1.957	4.51	0.236	0.056	0.069	0.25	0.846	0.957	0.769	0.685	0.844	0.816	0.936	0.819
Node169	0.026975	10	0	17	24	2	0	5	11	0.641	0.058	1.049	1.457	0.175	0.058	0.35	0.699	0.933	0.954	0.878	0.882	0.879	0.805	0.736	0.608
Node170	0.059942	58	3	83	258	30	5	18	29	1.547	0.105	2.203	6.792	0.813	0.157	0.498	0.787	0.838	0.918	0.741	0.57	0.602	0.521	0.656	0.577
Node171	0.015968	26	5	49	112	6	1	9	6	2.658	0.591	4.922	11.123	0.689	0.197	0.984	0.689	0.722	0.557	0.484	0.407	0.644	0.45	0.467	0.612
Node172	0.008238	34	5	26	60	24	4	9	32	6.678	1.145	5.152	11.639	4.77	0.954	1.908	6.296	0.392	0.297	0.467	0.391	0.127	0.095	0.27	0.062

		T_b				T_d				f_b				f_d				$P(f_b)$				P(f_d)			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC λ = 0.33**	**SC λ = 0.30**	**SN λ = 0.26**	**NR λ = 0.24**	**NC λ = 0.17**	**SC λ = -0.05**	**SN λ = 0.17**	**NR λ = 0.18**
Node173	0.012714	52	6	40	55	16	2	12	9	6.552	0.865	5.069	6.923	2.102	0.371	1.607	1.236	0.4	0.409	0.473	0.564	0.33	0.27	0.319	0.448
Node174	0.00622	101	12	115	154	92	20	51	58	25.774	3.285	29.312	39.167	23.5	5.306	13.14	14.909	0.017	0.027	0.014	0.062	0.002	0.006	0.005	0.007
Node175	0.016433	118	11	86	159	78	7	61	39	11.383	1.148	8.322	15.304	7.556	0.765	5.93	3.826	0.182	0.296	0.284	0.298	0.057	0.125	0.047	0.144
Node176	0.010719	75	14	53	130	27	3	19	21	11.144	2.199	7.918	19.209	4.106	0.587	2.933	3.226	0.19	0.089	0.302	0.226	0.157	0.17	0.161	0.181
Node177	0.036259	81	8	63	101	48	2	24	45	3.555	0.39	2.774	4.422	2.124	0.13	1.084	1.994	0.635	0.693	0.679	0.691	0.327	0.58	0.438	0.307
Node178	0.016658	37	8	47	75	22	10	23	25	3.586	0.849	4.529	7.171	2.17	1.038	2.265	2.453	0.632	0.417	0.515	0.553	0.321	0.085	0.224	0.25
Node179	0.022372	186	27	292	1023	55	10	42	44	13.138	1.967	20.586	71.944	3.934	0.773	3.021	3.162	0.137	0.116	0.046	0.01	0.167	0.123	0.155	0.186
Node180	0.032349	77	8	107	221	38	7	20	33	3.79	0.437	5.248	10.787	1.895	0.389	1.02	1.652	0.613	0.659	0.46	0.418	0.36	0.259	0.456	0.362
Node181	0.018174	44	6	23	57	28	1	17	12	3.892	0.605	2.076	5.016	2.508	0.173	1.557	1.124	0.604	0.548	0.756	0.658	0.28	0.491	0.329	0.476
Node182	0.010843	135	18	114	305	41	5	29	43	19.714	2.754	16.67	44.357	6.088	0.87	4.349	6.378	0.046	0.048	0.081	0.045	0.085	0.107	0.087	0.061
Node183	0.02146	71	11	47	107	23	3	11	10	5.274	0.879	3.516	7.91	1.758	0.293	0.879	0.806	0.489	0.403	0.605	0.522	0.382	0.333	0.501	0.57
Node184	0.031942	77	7	83	103	29	7	37	44	3.838	0.394	4.134	5.118	1.476	0.394	1.87	2.214	0.609	0.691	0.549	0.652	0.434	0.255	0.276	0.278
Node185	0.006311	26	9	22	28	8	1	6	11	6.725	2.491	5.729	7.223	2.242	0.498	1.744	2.989	0.389	0.065	0.426	0.551	0.312	0.202	0.296	0.199
Node186	0.154817	249	26	278	490	0	0	0	0	2.538	0.274	2.833	4.985	0.01	0.01	0.01	0.01	0.734	0.782	0.673	0.66	0.99	0.991	0.989	0.989
Node187	0.02055	132	21	167	362	41	6	28	38	10.173	1.683	12.85	27.765	3.212	0.535	2.218	2.983	0.222	0.161	0.142	0.126	0.214	0.187	0.229	0.2
Node188	0.035135	106	12	98	208	13	2	10	10	4.787	0.582	4.429	9.35	0.626	0.134	0.492	0.492	0.527	0.563	0.524	0.466	0.667	0.57	0.659	0.693
Node189	0.007692	80	7	74	128	40	4	31	35	16.552	1.635	15.326	26.361	8.378	1.022	6.539	7.357	0.078	0.17	0.098	0.139	0.046	0.087	0.038	0.045
Node190	0.016766	46	8	47	75	9	1	16	12	4.406	0.844	4.5	7.125	0.937	0.187	1.594	1.219	0.559	0.42	0.518	0.555	0.564	0.465	0.322	0.452
Node191	0.025855	108	5	97	158	22	1	22	15	6.626	0.365	5.958	9.666	1.398	0.122	1.398	0.973	0.395	0.712	0.411	0.455	0.45	0.601	0.361	0.518
Node192	0.0272	118	20	117	210	67	5	42	54	6.877	1.214	6.819	12.193	3.93	0.347	2.485	3.178	0.38	0.275	0.359	0.375	0.167	0.287	0.2	0.185
Node193	0.017997	126	11	161	307	40	13	29	30	11.092	1.048	14.149	26.9	3.581	1.223	2.62	2.707	0.191	0.332	0.117	0.134	0.188	0.068	0.187	0.224
Node194	0.026958	262	30	199	432	32	1	22	24	15.335	1.808	11.661	25.247	1.924	0.117	1.341	1.458	0.095	0.14	0.17	0.149	0.356	0.614	0.373	0.399
Node195	0.037903	110	15	107	156	23	4	12	13	4.603	0.664	4.479	6.511	0.995	0.207	0.539	0.581	0.542	0.514	0.52	0.583	0.548	0.434	0.636	0.655
Node196	0.011693	223	17	450	710	130	27	68	82	30.11	2.42	60.623	95.573	17.609	3.764	9.275	11.157	0.009	0.07	$\underline{3.26e^{-04}}$	0.003	0.006	0.012	0.016	0.016
Node197	0.113963	60	6	93	162	47	4	35	55	0.841	0.097	1.296	2.248	0.662	0.069	0.497	0.772	0.913	0.925	0.848	0.826	0.654	0.764	0.657	0.582

		T_b				T_d				f_b				f_d				$P(f_b)$				P(f_d)			
Site	**κ**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC** λ = 0.33	**SC** λ = 0.30	**SN** λ = 0.26	**NR** λ = 0.24	**NC** λ = 0.17	**SC** λ = -0.05	**SN** λ = 0.17	**NR** λ = 0.18
Node198	0.019885	210	34	200	381	39	7	36	33	16.678	2.767	15.888	30.195	3.162	0.632	2.925	2.688	0.076	0.048	0.09	0.108	0.219	0.156	0.162	0.226
Node199	0.131449	165	12	197	342	44	0	17	25	1.985	0.155	2.368	4.101	0.538	0.012	0.215	0.311	0.792	0.877	0.723	0.709	0.702	0.987	0.822	0.784
Node200	0.050559	27	7	15	20	5	0	7	5	0.87	0.249	0.497	0.653	0.187	0.031	0.249	0.187	0.91	0.802	0.944	0.943	0.872	0.918	0.799	0.858
Node201	0.02886	44	4	16	26	9	1	4	9	2.451	0.272	0.926	1.471	0.545	0.109	0.272	0.545	0.743	0.783	0.893	0.881	0.699	0.635	0.784	0.67
Node202	0.065903	151	24	167	371	39	6	23	29	3.625	0.596	4.007	8.872	0.954	0.167	0.572	0.716	0.628	0.554	0.56	0.484	0.56	0.502	0.621	0.602
Node203	0.00915	73	13	83	182	42	9	38	24	12.712	2.405	14.43	31.436	7.387	1.718	6.7	4.295	0.147	0.071	0.112	0.1	0.059	0.042	0.036	0.121
Node204	0.049373	159	18	207	360	7	1	10	6	5.094	0.605	6.622	11.493	0.255	0.064	0.35	0.223	0.503	0.549	0.37	0.396	0.834	0.784	0.736	0.835
Node205	0.128902	225	21	215	467	7	2	1	3	2.756	0.268	2.634	5.707	0.098	0.037	0.024	0.049	0.712	0.786	0.694	0.622	0.928	0.895	0.976	0.957
Node206	0.051792	277	32	339	1398	79	10	46	81	8.437	1.002	10.319	42.458	2.428	0.334	1.426	2.489	0.295	0.35	0.208	0.051	0.289	0.297	0.355	0.246
Node207	0.015061	63	9	77	97	34	4	32	41	6.679	1.044	8.14	10.227	3.653	0.522	3.444	4.383	0.392	0.334	0.292	0.436	0.183	0.193	0.127	0.117
Node208	0.013527	142	24	192	656	34	3	30	57	16.617	2.905	22.427	76.344	4.067	0.465	3.602	6.74	0.077	0.041	0.036	0.008	0.159	0.217	0.119	0.054
Node209	0.134154	107	22	161	205	21	1	14	21	1.265	0.269	1.898	2.414	0.258	0.023	0.176	0.258	0.868	0.785	0.776	0.815	0.832	0.949	0.851	0.814
Node210	0.063819	47	9	77	196	10	2	8	8	1.182	0.246	1.921	4.852	0.271	0.074	0.222	0.222	0.877	0.804	0.774	0.667	0.825	0.746	0.818	0.836
Node211	0.066565	114	4	169	356	12	0	6	5	2.716	0.118	4.014	8.43	0.307	0.024	0.165	0.142	0.716	0.908	0.559	0.501	0.807	0.948	0.858	0.888
Node212	0.218688	307	31	342	509	67	4	27	20	2.214	0.23	2.465	3.666	0.489	0.036	0.201	0.151	0.768	0.817	0.712	0.735	0.722	0.897	0.832	0.882
Node213	0.056002	148	18	221	873	75	11	45	44	4.182	0.533	6.231	24.531	2.133	0.337	1.291	1.263	0.578	0.594	0.394	0.157	0.326	0.295	0.385	0.442
Node214	0.011153	168	17	183	278	33	5	19	27	23.817	2.537	25.931	39.319	4.792	0.846	2.819	3.946	0.024	0.061	0.022	0.061	0.126	0.111	0.17	0.138
Node215	0.051531	160	18	179	370	55	12	36	55	4.911	0.58	5.49	11.316	1.708	0.397	1.129	1.708	0.517	0.564	0.443	0.401	0.391	0.254	0.426	0.352
Node216	0.052398	305	29	303	750	144	16	75	85	9.179	0.9	9.119	22.528	4.35	0.51	2.28	2.58	0.262	0.394	0.25	0.179	0.145	0.197	0.222	0.237

evidenced by the contrasting genome sizes and gene counts (n_{genes}) between these three species (Table 2.3.3.), where a genome size minimal difference (Δ_{min}) = 14.476% (5.38 Mbp) and n_{genes} Δ_{min} = 6.561% (n = 874) was observed between *F. graminearum* and *F. verticilloides* (which had a smaller genome and n_{genes} than *F. oxysporum*. These expansions are evidenced by considerably different chromosome numbers (n_{chr}) amongst these species, where *F. graminearum* possesses an n_{chr} of 4, *F. verticilloides* an n_{chr} of 11-12, and *F. oxysporum* an n_{chr} of 15 (Gale *et al*., 2005; Ma *et al*., 2010). The bursts observed during the speciation of *F. oxysporum* directly correspond to the HGT event between an unidentified *Fusarium* and *F. oxysporum* allowing for a broader host range (Ma *et al*., 2010).

2.3.7. Functional overrepresentations in remodelled fungal genes

2.3.7.1: Nested composite genes are likely to be involved in pathogenicty but unlikely to be involved in secondary metabolism

Nested composites were observed to be overrepresented for transmembrane transport process (BP) ontologs (GO:0055085; $P_B \leq 4.38e^{-05}$) (Table 2.3.11), signal transduction (GO:0007615; $P_B = 3.85e^{-03}$), and macromolecular modification (GO:0043412; $P_B = 7.60e^{-03}$). Overrepresentation was also observed for cellular component (CC) ontologs associated with the extracellular region (GO:0005576, GO:0044421; $P_B \leq 8.24e^{-06}$). These ontologies are commonly associated secreted as effectors (Vivek-Ananth *et al*., 2018). *A. fumigatus* genes from families with these overrepresented ontologs (*eg.* acetylxylan esterase (*axe1*; XP_748362.1; enzyme commission number (EC):3.1.1.72), cellobiohydrolase (AFUA_3G01910; XP_748511.1; E.C:3.2.1.91), and endo-1,4-β-xylanase (AFUA_6G13610;

XP_751237.1; E.C:3.2.1.8)) were previously reported to not only be secreted, but to play a role in phytopathogenicity through the degradation of plant cell walls (Rodriguez-Moreno *et al*., 2017; Uhse and Djamei, 2018). Conversely, while secondary metabolites are associated with pathogenic processes (Pusztahelyi *et al*., 2015), nested composite genes were observed to be significantly underrepresented for primary metabolism (GO:0044238; P_B = 0.0139) and secondary metabolism (P_B = $1.36e^{-03}$) (Table 2.3.12). Interestingly, phytopathogenic secretome-associated genes (such as those mentioned above) were commonly associated with primary metabolism (GO:0044238).

Interactions between pathogenic fungi and their hosts are complex, and are engaged in a constant evolutionary arms race (Möller and Stuckenbrock, 2017), where advanced invasion strategies emerged in response to advanced plant defences. Due to these factors, perhaps it is not surprising to observe secretome-related genes to be further remodelled. Genes involved in secondary metabolism are usually large, multidomain, and multifunctional (Pasek *et al*., 2006), and commonly share evolutionary histories with primary metabolic pathways (Smith and Tsai, 2007). Due to these factors, an overrepresentation of secondary metabolic genes was reasonably expected, but this was not the case.

2.3.7.2. Strict composite genes are likely to be involved in pathogenicity and metabolism but not secretion

In direct contrast to what was observed for nested composites, strict composites were statistically likely to be involved in primary metabolism (P_B = $4.59e^{-03}$) (Table 2.3.13) and secondary metabolism (P_B = $4.41e^{-04}$) and statistically unlikely to be involved in signal transduction (P_B = $1.35e^{-03}$) (Table 2.3.14) or transmembrane transport (P_B = $3.48e^{-03}$). *A. fumigatus* genes from families with overrepresented ontologs for primary metabolism (*eg.*

Table 2.3.14: Overrepresented GO-slims in nested composite genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly overrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0032196	BP	transposition	518/215494	537/401320	2	$5.11e^{-06}$	**$1.21e^{-03}$**
GO:0050789	BP	regulation of biological process	4016/215494	5515/401320	2	$1.62e^{-05}$	**$3.85e^{-03}$**
GO:0050794	BP	regulation of cellular process	4016/215494	5515/401320	3	$1.62e^{-05}$	**$3.85e^{-03}$**
GO:0007165	BP	signal transduction	4016/215494	5515/401320	4	$1.62e^{-05}$	**$3.85e^{-03}$**
GO:0065007	BP	biological regulation	5196/215494	7963/401320	1	$1.97e^{-05}$	**$4.67e^{-03}$**
GO:0043412	BP	macromolecule modification	15729/215494	21692/401320	4	$3.21e^{-05}$	**$7.60e^{-03}$**
GO:0036211	BP	protein modification process	15729/215494	21692/401320	5	$3.21e^{-05}$	**7.60e-03**
GO:0006464	BP	cellular protein modification process	15729/215494	21692/401320	6	$3.21e^{-05}$	**7.60e-03**
GO:0055085	BP	transmembrane transport	30576/215494	39724/401320	4	$4.38e^{-05}$	**0.0104**
GO:0051179	BP	localization	35756/215494	53205/401320	1	$4.99e^{-05}$	**0.0118**
GO:0051234	BP	establishment of localization	35756/215494	53205/401320	2	$4.99e^{-05}$	**0.0118**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0006810	BP	transport	35756/215494	53205/401320	3	$4.99e^{-05}$	**0.0118**
GO:0005773	CC	vacuole	51/215494	54/401320	5	$1.57e^{-06}$	$\mathbf{3.71e^{-04}}$
GO:0005615	CC	extracellular space	169/215494	222/401320	2	$2.84e^{-06}$	$\mathbf{6.72e^{-04}}$
GO:0044421	CC	extracellular region part	170/215494	226/401320	1	$3.51e^{-06}$	$\mathbf{8.31e^{-04}}$
GO:0005576	CC	extracellular region	957/215494	1368/401320	1	$8.24e^{-06}$	$\mathbf{1.95e^{-03}}$
GO:0005634	CC	nucleus	12453/215494	19392/401320	5	$3.14e^{-05}$	$\mathbf{7.43e^{-03}}$
GO:0043227	CC	membrane-bounded organelle	12913/215494	22063/401320	2	$3.28e^{-05}$	$\mathbf{7.78e^{-03}}$
GO:0043231	CC	intracellular membrane-bounded organelle	12913/215494	22063/401320	4	$3.28e^{-05}$	$\mathbf{7.78e^{-03}}$
GO:0044430	CC	cytoskeletal part	326/215494	557/401320	4	0.0241	1
GO:0005694	CC	chromosome	607/215494	1051/401320	5	0.00845	1
GO:0005815	CC	microtubule organizing center	326/215494	557/401320	5	0.0241	1
GO:0060090	MF	molecular adaptor activity	93/215494	99/401320	2	$1.36e^{-06}$	$\mathbf{3.23e^{-04}}$
GO:0030674	MF	protein binding, bridging	93/215494	99/401320	3	$1.36e^{-06}$	$\mathbf{3.23e^{-04}}$
GO:0032182	MF	ubiquitin-like protein binding	181/215494	189/401320	3	$2.47e^{-06}$	$\mathbf{5.85e^{-04}}$
GO:0004386	MF	helicase activity	1612/215494	2011/401320	7	$8.92e^{-06}$	$\mathbf{2.11e^{-03}}$
GO:0008289	MF	lipid binding	1159/215494	1961/401320	2	$1.16e^{-05}$	$\mathbf{2.76e^{-03}}$
GO:0008092	MF	cytoskeletal protein binding	1979/215494	2945/401320	3	$1.19e^{-05}$	$\mathbf{2.83e^{-03}}$
GO:0019899	MF	enzyme binding	2471/215494	3537/401320	3	$1.34e^{-05}$	$\mathbf{3.17e^{-03}}$

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0003924	MF	GTPase activity	5127/215494	5881/401320	7	$1.64e^{-05}$	**$3.90e^{-03}$**
GO:0016887	MF	ATPase activity	5203/215494	5630/401320	7	$1.65e^{-05}$	**$3.91e^{-03}$**
GO:0016746	MF	transferase activity, transferring acyl groups	2537/215494	4434/401320	3	$1.69e^{-05}$	**$4.01e^{-03}$**
GO:0005515	MF	protein binding	5050/215494	7948/401320	2	$1.93e^{-05}$	**$4.58e^{-03}$**
GO:0016798	MF	hydrolase activity, acting on glycosyl bonds	5383/215494	9306/401320	3	$2.14e^{-05}$	**$5.08e^{-03}$**
GO:0140110	MF	transcription regulator activity	7626/215494	10273/401320	1	$2.23e^{-05}$	**$5.30e^{-03}$**
GO:0003700	MF	DNA-binding transcription factor activity	7626/215494	10273/401320	2	$2.23e^{-05}$	**$5.30e^{-03}$**
GO:0016817	MF	hydrolase activity, acting on acid anhydrides	11503/215494	13032/401320	3	$2.49e^{-05}$	**$5.91e^{-03}$**
GO:0016818	MF	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	11503/215494	13032/401320	4	$2.49e^{-05}$	**$5.91e^{-03}$**
GO:0016462	MF	pyrophosphatase activity	11503/215494	13032/401320	5	$2.49e^{-05}$	**$5.91e^{-03}$**
GO:0017111	MF	nucleoside-triphosphatase activity	11503/215494	13032/401320	6	$2.49e^{-05}$	**$5.91e^{-03}$**
GO:0016301	MF	kinase activity	12881/215494	16743/401320	4	$2.81e^{-05}$	**$6.67e^{-03}$**
GO:0003677	MF	DNA binding	13048/215494	20257/401320	4	$3.22e^{-05}$	**$7.62e^{-03}$**
GO:0005215	MF	transporter activity	14310/215494	21900/401320	1	$3.27e^{-05}$	**$7.76e^{-03}$**
GO:0022857	MF	transmembrane transporter activity	14310/215494	21900/401320	2	$3.27e^{-05}$	**$7.76e^{-03}$**
GO:0016772	MF	transferase activity, transferring phosphorus-containing groups	14858/215494	21794/401320	3	$3.34e^{-05}$	**$7.91e^{-03}$**
GO:0016740	MF	transferase activity	21955/215494	38813/401320	2	$4.38e^{-05}$	**0.0104**
GO:0016787	MF	hydrolase activity	25514/215494	40915/401320	2	$4.50e^{-05}$	**0.0107**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0016491	MF	oxidoreductase activity	27272/215494	45921/401320	2	$4.65e^{-05}$	**0.011**
GO:0003674	MF	molecular function	180450/215494	325539/401320	0	$5.82e^{-05}$	**0.0138**
GO:0043167	MF	ion binding	77731/215494	107256/401320	2	$6.56e^{-05}$	**0.0155**
GO:0005488	MF	binding	88301/215494	132438/401320	1	$7.12e^{-05}$	**0.0169**
GO:0003824	MF	catalytic activity	81114/215494	141848/401320	1	$7.18e^{-05}$	**0.017**

Table 2.3.15: Underrepresented GO-slims in nested composite genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly underrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0048856	BP	anatomical structure development	3/215494	31/401320	2	$4.48e^{-07}$	**$1.06e^{-04}$**
GO:0051301	BP	cell division	0/215494	87/401320	2	$1.31e^{-06}$	**$3.10e^{-04}$**
GO:0007059	BP	chromosome segregation	44/215494	289/401320	2	$2.73e^{-06}$	**$6.47e^{-04}$**
GO:0000278	BP	mitotic cell cycle	24/215494	303/401320	3	$3.54e^{-06}$	**$8.39e^{-04}$**
GO:0007049	BP	cell cycle	54/215494	352/401320	2	$3.86e^{-06}$	**$9.15e^{-04}$**
GO:0034622	BP	cellular protein-containing complex assembly	4/215494	313/401320	5	$3.89e^{-06}$	**$9.22e^{-04}$**
GO:0022618	BP	ribonucleoprotein complex assembly	4/215494	313/401320	6	$3.89e^{-06}$	**$9.22e^{-04}$**
GO:0071826	BP	ribonucleoprotein complex subunit organization	4/215494	313/401320	4	$3.89e^{-06}$	**$9.22e^{-04}$**
GO:0006914	BP	autophagy	61/215494	478/401320	4	$4.12e^{-06}$	**$9.75e^{-04}$**
GO:0061919	BP	process utilizing autophagic mechanism	61/215494	478/401320	2	$4.12e^{-06}$	**$9.75e^{-04}$**
GO:0051169	BP	nuclear transport	186/215494	517/401320	5	$4.14e^{-06}$	**$9.81e^{-04}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0006913	BP	nucleocytoplasmic transport	186/215494	517/401320	6	$4.14e^{-06}$	**$9.81e^{-04}$**
GO:0007005	BP	mitochondrion organization	8/215494	542/401320	4	$4.62e^{-06}$	**$1.10e^{-03}$**
GO:0051604	BP	protein maturation	47/215494	513/401320	5	$4.72e^{-06}$	**$1.12e^{-03}$**
GO:0061024	BP	membrane organization	88/215494	516/401320	3	$5.02e^{-06}$	**$1.19e^{-03}$**
GO:0007034	BP	vacuolar transport	0/215494	802/401320	4	$5.47e^{-06}$	**$1.30e^{-03}$**
GO:0019748	BP	secondary metabolic process	313/215494	896/401320	2	$5.74e^{-06}$	**$1.36e^{-03}$**
GO:0007010	BP	cytoskeleton organization	316/215494	781/401320	4	$5.80e^{-06}$	**$1.38e^{-03}$**
GO:0044085	BP	cellular component biogenesis	195/215494	914/401320	2	$6.07e^{-06}$	**$1.44e^{-03}$**
GO:0022613	BP	ribonucleoprotein complex biogenesis	195/215494	914/401320	3	$6.07e^{-06}$	**$1.44e^{-03}$**
GO:0042254	BP	ribosome biogenesis	195/215494	914/401320	4	$6.07e^{-06}$	**$1.44e^{-03}$**
GO:0006886	BP	intracellular protein transport	135/215494	919/401320	8	$6.45e^{-06}$	**$1.53e^{-03}$**
GO:0006605	BP	protein targeting	135/215494	919/401320	9	$6.45e^{-06}$	**$1.53e^{-03}$**
GO:0006457	BP	protein folding	30/215494	1039/401320	2	$7.05e^{-06}$	**$1.67e^{-03}$**
GO:0051641	BP	cellular localization	321/215494	1440/401320	2	$7.51e^{-06}$	**$1.78e^{-03}$**
GO:0051649	BP	establishment of localization in cell	321/215494	1440/401320	3	$7.51e^{-06}$	**$1.78e^{-03}$**
GO:0046907	BP	intracellular transport	321/215494	1440/401320	4	$7.51e^{-06}$	**$1.78e^{-03}$**
GO:0044248	BP	cellular catabolic process	470/215494	1301/401320	3	$8.17e^{-06}$	**$1.94e^{-03}$**
GO:0006091	BP	generation of precursor metabolites and energy	88/215494	1417/401320	3	$8.17e^{-06}$	**$1.94e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0051276	BP	chromosome organization	699/215494	1844/401320	4	$9.54e^{-06}$	**$2.26e^{-03}$**
GO:0006790	BP	sulfur compound metabolic process	9/215494	1672/401320	3	$9.54e^{-06}$	**$2.26e^{-03}$**
GO:0032502	BP	developmental process	5/215494	33/401320	1	$1.02e^{-05}$	**$2.41e^{-03}$**
GO:0016071	BP	mRNA metabolic process	852/215494	2039/401320	7	$1.03e^{-05}$	**$2.44e^{-03}$**
GO:0006397	BP	mRNA processing	852/215494	2039/401320	8	$1.03e^{-05}$	**$2.44e^{-03}$**
GO:0006396	BP	RNA processing	852/215494	2039/401320	7	$1.03e^{-05}$	**$2.44e^{-03}$**
GO:0042592	BP	homeostatic process	1180/215494	2448/401320	3	$1.11e^{-05}$	**$2.64e^{-03}$**
GO:0065008	BP	regulation of biological quality	1180/215494	2448/401320	2	$1.11e^{-05}$	**$2.64e^{-03}$**
GO:0065003	BP	protein-containing complex assembly	482/215494	2472/401320	4	$1.12e^{-05}$	**$2.65e^{-03}$**
GO:0043933	BP	protein-containing complex subunit organization	482/215494	2472/401320	3	$1.12e^{-05}$	**$2.65e^{-03}$**
GO:0022607	BP	cellular component assembly	487/215494	2992/401320	3	$1.18e^{-05}$	**$2.79e^{-03}$**
GO:0051186	BP	cofactor metabolic process	310/215494	3046/401320	3	$1.19e^{-05}$	**$2.82e^{-03}$**
GO:0006996	BP	organelle organization	1023/215494	3167/401320	3	$1.26e^{-05}$	**$2.99e^{-03}$**
GO:0016192	BP	vesicle-mediated transport	2097/215494	5466/401320	4	$1.69e^{-05}$	**$4.00e^{-03}$**
GO:0034660	BP	ncRNA metabolic process	2486/215494	5991/401320	7	$1.74e^{-05}$	**$4.13e^{-03}$**
GO:0006399	BP	tRNA metabolic process	2486/215494	5991/401320	8	$1.74e^{-05}$	**$4.13e^{-03}$**
GO:0050896	BP	response to stimulus	2994/215494	6331/401320	1	$1.75e^{-05}$	**$4.15e^{-03}$**
GO:0006950	BP	response to stress	2994/215494	6331/401320	2	$1.75e^{-05}$	**$4.15e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016043	BP	cellular component organization	1592/215494	6188/401320	2	$1.77e^{-05}$	**$4.20e^{-03}$**
GO:0042886	BP	amide transport	2600/215494	5930/401320	5	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0045184	BP	establishment of protein localization	2600/215494	5930/401320	4	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0033036	BP	macromolecule localization	2600/215494	5930/401320	2	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0071705	BP	nitrogen compound transport	2600/215494	5930/401320	4	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0071702	BP	organic substance transport	2600/215494	5930/401320	4	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0015833	BP	peptide transport	2600/215494	5930/401320	6	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0008104	BP	protein localization	2600/215494	5930/401320	3	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0015031	BP	protein transport	2600/215494	5930/401320	7	$1.78e^{-05}$	**$4.22e^{-03}$**
GO:0019752	BP	carboxylic acid metabolic process	2550/215494	7482/401320	5	$1.84e^{-05}$	**$4.35e^{-03}$**
GO:0006520	BP	cellular amino acid metabolic process	2550/215494	7482/401320	6	$1.84e^{-05}$	**$4.35e^{-03}$**
GO:0006082	BP	organic acid metabolic process	2550/215494	7482/401320	3	$1.84e^{-05}$	**$4.35e^{-03}$**
GO:0043436	BP	oxoacid metabolic process	2550/215494	7482/401320	4	$1.84e^{-05}$	**$4.35e^{-03}$**
GO:0009056	BP	catabolic process	2171/215494	6927/401320	2	$1.88e^{-05}$	**$4.45e^{-03}$**
GO:0071840	BP	cellular component organization or biogenesis	1787/215494	7102/401320	1	$1.91e^{-05}$	**$4.54e^{-03}$**
GO:0016070	BP	RNA metabolic process	3338/215494	8030/401320	6	$1.92e^{-05}$	**$4.56e^{-03}$**
GO:0006259	BP	DNA metabolic process	3968/215494	7926/401320	6	$1.98e^{-05}$	**$4.70e^{-03}$**
GO:0006629	BP	lipid metabolic process	3440/215494	9168/401320	3	$2.15e^{-05}$	**$5.09e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0043604	BP	amide biosynthetic process	276/215494	9985/401320	5	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0043603	BP	cellular amide metabolic process	276/215494	9985/401320	4	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0044249	BP	cellular biosynthetic process	276/215494	9985/401320	3	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0034645	BP	cellular macromolecule biosynthetic process	276/215494	9985/401320	5	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0044271	BP	cellular nitrogen compound biosynthetic process	276/215494	9985/401320	4	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0009059	BP	macromolecule biosynthetic process	276/215494	9985/401320	4	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:1901576	BP	organic substance biosynthetic process	276/215494	9985/401320	3	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:1901566	BP	organonitrogen compound biosynthetic process	276/215494	9985/401320	4	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0043043	BP	peptide biosynthetic process	276/215494	9985/401320	6	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0006518	BP	peptide metabolic process	276/215494	9985/401320	5	$2.15e^{-05}$	**$5.10e^{-03}$**
GO:0006412	BP	translation	276/215494	9985/401320	7	$2^{.15e-05}$	**$5.10e^{-03}$**
GO:0090304	BP	nucleic acid metabolic process	7306/215494	15956/401320	5	$2.81e^{-05}$	**$6.66e^{-03}$**
GO:0005975	BP	carbohydrate metabolic process	8213/215494	16166/401320	3	$2.83e^{-05}$	**$6.71e^{-03}$**
GO:0006725	BP	cellular aromatic compound metabolic process	7591/215494	16648/401320	3	$2.87e^{-05}$	**$6.80e^{-03}$**
GO:0046483	BP	heterocycle metabolic process	7591/215494	16648/401320	3	$2.87e^{-05}$	**$6.80e^{-03}$**
GO:0006139	BP	nucleobase-containing compound metabolic process	7591/215494	16648/401320	4	$2.87e^{-05}$	**$6.80e^{-03}$**
GO:1901360	BP	organic cyclic compound metabolic process	7591/215494	16648/401320	3	$2.87e^{-05}$	**$6.80e^{-03}$**
GO:0044281	BP	small molecule metabolic process	5238/215494	17822/401320	2	$2.91e^{-05}$	**$6.90e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0044267	BP	cellular protein metabolic process	16003/215494	31675/401320	5	$3.86e^{-05}$	**$9.16e^{-03}$**
GO:0019538	BP	protein metabolic process	16017/215494	31990/401320	4	$3.98e^{-05}$	**$9.44e^{-03}$**
GO:1901564	BP	organonitrogen compound metabolic process	18566/215494	39471/401320	3	$4.36e^{-05}$	**0.0103**
GO:0044260	BP	cellular macromolecule metabolic process	19971/215494	39601/401320	4	$4.37e^{-05}$	**0.0104**
GO:0009058	BP	biosynthetic process	13170/215494	42776/401320	2	$4.54e^{-05}$	**0.0108**
GO:0043170	BP	macromolecule metabolic process	23211/215494	47834/401320	3	$4.71e^{-05}$	**0.0112**
GO:0034641	BP	cellular nitrogen compound metabolic process	18407/215494	52479/401320	3	$4.98e^{-05}$	**0.0118**
GO:0006807	BP	nitrogen compound metabolic process	34511/215494	76971/401320	2	$5.81e^{-05}$	**0.0138**
GO:0044237	BP	cellular metabolic process	34484/215494	79144/401320	2	$5.87e^{-05}$	**0.0139**
GO:0071704	BP	organic substance metabolic process	35368/215494	76658/401320	2	$5.85e^{-05}$	**0.0139**
GO:0044238	BP	primary metabolic process	35368/215494	76658/401320	2	$5.85e^{-05}$	**0.0139**
GO:0009987	BP	cellular process	39538/215494	90607/401320	1	$6.19e^{-05}$	**0.0147**
GO:0008152	BP	metabolic process	49860/215494	112739/401320	1	$6.70e^{-05}$	**0.0159**
GO:0007009	BP	plasma membrane organization	0/215494	11/401320	4	$2.10e^{-04}$	**0.0497**
GO:0000229	CC	cytoplasmic chromosome	6/215494	107/401320	6	$2.28e^{-06}$	**$5.40e^{-04}$**
GO:0005730	CC	nucleolus	200/215494	497/401320	5	$4.10e^{-06}$	**$9.73e^{-04}$**
GO:0042579	CC	microbody	201/215494	480/401320	5	$4.31e^{-06}$	**$1.02e^{-03}$**
GO:0005777	CC	peroxisome	201/215494	480/401320	6	$4.31e^{-06}$	**$1.02e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016020	CC	membrane	91/215494	502/401320	1	$4.59e^{-06}$	**$1.09e^{-03}$**
GO:0005886	CC	plasma membrane	91/215494	502/401320	2	$4.59e^{-06}$	**$1.09e^{-03}$**
GO:0005856	CC	cytoskeleton	10/215494	450/401320	5	$5.19e^{-06}$	**$1.23e^{-03}$**
GO:0005739	CC	mitochondrion	115/215494	1210/401320	5	$6.49e^{-06}$	**$1.54e^{-03}$**
GO:0005783	CC	endoplasmic reticulum	93/215494	927/401320	5	$7.06e^{-06}$	**$1.67e^{-03}$**
GO:0005737	CC	cytoplasm	1216/215494	4855/401320	3	$1.49e^{-05}$	**$3.53e^{-03}$**
GO:0005622	CC	intracellular	899/215494	7889/401320	2	$1.90e^{-05}$	**$4.50e^{-03}$**
GO:1990904	CC	ribonucleoprotein complex	201/215494	9786/401320	2	$2.15e^{-05}$	**$5.08e^{-03}$**
GO:0005840	CC	ribosome	201/215494	9786/401320	5	$2.15e^{-05}$	**$5.08e^{-03}$**
GO:0043232	CC	intracellular non-membrane-bounded organelle	1017/215494	11783/401320	4	$2.35e^{-05}$	**$5.56e^{-03}$**
GO:0043228	CC	non-membrane-bounded organelle	1017/215494	11783/401320	2	$2.35e^{-05}$	**$5.56e^{-03}$**
GO:0044444	CC	cytoplasmic part	667/215494	12564/401320	3	$2.47e^{-05}$	**$5.85e^{-03}$**
GO:0044428	CC	nuclear part	372/215494	803/401320	4	$3.31e^{-05}$	**$7.83e^{-03}$**
GO:0032991	CC	protein-containing complex	5360/215494	34280/401320	1	$4.03e^{-05}$	**$9.54e^{-03}$**
GO:0043229	CC	intracellular organelle	13927/215494	33803/401320	3	$4.07e^{-05}$	**$9.64e^{-03}$**
GO:0043226	CC	organelle	14009/215494	34799/401320	1	$4.12e^{-05}$	**$9.77e^{-03}$**
GO:0044424	CC	intracellular part	15468/215494	39026/401320	2	$4.27e^{-05}$	**0.0101**
GO:0044464	CC	cell part	16807/215494	42555/401320	1	$4.46e^{-05}$	**0.0106**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0005575	CC	cellular_component	59207/215494	122687/401320	0	$6.93e^{-05}$	**0.0164**
GO:0042393	MF	histone binding	55/215494	164/401320	3	$3.02e^{-06}$	$\mathbf{7.16e^{-04}}$
GO:0003729	MF	mRNA binding	21/215494	311/401320	5	$3.63e^{-06}$	$\mathbf{8.60e^{-04}}$
GO:0008134	MF	transcription factor binding	101/215494	317/401320	3	$4.47e^{-06}$	$\mathbf{1.06e^{-03}}$
GO:0019843	MF	rRNA binding	5/215494	616/401320	5	$5.61e^{-06}$	$\mathbf{1.33e^{-03}}$
GO:0051082	MF	unfolded protein binding	444/215494	1067/401320	3	$6.98e^{-06}$	$\mathbf{1.66e^{-03}}$
GO:0016765	MF	transferase activity, transferring alkyl or aryl (other than methyl) groups	369/215494	1381/401320	3	$8.03e^{-06}$	$\mathbf{1.90e^{-03}}$
GO:0016810	MF	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	643/215494	1820/401320	3	$8.34e^{-06}$	$\mathbf{1.98e^{-03}}$
GO:0016791	MF	phosphatase activity	532/215494	1699/401320	5	$8.99e^{-06}$	$\mathbf{2.13e^{-03}}$
GO:0042578	MF	phosphoric ester hydrolase activity	532/215494	1699/401320	4	$8.99e^{-06}$	$\mathbf{2.13e^{-03}}$
GO:0008135	MF	translation factor activity, RNA binding	164/215494	2158/401320	5	$9.94e^{-06}$	$\mathbf{2.36e^{-03}}$
GO:0045182	MF	translation regulator activity	164/215494	2158/401320	1	$9.94e^{-06}$	$\mathbf{2.36e^{-03}}$
GO:0090079	MF	translation regulator activity, nucleic acid binding	164/215494	2158/401320	4	$9.94e^{-06}$	$\mathbf{2.36e^{-03}}$
GO:0030234	MF	enzyme regulator activity	195/215494	2649/401320	2	$1.08e^{-05}$	$\mathbf{2.56e^{-03}}$
GO:0098772	MF	molecular function regulator	195/215494	2649/401320	1	$1.08e^{-05}$	$\mathbf{2.56e^{-03}}$
GO:0016853	MF	isomerase activity	1715/215494	4244/401320	2	$1.46e^{-05}$	$\mathbf{3.47e^{-03}}$
GO:0016779	MF	nucleotidyltransferase activity	1977/215494	5051/401320	4	$1.48e^{-05}$	$\mathbf{3.50e^{-03}}$
GO:0008168	MF	methyltransferase activity	1926/215494	5544/401320	4	$1.61e^{-05}$	$\mathbf{3.81e^{-03}}$

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016741	MF	transferase activity, transferring one-carbon groups	1926/215494	5544/401320	3	$1.61e^{-05}$	**$3.81e^{-03}$**
GO:0016788	MF	hydrolase activity, acting on ester bonds	2546/215494	5627/401320	3	$1.64e^{-05}$	**$3.89e^{-03}$**
GO:0016757	MF	transferase activity, transferring glycosyl groups	2272/215494	5667/401320	3	$1.65e^{-05}$	**$3.92e^{-03}$**
GO:0016874	MF	ligase activity	3199/215494	6591/401320	2	$1.73e^{-05}$	**$4.09e^{-03}$**
GO:0016829	MF	lyase activity	2042/215494	6527/401320	2	$1.76e^{-05}$	**$4.18e^{-03}$**
GO:0003723	MF	RNA binding	2404/215494	9993/401320	4	$2.17e^{-05}$	**$5.14e^{-03}$**
GO:0003735	MF	structural constituent of ribosome	207/215494	10102/401320	2	$2.19e^{-05}$	**$5.19e^{-03}$**
GO:0140096	MF	catalytic activity, acting on a protein	5649/215494	11340/401320	2	$2.31e^{-05}$	**$5.49e^{-03}$**
GO:0008233	MF	peptidase activity	5649/215494	11340/401320	3	$2.31e^{-05}$	**$5.49e^{-03}$**
GO:0005198	MF	structural molecule activity	960/215494	11671/401320	1	$2.44e^{-05}$	**$5.78e^{-03}$**
GO:1901363	MF	heterocyclic compound binding	15446/215494	30244/401320	2	$3.80e^{-05}$	**$8.99e^{-03}$**
GO:0003676	MF	nucleic acid binding	15446/215494	30244/401320	3	$3.80e^{-05}$	**$8.99e^{-03}$**
GO:0097159	MF	organic cyclic compound binding	15446/215494	30244/401320	2	$3.80e^{-05}$	**$8.99e^{-03}$**

malate synthase (*acuE*; XP_747723.1; E.C:2.3.3.9), α,α-trehalose phosphatase subunit Tps2 (*tps2*; XP_755036.1; E.C: 3.1.3.12), and chorismate mutase/prephenate dehydratase (AFUA_5G05690; XP_754063.1; E.C:5.4.99.5)) were, again, involved in pathogenicity (Olivas *et al*., 2008; Al-Bader *et al.*, 2010; Pérez *et al*., 2015; Xiaowei *et al*., 2019). Chorismate synthase is a constituent of the highly remodelled shikimate pathway (Richards *et al.*, 2006). These observations further highlight the evolutionary arms-race between pathogenic fungi and their hosts as mentioned above, and their prevalence amongst remodelled genes is not surprising.

2.3.7.3. Strict component genes are likely involved in mitosis and DNA repair

Strict component genes were observed to be significantly likely to be involved in stress responses (GO:0006950; $P_B = 3.35e^{-03}$) (Table 2.3.15) and in the mitotic cell cycle (GO:0000278, GO:0007059, GO:0051301; $P_B \leq 6.0e^{-03}$) and unlikely to be associated with regulatory processes (GO:0050789, GO:0050794, GO:0065007, GO:0140110; $P_B \leq 4.29e^{-03}$) (Table 2.3.16). Strict component *A. fumigatus* genes from families with overrepresented ontologs for stress response (*eg.* endonuclease III homolog (*ntg1*; XP_749248.1; E.C:4.2.99.15), mitochondrial genome maintenance protein mgm101 (AFUA_2G09560; XP_755290.1; (no assigned EC number)), and UV-endonuclease (*uve1*; XP_750978.1; E.C:3.-.-.-)) commonly facilitate the maintenance of nuclear and organelle genome integrity after exposure to physical stressors (Heijink *et al*., 2013; Verna and Idnrum, 2013; Bakhoum *et al*., 2017). Comparatively, the strict component *A. fumigatus* gene associated with mitosis (*eg.* CHL4 family chromosome segregation protein (AFUA_4G06540; XP_752172.1; (no assigned EC number))) aids in *de novo* kinetochore assembly and sister chromatid adhesion during replication (Mythreye and Bloom, 2003).

Table 2.3.16. Overrepresented GO-slims in strict composite genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly overrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0061919	BP	process utilizing autophagic mechanism	93/13168	478/401320	2	$8.67e^{-07}$	**$2.05e^{-04}$**
GO:0006914	BP	autophagy	93/13168	478/401320	4	$8.67e^{-07}$	**$2.05e^{-04}$**
GO:0051301	BP	cell division	16/13168	87/401320	2	$9.44e^{-07}$	**$2.24e^{-04}$**
GO:0007005	BP	mitochondrion organization	60/13168	542/401320	4	$1.28e^{-06}$	**$3.04e^{-04}$**
GO:0051169	BP	nuclear transport	100/13168	517/401320	5	$1.37e^{-06}$	**$3.24e^{-04}$**
GO:0006913	BP	nucleocytoplasmic transport	100/13168	517/401320	6	$1.37e^{-06}$	**$3.24e^{-04}$**
GO:0007059	BP	chromosome segregation	34/13168	289/401320	2	$1.44e^{-06}$	**$3.40e^{-04}$**
GO:0019748	BP	secondary metabolic process	249/13168	896/401320	2	$1.86e^{-06}$	**$4.41e^{-04}$**
GO:0044248	BP	cellular catabolic process	98/13168	1301/401320	3	$2.34e^{-06}$	**$5.54e^{-04}$**
GO:0046907	BP	intracellular transport	144/13168	1440/401320	4	$2.50e^{-06}$	**$5.94e^{-04}$**
GO:0051649	BP	establishment of localization in cell	144/13168	1440/401320	3	$2.50e^{-06}$	**$5.94e^{-04}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0051641	BP	cellular localization	144/13168	1440/401320	2	$2.50e^{-06}$	**$5.94e^{-04}$**
GO:0006457	BP	protein folding	74/13168	1039/401320	2	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0006091	BP	generation of precursor metabolites and energy	106/13168	1417/401320	3	$2.96e^{-06}$	**$7.00e^{-04}$**
GO:0006996	BP	organelle organization	161/13168	3167/401320	3	$3.77e^{-06}$	**$8.95e^{-04}$**
GO:0065003	BP	protein-containing complex assembly	150/13168	2472/401320	4	$4.03e^{-06}$	**$9.55e^{-04}$**
GO:0043933	BP	protein-containing complex subunit organization	150/13168	2472/401320	3	$4.03e^{-06}$	**$9.55e^{-04}$**
GO:0022607	BP	cellular component assembly	150/13168	2992/401320	3	$4.16e^{-06}$	**$9.87e^{-04}$**
GO:0016192	BP	vesicle-mediated transport	311/13168	5466/401320	4	$5.14e^{-06}$	**$1.22e^{-03}$**
GO:0016043	BP	cellular component organization	293/13168	6188/401320	2	$5.65e^{-06}$	**$1.34e^{-03}$**
GO:0034660	BP	ncRNA metabolic process	381/13168	5991/401320	7	$5.66e^{-06}$	**$1.34e^{-03}$**
GO:0006399	BP	tRNA metabolic process	381/13168	5991/401320	8	$5.66e^{-06}$	**$1.34e^{-03}$**
GO:0016070	BP	RNA metabolic process	409/13168	8030/401320	6	$5.87e^{-06}$	**$1.39e^{-03}$**
GO:0045184	BP	establishment of protein localization	316/13168	5930/401320	4	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0015833	BP	peptide transport	316/13168	5930/401320	6	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0071705	BP	nitrogen compound transport	316/13168	5930/401320	4	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0071702	BP	organic substance transport	316/13168	5930/401320	4	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0008104	BP	protein localization	316/13168	5930/401320	3	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0033036	BP	macromolecule localization	316/13168	5930/401320	2	$6.04e^{-06}$	**$1.43e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0042886	BP	amide transport	316/13168	5930/401320	5	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0015031	BP	protein transport	316/13168	5930/401320	7	$6.04e^{-06}$	**$1.43e^{-03}$**
GO:0043436	BP	oxoacid metabolic process	710/13168	7482/401320	4	$6.97e^{-06}$	**$1.65e^{-03}$**
GO:0006520	BP	cellular amino acid metabolic process	710/13168	7482/401320	6	$6.97e^{-06}$	**$1.65e^{-03}$**
GO:0006082	BP	organic acid metabolic process	710/13168	7482/401320	3	$6.97e^{-06}$	**$1.65e^{-03}$**
GO:0019752	BP	carboxylic acid metabolic process	710/13168	7482/401320	5	$6.97e^{-06}$	**$1.65e^{-03}$**
GO:0006629	BP	lipid metabolic process	641/13168	9168/401320	3	$7.38e^{-06}$	**$1.75e^{-03}$**
GO:0009056	BP	catabolic process	299/13168	6927/401320	2	$9.47e^{-06}$	**$2.24e^{-03}$**
GO:0005975	BP	carbohydrate metabolic process	670/13168	16166/401320	3	$9.90e^{-06}$	**$2.35e^{-03}$**
GO:0044281	BP	small molecule metabolic process	1145/13168	17822/401320	2	$1.02e^{-05}$	**$2.43e^{-03}$**
GO:0009058	BP	biosynthetic process	1933/13168	42776/401320	2	$1.44e^{-05}$	**$3.42e^{-03}$**
GO:0071704	BP	organic substance metabolic process	2887/13168	76658/401320	2	$1.94e^{-05}$	**$4.59e^{-03}$**
GO:0044238	BP	primary metabolic process	2887/13168	76658/401320	2	$1.94e^{-05}$	**$4.59e^{-03}$**
GO:0008152	BP	metabolic process	4406/13168	112739/401320	1	$2.34e^{-05}$	**$5.55e^{-03}$**
GO:0008150	BP	biological_process	8532/13168	247920/401320	0	$2.42e^{-05}$	**$5.74e^{-03}$**
GO:0007034	BP	vacuolar transport	48/13168	802/401320	4	$9.15e^{-05}$	**0.0217**
GO:0071840	BP	cellular component organization or biogenesis	293/13168	7102/401320	1	$1.09e^{-04}$	**0.0259**
GO:0007010	BP	cytoskeleton organization	47/13168	781/401320	4	$1.09e^{-04}$	**0.0259**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0006605	BP	protein targeting	44/13168	919/401320	9	0.0154	1
GO:0006886	BP	intracellular protein transport	44/13168	919/401320	8	0.0154	1
GO:0044428	CC	nuclear part	57/13168	803/401320	4	$1.26e^{-06}$	**$2.99e^{-04}$**
GO:0000228	CC	nuclear chromosome	54/13168	306/401320	6	$1.40e^{-06}$	**$3.31e^{-04}$**
GO:0016020	CC	membrane	57/13168	502/401320	1	$1.91e^{-06}$	**$4.52e^{-04}$**
GO:0005886	CC	plasma membrane	57/13168	502/401320	2	$1.91e^{-06}$	**$4.52e^{-04}$**
GO:0005694	CC	chromosome	54/13168	1051/401320	5	$1.68e^{-03}$	**0.398**
GO:0044422	CC	organelle part	59/13168	1360/401320	1	0.0325	1
GO:0044446	CC	intracellular organelle part	59/13168	1360/401320	3	0.0325	1
GO:0051082	MF	unfolded protein binding	74/13168	1067/401320	3	$1.65e^{-06}$	**$3.90e^{-04}$**
GO:0016810	MF	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	257/13168	1820/401320	3	$3.35e^{-06}$	**$7.95e^{-04}$**
GO:0016779	MF	nucleotidyltransferase activity	337/13168	5051/401320	4	$5.59e^{-06}$	**$1.32e^{-03}$**
GO:0016829	MF	lyase activity	363/13168	6527/401320	2	$6.00e^{-06}$	**$1.42e^{-03}$**
GO:0016757	MF	transferase activity, transferring glycosyl groups	343/13168	5667/401320	3	$6.32e^{-06}$	**$1.50e^{-03}$**
GO:0016874	MF	ligase activity	496/13168	6591/401320	2	$6.64e^{-06}$	**$1.57e^{-03}$**
GO:0003723	MF	RNA binding	471/13168	9993/401320	4	$6.87e^{-06}$	**$1.63e^{-03}$**
GO:0008233	MF	peptidase activity	608/13168	11340/401320	3	$7.87e^{-06}$	**$1.87e^{-03}$**
GO:0140096	MF	catalytic activity, acting on a protein	608/13168	11340/401320	2	$7.87e^{-06}$	**$1.87e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016740	MF	transferase activity	1448/13168	38813/401320	2	$1.57e^{-05}$	**$3.72e^{-03}$**
GO:0008168	MF	methyltransferase activity	240/13168	5544/401320	4	$2.90e^{-05}$	**$6.86e^{-03}$**
GO:0016741	MF	transferase activity, transferring one-carbon groups	240/13168	5544/401320	3	$2.90e^{-05}$	**$6.86e^{-03}$**
GO:0045182	MF	translation regulator activity	103/13168	2158/401320	1	$2.12e^{-04}$	0.0502
GO:0090079	MF	translation regulator activity, nucleic acid binding	103/13168	2158/401320	4	$2.12e^{-04}$	0.0502
GO:0008135	MF	translation factor activity, RNA binding	103/13168	2158/401320	5	$2.12e^{-04}$	0.0502
GO:0016746	MF	transferase activity, transferring acyl groups	181/13168	4434/401320	3	$3.42e^{-03}$	0.81
GO:0016765	MF	transferase activity, transferring alkyl or aryl (other than methyl) groups	63/13168	1381/401320	3	$9.94e^{-03}$	1

Table 2.3.17. Underrepresented GO-slims in strict composite genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly underrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0032196	BP	transposition	0/13168	537/401320	2	$1.12e^{-06}$	**$2.64e^{-04}$**
GO:0042254	BP	ribosome biogenesis	0/13168	914/401320	4	$1.50e^{-06}$	**$3.55e^{-04}$**
GO:0022613	BP	ribonucleoprotein complex biogenesis	0/13168	914/401320	3	$1.50e^{-06}$	**$3.55e^{-04}$**
GO:0044085	BP	cellular component biogenesis	0/13168	914/401320	2	$1.50e^{-06}$	**$3.55e^{-04}$**
GO:1901361	BP	organic cyclic compound catabolic process	5/13168	823/401320	4	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0046700	BP	heterocycle catabolic process	5/13168	823/401320	4	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0044270	BP	cellular nitrogen compound catabolic process	5/13168	823/401320	4	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0034655	BP	nucleobase-containing compound catabolic process	5/13168	823/401320	5	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0019439	BP	aromatic compound catabolic process	5/13168	823/401320	4	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:1901575	BP	organic substance catabolic process	5/13168	823/401320	3	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0006790	BP	sulfur compound metabolic process	3/13168	1672/401320	3	$3.46e^{-06}$	**$8.20e^{-04}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016071	BP	mRNA metabolic process	28/13168	2039/401320	7	$3.53e^{-06}$	**$8.37e^{-04}$**
GO:0006396	BP	RNA processing	28/13168	2039/401320	7	$3.53e^{-06}$	**$8.37e^{-04}$**
GO:0006397	BP	mRNA processing	28/13168	2039/401320	8	$3.53e^{-06}$	**$8.37e^{-04}$**
GO:0042592	BP	homeostatic process	13/13168	2448/401320	3	$3.62e^{-06}$	**$8.59e^{-04}$**
GO:0065008	BP	regulation of biological quality	13/13168	2448/401320	2	$3.62e^{-06}$	**$8.59e^{-04}$**
GO:0050789	BP	regulation of biological process	58/13168	5515/401320	2	$5.71e^{-06}$	**$1.35e^{-03}$**
GO:0007165	BP	signal transduction	58/13168	5515/401320	4	$5.71e^{-06}$	**$1.35e^{-03}$**
GO:0050794	BP	regulation of cellular process	58/13168	5515/401320	3	$5.71e^{-06}$	**$1.35e^{-03}$**
GO:0006950	BP	response to stress	133/13168	6331/401320	2	$6.11e^{-06}$	**$1.45e^{-03}$**
GO:0050896	BP	response to stimulus	133/13168	6331/401320	1	$6.11e^{-06}$	**$1.45e^{-03}$**
GO:0006259	BP	DNA metabolic process	134/13168	7926/401320	6	$6.94e^{-06}$	**$1.65e^{-03}$**
GO:0065007	BP	biological regulation	71/13168	7963/401320	1	$7.26e^{-06}$	**$1.72e^{-03}$**
GO:1901566	BP	organonitrogen compound biosynthetic process	157/13168	9985/401320	4	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0044271	BP	cellular nitrogen compound biosynthetic process	157/13168	9985/401320	4	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0009059	BP	macromolecule biosynthetic process	157/13168	9985/401320	4	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0006412	BP	translation	157/13168	9985/401320	7	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0043043	BP	peptide biosynthetic process	157/13168	9985/401320	6	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0006518	BP	peptide metabolic process	157/13168	9985/401320	5	$7.79e^{-06}$	**$1.85e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0044249	BP	cellular biosynthetic process	157/13168	9985/401320	3	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0034645	BP	cellular macromolecule biosynthetic process	157/13168	9985/401320	5	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0043604	BP	amide biosynthetic process	157/13168	9985/401320	5	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0043603	BP	cellular amide metabolic process	157/13168	9985/401320	4	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:1901576	BP	organic substance biosynthetic process	157/13168	9985/401320	3	$7.79e^{-06}$	**$1.85e^{-03}$**
GO:0061024	BP	membrane organization	2/13168	516/401320	3	$1.02e^{-05}$	**$2.43e^{-03}$**
GO:0043412	BP	macromolecule modification	447/13168	21692/401320	4	$1.03e^{-05}$	**$2.44e^{-03}$**
GO:0006464	BP	cellular protein modification process	447/13168	21692/401320	6	$1.03e^{-05}$	**$2.44e^{-03}$**
GO:0036211	BP	protein modification process	447/13168	21692/401320	5	$1.03e^{-05}$	**$2.44e^{-03}$**
GO:0044267	BP	cellular protein metabolic process	604/13168	31675/401320	5	$1.33e^{-05}$	**$3.14e^{-03}$**
GO:0019538	BP	protein metabolic process	607/13168	31990/401320	4	$1.34e^{-05}$	**$3.19e^{-03}$**
GO:0044260	BP	cellular macromolecule metabolic process	738/13168	39601/401320	4	$1.43e^{-05}$	**$3.39e^{-03}$**
GO:0055085	BP	transmembrane transport	1011/13168	39724/401320	4	$1.47e^{-05}$	**$3.48e^{-03}$**
GO:0043170	BP	macromolecule metabolic process	1150/13168	47834/401320	3	$1.55e^{-05}$	**$3.66e^{-03}$**
GO:0034641	BP	cellular nitrogen compound metabolic process	1456/13168	52479/401320	3	$1.72e^{-05}$	**$4.08e^{-03}$**
GO:0006807	BP	nitrogen compound metabolic process	2242/13168	76971/401320	2	$1.99e^{-05}$	**$4.72e^{-03}$**
GO:0022618	BP	ribonucleoprotein complex assembly	0/13168	313/401320	6	$4.72e^{-05}$	**0.0112**
GO:0034622	BP	cellular protein-containing complex assembly	0/13168	313/401320	5	$4.72e^{-05}$	**0.0112**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0071826	BP	ribonucleoprotein complex subunit organization	0/13168	313/401320	4	$4.72e^{-05}$	**0.0112**
GO:0000278	BP	mitotic cell cycle	0/13168	303/401320	3	$6.96e^{-05}$	**0.0165**
GO:0071941	BP	nitrogen cycle metabolic process	1/13168	386/401320	3	$7.58e^{-05}$	**0.018**
GO:0051604	BP	protein maturation	3/13168	513/401320	5	$8.91e^{-05}$	**0.0211**
GO:0071554	BP	cell wall organization or biogenesis	0/13168	250/401320	2	$4.98e^{-04}$	0.118
GO:0044237	BP	cellular metabolic process	2478/13168	79144/401320	2	$8.04e^{-03}$	1
GO:0005777	CC	peroxisome	0/13168	480/401320	6	$9.39e^{-07}$	**$2.23e^{-04}$**
GO:0042579	CC	microbody	0/13168	480/401320	5	$9.39e^{-07}$	**$2.23e^{-04}$**
GO:0030312	CC	external encapsulating structure	0/13168	902/401320	2	$2.06e^{-06}$	**$4.88e^{-04}$**
GO:0005618	CC	cell wall	0/13168	902/401320	3	$2.06e^{-06}$	**$4.88e^{-04}$**
GO:0005783	CC	endoplasmic reticulum	0/13168	927/401320	5	$2.84e^{-06}$	**$6.73e^{-04}$**
GO:0005576	CC	extracellular region	7/13168	1368/401320	1	$2.86e^{-06}$	**$6.78e^{-04}$**
GO:0044430	CC	cytoskeletal part	2/13168	557/401320	4	$3.67e^{-06}$	**$8.70e^{-04}$**
GO:0005815	CC	microtubule organizing center	2/13168	557/401320	5	$3.67e^{-06}$	**$8.70e^{-04}$**
GO:0005622	CC	intracellular	125/13168	7889/401320	2	$6.64e^{-06}$	**$1.57e^{-03}$**
GO:1990904	CC	ribonucleoprotein complex	157/13168	9786/401320	2	$7.24e^{-06}$	**$1.72e^{-03}$**
GO:0005840	CC	ribosome	157/13168	9786/401320	5	$7.24e^{-06}$	**$1.72e^{-03}$**
GO:0043228	CC	non-membrane-bounded organelle	215/13168	11783/401320	2	$8.00e^{-06}$	**$1.90e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0043232	CC	intracellular non-membrane-bounded organelle	215/13168	11783/401320	4	$8.00e^{-06}$	**$1.90e^{-03}$**
GO:0044444	CC	cytoplasmic part	193/13168	12564/401320	3	$8.35e^{-06}$	**$1.98e^{-03}$**
GO:0005634	CC	nucleus	421/13168	19392/401320	5	$9.69e^{-06}$	**$2.30e^{-03}$**
GO:0043227	CC	membrane-bounded organelle	457/13168	22063/401320	2	$1.11e^{-05}$	**$2.63e^{-03}$**
GO:0043231	CC	intracellular membrane-bounded organelle	457/13168	22063/401320	4	$1.11e^{-05}$	**$2.63e^{-03}$**
GO:0005856	CC	cytoskeleton	1/13168	450/401320	5	$1.31e^{-05}$	**$3.10e^{-03}$**
GO:0043229	CC	intracellular organelle	672/13168	33803/401320	3	$1.32e^{-05}$	**$3.13e^{-03}$**
GO:0043226	CC	organelle	708/13168	34799/401320	1	$1.38e^{-05}$	**$3.28e^{-03}$**
GO:0032991	CC	protein-containing complex	801/13168	34280/401320	1	$1.40e^{-05}$	**$3.31e^{-03}$**
GO:0044424	CC	intracellular part	815/13168	39026/401320	2	$1.44e^{-05}$	**$3.41e^{-03}$**
GO:0044464	CC	cell part	875/13168	42555/401320	1	$1.57e^{-05}$	**$3.73e^{-03}$**
GO:0005575	CC	cellular_component	3467/13168	122687/401320	0	$2.32e^{-05}$	**$5.49e^{-03}$**
GO:0005730	CC	nucleolus	3/13168	497/401320	5	$1.21e^{-04}$	**0.0286**
GO:0019843	MF	rRNA binding	0/13168	616/401320	5	$1.09e^{-06}$	**$2.58e^{-04}$**
GO:0004386	MF	helicase activity	0/13168	2011/401320	7	$3.06e^{-06}$	**$7.26e^{-04}$**
GO:0008289	MF	lipid binding	20/13168	1961/401320	2	$3.07e^{-06}$	**$7.28e^{-04}$**
GO:0004518	MF	nuclease activity	3/13168	3928/401320	4	$4.14e^{-06}$	**$9.81e^{-04}$**
GO:0008092	MF	cytoskeletal protein binding	52/13168	2945/401320	3	$4.49e^{-06}$	**$1.06e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016853	MF	isomerase activity	49/13168	4244/401320	2	$5.09e^{-06}$	**$1.21e^{-03}$**
GO:0003924	MF	GTPase activity	11/13168	5881/401320	7	$5.55e^{-06}$	**$1.32e^{-03}$**
GO:0016788	MF	hydrolase activity, acting on ester bonds	44/13168	5627/401320	3	$5.87e^{-06}$	**$1.39e^{-03}$**
GO:0016887	MF	ATPase activity	0/13168	5630/401320	7	$5.89e^{-06}$	**$1.40e^{-03}$**
GO:0005515	MF	protein binding	168/13168	7948/401320	2	$6.12e^{-06}$	**$1.45e^{-03}$**
GO:0016798	MF	hydrolase activity, acting on glycosyl bonds	164/13168	9306/401320	3	$6.49e^{-06}$	**$1.54e^{-03}$**
GO:0003735	MF	structural constituent of ribosome	157/13168	10102/401320	2	$6.72e^{-06}$	**$1.59e^{-03}$**
GO:0140110	MF	transcription regulator activity	156/13168	10273/401320	1	$7.00e^{-06}$	**$1.66e^{-03}$**
GO:0003700	MF	DNA-binding transcription factor activity	156/13168	10273/401320	2	$7.00e^{-06}$	**$1.66e^{-03}$**
GO:0005198	MF	structural molecule activity	157/13168	11671/401320	1	$8.22e^{-06}$	**$1.95e^{-03}$**
GO:0016818	MF	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	11/13168	13032/401320	4	$8.50e^{-06}$	**$2.01e^{-03}$**
GO:0016817	MF	hydrolase activity, acting on acid anhydrides	11/13168	13032/401320	3	$8.50e^{-06}$	**$2.01e^{-03}$**
GO:0016462	MF	pyrophosphatase activity	11/13168	13032/401320	5	$8.50e^{-06}$	**$2.01e^{-03}$**
GO:0017111	MF	nucleoside-triphosphatase activity	11/13168	13032/401320	6	$8.50e^{-06}$	**$2.01e^{-03}$**
GO:0016301	MF	kinase activity	284/13168	16743/401320	4	$9.23e^{-06}$	**$2.19e^{-03}$**
GO:0003677	MF	DNA binding	322/13168	20257/401320	4	$1.11e^{-05}$	**$2.64e^{-03}$**
GO:1901363	MF	heterocyclic compound binding	793/13168	30244/401320	2	$1.27e^{-05}$	**$3.01e^{-03}$**
GO:0003676	MF	nucleic acid binding	793/13168	30244/401320	3	$1.27e^{-05}$	**$3.01e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0097159	MF	organic cyclic compound binding	793/13168	30244/401320	2	$1.27e^{-05}$	**$3.01e^{-03}$**
GO:0016787	MF	hydrolase activity	1084/13168	40915/401320	2	$1.46e^{-05}$	**$3.45e^{-03}$**
GO:0003674	MF	molecular_function	10008/13168	325539/401320	0	$1.97e^{-05}$	**$4.66e^{-03}$**
GO:0043167	MF	ion binding	2216/13168	107256/401320	2	$2.18e^{-05}$	**$5.17e^{-03}$**
GO:0005488	MF	binding	3063/13168	132438/401320	1	$2.43e^{-05}$	**$5.75e^{-03}$**
GO:0008134	MF	transcription factor binding	0/13168	317/401320	3	$4.80e^{-05}$	**0.0114**
GO:0098772	MF	molecular function regulator	52/13168	2649/401320	1	$6.10e^{-05}$	**0.0144**
GO:0030234	MF	enzyme regulator activity	52/13168	2649/401320	2	$6.10e^{-05}$	**0.0144**
GO:0003729	MF	mRNA binding	0/13168	311/401320	5	$7.56e^{-05}$	**0.0179**
GO:0016772	MF	transferase activity, transferring phosphorus-containing groups	621/13168	21794/401320	3	$2.02e^{-04}$	**0.048**
GO:0032182	MF	ubiquitin-like protein binding	0/13168	189/401320	3	$3.38e^{-03}$	0.801
GO:0042393	MF	histone binding	0/13168	164/401320	3	$7.33e^{-03}$	1
GO:0019899	MF	enzyme binding	89/13168	3537/401320	3	$9.06e^{-03}$	1
GO:0042578	MF	phosphoric ester hydrolase activity	41/13168	1699/401320	4	0.0473	1
GO:0016791	MF	phosphatase activity	41/13168	1699/401320	5	0.0473	1

Table 2.3.18. Overrepresented GO-slims in strict component genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly overrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0051301	BP	cell division	63/86709	87/401320	2	$1.35e^{-06}$	**$3.20e^{-04}$**
GO:0071826	BP	ribonucleoprotein complex subunit organization	134/86709	313/401320	4	$2.48e^{-06}$	**$5.87e^{-04}$**
GO:0034622	BP	cellular protein-containing complex assembly	134/86709	313/401320	5	$2.48e^{-06}$	**$5.87e^{-04}$**
GO:0022618	BP	ribonucleoprotein complex assembly	134/86709	313/401320	6	$2.48e^{-06}$	**$5.87e^{-04}$**
GO:0007059	BP	chromosome segregation	113/86709	289/401320	2	$2.80e^{-06}$	**$6.64e^{-04}$**
GO:0051169	BP	nuclear transport	172/86709	517/401320	5	$3.84e^{-06}$	**$9.10e^{-04}$**
GO:0006913	BP	nucleocytoplasmic transport	172/86709	517/401320	6	$3.84e^{-06}$	**$9.10e^{-04}$**
GO:0061024	BP	membrane organization	159/86709	516/401320	3	$4.89e^{-06}$	**$1.16e^{-03}$**
GO:0006886	BP	intracellular protein transport	303/86709	919/401320	8	$5.50e^{-06}$	**$1.30e^{-03}$**
GO:0006605	BP	protein targeting	303/86709	919/401320	9	$5.50e^{-06}$	**$1.30e^{-03}$**
GO:0006457	BP	protein folding	318/86709	1039/401320	2	$5.70e^{-06}$	**$1.35e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0051641	BP	cellular localization	475/86709	1440/401320	2	$6.90e^{-06}$	**$1.63e^{-03}$**
GO:0051649	BP	establishment of localization in cell	475/86709	1440/401320	3	$6.90e^{-06}$	**$1.63e^{-03}$**
GO:0046907	BP	intracellular transport	475/86709	1440/401320	4	$6.90e^{-06}$	**$1.63e^{-03}$**
GO:0022607	BP	cellular component assembly	765/86709	2992/401320	3	$9.00e^{-06}$	**$2.13e^{-03}$**
GO:0043933	BP	protein-containing complex subunit organization	678/86709	2472/401320	3	$9.00e^{-06}$	**$2.13e^{-03}$**
GO:0065003	BP	protein-containing complex assembly	678/86709	2472/401320	4	$9.00e^{-06}$	**$2.13e^{-03}$**
GO:0065008	BP	regulation of biological quality	717/86709	2448/401320	2	$9.10e^{-06}$	**$2.16e^{-03}$**
GO:0042592	BP	homeostatic process	717/86709	2448/401320	3	$9.10e^{-06}$	**$2.16e^{-03}$**
GO:0051186	BP	cofactor metabolic process	847/86709	3046/401320	3	$9.76e^{-06}$	**$2.31e^{-03}$**
GO:0016192	BP	vesicle-mediated transport	1455/86709	5466/401320	4	$1.24e^{-05}$	**$2.95e^{-03}$**
GO:0050896	BP	response to stimulus	1706/86709	6331/401320	1	$1.41e^{-05}$	**$3.35e^{-03}$**
GO:0006950	BP	response to stress	1706/86709	6331/401320	2	$1.41e^{-05}$	**$3.35e^{-03}$**
GO:0009056	BP	catabolic process	1799/86709	6927/401320	2	$1.49e^{-05}$	**$3.53e^{-03}$**
GO:0006259	BP	DNA metabolic process	1939/86709	7926/401320	6	$1.59e^{-05}$	**$3.78e^{-03}$**
GO:0006629	BP	lipid metabolic process	2461/86709	9168/401320	3	$1.70e^{-05}$	**$4.03e^{-03}$**
GO:0044249	BP	cellular biosynthetic process	2573/86709	9985/401320	3	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:1901576	BP	organic substance biosynthetic process	2573/86709	9985/401320	3	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0009059	BP	macromolecule biosynthetic process	2573/86709	9985/401320	4	$1.82e^{-05}$	**$4.32e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0043603	BP	cellular amide metabolic process	2573/86709	9985/401320	4	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0044271	BP	cellular nitrogen compound biosynthetic process	2573/86709	9985/401320	4	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:1901566	BP	organonitrogen compound biosynthetic process	2573/86709	9985/401320	4	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0006518	BP	peptide metabolic process	2573/86709	9985/401320	5	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0034645	BP	cellular macromolecule biosynthetic process	2573/86709	9985/401320	5	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0043604	BP	amide biosynthetic process	2573/86709	9985/401320	5	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0043043	BP	peptide biosynthetic process	2573/86709	9985/401320	6	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0006412	BP	translation	2573/86709	9985/401320	7	$1.82e^{-05}$	**$4.32e^{-03}$**
GO:0044281	BP	small molecule metabolic process	4522/86709	17822/401320	2	$2.47e^{-05}$	**$5.85e^{-03}$**
GO:0000278	BP	mitotic cell cycle	97/86709	303/401320	3	$2.53e^{-05}$	**$6.00e^{-03}$**
GO:0009058	BP	biosynthetic process	10877/86709	42776/401320	2	$3.68e^{-05}$	**$8.72e^{-03}$**
GO:0034641	BP	cellular nitrogen compound metabolic process	12336/86709	52479/401320	3	$4.01e^{-05}$	**$9.50e^{-03}$**
GO:0008152	BP	metabolic process	25649/86709	112739/401320	1	$5.52e^{-05}$	**0.0131**
GO:0016043	BP	cellular component organization	1461/86709	6188/401320	2	$1.46e^{-04}$	**0.0345**
GO:0007049	BP	cell cycle	105/86709	352/401320	2	$3.46e^{-04}$	0.082
GO:0090304	BP	nucleic acid metabolic process	3623/86709	15956/401320	5	$6.36e^{-04}$	0.151
GO:0051276	BP	chromosome organization	459/86709	1844/401320	4	$7.34e^{-04}$	0.174
GO:0033036	BP	macromolecule localization	1383/86709	5930/401320	2	$1.32e^{-03}$	0.313

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0008104	BP	protein localization	1383/86709	5930/401320	3	$1.32e^{-03}$	0.313
GO:0045184	BP	establishment of protein localization	1383/86709	5930/401320	4	$1.32e^{-03}$	0.313
GO:0071702	BP	organic substance transport	1383/86709	5930/401320	4	$1.32e^{-03}$	0.313
GO:0071705	BP	nitrogen compound transport	1383/86709	5930/401320	4	$1.32e^{-03}$	0.313
GO:0042886	BP	amide transport	1383/86709	5930/401320	5	$1.32e^{-03}$	0.313
GO:0015833	BP	peptide transport	1383/86709	5930/401320	6	$1.32e^{-03}$	0.313
GO:0015031	BP	protein transport	1383/86709	5930/401320	7	$1.32e^{-03}$	0.313
GO:0032502	BP	developmental process	13/86709	33/401320	1	0.0189	1
GO:0048856	BP	anatomical structure development	13/86709	31/401320	2	0.0137	1
GO:0006725	BP	cellular aromatic compound metabolic process	3704/86709	16648/401320	3	0.0405	1
GO:0046483	BP	heterocycle metabolic process	3704/86709	16648/401320	3	0.0405	1
GO:1901360	BP	organic cyclic compound metabolic process	3704/86709	16648/401320	3	0.0405	1
GO:0006139	BP	nucleobase-containing compound metabolic process	3704/86709	16648/401320	4	0.0405	1
GO:0051604	BP	protein maturation	137/86709	513/401320	5	$6.12e^{-03}$	1
GO:0016020	CC	membrane	213/86709	502/401320	1	$3.60e^{-06}$	**$8.53e^{-04}$**
GO:0005886	CC	plasma membrane	213/86709	502/401320	2	$3.60e^{-06}$	**$8.53e^{-04}$**
GO:0042579	CC	microbody	177/86709	480/401320	5	$3.78e^{-06}$	**$8.97e^{-04}$**
GO:0005777	CC	peroxisome	177/86709	480/401320	6	$3.78e^{-06}$	**$8.97e^{-04}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0030312	CC	external encapsulating structure	376/86709	902/401320	2	$5.43e^{-06}$	**$1.29e^{-03}$**
GO:0005618	CC	cell wall	376/86709	902/401320	3	$5.43e^{-06}$	**$1.29e^{-03}$**
GO:0005622	CC	intracellular	2174/86709	7889/401320	2	$1.52e^{-05}$	**$3.61e^{-03}$**
GO:1990904	CC	ribonucleoprotein complex	2558/86709	9786/401320	2	$1.78e^{-05}$	**$4.23e^{-03}$**
GO:0005840	CC	ribosome	2558/86709	9786/401320	5	$1.78e^{-05}$	**$4.23e^{-03}$**
GO:0043228	CC	non-membrane-bounded organelle	2985/86709	11783/401320	2	$1.94e^{-05}$	**$4.59e^{-03}$**
GO:0043232	CC	intracellular non-membrane-bounded organelle	2985/86709	11783/401320	4	$1.94e^{-05}$	**$4.59e^{-03}$**
GO:0044444	CC	cytoplasmic part	3049/86709	12564/401320	3	$2.02e^{-05}$	**$4.79e^{-03}$**
GO:0032991	CC	protein-containing complex	9227/86709	34280/401320	1	$3.41e^{-05}$	**$8.07e^{-03}$**
GO:0044464	CC	cell part	9452/86709	42555/401320	1	$1.40e^{-03}$	**0.331**
GO:0005694	CC	chromosome	269/86709	1051/401320	5	$2.07e^{-03}$	**0.491**
GO:0005856	CC	cytoskeleton	124/86709	450/401320	5	$2.83e^{-03}$	0.672
GO:0042393	MF	histone binding	105/86709	164/401320	3	$2.39e^{-06}$	**$5.67e^{-04}$**
GO:0051082	MF	unfolded protein binding	321/86709	1067/401320	3	$5.30e^{-06}$	**$1.26e^{-03}$**
GO:0042578	MF	phosphoric ester hydrolase activity	657/86709	1699/401320	4	$6.95e^{-06}$	**$1.65e^{-03}$**
GO:0016791	MF	phosphatase activity	657/86709	1699/401320	5	$6.95e^{-06}$	**$1.65e^{-03}$**
GO:0016765	MF	transferase activity, transferring alkyl or aryl (other than methyl) groups	412/86709	1381/401320	3	$6.99e^{-06}$	**$1.66e^{-03}$**
GO:0016810	MF	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	486/86709	1820/401320	3	$7.24e^{-06}$	**$1.71e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0045182	MF	translation regulator activity	656/86709	2158/401320	1	$8.12e^{-06}$	**$1.93e^{-03}$**
GO:0090079	MF	translation regulator activity, nucleic acid binding	656/86709	2158/401320	4	$8.12e^{-06}$	**$1.93e^{-03}$**
GO:0008135	MF	translation factor activity, RNA binding	656/86709	2158/401320	5	$8.12e^{-06}$	**$1.93e^{-03}$**
GO:0008289	MF	lipid binding	540/86709	1961/401320	2	$8.19e^{-06}$	**$1.94e^{-03}$**
GO:0098772	MF	molecular function regulator	984/86709	2649/401320	1	$9.32e^{-06}$	**$2.21e^{-03}$**
GO:0030234	MF	enzyme regulator activity	984/86709	2649/401320	2	$9.32e^{-06}$	**$2.21e^{-03}$**
GO:0019899	MF	enzyme binding	887/86709	3537/401320	3	$1.12e^{-05}$	**$2.64e^{-03}$**
GO:0016788	MF	hydrolase activity, acting on ester bonds	1376/86709	5627/401320	3	$1.32e^{-05}$	**$3.13e^{-03}$**
GO:0016741	MF	transferase activity, transferring one-carbon groups	1413/86709	5544/401320	3	$1.33e^{-05}$	**$3.15e^{-03}$**
GO:0008168	MF	methyltransferase activity	1413/86709	5544/401320	4	$1.33e^{-05}$	**$3.15e^{-03}$**
GO:0016779	MF	nucleotidyltransferase activity	1238/86709	5051/401320	4	$1.34e^{-05}$	**$3.17e^{-03}$**
GO:0016874	MF	ligase activity	1705/86709	6591/401320	2	$1.38e^{-05}$	**$3.27e^{-03}$**
GO:0003735	MF	structural constituent of ribosome	2546/86709	10102/401320	2	$1.72e^{-05}$	**$4.08e^{-03}$**
GO:0016798	MF	hydrolase activity, acting on glycosyl bonds	2227/86709	9306/401320	3	$1.75e^{-05}$	**$4.14e^{-03}$**
GO:0003723	MF	RNA binding	2532/86709	9993/401320	4	$1.84e^{-05}$	**$4.35e^{-03}$**
GO:0005198	MF	structural molecule activity	3024/86709	11671/401320	1	$1.86e^{-05}$	**$4.42e^{-03}$**
GO:0016853	MF	isomerase activity	1030/86709	4244/401320	2	$4.00e^{-05}$	**$9.48e^{-03}$**
GO:0003674	MF	molecular function	72609/86709	325539/401320	0	$4.66e^{-05}$	**0.011**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0016829	MF	lyase activity	1524/86709	6527/401320	2	$6.45e^{-04}$	0.153
GO:0140096	MF	catalytic activity, acting on a protein	2556/86709	11340/401320	2	0.0146	1
GO:0008233	MF	peptidase activity	2556/86709	11340/401320	3	0.0146	1

Table 2.3.19. Underrepresented GO-slims in strict component genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly underrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0071941	BP	nitrogen cycle metabolic process	17/86709	386/401320	3	$3.03e^{-06}$	**$7.17e^{-04}$**
GO:0032196	BP	transposition	17/86709	537/401320	2	$3.55e^{-06}$	**$8.42e^{-04}$**
GO:0007005	BP	mitochondrion organization	35/86709	542/401320	4	$3.97e^{-06}$	**$9.40e^{-04}$**
GO:1901575	BP	organic substance catabolic process	88/86709	823/401320	3	$4.39e^{-06}$	**$1.04e^{-03}$**
GO:0019439	BP	aromatic compound catabolic process	88/86709	823/401320	4	$4.39e^{-06}$	**$1.04e^{-03}$**
GO:0044270	BP	cellular nitrogen compound catabolic process	88/86709	823/401320	4	$4.39e^{-06}$	**$1.04e^{-03}$**
GO:0046700	BP	heterocycle catabolic process	88/86709	823/401320	4	$4.39e^{-06}$	**$1.04e^{-03}$**
GO:1901361	BP	organic cyclic compound catabolic process	88/86709	823/401320	4	$4.39e^{-06}$	**$1.04e^{-03}$**
GO:0034655	BP	nucleobase-containing compound catabolic process	88/86709	823/401320	5	$4.39e^{-06}$	**$1.04e^{-03}$**
GO:0044085	BP	cellular component biogenesis	15/86709	914/401320	2	$5.18e^{-06}$	**$1.23e^{-03}$**
GO:0022613	BP	ribonucleoprotein complex biogenesis	15/86709	914/401320	3	$5.18e^{-06}$	**$1.23e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0042254	BP	ribosome biogenesis	15/86709	914/401320	4	$5.18e^{-06}$	**$1.23e^{-03}$**
GO:0007010	BP	cytoskeleton organization	111/86709	781/401320	4	$5.65e^{-06}$	**$1.34e^{-03}$**
GO:0044248	BP	cellular catabolic process	151/86709	1301/401320	3	$5.95e^{-06}$	**$1.41e^{-03}$**
GO:0061919	BP	process utilizing autophagic mechanism	63/86709	478/401320	2	$6.01e^{-06}$	**$1.43e^{-03}$**
GO:0006914	BP	autophagy	63/86709	478/401320	4	$6.01e^{-06}$	**$1.43e^{-03}$**
GO:0006790	BP	sulfur compound metabolic process	177/86709	1672/401320	3	$6.74e^{-06}$	**$1.60e^{-03}$**
GO:0006091	BP	generation of precursor metabolites and energy	125/86709	1417/401320	3	$6.77e^{-0}6$	**$1.60e^{-03}$**
GO:0050789	BP	regulation of biological process	811/86709	5515/401320	2	$1.26e^{-05}$	**$2.98e^{-03}$**
GO:0050794	BP	regulation of cellular process	811/86709	5515/401320	3	$1.26e^{-05}$	**$2.98e^{-03}$**
GO:0007165	BP	signal transduction	811/86709	5515/401320	4	$1.26e^{-05}$	**$2.98e^{-03}$**
GO:0006396	BP	RNA processing	359/86709	2039/401320	7	$1.38e^{-05}$	**$3.27e^{-03}$**
GO:0016071	BP	mRNA metabolic process	359/86709	2039/401320	7	$1.38e^{-05}$	**$3.27e^{-03}$**
GO:0006397	BP	mRNA processing	359/86709	2039/401320	8	$1.38e^{-05}$	**$3.27e^{-03}$**
GO:0065007	BP	biological regulation	1528/86709	7963/401320	1	$1.67e^{-05}$	**$3.95e^{-03}$**
GO:0043412	BP	macromolecule modification	3063/86709	21692/401320	4	$2.69e^{-05}$	**$6.37e^{-03}$**
GO:0036211	BP	protein modification process	3063/86709	21692/401320	5	$2.69e^{-05}$	**$6.37e^{-03}$**
GO:0006464	BP	cellular protein modification process	3063/86709	21692/401320	6	$2.69e^{-05}$	**$6.37e^{-03}$**
GO:0044267	BP	cellular protein metabolic process	5636/86709	31675/401320	5	$3.17e^{-05}$	**$7.50e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0019538	BP	protein metabolic process	5720/86709	31990/401320	4	$3.23e^{-05}$	**$7.66e^{-03}$**
GO:0044260	BP	cellular macromolecule metabolic process	7575/86709	39601/401320	4	$3.55e^{-05}$	**$8.41e^{-03}$**
GO:1901564	BP	organonitrogen compound metabolic process	7367/86709	39471/401320	3	$3.56e^{-05}$	**$8.43e^{-03}$**
GO:0055085	BP	transmembrane transport	4052/86709	39724/401320	4	$3.64e^{-05}$	**$8.62e^{-03}$**
GO:0043170	BP	macromolecule metabolic process	9343/86709	47834/401320	3	$3.89e^{-05}$	**$9.23e^{-03}$**
GO:0051179	BP	localization	7334/86709	53205/401320	1	$4.03e^{-05}$	**$9.55e^{-03}$**
GO:0051234	BP	establishment of localization	7334/86709	53205/401320	2	$4.03e^{-05}$	**$9.55e^{-03}$**
GO:0006810	BP	transport	7334/86709	53205/401320	3	$4.03e^{-05}$	**$9.55e^{-03}$**
GO:0009987	BP	cellular process	19029/86709	90607/401320	1	$5.11e^{-05}$	**0.0121**
GO:0044238	BP	primary metabolic process	16098/86709	76658/401320	2	$5.36e^{-05}$	**0.0127**
GO:0071704	BP	organic substance metabolic process	16098/86709	76658/401320	2	$5.36e^{-05}$	**0.0127**
GO:0008150	BP	biological_process	50281/86709	247920/401320	0	$5.98e^{-05}$	**0.0142**
GO:0044237	BP	cellular metabolic process	16662/86709	79144/401320	2	$7.43e^{-05}$	**0.0176**
GO:0006996	BP	organelle organization	605/86709	3167/401320	3	$5.23e^{-04}$	0.124
GO:0006807	BP	nitrogen compound metabolic process	16288/86709	76971/401320	2	$9.09e^{-04}$	0.216
GO:0019748	BP	secondary metabolic process	154/86709	896/401320	2	$1.14e^{-03}$	0.269
GO:0007034	BP	vacuolar transport	138/86709	802/401320	4	$2.27e^{-03}$	0.537
GO:0005730	CC	nucleolus	34/86709	497/401320	5	$3.20e^{-06}$	**$7.58e^{-04}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0044428	CC	nuclear part	114/86709	803/401320	4	$5.09e^{-06}$	**$1.21e^{-03}$**
GO:0005783	CC	endoplasmic reticulum	87/86709	927/401320	5	$6.02e^{-06}$	**$1.43e^{-03}$**
GO:0005576	CC	extracellular region	210/86709	1368/401320	1	$6.33e^{-06}$	**$1.50e^{-03}$**
GO:0044422	CC	organelle part	221/86709	1360/401320	1	$6.94e^{-06}$	**$1.64e^{-03}$**
GO:0044446	CC	intracellular organelle part	221/86709	1360/401320	3	$6.94e^{-06}$	**$1.64e^{-03}$**
GO:0005739	CC	mitochondrion	194/86709	1210/401320	5	$7.34e^{-06}$	**$1.74e^{-03}$**
GO:0005634	CC	nucleus	3799/86709	19392/401320	5	$2.52e^{-05}$	**$5.98e^{-03}$**
GO:0043227	CC	membrane-bounded organelle	4260/86709	22063/401320	2	$2.68e^{-05}$	**$6.36e^{-03}$**
GO:0043231	CC	intracellular membrane-bounded organelle	4260/86709	22063/401320	4	$2.68e^{-05}$	**$6.36e^{-03}$**
GO:0005575	CC	cellular_component	22547/86709	122687/401320	0	$5.69e^{-05}$	**0.0135**
GO:0005773	CC	vacuole	3/86709	54/401320	5	$2.43e^{-03}$	0.576
GO:0044424	CC	intracellular part	8279/86709	39026/401320	2	0.0484	1
GO:0060090	MF	molecular adaptor activity	2/86709	99/401320	2	$1.83e^{-06}$	**$4.34e^{-04}$**
GO:0030674	MF	protein binding, bridging	2/86709	99/401320	3	$1.83e^{-06}$	**$4.34e^{-04}$**
GO:0032182	MF	ubiquitin-like protein binding	2/86709	189/401320	3	$2.08e^{-06}$	**$4.94e^{-04}$**
GO:0008134	MF	transcription factor binding	0/86709	317/401320	3	$2.91e^{-06}$	**$6.89e^{-04}$**
GO:0004386	MF	helicase activity	173/86709	2011/401320	7	$7.24e^{-06}$	**$1.72e^{-03}$**
GO:0008092	MF	cytoskeletal protein binding	384/86709	2945/401320	3	$9.43e^{-06}$	**$2.24e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0004518	MF	nuclease activity	719/86709	3928/401320	4	$1.16e^{-05}$	**$2.74e^{-03}$**
GO:0019843	MF	rRNA binding	89/86709	616/401320	5	$1.23e^{-05}$	**$2.91e^{-03}$**
GO:0003924	MF	GTPase activity	719/86709	5881/401320	7	$1.31e^{-05}$	**$3.10e^{-03}$**
GO:0016887	MF	ATPase activity	391/86709	5630/401320	7	$1.33e^{-05}$	**$3.15e^{-03}$**
GO:0016757	MF	transferase activity, transferring glycosyl groups	908/86709	5667/401320	3	$1.42e^{-05}$	**$3.36e^{-03}$**
GO:0140110	MF	transcription regulator activity	1781/86709	10273/401320	1	$1.81e^{-05}$	**$4.29e^{-03}$**
GO:0003700	MF	DNA-binding transcription factor activity	1781/86709	10273/401320	2	$1.81e^{-05}$	**$4.29e^{-03}$**
GO:0016817	MF	hydrolase activity, acting on acid anhydrides	1238/86709	13032/401320	3	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0016818	MF	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	1238/86709	13032/401320	4	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0016462	MF	pyrophosphatase activity	1238/86709	13032/401320	5	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0017111	MF	nucleoside-triphosphatase activity	1238/86709	13032/401320	6	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0016301	MF	kinase activity	2479/86709	16743/401320	4	$2.33e^{-05}$	**$5.53e^{-03}$**
GO:0003677	MF	DNA binding	3790/86709	20257/401320	4	$2.51e^{-05}$	**$5.95e^{-03}$**
GO:0016772	MF	transferase activity, transferring phosphorus-containing groups	3717/86709	21794/401320	3	$2.60e^{-05}$	**$6.15e^{-03}$**
GO:0005215	MF	transporter activity	3008/86709	21900/401320	1	$2.71e^{-05}$	**$6.42e^{-03}$**
GO:0022857	MF	transmembrane transporter activity	3008/86709	21900/401320	2	$2.71e^{-05}$	**$6.42e^{-03}$**
GO:0016740	MF	transferase activity	7356/86709	38813/401320	2	$3.58e^{-05}$	**$8.47e^{-03}$**
GO:0016787	MF	hydrolase activity	7883/86709	40915/401320	2	$3.69e^{-05}$	**$8.74e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0043167	MF	ion binding	18248/86709	107256/401320	2	$5.37e^{-05}$	**0.0127**
GO:0005488	MF	binding	24200/86709	132438/401320	1	$5.78e^{-05}$	**0.0137**
GO:0003824	MF	catalytic activity	28917/86709	141848/401320	1	$5.93e^{-05}$	**0.014**
GO:0016491	MF	oxidoreductase activity	9664/86709	45921/401320	2	$1.92e^{-03}$	0.454
GO:0097159	MF	organic cyclic compound binding	6322/86709	30244/401320	2	$2.02e^{-03}$	0.478
GO:1901363	MF	heterocyclic compound binding	6322/86709	30244/401320	2	$2.02e^{-03}$	0.478
GO:0003676	MF	nucleic acid binding	6322/86709	30244/401320	3	$2.02e^{-03}$	0.478

DNA repair and mitosis are both highly conserved and tightly coordinated processes which can be significantly impaired by mutation (Hakem, 2008). Due to these factors, it is surprising to observe these ontologies in any remodelling category.

2.3.7.4. Non-remodelled genes are likely to possess conserved housekeeping functions

As observed in overrepresented strict component ontologs, non-remodelled ontologs were enriched for mitotic processes ($P_B \leq 8.33e^{-03}$) and for ribosomal biogenesis (GO:0042254; $P_B = 1.33e^{-03}$) (Tables 2.3.17-18.). As observed in strict components, non-remodelled ontologs were also underrepresented for signalling. Due to the highly conserved and coordinated role of the ribosome in protein synthesis, it is not surprising that these genes are not remodelled.

2.3.7.5. Functions of remodelled genes emerging at the root of Pezizomycotina correlate to phenotype

With 53 of 107 species, Pezizomycotina constitute the most speciose clade in our dataset (49.53%; Tables 2.2.2.-2.2.3). The economic value of Pezizomycotina is reflected by their relative volume of sequencing projects compared to other fungal lineages (Geiser *et al.*, 2006). The divergence of Pezizomycotina from Saccharomycotina is marked by considerable changes in phenotype, such as the transition from a predominantly anamorphic lifecycle to a predominantly teleomorphic life cycle, a transition towards predominant multicellularity, and the expansion of secondary metabolism pathways (Spatafora and Bushley, 2015). When remodelled gene families annotated to the root of Pezizomycotina ("Node_115") were tested for functional overrepresentation, considerable coincidences between function and phenotype were observed (Table 2.3.19.). Strict composites displayed significant overrepresentation for

Table 2.3.20: Overrepresented GO-slims in non-remodelled genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly overrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0007009	BP	plasma membrane organization	11/85510	11/401320	4	$4.11e^{-08}$	**9.73*e*-06**
GO:0071826	BP	ribonucleoprotein complex subunit organization	175/85510	313/401320	4	$2.20e^{-06}$	**5.20*e*-04**
GO:0034622	BP	cellular protein-containing complex assembly	175/85510	313/401320	5	$2.20e^{-06}$	**5.20*e*-04**
GO:0022618	BP	ribonucleoprotein complex assembly	175/85510	313/401320	6	$2.20e^{-06}$	**5.20*e*-04**
GO:0000278	BP	mitotic cell cycle	182/85510	303/401320	3	$3.01e^{-06}$	**7.13*e*-04**
GO:0007049	BP	cell cycle	182/85510	352/401320	2	$3.24e^{-06}$	**7.69*e*-04**
GO:0061024	BP	membrane organization	267/85510	516/401320	3	$3.37e^{-06}$	**7.99*e*-04**
GO:0007005	BP	mitochondrion organization	439/85510	542/401320	4	$3.51e^{-06}$	**8.32*e*-04**
GO:0007059	BP	chromosome segregation	98/85510	289/401320	2	$3.52e^{-06}$	**8.33*e*-04**
GO:0051604	BP	protein maturation	326/85510	513/401320	5	$4.08e^{-06}$	**9.66*e*-04**
GO:0007010	BP	cytoskeleton organization	307/85510	781/401320	4	$4.10e^{-06}$	**9.72*e*-04**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0061919	BP	process utilizing autophagic mechanism	261/85510	478/401320	2	$4.11e^{-06}$	**9.73e-04**
GO:0006914	BP	autophagy	261/85510	478/401320	4	$4.11e^{-06}$	**9.73e-04**
GO:0007034	BP	vacuolar transport	616/85510	802/401320	4	$4.46e^{-06}$	**1.06e-03**
GO:1901575	BP	organic substance catabolic process	321/85510	823/401320	3	$4.82e^{-06}$	**1.14e-03**
GO:0019439	BP	aromatic compound catabolic process	321/85510	823/401320	4	$4.82e^{-06}$	**1.14e-03**
GO:0044270	BP	cellular nitrogen compound catabolic process	321/85510	823/401320	4	$4.82e^{-06}$	**1.14e-03**
GO:0046700	BP	heterocycle catabolic process	321/85510	823/401320	4	$4.82e^{-06}$	**1.14e-03**
GO:1901361	BP	organic cyclic compound catabolic process	321/85510	823/401320	4	$4.82e^{-06}$	**1.14e-03**
GO:0034655	BP	nucleobase-containing compound catabolic process	321/85510	823/401320	5	$4.82e^{-06}$	**1.14e-03**
GO:0044085	BP	cellular component biogenesis	704/85510	914/401320	2	$5.61e^{-06}$	**1.33e-03**
GO:0022613	BP	ribonucleoprotein complex biogenesis	704/85510	914/401320	3	$5.61e^{-06}$	**1.33e-03**
GO:0042254	BP	ribosome biogenesis	704/85510	914/401320	4	$5.61e^{-06}$	**1.33e-03**
GO:0006886	BP	intracellular protein transport	437/85510	919/401320	8	$5.95e^{-06}$	**1.41e-03**
GO:0006605	BP	protein targeting	437/85510	919/401320	9	$5.95e^{-06}$	**1.41e-03**
GO:0006457	BP	protein folding	617/85510	1039/401320	2	$6.07e^{-06}$	**1.44e-03**
GO:0051641	BP	cellular localization	500/85510	1440/401320	2	$6.17e^{-06}$	**1.46e-03**
GO:0051649	BP	establishment of localization in cell	500/85510	1440/401320	3	$6.17e^{-06}$	**1.46e-03**
GO:0046907	BP	intracellular transport	500/85510	1440/401320	4	$6.17e^{-06}$	**1.46e-03**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0044248	BP	cellular catabolic process	582/85510	1301/401320	3	$6.25e^{-06}$	**1.48*e*-03**
GO:0006790	BP	sulfur compound metabolic process	1483/85510	1672/401320	3	$7.00e^{-06}$	**1.66*e*-03**
GO:0006091	BP	generation of precursor metabolites and energy	1098/85510	1417/401320	3	$7.07e^{-06}$	**1.67*e*-03**
GO:0051276	BP	chromosome organization	632/85510	1844/401320	4	$7.51e^{-06}$	**1.78*e*-03**
GO:0006396	BP	RNA processing	800/85510	2039/401320	7	$7.53e^{-06}$	**1.78*e*-03**
GO:0016071	BP	mRNA metabolic process	800/85510	2039/401320	7	$7.53e^{-06}$	**1.78*e*-03**
GO:0006397	BP	mRNA processing	800/85510	2039/401320	8	$7.53e^{-06}$	**1.78*e*-03**
GO:0043933	BP	protein-containing complex subunit organization	1162/85510	2472/401320	3	$9.07e^{-06}$	**2.15*e*-03**
GO:0065003	BP	protein-containing complex assembly	1162/85510	2472/401320	4	$9.07e^{-06}$	**2.15*e*-03**
GO:0006996	BP	organelle organization	1378/85510	3167/401320	3	$9.66e^{-06}$	**2.29*e*-03**
GO:0051186	BP	cofactor metabolic process	1804/85510	3046/401320	3	$9.78e^{-06}$	**2.32*e*-03**
GO:0022607	BP	cellular component assembly	1590/85510	2992/401320	3	$9.99e^{-06}$	**2.37*e*-03**
GO:0033036	BP	macromolecule localization	1631/85510	5930/401320	2	$1.28e^{-05}$	**3.04*e*-03**
GO:0008104	BP	protein localization	1631/85510	5930/401320	3	$1.28e^{-05}$	**3.04*e*-03**
GO:0045184	BP	establishment of protein localization	1631/85510	5930/401320	4	$1.28e^{-05}$	**3.04*e*-03**
GO:0071702	BP	organic substance transport	1631/85510	5930/401320	4	$1.28e^{-05}$	**3.04*e*-03**
GO:0071705	BP	nitrogen compound transport	1631/85510	5930/401320	4	$1.28e^{-05}$	**3.04*e*-03**
GO:0042886	BP	amide transport	1631/85510	5930/401320	5	$1.28e^{-05}$	**3.04*e*-03**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0015833	BP	peptide transport	1631/85510	5930/401320	6	$1.28e^{-05}$	**3.04e-03**
GO:0015031	BP	protein transport	1631/85510	5930/401320	7	$1.28e^{-05}$	**3.04e-03**
GO:0016192	BP	vesicle-mediated transport	1603/85510	5466/401320	4	$1.31e^{-05}$	**3.11e-03**
GO:0034660	BP	ncRNA metabolic process	1799/85510	5991/401320	7	$1.42e^{-05}$	**3.36e-03**
GO:0006399	BP	tRNA metabolic process	1799/85510	5991/401320	8	$1.42e^{-05}$	**3.36e-03**
GO:0016043	BP	cellular component organization	2842/85510	6188/401320	2	$1.46e^{-05}$	**3.46e-03**
GO:0071840	BP	cellular component organization or biogenesis	3546/85510	7102/401320	1	$1.50e^{-05}$	**3.56e-03**
GO:0016070	BP	RNA metabolic process	2599/85510	8030/401320	6	$1.54e^{-05}$	**3.64e-03**
GO:0009056	BP	catabolic process	2658/85510	6927/401320	2	$1.54e^{-05}$	**3.65e-03**
GO:0006082	BP	organic acid metabolic process	2575/85510	7482/401320	3	$1.57e^{-05}$	**3.72e-03**
GO:0043436	BP	oxoacid metabolic process	2575/85510	7482/401320	4	$1.57e^{-05}$	**3.72e-03**
GO:0019752	BP	carboxylic acid metabolic process	2575/85510	7482/401320	5	$1.57e^{-05}$	**3.72e-03**
GO:0006520	BP	cellular amino acid metabolic process	2575/85510	7482/401320	6	$1.57e^{-05}$	**3.72e-03**
GO:0006259	BP	DNA metabolic process	1885/85510	7926/401320	6	$1.64e^{-05}$	**3.88e-03**
GO:0006629	BP	lipid metabolic process	2626/85510	9168/401320	3	$1.73e^{-05}$	**4.10e-03**
GO:0044249	BP	cellular biosynthetic process	6979/85510	9985/401320	3	$1.75e^{-05}$	**4.14e-03**
GO:1901576	BP	organic substance biosynthetic process	6979/85510	9985/401320	3	$1.75e^{-05}$	**4.14e-03**
GO:0009059	BP	macromolecule biosynthetic process	6979/85510	9985/401320	4	$1.75e^{-05}$	**4.14e-03**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0043603	BP	cellular amide metabolic process	6979/85510	9985/401320	4	$1.75e^{-05}$	**4.14e-03**
GO:0044271	BP	cellular nitrogen compound biosynthetic process	6979/85510	9985/401320	4	$1.75e^{-05}$	**4.14e-03**
GO:1901566	BP	organonitrogen compound biosynthetic process	6979/85510	9985/401320	4	$1.75e^{-05}$	**4.14e-03**
GO:0006518	BP	peptide metabolic process	6979/85510	9985/401320	5	$1.75e^{-05}$	**4.14e-03**
GO:0034645	BP	cellular macromolecule biosynthetic process	6979/85510	9985/401320	5	$1.75e^{-05}$	**4.14e-03**
GO:0043604	BP	amide biosynthetic process	6979/85510	9985/401320	5	$1.75e^{-05}$	**4.14e-03**
GO:0043043	BP	peptide biosynthetic process	6979/85510	9985/401320	6	$1.75e^{-05}$	**4.14e-03**
GO:0006412	BP	translation	6979/85510	9985/401320	7	$1.75e^{-05}$	**4.14e-03**
GO:0050896	BP	response to stimulus	1498/85510	6331/401320	1	$1.95e^{-05}$	**4.62e-03**
GO:0006950	BP	response to stress	1498/85510	6331/401320	2	$1.95e^{-05}$	**4.62e-03**
GO:0005975	BP	carbohydrate metabolic process	3743/85510	16166/401320	3	$2.28e^{-05}$	**5.41e-03**
GO:0006725	BP	cellular aromatic compound metabolic process	4805/85510	16648/401320	3	$2.29e^{-05}$	**5.43e-03**
GO:0046483	BP	heterocycle metabolic process	4805/85510	16648/401320	3	$2.29e^{-05}$	**5.43e-03**
GO:1901360	BP	organic cyclic compound metabolic process	4805/85510	16648/401320	3	$2.29e^{-05}$	**5.43e-03**
GO:0006139	BP	nucleobase-containing compound metabolic process	4805/85510	16648/401320	4	$2.29e^{-05}$	**5.43e-03**
GO:0090304	BP	nucleic acid metabolic process	4484/85510	15956/401320	5	$2.31e^{-05}$	**5.48e-03**
GO:0044281	BP	small molecule metabolic process	6917/85510	17822/401320	2	$2.34e^{-05}$	**5.55e-03**
GO:0044267	BP	cellular protein metabolic process	9432/85510	31675/401320	5	$3.17e^{-05}$	**7.52e-03**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0019538	BP	protein metabolic process	9646/85510	31990/401320	4	$3.23e^{-05}$	**7.67e-03**
GO:1901564	BP	organonitrogen compound metabolic process	12221/85510	39471/401320	3	$3.53e^{-05}$	**8.37e-03**
GO:0044260	BP	cellular macromolecule metabolic process	11317/85510	39601/401320	4	$3.63e^{-05}$	**8.59e-03**
GO:0009058	BP	biosynthetic process	16796/85510	42776/401320	2	$3.75e^{-05}$	**8.88e-03**
GO:0043170	BP	macromolecule metabolic process	14130/85510	47834/401320	3	$3.94e^{-05}$	**9.35e-03**
GO:0034641	BP	cellular nitrogen compound metabolic process	20280/85510	52479/401320	3	$4.04e^{-05}$	**9.58e-03**
GO:0044238	BP	primary metabolic process	22305/85510	76658/401320	2	$4.70e^{-05}$	**0.0111**
GO:0071704	BP	organic substance metabolic process	22305/85510	76658/401320	2	$4.70e^{-05}$	**0.0111**
GO:0006807	BP	nitrogen compound metabolic process	23930/85510	76971/401320	2	$4.73e^{-05}$	**0.0112**
GO:0044237	BP	cellular metabolic process	25520/85510	79144/401320	2	$4.81e^{-05}$	**0.0114**
GO:0009987	BP	cellular process	29121/85510	90607/401320	1	$5.17e^{-05}$	**0.0122**
GO:0008152	BP	metabolic process	32824/85510	112739/401320	1	$5.45e^{-05}$	**0.0129**
GO:0008150	BP	biological_process	56460/85510	247920/401320	0	$5.96e^{-05}$	**0.0141**
GO:0071554	BP	cell wall organization or biogenesis	79/85510	250/401320	2	$1.42e^{-04}$	**0.0336**
GO:0048856	BP	anatomical structure development	15/85510	31/401320	2	$7.44e^{-04}$	0.176
GO:0032502	BP	developmental process	15/85510	33/401320	1	$2.03e^{-03}$	0.481
GO:0030705	BP	cytoskeleton-dependent intracellular transport	4/85510	4/401320	5	$2.06e^{-03}$	0.488
GO:0040011	BP	locomotion	2/85510	2/401320	1	0.0454	1

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0006928	BP	movement of cell or subcellular component	2/85510	2/401320	2	0.0454	1
GO:0008219	BP	cell death	3/85510	3/401320	2	$9.67e^{-03}$	1
GO:0048870	BP	cell motility	2/85510	2/401320	3	0.0454	1
GO:0000229	CC	cytoplasmic chromosome	71/85510	107/401320	6	$1.32e^{-06}$	**$3.13e^{-04}$**
GO:0005856	CC	cytoskeleton	315/85510	450/401320	5	$3.62e^{-06}$	**$8.57e^{-04}$**
GO:0005730	CC	nucleolus	260/85510	497/401320	5	$3.73e^{-06}$	**$8.83e^{-04}$**
GO:0044428	CC	nuclear part	260/85510	803/401320	4	$4.56e^{-06}$	**$1.08e^{-03}$**
GO:0005783	CC	endoplasmic reticulum	747/85510	927/401320	5	$5.41e^{-06}$	**$1.28e^{-03}$**
GO:0044422	CC	organelle part	382/85510	1360/401320	1	$6.23e^{-06}$	**$1.48e^{-03}$**
GO:0044446	CC	intracellular organelle part	382/85510	1360/401320	3	$6.23e^{-06}$	**$1.48e^{-03}$**
GO:0005739	CC	mitochondrion	865/85510	1210/401320	5	$6.52e^{-06}$	**$1.55e^{-03}$**
GO:0005737	CC	cytoplasm	2490/85510	4855/401320	3	$1.32e^{-05}$	**$3.13e^{-03}$**
GO:0005622	CC	intracellular	4691/85510	7889/401320	2	$1.57e^{-05}$	**$3.71e^{-03}$**
GO:1990904	CC	ribonucleoprotein complex	6870/85510	9786/401320	2	$1.71e^{-05}$	**$4.05e^{-03}$**
GO:0005840	CC	ribosome	6870/85510	9786/401320	5	$1.71e^{-05}$	**$4.05e^{-03}$**
GO:0044444	CC	cytoplasmic part	8655/85510	12564/401320	3	$2.03e^{-05}$	**$4.81e^{-03}$**
GO:0043228	CC	non-membrane-bounded organelle	7566/85510	11783/401320	2	$2.05e^{-05}$	**$4.85e^{-03}$**
GO:0043232	CC	intracellular non-membrane-bounded organelle	7566/85510	11783/401320	4	$2.05e^{-05}$	**$4.85e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0032991	CC	protein-containing complex	18892/85510	34280/401320	1	$3.30e^{-05}$	**$7.82e^{-03}$**
GO:0043226	CC	organelle	12553/85510	34799/401320	1	$3.31e^{-05}$	**$7.85e^{-03}$**
GO:0043229	CC	intracellular organelle	11998/85510	33803/401320	3	$3.32e^{-05}$	**$7.86e^{-03}$**
GO:0044424	CC	intracellular part	14464/85510	39026/401320	2	$3.51e^{-05}$	**$8.32e^{-03}$**
GO:0044464	CC	cell part	15421/85510	42555/401320	1	$3.70e^{-05}$	**$8.76e^{-03}$**
GO:0005575	CC	cellular_component	37466/85510	122687/401320	0	$5.60e^{-05}$	**0.0133**
GO:0016020	CC	membrane	141/85510	502/401320	1	$3.11e^{-04}$	0.0736
GO:0005886	CC	plasma membrane	141/85510	502/401320	2	$3.11e^{-04}$	0.0736
GO:0031012	CC	extracellular matrix	3/85510	4/401320	2	0.0325	1
GO:0042995	CC	cell projection	2/85510	2/401320	2	0.0454	1
GO:0120025	CC	plasma membrane bounded cell projection	2/85510	2/401320	3	0.0454	1
GO:0005929	CC	cilium	2/85510	2/401320	4	0.0454	1
GO:0008134	MF	transcription factor binding	216/85510	317/401320	3	$2.56e^{-06}$	**$6.07e^{-04}$**
GO:0003729	MF	mRNA binding	220/85510	311/401320	5	$2.88e^{-06}$	**$6.83e^{-04}$**
GO:0019843	MF	rRNA binding	522/85510	616/401320	5	$4.28e^{-06}$	**$1.02e^{-03}$**
GO:0016765	MF	transferase activity, transferring alkyl or aryl (other than methyl) groups	537/85510	1381/401320	3	$6.28e^{-06}$	**$1.49e^{-03}$**
GO:0042578	MF	phosphoric ester hydrolase activity	469/85510	1699/401320	4	$7.20e^{-06}$	**$1.71e^{-03}$**
GO:0016791	MF	phosphatase activity	469/85510	1699/401320	5	$7.20e^{-06}$	**$1.71e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0045182	MF	translation regulator activity	1235/85510	2158/401320	1	$8.28e^{-06}$	**$1.96e^{-03}$**
GO:0090079	MF	translation regulator activity, nucleic acid binding	1235/85510	2158/401320	4	$8.28e^{-06}$	**$1.96e^{-03}$**
GO:0008135	MF	translation factor activity, RNA binding	1235/85510	2158/401320	5	$8.28e^{-06}$	**$1.96e^{-03}$**
GO:0098772	MF	molecular function regulator	1418/85510	2649/401320	1	$8.39e^{-06}$	**$1.99e^{-03}$**
GO:0030234	MF	enzyme regulator activity	1418/85510	2649/401320	2	$8.39e^{-06}$	**$1.99e^{-03}$**
GO:0016853	MF	isomerase activity	1450/85510	4244/401320	2	$1.12e^{-05}$	**$2.65e^{-03}$**
GO:0004518	MF	nuclease activity	1192/85510	3928/401320	4	$1.14e^{-05}$	**$2.71e^{-03}$**
GO:0016757	MF	transferase activity, transferring glycosyl groups	2144/85510	5667/401320	3	$1.28e^{-05}$	**$3.03e^{-03}$**
GO:0016788	MF	hydrolase activity, acting on ester bonds	1661/85510	5627/401320	3	$1.29e^{-05}$	**$3.05e^{-03}$**
GO:0016741	MF	transferase activity, transferring one-carbon groups	1965/85510	5544/401320	3	$1.29e^{-05}$	**$3.07e^{-03}$**
GO:0008168	MF	methyltransferase activity	1965/85510	5544/401320	4	$1.29e^{-05}$	**$3.07e^{-03}$**
GO:0016779	MF	nucleotidyltransferase activity	1499/85510	5051/401320	4	$1.31e^{-05}$	**$3.10e^{-03}$**
GO:0016829	MF	lyase activity	2598/85510	6527/401320	2	$1.40e^{-05}$	**$3.31e^{-03}$**
GO:0003735	MF	structural constituent of ribosome	7192/85510	10102/401320	2	$1.75e^{-05}$	**$4.14e^{-03}$**
GO:0003723	MF	RNA binding	4586/85510	9993/401320	4	$1.86e^{-05}$	**$4.41e^{-03}$**
GO:0005198	MF	structural molecule activity	7530/85510	11671/401320	1	$1.87e^{-05}$	**$4.44e^{-03}$**
GO:0097159	MF	organic cyclic compound binding	7683/85510	30244/401320	2	$3.15e^{-05}$	**$7.47e^{-03}$**
GO:1901363	MF	heterocyclic compound binding	7683/85510	30244/401320	2	$3.15e^{-05}$	**$7.47e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0003676	MF	nucleic acid binding	7683/85510	30244/401320	3	$3.15e^{-05}$	**$7.47e^{-03}$**
GO:0140096	MF	catalytic activity, acting on a protein	2527/85510	11340/401320	2	0.0105	1
GO:0008233	MF	peptidase activity	2527/85510	11340/401320	3	0.0105	1
GO:0016810	MF	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	434/85510	1820/401320	3	$9.01e^{-03}$	1

Table 2.3.21: Underrepresented GO-slims in non-remodelled genes

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly overrepresented.

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
GO:0032196	BP	transposition	2/85510	537/401320	2	$4.07e^{-06}$	**$9.65e^{-04}$**
GO:0051169	BP	nuclear transport	59/85510	517/401320	5	$4.45e^{-06}$	**$1.06e^{-03}$**
GO:0006913	BP	nucleocytoplasmic transport	59/85510	517/401320	6	$4.45e^{-06}$	**$1.06e^{-03}$**
GO:0050789	BP	regulation of biological process	630/85510	5515/401320	2	$1.33e^{-05}$	**$3.14e^{-03}$**
GO:0050794	BP	regulation of cellular process	630/85510	5515/401320	3	$1.33e^{-05}$	**$3.14e^{-03}$**
GO:0007165	BP	signal transduction	630/85510	5515/401320	4	$1.33e^{-05}$	**$3.14e^{-03}$**
GO:0065007	BP	biological regulation	1168/85510	7963/401320	1	$1.61e^{-05}$	**$3.81e^{-03}$**
GO:0043412	BP	macromolecule modification	2453/85510	21692/401320	4	$2.64e^{-05}$	**$6.25e^{-03}$**
GO:0036211	BP	protein modification process	2453/85510	21692/401320	5	$2.64e^{-05}$	**$6.25e^{-03}$**
GO:0006464	BP	cellular protein modification process	2453/85510	21692/401320	6	$2.64e^{-05}$	**$6.25e^{-03}$**
GO:0055085	BP	transmembrane transport	4085/85510	39724/401320	4	$3.51e^{-05}$	**$8.33e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0051179	BP	localization	8347/85510	53205/401320	1	$4.06e^{-05}$	**$9.63e^{-03}$**
GO:0051234	BP	establishment of localization	8347/85510	53205/401320	2	$4.06e^{-05}$	**$9.63e^{-03}$**
GO:0006810	BP	transport	8347/85510	53205/401320	3	$4.06e^{-05}$	**$9.63e^{-03}$**
GO:0051301	BP	cell division	8/85510	87/401320	2	$3.78e^{-03}$	0.897
GO:0005615	CC	extracellular space	3/85510	222/401320	2	$2.28e^{-06}$	$5.41e^{-04}$
GO:0000228	CC	nuclear chromosome	0/85510	306/401320	6	$2.37e^{-06}$	**$5.63e^{-04}$**
GO:0044421	CC	extracellular region part	6/85510	226/401320	1	$2.80e^{-06}$	**$6.64e^{-04}$**
GO:0005773	CC	vacuole	0/85510	54/401320	5	$5.40e^{-06}$	**$1.28e^{-03}$**
GO:0005576	CC	extracellular region	194/85510	1368/401320	1	$5.67e^{-06}$	**$1.34e^{-03}$**
GO:0005694	CC	chromosome	121/85510	1051/401320	5	$5.74e^{-06}$	**$1.36e^{-03}$**
GO:0030312	CC	external encapsulating structure	59/85510	902/401320	2	$5.91e^{-06}$	**$1.40e^{-03}$**
GO:0005618	CC	cell wall	59/85510	902/401320	3	$5.91e^{-06}$	**$1.40e^{-03}$**
GO:0005634	CC	nucleus	2719/85510	19392/401320	5	$2.49e^{-05}$	**$5.89e^{-03}$**
GO:0043227	CC	membrane-bounded organelle	4433/85510	22063/401320	2	$3.29e^{-05}$	**$7.80e^{-03}$**
GO:0043231	CC	intracellular membrane-bounded organelle	4433/85510	22063/401320	4	$3.29e^{-05}$	**$7.80e^{-03}$**
GO:0032182	MF	ubiquitin-like protein binding	6/85510	189/401320	3	$1.81e^{-06}$	**$4.29e^{-04}$**
GO:0042393	MF	histone binding	4/85510	164/401320	3	$2.04e^{-06}$	**$4.84e^{-04}$**
GO:0060090	MF	molecular adaptor activity	4/85510	99/401320	2	$3.03e^{-06}$	**$7.19e^{-04}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0030674	MF	protein binding, bridging	4/85510	99/401320	3	$3.03e^{-06}$	**$7.19e^{-04}$**
GO:0008289	MF	lipid binding	242/85510	1961/401320	2	$7.35e^{-06}$	**$1.74e^{-03}$**
GO:0004386	MF	helicase activity	226/85510	2011/401320	7	$8.41e^{-06}$	**$1.99e^{-03}$**
GO:0019899	MF	enzyme binding	90/85510	3537/401320	3	$1.11e^{-05}$	**$2.62e^{-03}$**
GO:0016746	MF	transferase activity, transferring acyl groups	810/85510	4434/401320	3	$1.23e^{-05}$	**$2.92e^{-03}$**
GO:0003924	MF	GTPase activity	24/85510	5881/401320	7	$1.37e^{-05}$	**$3.25e^{-03}$**
GO:0016887	MF	ATPase activity	36/85510	5630/401320	7	$1.39e^{-05}$	**$3.30e^{-03}$**
GO:0016874	MF	ligase activity	1191/85510	6591/401320	2	$1.43e^{-05}$	**$3.40e^{-03}$**
GO:0005515	MF	protein binding	1072/85510	7948/401320	2	$1.68e^{-05}$	**$3.98e^{-03}$**
GO:0008092	MF	cytoskeletal protein binding	530/85510	2945/401320	3	$1.76e^{-05}$	**$4.18e^{-03}$**
GO:0016798	MF	hydrolase activity, acting on glycosyl bonds	1532/85510	9306/401320	3	$1.78e^{-05}$	**$4.21e^{-03}$**
GO:0140110	MF	transcription regulator activity	710/85510	10273/401320	1	$1.83e^{-05}$	**$4.34e^{-03}$**
GO:0003700	MF	DNA-binding transcription factor activity	710/85510	10273/401320	2	$1.83e^{-05}$	**$4.34e^{-03}$**
GO:0016817	MF	hydrolase activity, acting on acid anhydrides	280/85510	13032/401320	3	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0016818	MF	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	280/85510	13032/401320	4	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0016462	MF	pyrophosphatase activity	280/85510	13032/401320	5	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0017111	MF	nucleoside-triphosphatase activity	280/85510	13032/401320	6	$2.11e^{-05}$	**$5.01e^{-03}$**
GO:0016301	MF	kinase activity	1099/85510	16743/401320	4	$2.31e^{-05}$	**$5.48e^{-03}$**

GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
GO:0003677	MF	DNA binding	3097/85510	20257/401320	4	$2.57e^{-05}$	**$6.08e^{-03}$**
GO:0016772	MF	transferase activity, transferring phosphorus-containing groups	2598/85510	21794/401320	3	$2.75e^{-05}$	**$6.51e^{-03}$**
GO:0005215	MF	transporter activity	3857/85510	21900/401320	1	$2.76e^{-05}$	**$6.54e^{-03}$**
GO:0022857	MF	transmembrane transporter activity	3857/85510	21900/401320	2	$2.76e^{-05}$	**$6.54e^{-03}$**
GO:0016787	MF	hydrolase activity	6434/85510	40915/401320	2	$3.56e^{-05}$	**$8.43e^{-03}$**
GO:0016491	MF	oxidoreductase activity	7483/85510	45921/401320	2	$3.83e^{-05}$	**$9.07e^{-03}$**
GO:0003674	MF	molecular_function	62472/85510	325539/401320	0	$4.74e^{-05}$	**0.0112**
GO:0043167	MF	ion binding	9061/85510	107256/401320	2	$5.40e^{-05}$	**0.0128**
GO:0005488	MF	binding	16874/85510	132438/401320	1	$5.79e^{-05}$	**0.0137**
GO:0003824	MF	catalytic activity	27099/85510	141848/401320	1	$5.82e^{-05}$	**0.0138**
GO:0016740	MF	transferase activity	8054/85510	38813/401320	2	$4.84e^{-03}$	1

Table 2.3.22: Functional overrepresentations at the root of Pezizomycotina

Each GO-slim entry is annotated with its ontology type (biological process (BP), or cellular component (CC), or molecular function (MF)), its proportion within the sample ($\hat{p}(n)$), its proportion within the background ($\hat{p}(N)$), its annotation depth, its RC, and its uncorrected and corrected *P*-values. Instances where $P_B \leq 0.05$ were considered to be significantly overrepresented. RCs are separated by a solid line and ontology types (within RCs) are separated by a dashed line.

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	*P*	P_B
NC	GO:0050794	BP	regulation of cellular process	467/29738	2629/455713	3	$5.00e^{-06}$	**$8.21e^{-04}$**
NC	GO:0007165	BP	signal transduction	467/29738	2629/455713	4	$5.00e^{-06}$	**$8.21e^{-04}$**
NC	GO:0043412	BP	macromolecule modification	822/29738	9933/455713	4	$1.09e^{-05}$	**$1.79e^{-03}$**
NC	GO:0036211	BP	protein modification process	822/29738	9933/455713	5	$1.09e^{-05}$	**$1.79e^{-03}$**
NC	GO:0006464	BP	cellular protein modification process	822/29738	9933/455713	6	$1.09e^{-05}$	**$1.79e^{-03}$**
NC	GO:0005975	BP	carbohydrate metabolic process	911/29738	9383/455713	3	$1.11e^{-05}$	**$1.81e^{-03}$**
NC	GO:0051179	BP	localization	2542/29738	30759/455713	1	$1.87e^{-05}$	**$3.06e^{-03}$**
NC	GO:0051234	BP	establishment of localization	2542/29738	30759/455713	2	$1.87e^{-05}$	**$3.06e^{-03}$**
NC	GO:0006810	BP	transport	2542/29738	30759/455713	3	$1.87e^{-05}$	**$3.06e^{-03}$**
NC	GO:0008150	BP	biological_process	12220/29738	136382/455713	0	$3.49e^{-05}$	**$5.72e^{-03}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
NC	GO:0042579	CC	microbody	41/29738	216/455713	5	$1.12e^{-06}$	**$1.83e^{-04}$**
NC	GO:0005777	CC	peroxisome	41/29738	216/455713	6	$1.12e^{-06}$	**$1.83e^{-04}$**
NC	GO:0030312	CC	external encapsulating structure	60/29738	261/455713	2	$1.72e^{-06}$	**$2.81e^{-04}$**
NC	GO:0005618	CC	cell wall	60/29738	261/455713	3	$1.72e^{-06}$	**$2.81e^{-04}$**
NC	GO:0044422	CC	organelle part	83/29738	249/455713	1	$1.96e^{-06}$	**$3.21e^{-04}$**
NC	GO:0044446	CC	intracellular organelle part	83/29738	249/455713	3	$1.96e^{-06}$	**$3.21e^{-04}$**
NC	GO:0044428	CC	nuclear part	83/29738	249/455713	4	$1.96e^{-06}$	**$3.21e^{-04}$**
NC	GO:0005730	CC	nucleolus	83/29738	249/455713	5	$1.96e^{-06}$	**$3.21e^{-04}$**
NC	GO:0005576	CC	extracellular region	166/29738	916/455713	1	$2.33e^{-06}$	**$3.82e^{-04}$**
NC	GO:0016298	MF	lipase activity	111/29738	519/455713	4	$1.79e^{-06}$	**$2.94e^{-04}$**
NC	GO:0003774	MF	motor activity	101/29738	798/455713	7	$2.58e^{-06}$	**$4.22e^{-04}$**
NC	GO:0016788	MF	hydrolase activity, acting on ester bonds	191/29738	1347/455713	3	$3.12e^{-06}$	**$5.11e^{-04}$**
NC	GO:0016779	MF	nucleotidyltransferase activity	276/29738	2466/455713	4	$4.42e^{-06}$	**$7.26e^{-04}$**
NC	GO:0008233	MF	peptidase activity	839/29738	5882/455713	3	$7.90e^{-06}$	**$1.30e^{-03}$**
NC	GO:0016301	MF	kinase activity	709/29738	6491/455713	4	$9.29e^{-06}$	**$1.52e^{-03}$**
NC	GO:0016773	MF	phosphotransferase activity, alcohol group as acceptor	709/29738	6491/455713	4	$9.29e^{-06}$	**$1.52e^{-03}$**
NC	GO:0004672	MF	protein kinase activity	709/29738	6491/455713	5	$9.29e^{-06}$	**$1.52e^{-03}$**
NC	GO:0016772	MF	transferase activity, transferring phosphorus-containing groups	985/29738	8957/455713	3	$1.03e^{-05}$	**$1.70e^{-03}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
NC	GO:0140096	MF	catalytic activity, acting on a protein	1548/29738	12373/455713	2	$1.11e^{-05}$	**$1.83e^{-03}$**
NC	GO:0005215	MF	transporter activity	1053/29738	13052/455713	1	$1.24e^{-05}$	**$2.03e^{-03}$**
NC	GO:0016740	MF	transferase activity	1795/29738	21935/455713	2	$1.59e^{-05}$	**$2.60e^{-03}$**
NC	GO:0005515	MF	protein binding	1875/29738	21689/455713	2	$1.62e^{-05}$	**$2.66e^{-03}$**
NC	GO:0016491	MF	oxidoreductase activity	4448/29738	26404/455713	2	$1.70e^{-05}$	**$2.79e^{-03}$**
NC	GO:0016787	MF	hydrolase activity	3929/29738	31846/455713	2	$1.95e^{-05}$	**$3.20e^{-03}$**
NC	GO:0005488	MF	binding	2912/29738	37846/455713	1	$2.04e^{-05}$	**$3.34e^{-03}$**
NC	GO:0003824	MF	catalytic activity	10408/29738	88637/455713	1	$2.98e^{-05}$	**$4.89e^{-03}$**
NC	GO:0003674	MF	molecular_function	16178/29738	174397/455713	0	$3.66e^{-05}$	**$6.00e^{-03}$**
SC	GO:0019748	BP	secondary metabolic process	92/2548	594/455713	2	$4.00e^{-07}$	**$6.56e^{-05}$**
SC	GO:0009404	BP	toxin metabolic process	92/2548	484/455713	3	$9.09e^{-07}$	**$1.49e^{-04}$**
SC	GO:0044281	BP	small molecule metabolic process	88/2548	4313/455713	2	$1.49e^{-06}$	**$2.45e^{-04}$**
SC	GO:0006082	BP	organic acid metabolic process	88/2548	3885/455713	3	$2.07e^{-06}$	**$3.39e^{-04}$**
SC	GO:0043436	BP	oxoacid metabolic process	88/2548	3885/455713	4	$2.07e^{-06}$	**$3.39e^{-04}$**
SC	GO:0019752	BP	carboxylic acid metabolic process	88/2548	3885/455713	5	$2.07e^{-06}$	**$3.39e^{-04}$**
SC	GO:0006520	BP	cellular amino acid metabolic process	88/2548	3885/455713	6	$2.07e^{-06}$	**$3.39e^{-04}$**
SC	GO:0006629	BP	lipid metabolic process	74/2548	5010/455713	3	$2.16e^{-06}$	**$3.54e^{-04}$**
SC	GO:0006950	BP	response to stress	41/2548	3243/455713	2	$4.51e^{-06}$	**$7.40e^{-04}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
SC	GO:0050896	BP	response to stimulus	41/2548	3321/455713	1	$5.51e^{-06}$	**$9.04e^{-04}$**
SC	GO:0044237	BP	cellular metabolic process	306/2548	34688/455713	2	$5.63e^{-06}$	**$9.23e^{-04}$**
SC	GO:0009987	BP	cellular process	306/2548	40381/455713	1	$6.27e^{-06}$	**$1.03e^{-03}$**
SC	GO:0008152	BP	metabolic process	380/2548	49074/455713	1	$7.02e^{-06}$	**$1.15e^{-03}$**
SC	GO:0006259	BP	DNA metabolic process	41/2548	3658/455713	6	$4.65e^{-05}$	**$7.63e^{-03}$**
SC	GO:0016020	CC	membrane	93/2548	11060/455713	1	$1.69e^{-04}$	**0.0277**
SC	GO:0003723	MF	RNA binding	127/2548	4823/455713	4	$2.29e^{-06}$	**$3.76e^{-04}$**
SC	GO:0016491	MF	oxidoreductase activity	221/2548	26404/455713	2	$5.09e^{-06}$	**$8.35e^{-04}$**
SC	GO:0016787	MF	hydrolase activity	396/2548	31846/455713	2	$5.90e^{-06}$	**$9.67e^{-04}$**
SC	GO:0003824	MF	catalytic activity	654/2548	88637/455713	1	$8.76e^{-06}$	**$1.44e^{-03}$**
SC	GO:0016829	MF	lyase activity	43/2548	3814/455713	2	$2.57e^{-05}$	**$4.22e^{-03}$**
SN	GO:0007010	BP	cytoskeleton organization	45/11880	425/455713	4	$1.25e^{-06}$	**$2.05e^{-04}$**
SN	GO:0016043	BP	cellular component organization	94/11880	2166/455713	2	$5.18e^{-06}$	**$8.49e^{-04}$**
SN	GO:0006996	BP	organelle organization	94/11880	2166/455713	3	$5.18e^{-06}$	**$8.49e^{-04}$**
SN	GO:0005975	BP	carbohydrate metabolic process	363/11880	9383/455713	3	$6.07e^{-06}$	**$9.96e^{-04}$**
SN	GO:0030312	CC	external encapsulating structure	31/11880	261/455713	2	$5.06e^{-07}$	**$8.30e^{-05}$**
SN	GO:0005618	CC	cell wall	31/11880	261/455713	3	$5.06e^{-07}$	**$8.30e^{-05}$**
SN	GO:0005886	CC	plasma membrane	46/11880	269/455713	2	$8.56e^{-07}$	**$1.40e^{-04}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
SN	GO:0005694	CC	chromosome	49/11880	532/455713	5	$1.67e^{-06}$	**$2.73e^{-04}$**
SN	GO:0005576	CC	extracellular region	62/11880	916/455713	1	$1.91e^{-06}$	**$3.13e^{-04}$**
SN	GO:0016853	MF	isomerase activity	119/11880	2125/455713	2	$2.95e^{-06}$	**$4.83e^{-04}$**
SN	GO:0008233	MF	peptidase activity	294/11880	5882/455713	3	$4.59e^{-06}$	**$7.52e^{-04}$**
SN	GO:0016491	MF	oxidoreductase activity	956/11880	26404/455713	2	$1.09e^{-05}$	**$1.78e^{-03}$**
SN	GO:0016787	MF	hydrolase activity	1138/11880	31846/455713	2	$1.20e^{-05}$	**$1.97e^{-03}$**
SN	GO:0042578	MF	phosphoric ester hydrolase activity	44/11880	828/455713	4	$1.46e^{-05}$	**$2.40e^{-03}$**
SN	GO:0016791	MF	phosphatase activity	44/11880	828/455713	5	$1.46e^{-05}$	**$2.40e^{-03}$**
SN	GO:0003824	MF	catalytic activity	2706/11880	88637/455713	1	$1.91e^{-05}$	**$3.13e^{-03}$**
SN	GO:0003674	MF	molecular_function	4752/11880	174397/455713	0	$1.16e^{-04}$	**0.019**
NR	GO:0006457	BP	protein folding	44/12198	467/455713	2	$6.60e^{-07}$	**$1.08e^{-04}$**
NR	GO:0044085	BP	cellular component biogenesis	49/12198	452/455713	2	$9.26e^{-07}$	**$1.52e^{-04}$**
NR	GO:0022613	BP	ribonucleoprotein complex biogenesis	49/12198	452/455713	3	$9.26e^{-07}$	**$1.52e^{-04}$**
NR	GO:0042254	BP	ribosome biogenesis	49/12198	452/455713	4	$9.26e^{-07}$	**$1.52e^{-04}$**
NR	GO:0071840	BP	cellular component organization or biogenesis	145/12198	2618/455713	1	$2.56e^{-06}$	**$4.19e^{-04}$**
NR	GO:0016192	BP	vesicle-mediated transport	225/12198	2696/455713	4	$3.59e^{-06}$	**$5.88e^{-04}$**
NR	GO:1901566	BP	organonitrogen compound biosynthetic process	389/12198	4535/455713	4	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0043603	BP	cellular amide metabolic process	389/12198	4535/455713	4	$4.40e^{-06}$	**$7.22e^{-04}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
NR	GO:0006518	BP	peptide metabolic process	389/12198	4535/455713	5	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0043604	BP	amide biosynthetic process	389/12198	4535/455713	5	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0043043	BP	peptide biosynthetic process	389/12198	4535/455713	6	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0006412	BP	translation	389/12198	4535/455713	7	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0016043	BP	cellular component organization	96/12198	2166/455713	2	$4.99e^{-06}$	**$8.18e^{-04}$**
NR	GO:0006996	BP	organelle organization	96/12198	2166/455713	3	$4.99e^{-06}$	**$8.18e^{-04}$**
NR	GO:0009058	BP	biosynthetic process	567/12198	11413/455713	2	$6.83e^{-06}$	**$1.12e^{-03}$**
NR	GO:0044249	BP	cellular biosynthetic process	567/12198	11413/455713	3	$6.83e^{-06}$	**$1.12e^{-03}$**
NR	GO:1901576	BP	organic substance biosynthetic process	567/12198	11413/455713	3	$6.83e^{-06}$	**$1.12e^{-03}$**
NR	GO:0044271	BP	cellular nitrogen compound biosynthetic process	567/12198	11413/455713	4	$6.83e^{-06}$	**$1.12e^{-03}$**
NR	GO:0009059	BP	macromolecule biosynthetic process	567/12198	11413/455713	4	$6.83e^{-06}$	**$1.12e^{-03}$**
NR	GO:0034645	BP	cellular macromolecule biosynthetic process	567/12198	11413/455713	5	$6.83e^{-06}$	**$1.12e^{-03}$**
NR	GO:0016070	BP	RNA metabolic process	641/12198	14160/455713	6	$8.68e^{-06}$	**$1.42e^{-03}$**
NR	GO:0046483	BP	heterocycle metabolic process	723/12198	17717/455713	3	$8.90e^{-06}$	**$1.46e^{-03}$**
NR	GO:1901360	BP	organic cyclic compound metabolic process	723/12198	17717/455713	3	$8.90e^{-06}$	**$1.46e^{-03}$**
NR	GO:0006725	BP	cellular aromatic compound metabolic process	723/12198	17717/455713	3	$8.90e^{-06}$	**$1.46e^{-03}$**
NR	GO:0006139	BP	nucleobase-containing compound metabolic process	723/12198	17717/455713	4	$8.90e^{-06}$	**$1.46e^{-03}$**
NR	GO:0090304	BP	nucleic acid metabolic process	723/12198	17717/455713	5	$8.90e^{-06}$	**$1.46e^{-03}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
NR	GO:0034641	BP	cellular nitrogen compound metabolic process	1111/12198	22247/455713	3	$1.03e^{-05}$	**$1.69e^{-03}$**
NR	GO:0043170	BP	macromolecule metabolic process	1111/12198	32170/455713	3	$1.22e$-^{05}	**$2.01e^{-03}$**
NR	GO:0006807	BP	nitrogen compound metabolic process	1187/12198	34304/455713	2	$1.26e^{-05}$	**$2.06e^{-03}$**
NR	GO:0044237	BP	cellular metabolic process	1187/12198	34688/455713	2	$1.29e^{-05}$	**$2.11e^{-03}$**
NR	GO:0009987	BP	cellular process	1327/12198	40381/455713	1	$1.39e^{-05}$	**$2.28e^{-03}$**
NR	GO:0071704	BP	organic substance metabolic process	1433/12198	48007/455713	2	$2.83e^{-05}$	**$4.65e^{-03}$**
NR	GO:0044238	BP	primary metabolic process	1433/12198	48007/455713	2	$2.83e^{-05}$	**$4.65e^{-03}$**
NR	GO:0005783	CC	endoplasmic reticulum	48/12198	424/455713	5	$6.90e^{-07}$	**$1.13e^{-04}$**
NR	GO:0044422	CC	organelle part	49/12198	249/455713	1	$1.30e^{-06}$	**$2.13e^{-04}$**
NR	GO:0044446	CC	intracellular organelle part	49/12198	249/455713	3	$1.30e^{-06}$	**$2.13e^{-04}$**
NR	GO:0044428	CC	nuclear part	49/12198	249/455713	4	$1.30e^{-06}$	**$2.13e^{-04}$**
NR	GO:0005730	CC	nucleolus	49/12198	249/455713	5	$1.30e^{-06}$	**$2.13e^{-04}$**
NR	GO:0005739	CC	mitochondrion	261/12198	622/455713	5	$2.21e^{-06}$	**$3.62e^{-04}$**
NR	GO:0032991	CC	protein-containing complex	389/12198	4443/455713	1	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:1990904	CC	ribonucleoprotein complex	389/12198	4443/455713	2	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0005840	CC	ribosome	389/12198	4443/455713	5	$4.40e^{-06}$	**$7.22e^{-04}$**
NR	GO:0044444	CC	cytoplasmic part	698/12198	5839/455713	3	$5.61e^{-06}$	**$9.20e^{-04}$**
NR	GO:0043228	CC	non-membrane-bounded organelle	438/12198	5423/455713	2	$5.62e^{-06}$	**$9.21e^{-04}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
NR	GO:0043232	CC	intracellular non-membrane-bounded organelle	438/12198	5423/455713	4	$5.62e^{-06}$	**$9.21e^{-04}$**
NR	GO:0016020	CC	membrane	434/12198	11060/455713	1	$7.50e^{-06}$	**$1.23e^{-03}$**
NR	GO:0043227	CC	membrane-bounded organelle	639/12198	14258/455713	2	$8.23e^{-06}$	**$1.35e^{-03}$**
NR	GO:0043231	CC	intracellular membrane-bounded organelle	639/12198	14258/455713	4	$8.23e^{-06}$	**$1.35e^{-03}$**
NR	GO:0044464	CC	cell part	1076/12198	20268/455713	1	$9.29e^{-06}$	**$1.52e^{-03}$**
NR	GO:0043226	CC	organelle	1076/12198	19648/455713	1	$9.37e^{-06}$	**$1.54e^{-03}$**
NR	GO:0043229	CC	intracellular organelle	1076/12198	19648/455713	3	$9.37e^{-06}$	**$1.54e^{-03}$**
NR	GO:0044424	CC	intracellular part	1076/12198	19738/455713	2	$9.80e^{-06}$	**$1.61e^{-03}$**
NR	GO:0005575	CC	cellular_component	3153/12198	69090/455713	0	$1.81e^{-05}$	**$2.96e^{-03}$**
NR	GO:0042578	MF	phosphoric ester hydrolase activity	96/12198	828/455713	4	$1.25e^{-06}$	**$2.05e^{-04}$**
NR	GO:0016791	MF	phosphatase activity	96/12198	828/455713	5	$1.25e^{-06}$	**$2.05e^{-04}$**
NR	GO:0045182	MF	translation regulator activity	224/12198	1067/455713	1	$2.42e^{-06}$	**$3.97e^{-04}$**
NR	GO:0016788	MF	hydrolase activity, acting on ester bonds	96/12198	1347/455713	3	$2.53e^{-06}$	**$4.15e^{-04}$**
NR	GO:0016853	MF	isomerase activity	118/12198	2125/455713	2	$2.84e^{-06}$	**$4.66e^{-04}$**
NR	GO:0003723	MF	RNA binding	448/12198	4823/455713	4	$4.53e^{-06}$	**$7.43e^{-04}$**
NR	GO:0005198	MF	structural molecule activity	434/12198	5132/455713	1	$4.67e^{-06}$	**$7.65e^{-04}$**
NR	GO:1901363	MF	heterocyclic compound binding	583/12198	16914/455713	2	$8.75e^{-06}$	**$1.44e^{-03}$**
NR	GO:0097159	MF	organic cyclic compound binding	583/12198	16914/455713	2	$8.75e^{-06}$	**$1.44e^{-03}$**

RC	GO ID	Type	GO term	$\hat{p}(n)$	$\hat{p}(N)$	Depth	P	P_B
NR	GO:0003676	MF	nucleic acid binding	583/12198	16914/455713	3	$8.75e^{-06}$	**$1.44e^{-03}$**

secondary metabolism and toxin production processes (GO:0019748, GO:0009404; $P_B \leq 1.49e^{-04}$). A representative *A. fumigatus* gene from this this subset (AFUA_6G00680A; XP_731488.2) was reported to be a mycotoxin synthase with two oxidase ustYa domains (IPR021765). Oxidase ustYa domains catalyse the cyclisation of peptide secondary metabolites such as ustiloxins and cyclochlorotines (Umemura *et al*., 2014).

Secondary metabolite production is often expressed during the sexual cycle in Pezizomycotina, and are believed to protect spores from predation (Calvo *et al*., 2002). Interestingly, two nested composite *A. fumigatus* genes overrepresented for signal transduction and assigned to "Node_115", serine-threonine kinase (*rim1*, XP_755834.1, E.C:2.7.11.1.) and adenylate cyclase (*AcyA*, XP_750741.2, E.C: 4.6.1.1.), are involved in meiosis (Li and Mitchell, 1997; Kang *et al*., 2016). These results highlight the correlation between gene remodelling and the major phenotypic transition observed during the divergence of Pezizomycotina from Saccharomycotina.

2.3.8: Genes of bacterial origin are statistically more likely to be remodelled than eukaryotic-originating genes

Families of archaeal origin were not significantly enriched ($P \leq \alpha_B \leq 4.16e^{-03}$) for any RC (Table 2.3.22.). Bacterial originating ($P \leq 2.0e^{-03}$) and undefined prokaryote originating ($P \leq 0.999$) families were reported to be as significantly enriched as all RCs except NR. Families of eukaryote origin were observed to be enriched for NR ($P < 1.0e^{-04}$). Due to the considerable rarity of horizontal gene transfer between bacteria and eukaryotes (McCarthy and Fitzpatrick, 2016), it is likely that the vast majority of gene families with a bacterial or undefined prokaryote DO were vertically inherited. Therefore, these results suggest that "more ancient" gene families are more much statistically likely to be subjected to remodelling due to their persistence in the

Table 2.3.23: "Domains-of-Origin" for each remodelling category

Gene families (as defined by CompositeSearch) were assigned to a "Domain-of-Origin" (DO). The sum of genes (*n*) from each RC and their associated significance (*P*) for each DO are presented. Significant observations ($P \leq \alpha_B \leq 4.16e^{-03}$) are emboldened.

	Archaea		**Bacteria**		**Eukaryote**		**Undefined prokaryote**	
	n	***P***	***n***	***P***	***n***	***P***	***n***	***P***
NC	141	0.654	1319	**<1.0*e*$^{-04}$**	11425	>0.999	3917	**<1.0*e*$^{-04}$**
SC	27	0.029	162	**2.0*e*$^{-03}$**	1592	>0.999	326	**<1.0*e*$^{-04}$**
SN	208	0.043	1660	**<1.0*e*$^{-04}$**	16158	>0.999	3701	**<1.0*e*$^{-04}$**
NR	326	0.977	1879	>0.999	36000	**<1.0*e*$^{-04}$**	2605	>0.999

genome over time. It is important to state, however, that 65,175 of 81,446 families (80.02%) of genes were of eukaryote origin, and considerably more families of eukaryote origin were detected in each RC than in any other DO. These results suggest that while "older" gene families are more likely to remodel, considerable remodelling is observed in "younger" families also.

2.3.9. Trends between genomic characteristics and remodelling extent in genomes

RCPs were calculated (Tables 2.3.23.-24.) for each genome and compared to genomic characteristics (Table 2.3.3; Figure 2.3.5.). For GRCP, significant positive correlations ($P \leq \alpha_B \leq 0.01$) were observed between genome completeness (as defined by BUSCO) and (*a*) SN genomic proportions ($\rho = 0.4105$; $P < 1.0e^{-04}$) and (*b*) NR proportions ($\rho = 0.5298$; $P < 1.0e^{-04}$) respectively (Figure 2.3.6). A significant negative correlation was also observed between genome completeness and excluded proportions ($\rho = 0.5298$; $P < 1.0e^{-04}$) and between genome size and SN proportions ($\rho = -0.3713$; $P < 1.0e^{-04}$). These results suggest that using higher quality assemblies for gene remodelling analyses results in a greater number genes being assigned to strict component or non-remodelling families as opposed to being excluded (due to being defined as a singleton by CompositeSearch). This suggests that incomplete genome assemblies may result in Type II errors (false negative) in remodelled gene detection.

For IRCP, significant positive correlations ($P \leq \alpha_B \leq 0.01$) were observed between genome size and each RCP ($\rho = 0.3298$-0.6662; $P \leq 1.0e^{-04}$) except the "excluded" proportion where a significant negative correlation was observed ($\rho = -0.4591$; $P < 1.0e^{-04}$). This is likely due to genetic redundancy from events such as WGD and genomic expansions, which would promote the clustering of more non-singleton families into non-excluded families. Conversely,

Table 2.3.24. GRCPs for each fungal genome

GRCPs (%) were calculated for each genome by dividing the number of remodelled genes per RC (*n*) by the total number of genes in its respective genome.

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Acremonium alcalophilum	2818	291	1580	2003	2829	29.598	3.056	16.595	21.038	29.713
Agaricus bisporus	3721	265	2035	2624	2644	32.961	2.347	18.026	23.244	23.421
Allomyces macrogynus	3009	285	2668	5331	6307	17.097	1.619	15.159	30.290	35.835
Alternaria brassicicola	2820	260	1672	2318	3618	26.385	2.433	15.644	21.688	33.851
Ashbya gossypii	1323	119	974	1359	942	28.047	2.523	20.649	28.811	19.970
Aspergillus aculeatus	4487	381	1906	2687	1367	41.439	3.519	17.603	24.815	12.625
Aspergillus carbonarius	4778	378	2043	2825	1600	41.105	3.252	17.576	24.303	13.765
Aspergillus clavatus	3616	317	1832	2530	825	39.649	3.476	20.088	27.741	9.046
Aspergillus flavus	4643	374	2146	2963	2461	36.887	2.971	17.049	23.540	19.552
Aspergillus fumigatus	3792	322	1820	2441	1512	38.353	3.257	18.408	24.689	15.293
Aspergillus nidulans	4237	342	1928	2472	1581	40.123	3.239	18.258	23.409	14.972
Aspergillus oryzae	4569	312	2232	2665	2285	37.876	2.586	18.503	22.092	18.942
Aspergillus terreus	4058	323	1857	2266	1902	38.997	3.104	17.845	21.776	18.278
Auricularia delicata	6148	465	3527	4810	8627	26.076	1.972	14.959	20.401	36.591
Batrachochytrium dendrobatidis	1889	224	1291	1801	3527	21.633	2.565	14.785	20.625	40.392
Baudoinia compniacensis	3142	295	1740	2190	3146	29.887	2.806	16.551	20.831	29.925
Bjerkandera adusta	4729	366	2378	3608	4392	30.563	2.365	15.369	23.318	28.385
Blastomyces dermatitidis	3017	305	1677	2241	2282	31.685	3.203	17.612	23.535	23.966
Botryotinia cinerea	3212	252	1776	2641	8567	19.528	1.532	10.798	16.057	52.085
Candida albicans	1741	180	1256	1917	1111	28.058	2.901	20.242	30.894	17.905
Candida caseinolytica	1340	127	832	936	1422	28.774	2.727	17.866	20.099	30.535
Candida glabrata	1444	128	1066	1627	937	27.759	2.461	20.492	31.276	18.012
Candida tenuis	1716	125	1164	1756	772	31.014	2.259	21.037	31.737	13.953
Ceriporiopsis subvermispora	4053	292	1941	2601	3238	33.427	2.408	16.008	21.452	26.705
Chaetomium globosum	3460	285	1674	2249	3456	31.104	2.562	15.049	20.218	31.068
Coccidioides immitis	2975	306	1773	2960	2640	27.924	2.872	16.642	27.783	24.779
Coccidioides posadasii	2672	275	1719	2852	2606	26.393	2.716	16.979	28.171	25.741
Cochliobolus heterostrophus	3848	338	1976	2702	769	39.946	3.509	20.513	28.049	7.983
Cochliobolus sativus	4247	390	2236	3420	1957	34.669	3.184	18.253	27.918	15.976
Coniophora putinea	4683	340	2170	2875	3693	34.031	2.471	15.769	20.892	26.837
Coprinopsis cinerea	3501	296	1979	2926	4692	26.139	2.210	14.775	21.846	35.031

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Cryphonectria parasitica	3969	333	1907	2460	2515	35.488	2.977	17.051	21.996	22.487
Cryptococcus neoformans	1916	161	1141	1632	2117	27.501	2.311	16.377	23.425	30.386
Debaryomyces hansenii	1855	154	1283	2032	4918	18.112	1.504	12.527	19.840	48.018
Dichomitus squalens	4379	321	2195	2974	2421	69.818	5.118	34.997	47.417	19.699
Dothistroma septosporum	3580	332	1956	2632	4918	29.129	2.701	15.915	21.416	32.432
Dsppyopinax sp.	2907	235	1507	1903	1903	23.108	1.868	11.979	15.127	36.028
Fomitiporia mediterranea	3600	286	1799	2412	3236	31.766	2.524	15.874	21.283	28.554
Fomitopsis pinicola	4744	390	2570	3374	3646	32.220	2.649	17.454	22.915	24.762
Fusarium graminearum	4510	387	2277	3241	2906	33.856	2.905	17.093	24.330	21.815
Fusarium oxysporum	5777	457	3389	4790	3195	32.809	2.595	19.247	27.204	18.145
Fusarium verticillioides	4640	355	2603	3638	2959	32.688	2.501	18.337	25.629	20.845
Ganoderma sp.	4431	357	2185	3178	2759	34.322	2.765	16.925	24.617	21.371
Gloeophyllum trabeum	3713	345	2200	2818	2770	31.344	2.912	18.572	23.789	23.383
Hansenula polymorpha	1556	124	970	1213	1314	30.056	2.395	18.737	23.431	25.381
Heterobasidion annosum	3466	261	1901	2333	4338	28.181	2.122	15.457	18.969	35.271
Histoplasma capsulatum	2393	256	1445	1935	3222	25.867	2.767	15.620	20.917	34.829
Hysterium pulicare	3898	340	1846	2485	3783	31.558	2.753	14.945	20.118	30.627
Laccaria bicolor	3852	367	2727	3858	8232	20.235	1.928	14.325	20.267	43.244
Leptosphaeria maculans	3000	285	1671	2300	5213	24.060	2.286	13.401	18.446	41.808
Lipomyces starkeyi	2215	232	1254	1571	2920	27.039	2.832	15.308	19.177	35.645
Magnaporthe grisea	3412	297	1651	2092	3657	30.714	2.674	14.862	18.832	32.919
Melampsora laricis_populina	3379	356	2385	3562	7149	20.076	2.115	14.170	21.163	42.475
Microsporum canis	3126	314	1680	2386	1259	35.665	3.582	19.167	27.222	14.364
Microsporum gypseum	3045	301	1649	2442	1439	34.306	3.391	18.578	27.512	16.212
Mucor circinelloides	2989	234	1845	2766	3096	27.347	2.141	16.880	25.306	28.326
Mycosphaerella fijiensis	3344	292	1974	2395	2308	32.425	2.831	19.141	23.223	22.380
Mycosphaerella graminicola	3414	300	1836	2301	3082	31.227	2.744	16.793	21.046	28.190
Nectria haematococca	6447	484	2634	3556	2586	41.045	3.081	16.770	22.640	16.464
Neosartorya fischeri	4127	356	2104	2871	945	39.671	3.422	20.225	27.598	9.084
Neurospora crassa	3147	324	1858	3349	1230	31.762	3.270	18.753	33.801	12.414
Neurospora tetrasperma	2958	292	1736	3164	2490	27.801	2.744	16.316	29.737	23.402
Paracoccidioides brasiliensis	2458	260	1399	1855	3164	26.905	2.846	15.313	20.304	34.632
Phanerochaete chrysosporium	3428	276	1790	2249	2305	34.116	2.747	17.814	22.383	22.940
Phlebia brevispora	5092	403	2698	3604	4373	31.490	2.492	16.685	22.288	27.044
Phlebiopsis gigantea	3630	296	2057	2720	3188	30.527	2.489	17.299	22.874	26.810
Phycomyces blakesleeanus	5109	299	2488	3075	5557	30.911	1.809	15.053	18.605	33.622
Pichia membranifaciens	1422	132	921	1231	1840	25.640	2.380	16.607	22.196	33.177
Pichia stipitis	1885	151	1270	1912	589	32.461	2.600	21.870	32.926	10.143
Pleurotus ostreatus	3671	321	2051	2833	2727	31.638	2.767	17.676	24.416	23.503
Podospora anserina	3600	321	1799	2412	2469	33.959	3.028	16.970	22.753	23.290

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Puccinia graminis	3360	327	2139	3003	11737	16.338	1.590	10.401	14.602	57.070
Punctularia strigosozonata	3909	308	1950	2697	2674	33.879	2.669	16.901	23.375	23.176
Pyrenophora teres	3869	358	2242	3465	1865	32.791	3.034	19.002	29.367	15.806
Pyrenophora triticirepentis	3730	357	2179	3456	2447	30.652	2.934	17.906	28.400	20.108
Rhizopus oryzae	4914	272	2673	2971	6629	28.146	1.558	15.310	17.017	37.969
Rhodotorula graminis	1816	168	1163	1561	2575	24.935	2.307	15.969	21.433	35.356
Rhytidhysteron rufulum	4127	313	1881	2455	3341	34.060	2.583	15.524	20.261	27.573
Saccharomyces cerevisiae	1668	135	1228	1805	1049	28.343	2.294	20.867	30.671	17.825
Schizophyllum commune	3667	300	2091	2964	4159	27.820	2.276	15.864	22.487	31.553
Schizosaccharomyces cryophilus	1565	134	1209	1866	283	30.947	2.650	23.907	36.899	5.596
Schizosaccharomyces japonicus	1428	130	997	1390	869	29.663	2.700	20.710	28.874	18.052
Schizosaccharomyces octosporus	1568	156	1197	1905	99	31.838	3.168	24.305	38.680	2.010
Schizosaccharomyces pombe	1568	136	1171	1759	376	31.297	2.715	23.373	35.110	7.505
Sclerotinia sclerotiorum	3349	265	1798	2641	6469	23.062	1.825	12.381	18.186	44.546
Septoria musiva	3312	324	1995	3159	1443	32.366	3.166	19.496	30.871	14.101
Septoria populicola	3120	306	1980	3045	1288	32.036	3.142	20.331	31.266	13.225
Serpula lacrymans	4309	271	2404	2616	4895	29.727	1.870	16.585	18.048	33.770
Setosphaeria turcica	4055	375	2123	3253	1896	34.652	3.205	18.142	27.799	16.202
Spathaspora passalidarum	1717	141	1235	1883	1007	28.698	2.357	20.642	31.473	16.831
Spizellomyces punctatus	1616	153	1024	1548	4463	18.355	1.738	11.631	17.583	50.693
Sporobolomyces roseus	1493	126	902	1101	1914	26.969	2.276	16.293	19.888	34.574
Sporotrichum thermophile	3128	338	1819	2412	1109	35.521	3.838	20.656	27.390	12.594
Thielavia terrestris	3319	341	1832	2525	1798	33.816	3.474	18.665	25.726	18.319
Trametes versicolor	5031	355	2548	3455	2907	35.192	2.483	17.823	24.168	20.334
Tremella mesenterica	2047	162	1220	1734	3150	24.624	1.949	14.676	20.859	37.892
Trichoderma atroviride	4308	345	2161	2881	1405	38.811	3.108	19.468	25.955	12.658
Trichoderma reesei	3565	316	1853	2612	797	38.992	3.456	20.267	28.568	8.717
Trichoderma virens	4574	377	2283	3111	1298	39.285	3.238	19.608	26.720	11.148
Trichophyton equinum	2738	286	1563	2239	1734	31.986	3.341	18.259	26.157	20.257
Uncinocarpus reesii	2520	240	1417	1851	1770	32.316	3.078	18.171	23.737	22.698
Ustilago maydis	1393	116	735	918	3360	21.358	1.779	11.270	14.075	51.518
Verticillium alboatrum	3120	267	1721	2388	2724	30.528	2.613	16.840	23.366	26.654
Verticillium dahliae	3680	326	1939	2788	1802	34.931	3.094	18.405	26.464	17.105
Wickerhamomyces anomalus	1954	149	1198	1534	1588	30.422	2.320	18.652	23.883	24.724
Wolfiporia cocos	3993	323	2163	2859	3408	31.327	2.534	16.970	22.431	26.738
Yarrowia lipolytica	1737	169	961	1309	2272	26.939	2.621	14.904	20.301	35.236

Table 2.3.25. IRCPs for each fungal genome

IRCPs (%) were calculated for each genome by dividing the number of remodelled genes per RC (*n*) by the total number of genes in its respective genome.

	n					**%**				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Acremonium alcalophilum	19	18	41	898	8545	0.200	0.189	0.431	9.432	89.749
Agaricus bisporus	448	235	458	2499	7649	3.968	2.082	4.057	22.137	67.756
Allomyces macrogynus	624	183	909	8659	7225	3.545	1.040	5.165	49.199	41.051
Alternaria brassicicola	0	15	27	952	9694	0	0.140	0.253	8.907	90.700
Ashbya gossypii	0	0	0	415	4302	0	0	0	8.798	91.202
Aspergillus aculeatus	146	63	295	2508	7816	1.348	0.582	2.724	23.162	72.183
Aspergillus carbonarius	287	53	429	3297	7558	2.469	0.456	3.691	28.364	65.021
Aspergillus clavatus	37	25	119	1590	7349	0.406	0.274	1.305	17.434	80.581
Aspergillus flavus	240	71	275	2525	9476	1.907	0.564	2.185	20.060	75.284
Aspergillus fumigatus	24	52	139	1626	8046	0.243	0.526	1.406	16.446	81.380
Aspergillus nidulans	79	60	179	2123	8119	0.748	0.568	1.695	20.104	76.884
Aspergillus oryzae	183	138	377	2584	8781	1.517	1.144	3.125	21.421	72.793
Aspergillus terreus	79	67	182	1932	8146	0.759	0.644	1.749	18.566	78.282
Auricularia delicata	1486	443	1577	6548	13523	6.303	1.879	6.689	27.773	57.357
Batrachochytrium dendrobatidis	394	131	385	1367	6455	4.512	1.500	4.409	15.655	73.923
Baudoinia compniacensis	46	11	94	1240	9122	0.438	0.105	0.894	11.795	86.769
Bjerkandera adusta	355	281	447	3898	10492	2.294	1.816	2.889	25.192	67.808
Blastomyces dermatitidis	195	17	117	848	8345	2.048	0.179	1.229	8.906	87.639
Botryotinia cinerea	30	63	98	1317	14940	0.182	0.383	0.596	8.007	90.832
Candida albicans	0	33	18	924	5230	0	0.532	0.290	14.891	84.287
Candida caseinolytica	4	6	13	486	4148	0.086	0.129	0.279	10.436	89.070
Candida glabrata	0	0	0	801	4401	0	0	0	15.398	84.602
Candida tenuis	5	0	7	886	4635	0.090	0	0.127	16.013	83.770
Ceriporiopsis subvermispora	327	135	315	2870	8478	2.697	1.113	2.598	23.670	69.922
Chaetomium globosum	173	18	128	1115	9690	1.555	0.162	1.151	10.023	87.109
Coccidioides immitis	9	0	6	1225	9414	0.084	0	0.056	11.498	88.361
Coccidioides posadasii	0	6	24	850	9244	0	0.059	0.237	8.396	91.308
Cochliobolus heterostrophus	129	27	154	1892	7431	1.339	0.280	1.599	19.641	77.141
Cochliobolus sativus	122	40	200	2113	9775	0.996	0.327	1.633	17.249	79.796

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Coniophora putinea	428	175	509	3986	8663	3.110	1.272	3.699	28.966	62.953
Coprinopsis cinerea	245	57	241	2832	10019	1.829	0.426	1.799	21.144	74.802
Cryphonectria parasitica	119	38	150	2120	8757	1.064	0.340	1.341	18.956	78.299
Cryptococcus neoformans	21	13	48	799	6086	0.301	0.187	0.689	11.468	87.355
Debaryomyces hansenii	15	0	5	1054	9168	0.146	0	0.049	10.291	89.514
Dichomitus squalens	191	130	303	3110	2538	3.045	2.073	4.831	49.585	40.466
Dothistroma septosporum	50	24	106	1743	10367	0.407	0.195	0.862	14.182	84.353
Dsppyopinax sp.	57	51	107	2195	10170	0.453	0.405	0.851	17.448	80.843
Fomitiporia mediterranea	194	166	336	2519	8118	1.712	1.465	2.965	22.227	71.632
Fomitopsis pinicola	560	208	548	3792	9616	3.803	1.413	3.722	25.754	65.308
Fusarium graminearum	44	47	119	2545	10566	0.330	0.353	0.893	19.105	79.318
Fusarium oxysporum	533	272	741	4467	11595	3.027	1.545	4.208	25.369	65.851
Fusarium verticillioides	214	91	240	2423	11227	1.508	0.641	1.691	17.069	79.091
Ganoderma sp.	278	215	456	2910	9051	2.153	1.665	3.532	22.541	70.108
Gloeophyllum trabeum	123	148	254	2730	8591	1.038	1.249	2.144	23.046	72.522
Hansenula polymorpha	0	2	7	763	4405	0	0.039	0.135	14.738	85.088
Heterobasidion annosum	244	81	247	2311	9416	1.984	0.659	2.008	18.790	76.559
Histoplasma capsulatum	0	4	10	641	8596	0	0.043	0.108	6.929	92.920
Hysterium pulicare	81	54	150	1906	10161	0.656	0.437	1.214	15.431	82.262
Laccaria bicolor	485	175	548	4122	13706	2.548	0.919	2.879	21.654	72.000
Leptosphaeria maculans	15	11	24	960	11459	0.120	0.088	0.192	7.699	91.900
Lipomyces starkeyi	49	60	132	1468	6483	0.598	0.732	1.611	17.920	79.138
Magnaporthe grisea	312	48	284	1317	9148	2.809	0.432	2.556	11.855	82.348
Melampsora laricis_populina	1318	207	940	3818	10548	7.831	1.230	5.585	22.684	62.670
Microsporum canis	25	13	62	1248	7417	0.285	0.148	0.707	14.238	84.621
Microsporum gypseum	22	14	48	1171	7621	0.248	0.158	0.541	13.193	85.861
Mucor circinelloides	236	74	256	3409	6955	2.159	0.677	2.342	31.189	63.632
Mycosphaerella fijiensis	51	15	81	2061	8105	0.495	0.145	0.785	19.984	78.590
Mycosphaerella graminicola	46	28	132	1802	8925	0.421	0.256	1.207	16.482	81.634
Nectria haematococca	493	139	675	4423	9977	3.139	0.885	4.297	28.159	63.519
Neosartorya fischeri	77	51	165	2178	7932	0.740	0.490	1.586	20.936	76.247
Neurospora crassa	9	12	36	1193	8658	0.091	0.121	0.363	12.041	87.384
Neurospora tetrasperma	0	12	29	878	9721	0	0.113	0.273	8.252	91.363
Paracoccidioides brasiliensis	6	39	47	558	8486	0.066	0.427	0.514	6.108	92.885
Phanerochaete chrysosporium	272	81	318	2303	7074	2.707	0.806	3.165	22.920	70.402
Phlebia brevispora	814	343	794	3971	10248	5.034	2.121	4.910	24.558	63.377
Phlebiopsis gigantea	58	40	197	2462	9134	0.488	0.336	1.657	20.705	76.814

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Phycomyces blakesleeanus	2472	167	939	3598	9352	14.956	1.010	5.681	21.769	56.583
Pichia membranifaciens	0	0	0	689	4857	0	0	0	12.423	87.577
Pichia stipitis	5	6	16	1077	4703	0.086	0.103	0.276	18.547	80.988
Pleurotus ostreatus	65	91	231	3259	7957	0.560	0.784	1.991	28.088	68.577
Podospora anserina	57	32	129	1350	9033	0.538	0.302	1.217	12.735	85.209
Puccinia graminis	1698	194	1021	3070	14583	8.256	0.943	4.965	14.928	70.908
Punctularia strigosozonata	235	140	302	2853	8008	2.037	1.213	2.617	24.727	69.405
Pyrenophora teres	97	56	181	1711	9754	0.822	0.475	1.534	14.501	82.668
Pyrenophora triticirepentis	108	58	140	1652	10211	0.888	0.477	1.150	13.575	83.910
Rhizopus oryzae	2107	205	1351	4087	9709	12.068	1.174	7.738	23.409	55.610
Rhodotorula graminis	0	7	12	1070	6194	0	0.096	0.165	14.692	85.047
Rhytidhysteron rufulum	180	29	288	1876	9744	1.486	0.239	2.377	15.482	80.416
Saccharomyces cerevisiae	67	4	62	1225	4527	1.138	0.068	1.054	20.816	76.924
Schizophyllum commune	177	75	329	3432	9168	1.343	0.569	2.496	26.037	69.555
Schizosaccharomyces cryophilus	0	5	9	789	4254	0	0.099	0.178	15.602	84.121
Schizosaccharomyces japonicus	0	2	24	709	4079	0	0.042	0.499	14.728	84.732
Schizosaccharomyces octosporus	2	50	11	781	4081	0.041	1.015	0.223	15.858	82.863
Schizosaccharomyces pombe	0	0	0	861	4149	0	0	0	17.186	82.814
Sclerotinia sclerotiorum	173	29	214	1203	12903	1.191	0.200	1.474	8.284	88.851
Septoria musiva	53	30	62	1408	8680	0.518	0.293	0.606	13.759	84.824
Septoria populicola	0	11	24	1377	8327	0	0.113	0.246	14.139	85.502
Serpula lacrymans	940	154	693	2450	10258	6.485	1.062	4.781	16.902	70.769
Setosphaeria turcica	137	33	218	1902	9412	1.171	0.282	1.863	16.254	80.431
Spathaspora passalidarum	0	4	4	933	5042	0	0.067	0.067	15.594	84.272
Spizellomyces punctatus	23	19	52	890	7820	0.261	0.216	0.591	10.109	88.823
Sporobolomyces roseus	0	8	16	754	4758	0	0.145	0.289	13.620	85.947
Sporotrichum thermophile	34	38	119	1272	7343	0.386	0.432	1.351	14.445	83.386
Thielavia terrestris	43	6	85	1388	8293	0.438	0.061	0.866	14.142	84.493
Trametes versicolor	440	284	833	3829	8910	3.078	1.987	5.827	26.784	62.325
Tremella mesenterica	272	19	86	882	7054	3.272	0.229	1.035	10.610	84.855
Trichoderma atroviride	165	51	358	2185	8341	1.486	0.459	3.225	19.685	75.144
Trichoderma reesei	10	33	102	1530	7468	0.109	0.361	1.116	16.734	81.680
Trichoderma virens	141	87	226	2677	8512	1.211	0.747	1.941	22.992	73.108
Trichophyton equinum	11	2	19	956	7572	0.129	0.023	0.222	11.168	88.458
Uncinocarpus reesii	8	12	28	768	6982	0.103	0.154	0.359	9.849	89.536
Ustilago maydis	7	8	28	476	6003	0.107	0.123	0.429	7.298	92.042

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Verticillium alboatrum	67	35	121	1020	8977	0.656	0.342	1.184	9.980	87.838
Verticillium dahliae	59	44	157	1548	8727	0.560	0.418	1.490	14.694	82.838
Wickerhamomyces anomalus	11	11	40	1413	4948	0.171	0.171	0.623	21.999	77.036
Wolfiporia cocos	259	191	506	2687	9103	2.032	1.499	3.970	21.081	71.418
Yarrowia lipolytica	10	3	16	1118	5301	0.155	0.047	0.248	17.339	82.212

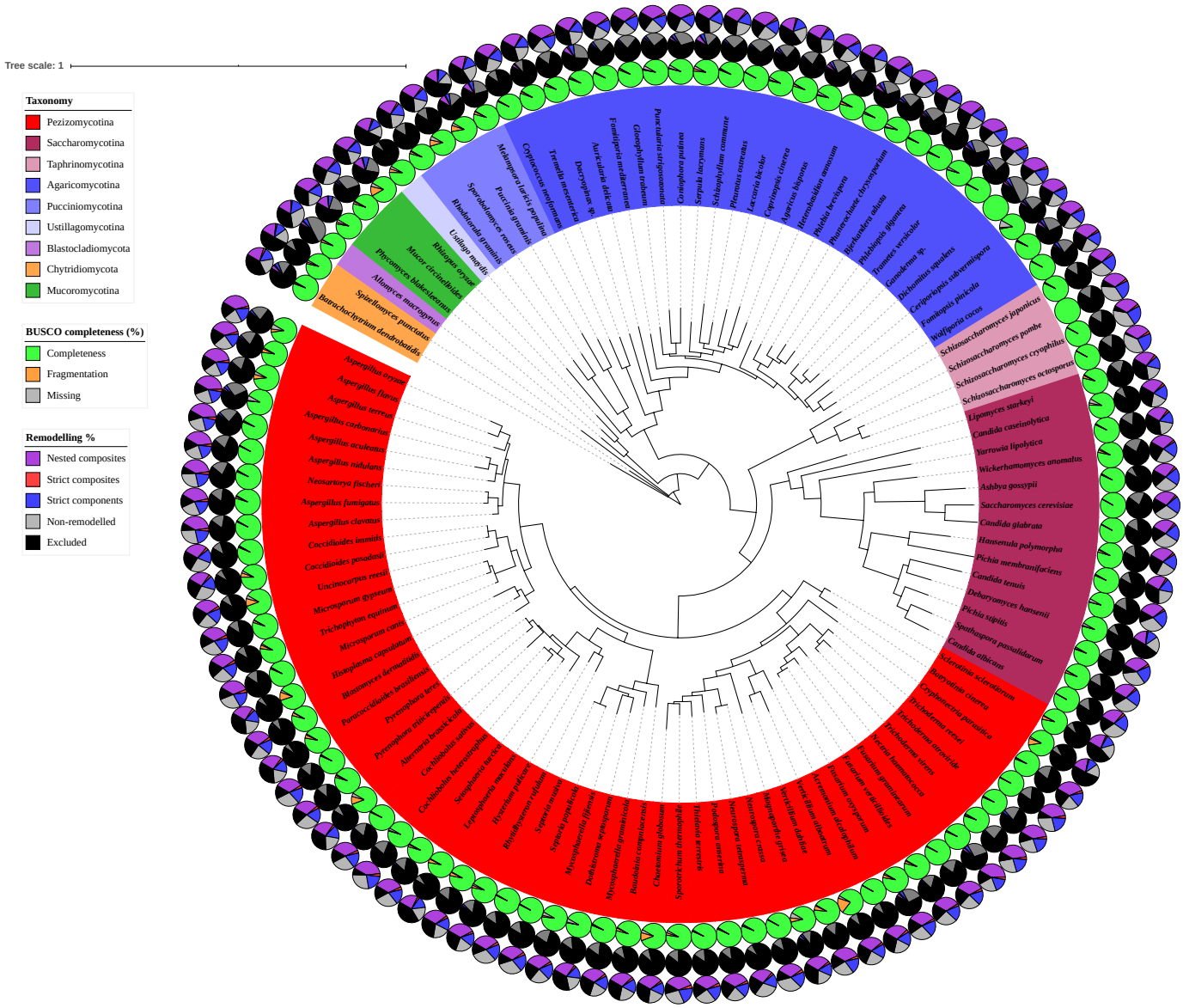

Figure 2.3.5. Representative fungal phylogeny annotated with genomic characteristics

The representative phylogeny (Figure 2.3.3.) was annotated with three datasets, each represented as a pie charts. The internal pie chart illustrates genomic completeness (Table 2.3.3.) for each species where completeness is annotated as green, fragmentation as orange, and missing BUSCO genes as grey. Middle pie charts represent IRCP for each species (Table 2.3.24) and external pie charts represent GRCPs for each species (Table 2.3.23). The 5 RCPs were annotated purple for NC, red for SC, blue for SN, grey for NR, and black for excluded.

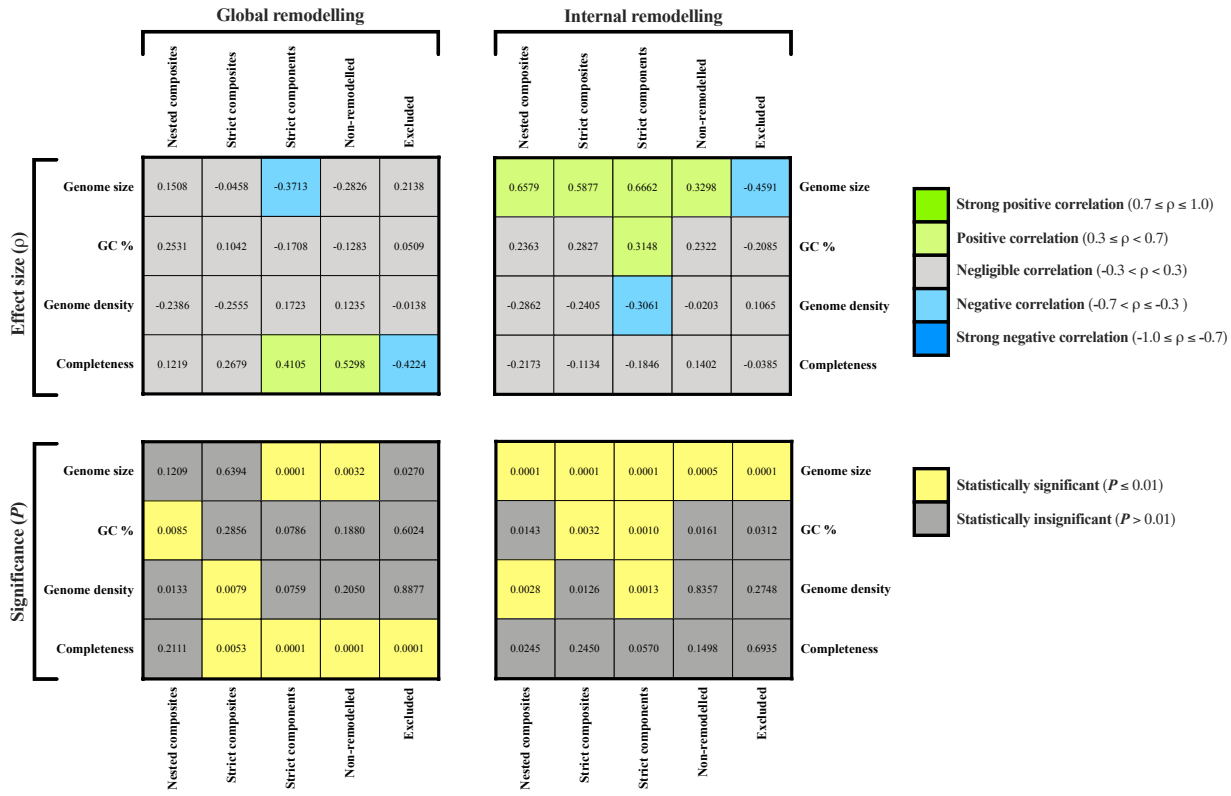

Figure 2.3.6. *Correlation matrix between genomic characteristics and genomic remodelling extent*

Illustration of effect sizes (Spearman's ρ) and their significance (*P*) between genomic characteristics for (*a*) the globally remodelled dataset and *(b*) the internally remodelled dataset. Each data category is annotated as per the legend. Data calculated from GRCP (Table 2.3.23.) is presented on the left, and from IRCP (Table 2.3.24.) on the right.

a negative correlation was observed between genome density and SN proportions ($\rho = -0.3061$; $P = 1.3e^{-04}$) when sampled from IRCP. This was not expected due to the positive correlations usually observed between genome size and genome expansion events (Fischer *et al*., 2014). However, as a significant negative correlation ($\rho = 0.3713$; $P < 1.0e^{-04}$) was observed between genome size and SN proportions from GRCP, it could be a case of larger fungal genomes are being likely to have undergone polyploidisation (Kellis *et al.*, 2004) which may promote the transition of strict components to nested composites. This would likely result in a negative correlation for these categories.

2.3.10: A case of gene remodelling in *Batrachochytrium dendrobatidis*

A cluster of family cliques (40 genes from 7 families) were detected during the speciation of *Batrachochytrium dendrobatidis* (Chytridiomycota; Figures 2.3.7-2.3.8). Of these, 4 were strict composites, and 3 were strict components. In each instance, where two families shared homology, a complete bipartite graph was observed (whether homology was detected as a consequence of remodelling or not). The completeness of this graph suggests that these genes were associated with rapid or recent remodelling. As only a single gene (from any family) was identified in another genome in our dataset (*Spizellomyces punctatus,* Chytridiomycota) and how all other genes were only identified in *B. dendrobatidis* it could be suggested that these genes were duplicated and expanded to fulfil a specific niche in its lifecycle. Each gene in the four composite families shared a peptidase family s41 domain (PF03572), which are reported to be potent virulence factors against amphibian hosts (Thekkiniath *et al.,* 2013). An identifiable PFAM domain was not detected in any strict component gene.

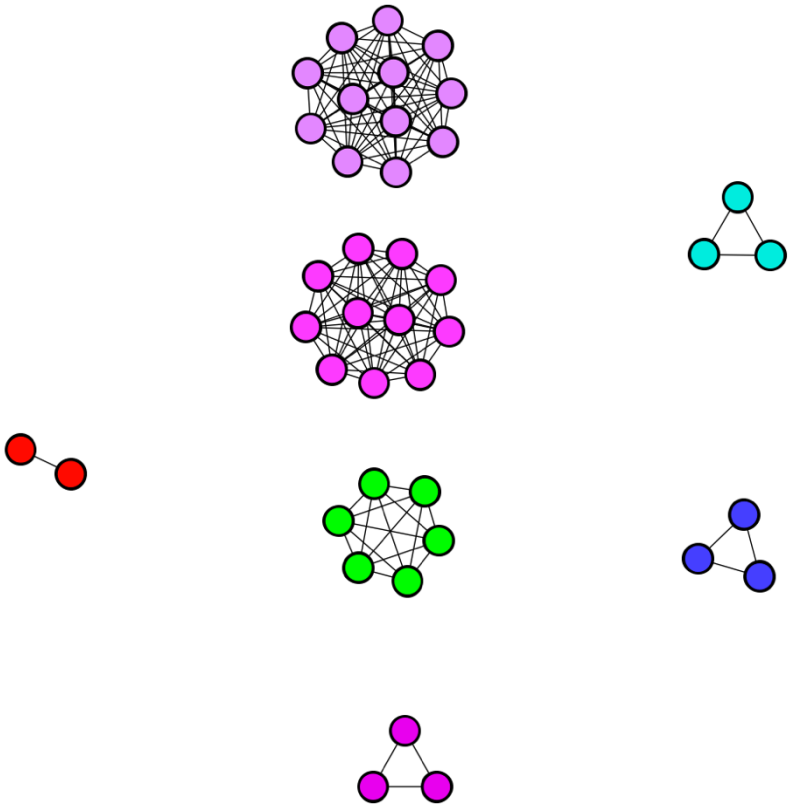

Figure 2.3.7: Inter-familial relationships between a subset of genes in *B. dendrobatidids*

Families that have full connectivity between each vertex (cliques) in *Batrachochytrium dendrobatidis*. The green composite family contains one gene from *Spizellomyces punctatus*, all other families are specific to *B. dendrobatidis*. Purple families are also composite families. Blue component families share homology between each other and to the C-terminus of each composite family. Red component families share homology with the N-terminus of each composite family.

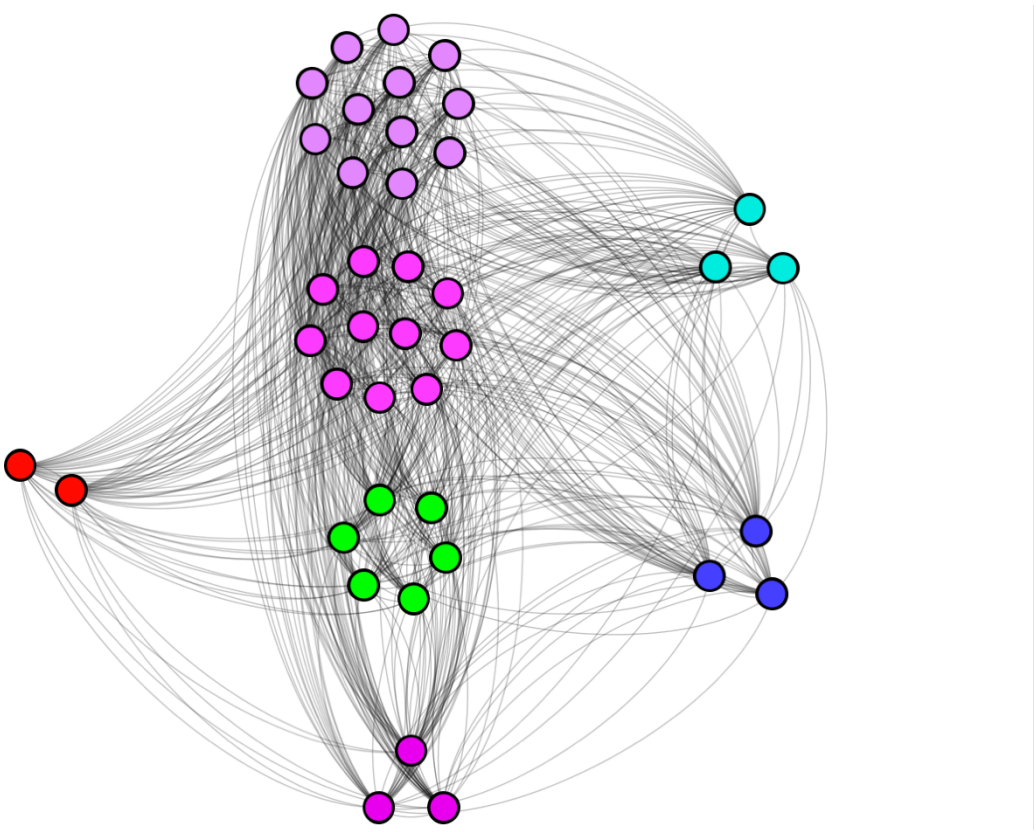

Figure 2.3.8. Relationship between remodelled *B. dendrobatidis* cliques.

Each composite family forms a clique with each other composite family and each component family. Both blue component families form a clique with each other. Red component families form a clique with each composite family. Composite families serve as articulation points between red and blue component families.

2.3.11: A potential case of fission mediated subneofunctionalsation in *Saccharomyces cerevisiae*

Trp1 (SPBC1539.09c) is an essential multifunctional enzyme in *Schizosaccharomyces pombe* where it is involved in tryptophan biosynthesis (Thuriaux et al., 1982). *Trp1* was reported as a composite by CompositeSearch where two non-homologous components were found in *S. cerevisiae* (*trp3* (YKL211C) and *trp1* (YDR007W)). This remodelling event was found to be a whole gene fusion/fission event by *fdf*BLAST. *S. pombe Trp1* was found to be a composite of YKL211C at along N-terminus and YDR007W at its C-terminus (Figure 2.3.9.). While this composite was also detected by Leonard & Richards in 2010 it is not amongst the list of fusion proteins on Pombase (McDowall et al., 2015). YKL211C and YDR007W are both non-essential, null mutants are tryptophan auxotrophic and perform distinct functions during tryptophan biosynthesis, all of which are performed by *Trp1* in *S. pombe* (Miozzari *et al*., 1978; Thuriaux *et al.*, 1982).

2.4. Discussion

2.4.1 Gene remodelling is extensive in fungi

Our results illustrate the extent of sequence remodelling throughout the evolution of fungi, with 49.94% of all sequences in our dataset having a history of some form of remodelling in that timeframe. In addition, we found that composite sequences constitute 32.76% of all sampled genes. The primary mechanisms of sequence evolution are duplication, recombination, and divergence (Dittmar and Liberles 2011; van Rijk and Bloemendal 2003; Leonard and Richards 2012; Przytycka *et al*. 2006; Patthy 1999) all of which are correlated

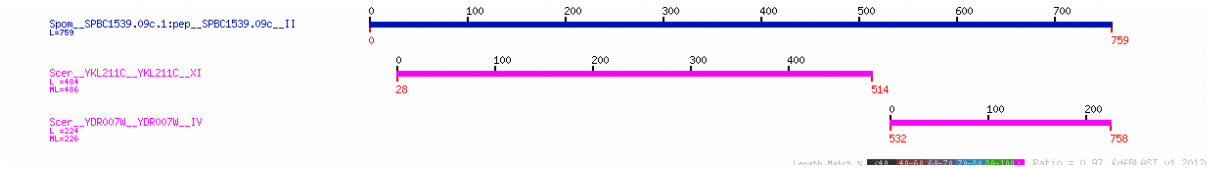

Figure 2.3.9: *Composite tryptophan biosynthesis gene*

Differential fusion gene (*Trp1*; SPB1539.09c) observed in *Schizosaccharomyces pombe*, each component gene (*trp3* (YKL211C) and *trp1* (YDR007W)) is observed in *Saccharomyces cerevisiae*. A copy of the fusion gene was not observed in *S. cerevisiae* and a copy of neither component gene was found in *S. pombe*. This image was produced using *fdf*BLAST (Leonard and Richards, 2012)

with each other to varying degrees (Vogel *et al.* 2005). To our knowledge, the only other published research to look at global scale remodelling was during the benchmarking of CompositeSearch and of FusedTriplets, where 21,623 genes out of 204,894 (10.6%) in a viral dataset were reported as composite genes (Pathmanathan *et al.* 2018; Jachiet *et al.* 2014). The sum of composites in fungi has previously been estimated to be approximately 4% and the sum of fusion components to be approximately 9% (Durrens *et al.* 2008; Enright & Ouzounis 2001). A study of nine Drosophila species reported approximately 9,000 sequences (approximately of fusion components to be approximately 9% (Durrens *et al.* 2008; Enright & Ouzounis 2001). A study of nine Drosophila species reported approximately 9,000 sequences (approximately 6.62%) that are the product of domain recombination (Wu *et al.* 2012). Here, using network analysis, we report levels of remodelling that are considerably greater than previously reported.

2.4.2. Composite genes are highly homoplastic

It was observed that remodelled families were likely to be homoplastic ($P \leq 0.002$), larger in comparison to other remodelling categories ($P \leq 7.10e^{-03}$), statistically enriched to be of prokaryote origin ($P \leq 2.0e^{-03}$) and unlikely to be of eukaryote origin ($P>0.999$). There are a few possible explanations for these two findings. First of all, the larger family size among composites could be due to the CompositeSearch software clustering epaktologs into families. Epaktologs are sequences that have independently acquired the same domain architecture, meaning that, strictly speaking they are neither orthologs nor paralogs (Haggerty et al. 2014; Nagy & Patthy 2011). Larger composite family size is an indication that particular kinds of protein more “successful” or “useful”, in other words, more likely to arise in the first place and then more likely to persist throughout evolution, once they arise. This would cause composite families to be larger than non-composite families on average and also more likely to arise more

than once. Composite families are also more likely to be multifunctional (Pasek et al. 2006), and therefore, there is less of a requirement to independently transcribe and translate two different genes, when instead they can be combined into a single gene. This might influence the retention of composite proteins over non-composites. The observation of prokaryote DO enrichments across all remodelled categories (except NR) is in agreement with other studies highlighting higher duplication rates and expendability of more ancient Prokaryote orthologs in Eukaryotes (Jordan *et al.*, 2002; Cotton and McInerney, 2010; Alvarez-Ponce *et al.*, 2013; Luo *et al.*, 2015). Synapomorphy gain and loss patterns were compared between leaves and branches using multiple Mann-Whitney *U* tests, where significant differences were observed for all categories ($P \leq 1.19e^{-03}$) suggesting that remodelling (and evolution) occurs at a more rapid rate during speciation. Only three branches were observed to have an evolutionary burst, the *C. immitis* speciation branch ($P \leq 2.23e^{-04}$), the *F. verticillioides* speciation branch ($P \leq 1.71e^{-04}$), and "Node_196" ($P = 3.26e^{-04}$). The burst at *C. immitis* is likely due to its exceptionally short branch length ($\kappa = 0.000517$) and the bursts in "Node_196" and *F. verticillioides* are likely due to large genomic expansions during the divergence from *F. graminearum*.

These results further corroborate with hypothesis that composites are born at a relatively clocklike rate, hence why an insignificant difference in birth rate is observed, and why greater proportions of composite families are observed at leaf nodes on average. These results are further compounded by the fact that remodelled genes are more likely to have arisen in ancient bacterial lineages ($P \leq 0.0001$) allowing for more time for homoplasy to occur.

2.4.3. Composite genes emerging at the root of Pezizomycotina are involved in typical Pezizomycotina phenotypes

Pezizomycotina were the largest clade in our dataset. GO slim terms from composite genes that were reported to have emerged at the root of Pezizomycotina were enriched against a subset of all genes from Pezizomycotina resulting in the detection of several significant ($P_B \leq 0.05$) biological process enrichments (GO:0005975 (carbohydrate metabolic process), GO:0007049 (cell cycle), GO:006259 (DNA metabolic process), GO:0019748 (secondary metabolic process), and GO:0009404 (toxin metabolic process)). During the emergence of Pezizomycotina, the shift to a predominantly multicellular clade, with a predominantly sexual lifecycle and a wide effector arsenal was observed (Liu and Hall, 2004; Arvas *et al.*, 2007). Cytokinesis ontologs were also enriched. Cytokinesis is the physical process of cell division and is essential for the proliferation of a cell line. Despite its importance, ubiquity and partial uniformity of steps involved, cytokinesis displays lineage specific differences in its co-ordination, highly dependent on organism complexity, specifically the development of the division plane (Canman *et al.*, 2003). During the evolution of Ascomycota, Taphrinomycotina and Sacchromycotina developed synapomorphic mechanisms to optimise nuclear segregation (Khmelinskii *et al.*, 2010). A paradigm shift in cytokinetic machinery seems to have coincided with the shift to a predominantly multicellular lifecycle during the evolution of Pezizomycotina. Enrichments of secondary metabolite processes (GO:0019748) and toxin metabolic processes (GO:0009404) coincide the expansions of effector arsenals during the evolution of early Pezizomycotina to potent pathogens (Cavalier-Smith, 1992; Demain and Fang, 2000; Arvas *et al.*, 2007).

2.4.4. Remodelled genes are likely to be involved in transport, whereas non-remodelled genes tend to be involved in housekeeping processes

When all genes in each remodelling category were searched for enrichment, eleven significantly enriched nested composite ontologies could be directly associated with small molecule transport.These results are unsurprising as these ontologies are all directly associated with the Major Facilitator Superfamily (MFS) (CL0015) of transporters or proteins that aid in their function (Madej, 2014; Yan, 2015). MFS is one of the two most abundant membrane bound transporter families in biology (Nelissen *et al.*, 2006) where it is ubiquitously distributed across cellular life and has been observed in a multitude of domain architecture combinations, where different combinations permit the transport of diverse small molecule repertoire (Madej, 2014). Significantly enriched non-remodelled ontologies were associated with housekeeping functions, specifically transcription, ribosomal structure, and base metabolic precursor generation. RNA transcription is a finely tuned, ubiquitous process. Transcription factor mutation is usually highly deleterious and altered dosage interferes with epigenetic controls (Poveda *et al.*, 2010). It is unsurprising that genes related to such a selectively pressurised process would be unaffected by remodelling.

2.4.5. Virulence factors are remodelled in *Batrachochytrium dendrobatidis*

Each gene in the four composite families in the cluster of complete bipartite graphs from *B. dendrobatidis* shared a peptidase family serine 41 domain (PF03572). The relevance of S41 peptidases in *Batrachochytrium dendrobatidis* have been previously described (Thekkiniath *et al.*, 2013), where it was concluded after an exhaustive search of approximately 6,000 proteomes that a considerable expansion of S41 peptidases was observed in *B. dendrobatidis.* The consequence of *B. dendrobatidis* infection on tadpoles is usually non-lethal keratinization of mouthparts. Metamorphosis in frogs is governed by 3, 5, 3′-triiodothyronine (T_3), after exposure to T_3 *Batrachochytrium dendrobatidis* zoospores display considerable

chemotaxic attraction to T_3 and exposure to T_3 exhibits increased production of serine proteases that function to degrade host defence peptides allowing for advanced pathogen invasion of the hosts dermal tissues (Thekkiniath *et al.*, 2013).

2.5. Conclusion

In conclusion, this analysis highlights the extent to which protein remodelling has been central to the diversification of fungi. While most studies of protein evolution, tend to focus on the treelike parts of evolutionary history, we show that the non-treelike parts can be equally interesting. More than seventy percent of the extant sequences we have examined in fungal genomes show evidence of remodelling and we have shown some of the factors that have influenced this remodelling. With increased sampling of genomes, we expect the proportion of proteins where we can detect remodelling in their history will increase – greater sampling of genes increases the likelihood of detecting remodelling events. Remodelling of proteins provides a rich supply of possibilities for all organisms, in terms of new functions, combinations of functions, co-expression of functions and separating functions from one another. It is perhaps not surprising that remodelling is so rampant.

Chapter III:

Bioinformatic and Biostatistical Analyses of Gene Remodelling in Viridiplantae

3.1: Introduction to botany

The cultivation of plants has possibly impacted human civilisation more than any other factor. Plants, alongside cyanobacteria, guided ancient ecosystem evolution by reducing atmospheric CO_2 levels and temperature, and by increasing O_2 and atmospheric pressure through photoautotrophic endeavours (Wallace *et al.*, 2017). Archaeplastida (Viridiplantae ('green' plants and algae), Rhodophyta ('red' algae), Glaucophyta (glaucophyte algae)), cyanobacteria ('blue-green' algae), and photosynthetic stramenopiles ('brown' algae) perform the majority of primary metabolic production in any exposed ecosystem and, as such, form the beginning of most food webs. The United Nations predict that world populations will approach 10 billion by 2050, with significant accelerated growth observed in underdeveloped regions of Africa and Asia, which is projected to continue indefinitely (https://population.un.org/wpp/Publications/). In light of such rapid acceleration in human population growth, one of the greatest challenges posed is how to adequately and sustainably feed an expanding population (Li and Zhang, 2007; Turner, 2009) despite continued devastating losses due to plant pathogens, pests and environmental disasters (Bebber, Ramotowski and Gurr, 2013; Bebber *et al.*, 2014; Bebber and Gurr, 2015; Lesk *et al*., 2016).

Comparative genomics between economically viable species, may unlock the potential for sustainable, higher yielding, more disease resistant, nutritious crop production without impacting local ecology or public health. Fossil fuels (such as oil, coal, and methane) are the anaerobically decomposed remnants of ancient plant and animal matter (Berner, 2003), and have been used for heat and energy production by humans since the Bronze Age (Dodson *et al.*, 2014). Fossil fuel consumption has exponentially increased during the course of the 20th century, and they have become a highly contentious, highly valuable economic asset with devastating environmental effects due to increased atmospheric CO_2 and other greenhouse

gasses (Cheng and Xiong, 2014; Olson and Lenzmann, 2016). The most striking effect observed is the steady increase of global temperatures (Rosenzweig *et al.*, 2008; Trenberth *et al.*, 2014) resulting in adverse environmental effects such as the increased instances of wildfires, cryosphere decay, rising sea levels, disruption of ecosystems, and ocean acidification (Botkin *et al.*, 2007; Hoegh-Guldberg *et al.*, 2007; Jolly *et al.*, 2015; Zhang and Wang, 2015). As mentioned above, plant metabolism catalysed the reduction of atmospheric CO_2 through photosynthetic metabolism. Deforestation has considerably impaired CO_2 bioremediation (Houghton, 1991; Song *et al.*, 2015). The devastating effects of global warming has prompted research into CO_2 bioremediation by plants and chlorophyte microalgae through comparative genomic mining and biotechnological augmentation (Maeda *et al.*, 1995; Zeiler *et al.*, 1995; Raeesossadati *et al.*, 2014; Gonçalves *et al.*, 2016). However, as fossil fuel resources are limited, the economic value of fossil fuel has sparked massive genetic engineering projects in chlorophytes and diatoms (Hill *et al.*, 2006; Cockerill and Martin, 2008; Gouveia and Oliveira, 2009; Olson and Lenzmann, 2016). These biofuels are long aliphatic hydrocarbons produced *via* the algal fatty acid synthesis pathway and are the subject of intense research (Miao and Wu, 2006; Chisti, 2007; Gouveia and Oliveira, 2009; Sani *et al*., 2013; Nayak *et al.*, 2016).

Plants, like fungi and bacteria, engage in secondary metabolism (Cavalier-Smith, 1992) and produce a wide array of therapeutic compounds and valuable industrial synthons (Aharoni and Galili, 2011). With the advent of advanced synthetic biology and systems biology tools, understanding the mechanisms underlying the structure and evolution of plant genomes is a highly important task if we wish to uncover hidden treasures in future biotechnological efforts.

Comparative analyses between plants have revealed that their evolutionary histories are replete with polyploidisation events (resulting in increased chromosome copy numbers) followed by massive chromosomal restructuring events (Jiao *et al.*, 2011; Li *et al.*, 2015). These events allow for the process of gene remodelling, specifically fusions, fissions, and exon

shuffling events (França *et al*., 2012; Leonard and Richards, 2012), gene subfunctionalisation, and neofunctionalisation (Rastogi and Liberles, 2005; Conant *et al.,* 2014), and the co-option of genes into alternative pathways (True and Carroll, 2002; Cock *et al.*, 2014) which can have drastic effects on the metabolic prowess, genotype, and phenotype of a given lineage.

Genomic analyses have identified a wealth of such remodelling events throughout the Tree of Life (Enright and Ouzounis, 2001; Froy and Gurevitz, 2003; Pasek *et al.*, 2006; Nakamura *et al.*, 2007; Nagy and Patthy, 2011; Leonard and Richards, 2012), however these analyses generally look at specific events, or specific types of events during plant evolution. The expression of fused plant genes in plant hosts has yielded some valuable phenotypes, specifically increased salt tolerance in *Lolium perenne,* an agriculturally important food source for grazing animals (Cen *et al.*, 2016), increased trehalose accumulation and decreased abiotic stress in *Oryza sativa* (Garg *et al.*, 2002), increased wax ester biosynthesis in *Nicotania tabacum* (Aslan *et al.*, 2014), and enriched formaldehyde bioremediation by *Pelargonium* sp. 'Frensham' (geranium) (Song *et al.*, 2010). Comparative analyses of plants and mining for remodelling events may provide invaluable knowledge by aiding in the elucidation of their complex biochemistries and evolutionary histories.

3.1.1: An overview of plant evolution

Photosynthesis was first achieved by ancient cyanobacteria over 3 billion years ago (gigayears ago; Ga; Blankenship, 2010; Betts *et al.,* 2018). Approximately 1033-1891 Ga an endosymbiotic event between a cyanobacteria and a non-photosynthetic eukaryote led to the evolution of the first photosynthetic eukaryotes, and the subsequent rise of the Archaeplastida (Keeling, 2010; Betts *et al.*, 2018). The combined cyanobacterial and archaeplastidian photoautotrophies led to a slow accumulation of atmospheric oxygen and ozone, a reduction in

CO_2 levels, and the subsequent cooling of Earth (Kenrick *et al.*, 2012; Martin and Allen, 2018). This ozone layer and thickened atmosphere reduced the levels of irradiating ultraviolet (UV) rays in the troposphere (Lowe and Tice, 2007), which is a key factor attributed to the eventual myco- and subsequent phytoterrestrialisation events (Bidartondo *et al.*, 2011; Kenrick *et al.*, 2012). Basal embryophytes (land plants) further expanded oxygen production (Wallace *et al.*, 2017).

Chlorophyta are the earliest known diverging Viridiplantae division (phylum) and, with the Charophyta, constitute the green algae (Martin *et al.*, 2002; Lewis and McCourt, 2004; Nishiyama *et al.*, 2018). The currently accepted hypothesis is that early marine chlorophytes adapted to limnic environments and transitioned into complex charophytes (Lewis and McCourt, 2004; McCourt *et al.,* 2004; Leliaert *et al.*, 2012; Delwiche and Cooper, 2015; Domozych *et al.*, 2016). The transition from aquatic to terrestrial environments led to the eventual transition toward the evolution of the multicellular embryophytes (Delwiche and Cooper, 2015). Streptophyta is a monophyletic clade and circumscribes charophytes and embryophytes (Becker and Marin, 2009).

Recent phylogenetic studies have grouped embryophytes as the sister taxa to three late diverging charophyte lineages, Charophyceae, Coleochaetophyceae, and Zygnematophyceae (CCZ clade), with Zygnematophyceae being the most likely closest relative to embryophytes (Karol *et al.*, 2001; Guiry, 2013; Lemieux *et al.*, 2016) (Figure 3.1.2.).

Phragmoplasophyta is a monophyletic clade that encompasses the CCZ clade and embryophytes (Wilhelmsson *et al.*, 2017). Phragmoplasophyta are defined by their synapomorphic phragmoplast, a structural component which aids in cell wall establishment after division cycles (Euteneuer and McIntosh, 1980; Smertenko *et al.*, 2018). The earliest branching embryophytes were non-vascular, and were likely very similar to extant bryophytes,

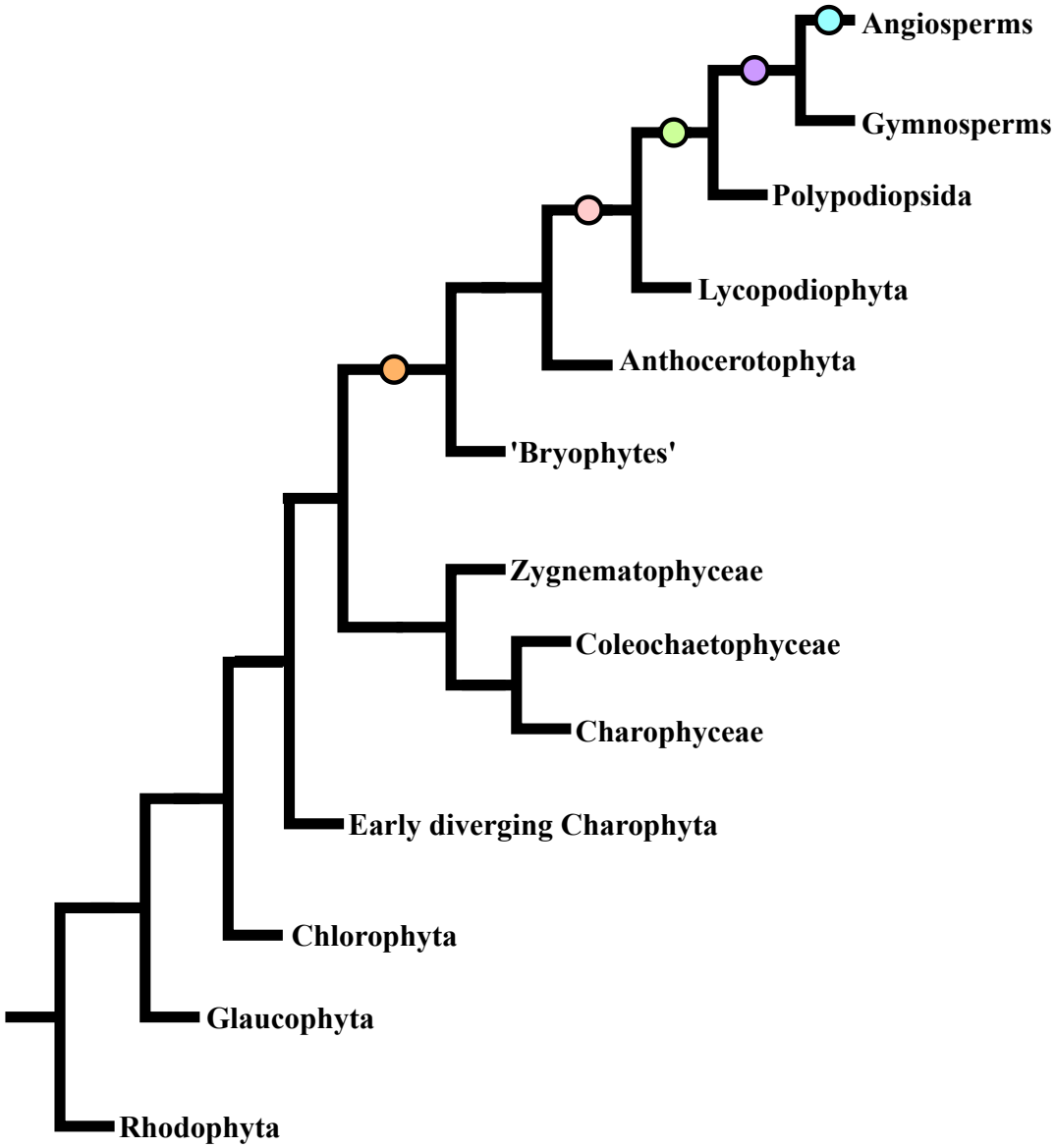

Figure 3.1.1. Phylogeny of the Archaeplastida

Rhodophyta and Glaucophyta are basal to Viridiplantae and are not “true plants”. Coloured circles indicate a major phenotype transition. The orange circle represents terrestrialisation, pink represents vascularisation, green represents the evolution of the modern leaf structure, purple represents the development of seeds, and blue represents the evolution of pollen (Baldwin and Husband, 2011; Bidartondo *et al.*, 2011).

represented by three phyla: Bryophyta (mosses), Anthocerotophyta (hornworts) and Marchantiophyta (liverworts) (Mishler and Churchill, 1984; Kenrick and Crane, 1997; Nickrent *et al.*, 2000). The exact evolutionary radiation of these divisions within the bryophytes is unresolved, with one hypothesis suggesting that bryophytes are, themselves, monophyletic and sister to the monophyletic tracheophyte (vascular plants) clade, and the other, more widely accepted hypothesis suggesting that Bryophyta and Marchantiophyta constitute a monophyletic clade basal to Anthocerotophyta, which, itself, is sister to the vascular tracheophytes (Nishiyama and Kato, 1999; Duff *et al.*, 2009; Stotler and Crandall-Stotler, 2009; Ligrone *et al.*, 2012; Delwiche and Cooper, 2015; Morris *et al.*, 2018).

Tracheophytes are divided into two clades, the phylum Lycopodiophyta (club mosses), and the superphylum Euphyllophyta (Tomescu, 2009; Vasco *et al*., 2013). Lycopodiophyta and Euphyllophyta are differentiated by leaf vasculature complexity, Lycopodiophyta possess a single vascular trace whereas Euphyllophyta possess a complex vascular network. Euphyllophyta consist of the divisions Pteridophyta (sporogenic ferns) and Spermatophyta (seed plants) (Bennici, 2008; Clarke *et al*., 2011).

Ancestral polyploidisation events coincided with significant reproductive strategy innovations: the evolution of pollen mediated reproduction, and the development of the coated seed. This innovation allowed for the development of seed dormancy, protection against desiccation, alternative dispersal mechanisms, and removed the dependence on aquatic fertilisation (Baldwin and Husband, 2011; Jiao *et al.*, 2011; Li *et al.*, 2015). A second round of polyploidisation led to the divergence of angiosperms (flowering plants) from gymnosperms (softwoods) (Van de Peer *et al.*, 2009; Conant *et al.*, 2014). This whole genome duplication event coincided with a third significant reproductive innovation: the evolution of flowers (Chanderbali *et al.*, 2016). This led to complex symbiotic relationships with arthropods and an explosion in plant diversity (Labandeira, 2013).

3.1.2 Genome architecture evolution in Viridiplantae

Plant evolutionary history is abound with polyploidisation events (PE) (Blanc et al., 2003; Adams and Wendel, 2005; Doyle and Egan, 2010; Jiao *et al.*, 2011; Weiss-Schneeweiss *et al.*, 2013). Polyploidy can result *via* intragenomic events, autopolyploidic hybridization, or allopolyploidic hybridisation (Arrigo and Barker, 2012). Plant comparative genomics have concluded that all angiosperms, and likely all tracheophytes, are ancestrally polyploid, and that angiosperms freely undergo polyploidisation (Jiao *et al.*, 2011), whereas PE are less common in gymnosperms (Bennett, 2004). Recent PE in angiosperm lineages can be observed alongside genomic remnants of the PE ancestral to all angiosperms, and of the PE hypothesised to be ancestral to tracheophytes (Soltis *et al.*, 2009).

Plant genome evolution appears to follow a cyclic PE model (CycPE), where rounds of polyploidy occur and are followed by subsequent loss of redundant paralogs, chromosomal rearrangements, and the eventual decay of the redundant chromosomes before entering another cycle (Wendel *et al.*, 2016). CycPE was first hypothesised after multispecies expressed sequence tags (ESTs) comparative analyses, revealed bursts of sequence similarity, duplication patterns, and micro- and macrosynteny between distantly related organisms, suggesting an ancestral origin (Leitch and Bennett, 2004; Fawcett *et al.*, 2009; Jiao *et al.*, 2011; Wendel *et al.*, 2016). A wide range of genomic responses to PE are observed. Early genomic responses pertain to the molecular responses to individual genes and to their expression (Jackson and Chen, 2010; Arrigo and Barker, 2012). Molecular genetic responses may result in pseudogenisation, homologous exchange, or the expansion of incorporated, activated transposable elements. Expression level responses likely result in expression bias eventually progressing to subfunctionalisation, neofunctionalization, subneofunctionalisation, or pseudogenisation, accompanied by chromosomal rearrangements (Lynch and Conery, 2000;

Rastogi and Liberles, 2005*b*; Marques *et al.*, 2008; Albertin and Marullo, 2012). Abiding responses, usually result in genomic subfunctionalisation and neofunctionalization, and significant chromosomal rearrangement and decay, further resulting in chromosomal number reduction (Leitch and Leitch, 2008; Wendel *et al.*, 2016). New species emerging from multiple CycPE ultimately display considerable chromosomal architecture changes compared to their pre-polyploid progenitor (Jiao *et al.*, 2011; Arrigo and Barker, 2012). Emergent species are diploidised by these chromosomal events, while maintaining ohnologs and vestigial PE genomic architecture (Conant *et al.*, 2014; Estep *et al.*, 2014; Clark and Donoghue, 2018).

The fate of duplicate genes, especially ohnologs (homologs arising from a WGD event), has intrigued biologists since their discovery. Observations of genomic reduction following CycPE may suggest a non-random, selective fate for ohnologs (Conant and Wolfe, 2008; De Smet *et al.*, 2013; Panchy *et al.*, 2016). Restored singleton genes (genes that once had an ohnolog which has been subsequently deleted) have higher expression patterns and wider expression domains than those in duplicated pairs, and are statistically enriched for DNA replication and repair, chloroplast function, and other essential housekeeping functions (Conant and Wolfe, 2008; Panchy *et al.*, 2016). The selection for these genes to return to a singleton state may be due to gene dosage toxicity, or the stoichiometries of protein-protein interaction or protein complex assembly (Freeling, 2009; Birchler and Veitia, 2012). For example, genes producing monomeric products with few interaction partners or those that are non-essential in biological pathways would be under less selection pressure than multimeric product genes (those involved in a number of different pathways) or more essential products (Bashton and Chothia, 2007).

Gene biased fractionation (the non-random selection of lost and retained ohnologs from each respective donor in a hybridization event) has been observed throughout the angiosperms (Wang *et al.*, 2011; Cheng *et al.*, 2012; Freeling *et al.*, 2012; Conant *et al.*, 2014). An ancient

example of differential genomic fractionation can be observed from an early Paleozoic (~65 Ma) intergeneric alloploidization event in ancestral *Gossypium* sp. (cotton), evidence of which is observed in modern lineages (Renny-Byfield *et al.*, 2015). The selection processes governing biased fractionation aren't completely understood but may involve repressibility due to adjacency of transposable elements (TEs), leading to fewer constraints and varying levels of expendability (Renny-Byfield *et al.*, 2015; Wendel *et al.*, 2016).

Despite rounds of genome reduction following PE, considerable genome size variation and variations in complexity exist between even relatively closely related Viridiplantae genomes (Lang *et al.*, 2010; Marroni, Pinosio and Morgante, 2014; Ruhfel *et al.*, 2014), for example, the *Hordeum vulgare* (barley) genome is over 11 times larger than another Poeceae grass genome, *Oryza sativa* (rice; 5.1 Gbp *vs.* 0.43 Gbp respectively) despite having less chromosomes (7 *vs.* 12 respectively; Goff, 2002; Mayer *et al.*, 2011). Alongside PE, genomes may undergo expansion *via* rapid TE propagation if removal efforts are overwhelmed by unequal or illegitimate recombination events (Devos, 2002). Occurrences of lineage specific TE amplifications have been observed in plant lineages. Comparative genomics between *Oryza* reveal an approximate 400 Mbp increase, almost doubling in genome size, between *O. australiensis* and *O. sativa*, mostly due to three retro-TE families (Turcotte *et al.*, 2001; Ma and Bennetzen, 2004; Piegu *et al.*, 2006). Interestingly, the TEs responsible for the genomic expansion in *O. australiensis* are present in all other *Oryza* species, but, with the exception of an approximate 200 Mbp genomic expansion in *O. granulata* estimated to have occurred approximately 2 Ma, have remained comparatively inactive (Ma and Bennetzen, 2004; Ammiraju *et al.*, 2007), which may suggest an environmental cause for such evolutionary bursts. TE amplification is also reported to have been responsible for the tripling in genome size between *Gossypium* species (Hawkins *et al.*, 2006). The genome size variations between

members of *Oryza* and *Gossypium* represent the dynamics imposed by TE proliferation and decay in conjunction with CycPEs.

Comparative genomic analyses between the TE poor *Genlisea aurea* and the TE rich octopolyploid *Paris japonica* revealed little variation in gene content despite massive variation in genome size (~60 Mbp and >150Gbp respectively) (Pellicer *et al.,* 2010). This observation illustrates the enormous effect of TE propagation in genomic expansion, which comparatively depreciate the effects of tandem gene duplication in genome expansion and can virtually negate the effect of gene loss *via* genomic fractionation following a PE (Wendel *et al.*, 2016). TEs play an important gene regulatory role despite observed gene content consistency and a relatively constant TE accumulation and decay rate (Leitch and Leitch, 2008; Schrader *et al.*, 2014).

Despite the vast information attained from reference genomes, organisms are constantly evolving and adapting such that a single genome does not represent the dynamic genetic variation within a species, however these sequences have proved invaluable in resequencing efforts, leading to considerably greater insight into genomic variation within resampled species (Huang *et al.*, 2009; Huang *et al.*, 2013). Although considerably valuable, resequencing efforts are limited by poor quality mapping of short read sequences in species with high TE activity or other genomic variations resulting in small indels and intergenic sequences not being detected (Kircher and Kelso, 2010; Alkan *et al.*, 2011).

High throughput systems biology approaches, such as pangenomic and pantranscriptomic analyses are used in an attempt to capture as much variation information as is possible, and have been used to map variances in a multitude of model plant systems (Hirsch *et al.*, 2014; Golicz *et al.* , 2016; Contreras-Moreira *et al.*, 2017; Sun *et al.*, 2017).

3.1.2.1. Evolutionary development via transcription factor co-option

Gene co-option (exaptation) is an evolutionary process where a gene is incorporated into a new functional pathway without initially altering its functional biochemistry (True and Carrol, 2002). Co-opted genes are often paralogs (or ohnologs) and may be subjected to subfunctionalisation or neofunctionalisation after co-option has taken place (McLennan, 2008; Hilgers *et al*., 2018). These genes may also be subjected to gene remodelling which may further increase their fitness effects. Gene co-option has been extensively studies in plants, especially through transcription factor (TF) co-option during the transition from haplontic predominant to diplontic predominant life cycles (Szovenyi *et al.*, 2011; Pires and Dolan, 2012).

Embryophytes display "alternation of generations" (AG) (the alternation between a multicellular haploid gametophyte generation and a diploid sporophyte generation during their life cycles; Kenrick, 1994; Graham and Wilcox, 2000). A haplontic lifecycle is exhibited by all known green algae where a haploid stage is observed for the majority of the algal life cycle, with an ephemeral diploid zygote stage. Conversely, embryophytes display a haplodiplontic life cycle, where two mitotic stages with differential ploidy is observed (Niklas and Kutschera, 2010). In charophytes, only zygotes display diploidy (Haig, 2010; Nishiyama *et al.*, 2018). Two competing hypotheses compete to decipher the origin of AG in embryophytes, the "homologous theory" (modification theory) and the more widely accepted "antithetic theory" (intercalation theory; Bower, 1890; Bennici, 2008). The homologous theory suggests that land plants first displayed isomorphic AG, however there is no known genetic support for this and only sparse early Devonian fossils which appear to be isomorphic (Thornber, 2006). Comparatively, the antithetic theory suggests that that the sporophyte arose through successive mitotic division phase insertions during zygotic generation prior to meiotic division resulting in a diploidic (sporophytic) embryo on a gametophytic thallus (Bennici, 2008). This

protosporophyte is hypothesised to have transitioned from a gametophyte nutritional parasite to a metabolically independent sporophyte, followed by subsequent transitionary steps towards life cycle dominance (Qiu *et al.*, 2006; Kenrick, 2017). The antithetic theory is supported by the triad of (*a*.) sporophyte absence in charophytes, (*b*.) gametophyte dominance over sporophytes in bryophytes, and (*c*.) sporophyte dominance over gametophytes in tracheophytes (Niklas and Kutschera, 2010).

A hypothetical model of sporophyte evolution *via* the co-option of ancient gametophyte genes and their regulatory gene networks has been proposed (Szovenyi *et al.*, 2011; Pires and Dolan, 2012). This model is based on greater differential expression patterns between gametophytes and sporophytes of angiosperms when compared with bryophytes, and that a plethora of gametophyte-biased transcription factors are preferentially expressed in angiosperm sporophytes. One such example of gametophyte- specific regulatory TF co-option to sporophyte regulation can be observed in type II MADS box (MADS-II) TF (Henschel *et al.*, 2002; Singer *et al.*, 2007). MADS-II has a distinct, functional evolutionary origin in charophytes, where it regulates haploid germ cell differentiation (Tanabe *et al.*, 2005). During embryophyte divergence, MADS-II underwent a duplication and subsequent neofunctionalization, forming the MIKC^{C} and MIKC* TF families (Henschel *et al.*, 2002; Singer *et al.*, 2007). MIKC* retained MADS-II conserved function, regulating gametophyte development in embryophytes, whereas MIKC^{C} primarily regulates sporophyte development. MIKC* displays differential expression and regulatory functions between bryophytes and angiosperms, it is expressed in both bryophyte gametophyte and sporophyte generations, but is restricted to sporophyte expression in *Arabidopsis thaliana* (Quodt *et al.*, 2007). An almost complete set of the MIKC^{C} gene family is present in sampled gymnosperms, further expanded before the divergence of the basal angiosperm genus, *Amborella*, and further expanded again in crown angiosperms (Münster *et al.*, 1997; Gramzow and Theissen, 2010; Melzer *et al.*,

2010). These expansions and diversifications aided in the development of the sporophyte, and later, the flower (Tanahashi *et al.*, 2005; Singer, Krogan and Ashton, 2007). This evolutionary pattern likely follows an evolution by subneofunctionalisation model (He and Zhang, 2005). Comparative studies between *Physcomitrella patens*, a model Bryophyta, and spermatophytes have shed light on the evolution of both embryogenesis and AG (Kimura *et al.*, 2008; Lee *et al.*, 2008; Hay and Tsiantis, 2010). Homeodomain transcription factor (TF) control was found to be of particular importance in both of these pathways, KNOX family TF repression results in the repression of gametophyte body plan development, and BELL family TF repression results in either the repression of zygote development or zygote division. An RWP-RK family TF in the model liverwort (Marchantiophyta) *Marchantia polymorpha* was found to control dormancy in unfertilized ova and germ cell formation (Chardin *et al.*, 2014). Gametophyte reduction is observed in angiosperms when compared to other embryophyte clades (Brandes, 1973; Kerp, *et al.*, 2003; Borg *et al.*, 2009). MpRWP-RK TF orthologs in *A. thaliana* (*DUO1* and *DUO3*) were found to control development of pollen generation cells (Durbarry, 2005; Rövekamp *et al.*, 2016). As MpRWP-RK TF was found to complement *DUO1* and *DUO3* function in *A. thaliana* (Durbarry, 2005) it has been hypothesised that sexual reproductive control has been conserved since the emergence of early embryophytes. It is possible that gametophyte genes (and their regulatory networks) were co-opted or neofunctionalised to promote sporophyte evolution during the rise of the spermatophytes (True and Carroll, 2002; Teshima and Innan, 2008).

Some genes with ancient origins are transcriptionally restricted to one lifecycle generation, two examples of which are the TALE homeobox TF superclass (Hamant and Pautot, 2010; Hudry *et al.*, 2014) and *LFY* (Tanahashi, 2005). Initiation of the diploid phase in the model alga *Chlamydomonas reinhardtii* (Chlorophyta) lifecycle is achieved by heterodimerisation of two TALE proteins, Gsm1p and Gsp1p, whereas embryophytic

sporophyte development processes are also under control of TALE homeobox TF, specifically KNOX and BELL (Lee *et al.*, 2008). These results suggest that TALE genes may be restricted to diploid stages throughout Viridiplantae.

LFY is a floral meristem regulator in angiosperms, however it is expressed during the sporophytic generation throughout streptophyte lineages, where it regulates zygotic cell division in non-flowering embryophytes and charophytes (Tanahashi, 2005; Silva *et al.*, 2015). These examples illustrate generation phase exclusivity, and highlight the questions of "Where did generation phase genes first originate?" and "What were their original pathways (if any) before evolving to function as sporophyte regulators?"

Polycomb-group proteins (PGP) remodel chromatin by methylating lysine 27 on histone 3 (H3K27) resulting in epigenetic transcriptional repression (Okano *et al.*, 2009). Two interacting PGPs, CLFp and FIEp, have been identified as AG regulators in *P. patens* (Mosquna *et al.*, 2009). Deletion of either CLFp or FIEp in *P. patens* results in the initiation of apogamy (the fertilisation independent generation of sporophytes) (Mosquna *et al.*, 2009; Chen *et al.*, 2014). CLFp is homologous to gene products from the $E_{(Z)}$ gene cluster in animals (Schatlowski *et al.*, 2010) which are responsible for appropriate epigenetic control of embryonic development, suggesting an ancient role for these genes in sexual reproduction. CLFp in angiosperms, however, is not implicated in embryogenesis, instead, it regulates the floral homeotic gene *FLC*1 (Jiang *et al.*, 2008). Δ*CLF* mutants display altered flowering and curled leaf morphologies (Jiang *et al.*, 2008; Chen *et al.*, 2014). The recruitment of CLF as a floral repressor likely occurred during the angiosperm divergence from gymnosperms, where it may have originally prevented floral gene expression in leaf tissues (Cairney and Pullman, 2007).

Due to the dynamic genome architectures observed during the course of divergent Viridiplantae evolution, it is reasonable to hypothesise that these genes may have been

subneofunctionalised following a PE, subsequently leading to greater gene diversity, anatomical evolution, and the emergence of the sporophyte.

3.1.2.2. Metabolic evolution

Due to the "polyploidisation-rearrangement-reduction" model observed during CycPE throughout plant evolutionary history, it is likely that new genes may arise at these points through gene remodelling events. This is largely due to the favourable conditions for remodelling that are observed during the course of the CycPE, namely the influx of new genetic material and rampant chromosomal restructuring (Leonard and Richards, 2012). As mentioned previously, remodelled genes may be co-opted, subfunctionalised, neofunctionalised, or subneofunctionalised (Hellsten *et al.*, 2007; Xia *et al.*, 2016) leading to novel or augmented metabolic pathways (Richards *et al.*, 2006). Previous studies have investigated plant gene remodelling events, but most work investigates specific fusion genes (Wang, 2006; Aslan *et al.*, 2015; Méheust *et al.*, 2016), or comparative analyses that identify differential remodelling events between a small set of genomes (*eg.* Nakamura *et al,*, 2007). Plant gene fusion events have garnered considerable interest due to their influences on secondary metabolism, where both component genes (if present) are often co-localised and co-expressed (Hagel and Facchini, 2017). Gene fusion and fission events are the most well studied remodelling events, where they have been attributed to the expansions of protein domain complexities in bacteria (Pasek *et al.*, 2006). Differential gene fusions are likely to be involved in the same pathway as their components, and have been annotated as "extreme clustering", and may confer niche physiological advantages compared to species with unfused components in different environments, by ensuring extremely tight control of co- expression and co-localisation of two components of their respective pathways (Enright and Ouzounis, 2001; Richards *et al.*, 2006;

Fani *et al.*, 2007; Leonard and Richards, 2012; Avelar *et al.*, 2014). As plant evolution is abound with CycPEs, and major metabolic and phenotypic innovations are associated with ancestral polyploidisation events, it is probable that gene remodelling played a role in their establishment.

3.2. Methodology

The methodology for this chapter largely follows the methodology used in Chapter II, any changes made to any protocol described in Chapter II are discussed in detail.

3.2.1. Database construction and quality control

A database of 50 non-canonical Viridiplantae proteomes were constructed from PLAZA *v*4.0 (Van Bel *et al*., 2018), Gymno-PLAZA *v*1.0 (Proost *et al*., 2015), and pico-PLAZA *v*2.0 (Vandepoele *et al*., 2013). The dataset consisted of six Chlorophyta, one Bryophyta, one Marchantophyta, one Lycopodiophyta, three Pinophyta (Gymnospermae), and 38 Angiospermae (Table 3.2.1.). This dataset contained a combined total of 1,672,377 sequences and is biased towards angiosperms as they are the most widely sampled species due to their economic importance. There are only a limited number of sequenced basal embryophytes, and only three gymnosperms from Gymno-PLAZA could be used as other assemblies were constructed from transcriptomic data, which is not suitable for gene remodelling analyses due to elevated levels of truncated sequences in such assemblies (Hara *et al*., 2015). As per *subsection 2.2.3.1.*, genome density, genome size and GC% were obtained from source. Genome quality was assessed using BUSCO *v*3 with default settings with the Viridiplantae ortholog dataset from OrthoDB *v*10 (Kriventseva et al., 2018). Genome size,

Table 3.2.1: Dataset of 50 Viridiplantae genomes

Each subset represents a major clade to which its species belongs. The subset 'Dicot' represents dicotyledons, and 'Monocot' represents monocotyledons. The majority of sequenced plant genomes are crops, leading to a bias towards angiosperm lineages. Common names are provided for each species (if applicable) and if a common name could not be ascertained, the reasoning behind the species inclusion is provided instead (*eg.* multicellular alga).

Common name/reason	Subset	Species	Source	Taxonomy
Model picoeukaryote	Alga	*Chlamydomonas reinhardtii*	PicoPlaza *v*2.0	Chlorophyta; Chlorophyceae; Chlamydomonadales; Chlamydomonadaceae; Chlamydomonas
Multicellular alga	Alga	*Volvox carteri*	PicoPlaza *v*2.0	Chlorophyta; Chlorophyceae; Chlamydomonadales; Volvocaceae; Volvox
Model picoeukaryote	Alga	*Micromonas commoda*	PicoPlaza *v*2.0	Chlorophyta; Mamiellophyceae; Mamiellales; Mamiellaceae; Micromonas
Model picoeukaryote	Alga	*Ostreococcus lucimarinus*	PicoPlaza *v*2.0	Chlorophyta; prasinophytes; Mamiellophyceae; Mamiellales; Bathycoccaceae; Ostreococcus
Model picoeukaryote	Alga	*Chlorella sp.* NC64A	PicoPlaza *v*2.0	Chlorophyta; Trebouxiophyceae; Chlorellales; Chlorellaceae; Chlorella
Model picoeukaryote	Alga	*Coccomyxa sp.* C169	PicoPlaza *v*2.0	Chlorophyta; Trebouxiophyceae; Trebouxiophyceae incertae sedis; Coccomyxaceae; Coccomyxa
Basal angiosperm	Basal angiosperm	*Amborella trichopoda*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; Amborellales; Amborellaceae; Amborella
Moss	Bryophyte	*Physcomitrella patens*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; Bryopsida; Funariidae; Funariales; Funariaceae; Physcomitrella
Potato	Dicot	*Solanum tuberosum*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; asterids; Solanales; Solanaceae; Solanoideae; Solaneae; Solanum
Tomato	Dicot	*Solanum lycopersicum*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; asterids; Solanales; Solanaceae; Solanoideae; Solaneae; Solanum; Lycopersicon
Pepper	Dicot	*Capsicum annuum*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; eudicotyledons; Gunneridae; Pentapetalae; asterids; lamiids; Solanales; Solanaceae; Solanoideae; Capsiceae; Capsicum
Beet	Dicot	*Beta vulgaris*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; eudicotyledons; Gunneridae; Pentapetalae; Caryophyllales; Chenopodiaceae; Betoideae; Beta
Clementine	Dicot	*Citrus clementina*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Sapindales; Rutaceae; Aurantioideae; Citrus
Rape	Dicot	*Brassica rapa*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Brassicales; Brassicaceae; Brassiceae; Brassica
Thale cress	Dicot	*Arabidopsis thaliana*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Papaya	Dicot	*Carica papaya*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Brassicales; Caricaceae; Carica
Watermelon	Dicot	*Citrullus lanatus*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Cucurbitales; Cucurbitaceae; Benincaseae; Citrullus

Common name/reason	Subset	Species	Source	Taxonomy
Cucumber	Dicot	*Cucumis melo*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Cucurbitales; Cucurbitaceae; Benincaseae; Cucumis
Soybean	Dicot	*Glycine max*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Fabales; Fabaceae; Papilionoideae; Phaseoleae; Glycine; Soja
Barrel clover	Dicot	*Medicago truncatula*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Fabales; Fabaceae; Papilionoideae; Trifolieae; Medicago
Castor bean	Dicot	*Ricinus communis*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Malpighiales; Euphorbiaceae; Acalyphoideae; Acalypheae; Ricinus
Cassava	Dicot	*Manihot esculenta*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Malpighiales; Euphorbiaceae; Crotonoideae; Manihoteae; Manihot
Poplar	Dicot	*Populus trichocarpa*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Malpighiales; Salicaceae; Saliceae; Populus
Cocoa	Dicot	*Theobroma cacao*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Malvales; Malvaceae; Byttnerioideae; Theobroma
Cotton	Dicot	*Gossypium raimondii*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Malvales; Malvaceae; Malvoideae; Gossypium
Eucalyptus	Dicot	*Eucalyptus grandis*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Myrtales; Myrtaceae; Myrtoideae; Eucalypteae; Eucalyptus
Peach	Dicot	*Prunus persica*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Rosales; Rosaceae; Amygdaloideae; Amygdaleae; Prunus
Apple	Dicot	*Malus domestica*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Rosales; Rosaceae; Amygdaloideae; Maleae; Malus
Strawberry	Dicot	*Fragaria vesca*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Rosales; Rosaceae; Rosoideae; Potentilleae; Fragariinae; Fragaria
Grape	Dicot	*Vitis vinifera*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; rosids; Vitales; Vitaceae; Vitis
Norway spruce	Gymnosperm	*Picea abies*	GymnoPlaza *v*1.0	Streptophyta; Streptophytina; Pinidae; Pinales; Pinaceae; Picea
White spruce	Gymnosperm	*Picea glauca*	GymnoPlaza *v*1.0	Streptophyta; Streptophytina; Pinidae; Pinales; Pinaceae; Picea
Loblolly pine	Gymnosperm	*Pinus taeda*	GymnoPlaza *v*1.0	Streptophyta; Streptophytina; Pinidae; Pinales; Pinaceae; Pinus
Spikemoss	Lycopodiophyte	*Selaginella moellendorffii*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; Lycopodiopsida; Selaginellales; Selaginellaceae; Selaginella
Liverwort	Marchantiophyte	*Marchantia polymorpha*	DicotPlaza *v*4.0	Streptophyta; Streptophytina; Marchantiopsida; Marchantiidae; Marchantiales; Marchantiaceae; Marchantia
Duckweed	Monocot	*Spirodela polyrhiza*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Alismatales; Araceae; Lemnoideae; Spirodela
Eelgrass	Monocot	*Zostera marina*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Alismatales; Zosteraceae; Zostera
Orchid	Monocot	*Phalaenopsis equestris*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Asparagales; Orchidaceae; Epidendroideae; Vandeae; Aeridinae; Phalaenopsis
Pineapple	Monocot	*Ananas comosus*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Bromeliaceae; Bromelioideae; Ananas
Bamboo	Monocot	*Phyllostachys edulis*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Bambusoideae; Arundinarieae; Arundinariinae; Phyllostachys
Lawngrass	Monocot	*Zoysia japonica*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Chloridoideae; Zoysieae; Zoysiinae; Zoysia
African rice	Monocot	*Oryza brachyantha*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Oryzoideae; Oryzeae; Oryzinae; Oryza
Rice	Monocot	*Oryza sativa ssp. japonica*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Oryzoideae; Oryzeae; Oryzinae; Oryza; Oryza sativa
Sorghum	Monocot	*Sorghum bicolor*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Panicoideae; Andropogoneae; Sorghinae; Sorghum
Maize	Monocot	*Zea mays*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Panicoideae; Andropogoneae; Tripsacinae; Zea
Millet	Monocot	*Setaria italica*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Panicoideae; Paniceae; Cenchrinae; Setaria
Wild grass	Monocot	*Brachypodium distachyon*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Pooideae; Brachypodieae; Brachypodium

Common name/reason	Subset	Species	Source	Taxonomy
Barley	Monocot	*Hordeum vulgare*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Poales; Poaceae; Pooideae; Triticeae; Hordeinae; Hordeum
Banana	Monocot	*Musa acuminata*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Petrosaviidae; Zingiberales; Musaceae; Musa
Wheat	Monocot	*Triticum aestivum*	MonocotPlaza *v*4.0	Streptophyta; Streptophytina; Liliopsida; Poales; Poaceae; BOP clade; Pooideae; Triticodae; Triticeae; Triticinae; Triticum

density and GC% were obtained from source (Table 3.3.1.). Descriptive statistics were computed for each data series presented in Table 3.3.1. (Table 3.3.2.).

3.2.2. CompositeSearch analysis, quality control, and annotation

A SSN was constructed using BLASTP (as per *subsection 2.2.3.2.*) resulting in $2.797e^{12}$ pairwise comparisons. CompositeSearch analysis, remodelled gene quality control, and remodelling category annotation was replicated as per *subsection 2.2.3.2.* The extent of remodelled genes and families was computed for the dataset (Table 3.3.3; Figure 3.3.1.) and compared to the fungal dataset (Table 2.3.5.) using a Fisher's exact test ($H_O:\pi(x)=\pi(X);H_A:\pi(x)\neq\pi(X)$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = |RC| = 4$; $\alpha_B = 0.0125$) and instances where $P \leq \alpha_B$ were considered statistically significant (Table 3.3.4.).

3.2.3. Trends in gene family size

As per subsection 2.2.3., descriptive statistics were computed for RC family sizes (Table 3.3.5.). Differences in family sizes between RCs were assessed using Mann-Whitney *U*-tests ($H_O:\eta_1=\eta_2;H_A:\eta_1\neq\eta_2$). Again, a Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$) and a $P \leq \alpha_B$ was considered statistically significant (Table 3.3.6.). Family sizes were compared for each specific RC (*eg.* NC *vs* NC) between fungal (Table 2.3.6) and plant datasets using Mann-Whitney *U*-tests ($H_O:\eta_1=\eta_2;H_A:\eta_1\neq\eta_2$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 4$; $\alpha_B = 0.0125$) and a $P \leq \alpha_B$ was considered statistically significant (Table 3.3.7.).

3.2.4. Phylogenetic and character state reconstruction

3.2.4.1. Phylogenetic reconstruction

Considerable difficulties are associated with resolving Viridiplantae phylogenies using molecular data due to the prevalence of ohnologs following frequent polyploidisation events (Ruhfel *et al.*, 2014; Van Bel *et al.*, 2018). We attempted phylogeny reconstruction using a ubiquitous ortholog gene tree superalignment. Ubiquitous genes from the Viridiplantae odb *v*10 dataset (as used in section 3.1.) were identified by searching against each proteome using reciprocal BLASTP ($E \leq 1e^{-20}$). A total of 99 genes were observed to be ubiquitous, however, as none were observed to be single copy, we selected the reciprocal BLAST best hit for each gene from each genome for further analysis. Genes were aligned in each of the 99 gene sets using MUSCLE *v*3.8 (Edgar, 2004) under default parameters. Uninformative alignments were removed using gBlocks (Castresana, 2000) with a minimum block size of 5, and concatenated into a superalignment using FASconCAT *v*1.11 (Kück and Meusemann, 2010). The best model for protein evolution was ascertained to be the LG+I+G model (Le and Gascuel, 2008) using ProtTest *v*3 (Abascal *et al.*, 2005). A total of 100 constrained bootstrap replicates were produced using PhyML v3.0 (Guindon *et al.*, 2010). A consensus tree was constructed from the PhyML produced replicates using 'majority rule' in PAUP* v3 (Swafford, 2002) and visualised using iTOL v3 (Letunic and Bork, 2016) (Figure 3.3.2.).

Due to the considerable disagreement between our phylogeny and the expected phylogeny, we reattempted reconstruction using the 31 ubiquitous orthologs identified by Ciccarelli and company (2006) using the same methodology as above. Again, no single copy orthologs were detected and the best model of protein evolution was determined to be LG+I+G (Figure 3.3.3.). This phylogeny was, again, highly disagreeable with the expected phylogeny.

To overcome the difficulties in resolving a defensible phylogeny, a "scaffold phylogeny" was inferred from PLAZA *v*4, Gymno-PLAZA *v*1.0, and Pico-PLAZA *v*2.0, and

used to constrain the construction of a phylogeny using the 99 ubiquitous genes from the Viridiplantae odb dataset as above (Figure 3.3.4.).

3.2.4.2. Character state reconstruction

Character state reconstruction was completed as per *subsection 2.2.3.4.2* using the phylogeny constructed in *subsection 3.2.4.1.* with two "pseudo-outgroups" appended to the root (Figure 3.3.5.).

3.2.4.3. Comparison of homoplastic proportions

Homoplastic proportions were derived using the formula described in *subsection 2.2.3.3.2.1.* and compared using a two-tailed Fisher's exact test (H_O:$\pi(x)=\pi(X)$;H_A:$\pi(x)\neq\pi(X)$). Again, a Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$). Instances where $P \leq \alpha_B$ was considered statistically significant (Table 3.3.8.). HPs were compared for each specific RC (*eg.* NC *vs* NC) between fungal (Table 2.3.7) and plant datasets using Fisher's exact tests (H_O:$\pi(x)=\pi(X)$;H_A:$\pi(x)\neq\pi(X)$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 4$; $\alpha_B = 0.0125$) and a $P \leq \alpha_B$ was considered statistically significant (Table 3.3.9.).

3.2.4.4. Evolutionary rate series construction

Evolutionary rate series (f_x) for each RC was constructed using the formulae described in *subsection 2.2.3.4.4.* Unlike fungi, a comprehensive collection of fossilised Viridiplantae structures have been curated (Harris and Davies, 2016), and two separate node calibration points have been presented that could be used for τ calculation (Betts *et al*., 2018):

(*i.*): The root of Embrophyta (Node 58)

Betts and company estimated Embryophyta to have emerged between 448.5 - 509 Ma. The farthest leaf from Node 58 was observed to be *Zoysia japonica* (FL = 0.63226874), from which in τ was calculated to be 709.35 - 805.04 Ma. The farthest leaf from the root was, again, observed to be *Z. japonica* (FL_R = 0.97834456). Adjusting τ for FL_R suggests that the common ancestor for all taxa in our dataset emerged between 725.05 - 822.86 Ma.

(*ii.*): The root of Angiospermae (Node 55)

Betts and company estimated Angiospermae to have emerged between 125.9 – 247.3 Ma, much more recent than their embryophyte ancestors. Again, farthest leaf from Node 58 was observed to be *Z. japonica* (LP = 0.44831458), from which τ was calculated to be 280.8296 - 551.6216 Ma. Adjusting τ by LP_R, the common ancestor of all taxa in our dataset emerged 287.05 - 563.83 Ma.

Fossil evidence was used to select the most appropriate τ (τ = 805.04 Ma) for rate series computation. Chlorophytes are hypothesised to have emerged during the Neoproterozoic Era, approximately 541 Ma – 1 Ga (Fang *et al*., 2017). The Neochlorophyta ("UTC") clade encompasses three chlorophyte classes, Ulvophyceae, Trebouxiophyceae, and Chlorophyceae (Derrien *et al*., 2009). The Ulvophyceae were absent in our dataset, however two trebouxiophycean taxa (*Chlorella* sp. NC64A and *Coccomyxa* sp. C169) and two chlorophycean taxa (*Chlamydomonas reinhardtii* and *Volvox carteri*) were represented. Evidence from the fossil record suggest that neochlorophytes had already diverged to three

distinct chlorophyte lineages by the Neoproterozoic Era specifically the Tonian Era, 720 Ma – 1 Ga (Butterfield *et al*., 1994; Arouri *et al*., 1999; Fang *et al*., 2017; Del Cortona *et al*., 2019). When FL_R is transformed to time using τ (derived from Embrophyta), the most recent common ancestor to all of our taxa is estimated to have existed between 725.05 - 822.86 Ma. The more ancient τ was selected to definitively place the MRCA within the Tonian era.

3.2.4.5. Comparison of rates between and within nodes and tips

Descriptive statistics were computed for RC evolutionary rates (Table 3.3.10). Evolutionary rates were compared for each RC between and within phylogenetic nodes and tips using Mann-Whitney *U* tests (H_O:$\eta_1=\eta_2$;H_A:$\eta_1\neq\eta_2$) as per *subsection 2.2.3.4.4* (Tables 3.3.11.-12.). HPs were compared for each specific RC (*eg.* NC *vs* NC) between fungal (Table 2.3.9) and plant datasets using Fisher's exact tests (H_O:$\pi(x)=\pi(X)$;H_A:$\pi(x)\neq\pi(X)$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 4$; $\alpha_B = 0.0125$) and a $P \leq \alpha_B$ was considered statistically significant (Table 3.3.13.).

3.2.4.6. Investigation for evolutionary bursts

Evolutionary bursts were detected using a *Q*-function (H_O:$Q(x)>1-\Phi(x)$;H_A:$Q(x)\leq 1-\Phi(x)$) as described in *subsection 2.2.3.3.2.2.* (Tables 3.3.13.-14; Figure 3.3.6.). Bonferroni corrections were applied to($\alpha = 0.05$; $|c_{(x)}| = 99$; $|c_{(y)}| = 50$; $\alpha_{B(x)} = 2.6e^{-04}$; $\alpha_{B(y)} = 4.8e^{-04}$) and instances where $P_{(x)} \leq \alpha_{B(x)}$ or $P_{(y)} \leq \alpha_{B(y)}$ were considered statistically significant.

3.2.5. Gene annotation (function and origin)

3.2.5.1. Functional gene annotation

Genes and gene families were assigned PFAM domains and gene ontologies as per *subsection 2.2.3.4.1.* Again, GO-slim subset groupings were completed using the curated generic ontology map so results could be easily compared to Chapter II later in this chapter. Bonferroni corrected over- and underrepresented GO-slim subsets per RC were detected using "find_enrichment.py" as per *subsection 2.2.3.4.1.* (Table 3.3.15.).

3.2.5.2. Gene origin annotation

Potential DOs were assigned to each gene and gene family using the methodology described in *subsection 2.2.3.4.3*. DOs were compared to all other DOs using a two-tailed Fisher's exact test (H_O:$\pi(x)=\pi(X)$;H_A:$\pi(x)\neq\pi(X)$) for each RC. A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = C_2^{|RC|} = 6$; $\alpha_B = 8.33e^{-03}$) and instances where $P \leq \alpha_B$ were considered statistically significant (Table 3.3.13.). HPs were compared for each specific RC (*eg.* NC *vs* NC) between fungal (Table 2.3.22) and plant datasets using Fisher's exact tests (H_O:$\pi(x)=\pi(X)$;H_A:$\pi(x)\neq\pi(X)$). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 4$; $\alpha_B = 0.0125$) and a $P \leq \alpha_B$ was considered statistically significant (Table 3.3.16.).

3.2.6: Trends between gene remodelling and genomic characteristics

GRCPs and IRCPs were calculated for each genome as per *subsection 2.2.3.5.1* (Tables 3.3.17.-18.). Data for four genomic characteristics (genome size, genome density, GC%, and BUSCO completeness) were collated for each sampled taxa as per *subsection 2.2.3.5.1.1.* Again, correlations were established using a Spearman's ρ correlation test

(H_O:$X_1 \propto X_2$;H_A:$X_1 \not\propto X_2$) between each RCP and each of the four genomic characteristics (Figure 3.3.7.). A Bonferroni correction was applied ($\alpha = 0.05$; $|c| = 5$; $\alpha_B = 0.01$) to each set and a $P \leq \alpha_B$ was considered statistically significant.

3.3. Results

3.3.1. Genome quality and characteristics

Genomic characteristics and quality (BUSCO completeness) were computed for each species (Table 3.3.1.). On average, a high level of genome completion (88.95±14.86%) was observed (Table 3.3.2.), however, a small subset of genomes ($n = 6$) exhibited completeness (C) below one standard deviation from the mean (C < 74.09%, ("below average")). All three Gymnospermae (*Picea abies, Picea glauca,* and *Pinus taeda*), *Hordeum vulgare*, *Malus domestica*, and *Zoysia japonica* were observed to possess below average completion. The three gymnosperm species have been used extensively in plant genome evolutionary analyses (Proost *et al.*, 2014; Van Bel *et al.*, 2018) without any reported quality issues, so it was hypothesised that their reported deficit in completeness (C ≤ 63.7) was due to one of two prominent reasons:

(*i.*): <u>Gene set bias in the orthologous database</u>

Gymnospermae are known to possess highly divergent gene sets when compared to other Viridiplantae lineages (De La Tore *et al.*, 2019), yet no

Table 3.3.1. Completeness and characteristics for 50 Viridiplantae genomes

Each taxon is annotated with its genome size (GS), GC content, number of genes, genome density, and genome completeness. Genome completeness is given as a percentage of expected orthologs where C is "completeness". Completeness is the cumulation of singleton (S) and duplicated (D) orthologs. Fragmented (F) and missing (M) orthologs detract from C.

	Genome characteristics				BUSCO completeness (%) (Viridiplantae_odb *v*10; *n* = 430)				
Species	**Genes (*n*)**	**GS (Mbp)**	**GC (%)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Amborella trichopoda	26846	706.495	38.1	38.00	97.4	96.5	0.9	2.1	0.5
Ananas comosus	27024	447.645	38.17	60.37	94.2	90.5	3.7	2.3	3.5
Arabidopsis thaliana	27655	119.669	36.05	231.10	98.6	98.4	0.2	1.2	0.2
Beta vulgaris	26920	566.55	37.31	47.52	98.8	97.4	1.4	0.9	0.3
Brachypodium distachyon	34310	271.299	46.41	126.47	97.7	97	0.7	2.3	0
Brassica rapa	40492	284.129	35.83	142.51	98.1	82.1	16	0.9	1
Capsicum annuum	35884	3063.86	35.36	11.71	92.3	90	2.3	3.3	4.4
Carica papaya	27768	370.419	39.01	74.96	78.6	76.3	2.3	14.2	7.2
Chlamydomonas reinhardtii	17741	107.1	61.95	165.65	97.2	96.3	0.9	1.2	1.6
Chlorella sp. NC64A	9761	46.16	65.5	211.46	84.9	83.7	1.2	8.4	6.7
Citrullus lanatus	23440	365.45	33.81	64.14	98.4	97.2	1.2	1.4	0.2
Citrus clementina	24533	301.365	35.2	81.41	90	88.6	1.4	7	3
Coccomyxa sp. C169	9994	48.8266	52.9	204.68	84.4	82.8	1.6	5.6	10
Cucumis melo	27427	374.928	34.9	73.15	91.2	90.5	0.7	5.1	3.7
Eucalyptus grandis	36349	691.43	39.99	52.57	97.2	93.7	3.5	1.9	0.9
Fragaria vesca	32381	214.373	38.91	151.05	91.9	90	1.9	2.8	5.3
Glycine max	56044	979.046	35.12	57.24	99.7	38.1	61.6	0.2	0.1
Gossypium raimondii	37505	761.565	33.53	49.25	99.8	92.1	7.7	0.2	0
Hordeum vulgare	24282	1779.49	44.9	13.65	63.3	60.5	2.8	24.9	11.8
Malus domestica	53922	703.358	39.359	76.66	64.6	55.8	8.8	22.8	12.6
Manihot esculenta	33033	582.279	38.01	56.73	99	92.3	6.7	0.7	0.3
Marchantia polymorpha	19287	225.761	42.5	85.43	97.2	96.5	0.7	1.2	1.6
Medicago truncatula	50894	412.924	34.05	123.25	97.7	91.9	5.8	0.7	1.6
Micromonas commoda	10103	21.1093	63.82	478.60	89.7	86	3.7	4.9	5.4
Musa acuminata	36528	472.231	40.73	77.35	95.5	88.8	6.7	4	0.5
Oryza brachyantha	32037	259.908	41.1	123.26	94.9	93.7	1.2	3.3	1.8
Oryza sativa ssp. *japonica*	42189	374.423	43.58	112.68	96.1	94.9	1.2	3.3	0.6
Ostreococcus lucimarinus	7805	13.2049	60.44	591.07	83.2	77.9	5.3	6	10.8

	Genome characteristics				BUSCO completeness (%) (Viridiplantae_odb *v*10; *n* = 430)				
Species	**Genes (*n*)**	**GS (Mbp)**	**GC (%)**	**Density (*n*/Mbp)**	**C**	**S**	**D**	**F**	**M**
Phalaenopsis equestris	29431	1064.2	35.1	27.66	87.9	85.1	2.8	9.1	3
Phyllostachys edulis	31987	2075	44.66	15.42	76.5	62.1	14.4	12.1	11.4
Physcomitrella patens	32926	472.081	33.89	69.75	97.9	84.4	13.5	0.9	1.2
Picea abies	66632	11961.4	37.9	5.57	34.4	29.5	4.9	40.9	24.7
Picea glauca	28909	24633.1	44.58	1.17	63.7	57.9	5.8	19.8	16.5
Pinus taeda	84446	22103.6	34.8	3.82	46.8	39.8	7	20.9	32.3
Populus trichocarpa	42950	434.29	34.15	98.90	97.2	79.1	18.1	2.3	0.5
Prunus persica	26873	227.569	37.67	118.09	99.5	97.4	2.1	0.2	0.3
Ricinus communis	31221	350.622	34.4	89.04	95.8	95.3	0.5	2.8	1.4
Selaginella moellendorffii	22285	212.315	45.3	104.96	94.4	87.4	7	2.1	3.5
Setaria italica	34584	405.868	46.17	85.21	97.7	95.6	2.1	1.6	0.7
Solanum lycopersicum	34725	828.349	35.7	41.92	97.5	97	0.5	1.9	0.6
Solanum tuberosum	39028	705.934	35.06	55.29	84.7	83.5	1.2	4.4	10.9
Sorghum bicolor	34211	709.345	44.16	48.23	96.3	95.1	1.2	3.5	0.2
Spirodela polyrhiza	19623	42.72	136.67	459.34	91.9	91.2	0.7	6.3	1.8
Theobroma cacao	29232	324.88	34.99	89.98	99.8	99.8	0	0	0.2
Triticum aestivum	103537	14547.3	46.05	7.12	99.1	9.3	89.8	0.9	0
Vitis vinifera	26346	486.197	35.03	54.19	96.5	95.3	1.2	3	0.5
Volvox carteri	15544	137.684	55.3	112.90	84.4	82.8	1.6	4.7	10.9
Zea mays	39498	2135.08	46.91	18.50	86.2	79.5	6.7	10	3.8
Zostera marina	20450	203.914	38.9	100.29	98.6	97	1.6	0.2	1.2
Zoysia japonica	53625	334.384	42.6	160.37	49.3	43.7	5.6	30.2	20.5

Table 3.3.2. Descriptive statistics for Viridiplantae genomic characteristics

The minimum (Min), maximum (Max), median (η), quartiles (η_x), mean (μ), standard error (SE), and coefficient of variation (CV) are presented for each genomic characteristic.

	Genome characteristics (*n* = 50)				BUSCO completeness (%) (Viridiplantae_odb *v*10; *n* = 430)				
	Genes (*n*)	**GS (Mbp)**	**GC (%)**	**Density (*n*/Mbp)**	**Complete**	**Singleton**	**Duplicated**	**Fragmented**	**Missing**
Min	7805	13.2	33.53	1.17	34.4	9.3	0	0	0
$\eta_{0.25}$	24470	227.1	35.12	48.05	84.85	79.4	1.2	1.2	0.5
η ($\eta_{0.50}$)	31604	409.4	38.91	77.01	95.65	90	2.2	2.9	1.6
$\eta_{0.75}$	37886	722.4	45	123.3	97.75	95.38	6.7	6.475	6.825
Max	103537	24633	136.7	591.1	99.8	99.8	89.8	40.9	32.3
μ	33604±17380	1979±5121	43.33±15.74	107±116.9	88.95±14.86	82.15±20.19	6.808±15.08	6.212±8.576	4.834±6.892
SE	2458	724.2	2.225	16.54	2.102	2.855	2.132	1.213	0.9746
CV (%)	51.72	258.7	36.32	109.3	16.71	24.57	221.5	138.1	142.6

Gymnospermae were used for the construction of Viridiplantae_odb *v.*10 (Kriventseva *et al*., 2018).

(*ii*.): Gene evolution patterns in Gymnospermae

Gene evolution in Gymnospermae is markedly different to other Viridiplantae lineages (De La Tore *et al*., 2017). In gymnosperms, genes slowly acquire insertions in their nucleotide and resultant amino acid sequences. These accumulations may result in the misidentification of a singleton BUSCO gene (S) as a fragment (F) if an ortholog exceeds two standard deviations from the average length used to construct the model BUSCO gene (Waterhouse *et al.*, 2018). This argument is strengthened when the mean fragmentation rate is gymnosperms (27.2±11.88%) is compared to non-gymnosperms (4.87±6.45%).

The reported lack of genome completeness observed in *H. vulgare* (C = 63.3%) could be due to the use of an older, perhaps incomplete or inferior quality assembly genome on PLAZA *v*4.0. The reported genome source was Ensembl Genomes ASM32608v1 (ftp://ftp.ensemblgenomes.org/pub/plants/release-35/). however, an updated *H. vulgare* (IBSC_v2) genome assembly has been uploaded to Ensembl Genomes (Mascher *et al.*, 2017). The sum of protein coding genes in ASM32608v1 was reported by PLAZA *v*4 to be 24,282 whereas a total of 39,841 were reported in IBSC_v2 (a difference of 15,559 genes (an inflation of 164.08%). A BUSCO analysis using the conserved eukaryote ortholog database reported 98% completion during reassembly (Mascher *et al.,* 2017).

The completion deficit (C = 49.3%) in *Z. japonica* is derived from high levels of fragmented and missing BUSCO sets (F = 30.2%; M = 20.5%). The *Z. japonica* sequencing

project reported genomic completeness of 94.7% (Tanaka *et al.*, 2016) using the CEGMA core eukaryote gene dataset (Parra *et al.*, 2007). Previous studies have placed *Z. japonica* as an early diverging member of the Poaceae PACMAD clade (*eg.* Van Bel *et al.,* 2018), and it displays a relatively long branch (κ = 0.274) in our phylogenetic analyses which provides support for divergence based fragmentation and losses. It may be the case, as with the gymnosperms, that the missing and fragmented BUSCO sets may reflect the evolutionary history of *Z. japonica*.

Finally, the completion deficit observed in *M. domestica* (C = 64.6%) is attributable to a considerable amount of fragmentation in BUSCO set hits (F = 22.8%). Again, as in the gymnosperms and *Z. japonica*, this may merely reflect the evolutionary history of *M. domestica*. Extensive WGD and genomic reorganisations have been reported during the recent evolutionary history of *M. domestica* and closely related species (*eg. Pyrae* spp.) compared to other, more distantly related, plant lineages (Xiang *et al.*, 2016). These evolutionary dynamics have resulted in significant differences in gene content, architecture and sizes between the genomes of *M. domestica* and other Angiospermae (Velasco *et al.*, 2010). Such events may have caused considerable difference in gene length resulting in higher rates of fractionation as expected in the Gymnospermae.

Genome incompleteness can be observed even in high quality assemblies of type species such as *Caenorhabditis elegans* (C = 85%) and *Homo sapiens* (C = 89%) (Simão *et al*., 2015). The six genomes with "below average" completion were retained for further analyses as they were determined to be of high quality (Proost *et al.,* 2015; Van Bel *et al.*, 2018) are of high economic importance, and (with the exception of *H. vulgare*) plausible evolutionary reasons could be determined to explain their reported incompleteness.

3.3.2. Extent of gene remodelling in Viridiplantae

In total, of the 1,672,377 sequences in this dataset, a total of 488,979 genes (29.28%) were excluded due to the removal of low complexity sequences or due to being a singleton, resulting in a sample of 1,183,398 genes within 81,112 families (Table 3.3.3; Figure 3.3.1.). These results are in stark contrast to what was observed in the fungi (Table 2.3.5.), where significant differences ($P \leq \alpha_B \leq 0.0125$ ($Genes_{(a)}$: $P \leq \alpha_B \leq 0.01$)) were observed between every comparison (Table 3.3.3.).

3.3.3. Variance in gene family sizes

As observed in the fungal dataset (section 2.3.3.), gene family sizes displayed considerable variation for each RC ($129\% \leq CV \leq 427\%$) with considerable bias observed towards smaller families ($\eta_{0.25} \leq 3$, $\eta_{0.50} \leq 11$, and $\eta_{0.75} \leq 31$ (Table 3.3.4.). Again, nested composites were observed to display the greatest variance (CV = 427%), the widest range ($n \in \{2, 3 \ldots, 9093\}$) and largest mean (37.99±162.1) when compared to other RCs ($129\% \leq CV \leq 229\%$); $n_{SC} \in \{2, 3 \ldots, 335\}$, $n_{SN} \in \{2, 3 \ldots, 442\}$, $n_{NR} \in \{2, 3 \ldots, 152\}$; and $\mu_{(SC, SN, NR)}$ = 19.11±24.71, 5.819±13.33, 6.698±10.62 respectively). Family sizes were reported to be significantly different ($P \leq \alpha_B \leq 8.33e^{-03}$) to each other when using a Mann-Whitney U-test ($P \leq 4.79e^{-121}$) (Table 3.3.5.). Again, stark contrasts ($P \leq 2.37e^{-208}$) were observed when families belonging to each specific RC were compared to their counterparts in the fungal dataset (Tables 2.3.7., 3.3.6). These contrasts are due to the observations of consistently larger means and medians observed in Viridiplantae family sizes when compared to the fungi (Tables 2.3.6., 3.3.4).

3.3.4. Comparison of evolutionary rates

Table 3.3.3. Extent of remodelled genes and families in Viridiplantae

The number (*n*) of genes in the entire dataset ($Genes_{(a)}$), the sum of genes in the sampled dataset ($Genes_{(s)}$; $Genes_{(a)}$ – Excluded), and the number of gene families attributed to each RC are presented with their associated proportion (%) within their respective populations. The "Excluded" category is only observed in $Genes_{(a)}$ as genes in this category were not used for sampling by CompositeSearch, and thus were excluded from $Genes_{(s)}$ and Families respectively.

	n			**%**		
	$Genes_{(a)}$	**$Genes_{(s)}$**	**Families**	**$Genes_{(a)}$**	**$Genes_{(s)}$**	**Families**
NC	774886	774886	20399	46.33	65.48	25.15
SC	48440	48440	2535	2.90	4.09	3.13
SN	196083	196083	33695	11.72	16.57	41.54
NR	163989	163989	244823	9.81	13.86	30.18
Excluded	488979	N/A		29.24	N/A	

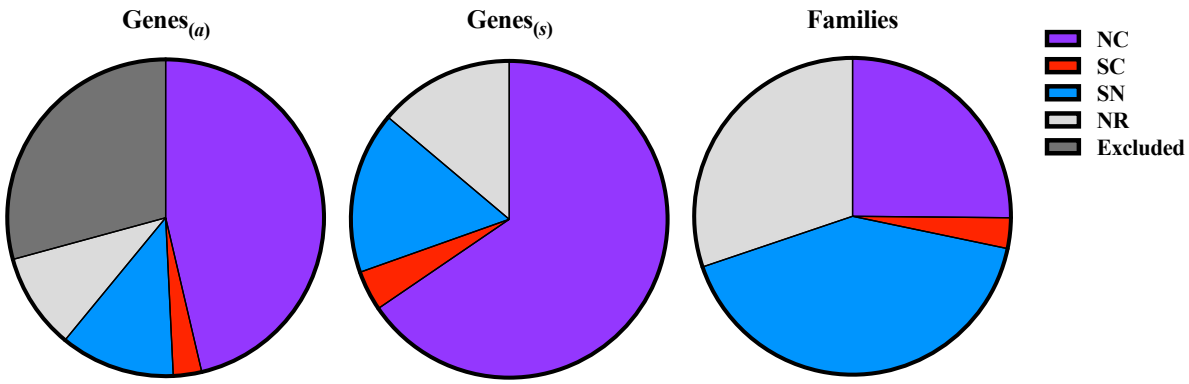

Figure 3.3.1. Extent of remodelled gene and family extent in Viridiplantae

Each pie chart represents one of three datasets (Genes$_{(a)}$, Genes$_{(s)}$, and Families) from Table 3.3.2. Again, the "Excluded" category is only observed in Genes$_{(a)}$ as genes in this category were not used for sampling by CompositeSearch, and thus were excluded from Genes$_{(s)}$ and Families respectively.

Table 3.3.4. Comparison of remodelling extent in fungi and plants

The remodelling extent is presented for plants ($\%_V$) and fungi ($\%_F$) for the 4 RCs in the "sampled" datasets ($Genes_{(s)}$ and Families) and for the 4 RCs and "excluded" genes in the entire dataset ($Genes_{(a)}$). All comparisons were observed to be significantly different ($Genes_{(s)}$, Families: $P \leq \alpha_B \leq 0.0125$; $Genes_{(a)}$: $P \leq \alpha_B \leq 0.01$)

		$\%_V$	$\%_F$	P
$Genes_{(a)}$	NC	46.33	30.67	0
	SC	2.90	2.62	$3.02e^{-45}$
	SN	11.72	16.90	0
	NR	9.81	23.57	0
	Excluded	29.24	26.24	0
$Genes_{(s)}$	NC	65.48	41.58	0
	SC	4.09	3.55	$3.77e^{-89}$
	SN	16.57	22.91	0
	NR	13.86	31.96	0
Families	NC	25.15	21.25	$2.64e^{-77}$
	SC	3.13	2.54	$8.44e^{-13}$
	SN	41.54	26.51	0
	NR	30.18	49.69	0

Table 3.3.5. Descriptive statistics for gene family sizes in Viridiplantae

Statistics describing family size distribution characteristics were tabulated for each RC. Each RC is assigned a mean (μ), median (η), quartiles ($\eta_{0.25, 0.50, 0.75}$), minima, maxima, and CV.

	NC	SC	SN	NR
n	20399	2535	33695	24483
Min	2	2	2	2
$\eta_{0.25}$	3	3	2	2
η ($\eta_{0.50}$)	8	11	2	3
$\eta_{0.75}$	31	25	4	6
Max	9093	335	442	192
μ	37.99±162.1	19.11±24.71	5.819±13.33	6.698±10.62
SE	1.135	0.491	0.072	0.0679
CV (%)	426.7	129.3	229.0	158.6

Table 3.3.6. Comparison of family sizes between RCs in Viridiplantae

$RC_{(a)}$ and $RC_{(b)}$ refer to the RCs being tested. The Mann-Whitney U statistic (U) is provided alongside a *P*-value for each comparison. All comparisons were considered statistically significant ($P \leq a_B \leq 8.33e^{-03}$).

$RC_{(a)}$	$RC_{(b)}$	U	*P*
NC	SC	$6.55e^{08}$	0
NC	SN	$5.30e^{08}$	0
NC	NR	$2.76e^{07}$	$5.35e^{-197}$
SC	SN	$2.23e^{08}$	$4.79e^{-121}$
SC	NR	$5.55e^{06}$	0
SN	NR	$7.37e^{06}$	0

Table 3.3.7. Comparison of RC family sizes between fungi and plants

Family sizes from each RC were compared between the fungal and plant datasets. A U statistic is provided alongside a *P*-value for each comparison. All comparisons were reported to be significantly different ($P \leq \alpha \leq 0.0125$).

	U	P
NC	$5.77e^{08}$	0
SC	$4.23e^{08}$	0
SN	$1.55e^{07}$	0
NR	$1.34e^{07}$	$2.37e^{-208}$

3.3.4.1. Phylogenetic annotation

After two failed attempts to reconstruct a phylogeny (Figures 3.3.2.-3.3.3.), a phylogeny was constructed using 99 ubiquitous gene family alignments from the OrthoDB Viridiplantae_odb 10 dataset (as used for BUSCO analyses) with a scaffold obtained from Van Bel *et al.*, 2018 (Figures 3.3.4.). Internal nodes were annotated as per the "-apo" function in TNT (Figure 3.3.5.).

3.3.4.2. Remodelled genes are more homoplastic than non-remodelled genes in Viridiplantae

As observed in the fungal dataset (*subsection 2.3.4.2.*), significant differences ($P \leq \alpha_B \leq 8.33e^{-03}$) were observed between all comparisons ($P \leq 4.60e^{-19}$) when sampled from across the entire phylogeny (Table 3.3.8.). As observed in fungal families, remodelled gene families were observed to be significantly more homoplastic ($0.458 \leq HP \leq 0.609$) than non-remodelled families ($HP_{NR} = 0.35$; $P \leq 1.69e^{-140}$). SC were observed to be the most homoplastic families ($HP_{SC} = 0.609$) amongst remodelled families ($HP_{NC} = 0.497$, $HP_{SN} = 0.458$; $P \leq 1.14e^{-26}$). In fungi, a supremum difference (sup_{HP}) of 0.54 is observed between RC, however, in Viridiplantae a sup_{HP} of 0.151 is observed. With the prominent exception of NC ($P = 0.044$), significant differences ($P \leq \alpha_B \leq 0.0125$) were observed for each specific RC comparison ($P \leq 2.16e^{-08}$) between the fungal and plant datasets when sampled from across entire phylogenies (Table 3.3.9.).

As observed when sampled from all nodes in the Viridiplantae phylogeny, all comparisons were significantly different ($P \leq 4.78e^{-14}$) when sampled from the subset of internal nodes. However, unlike what was observed when sampled from the entire phylogeny

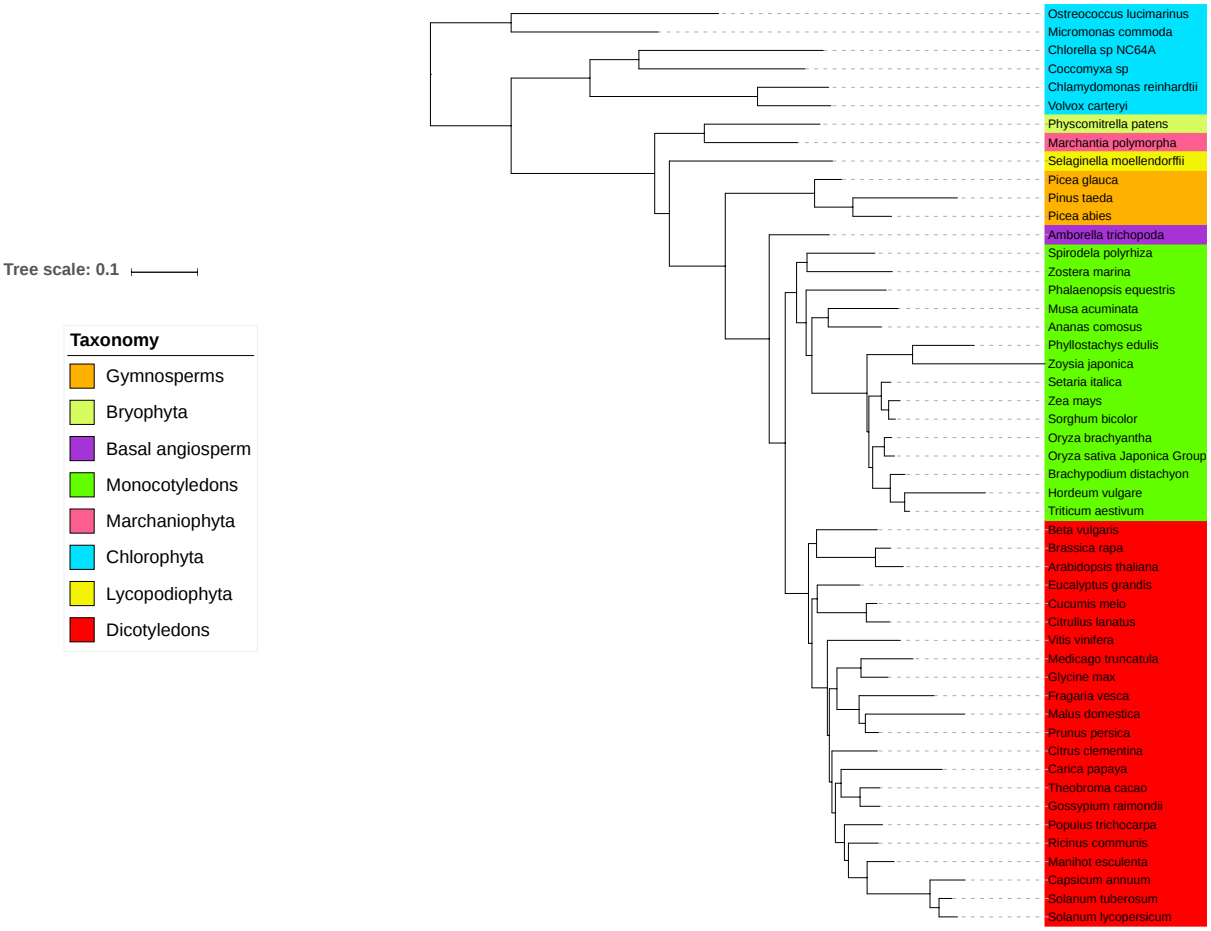

Figure 3.3.2. Unconstrained "BUSCO" phylogeny

A maximum likelihood superalignment of the 99 ubiquitous genes derived from the BUSCO viridiplantae_odb *v*10 database. All lineages are correctly placed, except for those within the dicots where all clades are incorrectly placed. This is likely due to dicots evolving through continuous rounds of WGD followed by degradation. As this tree resolved all other lineages so well, it was chosen as the basis for the constrained tree.

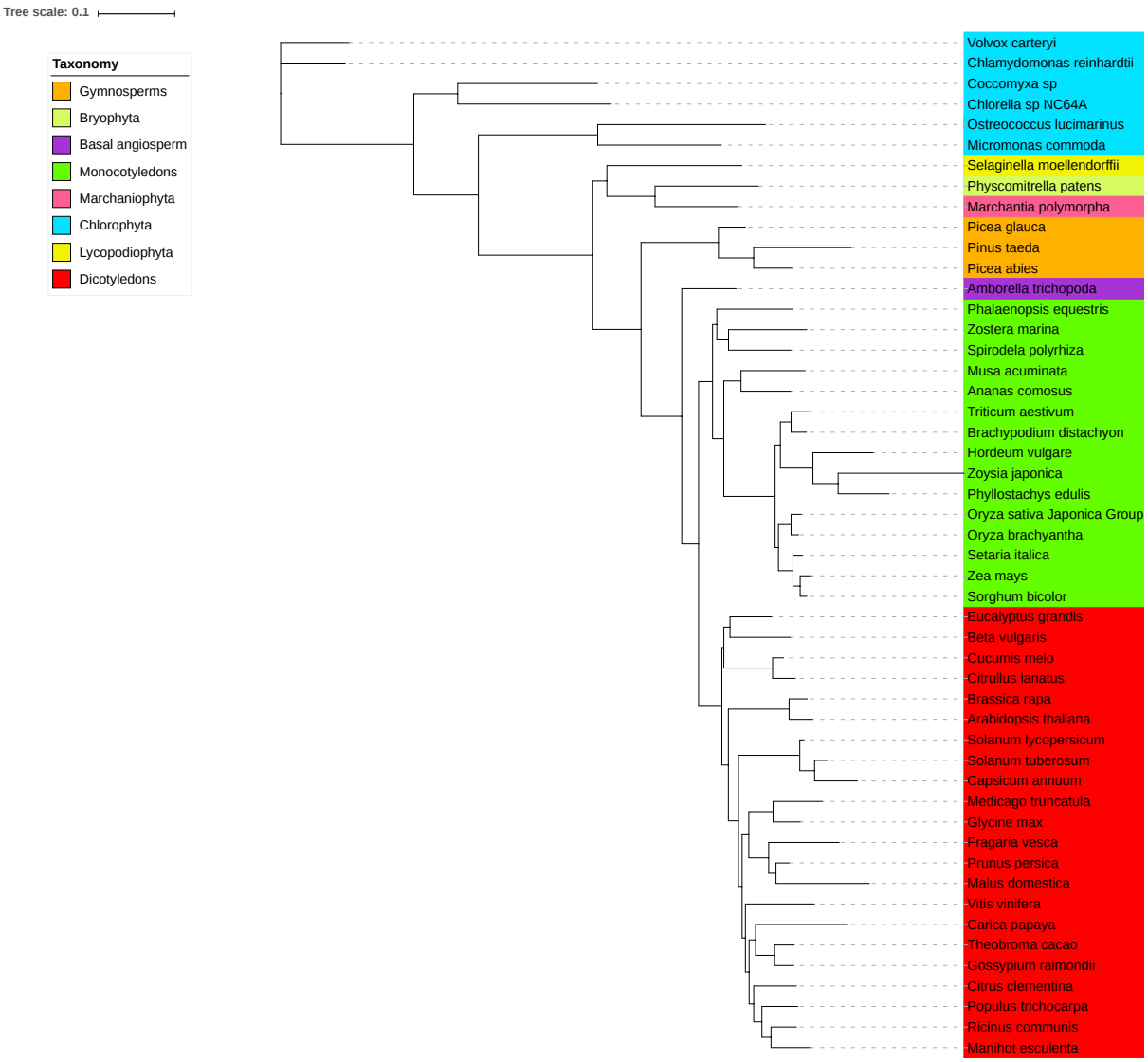

Figure 3.3.3. Unconstrained "Cicarelli" phylogeny

Maximum likelihood tree made from a superalignment of the universal orthologs described by Cicarelli *et al*., (2006) without constraints. In this phylogeny *Selaginella* is incorrectly placed as basal to all other embryophytes and the Chlorophyceae are incorrectly placed as the outgroup in place of the Chlorophyta, despite being reported as late diverging algae (Guiry and Guiry, 2018). Dicotyledon clades are highly misplaced, for example, *Vitis vinifera*, is an early diverging dicot, however, it is placed deep within the clade, and the mustard plants (Brassicaceae) are placed as ancestral to the nightshades (Solanaceae).

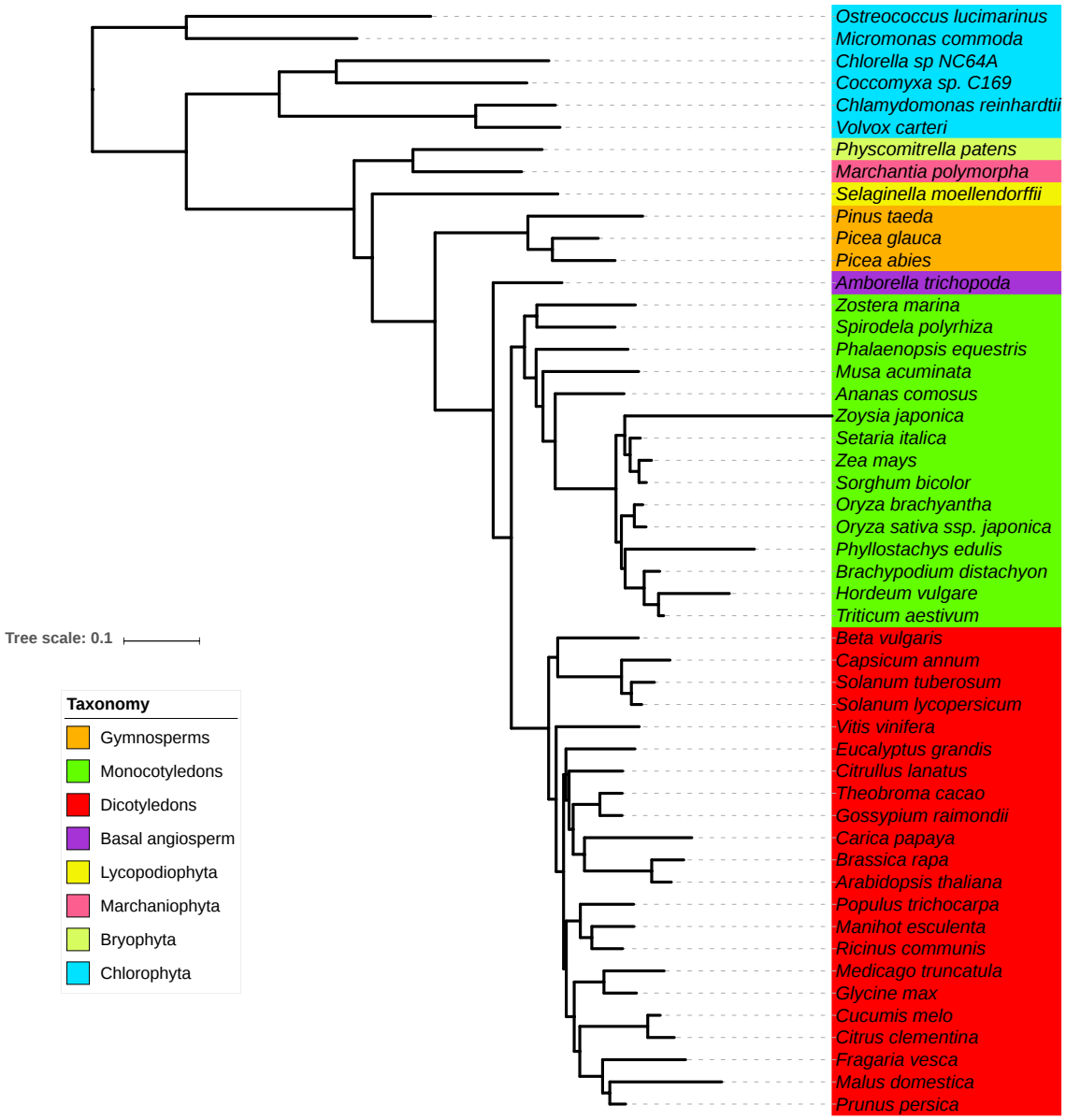

Figure 3.3.4. Constrained "BUSCO" phylogeny

Consensus phylogeny constructed from a concatenated superalignment of 99 ubiquitous orthologous genes with a manually constructed scaffold inferred from Van Bel *et al.* (2018). No bootstrap supports are present as a scaffold was used to restrict phylogenetic topologies. The root appears to bifurcate the Chlorophyta, however rerooting at the LCA of *Physcomitrella patens* and *Chlamydomonas reinhardtii* resolves the topology. We chose to leave the root where it is in support of new evidence that Mamiellophyceae is an outgroup to the rest of Chlorophyta (Guiry and Guiry, 2018).

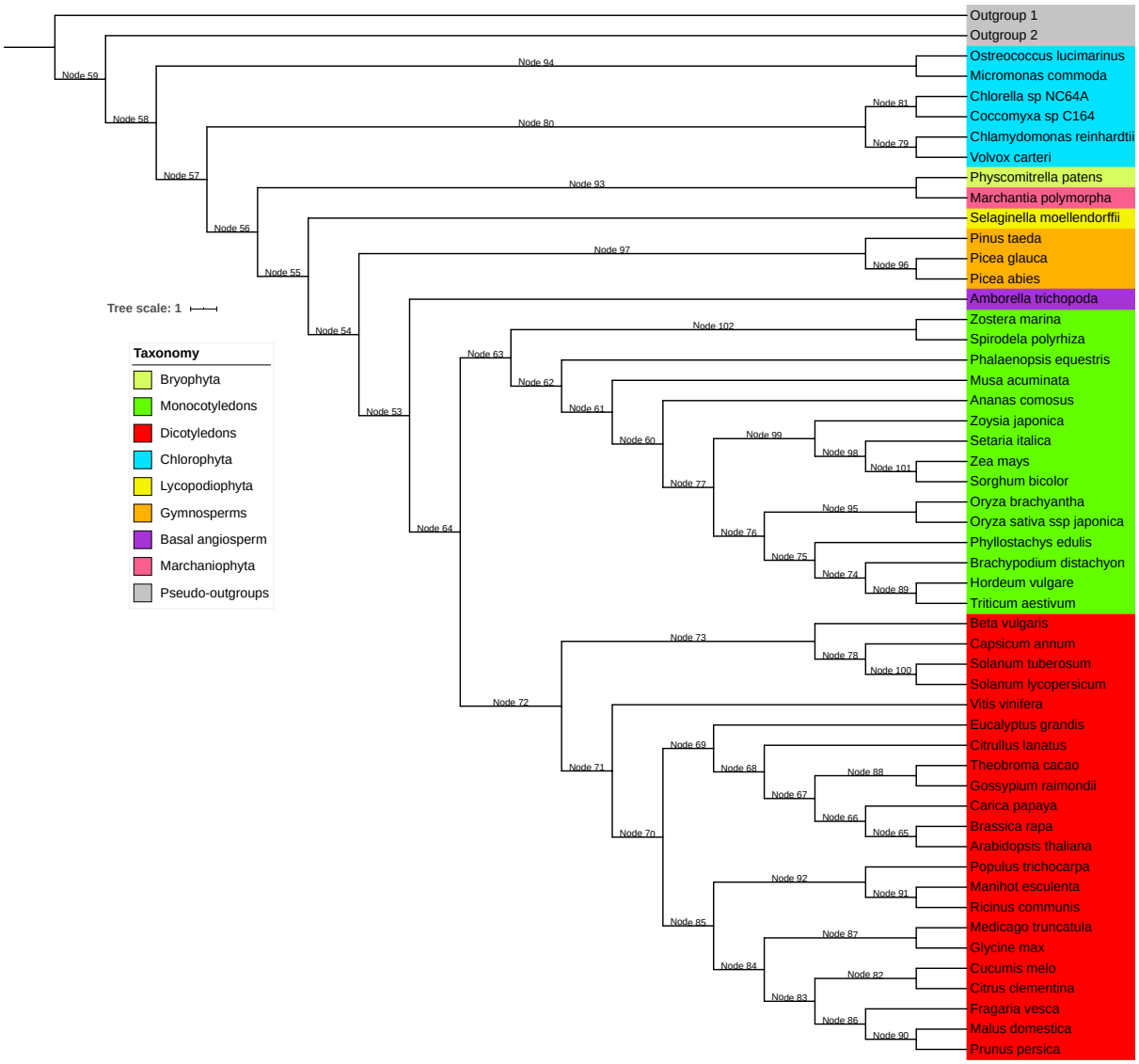

Figure 3.3.5. Annotated Viridiplantae phylogeny

Internal node IDs (as computed by TNT) were assigned to each node from the constrained "BUSCO" phylogeny (Figure 3.3.4.)

Table 3.3.8. Comparisons of homoplasy in Viridiplantae remodelling categories

HPs were computed for families in each RC when sampled from across the entire phylogeny and for the subset of exclusively internal nodes. The number of homoplastic families per RC per set (n_H) is presented alongside the total number of families per RC (n_{all}). All comparisons were observed to be statistically significant ($P \leq \alpha_B \leq 8.33e^{-03}$).

		RC$_{(a)}$			**RC$_{(b)}$**			***P***
Set	**Comparison**	n_H	n_{all}	**HP**	n_H	n_{all}	**HP**	
Entire phylogeny	NC *vs* SC	10144	20399	0.497	1545	2535	0.609	$1.14e^{-26}$
	NC *vs* SN	10144	20399	0.497	15423	33695	0.458	$4.60e^{-19}$
	NC *vs* NR	10144	20399	0.497	8559	24483	0.350	$2.58e^{-219}$
	SC *vs* SN	1545	2535	0.609	15423	33695	0.458	$2.49e^{-49}$
	SC *vs* NR	1545	2535	0.609	8559	24483	0.350	$1.69e^{-140}$
	SN *vs* NR	15423	33695	0.458	8559	24483	0.350	$8.23e^{-152}$
Exclusively internal branches	NC *vs* SC	1617	20399	0.079	319	2535	0.126	$4.78e^{-14}$
	NC *vs* SN	1617	20399	0.079	763	33695	0.023	$8.05e^{-205}$
	NC *vs* NR	1617	20399	0.079	1235	24483	0.050	$1.88e^{-35}$
	SC *vs* SN	319	2535	0.126	763	33695	0.023	$1.50e^{-115}$
	SC *vs* NR	319	2535	0.126	1235	24483	0.050	$3.56e^{-43}$
	SN *vs* NR	763	33695	0.023	1235	24483	0.050	$1.05e^{-72}$

Table 3.3.9. Comparison of homoplastic proportions between fungi and plants

HPs for each RC (per phylogenetic set) were computed between the Fungi (HP_F) and Viridiplantae (HP_V) respectively. Every comparison was observed to be significantly different ($P \leq \alpha_B \leq 0.0125$) with the exceptions of NC when sampled from across the entire phylogeny ($P = 0.044$) and for SC when sampled from across the subset of internal nodes ($P = 0.824$)

		HP_F	HP_V	*P*
Entire Phylogeny	**NC**	0.487	0.497	0.044
	SC	0.464	0.609	**$4.21e^{-23}$**
	SN	0.433	0.458	**$2.16e^{-08}$**
	NR	0.305	0.350	**$6.83e^{-32}$**
Exclusively internal nodes	**NC**	0.120	0.079	**$3.71e^{-40}$**
	SC	0.123	0.126	0.824
	SN	0.073	0.023	**$3.63e^{-178}$**
	NR	0.058	0.050	**$9.24e^{-05}$**

and in fungi, SN were significantly less homoplastic ($HP_{SN} = 0.023$; $P \leq 1.05e^{-72}$) than all other RCs ($0.05 \leq HP \leq 0.126$). Again, SC were observed to be more homoplastic than all other RCs ($HP_{SC} = 0.126$). The Viridiplantae HP_{SC} and HP_{NR} ($HP_{NR} = 0.05$) for this subset is comparable to their fungal counterparts ($HP_{SC} = 0.123$; $HP_{NR} = 0.053$), however HP_{NC} ($HP_{NC} = 0.079$) and HP_{SN} are considerably lower than their respective fungal counterparts ($HP_{NC} = 0.12$; $HP_{SN} = 0.073$). With the prominent exception of NC ($P = 0.824$), significant differences ($P \leq \alpha_B \leq 0.0125$) were observed for each specific RC comparison ($P \leq 9.24e^{-05}$) between the fungal and plant datasets when sampled from across phylogenetic internal node subsets.

These results are in agreement with those observed for fungi and, as such, suggest that remodelled genes, particularly composite genes, are either highly likely to disobey Dollo's Law of Irreversibility or more likely to be epaktologous as first postulated in *subsection 2.3.4.2.*

3.3.4.3. Evolutionary rate dynamics between remodelling categories

When sampled from across the phylogeny, birth rates displayed considerable variation within each RC (*eg.* $0.215 \leq f_b \leq 227$; $139 \leq CV \leq 173$) (Table 3.3.10.). Significant differences ($P \leq \alpha \leq 8.33e^{-03}$) were observed between each SC and each other RC, ($P \leq 7.57e^{-18}$), where SC were observed to have a much lower f_b, but not between any other comparison ($P \geq 0.56$) (Table 3.3.11.). High variance rates were observed for decay rates within each RC (*eg.* $0.002 \leq f_d \leq 50.2$; $112 \leq CV \leq 140$), with significant rate differences observed between each comparison ($P \leq 5.27e^{-04}$) with the exception of SN *vs* NR ($P = 0.28$). While these rates are high, they are considerably low when compared to the fungi (f_b: $225.1 \leq CV \leq 276.1$; f_d: $315.8 \leq 378.8$) where a greater range in rates (*eg.* $0.00846 \leq f_b \leq 416.4$; $0.00846 \leq f_d \leq 322$) was also observed (*subsection 2.3.4.8*; Table 2.3.7.). In contrast to the fungi, lower f_d were observed in Viridiplantae when compared to f_b.

Table 3.3.10. Descriptive statistics for RC evolutionary rates in Viridiplantae

Descriptive statistics for evolutionary birth (f_b) and decay (f_d) rates were calculated for the entire phylogeny and for each phylogenetic subset (exclusively leaf nodes and exclusively internal nodes).

		f_b				f_d			
		NC	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**
Entire phylogeny (*n* = 99)	Min.	0.351	0.078	0.215	0.298	0.002	0.002	0.002	0.002
	$\eta_{0.25}$	4.25	0.765	3	4.06	1.17	0.25	0.403	0.486
	η ($\eta_{0.50}$)	8.45	1.46	8.07	7.89	3.88	0.646	1.37	1.54
	$\eta_{0.75}$	13.7	2.37	19	13.9	7.09	1.12	2.07	3.01
	Max.	170	25.7	227	159	50.7	6.67	20.1	25.8
	μ	13±19.3	2.1±2.92	15.4±26.8	12.6±18.4	5.41±6.87	0.903±1.01	1.89±2.56	2.39±3.36
	SE	1.94	0.293	2.69	1.85	0.69	0.102	0.258	0.337
	CV (%)	149	139	173	146	127	112	135	140
Exclusively internal nodes (*n* = 49)	Min.	0.351	0.078	0.215	0.298	0.002	0.002	0.002	0.002
	$\eta_{0.25}$	3.24	0.511	1.89	2.55	0.713	0.171	0.306	0.438
	η ($\eta_{0.50}$)	7.49	1.12	3.34	5.07	2.21	0.489	0.787	1.08
	$\eta_{0.75}$	9.74	2.01	7.29	11.2	4.21	0.808	1.69	1.94
	Max.	21.9	3.34	26.6	33.4	9.58	2.21	4.7	6.5
	μ	7.43±5.06	1.27±0.851	5.45±5.69	7.34±6.54	2.93±2.65	0.607±0.562	1.14±1.14	1.47±1.47
	SE	0.723	0.122	0.814	0.934	0.379	0.0802	0.163	0.21
	CV (%)	68.2	67.2	104	89.1	90.8	92.5	99.5	100
Exclusively leaf nodes (*n* = 50)	Min.	1.33	0.261	1.11	1.6	0.004	0.004	0.004	0.004
	$\eta_{0.25}$	6.82	1.04	7.98	5.73	2.66	0.352	0.868	0.681
	η ($\eta_{0.50}$)	10.7	1.67	16	9.34	5.69	0.863	1.76	2.01
	$\eta_{0.75}$	19.3	3.6	28.7	19.2	9.51	1.51	3	3.95
	Max.	170	25.7	227	159	50.7	6.67	20.1	25.8
	μ	18.4±25.7	2.93±3.86	25.2±34.7	17.7±24.1	7.85±8.67	1.19±1.25	2.63±3.28	3.3±4.33
	SE	3.63	0.546	4.91	3.41	1.23	0.177	0.464	0.612
	CV (%)	140	132	137	136	111	105	125	131

Table 3.3.11. Comparison of RC evolutionary rates in Viridiplantae

$RC_{(a)}$ and $RC_{(b)}$ refers to the RCs being compared for each rate in each phylogenetic subset. At least 3 significant differences ($P \leq \alpha_B \leq 8.33e^{-03}$) were observed in each analysis run. Interestingly, comparisons with SC were observed to be significantly different in every instance ($P \leq 1.00e^{-03}$) with the exception of SC *vs* SN when sampled from across internal nodes ($P = 0.02$).

		$RC_{(a)}$	$RC_{(b)}$	U	*P*
Entire phylogeny	f_b	NC	SC	8927	**$1.75e^{-23}$**
		NC	SN	5137	0.56
		NC	NR	5123.5	0.58
		SC	SN	1430.5	**$7.57e^{-18}$**
		SC	NR	1131	**$8.87e^{-21}$**
		SN	NR	4834.5	0.87
	f_d	NC	SC	7894	**$1.14e^{-13}$**
		NC	SN	6990.5	**$2.18e^{-07}$**
		NC	NR	6639.5	**$1.62e^{-05}$**
		SC	SN	3502.5	**$5.27e^{-04}$**
		SC	NR	3111	**$9.10e^{-06}$**
		SN	NR	4468	0.28
Exclusive internal nodes	f_b	NC	SC	2323	**$1.43e^{-13}$**
		NC	SN	1625	$9.83e^{-03}$
		NC	NR	1346.5	0.51
		SC	SN	384	**$2.42e^{-09}$**
		SC	NR	295	**$4.70e^{-11}$**
		SN	NR	999.5	0.08
	f_d	NC	SC	1970	**$7.02e^{-07}$**
		NC	SN	1759.5	**$4.49e^{-04}$**
		NC	NR	1663.5	**$4.41e^{-03}$**
		SC	SN	918.5	0.02
		SC	NR	763.5	**$8.06e^{-04}$**
		SN	NR	1095.5	0.29
Exclusive leaf nodes	f_b	NC	SC	2204	**$1.03e^{-12}$**
		NC	SN	981	0.12
		NC	NR	1244	0.76
		SC	SN	194	**$8.82e^{-13}$**
		SC	NR	199	**$1.14e^{-12}$**
		SN	NR	1437	0.09
	f_d	NC	SC	2011	**$8.65e^{-09}$**
		NC	SN	1793	**$2.60e^{-05}$**
		NC	NR	1715	**$2.60e^{-04}$**
		SC	SN	791.5	**$3.70e^{-03}$**
		SC	NR	737	**$1.00e^{-03}$**
		SN	NR	1083.5	0.41

When sampled from the subset of exclusively internal nodes, considerably less variance was observed within both f_b (*eg.* $0.298 \leq f_b \leq 33.4$; $67.2 \leq CV \leq 104$) and f_d (*eg.* $0.002 \leq f_d \leq 9.58$; $90.8 \leq CV \leq 100$). Again, significant differences ($P \leq \alpha \leq 8.33e^{-03}$) between RC f_b were only observed when compared to SC ($P \leq 2.42e^{-09}$) but not between other RCs ($P \geq 9.83e^{-03}$). Comparatively, significant differences between RC f_d were observed when compared to NC ($P \leq 4.41e^{-03}$) and between SC and NR ($P = 8.06e^{-04}$) but not for SC *vs* SN or SN *vs* NR ($P \geq 0.02$). These trends are markedly different to what was observed in fungi, where significant differences between all comparisons (except SN *vs* NR) were observed for f_b and only when compared to SC for f_d.

Finally, when sampled from the subset of exclusively leaf nodes (speciation event nodes), a relatively high degree of variance was observed for both f_b (*eg.* $1.11 \leq f_b \leq 227$; $132 \leq CV \leq 140$) and f_d (*eg.* $0.004 \leq f_d \leq 50.7$; $105 \leq CV \leq 131$) respectively. Significant RC f_b differences were only observed in comparisons with SC ($P \leq 1.03e^{-12}$), however, significant f_d differences were observed between all comparisons ($P \leq 3.70e^{-03}$) except SN *vs* NR ($P = 0.41$). Birth rate differences are comparable to the fungi where significant differences were also observed between all comparisons to SC ($P \leq 1.93e^{-21}$) but also for NC *vs* NR ($4.73e^{-03}$). Decay rates were quite different to fungi. As mentioned above, in Viridiplantae significant differences in RC f_d during speciation were observed between all comparisons except SN *vs* NR, however in fungi significant differences were only observed when compared to SC.

Significant differences ($P \leq \alpha_B \leq 0.0125$) were observed for every comparison when rates for each specific RC (*eg.* NC *vs* NC) were compared between internal and leaf nodes ($P \leq 8.95e^{-03}$) (Table 3.3.12.) where higher medians were observed for leaf nodes in each comparison (Table 3.3.7.).

As observed in fungi, these results suggest that despite the considerable variance observed within RCs, with the notable exception of SC, their birth rates are relatively constant.

Table 3.3.12. Comparison of evolutionary rates between leaf and internal nodes in Viridipantae

Evolutionary rates were compared between leaf and internal node subsets from the Viridiplantae phylogeny. Every comparison was observed to be significantly different ($P \leq \alpha_B \leq 0.0125$).

	RC	U	*P*
f_b	NC	686.5	$1.66e^{-04}$
	SC	725	$4.73e^{-04}$
	SN	393	$5.91e^{-09}$
	NR	665	$9.01e^{-05}$
f_d	NC	703.5	$2.66e^{-04}$
	SC	851	$8.95e^{-03}$
	SN	797.5	$2.80e^{-03}$
	NR	851	$8.95e^{-03}$

The significant evolutionary rate increases observed between internal nodes and leaf nodes in each RC is likely due to genomic innovation during speciation (Gogarten and Townsend, 2005). These results also support the hypotheses discussed in *subsection 3.3.4.2*. which posits that remodelling events are likely to flout Dollo's Law and that they are more likely to be epaktologous during speciation. Alongside instances of convergent evolution, the simultaneous increase in both birth rate and homoplasy suggest that homoplasy is a major driving force behind these phenomena.

Synapomorphic families observed at internal nodes were evolutionarily "successful", meaning they were retained post speciation events. Such differences were expected as genome sequencing provides a "snapshot in time" (Klimke *et al*., 2011) of a given genome, with no guarantee that an innovation will persist. The evolutionary rate increase is likely influenced by homoplasy and epaktologous events which would be consistent with the HP differences observed in Table 3.3.7.

Evolutionary rates for each RC (from each phylogenetic set) were compared to each of their respective counterparts from the fungal dataset (Table 3.3.13.). Rates were found to be remarkably similar when sampled from the entire phylogenies and from the subsets of internal nodes. When sampled from across the phylogeny, significant differences ($P \leq \alpha_B \leq 0.0125$) were only observed in SC birth and decay rates ($P \leq 9.54e^{-04}$), and when sampled from the internal node subsets, significant differences were only observed in SC decay rates ($P = 1.89e^{-03}$). When sampled from the subset of leaf nodes however, significant differences were observed, again, for SC birth rates ($P = 2.06e^{-05}$), for NC decay rates ($P = 5.33e^{-03}$), and for both SN birth and decay rates ($P \leq 0.01$). These results suggest that, with the exception of SC, remodelled gene families are retained at a consistent rate within both plants and fungi. The significant differences observed between the leaf node subsets (speciation events) are likely due to the increased number of polyploidisation events during Viridiplantae speciation

Table 3.3.13. Comparison of RC evolutionary rates between fungi and plants

Each RC evolutionary rate (from each phylogenetic set) was compared to its counterpart from the fungal dataset. SC are the only RC to display significant rate differences ($P \leq \alpha_B \leq 0.0125$) between fungi and plants when sampled from either (*a.*) the entire phylogeny or (*b.*) the subset of internal nodes. Significant differences are also observed in SN and NR rate evolution when sampled from just leaf nodes.

			U	*P*
Entire phylogeny	f_b	**NC**	8675	0.014
		SC	7348	**$2.06e^{-05}$**
		SN	9517	0.186
		NR	11884	0.06
	f_d	**NC**	8848	0.026
		SC	8053	**$9.54e^{-04}$**
		SN	12126	0.027
		NR	12147	0.025
Exclusively internal nodes	f_b	**NC**	2216	0.195
		SC	1934	0.016
		SN	2902	0.167
		NR	3127	0.024
	f_d	**NC**	2073	0.064
		SC	1753	**$1.89e^{-03}$**
		SN	2793	0.339
		NR	2698	0.559
Exclusively leaf nodes	f_b	**NC**	2092	0.028
		SC	1696	**$2.27e^{-04}$**
		SN	1794	**$9.08e^{-04}$**
		NR	2815	0.5992
	f_d	**NC**	2289	0.1464
		SC	2204	0.076
		SN	3360	**0.01**
		NR	3415	**$5.33e^{-03}$**

resulting in greater family genesis when compared to the fungi (Albertin and Marullo, 2012; Clark and Donohue, 2018)

3.3.4.4. Gene remodelling is 'clocklike' in Viridiplantae

Only one site of evolutionary bursts ($P \leq \alpha_B \leq 5.05e^{-04}$), the speciation of *Triticum aestivum*, was reported across the Viridiplantae phylogeny where bursts in NC ($f_b = 170.35$/Ma; $P = 4.87e^{-04}$) and NR ($f_b = 158.98$/Ma; $P = 4.65e^{-04}$) were observed (Table 3.3.13.), no bursts in f_d were observed at any site along the phylogeny. Similarly to these conservative results, no bursts ($P \leq 1.02e^{-03}$) were observed when sampled from the subset of internal branches (Table 3.3.14.). These sparse results are consistent with those observed in fungi (*subsection 2.3.4.3.*) suggesting that evolution *via* gene remodelling is relatively clocklike. The bursts observed during the speciation of *T. aestivum* are consistent with the two rounds of allopolyploidisation attributed to its genome evolution (Matsuoka, 2011), where these massive redundant genetic influxes promote rampant remodelling events (Leonard and Richards, 2012).

3.3.4.5. Functional overrepresentations in remodelling categories

(*i.*): Overrepresentations exclusive to nested composites

NC exclusive BP ontologs were observed to be significantly overrepresented ($P_B \leq 0.05$) for growth (GO:0040007; $P_B = 1.04e^{-03}$), macromolecule modification (GO:0005975, GO:0019538, GO:0043412, GO:0044260, GO:0044267, GO:1901564; $P_B \leq 0.0127$), biological quality regulation (GO:0007165, GO:0042592, GO:0050794, GO:0065008; $P_B \leq 4.76e^{-}$

Table 3.3.14. Evolutionary rates in Viridiplantae

For each RC at each node, the sum of gained (T_b) and lost (T_d) synapomorphies are presented alongside their rates (f_x) and the probability that a burst was observed ($P(f_x)$). λ values used for Box-Cox transformations and branch lengths (κ) used to calculate rates are also provided.

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 0.037	SC λ = 0.026	SN λ = 0.044	NR λ = 0.106	NC λ = 0.295	SC λ = 0.295	SN λ = 0.27	NR λ = 0.261
Amborella trichopoda	0.09600	830	149	1224	1018	142	19	31	23	10.753	1.941	15.851	13.185	1.850	0.259	0.414	0.311	0.386	0.350	0.272	0.304	0.643	0.717	0.727	0.800
Ananas comosus	0.09610	710	134	793	697	424	62	107	124	9.190	1.745	10.263	9.022	5.493	0.814	1.396	1.616	0.450	0.393	0.397	0.440	0.338	0.389	0.427	0.445
Arabidopsis thaliana	0.03099	342	58	314	336	192	27	59	98	13.750	2.365	12.628	13.509	7.737	1.122	2.405	3.969	0.292	0.276	0.336	0.295	0.236	0.281	0.265	0.190
Beta vulgaris	0.11180	593	130	626	763	283	35	87	66	6.600	1.455	6.966	8.488	3.155	0.400	0.978	0.744	0.584	0.468	0.517	0.463	0.505	0.611	0.530	0.646
Brachypodium distachyon	0.02567	260	60	396	389	170	17	29	50	12.632	2.952	19.214	18.875	8.276	0.871	1.452	2.468	0.324	0.202	0.223	0.192	0.217	0.366	0.415	0.321
Brassica rapa	0.04744	662	117	714	520	147	19	54	63	17.358	3.089	18.720	13.640	3.875	0.524	1.440	1.676	0.214	0.189	0.230	0.292	0.445	0.532	0.418	0.435
Capsicum annum	0.06913	759	91	1237	466	366	59	104	153	13.657	1.653	22.247	8.392	6.595	1.078	1.887	2.767	0.295	0.415	0.190	0.467	0.283	0.294	0.336	0.288
Carica papaya	0.14703	815	101	1492	605	1036	197	338	374	6.894	0.862	12.613	5.120	8.761	1.673	2.864	3.168	0.567	0.681	0.336	0.643	0.201	0.161	0.216	0.250
Chlamydomonas reinhardtii	0.11045	355	67	315	623	83	7	20	18	4.004	0.765	3.554	7.018	0.945	0.090	0.236	0.214	0.764	0.724	0.711	0.533	0.770	0.874	0.813	0.844
Chlorella sp. NC64A	0.28594	370	82	384	508	147	26	46	38	1.612	0.361	1.673	2.211	0.643	0.117	0.204	0.169	0.947	0.914	0.867	0.859	0.821	0.845	0.831	0.866
Citrullus lanatus	0.07601	537	86	637	482	281	34	95	137	8.792	1.422	10.427	7.893	4.609	0.572	1.569	2.255	0.468	0.478	0.393	0.490	0.392	0.505	0.392	0.347
Citrus clementina	0.04019	555	112	805	589	399	71	143	197	17.185	3.493	24.913	18.236	12.364	2.225	4.451	6.120	0.217	0.155	0.166	0.201	0.117	0.094	0.111	0.097
Coccomyxa sp. C169	0.25699	469	82	327	431	73	12	19	30	2.272	0.401	1.585	2.088	0.358	0.063	0.097	0.150	0.900	0.896	0.875	0.869	0.877	0.903	0.897	0.876
Cucumis melo	0.02156	638	104	1044	661	372	61	97	131	36.817	6.050	60.210	38.142	21.491	3.572	5.646	7.605	0.055	0.053	0.044	0.053	0.034	0.028	0.070	0.063
Eucalyptus grandis	0.09636	807	128	942	652	407	79	157	161	10.416	1.663	12.156	8.418	5.260	1.031	2.037	2.088	0.399	0.413	0.347	0.466	0.352	0.309	0.313	0.370
Fragaria vesca	0.11475	518	76	666	473	461	78	183	242	5.618	0.833	7.220	5.131	5.001	0.855	1.992	2.630	0.647	0.693	0.506	0.642	0.367	0.372	0.320	0.302

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 0.037	SC λ = 0.026	SN λ = 0.044	NR λ = 0.106	NC λ = 0.295	SC λ = 0.295	SN λ = 0.27	NR λ = 0.261
Glycine max	0.04811	964	151	1277	865	108	24	51	64	24.917	3.925	32.998	22.360	2.814	0.646	1.343	1.678	0.120	0.127	0.114	0.148	0.537	0.466	0.439	0.434
Gossypium raimondii	0.03441	698	112	855	732	162	35	59	95	25.235	4.079	30.903	26.462	5.885	1.300	2.166	3.466	0.117	0.118	0.125	0.111	0.317	0.234	0.295	0.225
Hordeum vulgare	0.09878	696	84	1507	442	1539	259	548	712	8.765	1.069	18.963	5.571	19.366	3.270	6.904	8.966	0.469	0.597	0.227	0.615	0.044	0.036	0.044	0.043
Malus domestica	0.15309	1481	183	2846	1014	1430	178	420	379	12.025	1.493	23.100	8.236	11.611	1.452	3.416	3.083	0.342	0.458	0.181	0.474	0.131	0.201	0.170	0.257
Manihot esculenta	0.06078	411	84	558	483	237	42	81	93	8.420	1.737	11.424	9.891	4.864	0.879	1.676	1.921	0.485	0.395	0.365	0.406	0.376	0.363	0.372	0.395
Marchantia polymorpha	0.14886	599	138	462	689	65	12	17	25	5.007	1.160	3.864	5.758	0.551	0.108	0.150	0.217	0.689	0.563	0.689	0.603	0.838	0.854	0.862	0.842
Medicago truncatula	0.08460	720	117	1101	817	323	60	119	175	10.587	1.733	16.181	12.011	4.757	0.896	1.762	2.584	0.392	0.396	0.267	0.336	0.383	0.356	0.357	0.308
Micromonas commoda	0.23061	342	81	287	484	0	0	0	0	1.848	0.442	1.551	2.612	0.005	0.005	0.005	0.005	0.931	0.877	0.879	0.827	0.981	0.978	0.977	0.979
Musa acuminata	0.13163	878	129	1130	637	230	38	56	51	8.295	1.227	10.673	6.021	2.180	0.368	0.538	0.491	0.492	0.540	0.385	0.588	0.604	0.633	0.676	0.729
Node100	0.01798	306	42	384	483	119	30	39	56	21.211	2.971	26.600	33.440	8.291	2.142	2.764	3.938	0.158	0.200	0.153	0.071	0.216	0.102	0.225	0.192
Node101	0.01674	173	34	133	220	53	13	25	23	12.913	2.598	9.945	16.401	4.008	1.039	1.930	1.781	0.315	0.243	0.407	0.233	0.435	0.307	0.330	0.417
Node102	0.02130	130	26	67	130	121	11	25	22	7.641	1.575	3.966	7.641	7.116	0.700	1.516	1.341	0.525	0.435	0.682	0.502	0.260	0.439	0.402	0.498
Node53	0.08190	655	89	212	212	27	3	10	15	9.950	1.365	3.231	3.231	0.425	0.061	0.167	0.243	0.417	0.495	0.735	0.778	0.863	0.906	0.852	0.830
Node54	0.08793	534	27	102	54	27	11	18	30	7.558	0.396	1.455	0.777	0.396	0.170	0.268	0.438	0.530	0.899	0.888	0.970	0.869	0.795	0.796	0.749
Node55	0.02909	206	20	34	24	7	3	4	8	8.840	0.897	1.495	1.068	0.342	0.171	0.214	0.384	0.465	0.666	0.884	0.950	0.881	0.794	0.825	0.769
Node56	0.22691	710	67	123	106	0	0	0	0	3.892	0.372	0.679	0.586	0.005	0.005	0.005	0.005	0.772	0.909	0.960	0.981	0.980	0.978	0.977	0.979
Node57	0.12906	143	19	29	30	0	0	0	0	1.386	0.192	0.289	0.298	0.010	0.010	0.010	0.010	0.961	0.977	0.990	0.994	0.977	0.971	0.971	0.973
Node58	0.50451	615	85	141	133	0	0	0	0	1.517	0.212	0.350	0.330	0.002	0.002	0.002	0.002	0.953	0.972	0.987	0.993	0.984	0.985	0.983	0.984
Node60	0.02119	200	37	102	178	99	21	44	65	11.784	2.228	6.038	10.494	5.862	1.290	2.638	3.869	0.350	0.297	0.561	0.385	0.318	0.237	0.238	0.196
Node61	0.01405	180	25	96	141	53	17	12	33	16.003	2.299	8.576	12.555	4.774	1.591	1.149	3.006	0.240	0.286	0.452	0.321	0.382	0.175	0.484	0.264
Node62	0.02035	131	20	51	82	49	8	20	19	8.058	1.282	3.174	5.067	3.052	0.549	1.282	1.221	0.503	0.522	0.739	0.646	0.515	0.518	0.452	0.524
Node63	0.02243	56	6	38	41	30	6	10	18	3.157	0.388	2.160	2.326	1.717	0.388	0.609	1.052	0.830	0.902	0.823	0.850	0.660	0.619	0.648	0.563
Node64	0.02900	244	28	69	96	38	9	7	18	10.496	1.242	2.999	4.155	1.671	0.428	0.343	0.814	0.396	0.535	0.753	0.709	0.666	0.592	0.760	0.626
Node65	0.09392	505	84	489	801	245	36	123	158	6.692	1.124	6.481	10.607	3.254	0.489	1.640	2.103	0.579	0.576	0.539	0.381	0.497	0.553	0.379	0.368

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 0.037	SC λ = 0.026	SN λ = 0.044	NR λ = 0.106	NC λ = 0.295	SC λ = 0.295	SN λ = 0.27	NR λ = 0.261
Node66	0.01998	44	7	22	57	139	21	55	75	2.797	0.497	1.430	3.606	8.703	1.368	3.481	4.725	0.859	0.851	0.890	0.749	0.203	0.219	0.166	0.149
Node67	0.01040	35	5	15	33	31	4	15	14	4.301	0.717	1.911	4.062	3.823	0.597	1.911	1.792	0.741	0.746	0.845	0.715	0.449	0.491	0.332	0.415
Node68	0.00912	61	17	30	72	37	8	18	13	8.446	2.452	4.223	9.945	5.177	1.226	2.588	1.907	0.484	0.263	0.665	0.404	0.357	0.253	0.244	0.397
Node69	0.00813	48	12	19	32	30	5	9	17	7.488	1.987	3.056	5.043	4.737	0.917	1.528	2.751	0.533	0.341	0.748	0.648	0.384	0.348	0.400	0.290
Node70	0.01482	132	31	98	140	17	8	6	16	11.149	2.682	8.299	11.820	1.509	0.754	0.587	1.425	0.372	0.232	0.463	0.342	0.687	0.414	0.657	0.481
Node71	0.01493	100	19	58	72	13	1	9	12	8.405	1.664	4.910	6.075	1.165	0.166	0.832	1.082	0.486	0.412	0.622	0.585	0.736	0.798	0.573	0.556
Node72	0.05433	185	36	96	109	12	3	5	10	4.252	0.846	2.218	2.515	0.297	0.091	0.137	0.251	0.745	0.688	0.818	0.835	0.890	0.872	0.870	0.826
Node73	0.01699	115	20	67	109	34	7	10	12	8.483	1.536	4.973	8.044	2.559	0.585	0.804	0.951	0.482	0.446	0.618	0.483	0.563	0.498	0.581	0.589
Node74	0.03014	182	27	140	152	36	5	15	24	7.543	1.154	5.812	6.306	1.525	0.247	0.659	1.030	0.530	0.565	0.572	0.571	0.685	0.726	0.630	0.568
Node75	0.01000	72	12	36	53	31	10	14	16	9.069	1.615	4.597	6.709	3.976	1.367	1.864	2.112	0.455	0.425	0.641	0.549	0.438	0.219	0.340	0.367
Node76	0.01214	38	8	22	25	9	6	3	4	3.989	0.921	2.353	2.659	1.023	0.716	0.409	0.511	0.765	0.656	0.806	0.823	0.758	0.432	0.729	0.722
Node77	0.08536	246	39	165	261	188	31	93	113	3.594	0.582	2.416	3.813	2.750	0.466	1.368	1.659	0.796	0.810	0.801	0.734	0.544	0.568	0.433	0.438
Node78	0.08946	319	48	245	254	87	17	28	34	4.443	0.680	3.416	3.541	1.222	0.250	0.403	0.486	0.730	0.763	0.721	0.754	0.727	0.724	0.732	0.731
Node79	0.26466	185	36	334	867	28	6	2	3	0.873	0.174	1.572	4.074	0.136	0.033	0.014	0.019	0.986	0.982	0.877	0.714	0.930	0.939	0.965	0.963
Node80	0.12727	35	7	21	37	0	0	0	0	0.351	0.078	0.215	0.371	0.010	0.010	0.010	0.010	0.999	0.998	0.994	0.992	0.977	0.971	0.971	0.973
Node81	0.08041	117	24	120	166	14	7	6	3	1.823	0.386	1.869	2.580	0.232	0.124	0.108	0.062	0.933	0.903	0.849	0.829	0.906	0.839	0.889	0.929
Node82	0.09485	245	42	282	342	260	37	99	88	3.222	0.563	3.706	4.492	3.418	0.498	1.310	1.166	0.825	0.819	0.700	0.685	0.482	0.548	0.446	0.536
Node83	0.01238	21	4	13	13	83	21	44	54	2.207	0.502	1.404	1.404	8.427	2.207	4.514	5.518	0.905	0.849	0.892	0.925	0.212	0.096	0.109	0.116
Node84	0.01549	18	5	8	24	29	8	11	16	1.524	0.481	0.722	2.005	2.406	0.722	0.962	1.363	0.953	0.858	0.957	0.876	0.579	0.429	0.534	0.493
Node85	0.00857	49	10	22	44	10	2	2	7	7.249	1.595	3.335	6.524	1.595	0.435	0.435	1.160	0.547	0.430	0.727	0.559	0.676	0.587	0.718	0.538
Node86	0.03514	50	8	35	56	56	14	33	55	1.803	0.318	1.273	2.015	2.015	0.530	1.202	1.980	0.934	0.933	0.905	0.875	0.623	0.529	0.471	0.386
Node87	0.04433	316	75	299	463	78	8	25	34	8.882	2.129	8.406	13.001	2.214	0.252	0.729	0.981	0.463	0.314	0.459	0.309	0.600	0.722	0.606	0.581
Node88	0.04035	201	25	152	261	47	9	21	32	6.219	0.801	4.711	8.067	1.478	0.308	0.677	1.016	0.608	0.708	0.634	0.482	0.691	0.678	0.624	0.572
Node89	0.02399	422	46	401	228	76	16	25	40	21.904	2.434	20.817	11.858	3.987	0.880	1.346	2.123	0.149	0.266	0.205	0.341	0.437	0.362	0.438	0.365

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 0.037	SC λ = 0.026	SN λ = 0.044	NR λ = 0.106	NC λ = 0.295	SC λ = 0.295	SN λ = 0.27	NR λ = 0.261
Node90	0.01453	152	23	156	181	111	23	54	75	13.081	2.052	13.423	15.561	9.576	2.052	4.702	6.498	0.311	0.329	0.318	0.249	0.177	0.111	0.101	0.086
Node91	0.02020	231	40	189	236	130	13	49	58	14.268	2.522	11.685	14.576	8.057	0.861	3.075	3.629	0.279	0.253	0.358	0.270	0.224	0.370	0.197	0.213
Node92	0.02361	94	19	55	129	18	4	9	13	4.999	1.052	2.947	6.841	1.000	0.263	0.526	0.737	0.690	0.603	0.757	0.542	0.761	0.713	0.680	0.648
Node93	0.08282	555	122	332	505	23	4	7	9	8.339	1.845	4.994	7.589	0.360	0.075	0.120	0.150	0.489	0.370	0.617	0.504	0.877	0.890	0.881	0.876
Node94	0.12911	255	53	282	488	0	0	0	0	2.463	0.520	2.723	4.705	0.010	0.010	0.010	0.010	0.885	0.840	0.775	0.670	0.977	0.971	0.971	0.973
Node95	0.02211	212	35	189	301	51	11	13	20	11.967	2.023	10.675	16.968	2.922	0.674	0.787	1.180	0.344	0.334	0.385	0.223	0.527	0.452	0.587	0.533
Node96	0.03743	286	35	688	505	197	17	56	21	9.523	1.195	22.863	16.790	6.570	0.597	1.891	0.730	0.435	0.551	0.184	0.226	0.284	0.491	0.336	0.650
Node97	0.12781	335	58	300	230	205	43	44	44	3.265	0.573	2.925	2.245	2.002	0.428	0.437	0.437	0.822	0.814	0.758	0.857	0.625	0.592	0.717	0.749
Node98	0.01674	233	44	156	301	41	3	4	6	17.365	3.339	11.651	22.412	3.117	0.297	0.371	0.519	0.214	0.167	0.359	0.147	0.509	0.686	0.746	0.719
Node99	0.01212	72	9	78	121	42	6	16	15	7.484	1.025	8.099	12.507	4.408	0.718	1.743	1.640	0.534	0.614	0.470	0.322	0.406	0.431	0.360	0.441
Oryza brachyantha	0.01621	660	114	847	658	660	86	261	336	50.654	8.813	64.985	50.501	50.654	6.667	20.078	25.825	0.026	0.021	0.038	0.027	0.001	0.002	0.001	0.001
Oryza sativa ssp. *japonica*	0.02049	846	105	1028	917	263	41	84	76	51.349	6.426	62.382	55.653	16.005	2.546	5.153	4.668	0.025	0.047	0.041	0.021	0.070	0.070	0.085	0.151
Ostreococcus lucimarinus	0.32801	349	68	292	422	0	0	0	0	1.325	0.261	1.110	1.602	0.004	0.004	0.004	0.004	0.964	0.955	0.921	0.909	0.982	0.982	0.980	0.981
Phalaenopsis equestris	0.12656	853	145	1512	925	604	69	203	156	8.382	1.433	14.850	9.089	5.938	0.687	2.002	1.541	0.487	0.475	0.290	0.438	0.314	0.445	0.319	0.459
Phyllostachys edulis	0.17589	928	111	1304	548	859	120	322	355	6.561	0.791	9.216	3.877	6.073	0.855	2.281	2.514	0.587	0.712	0.430	0.729	0.308	0.372	0.280	0.316
Physcomitrella patens	0.17592	1068	200	1092	1358	38	9	6	19	7.548	1.419	7.718	9.596	0.275	0.071	0.049	0.141	0.530	0.479	0.485	0.418	0.895	0.895	0.932	0.881
Picea abies	0.08793	986	98	2944	1294	290	59	102	77	13.944	1.399	41.606	18.295	4.111	0.848	1.455	1.102	0.287	0.485	0.081	0.201	0.427	0.375	0.415	0.551
Picea glauca	0.06567	918	213	1478	965	487	15	89	52	17.382	4.048	27.974	18.271	9.230	0.303	1.702	1.002	0.214	0.120	0.143	0.201	0.187	0.682	0.367	0.575
Pinus taeda	0.15683	1336	117	3323	712	96	16	28	28	10.590	0.935	26.328	5.647	0.768	0.135	0.230	0.230	0.392	0.650	0.154	0.610	0.799	0.828	0.817	0.836
Populus trichocarpa	0.07555	810	129	1096	804	197	37	74	85	13.334	2.137	18.036	13.235	3.255	0.625	1.233	1.414	0.304	0.313	0.239	0.302	0.497	0.477	0.464	0.483
Prunus persica	0.02606	578	140	503	614	149	27	36	68	27.600	6.721	24.025	29.316	7.150	1.335	1.764	3.289	0.099	0.042	0.173	0.092	0.259	0.226	0.357	0.239
Ricinus communis	0.04582	615	107	721	510	583	103	264	335	16.701	2.928	19.575	13.854	15.833	2.820	7.185	9.110	0.226	0.205	0.219	0.287	0.072	0.054	0.040	0.041
Selaginella moellendorffii	0.25038	817	190	904	1016	53	10	11	11	4.058	0.948	4.490	5.045	0.268	0.055	0.060	0.060	0.760	0.645	0.648	0.648	0.897	0.913	0.924	0.930
Setaria italica	0.01857	358	83	517	548	80	17	30	58	24.015	5.619	34.651	36.725	5.418	1.204	2.074	3.947	0.128	0.063	0.107	0.058	0.343	0.258	0.308	0.191

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**	**NC** $\lambda = 0.037$	**SC** $\lambda = 0.026$	**SN** $\lambda = 0.044$	**NR** $\lambda = 0.106$	**NC** $\lambda = 0.295$	**SC** $\lambda = 0.295$	**SN** $\lambda = 0.27$	**NR** $\lambda = 0.261$
Solanum lycopersicum	0.01877	667	95	1036	505	215	31	78	94	44.208	6.353	68.628	33.487	14.295	2.118	5.228	6.287	0.036	0.048	0.034	0.071	0.089	0.105	0.082	0.092
Solanum tuberosum	0.03555	643	83	1153	572	547	77	162	143	22.503	2.935	40.324	20.022	19.149	2.726	5.696	5.032	0.143	0.204	0.085	0.176	0.046	0.059	0.069	0.135
Sorghum bicolor	0.01486	423	61	456	489	205	27	73	107	35.445	5.183	38.203	40.962	17.221	2.341	6.186	9.028	0.060	0.074	0.093	0.045	0.059	0.085	0.057	0.042
Spirodela polyrhiza	0.10855	768	145	894	741	612	84	158	167	8.800	1.671	10.242	8.491	7.015	0.973	1.819	1.922	0.467	0.411	0.398	0.463	0.264	0.329	0.347	0.394
Theobroma cacao	0.03461	506	63	555	450	287	58	119	172	18.198	2.297	19.956	16.188	10.337	2.118	4.307	6.209	0.200	0.286	0.214	0.237	0.158	0.105	0.118	0.094
Triticum aestivum	0.01182	1620	244	2160	1511	69	8	17	31	170.354	25.747	227.103	158.899	7.356	0.946	1.892	3.363	**$4.65e^{-04}$**	$6.16e^{-03}$	$2.05e^{-03}$	**$4.87e^{-04}$**	0.250	0.338	0.336	0.233
Vitis vinifera	0.11489	1046	148	1105	687	655	89	203	162	11.320	1.611	11.957	7.438	7.092	0.973	2.206	1.762	0.366	0.426	0.352	0.512	0.261	0.328	0.290	0.420
Volvox carteri	0.11651	338	64	355	445	94	15	24	23	3.614	0.693	3.795	4.755	1.013	0.171	0.267	0.256	0.794	0.757	0.694	0.667	0.759	0.794	0.797	0.824
Zea mays	0.02163	1004	106	1404	662	440	64	151	227	57.717	6.145	80.688	38.076	25.326	3.733	8.729	13.094	0.018	0.052	0.025	0.054	0.021	0.024	0.023	0.015
Zostera marina	0.13537	777	170	878	671	433	45	111	118	7.139	1.569	8.066	6.167	3.983	0.422	1.028	1.092	0.553	0.437	0.471	0.579	0.437	0.596	0.516	0.553
Zoysia japonica	0.27919	899	70	1456	730	1382	170	394	405	4.004	0.316	6.482	3.252	6.153	0.761	1.757	1.806	0.764	0.934	0.539	0.776	0.304	0.411	0.358	0.413

Table 3.3.15. Evolutionary rates across internal nodes of the Viridiplantae phylogeny

For each RC at each node, the sum of gained (T_b) and lost (T_d) synapomorphies are presented alongside their rates (f_x) and the probability that a burst was observed ($P(f_x)$). λ values used for Box-Cox transformations and branch lengths (κ) used to calculate rates are also provided.

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 2.0	SC λ = 1.9	SN λ = 0.12	NR λ = 0.08	NC λ = 0.34	SC λ = 0.4	SN λ = 0.32	NR λ = 0.35
Node100	0.01798	306	42	384	483	119	30	39	56	21.211	2.971	26.600	33.440	8.291	2.142	2.764	3.938	0.020	0.053	0.013	0.008	0.066	0.029	0.102	0.078
Node101	0.01674	173	34	133	220	53	13	25	23	12.913	2.598	9.945	16.401	4.008	1.039	1.930	1.781	0.140	0.086	0.147	0.096	0.264	0.186	0.195	0.301
Node102	0.0213	130	26	67	130	121	11	25	22	7.641	1.575	3.966	7.641	7.116	0.700	1.516	1.341	0.401	0.298	0.469	0.360	0.097	0.325	0.270	0.398
Node53	0.0819	655	89	212	212	27	3	10	15	9.950	1.365	3.231	3.231	0.425	0.061	0.167	0.243	0.259	0.374	0.548	0.688	0.827	0.886	0.818	0.807
Node54	0.08793	534	27	102	54	27	11	18	30	7.558	0.396	1.455	0.777	0.396	0.170	0.268	0.438	0.407	0.861	0.800	0.934	0.835	0.754	0.749	0.710
Node55	0.02909	206	20	34	24	7	3	4	8	8.840	0.897	1.495	1.068	0.342	0.171	0.214	0.384	0.322	0.592	0.794	0.906	0.849	0.752	0.785	0.735
Node56	0.22691	710	67	123	106	0	0	0	0	3.892	0.372	0.679	0.586	0.005	0.005	0.005	0.005	0.722	0.872	0.928	0.952	0.969	0.969	0.967	0.968
Node57	0.12906	143	19	29	30	0	0	0	0	1.386	0.192	0.289	0.298	0.010	0.010	0.010	0.010	0.934	0.949	0.981	0.977	0.966	0.961	0.960	0.963
Node58	0.50451	615	85	141	133	0	0	0	0	1.517	0.212	0.350	0.330	0.002	0.002	0.002	0.002	0.925	0.942	0.974	0.974	0.973	0.976	0.974	0.974
Node60	0.02119	200	37	102	178	99	21	44	65	11.784	2.228	6.038	10.494	5.862	1.290	2.638	3.869	0.178	0.137	0.306	0.234	0.145	0.122	0.112	0.082
Node61	0.01405	180	25	96	141	53	17	12	33	16.003	2.299	8.576	12.555	4.774	1.591	1.149	3.006	0.070	0.126	0.189	0.171	0.206	0.074	0.362	0.139
Node62	0.02035	131	20	51	82	49	8	20	19	8.058	1.282	3.174	5.067	3.052	0.549	1.282	1.221	0.372	0.408	0.555	0.529	0.360	0.416	0.325	0.430
Node63	0.02243	56	6	38	41	30	6	10	18	3.157	0.388	2.160	2.326	1.717	0.388	0.609	1.052	0.790	0.865	0.690	0.777	0.550	0.538	0.562	0.478
Node64	0.029	244	28	69	96	38	9	7	18	10.496	1.242	2.999	4.155	1.671	0.428	0.343	0.814	0.232	0.425	0.576	0.604	0.558	0.505	0.702	0.557
Node65	0.09392	505	84	489	801	245	36	123	158	6.692	1.124	6.481	10.607	3.254	0.489	1.640	2.103	0.474	0.478	0.281	0.230	0.337	0.458	0.244	0.245
Node66	0.01998	44	7	22	57	139	21	55	75	2.797	0.497	1.430	3.606	8.703	1.368	3.481	4.725	0.822	0.809	0.805	0.653	0.058	0.107	0.059	0.048

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 2.0	SC λ = 1.9	SN λ = 0.12	NR λ = 0.08	NC λ = 0.34	SC λ = 0.4	SN λ = 0.32	NR λ = 0.35
Node67	0.0104	35	5	15	33	31	4	15	14	4.301	0.717	1.911	4.062	3.823	0.597	1.911	1.792	0.684	0.689	0.728	0.612	0.281	0.384	0.197	0.299
Node68	0.00912	61	17	30	72	37	8	18	13	8.446	2.452	4.223	9.945	5.177	1.226	2.588	1.907	0.346	0.104	0.444	0.254	0.181	0.136	0.117	0.278
Node69	0.00813	48	12	19	32	30	5	9	17	7.488	1.987	3.056	5.043	4.737	0.917	1.528	2.751	0.412	0.184	0.569	0.531	0.209	0.227	0.267	0.163
Node70	0.01482	132	31	98	140	17	8	6	16	11.149	2.682	8.299	11.820	1.509	0.754	0.587	1.425	0.203	0.077	0.199	0.191	0.588	0.297	0.573	0.377
Node71	0.01493	100	19	58	72	13	1	9	12	8.405	1.664	4.910	6.075	1.165	0.166	0.832	1.082	0.349	0.269	0.385	0.456	0.655	0.757	0.468	0.469
Node72	0.05433	185	36	96	109	12	3	5	10	4.252	0.846	2.218	2.515	0.297	0.091	0.137	0.251	0.689	0.619	0.681	0.758	0.862	0.847	0.840	0.802
Node73	0.01699	115	20	67	109	34	7	10	12	8.483	1.536	4.973	8.044	2.559	0.585	0.804	0.951	0.344	0.311	0.380	0.339	0.421	0.392	0.479	0.511
Node74	0.03014	182	27	140	152	36	5	15	24	7.543	1.154	5.812	6.306	1.525	0.247	0.659	1.030	0.408	0.465	0.321	0.440	0.585	0.670	0.539	0.485
Node75	0.01	72	12	36	53	31	10	14	16	9.069	1.615	4.597	6.709	3.976	1.367	1.864	2.112	0.308	0.284	0.411	0.415	0.267	0.107	0.205	0.244
Node76	0.01214	38	8	22	25	9	6	3	4	3.989	0.921	2.353	2.659	1.023	0.716	0.409	0.511	0.713	0.580	0.662	0.743	0.685	0.317	0.664	0.677
Node77	0.08536	246	39	165	261	188	31	93	113	3.594	0.582	2.416	3.813	2.750	0.466	1.368	1.659	0.750	0.763	0.653	0.634	0.396	0.475	0.303	0.325
Node78	0.08946	319	48	245	254	87	17	28	34	4.443	0.680	3.416	3.541	1.222	0.250	0.403	0.486	0.671	0.709	0.527	0.659	0.643	0.667	0.668	0.688
Node79	0.26466	185	36	334	867	28	6	2	3	0.873	0.174	1.572	4.074	0.136	0.033	0.014	0.019	0.964	0.956	0.781	0.611	0.911	0.924	0.953	0.952
Node80	0.12727	35	7	21	37	0	0	0	0	0.351	0.078	0.215	0.371	0.010	0.010	0.010	0.010	0.988	0.984	0.989	0.971	0.966	0.961	0.960	0.962
Node81	0.08041	117	24	120	166	14	7	6	3	1.823	0.386	1.869	2.580	0.232	0.124	0.108	0.062	0.903	0.865	0.734	0.751	0.881	0.807	0.863	0.917
Node82	0.09485	245	42	282	342	260	37	99	88	3.222	0.563	3.706	4.492	3.418	0.498	1.310	1.166	0.784	0.773	0.495	0.575	0.320	0.452	0.318	0.445
Node83	0.01238	21	4	13	13	83	21	44	54	2.207	0.502	1.404	1.404	8.427	2.207	4.514	5.518	0.873	0.806	0.809	0.871	0.063	0.026	0.028	0.030
Node84	0.01549	18	5	8	24	29	8	11	16	1.524	0.481	0.722	2.005	2.406	0.722	0.962	1.363	0.924	0.817	0.921	0.810	0.442	0.314	0.421	0.392
Node85	0.00857	49	10	22	44	10	2	2	7	7.249	1.595	3.335	6.524	1.595	0.435	0.435	1.160	0.430	0.291	0.536	0.426	0.572	0.499	0.650	0.447
Node86	0.03514	50	8	35	56	56	14	33	55	1.803	0.318	1.273	2.015	2.015	0.530	1.202	1.980	0.904	0.898	0.831	0.809	0.501	0.429	0.347	0.265
Node87	0.04433	316	75	299	463	78	8	25	34	8.882	2.129	8.406	13.001	2.214	0.252	0.729	0.981	0.319	0.155	0.195	0.160	0.470	0.665	0.509	0.501
Node88	0.04035	201	25	152	261	47	9	21	32	6.219	0.801	4.711	8.067	1.478	0.308	0.677	1.016	0.512	0.644	0.401	0.338	0.593	0.610	0.531	0.490
Node89	0.02399	422	46	401	228	76	16	25	40	21.904	2.434	20.817	11.858	3.987	0.880	1.346	2.123	0.017	0.106	0.027	0.190	0.266	0.242	0.309	0.242
Node90	0.01453	152	23	156	181	111	23	54	75	13.081	2.052	13.423	15.561	9.576	2.052	4.702	6.498	0.135	0.170	0.082	0.109	0.044	0.034	0.024	0.017
Node91	0.0202	231	40	189	236	130	13	49	58	14.268	2.522	11.685	14.576	8.057	0.861	3.075	3.629	0.104	0.095	0.109	0.126	0.071	0.250	0.081	0.095

	κ	T_b				T_d				f_b				f_d				$P(f_b)$				$P(f_d)$			
		NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC	SC	SN	NR	NC λ = 2.0	SC λ = 1.9	SN λ = 0.12	NR λ = 0.08	NC λ = 0.34	SC λ = 0.4	SN λ = 0.32	NR λ = 0.35
Node92	0.02361	94	19	55	129	18	4	9	13	4.999	1.052	2.947	6.841	1.000	0.263	0.526	0.737	0.620	0.513	0.583	0.406	0.689	0.654	0.602	0.586
Node93	0.08282	555	122	332	505	23	4	7	9	8.339	1.845	4.994	7.589	0.360	0.075	0.120	0.150	0.353	0.218	0.378	0.363	0.844	0.868	0.854	0.860
Node94	0.12911	255	53	282	488	0	0	0	0	2.463	0.520	2.723	4.705	0.010	0.010	0.010	0.010	0.851	0.797	0.611	0.558	0.966	0.961	0.960	0.963
Node95	0.02211	212	35	189	301	51	11	13	20	11.967	2.023	10.675	16.968	2.922	0.674	0.787	1.180	0.171	0.177	0.130	0.088	0.375	0.339	0.486	0.441
Node96	0.03743	286	35	688	505	197	17	56	21	9.523	1.195	22.863	16.790	6.570	0.597	1.891	0.730	0.282	0.446	0.021	0.091	0.115	0.385	0.201	0.588
Node97	0.12781	335	58	300	230	205	43	44	44	3.265	0.573	2.925	2.245	2.002	0.428	0.437	0.437	0.780	0.768	0.585	0.785	0.503	0.505	0.649	0.710
Node98	0.01674	233	44	156	301	41	3	4	6	17.365	3.339	11.651	22.412	3.117	0.297	0.371	0.519	0.051	0.032	0.110	0.040	0.352	0.621	0.686	0.674
Node99	0.01212	72	9	78	121	42	6	16	15	7.484	1.025	8.099	12.507	4.408	0.718	1.743	1.640	0.413	0.526	0.206	0.172	0.232	0.316	0.225	0.329

Table 3.3.16. Functional over- and underrepresentation in Viridiplantae remodelling categories

For each RC, representations (Rep; overrepresentation (enrichment; e) and underrepresentation (purification; p)) are presented for each GO. The GO depth and ratios used to calculate P (and P_B) are provided. Only significant ($P_B \leq 0.05$) representations are displayed.

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	e	GO:0040007	growth	498/480185	522/618016	1	4.30E-06	1.04E-03
NC	BP	e	GO:0043062	extracellular structure organization	498/480185	522/618016	3	4.30E-06	1.04E-03
NC	BP	e	GO:0030198	extracellular matrix organization	498/480185	522/618016	4	4.30E-06	1.04E-03
NC	BP	e	GO:0071554	cell wall organization or biogenesis	2810/480185	2873/618016	2	9.51E-06	2.30E-03
NC	BP	e	GO:0022610	biological adhesion	48/480185	48/618016	1	1.18E-05	2.86E-03
NC	BP	e	GO:0007155	cell adhesion	48/480185	48/618016	2	1.18E-05	2.86E-03
NC	BP	e	GO:0065008	regulation of biological quality	5139/480185	5989/618016	2	1.34E-05	3.25E-03
NC	BP	e	GO:0042592	homeostatic process	5139/480185	5989/618016	3	1.34E-05	3.25E-03
NC	BP	e	GO:0050789	regulation of biological process	10200/480185	11532/618016	2	1.97E-05	4.76E-03
NC	BP	e	GO:0050794	regulation of cellular process	10200/480185	11532/618016	3	1.97E-05	4.76E-03
NC	BP	e	GO:0007165	signal transduction	10200/480185	11532/618016	4	1.97E-05	4.76E-03
NC	BP	e	GO:0065007	biological regulation	15252/480185	17434/618016	1	2.37E-05	5.73E-03
NC	BP	e	GO:0005975	carbohydrate metabolic process	19471/480185	24039/618016	3	2.81E-05	6.80E-03
NC	BP	e	GO:0055085	transmembrane transport	21434/480185	26703/618016	4	3.04E-05	7.35E-03
NC	BP	e	GO:0043412	macromolecule modification	60159/480185	64734/618016	4	4.67E-05	0.0113

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	e	GO:0036211	protein modification process	60159/480185	64734/618016	5	4.67E-05	0.0113
NC	BP	e	GO:0006464	cellular protein modification process	60159/480185	64734/618016	6	4.67E-05	0.0113
NC	BP	e	GO:0044267	cellular protein metabolic process	64238/480185	75999/618016	5	4.97E-05	0.012
NC	BP	e	GO:0019538	protein metabolic process	64331/480185	76497/618016	4	5.01E-05	0.0121
NC	BP	e	GO:1901564	organonitrogen compound metabolic process	67765/480185	82805/618016	3	5.21E-05	0.0126
NC	BP	e	GO:0044260	cellular macromolecule metabolic process	67394/480185	81587/618016	4	5.25E-05	0.0127
NC	BP	e	GO:0043170	macromolecule metabolic process	71080/480185	87795/618016	3	5.35E-05	0.013
NC	BP	e	GO:0044238	primary metabolic process	101378/480185	129085/618016	2	6.35E-05	0.0154
NC	BP	e	GO:0071704	organic substance metabolic process	101378/480185	129085/618016	2	6.35E-05	0.0154
NC	BP	p	GO:0032196	transposition	4/480185	29/618016	2	7.76E-07	1.88E-04
NC	BP	p	GO:0019748	secondary metabolic process	39/480185	84/618016	2	1.09E-06	2.63E-04
NC	BP	p	GO:0040011	locomotion	0/480185	9/618016	1	1.36E-06	3.30E-04
NC	BP	p	GO:0006928	movement of cell or subcellular component	0/480185	9/618016	2	1.36E-06	3.30E-04
NC	BP	p	GO:0048870	cell motility	0/480185	9/618016	3	1.36E-06	3.30E-04
NC	BP	p	GO:0000278	mitotic cell cycle	12/480185	144/618016	3	1.63E-06	3.95E-04
NC	BP	p	GO:0051301	cell division	0/480185	151/618016	2	1.79E-06	4.32E-04
NC	BP	p	GO:0007059	chromosome segregation	0/480185	158/618016	2	1.92E-06	4.65E-04
NC	BP	p	GO:0007005	mitochondrion organization	37/480185	123/618016	4	2.04E-06	4.94E-04
NC	BP	p	GO:0032501	multicellular organismal process	0/480185	159/618016	1	2.16E-06	5.24E-04
NC	BP	p	GO:0007275	multicellular organism development	0/480185	159/618016	3	2.16E-06	5.24E-04
NC	BP	p	GO:0009790	embryo development	0/480185	159/618016	4	2.16E-06	5.24E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	p	GO:0061024	membrane organization	78/480185	209/618016	3	2.47E-06	5.99E-04
NC	BP	p	GO:0061919	process utilizing autophagic mechanism	146/480185	289/618016	2	2.58E-06	6.24E-04
NC	BP	p	GO:0006914	autophagy	146/480185	289/618016	4	2.58E-06	6.24E-04
NC	BP	p	GO:0008283	cell population proliferation	0/480185	257/618016	1	2.83E-06	6.86E-04
NC	BP	p	GO:0071826	ribonucleoprotein complex subunit organization	154/480185	251/618016	4	3.04E-06	7.36E-04
NC	BP	p	GO:0034622	cellular protein-containing complex assembly	154/480185	251/618016	5	3.04E-06	7.36E-04
NC	BP	p	GO:0022618	ribonucleoprotein complex assembly	154/480185	251/618016	6	3.04E-06	7.36E-04
NC	BP	p	GO:0071941	nitrogen cycle metabolic process	97/480185	245/618016	3	3.30E-06	7.99E-04
NC	BP	p	GO:0007049	cell cycle	73/480185	468/618016	2	3.59E-06	8.69E-04
NC	BP	p	GO:0044085	cellular component biogenesis	301/480185	654/618016	2	3.80E-06	9.19E-04
NC	BP	p	GO:0022613	ribonucleoprotein complex biogenesis	301/480185	654/618016	3	3.80E-06	9.19E-04
NC	BP	p	GO:0042254	ribosome biogenesis	301/480185	654/618016	4	3.80E-06	9.19E-04
NC	BP	p	GO:0007034	vacuolar transport	348/480185	661/618016	4	4.35E-06	1.05E-03
NC	BP	p	GO:0006886	intracellular protein transport	258/480185	663/618016	8	4.43E-06	1.07E-03
NC	BP	p	GO:0006605	protein targeting	258/480185	663/618016	9	4.43E-06	1.07E-03
NC	BP	p	GO:0051169	nuclear transport	247/480185	370/618016	5	4.48E-06	1.08E-03
NC	BP	p	GO:0006913	nucleocytoplasmic transport	247/480185	370/618016	6	4.48E-06	1.08E-03
NC	BP	p	GO:0007010	cytoskeleton organization	383/480185	869/618016	4	4.67E-06	1.13E-03
NC	BP	p	GO:0051604	protein maturation	93/480185	564/618016	5	4.81E-06	1.16E-03
NC	BP	p	GO:0051641	cellular localization	505/480185	1035/618016	2	4.90E-06	1.19E-03
NC	BP	p	GO:0051649	establishment of localization in cell	505/480185	1035/618016	3	4.90E-06	1.19E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	p	GO:0046907	intracellular transport	505/480185	1035/618016	4	4.90E-06	1.19E-03
NC	BP	p	GO:1901575	organic substance catabolic process	700/480185	1074/618016	3	6.14E-06	1.49E-03
NC	BP	p	GO:0019439	aromatic compound catabolic process	700/480185	1074/618016	4	6.14E-06	1.49E-03
NC	BP	p	GO:0044270	cellular nitrogen compound catabolic process	700/480185	1074/618016	4	6.14E-06	1.49E-03
NC	BP	p	GO:0046700	heterocycle catabolic process	700/480185	1074/618016	4	6.14E-06	1.49E-03
NC	BP	p	GO:1901361	organic cyclic compound catabolic process	700/480185	1074/618016	4	6.14E-06	1.49E-03
NC	BP	p	GO:0034655	nucleobase-containing compound catabolic process	700/480185	1074/618016	5	6.14E-06	1.49E-03
NC	BP	p	GO:0044248	cellular catabolic process	846/480185	1363/618016	3	6.73E-06	1.63E-03
NC	BP	p	GO:0006396	RNA processing	1097/480185	1614/618016	7	6.90E-06	1.67E-03
NC	BP	p	GO:0016071	mRNA metabolic process	1097/480185	1614/618016	7	6.90E-06	1.67E-03
NC	BP	p	GO:0006397	mRNA processing	1097/480185	1614/618016	8	6.90E-06	1.67E-03
NC	BP	p	GO:0051276	chromosome organization	1288/480185	2021/618016	4	7.69E-06	1.86E-03
NC	BP	p	GO:0006790	sulfur compound metabolic process	899/480185	1865/618016	3	8.11E-06	1.96E-03
NC	BP	p	GO:0043933	protein-containing complex subunit organization	1110/480185	2267/618016	3	8.52E-06	2.06E-03
NC	BP	p	GO:0065003	protein-containing complex assembly	1110/480185	2267/618016	4	8.52E-06	2.06E-03
NC	BP	p	GO:0006996	organelle organization	1708/480185	3013/618016	3	9.24E-06	2.24E-03
NC	BP	p	GO:0006091	generation of precursor metabolites and energy	1849/480185	2901/618016	3	9.46E-06	2.29E-03
NC	BP	p	GO:0015979	photosynthesis	281/480185	2478/618016	3	9.50E-06	2.30E-03
NC	BP	p	GO:0051186	cofactor metabolic process	1219/480185	3066/618016	3	1.00E-05	2.42E-03
NC	BP	p	GO:0022607	cellular component assembly	1404/480185	3012/618016	3	1.02E-05	2.47E-03
NC	BP	p	GO:0034660	ncRNA metabolic process	2618/480185	4249/618016	7	1.16E-05	2.81E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	p	GO:0006399	tRNA metabolic process	2618/480185	4249/618016	8	1.16E-05	2.81E-03
NC	BP	p	GO:0006457	protein folding	2490/480185	3697/618016	2	1.18E-05	2.85E-03
NC	BP	p	GO:0016192	vesicle-mediated transport	3267/480185	5271/618016	4	1.23E-05	2.99E-03
NC	BP	p	GO:0033036	macromolecule localization	3650/480185	6129/618016	2	1.36E-05	3.30E-03
NC	BP	p	GO:0008104	protein localization	3650/480185	6129/618016	3	1.36E-05	3.30E-03
NC	BP	p	GO:0045184	establishment of protein localization	3650/480185	6129/618016	4	1.36E-05	3.30E-03
NC	BP	p	GO:0071702	organic substance transport	3650/480185	6129/618016	4	1.36E-05	3.30E-03
NC	BP	p	GO:0071705	nitrogen compound transport	3650/480185	6129/618016	4	1.36E-05	3.30E-03
NC	BP	p	GO:0042886	amide transport	3650/480185	6129/618016	5	1.36E-05	3.30E-03
NC	BP	p	GO:0015833	peptide transport	3650/480185	6129/618016	6	1.36E-05	3.30E-03
NC	BP	p	GO:0015031	protein transport	3650/480185	6129/618016	7	1.36E-05	3.30E-03
NC	BP	p	GO:0006259	DNA metabolic process	3156/480185	5589/618016	6	1.40E-05	3.40E-03
NC	BP	p	GO:0016043	cellular component organization	3627/480185	6391/618016	2	1.45E-05	3.51E-03
NC	BP	p	GO:0016070	RNA metabolic process	3715/480185	5863/618016	6	1.45E-05	3.51E-03
NC	BP	p	GO:0006082	organic acid metabolic process	3437/480185	6311/618016	3	1.47E-05	3.56E-03
NC	BP	p	GO:0043436	oxoacid metabolic process	3437/480185	6311/618016	4	1.47E-05	3.56E-03
NC	BP	p	GO:0019752	carboxylic acid metabolic process	3437/480185	6311/618016	5	1.47E-05	3.56E-03
NC	BP	p	GO:0006520	cellular amino acid metabolic process	3437/480185	6311/618016	6	1.47E-05	3.56E-03
NC	BP	p	GO:0071840	cellular component organization or biogenesis	3928/480185	7045/618016	1	1.57E-05	3.79E-03
NC	BP	p	GO:0009056	catabolic process	6799/480185	9013/618016	2	1.84E-05	4.46E-03
NC	BP	p	GO:0090304	nucleic acid metabolic process	6870/480185	11451/618016	5	1.94E-05	4.69E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	p	GO:0044249	cellular biosynthetic process	4083/480185	11269/618016	3	1.94E-05	4.70E-03
NC	BP	p	GO:1901576	organic substance biosynthetic process	4083/480185	11269/618016	3	1.94E-05	4.70E-03
NC	BP	p	GO:0009059	macromolecule biosynthetic process	4083/480185	11269/618016	4	1.94E-05	4.70E-03
NC	BP	p	GO:0043603	cellular amide metabolic process	4083/480185	11269/618016	4	1.94E-05	4.70E-03
NC	BP	p	GO:0044271	cellular nitrogen compound biosynthetic process	4083/480185	11269/618016	4	1.94E-05	4.70E-03
NC	BP	p	GO:1901566	organonitrogen compound biosynthetic process	4083/480185	11269/618016	4	1.94E-05	4.70E-03
NC	BP	p	GO:0006518	peptide metabolic process	4083/480185	11269/618016	5	1.94E-05	4.70E-03
NC	BP	p	GO:0034645	cellular macromolecule biosynthetic process	4083/480185	11269/618016	5	1.94E-05	4.70E-03
NC	BP	p	GO:0043604	amide biosynthetic process	4083/480185	11269/618016	5	1.94E-05	4.70E-03
NC	BP	p	GO:0043043	peptide biosynthetic process	4083/480185	11269/618016	6	1.94E-05	4.70E-03
NC	BP	p	GO:0006412	translation	4083/480185	11269/618016	7	1.94E-05	4.70E-03
NC	BP	p	GO:0006725	cellular aromatic compound metabolic process	7450/480185	12346/618016	3	2.05E-05	4.96E-03
NC	BP	p	GO:0046483	heterocycle metabolic process	7450/480185	12346/618016	3	2.05E-05	4.96E-03
NC	BP	p	GO:1901360	organic cyclic compound metabolic process	7450/480185	12346/618016	3	2.05E-05	4.96E-03
NC	BP	p	GO:0006139	nucleobase-containing compound metabolic process	7450/480185	12346/618016	4	2.05E-05	4.96E-03
NC	BP	p	GO:0006629	lipid metabolic process	8970/480185	13507/618016	3	2.20E-05	5.33E-03
NC	BP	p	GO:0044281	small molecule metabolic process	10669/480185	19428/618016	2	2.53E-05	6.12E-03
NC	BP	p	GO:0009058	biosynthetic process	21656/480185	40964/618016	2	3.77E-05	9.12E-03
NC	BP	p	GO:0034641	cellular nitrogen compound metabolic process	21714/480185	43601/618016	3	3.94E-05	9.53E-03
NC	BP	p	GO:0051179	localization	37852/480185	50086/618016	1	4.16E-05	0.0101
NC	BP	p	GO:0051234	establishment of localization	37852/480185	50086/618016	2	4.16E-05	0.0101

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	BP	p	GO:0006810	transport	37852/480185	50086/618016	3	4.16E-05	0.0101
NC	BP	p	GO:0032502	developmental process	811/480185	1118/618016	1	5.49E-05	0.0133
NC	BP	p	GO:0006807	nitrogen compound metabolic process	83482/480185	111424/618016	2	5.96E-05	0.0144
NC	BP	p	GO:0044237	cellular metabolic process	84842/480185	115940/618016	2	6.03E-05	0.0146
NC	BP	p	GO:0048856	anatomical structure development	807/480185	1112/618016	2	6.13E-05	0.0148
NC	BP	p	GO:0009987	cellular process	100289/480185	136120/618016	1	6.43E-05	0.0156
NC	BP	p	GO:0008152	metabolic process	119987/480185	162190/618016	1	6.90E-05	0.0167
NC	CC	e	GO:0042579	microbody	255/480185	281/618016	5	3.72E-06	9.01E-04
NC	CC	e	GO:0005777	peroxisome	255/480185	281/618016	6	3.72E-06	9.01E-04
NC	CC	e	GO:0030312	external encapsulating structure	4014/480185	4071/618016	2	1.16E-05	2.82E-03
NC	CC	e	GO:0005618	cell wall	4014/480185	4071/618016	3	1.16E-05	2.82E-03
NC	CC	p	GO:0042995	cell projection	0/480185	9/618016	2	1.36E-06	3.30E-04
NC	CC	p	GO:0120025	plasma membrane bounded cell projection	0/480185	9/618016	3	1.36E-06	3.30E-04
NC	CC	p	GO:0005929	cilium	0/480185	9/618016	4	1.36E-06	3.30E-04
NC	CC	p	GO:0016020	membrane	14/480185	97/618016	1	1.85E-06	4.48E-04
NC	CC	p	GO:0005886	plasma membrane	14/480185	97/618016	2	1.85E-06	4.48E-04
NC	CC	p	GO:0009579	thylakoid	22/480185	200/618016	3	2.36E-06	5.71E-04
NC	CC	p	GO:0005730	nucleolus	101/480185	169/618016	5	2.72E-06	6.58E-04
NC	CC	p	GO:0009536	plastid	72/480185	378/618016	5	3.08E-06	7.46E-04
NC	CC	p	GO:0044421	extracellular region part	118/480185	311/618016	1	3.12E-06	7.55E-04
NC	CC	p	GO:0031012	extracellular matrix	118/480185	311/618016	2	3.12E-06	7.55E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	CC	p	GO:0005856	cytoskeleton	2/480185	313/618016	5	3.25E-06	7.87E-04
NC	CC	p	GO:0005811	lipid droplet	1/480185	269/618016	5	3.26E-06	7.89E-04
NC	CC	p	GO:0005739	mitochondrion	9/480185	496/618016	5	4.11E-06	9.95E-04
NC	CC	p	GO:0005694	chromosome	357/480185	718/618016	5	4.47E-06	1.08E-03
NC	CC	p	GO:0005783	endoplasmic reticulum	300/480185	854/618016	5	4.70E-06	1.14E-03
NC	CC	p	GO:0005737	cytoplasm	1997/480185	4633/618016	3	1.20E-05	2.91E-03
NC	CC	p	GO:0005622	intracellular	4344/480185	10036/618016	2	1.87E-05	4.52E-03
NC	CC	p	GO:1990904	ribonucleoprotein complex	3959/480185	11026/618016	2	1.94E-05	4.70E-03
NC	CC	p	GO:0005840	ribosome	3959/480185	11026/618016	5	1.94E-05	4.70E-03
NC	CC	p	GO:0043228	non-membrane-bounded organelle	4420/480185	12495/618016	2	1.98E-05	4.79E-03
NC	CC	p	GO:0043232	intracellular non-membrane-bounded organelle	4420/480185	12495/618016	4	1.98E-05	4.79E-03
NC	CC	p	GO:0044444	cytoplasmic part	4594/480185	13030/618016	3	2.09E-05	5.07E-03
NC	CC	p	GO:0043227	membrane-bounded organelle	12508/480185	17301/618016	2	2.40E-05	5.81E-03
NC	CC	p	GO:0043231	intracellular membrane-bounded organelle	12508/480185	17301/618016	4	2.40E-05	5.81E-03
NC	CC	p	GO:0043226	organelle	16976/480185	29938/618016	1	3.14E-05	7.59E-03
NC	CC	p	GO:0043229	intracellular organelle	16925/480185	29789/618016	3	3.20E-05	7.74E-03
NC	CC	p	GO:0044428	nuclear part	243/480185	357/618016	4	3.24E-05	7.84E-03
NC	CC	p	GO:0032991	protein-containing complex	13150/480185	31947/618016	1	3.36E-05	8.13E-03
NC	CC	p	GO:0044424	intracellular part	18927/480185	34591/618016	2	3.46E-05	8.36E-03
NC	CC	p	GO:0044464	cell part	25271/480185	42317/618016	1	3.79E-05	9.17E-03
NC	MF	e	GO:0003729	mRNA binding	218/480185	238/618016	5	2.43E-06	5.88E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	MF	e	GO:0042393	histone binding	818/480185	877/618016	3	5.19E-06	1.26E-03
NC	MF	e	GO:0019899	enzyme binding	1905/480185	2246/618016	3	7.71E-06	1.86E-03
NC	MF	e	GO:0004386	helicase activity	2343/480185	2693/618016	7	9.25E-06	2.24E-03
NC	MF	e	GO:0003924	GTPase activity	2292/480185	2773/618016	7	1.01E-05	2.44E-03
NC	MF	e	GO:0016887	ATPase activity	6668/480185	7330/618016	7	1.56E-05	3.78E-03
NC	MF	e	GO:0016817	hydrolase activity, acting on acid anhydrides	11023/480185	12352/618016	3	2.06E-05	4.99E-03
NC	MF	e	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	11023/480185	12352/618016	4	2.06E-05	4.99E-03
NC	MF	e	GO:0016462	pyrophosphatase activity	11023/480185	12352/618016	5	2.06E-05	4.99E-03
NC	MF	e	GO:0017111	nucleoside-triphosphatase activity	11023/480185	12352/618016	6	2.06E-05	4.99E-03
NC	MF	e	GO:0016798	hydrolase activity, acting on glycosyl bonds	14624/480185	17114/618016	3	2.46E-05	5.95E-03
NC	MF	e	GO:0140110	transcription regulator activity	15248/480185	18464/618016	1	2.56E-05	6.19E-03
NC	MF	e	GO:0003700	DNA-binding transcription factor activity	15248/480185	18464/618016	2	2.56E-05	6.19E-03
NC	MF	e	GO:0016757	transferase activity, transferring glycosyl groups	16680/480185	19461/618016	3	2.67E-05	6.46E-03
NC	MF	e	GO:0005215	transporter activity	18422/480185	22509/618016	1	2.76E-05	6.68E-03
NC	MF	e	GO:0022857	transmembrane transporter activity	18422/480185	22509/618016	2	2.76E-05	6.68E-03
NC	MF	e	GO:0016787	hydrolase activity	43156/480185	53542/618016	2	4.31E-05	0.0104
NC	MF	e	GO:0016491	oxidoreductase activity	50514/480185	59827/618016	2	4.46E-05	0.0108
NC	MF	e	GO:0016301	kinase activity	55373/480185	58982/618016	4	4.46E-05	0.0108
NC	MF	e	GO:0016772	transferase activity, transferring phosphorus-containing groups	58527/480185	63814/618016	3	4.66E-05	0.0113
NC	MF	e	GO:0016740	transferase activity	90490/480185	105241/618016	2	5.82E-05	0.0141
NC	MF	e	GO:0043167	ion binding	144810/480185	169425/618016	2	6.99E-05	0.0169

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	MF	e	GO:0005488	binding	183234/480185	222950/618016	1	7.46E-05	0.018
NC	MF	e	GO:0003824	catalytic activity	194092/480185	234033/618016	1	7.61E-05	0.0184
NC	MF	p	GO:0032182	ubiquitin-like protein binding	27/480185	110/618016	3	1.69E-06	4.09E-04
NC	MF	p	GO:0019843	rRNA binding	316/480185	858/618016	5	4.94E-06	1.20E-03
NC	MF	p	GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	475/480185	1221/618016	3	6.31E-06	1.53E-03
NC	MF	p	GO:0051082	unfolded protein binding	840/480185	1302/618016	3	6.59E-06	1.60E-03
NC	MF	p	GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	807/480185	1578/618016	3	7.22E-06	1.75E-03
NC	MF	p	GO:0042578	phosphoric ester hydrolase activity	1725/480185	2393/618016	4	8.32E-06	2.01E-03
NC	MF	p	GO:0016791	phosphatase activity	1725/480185	2393/618016	5	8.32E-06	2.01E-03
NC	MF	p	GO:0045182	translation regulator activity	972/480185	2397/618016	1	8.46E-06	2.05E-03
NC	MF	p	GO:0090079	translation regulator activity, nucleic acid binding	972/480185	2397/618016	4	8.46E-06	2.05E-03
NC	MF	p	GO:0008135	translation factor activity, RNA binding	972/480185	2397/618016	5	8.46E-06	2.05E-03
NC	MF	p	GO:0004518	nuclease activity	2332/480185	3725/618016	4	1.06E-05	2.57E-03
NC	MF	p	GO:0008092	cytoskeletal protein binding	3334/480185	4676/618016	3	1.22E-05	2.94E-03
NC	MF	p	GO:0016779	nucleotidyltransferase activity	3155/480185	4833/618016	4	1.23E-05	2.97E-03
NC	MF	p	GO:0016874	ligase activity	2850/480185	4776/618016	2	1.28E-05	3.09E-03
NC	MF	p	GO:0016853	isomerase activity	3751/480185	5524/618016	2	1.35E-05	3.26E-03
NC	MF	p	GO:0016788	hydrolase activity, acting on ester bonds	4057/480185	6118/618016	3	1.44E-05	3.48E-03
NC	MF	p	GO:0098772	molecular function regulator	4421/480185	7248/618016	1	1.49E-05	3.61E-03
NC	MF	p	GO:0030234	enzyme regulator activity	4421/480185	7248/618016	2	1.49E-05	3.61E-03
NC	MF	p	GO:0016829	lyase activity	5377/480185	7558/618016	2	1.53E-05	3.70E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NC	MF	p	GO:0005515	protein binding	7116/480185	9464/618016	2	1.76E-05	4.26E-03
NC	MF	p	GO:0016741	transferase activity, transferring one-carbon groups	6593/480185	9655/618016	3	1.80E-05	4.35E-03
NC	MF	p	GO:0008168	methyltransferase activity	6593/480185	9655/618016	4	1.80E-05	4.35E-03
NC	MF	p	GO:0016746	transferase activity, transferring acyl groups	7923/480185	10773/618016	3	1.82E-05	4.41E-03
NC	MF	p	GO:0003735	structural constituent of ribosome	4083/480185	11389/618016	2	1.94E-05	4.68E-03
NC	MF	p	GO:0005198	structural molecule activity	4407/480185	12200/618016	1	2.02E-05	4.90E-03
NC	MF	p	GO:0003723	RNA binding	6887/480185	12346/618016	4	2.05E-05	4.96E-03
NC	MF	p	GO:0003677	DNA binding	30109/480185	39471/618016	4	3.62E-05	8.76E-03
NC	MF	p	GO:0097159	organic cyclic compound binding	36995/480185	51816/618016	2	4.20E-05	0.0102
NC	MF	p	GO:1901363	heterocyclic compound binding	36995/480185	51816/618016	2	4.20E-05	0.0102
NC	MF	p	GO:0003676	nucleic acid binding	36995/480185	51816/618016	3	4.20E-05	0.0102
NR	BP	e	GO:0040011	locomotion	9/40223	9/618016	1	1.58E-07	3.83E-05
NR	BP	e	GO:0006928	movement of cell or subcellular component	9/40223	9/618016	2	1.58E-07	3.83E-05
NR	BP	e	GO:0048870	cell motility	9/40223	9/618016	3	1.58E-07	3.83E-05
NR	BP	e	GO:0032501	multicellular organismal process	159/40223	159/618016	1	4.51E-07	1.09E-04
NR	BP	e	GO:0007275	multicellular organism development	159/40223	159/618016	3	4.51E-07	1.09E-04
NR	BP	e	GO:0009790	embryo development	159/40223	159/618016	4	4.51E-07	1.09E-04
NR	BP	e	GO:0051301	cell division	151/40223	151/618016	2	4.98E-07	1.20E-04
NR	BP	e	GO:0000278	mitotic cell cycle	40/40223	144/618016	3	6.27E-07	1.52E-04
NR	BP	e	GO:0071941	nitrogen cycle metabolic process	89/40223	245/618016	3	7.15E-07	1.73E-04
NR	BP	e	GO:0032196	transposition	21/40223	29/618016	2	1.02E-06	2.46E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	BP	e	GO:0007059	chromosome segregation	93/40223	158/618016	2	1.28E-06	3.10E-04
NR	BP	e	GO:0051169	nuclear transport	103/40223	370/618016	5	1.28E-06	3.10E-04
NR	BP	e	GO:0006913	nucleocytoplasmic transport	103/40223	370/618016	6	1.28E-06	3.10E-04
NR	BP	e	GO:0008283	cell population proliferation	196/40223	257/618016	1	1.40E-06	3.38E-04
NR	BP	e	GO:0061919	process utilizing autophagic mechanism	107/40223	289/618016	2	1.82E-06	4.40E-04
NR	BP	e	GO:0006914	autophagy	107/40223	289/618016	4	1.82E-06	4.40E-04
NR	BP	e	GO:0007049	cell cycle	119/40223	468/618016	2	1.94E-06	4.70E-04
NR	BP	e	GO:0044085	cellular component biogenesis	325/40223	654/618016	2	2.37E-06	5.73E-04
NR	BP	e	GO:0022613	ribonucleoprotein complex biogenesis	325/40223	654/618016	3	2.37E-06	5.73E-04
NR	BP	e	GO:0042254	ribosome biogenesis	325/40223	654/618016	4	2.37E-06	5.73E-04
NR	BP	e	GO:0007034	vacuolar transport	86/40223	661/618016	4	2.58E-06	6.24E-04
NR	BP	e	GO:0006886	intracellular protein transport	131/40223	663/618016	8	2.65E-06	6.42E-04
NR	BP	e	GO:0006605	protein targeting	131/40223	663/618016	9	2.65E-06	6.42E-04
NR	BP	e	GO:0051604	protein maturation	226/40223	564/618016	5	3.03E-06	7.34E-04
NR	BP	e	GO:0048856	anatomical structure development	166/40223	1112/618016	2	3.39E-06	8.21E-04
NR	BP	e	GO:0044248	cellular catabolic process	266/40223	1363/618016	3	3.50E-06	8.46E-04
NR	BP	e	GO:0032502	developmental process	168/40223	1118/618016	1	3.50E-06	8.47E-04
NR	BP	e	GO:0051641	cellular localization	236/40223	1035/618016	2	3.61E-06	8.73E-04
NR	BP	e	GO:0051649	establishment of localization in cell	236/40223	1035/618016	3	3.61E-06	8.73E-04
NR	BP	e	GO:0046907	intracellular transport	236/40223	1035/618016	4	3.61E-06	8.73E-04
NR	BP	e	GO:0006396	RNA processing	160/40223	1614/618016	7	3.95E-06	9.57E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	BP	e	GO:0016071	mRNA metabolic process	160/40223	1614/618016	7	3.95E-06	9.57E-04
NR	BP	e	GO:0006397	mRNA processing	160/40223	1614/618016	8	3.95E-06	9.57E-04
NR	BP	e	GO:1901575	organic substance catabolic process	159/40223	1074/618016	3	4.04E-06	9.79E-04
NR	BP	e	GO:0019439	aromatic compound catabolic process	159/40223	1074/618016	4	4.04E-06	9.79E-04
NR	BP	e	GO:0044270	cellular nitrogen compound catabolic process	159/40223	1074/618016	4	4.04E-06	9.79E-04
NR	BP	e	GO:0046700	heterocycle catabolic process	159/40223	1074/618016	4	4.04E-06	9.79E-04
NR	BP	e	GO:1901361	organic cyclic compound catabolic process	159/40223	1074/618016	4	4.04E-06	9.79E-04
NR	BP	e	GO:0034655	nucleobase-containing compound catabolic process	159/40223	1074/618016	5	4.04E-06	9.79E-04
NR	BP	e	GO:0006790	sulfur compound metabolic process	194/40223	1865/618016	3	4.75E-06	1.15E-03
NR	BP	e	GO:0043933	protein-containing complex subunit organization	630/40223	2267/618016	3	4.74E-06	1.15E-03
NR	BP	e	GO:0065003	protein-containing complex assembly	630/40223	2267/618016	4	4.74E-06	1.15E-03
NR	BP	e	GO:0015979	photosynthesis	1370/40223	2478/618016	3	5.63E-06	1.36E-03
NR	BP	e	GO:0034660	ncRNA metabolic process	583/40223	4249/618016	7	5.87E-06	1.42E-03
NR	BP	e	GO:0006399	tRNA metabolic process	583/40223	4249/618016	8	5.87E-06	1.42E-03
NR	BP	e	GO:0022607	cellular component assembly	788/40223	3012/618016	3	6.00E-06	1.45E-03
NR	BP	e	GO:0051186	cofactor metabolic process	734/40223	3066/618016	3	6.10E-06	1.48E-03
NR	BP	e	GO:0016192	vesicle-mediated transport	476/40223	5271/618016	4	7.07E-06	1.71E-03
NR	BP	e	GO:0006259	DNA metabolic process	648/40223	5589/618016	6	7.79E-06	1.88E-03
NR	BP	e	GO:0006457	protein folding	315/40223	3697/618016	2	8.41E-06	2.04E-03
NR	BP	e	GO:0033036	macromolecule localization	671/40223	6129/618016	2	8.45E-06	2.05E-03
NR	BP	e	GO:0008104	protein localization	671/40223	6129/618016	3	8.45E-06	2.05E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	BP	e	GO:0045184	establishment of protein localization	671/40223	6129/618016	4	8.45E-06	2.05E-03
NR	BP	e	GO:0071702	organic substance transport	671/40223	6129/618016	4	8.45E-06	2.05E-03
NR	BP	e	GO:0071705	nitrogen compound transport	671/40223	6129/618016	4	8.45E-06	2.05E-03
NR	BP	e	GO:0042886	amide transport	671/40223	6129/618016	5	8.45E-06	2.05E-03
NR	BP	e	GO:0015833	peptide transport	671/40223	6129/618016	6	8.45E-06	2.05E-03
NR	BP	e	GO:0015031	protein transport	671/40223	6129/618016	7	8.45E-06	2.05E-03
NR	BP	e	GO:0071840	cellular component organization or biogenesis	1192/40223	7045/618016	1	8.65E-06	2.09E-03
NR	BP	e	GO:0016070	RNA metabolic process	743/40223	5863/618016	6	8.73E-06	2.11E-03
NR	BP	e	GO:0006082	organic acid metabolic process	915/40223	6311/618016	3	8.95E-06	2.17E-03
NR	BP	e	GO:0043436	oxoacid metabolic process	915/40223	6311/618016	4	8.95E-06	2.17E-03
NR	BP	e	GO:0019752	carboxylic acid metabolic process	915/40223	6311/618016	5	8.95E-06	2.17E-03
NR	BP	e	GO:0006520	cellular amino acid metabolic process	915/40223	6311/618016	6	8.95E-06	2.17E-03
NR	BP	e	GO:0016043	cellular component organization	867/40223	6391/618016	2	9.05E-06	2.19E-03
NR	BP	e	GO:0090304	nucleic acid metabolic process	1391/40223	11451/618016	5	1.08E-05	2.62E-03
NR	BP	e	GO:0044249	cellular biosynthetic process	2905/40223	11269/618016	3	1.10E-05	2.67E-03
NR	BP	e	GO:1901576	organic substance biosynthetic process	2905/40223	11269/618016	3	1.10E-05	2.67E-03
NR	BP	e	GO:0009059	macromolecule biosynthetic process	2905/40223	11269/618016	4	1.10E-05	2.67E-03
NR	BP	e	GO:0043603	cellular amide metabolic process	2905/40223	11269/618016	4	1.10E-05	2.67E-03
NR	BP	e	GO:0044271	cellular nitrogen compound biosynthetic process	2905/40223	11269/618016	4	1.10E-05	2.67E-03
NR	BP	e	GO:1901566	organonitrogen compound biosynthetic process	2905/40223	11269/618016	4	1.10E-05	2.67E-03
NR	BP	e	GO:0006518	peptide metabolic process	2905/40223	11269/618016	5	1.10E-05	2.67E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	BP	e	GO:0034645	cellular macromolecule biosynthetic process	2905/40223	11269/618016	5	1.10E-05	2.67E-03
NR	BP	e	GO:0043604	amide biosynthetic process	2905/40223	11269/618016	5	1.10E-05	2.67E-03
NR	BP	e	GO:0043043	peptide biosynthetic process	2905/40223	11269/618016	6	1.10E-05	2.67E-03
NR	BP	e	GO:0006412	translation	2905/40223	11269/618016	7	1.10E-05	2.67E-03
NR	BP	e	GO:0006725	cellular aromatic compound metabolic process	1550/40223	12346/618016	3	1.15E-05	2.77E-03
NR	BP	e	GO:0046483	heterocycle metabolic process	1550/40223	12346/618016	3	1.15E-05	2.77E-03
NR	BP	e	GO:1901360	organic cyclic compound metabolic process	1550/40223	12346/618016	3	1.15E-05	2.77E-03
NR	BP	e	GO:0006139	nucleobase-containing compound metabolic process	1550/40223	12346/618016	4	1.15E-05	2.77E-03
NR	BP	e	GO:0006629	lipid metabolic process	1519/40223	13507/618016	3	1.21E-05	2.94E-03
NR	BP	e	GO:0044281	small molecule metabolic process	2860/40223	19428/618016	2	1.49E-05	3.60E-03
NR	BP	e	GO:0009058	biosynthetic process	7686/40223	40964/618016	2	2.24E-05	5.42E-03
NR	BP	e	GO:0034641	cellular nitrogen compound metabolic process	8514/40223	43601/618016	3	2.24E-05	5.42E-03
NR	BP	e	GO:0006807	nitrogen compound metabolic process	9969/40223	111424/618016	2	3.40E-05	8.22E-03
NR	BP	e	GO:0044237	cellular metabolic process	11833/40223	115940/618016	2	3.45E-05	8.36E-03
NR	BP	e	GO:0009987	cellular process	13189/40223	136120/618016	1	3.71E-05	8.97E-03
NR	BP	e	GO:0008152	metabolic process	15054/40223	162190/618016	1	3.98E-05	9.64E-03
NR	BP	p	GO:0040007	growth	0/40223	522/618016	1	1.85E-06	4.47E-04
NR	BP	p	GO:0043062	extracellular structure organization	0/40223	522/618016	3	1.85E-06	4.47E-04
NR	BP	p	GO:0030198	extracellular matrix organization	0/40223	522/618016	4	1.85E-06	4.47E-04
NR	BP	p	GO:0071826	ribonucleoprotein complex subunit organization	0/40223	251/618016	4	2.04E-06	4.93E-04
NR	BP	p	GO:0034622	cellular protein-containing complex assembly	0/40223	251/618016	5	2.04E-06	4.93E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	BP	p	GO:0022618	ribonucleoprotein complex assembly	0/40223	251/618016	6	2.04E-06	4.93E-04
NR	BP	p	GO:0007010	cytoskeleton organization	16/40223	869/618016	4	2.41E-06	5.83E-04
NR	BP	p	GO:0071554	cell wall organization or biogenesis	6/40223	2873/618016	2	5.15E-06	1.25E-03
NR	BP	p	GO:0065008	regulation of biological quality	24/40223	5989/618016	2	8.49E-06	2.06E-03
NR	BP	p	GO:0042592	homeostatic process	24/40223	5989/618016	3	8.49E-06	2.06E-03
NR	BP	p	GO:0050789	regulation of biological process	227/40223	11532/618016	2	1.16E-05	2.82E-03
NR	BP	p	GO:0050794	regulation of cellular process	227/40223	11532/618016	3	1.16E-05	2.82E-03
NR	BP	p	GO:0007165	signal transduction	227/40223	11532/618016	4	1.16E-05	2.82E-03
NR	BP	p	GO:0065007	biological regulation	251/40223	17434/618016	1	1.39E-05	3.35E-03
NR	BP	p	GO:0005975	carbohydrate metabolic process	1073/40223	24039/618016	3	1.65E-05	3.98E-03
NR	BP	p	GO:0055085	transmembrane transport	1094/40223	26703/618016	4	1.71E-05	4.14E-03
NR	BP	p	GO:0051179	localization	2952/40223	50086/618016	1	2.40E-05	5.82E-03
NR	BP	p	GO:0051234	establishment of localization	2952/40223	50086/618016	2	2.40E-05	5.82E-03
NR	BP	p	GO:0006810	transport	2952/40223	50086/618016	3	2.40E-05	5.82E-03
NR	BP	p	GO:0043412	macromolecule modification	840/40223	64734/618016	4	2.65E-05	6.41E-03
NR	BP	p	GO:0036211	protein modification process	840/40223	64734/618016	5	2.65E-05	6.41E-03
NR	BP	p	GO:0006464	cellular protein modification process	840/40223	64734/618016	6	2.65E-05	6.41E-03
NR	BP	p	GO:0044267	cellular protein metabolic process	3745/40223	75999/618016	5	2.85E-05	6.90E-03
NR	BP	p	GO:0019538	protein metabolic process	3910/40223	76497/618016	4	2.90E-05	7.02E-03
NR	BP	p	GO:0044260	cellular macromolecule metabolic process	4393/40223	81587/618016	4	2.99E-05	7.23E-03
NR	BP	p	GO:1901564	organonitrogen compound metabolic process	4825/40223	82805/618016	3	3.03E-05	7.34E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	BP	p	GO:0043170	macromolecule metabolic process	5301/40223	87795/618016	3	3.10E-05	7.51E-03
NR	BP	p	GO:0050896	response to stimulus	896/40223	15747/618016	1	3.22E-05	7.79E-03
NR	BP	p	GO:0006950	response to stress	896/40223	15747/618016	2	3.22E-05	7.79E-03
NR	CC	e	GO:0042995	cell projection	9/40223	9/618016	2	1.58E-07	3.83E-05
NR	CC	e	GO:0120025	plasma membrane bounded cell projection	9/40223	9/618016	3	1.58E-07	3.83E-05
NR	CC	e	GO:0005929	cilium	9/40223	9/618016	4	1.58E-07	3.83E-05
NR	CC	e	GO:0005730	nucleolus	54/40223	169/618016	5	5.28E-07	1.28E-04
NR	CC	e	GO:0000228	nuclear chromosome	42/40223	188/618016	6	5.98E-07	1.45E-04
NR	CC	e	GO:0009579	thylakoid	37/40223	200/618016	3	8.33E-07	2.02E-04
NR	CC	e	GO:0016020	membrane	62/40223	97/618016	1	9.49E-07	2.30E-04
NR	CC	e	GO:0005886	plasma membrane	62/40223	97/618016	2	9.49E-07	2.30E-04
NR	CC	e	GO:0005811	lipid droplet	190/40223	269/618016	5	1.28E-06	3.10E-04
NR	CC	e	GO:0044428	nuclear part	96/40223	357/618016	4	1.45E-06	3.51E-04
NR	CC	e	GO:0005856	cytoskeleton	120/40223	313/618016	5	1.62E-06	3.92E-04
NR	CC	e	GO:0009536	plastid	191/40223	378/618016	5	1.91E-06	4.61E-04
NR	CC	e	GO:0005739	mitochondrion	285/40223	496/618016	5	2.48E-06	5.99E-04
NR	CC	e	GO:0044422	organelle part	101/40223	598/618016	1	2.66E-06	6.44E-04
NR	CC	e	GO:0044446	intracellular organelle part	101/40223	598/618016	3	2.66E-06	6.44E-04
NR	CC	e	GO:0005783	endoplasmic reticulum	125/40223	854/618016	5	2.84E-06	6.87E-04
NR	CC	e	GO:0005576	extracellular region	312/40223	2060/618016	1	4.73E-06	1.14E-03
NR	CC	e	GO:0005737	cytoplasm	1024/40223	4633/618016	3	6.19E-06	1.50E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	CC	e	GO:0005622	intracellular	2729/40223	10036/618016	2	1.04E-05	2.51E-03
NR	CC	e	GO:1990904	ribonucleoprotein complex	2831/40223	11026/618016	2	1.16E-05	2.81E-03
NR	CC	e	GO:0005840	ribosome	2831/40223	11026/618016	5	1.16E-05	2.81E-03
NR	CC	e	GO:0044444	cytoplasmic part	3431/40223	13030/618016	3	1.18E-05	2.86E-03
NR	CC	e	GO:0043228	non-membrane-bounded organelle	3247/40223	12495/618016	2	1.19E-05	2.88E-03
NR	CC	e	GO:0043232	intracellular non-membrane-bounded organelle	3247/40223	12495/618016	4	1.19E-05	2.88E-03
NR	CC	e	GO:0043227	membrane-bounded organelle	1434/40223	17301/618016	2	1.38E-05	3.35E-03
NR	CC	e	GO:0043231	intracellular membrane-bounded organelle	1434/40223	17301/618016	4	1.38E-05	3.35E-03
NR	CC	e	GO:0043229	intracellular organelle	4677/40223	29789/618016	3	1.82E-05	4.40E-03
NR	CC	e	GO:0043226	organelle	4765/40223	29938/618016	1	1.90E-05	4.60E-03
NR	CC	e	GO:0032991	protein-containing complex	8647/40223	31947/618016	1	1.91E-05	4.63E-03
NR	CC	e	GO:0044424	intracellular part	5691/40223	34591/618016	2	2.02E-05	4.88E-03
NR	CC	e	GO:0044464	cell part	6349/40223	42317/618016	1	2.26E-05	5.47E-03
NR	CC	p	GO:0030312	external encapsulating structure	0/40223	4071/618016	2	5.81E-06	1.41E-03
NR	CC	p	GO:0005618	cell wall	0/40223	4071/618016	3	5.81E-06	1.41E-03
NR	CC	p	GO:0005634	nucleus	830/40223	15292/618016	5	1.36E-05	3.29E-03
NR	CC	p	GO:0042579	microbody	3/40223	281/618016	5	1.62E-05	3.93E-03
NR	CC	p	GO:0005777	peroxisome	3/40223	281/618016	6	1.62E-05	3.93E-03
NR	MF	e	GO:0032182	ubiquitin-like protein binding	55/40223	110/618016	3	5.68E-07	1.37E-04
NR	MF	e	GO:0019843	rRNA binding	203/40223	858/618016	5	3.03E-06	7.33E-04
NR	MF	e	GO:0051082	unfolded protein binding	149/40223	1302/618016	3	3.58E-06	8.66E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	MF	e	GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	457/40223	1578/618016	3	3.90E-06	9.43E-04
NR	MF	e	GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	395/40223	1221/618016	3	4.00E-06	9.68E-04
NR	MF	e	GO:0042578	phosphoric ester hydrolase activity	402/40223	2393/618016	4	4.68E-06	1.13E-03
NR	MF	e	GO:0016791	phosphatase activity	402/40223	2393/618016	5	4.68E-06	1.13E-03
NR	MF	e	GO:0045182	translation regulator activity	637/40223	2397/618016	1	4.76E-06	1.15E-03
NR	MF	e	GO:0090079	translation regulator activity, nucleic acid binding	637/40223	2397/618016	4	4.76E-06	1.15E-03
NR	MF	e	GO:0008135	translation factor activity, RNA binding	637/40223	2397/618016	5	4.76E-06	1.15E-03
NR	MF	e	GO:0004518	nuclease activity	384/40223	3725/618016	4	6.65E-06	1.61E-03
NR	MF	e	GO:0016779	nucleotidyltransferase activity	817/40223	4833/618016	4	7.32E-06	1.77E-03
NR	MF	e	GO:0016853	isomerase activity	561/40223	5524/618016	2	7.81E-06	1.89E-03
NR	MF	e	GO:0016788	hydrolase activity, acting on ester bonds	786/40223	6118/618016	3	8.32E-06	2.01E-03
NR	MF	e	GO:0016829	lyase activity	905/40223	7558/618016	2	9.12E-06	2.21E-03
NR	MF	e	GO:0098772	molecular function regulator	654/40223	7248/618016	1	9.24E-06	2.23E-03
NR	MF	e	GO:0030234	enzyme regulator activity	654/40223	7248/618016	2	9.24E-06	2.23E-03
NR	MF	e	GO:0016741	transferase activity, transferring one-carbon groups	1605/40223	9655/618016	3	1.03E-05	2.50E-03
NR	MF	e	GO:0008168	methyltransferase activity	1605/40223	9655/618016	4	1.03E-05	2.50E-03
NR	MF	e	GO:0016746	transferase activity, transferring acyl groups	990/40223	10773/618016	3	1.11E-05	2.68E-03
NR	MF	e	GO:0003723	RNA binding	1970/40223	12346/618016	4	1.15E-05	2.77E-03
NR	MF	e	GO:0005198	structural molecule activity	3124/40223	12200/618016	1	1.21E-05	2.92E-03
NR	MF	e	GO:0003735	structural constituent of ribosome	3025/40223	11389/618016	2	1.22E-05	2.96E-03
NR	MF	p	GO:0003729	mRNA binding	0/40223	238/618016	5	9.58E-07	2.32E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	MF	p	GO:0042393	histone binding	0/40223	877/618016	3	2.75E-06	6.67E-04
NR	MF	p	GO:0019899	enzyme binding	76/40223	2246/618016	3	5.27E-06	1.27E-03
NR	MF	p	GO:0004386	helicase activity	41/40223	2693/618016	7	5.36E-06	1.30E-03
NR	MF	p	GO:0008289	lipid binding	125/40223	3286/618016	2	5.45E-06	1.32E-03
NR	MF	p	GO:0008092	cytoskeletal protein binding	217/40223	4676/618016	3	6.84E-06	1.66E-03
NR	MF	p	GO:0016887	ATPase activity	81/40223	7330/618016	7	9.27E-06	2.24E-03
NR	MF	p	GO:0005515	protein binding	502/40223	9464/618016	2	1.13E-05	2.73E-03
NR	MF	p	GO:0016817	hydrolase activity, acting on acid anhydrides	209/40223	12352/618016	3	1.15E-05	2.79E-03
NR	MF	p	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	209/40223	12352/618016	4	1.15E-05	2.79E-03
NR	MF	p	GO:0016462	pyrophosphatase activity	209/40223	12352/618016	5	1.15E-05	2.79E-03
NR	MF	p	GO:0017111	nucleoside-triphosphatase activity	209/40223	12352/618016	6	1.15E-05	2.79E-03
NR	MF	p	GO:0016798	hydrolase activity, acting on glycosyl bonds	320/40223	17114/618016	3	1.34E-05	3.23E-03
NR	MF	p	GO:0140110	transcription regulator activity	72/40223	18464/618016	1	1.47E-05	3.57E-03
NR	MF	p	GO:0003700	DNA-binding transcription factor activity	72/40223	18464/618016	2	1.47E-05	3.57E-03
NR	MF	p	GO:0016757	transferase activity, transferring glycosyl groups	767/40223	19461/618016	3	1.51E-05	3.66E-03
NR	MF	p	GO:0005215	transporter activity	1196/40223	22509/618016	1	1.64E-05	3.97E-03
NR	MF	p	GO:0022857	transmembrane transporter activity	1196/40223	22509/618016	2	1.64E-05	3.97E-03
NR	MF	p	GO:0003677	DNA binding	1054/40223	39471/618016	4	2.14E-05	5.17E-03
NR	MF	p	GO:0097159	organic cyclic compound binding	3024/40223	51816/618016	2	2.40E-05	5.81E-03
NR	MF	p	GO:1901363	heterocyclic compound binding	3024/40223	51816/618016	2	2.40E-05	5.81E-03
NR	MF	p	GO:0003676	nucleic acid binding	3024/40223	51816/618016	3	2.40E-05	5.81E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
NR	MF	p	GO:0016787	hydrolase activity	2715/40223	53542/618016	2	2.48E-05	6.00E-03
NR	MF	p	GO:0016301	kinase activity	563/40223	58982/618016	4	2.52E-05	6.11E-03
NR	MF	p	GO:0016491	oxidoreductase activity	2783/40223	59827/618016	2	2.63E-05	6.36E-03
NR	MF	p	GO:0016772	transferase activity, transferring phosphorus-containing groups	1380/40223	63814/618016	3	2.73E-05	6.61E-03
NR	MF	p	GO:0003924	GTPase activity	128/40223	2773/618016	7	2.98E-05	7.22E-03
NR	MF	p	GO:0016740	transferase activity	5199/40223	105241/618016	2	3.32E-05	8.04E-03
NR	MF	p	GO:0043167	ion binding	4906/40223	169425/618016	2	3.96E-05	9.58E-03
NR	MF	p	GO:0003824	catalytic activity	12477/40223	234033/618016	1	4.35E-05	0.0105
NR	MF	p	GO:0005488	binding	8365/40223	222950/618016	1	4.39E-05	0.0106
SC	BP	e	GO:0061024	membrane organization	105/20061	209/618016	3	4.23E-07	1.02E-04
SC	BP	e	GO:0019748	secondary metabolic process	33/20061	84/618016	2	5.36E-07	1.30E-04
SC	BP	e	GO:0007005	mitochondrion organization	76/20061	123/618016	4	5.50E-07	1.33E-04
SC	BP	e	GO:0071826	ribonucleoprotein complex subunit organization	76/20061	251/618016	4	8.98E-07	2.17E-04
SC	BP	e	GO:0034622	cellular protein-containing complex assembly	76/20061	251/618016	5	8.98E-07	2.17E-04
SC	BP	e	GO:0022618	ribonucleoprotein complex assembly	76/20061	251/618016	6	8.98E-07	2.17E-04
SC	BP	e	GO:0051604	protein maturation	149/20061	564/618016	5	1.03E-06	2.50E-04
SC	BP	e	GO:0007059	chromosome segregation	19/20061	158/618016	2	1.08E-06	2.62E-04
SC	BP	e	GO:0007049	cell cycle	135/20061	468/618016	2	1.12E-06	2.72E-04
SC	BP	e	GO:0000278	mitotic cell cycle	80/20061	144/618016	3	1.17E-06	2.82E-04
SC	BP	e	GO:0008283	cell population proliferation	53/20061	257/618016	1	1.38E-06	3.35E-04
SC	BP	e	GO:1901575	organic substance catabolic process	90/20061	1074/618016	3	1.57E-06	3.80E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	BP	e	GO:0019439	aromatic compound catabolic process	90/20061	1074/618016	4	1.57E-06	3.80E-04
SC	BP	e	GO:0044270	cellular nitrogen compound catabolic process	90/20061	1074/618016	4	1.57E-06	3.80E-04
SC	BP	e	GO:0046700	heterocycle catabolic process	90/20061	1074/618016	4	1.57E-06	3.80E-04
SC	BP	e	GO:1901361	organic cyclic compound catabolic process	90/20061	1074/618016	4	1.57E-06	3.80E-04
SC	BP	e	GO:0034655	nucleobase-containing compound catabolic process	90/20061	1074/618016	5	1.57E-06	3.80E-04
SC	BP	e	GO:0007010	cytoskeleton organization	127/20061	869/618016	4	1.95E-06	4.71E-04
SC	BP	e	GO:0007034	vacuolar transport	109/20061	661/618016	4	2.43E-06	5.88E-04
SC	BP	e	GO:0044248	cellular catabolic process	96/20061	1363/618016	3	2.49E-06	6.02E-04
SC	BP	e	GO:0051276	chromosome organization	218/20061	2021/618016	4	2.84E-06	6.88E-04
SC	BP	e	GO:0006396	RNA processing	215/20061	1614/618016	7	2.90E-06	7.01E-04
SC	BP	e	GO:0016071	mRNA metabolic process	215/20061	1614/618016	7	2.90E-06	7.01E-04
SC	BP	e	GO:0006397	mRNA processing	215/20061	1614/618016	8	2.90E-06	7.01E-04
SC	BP	e	GO:0043933	protein-containing complex subunit organization	203/20061	2267/618016	3	3.21E-06	7.77E-04
SC	BP	e	GO:0065003	protein-containing complex assembly	203/20061	2267/618016	4	3.21E-06	7.77E-04
SC	BP	e	GO:0015979	photosynthesis	143/20061	2478/618016	3	3.82E-06	9.24E-04
SC	BP	e	GO:0051186	cofactor metabolic process	528/20061	3066/618016	3	3.92E-06	9.48E-04
SC	BP	e	GO:0006790	sulfur compound metabolic process	423/20061	1865/618016	3	4.01E-06	9.69E-04
SC	BP	e	GO:0006996	organelle organization	421/20061	3013/618016	3	4.20E-06	1.02E-03
SC	BP	e	GO:0022607	cellular component assembly	383/20061	3012/618016	3	4.19E-06	1.02E-03
SC	BP	e	GO:0034660	ncRNA metabolic process	382/20061	4249/618016	7	4.83E-06	1.17E-03
SC	BP	e	GO:0006399	tRNA metabolic process	382/20061	4249/618016	8	4.83E-06	1.17E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	BP	e	GO:0016070	RNA metabolic process	597/20061	5863/618016	6	5.04E-06	1.22E-03
SC	BP	e	GO:0071840	cellular component organization or biogenesis	713/20061	7045/618016	1	5.68E-06	1.38E-03
SC	BP	e	GO:0016192	vesicle-mediated transport	389/20061	5271/618016	4	5.72E-06	1.38E-03
SC	BP	e	GO:0033036	macromolecule localization	627/20061	6129/618016	2	5.77E-06	1.40E-03
SC	BP	e	GO:0008104	protein localization	627/20061	6129/618016	3	5.77E-06	1.40E-03
SC	BP	e	GO:0045184	establishment of protein localization	627/20061	6129/618016	4	5.77E-06	1.40E-03
SC	BP	e	GO:0071702	organic substance transport	627/20061	6129/618016	4	5.77E-06	1.40E-03
SC	BP	e	GO:0071705	nitrogen compound transport	627/20061	6129/618016	4	5.77E-06	1.40E-03
SC	BP	e	GO:0042886	amide transport	627/20061	6129/618016	5	5.77E-06	1.40E-03
SC	BP	e	GO:0015833	peptide transport	627/20061	6129/618016	6	5.77E-06	1.40E-03
SC	BP	e	GO:0015031	protein transport	627/20061	6129/618016	7	5.77E-06	1.40E-03
SC	BP	e	GO:0006259	DNA metabolic process	494/20061	5589/618016	6	6.15E-06	1.49E-03
SC	BP	e	GO:0016043	cellular component organization	713/20061	6391/618016	2	6.37E-06	1.54E-03
SC	BP	e	GO:0006082	organic acid metabolic process	681/20061	6311/618016	3	6.59E-06	1.60E-03
SC	BP	e	GO:0043436	oxoacid metabolic process	681/20061	6311/618016	4	6.59E-06	1.60E-03
SC	BP	e	GO:0019752	carboxylic acid metabolic process	681/20061	6311/618016	5	6.59E-06	1.60E-03
SC	BP	e	GO:0006520	cellular amino acid metabolic process	681/20061	6311/618016	6	6.59E-06	1.60E-03
SC	BP	e	GO:0090304	nucleic acid metabolic process	1091/20061	11451/618016	5	7.63E-06	1.85E-03
SC	BP	e	GO:0006725	cellular aromatic compound metabolic process	1130/20061	12346/618016	3	7.93E-06	1.92E-03
SC	BP	e	GO:0046483	heterocycle metabolic process	1130/20061	12346/618016	3	7.93E-06	1.92E-03
SC	BP	e	GO:1901360	organic cyclic compound metabolic process	1130/20061	12346/618016	3	7.93E-06	1.92E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	BP	e	GO:0006139	nucleobase-containing compound metabolic process	1130/20061	12346/618016	4	7.93E-06	1.92E-03
SC	BP	e	GO:0044249	cellular biosynthetic process	1437/20061	11269/618016	3	8.30E-06	2.01E-03
SC	BP	e	GO:1901576	organic substance biosynthetic process	1437/20061	11269/618016	3	8.30E-06	2.01E-03
SC	BP	e	GO:0009059	macromolecule biosynthetic process	1437/20061	11269/618016	4	8.30E-06	2.01E-03
SC	BP	e	GO:0043603	cellular amide metabolic process	1437/20061	11269/618016	4	8.30E-06	2.01E-03
SC	BP	e	GO:0044271	cellular nitrogen compound biosynthetic process	1437/20061	11269/618016	4	8.30E-06	2.01E-03
SC	BP	e	GO:1901566	organonitrogen compound biosynthetic process	1437/20061	11269/618016	4	8.30E-06	2.01E-03
SC	BP	e	GO:0006518	peptide metabolic process	1437/20061	11269/618016	5	8.30E-06	2.01E-03
SC	BP	e	GO:0034645	cellular macromolecule biosynthetic process	1437/20061	11269/618016	5	8.30E-06	2.01E-03
SC	BP	e	GO:0043604	amide biosynthetic process	1437/20061	11269/618016	5	8.30E-06	2.01E-03
SC	BP	e	GO:0043043	peptide biosynthetic process	1437/20061	11269/618016	6	8.30E-06	2.01E-03
SC	BP	e	GO:0006412	translation	1437/20061	11269/618016	7	8.30E-06	2.01E-03
SC	BP	e	GO:0006629	lipid metabolic process	808/20061	13507/618016	3	8.72E-06	2.11E-03
SC	BP	e	GO:0044281	small molecule metabolic process	1676/20061	19428/618016	2	9.66E-06	2.34E-03
SC	BP	e	GO:0009056	catabolic process	373/20061	9013/618016	2	1.05E-05	2.54E-03
SC	BP	e	GO:0055085	transmembrane transport	1269/20061	26703/618016	4	1.29E-05	3.12E-03
SC	BP	e	GO:0009058	biosynthetic process	3677/20061	40964/618016	2	1.50E-05	3.63E-03
SC	BP	e	GO:0034641	cellular nitrogen compound metabolic process	4535/20061	43601/618016	3	1.58E-05	3.81E-03
SC	BP	e	GO:0051179	localization	2802/20061	50086/618016	1	1.70E-05	4.11E-03
SC	BP	e	GO:0051234	establishment of localization	2802/20061	50086/618016	2	1.70E-05	4.11E-03
SC	BP	e	GO:0006810	transport	2802/20061	50086/618016	3	1.70E-05	4.11E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	BP	e	GO:0043170	macromolecule metabolic process	3105/20061	87795/618016	3	2.28E-05	5.51E-03
SC	BP	e	GO:0006807	nitrogen compound metabolic process	5350/20061	111424/618016	2	2.44E-05	5.91E-03
SC	BP	e	GO:0044237	cellular metabolic process	5539/20061	115940/618016	2	2.44E-05	5.91E-03
SC	BP	e	GO:0044238	primary metabolic process	4900/20061	129085/618016	2	2.64E-05	6.38E-03
SC	BP	e	GO:0071704	organic substance metabolic process	4900/20061	129085/618016	2	2.64E-05	6.38E-03
SC	BP	e	GO:0009987	cellular process	6235/20061	136120/618016	1	2.66E-05	6.44E-03
SC	BP	e	GO:0008152	metabolic process	7145/20061	162190/618016	1	2.75E-05	6.66E-03
SC	BP	e	GO:0050896	response to stimulus	603/20061	15747/618016	1	5.49E-05	0.0133
SC	BP	e	GO:0006950	response to stress	603/20061	15747/618016	2	5.49E-05	0.0133
SC	BP	p	GO:0040007	growth	0/20061	522/618016	1	1.37E-06	3.32E-04
SC	BP	p	GO:0043062	extracellular structure organization	0/20061	522/618016	3	1.37E-06	3.32E-04
SC	BP	p	GO:0030198	extracellular matrix organization	0/20061	522/618016	4	1.37E-06	3.32E-04
SC	BP	p	GO:0044085	cellular component biogenesis	0/20061	654/618016	2	2.15E-06	5.21E-04
SC	BP	p	GO:0022613	ribonucleoprotein complex biogenesis	0/20061	654/618016	3	2.15E-06	5.21E-04
SC	BP	p	GO:0042254	ribosome biogenesis	0/20061	654/618016	4	2.15E-06	5.21E-04
SC	BP	p	GO:0048856	anatomical structure development	0/20061	1112/618016	2	2.43E-06	5.88E-04
SC	BP	p	GO:0032502	developmental process	0/20061	1118/618016	1	2.67E-06	6.45E-04
SC	BP	p	GO:0071554	cell wall organization or biogenesis	0/20061	2873/618016	2	4.03E-06	9.74E-04
SC	BP	p	GO:0006091	generation of precursor metabolites and energy	27/20061	2901/618016	3	4.48E-06	1.08E-03
SC	BP	p	GO:0006457	protein folding	41/20061	3697/618016	2	4.77E-06	1.15E-03
SC	BP	p	GO:0050789	regulation of biological process	71/20061	11532/618016	2	8.22E-06	1.99E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	BP	p	GO:0050794	regulation of cellular process	71/20061	11532/618016	3	8.22E-06	1.99E-03
SC	BP	p	GO:0007165	signal transduction	71/20061	11532/618016	4	8.22E-06	1.99E-03
SC	BP	p	GO:0065007	biological regulation	218/20061	17434/618016	1	9.66E-06	2.34E-03
SC	BP	p	GO:0005975	carbohydrate metabolic process	505/20061	24039/618016	3	1.17E-05	2.84E-03
SC	BP	p	GO:0043412	macromolecule modification	452/20061	64734/618016	4	1.88E-05	4.55E-03
SC	BP	p	GO:0036211	protein modification process	452/20061	64734/618016	5	1.88E-05	4.55E-03
SC	BP	p	GO:0006464	cellular protein modification process	452/20061	64734/618016	6	1.88E-05	4.55E-03
SC	BP	p	GO:0044267	cellular protein metabolic process	1889/20061	75999/618016	5	2.02E-05	4.89E-03
SC	BP	p	GO:0019538	protein metabolic process	2038/20061	76497/618016	4	2.03E-05	4.91E-03
SC	BP	p	GO:0044260	cellular macromolecule metabolic process	2383/20061	81587/618016	4	2.12E-05	5.13E-03
SC	CC	e	GO:0044421	extracellular region part	152/20061	311/618016	1	5.06E-07	1.22E-04
SC	CC	e	GO:0031012	extracellular matrix	152/20061	311/618016	2	5.06E-07	1.22E-04
SC	CC	e	GO:0005856	cytoskeleton	37/20061	313/618016	5	5.77E-07	1.40E-04
SC	CC	e	GO:0009536	plastid	58/20061	378/618016	5	1.14E-06	2.77E-04
SC	CC	e	GO:0005783	endoplasmic reticulum	275/20061	854/618016	5	1.47E-06	3.56E-04
SC	CC	e	GO:0005694	chromosome	115/20061	718/618016	5	2.26E-06	5.47E-04
SC	CC	e	GO:0005737	cytoplasm	703/20061	4633/618016	3	5.44E-06	1.32E-03
SC	CC	e	GO:0005622	intracellular	956/20061	10036/618016	2	7.07E-06	1.71E-03
SC	CC	e	GO:0043228	non-membrane-bounded organelle	1591/20061	12495/618016	2	7.95E-06	1.92E-03
SC	CC	e	GO:0043232	intracellular non-membrane-bounded organelle	1591/20061	12495/618016	4	7.95E-06	1.92E-03
SC	CC	e	GO:1990904	ribonucleoprotein complex	1437/20061	11026/618016	2	8.54E-06	2.07E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	CC	e	GO:0044444	cytoplasmic part	1791/20061	13030/618016	3	8.56E-06	2.07E-03
SC	CC	e	GO:0005840	ribosome	1437/20061	11026/618016	5	8.54E-06	2.07E-03
SC	CC	e	GO:0005634	nucleus	640/20061	15292/618016	5	9.06E-06	2.19E-03
SC	CC	e	GO:0043227	membrane-bounded organelle	994/20061	17301/618016	2	9.89E-06	2.39E-03
SC	CC	e	GO:0043231	intracellular membrane-bounded organelle	994/20061	17301/618016	4	9.89E-06	2.39E-03
SC	CC	e	GO:0043226	organelle	2591/20061	29938/618016	1	1.27E-05	3.08E-03
SC	CC	e	GO:0044424	intracellular part	3299/20061	34591/618016	2	1.32E-05	3.18E-03
SC	CC	e	GO:0043229	intracellular organelle	2585/20061	29789/618016	3	1.31E-05	3.18E-03
SC	CC	e	GO:0032991	protein-containing complex	3549/20061	31947/618016	1	1.39E-05	3.36E-03
SC	CC	e	GO:0044464	cell part	3421/20061	42317/618016	1	1.55E-05	3.76E-03
SC	CC	p	GO:0030312	external encapsulating structure	0/20061	4071/618016	2	4.57E-06	1.11E-03
SC	CC	p	GO:0005618	cell wall	0/20061	4071/618016	3	4.57E-06	1.11E-03
SC	CC	p	GO:0044422	organelle part	5/20061	598/618016	1	1.83E-04	0.0443
SC	CC	p	GO:0044446	intracellular organelle part	5/20061	598/618016	3	1.83E-04	0.0443
SC	MF	e	GO:0019843	rRNA binding	214/20061	858/618016	5	1.56E-06	3.78E-04
SC	MF	e	GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	202/20061	1221/618016	3	2.16E-06	5.23E-04
SC	MF	e	GO:0045182	translation regulator activity	168/20061	2397/618016	1	3.44E-06	8.33E-04
SC	MF	e	GO:0090079	translation regulator activity, nucleic acid binding	168/20061	2397/618016	4	3.44E-06	8.33E-04
SC	MF	e	GO:0008135	translation factor activity, RNA binding	168/20061	2397/618016	5	3.44E-06	8.33E-04
SC	MF	e	GO:0004518	nuclease activity	298/20061	3725/618016	4	4.14E-06	1.00E-03
SC	MF	e	GO:0016874	ligase activity	420/20061	4776/618016	2	5.11E-06	1.24E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	MF	e	GO:0008289	lipid binding	163/20061	3286/618016	2	5.19E-06	1.26E-03
SC	MF	e	GO:0016853	isomerase activity	382/20061	5524/618016	2	5.42E-06	1.31E-03
SC	MF	e	GO:0016788	hydrolase activity, acting on ester bonds	413/20061	6118/618016	3	5.64E-06	1.37E-03
SC	MF	e	GO:0016741	transferase activity, transferring one-carbon groups	413/20061	9655/618016	3	7.08E-06	1.71E-03
SC	MF	e	GO:0008168	methyltransferase activity	413/20061	9655/618016	4	7.08E-06	1.71E-03
SC	MF	e	GO:0016829	lyase activity	574/20061	7558/618016	2	7.45E-06	1.80E-03
SC	MF	e	GO:0016746	transferase activity, transferring acyl groups	596/20061	10773/618016	3	7.66E-06	1.85E-03
SC	MF	e	GO:0003723	RNA binding	895/20061	12346/618016	4	7.93E-06	1.92E-03
SC	MF	e	GO:0003735	structural constituent of ribosome	1437/20061	11389/618016	2	8.17E-06	1.98E-03
SC	MF	e	GO:0005198	structural molecule activity	1451/20061	12200/618016	1	8.92E-06	2.16E-03
SC	MF	e	GO:0005215	transporter activity	905/20061	22509/618016	1	1.14E-05	2.75E-03
SC	MF	e	GO:0022857	transmembrane transporter activity	905/20061	22509/618016	2	1.14E-05	2.75E-03
SC	MF	e	GO:0097159	organic cyclic compound binding	1981/20061	51816/618016	2	1.71E-05	4.14E-03
SC	MF	e	GO:1901363	heterocyclic compound binding	1981/20061	51816/618016	2	1.71E-05	4.14E-03
SC	MF	e	GO:0003676	nucleic acid binding	1981/20061	51816/618016	3	1.71E-05	4.14E-03
SC	MF	e	GO:0042578	phosphoric ester hydrolase activity	115/20061	2393/618016	4	4.90E-05	0.0119
SC	MF	e	GO:0016791	phosphatase activity	115/20061	2393/618016	5	4.90E-05	0.0119
SC	MF	p	GO:0051082	unfolded protein binding	3/20061	1302/618016	3	2.09E-06	5.07E-04
SC	MF	p	GO:0042393	histone binding	0/20061	877/618016	3	2.27E-06	5.49E-04
SC	MF	p	GO:0019899	enzyme binding	3/20061	2246/618016	3	3.90E-06	9.45E-04
SC	MF	p	GO:0003924	GTPase activity	2/20061	2773/618016	7	4.56E-06	1.10E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	MF	p	GO:0008092	cytoskeletal protein binding	9/20061	4676/618016	3	4.93E-06	1.19E-03
SC	MF	p	GO:0016887	ATPase activity	84/20061	7330/618016	7	6.30E-06	1.53E-03
SC	MF	p	GO:0098772	molecular function regulator	149/20061	7248/618016	1	6.55E-06	1.59E-03
SC	MF	p	GO:0030234	enzyme regulator activity	149/20061	7248/618016	2	6.55E-06	1.59E-03
SC	MF	p	GO:0005515	protein binding	16/20061	9464/618016	2	7.57E-06	1.83E-03
SC	MF	p	GO:0016817	hydrolase activity, acting on acid anhydrides	105/20061	12352/618016	3	7.96E-06	1.93E-03
SC	MF	p	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	105/20061	12352/618016	4	7.96E-06	1.93E-03
SC	MF	p	GO:0016462	pyrophosphatase activity	105/20061	12352/618016	5	7.96E-06	1.93E-03
SC	MF	p	GO:0017111	nucleoside-triphosphatase activity	105/20061	12352/618016	6	7.96E-06	1.93E-03
SC	MF	p	GO:0016757	transferase activity, transferring glycosyl groups	382/20061	19461/618016	3	9.82E-06	2.38E-03
SC	MF	p	GO:0016798	hydrolase activity, acting on glycosyl bonds	365/20061	17114/618016	3	9.86E-06	2.39E-03
SC	MF	p	GO:0140110	transcription regulator activity	62/20061	18464/618016	1	1.04E-05	2.52E-03
SC	MF	p	GO:0003700	DNA-binding transcription factor activity	62/20061	18464/618016	2	1.04E-05	2.52E-03
SC	MF	p	GO:0140096	catalytic activity, acting on a protein	405/20061	17082/618016	2	1.07E-05	2.58E-03
SC	MF	p	GO:0008233	peptidase activity	405/20061	17082/618016	3	1.07E-05	2.58E-03
SC	MF	p	GO:0003677	DNA binding	1086/20061	39471/618016	4	1.49E-05	3.59E-03
SC	MF	p	GO:0016787	hydrolase activity	1489/20061	53542/618016	2	1.70E-05	4.12E-03
SC	MF	p	GO:0016491	oxidoreductase activity	1708/20061	59827/618016	2	1.83E-05	4.42E-03
SC	MF	p	GO:0016772	transferase activity, transferring phosphorus-containing groups	380/20061	63814/618016	3	1.85E-05	4.48E-03
SC	MF	p	GO:0016301	kinase activity	200/20061	58982/618016	4	1.90E-05	4.61E-03
SC	MF	p	GO:0016740	transferase activity	1807/20061	105241/618016	2	2.34E-05	5.65E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SC	MF	p	GO:0043167	ion binding	2320/20061	169425/618016	2	2.87E-05	6.93E-03
SC	MF	p	GO:0005488	binding	4353/20061	222950/618016	1	3.04E-05	7.36E-03
SC	MF	p	GO:0003824	catalytic activity	6161/20061	234033/618016	1	3.10E-05	7.51E-03
SC	MF	p	GO:0004386	helicase activity	55/20061	2693/618016	7	2.04E-04	0.0495
SN	BP	e	GO:0007059	chromosome segregation	46/77545	158/618016	2	1.69E-06	4.09E-04
SN	BP	e	GO:0007049	cell cycle	141/77545	468/618016	2	2.42E-06	5.86E-04
SN	BP	e	GO:0071941	nitrogen cycle metabolic process	59/77545	245/618016	3	2.65E-06	6.41E-04
SN	BP	e	GO:0007010	cytoskeleton organization	343/77545	869/618016	4	3.64E-06	8.82E-04
SN	BP	e	GO:0006886	intracellular protein transport	235/77545	663/618016	8	4.22E-06	1.02E-03
SN	BP	e	GO:0006605	protein targeting	235/77545	663/618016	9	4.22E-06	1.02E-03
SN	BP	e	GO:0051641	cellular localization	243/77545	1035/618016	2	4.86E-06	1.18E-03
SN	BP	e	GO:0051649	establishment of localization in cell	243/77545	1035/618016	3	4.86E-06	1.18E-03
SN	BP	e	GO:0046907	intracellular transport	243/77545	1035/618016	4	4.86E-06	1.18E-03
SN	BP	e	GO:0006790	sulfur compound metabolic process	349/77545	1865/618016	3	6.08E-06	1.47E-03
SN	BP	e	GO:0051276	chromosome organization	375/77545	2021/618016	4	6.37E-06	1.54E-03
SN	BP	e	GO:0015979	photosynthesis	684/77545	2478/618016	3	7.73E-06	1.87E-03
SN	BP	e	GO:0006091	generation of precursor metabolites and energy	789/77545	2901/618016	3	7.76E-06	1.88E-03
SN	BP	e	GO:0006996	organelle organization	728/77545	3013/618016	3	7.82E-06	1.89E-03
SN	BP	e	GO:0006457	protein folding	851/77545	3697/618016	2	8.15E-06	1.97E-03
SN	BP	e	GO:0051186	cofactor metabolic process	585/77545	3066/618016	3	8.23E-06	1.99E-03
SN	BP	e	GO:0034660	ncRNA metabolic process	666/77545	4249/618016	7	8.94E-06	2.16E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	BP	e	GO:0006399	tRNA metabolic process	666/77545	4249/618016	8	8.94E-06	2.16E-03
SN	BP	e	GO:0016192	vesicle-mediated transport	1139/77545	5271/618016	4	1.04E-05	2.53E-03
SN	BP	e	GO:0006259	DNA metabolic process	1291/77545	5589/618016	6	1.06E-05	2.58E-03
SN	BP	e	GO:0033036	macromolecule localization	1181/77545	6129/618016	2	1.08E-05	2.61E-03
SN	BP	e	GO:0008104	protein localization	1181/77545	6129/618016	3	1.08E-05	2.61E-03
SN	BP	e	GO:0045184	establishment of protein localization	1181/77545	6129/618016	4	1.08E-05	2.61E-03
SN	BP	e	GO:0071702	organic substance transport	1181/77545	6129/618016	4	1.08E-05	2.61E-03
SN	BP	e	GO:0071705	nitrogen compound transport	1181/77545	6129/618016	4	1.08E-05	2.61E-03
SN	BP	e	GO:0042886	amide transport	1181/77545	6129/618016	5	1.08E-05	2.61E-03
SN	BP	e	GO:0015833	peptide transport	1181/77545	6129/618016	6	1.08E-05	2.61E-03
SN	BP	e	GO:0015031	protein transport	1181/77545	6129/618016	7	1.08E-05	2.61E-03
SN	BP	e	GO:0006082	organic acid metabolic process	1278/77545	6311/618016	3	1.11E-05	2.69E-03
SN	BP	e	GO:0043436	oxoacid metabolic process	1278/77545	6311/618016	4	1.11E-05	2.69E-03
SN	BP	e	GO:0019752	carboxylic acid metabolic process	1278/77545	6311/618016	5	1.11E-05	2.69E-03
SN	BP	e	GO:0006520	cellular amino acid metabolic process	1278/77545	6311/618016	6	1.11E-05	2.69E-03
SN	BP	e	GO:0016043	cellular component organization	1184/77545	6391/618016	2	1.15E-05	2.79E-03
SN	BP	e	GO:0071840	cellular component organization or biogenesis	1212/77545	7045/618016	1	1.20E-05	2.89E-03
SN	BP	e	GO:0090304	nucleic acid metabolic process	2099/77545	11451/618016	5	1.49E-05	3.62E-03
SN	BP	e	GO:0006725	cellular aromatic compound metabolic process	2216/77545	12346/618016	3	1.53E-05	3.71E-03
SN	BP	e	GO:0046483	heterocycle metabolic process	2216/77545	12346/618016	3	1.53E-05	3.71E-03
SN	BP	e	GO:1901360	organic cyclic compound metabolic process	2216/77545	12346/618016	3	1.53E-05	3.71E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	BP	e	GO:0006139	nucleobase-containing compound metabolic process	2216/77545	12346/618016	4	1.53E-05	3.71E-03
SN	BP	e	GO:0044249	cellular biosynthetic process	2844/77545	11269/618016	3	1.56E-05	3.77E-03
SN	BP	e	GO:1901576	organic substance biosynthetic process	2844/77545	11269/618016	3	1.56E-05	3.77E-03
SN	BP	e	GO:0009059	macromolecule biosynthetic process	2844/77545	11269/618016	4	1.56E-05	3.77E-03
SN	BP	e	GO:0043603	cellular amide metabolic process	2844/77545	11269/618016	4	1.56E-05	3.77E-03
SN	BP	e	GO:0044271	cellular nitrogen compound biosynthetic process	2844/77545	11269/618016	4	1.56E-05	3.77E-03
SN	BP	e	GO:1901566	organonitrogen compound biosynthetic process	2844/77545	11269/618016	4	1.56E-05	3.77E-03
SN	BP	e	GO:0006518	peptide metabolic process	2844/77545	11269/618016	5	1.56E-05	3.77E-03
SN	BP	e	GO:0034645	cellular macromolecule biosynthetic process	2844/77545	11269/618016	5	1.56E-05	3.77E-03
SN	BP	e	GO:0043604	amide biosynthetic process	2844/77545	11269/618016	5	1.56E-05	3.77E-03
SN	BP	e	GO:0043043	peptide biosynthetic process	2844/77545	11269/618016	6	1.56E-05	3.77E-03
SN	BP	e	GO:0006412	translation	2844/77545	11269/618016	7	1.56E-05	3.77E-03
SN	BP	e	GO:0006629	lipid metabolic process	2210/77545	13507/618016	3	1.65E-05	3.98E-03
SN	BP	e	GO:0044281	small molecule metabolic process	4223/77545	19428/618016	2	1.95E-05	4.73E-03
SN	BP	e	GO:0009058	biosynthetic process	7945/77545	40964/618016	2	2.87E-05	6.95E-03
SN	BP	e	GO:0034641	cellular nitrogen compound metabolic process	8838/77545	43601/618016	3	3.06E-05	7.40E-03
SN	BP	e	GO:0007034	vacuolar transport	118/77545	661/618016	4	9.91E-05	0.024
SN	BP	p	GO:0032501	multicellular organismal process	0/77545	159/618016	1	8.49E-07	2.06E-04
SN	BP	p	GO:0007275	multicellular organism development	0/77545	159/618016	3	8.49E-07	2.06E-04
SN	BP	p	GO:0009790	embryo development	0/77545	159/618016	4	8.49E-07	2.06E-04
SN	BP	p	GO:0051301	cell division	0/77545	151/618016	2	1.52E-06	3.68E-04

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	BP	p	GO:0008283	cell population proliferation	8/77545	257/618016	1	1.77E-06	4.28E-04
SN	BP	p	GO:0051169	nuclear transport	8/77545	370/618016	5	2.90E-06	7.01E-04
SN	BP	p	GO:0006913	nucleocytoplasmic transport	8/77545	370/618016	6	2.90E-06	7.01E-04
SN	BP	p	GO:0040007	growth	24/77545	522/618016	1	3.24E-06	7.83E-04
SN	BP	p	GO:0043062	extracellular structure organization	24/77545	522/618016	3	3.24E-06	7.83E-04
SN	BP	p	GO:0030198	extracellular matrix organization	24/77545	522/618016	4	3.24E-06	7.83E-04
SN	BP	p	GO:0044085	cellular component biogenesis	28/77545	654/618016	2	3.62E-06	8.76E-04
SN	BP	p	GO:0022613	ribonucleoprotein complex biogenesis	28/77545	654/618016	3	3.62E-06	8.76E-04
SN	BP	p	GO:0042254	ribosome biogenesis	28/77545	654/618016	4	3.62E-06	8.76E-04
SN	BP	p	GO:0071554	cell wall organization or biogenesis	57/77545	2873/618016	2	8.01E-06	1.94E-03
SN	BP	p	GO:0006396	RNA processing	142/77545	1614/618016	7	8.65E-06	2.09E-03
SN	BP	p	GO:0016071	mRNA metabolic process	142/77545	1614/618016	7	8.65E-06	2.09E-03
SN	BP	p	GO:0006397	mRNA processing	142/77545	1614/618016	8	8.65E-06	2.09E-03
SN	BP	p	GO:0050789	regulation of biological process	1034/77545	11532/618016	2	1.50E-05	3.63E-03
SN	BP	p	GO:0050794	regulation of cellular process	1034/77545	11532/618016	3	1.50E-05	3.63E-03
SN	BP	p	GO:0007165	signal transduction	1034/77545	11532/618016	4	1.50E-05	3.63E-03
SN	BP	p	GO:0065007	biological regulation	1713/77545	17434/618016	1	1.97E-05	4.77E-03
SN	BP	p	GO:0055085	transmembrane transport	2906/77545	26703/618016	4	2.43E-05	5.87E-03
SN	BP	p	GO:0043412	macromolecule modification	3283/77545	64734/618016	4	3.69E-05	8.94E-03
SN	BP	p	GO:0036211	protein modification process	3283/77545	64734/618016	5	3.69E-05	8.94E-03
SN	BP	p	GO:0006464	cellular protein modification process	3283/77545	64734/618016	6	3.69E-05	8.94E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	BP	p	GO:0019538	protein metabolic process	6218/77545	76497/618016	4	3.95E-05	9.56E-03
SN	BP	p	GO:0044267	cellular protein metabolic process	6127/77545	75999/618016	5	3.95E-05	9.57E-03
SN	BP	p	GO:0044260	cellular macromolecule metabolic process	7417/77545	81587/618016	4	4.01E-05	9.71E-03
SN	BP	p	GO:0043170	macromolecule metabolic process	8309/77545	87795/618016	3	4.18E-05	0.0101
SN	BP	p	GO:1901564	organonitrogen compound metabolic process	7496/77545	82805/618016	3	4.17E-05	0.0101
SN	BP	p	GO:0006807	nitrogen compound metabolic process	12623/77545	111424/618016	2	4.69E-05	0.0113
SN	BP	p	GO:0044237	cellular metabolic process	13726/77545	115940/618016	2	4.65E-05	0.0113
SN	BP	p	GO:0044238	primary metabolic process	14319/77545	129085/618016	2	4.96E-05	0.012
SN	BP	p	GO:0071704	organic substance metabolic process	14319/77545	129085/618016	2	4.96E-05	0.012
SN	BP	p	GO:0009987	cellular process	16407/77545	136120/618016	1	5.14E-05	0.0124
SN	CC	e	GO:0009579	thylakoid	132/77545	200/618016	3	1.62E-06	3.93E-04
SN	CC	e	GO:0005811	lipid droplet	78/77545	269/618016	5	1.93E-06	4.67E-04
SN	CC	e	GO:0005856	cytoskeleton	154/77545	313/618016	5	2.03E-06	4.91E-04
SN	CC	e	GO:0005694	chromosome	194/77545	718/618016	5	3.39E-06	8.20E-04
SN	CC	e	GO:0005739	mitochondrion	196/77545	496/618016	5	3.39E-06	8.20E-04
SN	CC	e	GO:0005783	endoplasmic reticulum	154/77545	854/618016	5	8.12E-06	1.97E-03
SN	CC	e	GO:0005737	cytoplasm	909/77545	4633/618016	3	1.03E-05	2.50E-03
SN	CC	e	GO:0005622	intracellular	2007/77545	10036/618016	2	1.49E-05	3.60E-03
SN	CC	e	GO:0005840	ribosome	2799/77545	11026/618016	5	1.53E-05	3.71E-03
SN	CC	e	GO:0043228	non-membrane-bounded organelle	3237/77545	12495/618016	2	1.62E-05	3.92E-03
SN	CC	e	GO:0043232	intracellular non-membrane-bounded organelle	3237/77545	12495/618016	4	1.62E-05	3.92E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	CC	e	GO:0044444	cytoplasmic part	3214/77545	13030/618016	3	1.67E-05	4.05E-03
SN	CC	e	GO:0043229	intracellular organelle	5602/77545	29789/618016	3	2.51E-05	6.08E-03
SN	CC	e	GO:0043226	organelle	5606/77545	29938/618016	1	2.53E-05	6.11E-03
SN	CC	e	GO:0032991	protein-containing complex	6601/77545	31947/618016	1	2.60E-05	6.28E-03
SN	CC	e	GO:0044424	intracellular part	6674/77545	34591/618016	2	2.65E-05	6.41E-03
SN	CC	e	GO:0043227	membrane-bounded organelle	2365/77545	17301/618016	2	2.78E-05	6.72E-03
SN	CC	e	GO:0043231	intracellular membrane-bounded organelle	2365/77545	17301/618016	4	2.78E-05	6.72E-03
SN	CC	e	GO:0044464	cell part	7276/77545	42317/618016	1	2.99E-05	7.24E-03
SN	CC	p	GO:0000228	nuclear chromosome	4/77545	188/618016	6	1.30E-06	3.15E-04
SN	CC	p	GO:0042579	microbody	8/77545	281/618016	5	2.07E-06	5.01E-04
SN	CC	p	GO:0005777	peroxisome	8/77545	281/618016	6	2.07E-06	5.01E-04
SN	CC	p	GO:0044428	nuclear part	16/77545	357/618016	4	2.88E-06	6.96E-04
SN	CC	p	GO:0005576	extracellular region	94/77545	2060/618016	1	5.68E-06	1.37E-03
SN	CC	p	GO:0030312	external encapsulating structure	57/77545	4071/618016	2	8.91E-06	2.16E-03
SN	CC	p	GO:0005618	cell wall	57/77545	4071/618016	3	8.91E-06	2.16E-03
SN	MF	e	GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	278/77545	1578/618016	3	5.70E-06	1.38E-03
SN	MF	e	GO:0051082	unfolded protein binding	310/77545	1302/618016	3	5.76E-06	1.39E-03
SN	MF	e	GO:0045182	translation regulator activity	620/77545	2397/618016	1	7.28E-06	1.76E-03
SN	MF	e	GO:0090079	translation regulator activity, nucleic acid binding	620/77545	2397/618016	4	7.28E-06	1.76E-03
SN	MF	e	GO:0008135	translation factor activity, RNA binding	620/77545	2397/618016	5	7.28E-06	1.76E-03
SN	MF	e	GO:0004518	nuclease activity	711/77545	3725/618016	4	7.85E-06	1.90E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	MF	e	GO:0016874	ligase activity	1166/77545	4776/618016	2	9.33E-06	2.26E-03
SN	MF	e	GO:0008092	cytoskeletal protein binding	1116/77545	4676/618016	3	1.02E-05	2.48E-03
SN	MF	e	GO:0016853	isomerase activity	830/77545	5524/618016	2	1.14E-05	2.76E-03
SN	MF	e	GO:0098772	molecular function regulator	2024/77545	7248/618016	1	1.24E-05	2.99E-03
SN	MF	e	GO:0030234	enzyme regulator activity	2024/77545	7248/618016	2	1.24E-05	2.99E-03
SN	MF	e	GO:0005515	protein binding	1830/77545	9464/618016	2	1.37E-05	3.32E-03
SN	MF	e	GO:0003735	structural constituent of ribosome	2844/77545	11389/618016	2	1.51E-05	3.66E-03
SN	MF	e	GO:0003723	RNA binding	2594/77545	12346/618016	4	1.53E-05	3.71E-03
SN	MF	e	GO:0005198	structural molecule activity	3218/77545	12200/618016	1	1.65E-05	4.00E-03
SN	MF	e	GO:0140096	catalytic activity, acting on a protein	2381/77545	17082/618016	2	1.91E-05	4.61E-03
SN	MF	e	GO:0008233	peptidase activity	2381/77545	17082/618016	3	1.91E-05	4.61E-03
SN	MF	e	GO:0140110	transcription regulator activity	3082/77545	18464/618016	1	2.01E-05	4.86E-03
SN	MF	e	GO:0003700	DNA-binding transcription factor activity	3082/77545	18464/618016	2	2.01E-05	4.86E-03
SN	MF	e	GO:0003677	DNA binding	7222/77545	39471/618016	4	2.84E-05	6.88E-03
SN	MF	e	GO:0097159	organic cyclic compound binding	9816/77545	51816/618016	2	3.33E-05	8.05E-03
SN	MF	e	GO:1901363	heterocyclic compound binding	9816/77545	51816/618016	2	3.33E-05	8.05E-03
SN	MF	e	GO:0003676	nucleic acid binding	9816/77545	51816/618016	3	3.33E-05	8.05E-03
SN	MF	e	GO:0008134	transcription factor binding	55/77545	253/618016	3	5.49E-05	0.0133
SN	MF	p	GO:0042393	histone binding	59/77545	877/618016	3	4.04E-06	9.79E-04
SN	MF	p	GO:0042578	phosphoric ester hydrolase activity	151/77545	2393/618016	4	6.18E-06	1.49E-03
SN	MF	p	GO:0016791	phosphatase activity	151/77545	2393/618016	5	6.18E-06	1.49E-03

RC	GO	Rep.	GO ID	Name	Sample ratio	Population ratio	Depth	*P*	P_B
SN	MF	p	GO:0004386	helicase activity	254/77545	2693/618016	7	7.98E-06	1.93E-03
SN	MF	p	GO:0016829	lyase activity	702/77545	7558/618016	2	1.25E-05	3.02E-03
SN	MF	p	GO:0016887	ATPase activity	497/77545	7330/618016	7	1.27E-05	3.08E-03
SN	MF	p	GO:0016741	transferase activity, transferring one-carbon groups	1044/77545	9655/618016	3	1.45E-05	3.50E-03
SN	MF	p	GO:0008168	methyltransferase activity	1044/77545	9655/618016	4	1.45E-05	3.50E-03
SN	MF	p	GO:0016817	hydrolase activity, acting on acid anhydrides	1015/77545	12352/618016	3	1.64E-05	3.97E-03
SN	MF	p	GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	1015/77545	12352/618016	4	1.64E-05	3.97E-03
SN	MF	p	GO:0016462	pyrophosphatase activity	1015/77545	12352/618016	5	1.64E-05	3.97E-03
SN	MF	p	GO:0017111	nucleoside-triphosphatase activity	1015/77545	12352/618016	6	1.64E-05	3.97E-03
SN	MF	p	GO:0016798	hydrolase activity, acting on glycosyl bonds	1805/77545	17114/618016	3	1.94E-05	4.69E-03
SN	MF	p	GO:0016757	transferase activity, transferring glycosyl groups	1632/77545	19461/618016	3	1.99E-05	4.81E-03
SN	MF	p	GO:0005215	transporter activity	1986/77545	22509/618016	1	2.10E-05	5.08E-03
SN	MF	p	GO:0022857	transmembrane transporter activity	1986/77545	22509/618016	2	2.10E-05	5.08E-03
SN	MF	p	GO:0016787	hydrolase activity	6182/77545	53542/618016	2	3.32E-05	8.04E-03
SN	MF	p	GO:0016301	kinase activity	2846/77545	58982/618016	4	3.52E-05	8.53E-03
SN	MF	p	GO:0016491	oxidoreductase activity	4822/77545	59827/618016	2	3.53E-05	8.54E-03
SN	MF	p	GO:0016772	transferase activity, transferring phosphorus-containing groups	3527/77545	63814/618016	3	3.68E-05	8.90E-03
SN	MF	p	GO:0016740	transferase activity	7745/77545	105241/618016	2	4.58E-05	0.0111
SN	MF	p	GO:0043167	ion binding	17389/77545	169425/618016	2	5.50E-05	0.0133
SN	MF	p	GO:0005488	binding	26998/77545	222950/618016	1	5.86E-05	0.0142
SN	MF	p	GO:0003824	catalytic activity	21303/77545	234033/618016	1	5.97E-05	0.0145

^{03}), and extracellular matrix organisation (GO:0030198, GO:0043062; $P_B \leq 1.04e^{-03}$; Table 3.3.15.). These ontologs coupled with NC exclusive CC overrepresentations for microbodies (GO:0005777, GO:0042579; $P_B \leq 9.01e^{-04}$) and extracellular encapsulating structures (GO:0005777, GO:0042579; $P_B \leq 2.82e^{-03}$) are strongly attributed to plant defences against insect attack and fungal infection (Gols, 2014; Duan *et al*., 2016).

(*ii.*): Overrepresentations exclusive to strict composites

SC exclusive BP ontologs were observed to be significantly overrepresented for response to external stimuli (stress and stimulus response (GO:0006950, GO:0050896; $P_B \leq 0.0133$), secondary metabolism (GO:0009056, GO:0019748; $P \leq 2.54e^{-03}$), transport (GO:0006810, GO:0051179, GO:0051234; $P_B \leq 4.11e^{-03}$)). The appearance of secondary metabolism, stress response, and transport ontologs is highly complementary to the results observed for NC. Secondary metabolites are often produced in response to environmental stressors (*eg.* infection or animal attack) in plants (Wink, 2003, 2018; Hartmann, 2004) As secondary metabolites are often produced by large, multidomain and multifunctional genes (Pasek *et al*., 2006), an overrepresentation of secondary metabolism ontologs in SC is not surprising. These results are highly complementary to what was observed for NC, suggesting that the plant defence system evolved through rounds of gene remodelling.

(*iii.*): Overrepresentations exclusive to strict components

SN exclusive overrepresented BP ontologs were found to be significantly overrepresented ($P_B \leq 0.05$) for growth (GO:0040007; $P_B = 1.04e^{-03}$), transcription, specifically DNA and transcription factor binding (GO:0003677, GO:0008134; $P_B \leq 0.0133$), and for protein modification (GO:0036211, GO:0006464; $P \leq 0.011$) (Table 3.3.15). Protein modification is often mediated by small, non-specific, and often monomeric biocatalysts (Beltrao *et al.*, 2013). As discussed in Section 3.1., gene co-option is relatively common in plants, and these findings will be further discussed in Section 3.4. Due to their broad functionality and their co-option into biochemical pathways (Prabakaran *et al.*, 2012), it is not surprising to observe their overrepresentation as strict components.

(*iv.*): Overrepresentation exclusive to non-remodelled genes

NR exclusive BP ontologs were observed to be significantly overrepresented for embryonic developmental processes (GO:0007275, GO:0009790, GO:0032501, GO:0032502, GO:0048856; $P_B \leq 8.47e^{-04}$), ribosomal biogenesis (GO:0042254; $P_B = 5.73e^{-04}$), and cell division (GO:0051301; $P = 1.2e^{-04}$; Table 3.3.15.). These are all essential housekeeping functions in multicellular eukaryotes, and mutation in such processes often proves to be highly deleterious (Kasarskis *et al.*, 1998; Sparkes *et al.*, 2003) so the fact that they are overrepresented in non-remodelled families is not surprising

3.3.5. Ancient genes are more likely to be remodelled in Viridiplantae

Families of archaeal, bacterial, and undefined prokaryote origin were observed to be significantly overrepresented ($P \leq \alpha_B \leq 3.125e^{-03}$) for NC ($P \leq 8.97e^{-15}$) and SC ($2.98e^{-08}$) categories (Table 3.3.16.). Eukaryote specific families were overrepresented for NR ($1.20e^{-24}$). These results are similar to those observed in fungi where significant overrepresentations were observed for SN comparisons. Insignificant differences ($P \geq 0.32$) were observed for SN families of bacterial, eukaryotic and UP origin and for NR families of archaeal origin.

For NC, families of bacterial, archaeal, and UP origin were observed significantly more in Viridiplantae than fungi ($P \leq 1.69e^{-19}$; Table 3.3.17.), whereas a greater proportion of eukaryote originating families were observed in fungi ($P = 7.65e^{-48}$). Again, Viridiplantae were observed to have greater proportions of SC families in every DO comparison ($P \leq 3.33e^{-05}$) except for eukaryote originating families which were greater in fungi ($P = 3.33e^{-20}$). SN families of archaeal origin were observed in significantly greater proportions in Viridiplantae ($P = 6.02e^{-04}$). Finally, NR families of bacterial and UP origin were observed in significantly greater proportions in Viridiplantae ($P \leq 1.26e^{-19}$), and again, eukaryote specific NR family proportions were observed to be significantly greater in fungi ($P = 3.87e^{-62}$).

These results suggest that Viridiplantae possess greater proportions of more ancient gene families for RC than their fungal counterparts. While both datasets were observed to be highly plastic, Viridiplantae were observed to display more relative plasticity in their ancient families when compared to fungi which are more likely to subject “newer” families to molecular innovation.

3.3.6. Gene remodelling is more prominent in genomes that undergo frequent WGD

Table 3.3.17. Domains-of-Origin for each remodelling category in Viridiplantae

The count (*n*) of each RC per DO is presented on the left and *P*-values on the right. Significant observations ($P \leq \alpha_B \leq 4.16e^{-03}$) are annotated in bold. No significant comparisons were observed in SN

	Counts (*n*)				***P***			
	NC	**SC**	**SN**	**NR**	**NC**	**SC**	**SN**	**NR**
Archaea	401	85	429	213	$\mathbf{8.97e^{-15}}$	$\mathbf{5.06e^{-13}}$	0.992	1
Bacteria	2015	287	2500	1678	$\mathbf{2.33e^{-25}}$	$\mathbf{2.98e^{-08}}$	1	1
Eukaryote	12391	1598	25135	20564	1	1	0.016	$\mathbf{1.20e^{-64}}$
Prokaryote	5592	565	5631	2028	$\mathbf{1.78e^{-286}}$	$\mathbf{6.12e^{-09}}$	0.959	1

Table 3.3.18. Domain-of-Origin comparisons between fungi and Viridiplantae

For each RC, each DO is annotated with its count (n_{RC}), the count of the background (all families not annotated to the same RC; n_B) and the percentage (of total families; %) for each taxonomic kingdom. Comparisons between kingdoms are denoted by *P*. Significant comparisons ($P \leq \alpha_B \leq 0.0125$) are emboldened. The greater % value denotes which kingdom was significantly more likely to have gene families of a specific RC originating in a specific DO.

		Fungi			Viridiplantae			
RC	**DO**	n_{RC}	n_B	**%**	n_{RC}	n_B	**%**	***P***
NC	**Archaea**	141	16661	0.84	401	19998	1.97	**$2.38e^{-20}$**
	Bacteria	1319	15483	7.85	2015	18384	9.88	**$8.67e^{-12}$**
	Eukaryote	11425	5377	68.00	12391	8008	60.74	**$7.65e^{-48}$**
	Prokaryote	3917	12885	23.31	5592	14807	27.41	**$1.69e^{-19}$**
SC	**Archaea**	27	2080	1.28	85	2450	3.35	**$3.02e^{-06}$**
	Bacteria	162	1945	7.69	287	2248	11.32	**$3.33e^{-05}$**
	Eukaryote	1592	515	75.56	1598	937	63.04	**$3.33e^{-20}$**
	Prokaryote	326	1781	15.47	565	1970	22.29	**$3.92e^{-09}$**
SN	**Archaea**	208	21519	0.96	429	33266	1.27	**$6.02e^{-04}$**
	Bacteria	1660	20067	7.64	2500	31195	7.42	0.34
	Eukaryote	16158	5569	74.37	25135	8560	74.60	0.55
	Prokaryote	3701	18026	17.03	5631	28064	16.71	0.32
NR	**Archaea**	326	40484	0.80	213	24270	0.87	0.35
	Bacteria	1879	38931	4.60	1678	22805	6.85	**$1.11e^{-33}$**
	Eukaryote	36000	4810	88.21	20564	3919	83.99	**$3.87e^{-52}$**
	Prokaryote	2605	38205	6.38	2028	22455	8.28	**$1.26e^{-19}$**

Relative "global remodelling" and "internal (retained) remodelling" proportions were calculated (Tables 3.3.18.-19; Figure 3.3.6.) and compared to genomic characteristics for each taxon. When relative "globally remodelled" RC genomic proportions were compared to genomic characteristics, significant positive correlations ($P \leq \alpha_B \leq 0.01$; $0.3 < \rho \leq 0.7$) were observed between genome size and the SN proportion ($\rho = 0.5086$; $P = 0.002$), between genome density and the NR proportion ($\rho = 0.3878$; $P = 0.0054$), and between genomic completeness and both NC ($\rho = 0.5774$; $P < 0.0001$) and SC ($\rho = 0.5591$; $P < 0.0001$) proportions respectively (Figure 3.3.7.). Significant negative correlations ($-0.7 < \rho \leq -0.3$) were observed between genome size and the non-remodelled genomic proportion ($\rho = -0.4650$; $P = 0.0007$, between genome density and the strict component genomic proportion ($\rho = -0.5357$; $P < 0.0001$), and between completeness and the excluded genomic proportion ($\rho = 0.5591$; $P < 0.0001$). Positive correlations between completeness and genomic proportions are unsurprising as it can be reasonably expected that higher quality genomes would make better candidates for remodelled gene studies, negative correlations between completeness and excluded proportions are unsurprising for the same reason. Negative correlations between genome size and NR proportions are also unsurprising. As discussed in Section 3.1., Viridiplantae genomes evolve through cyclic polyploidisation which has the dual effect of increasing genome sizes and promoting remodelling events (Leonard and Richards, 2012; Madlung, 2013). The negative correlation between genome density and SN proportions was also unsurprising. Positive correlations have previously been established between angiosperm genome size and polyploidisation frequency (Adams and Wendel, 2005). While the reverse is observed for Gymnosperms, which have extremely sparse genomes but do not readily undergo polyploidisation (Wan *et al.*, 2018), they only contribute 3 genomes (6%) to these analyses, whereas angiosperms contribute 38 genomes (76%). As angiosperms constitute the vast majority of our dataset, this correlation was not surprising.

Table 3.3.19. Relative GRCPs in Viridiplantae

Each genome is annotated with the sum of genes (*n*) (and their proportion (%)) ascribed to each RC and those excluded from further analyses (E) for the "globally remodelled" dataset.

	n					**%**				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Amborella trichopoda	10761	826	3109	2632	9518	40.08	3.08	11.58	9.8	35.45
Ananas comosus	12179	879	2893	2626	8447	45.07	3.25	10.71	9.72	31.26
Arabidopsis thaliana	16184	1105	3471	3748	3147	58.52	4.00	12.55	13.55	11.38
Beta vulgaris	11991	896	2581	2795	8657	44.54	3.33	9.59	10.38	32.16
Brachypodium distachyon	17775	1053	3829	3536	8117	51.81	3.07	11.16	10.31	23.66
Brassica rapa	24347	1665	5904	5320	3256	60.13	4.11	14.58	13.14	8.04
Capsicum annum	18816	975	4914	2781	8398	52.44	2.72	13.69	7.75	23.4
Carica papaya	11370	683	3392	2357	9966	40.95	2.46	12.22	8.49	35.89
Chlamydomonas reinhardtii	2604	380	1431	2666	10660	14.68	2.14	8.07	15.03	60.09
Chlorella sp NC64A	2296	308	1014	1289	4854	23.52	3.16	10.39	13.21	49.73
Citrullus lanatus	12497	960	3270	2931	3782	53.31	4.1	13.95	12.5	16.13
Citrus clementina	15328	953	2982	2614	2656	62.48	3.88	12.16	10.66	10.83
Coccomyxa sp. C169	2609	331	889	1224	4941	26.11	3.31	8.90	12.25	49.44
Cucumis melo	12561	854	3615	3225	7172	45.8	3.11	13.18	11.76	26.15
Eucalyptus grandis	22096	1041	3697	2965	6550	60.79	2.86	10.17	8.16	18.02
Fragaria vesca	12667	757	2850	2311	13796	39.12	2.34	8.80	7.14	42.61
Glycine max	32182	2035	7119	6107	8601	57.42	3.63	12.7	10.9	15.35
Gossypium raimondii	22676	1428	4834	4306	4261	60.46	3.81	12.89	11.48	11.36
Hordeum vulgare	13063	601	3927	2176	4515	53.8	2.48	16.17	8.96	18.59
Malus domestica	12925	819	5442	3132	31604	23.97	1.52	10.09	5.81	58.61
Manihot esculenta	19389	1231	3995	3818	4600	58.7	3.73	12.09	11.56	13.93
Marchantia polymorpha	6408	680	1802	2319	8078	33.22	3.53	9.34	12.02	41.88
Medicago truncatula	22468	1297	5193	5037	16899	44.15	2.55	10.2	9.9	33.2
Micromonas commoda	2232	294	982	1382	5213	22.09	2.91	9.72	13.68	51.6
Musa acuminata	18730	1176	4074	2844	9704	51.28	3.22	11.15	7.79	26.57
Oryza brachyantha	13871	900	3082	2902	11282	43.3	2.81	9.62	9.06	35.22
Oryza sativa ssp. *japonica*	19516	1039	4239	4166	13229	46.26	2.46	10.05	9.87	31.36
Ostreococcus lucimarinus	2204	295	921	1377	3008	28.24	3.78	11.8	17.64	38.54
Phalaenopsis equestris	12081	799	3840	2499	10212	41.05	2.71	13.05	8.49	34.7
Phyllostachys edulis	15129	936	3714	2598	9610	47.3	2.93	11.61	8.12	30.04
Physcomitrella patens	10566	1104	3762	4231	13263	32.09	3.35	11.43	12.85	40.28

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Picea abies	20586	873	8799	4906	31468	30.9	1.31	13.21	7.36	47.23
Picea glauca	11898	990	4232	3110	8679	41.16	3.42	14.64	10.76	30.02
Pinus taeda	30962	1086	10511	3276	38611	36.66	1.29	12.45	3.88	45.72
Populus trichocarpa	22516	1369	4829	4397	9839	52.42	3.19	11.24	10.24	22.91
Prunus persica	15645	974	2926	2798	4530	58.22	3.62	10.89	10.41	16.86
Ricinus communis	12928	889	2798	2527	12079	41.41	2.85	8.96	8.09	38.69
Selaginella moellendorffii	9602	724	2380	2592	6987	43.09	3.25	10.68	11.63	31.35
Setaria italica	19435	1151	4158	4433	5407	56.2	3.33	12.02	12.82	15.63
Solanum lycopersicum	17300	1093	4653	3424	8255	49.82	3.15	13.4	9.86	23.77
Solanum tuberosum	19600	1027	4991	3580	9830	50.22	2.63	12.79	9.17	25.19
Sorghum bicolor	18297	1094	3774	4239	6807	53.48	3.2	11.03	12.39	19.9
Spirodela polyrhiza	9870	712	2375	2020	4646	50.3	3.63	12.1	10.29	23.68
Theobroma cacao	14910	899	3063	2821	7539	51.01	3.08	10.48	9.65	25.79
Triticum aestivum	63441	3294	13059	11950	11793	61.27	3.18	12.61	11.54	11.39
Vitis vinifera	12962	833	2869	2256	7426	49.2	3.16	10.89	8.56	28.19
Volvox carteri	2658	357	1556	2269	8704	17.1	2.3	10.01	14.6	56
Zea mays	19577	1161	5647	4483	8630	49.56	2.94	14.3	11.35	21.85
Zostera marina	10836	895	2791	2118	3810	52.99	4.38	13.65	10.36	18.63
Zoysia japonica	12342	719	3905	2876	33783	23.02	1.34	7.28	5.36	63

Table 3.3.20. Relative IRCPs in Viridiplantae

Each genome is annotated with the sum of genes (*n*) (and their proportion (%)) ascribed to each RC and those excluded from further analyses (E) for the "internally remodelled" dataset.

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Amborella trichopoda	1252	555	1519	5440	18080	4.66	2.07	5.66	20.26	67.35
Ananas comosus	1342	260	1542	7391	16489	4.97	0.96	5.71	27.35	61.02
Arabidopsis thaliana	2041	214	1810	12700	10890	7.38	0.77	6.54	45.92	39.38
Beta vulgaris	1079	240	970	7686	16945	4.01	0.89	3.6	28.55	62.95
Brachypodium distachyon	2792	318	2250	11174	17776	8.14	0.93	6.56	32.57	51.81
Brassica rapa	3807	896	4279	22271	9239	9.4	2.21	10.57	55	22.82
Capsicum annum	4731	623	3197	9877	17456	13.18	1.74	8.91	27.52	48.65
Carica papaya	1445	346	998	5785	19194	5.2	1.25	3.59	20.83	69.12
Chlamydomonas reinhardtii	247	76	370	2174	14874	1.39	0.43	2.09	12.25	83.84
Chlorella sp NC64A	44	12	116	1278	8311	0.45	0.12	1.19	13.09	85.14
Citrullus lanatus	1337	435	1434	7911	12323	5.7	1.86	6.12	33.75	52.57
Citrus clementina	2752	391	1454	8566	11370	11.22	1.59	5.93	34.92	46.35
Coccomyxa sp. C169	6	10	99	1510	8369	0.06	0.1	0.99	15.11	83.74
Cucumis melo	1161	469	1598	8124	16075	4.23	1.71	5.83	29.62	58.61
Eucalyptus grandis	6116	882	2215	11961	15175	16.83	2.43	6.09	32.91	41.75
Fragaria vesca	1901	218	2069	6205	21988	5.87	0.67	6.39	19.16	67.9
Glycine max	6902	1643	6610	27464	13425	12.32	2.93	11.79	49	23.95
Gossypium raimondii	3609	990	3736	18153	11017	9.62	2.64	9.96	48.4	29.37
Hordeum vulgare	1936	444	1425	6352	14125	7.97	1.83	5.87	26.16	58.17
Malus domestica	775	389	1559	8336	42863	1.44	0.72	2.89	15.46	79.49
Manihot esculenta	3133	383	2499	15592	11426	9.48	1.16	7.57	47.2	34.59
Marchantia polymorpha	424	366	446	3985	14066	2.2	1.9	2.31	20.66	72.93
Medicago truncatula	6272	743	4411	14736	24732	12.32	1.46	8.67	28.95	48.6
Micromonas commoda	1	31	50	1201	8820	0.01	0.31	0.49	11.89	87.3
Musa acuminata	2378	580	2853	13652	17065	6.51	1.59	7.81	37.37	46.72
Oryza brachyantha	1863	251	1107	7845	20971	5.82	0.78	3.46	24.49	65.46
Oryza sativa ssp. *japonica*	4321	598	2628	11215	23427	10.24	1.42	6.23	26.58	55.53
Ostreococcus lucimarinus	3	17	75	1528	6182	0.04	0.22	0.96	19.58	79.21
Phalaenopsis equestris	2093	588	1622	6856	18272	7.11	2	5.51	23.3	62.08
Phyllostachys edulis	1926	359	1691	9917	18094	6.02	1.12	5.29	31	56.57

	n					%				
	NC	**SC**	**SN**	**NR**	**E**	**NC**	**SC**	**SN**	**NR**	**E**
Physcomitrella patens	993	218	1867	12231	17617	3.02	0.66	5.67	37.15	53.5
Picea abies	8405	1278	5251	9859	41839	12.61	1.92	7.88	14.8	62.79
Picea glauca	1477	827	1550	7813	17242	5.11	2.86	5.36	27.03	59.64
Pinus taeda	15051	1409	10016	9207	48763	17.82	1.67	11.86	10.9	57.74
Populus trichocarpa	1956	420	1564	10290	28720	4.55	0.98	3.64	23.96	66.87
Prunus persica	2271	519	1674	10027	12382	8.45	1.93	6.23	37.31	46.08
Ricinus communis	1362	301	1069	7726	20763	4.36	0.96	3.42	24.75	66.5
Selaginella moellendorffii	2371	290	1390	7249	10985	10.64	1.3	6.24	32.53	49.29
Setaria italica	2615	580	2531	13697	15161	7.56	1.68	7.32	39.61	43.84
Solanum lycopersicum	1875	647	2429	11494	18280	5.4	1.86	6.99	33.1	52.64
Solanum tuberosum	5761	702	3381	10892	18292	14.76	1.8	8.66	27.91	46.87
Sorghum bicolor	2315	398	1938	12798	16762	6.77	1.16	5.66	37.41	49
Spirodela polyrhiza	1100	191	804	5243	12285	5.61	0.97	4.1	26.72	62.61
Theobroma cacao	2171	265	1661	8839	16296	7.43	0.91	5.68	30.24	55.75
Triticum aestivum	25082	2495	15040	43392	17528	24.23	2.41	14.53	41.91	16.93
Vitis vinifera	1481	235	1543	7222	15865	5.62	0.89	5.86	27.41	60.22
Volvox carteri	212	162	455	1835	12880	1.36	1.04	2.93	11.81	82.86
Zea mays	3721	778	4351	12862	17786	9.42	1.97	11.02	32.56	45.03
Zostera marina	1012	263	1085	6890	11200	4.95	1.29	5.31	33.69	54.77
Zoysia japonica	1052	377	1805	8226	42165	1.96	0.7	3.37	15.34	78.63

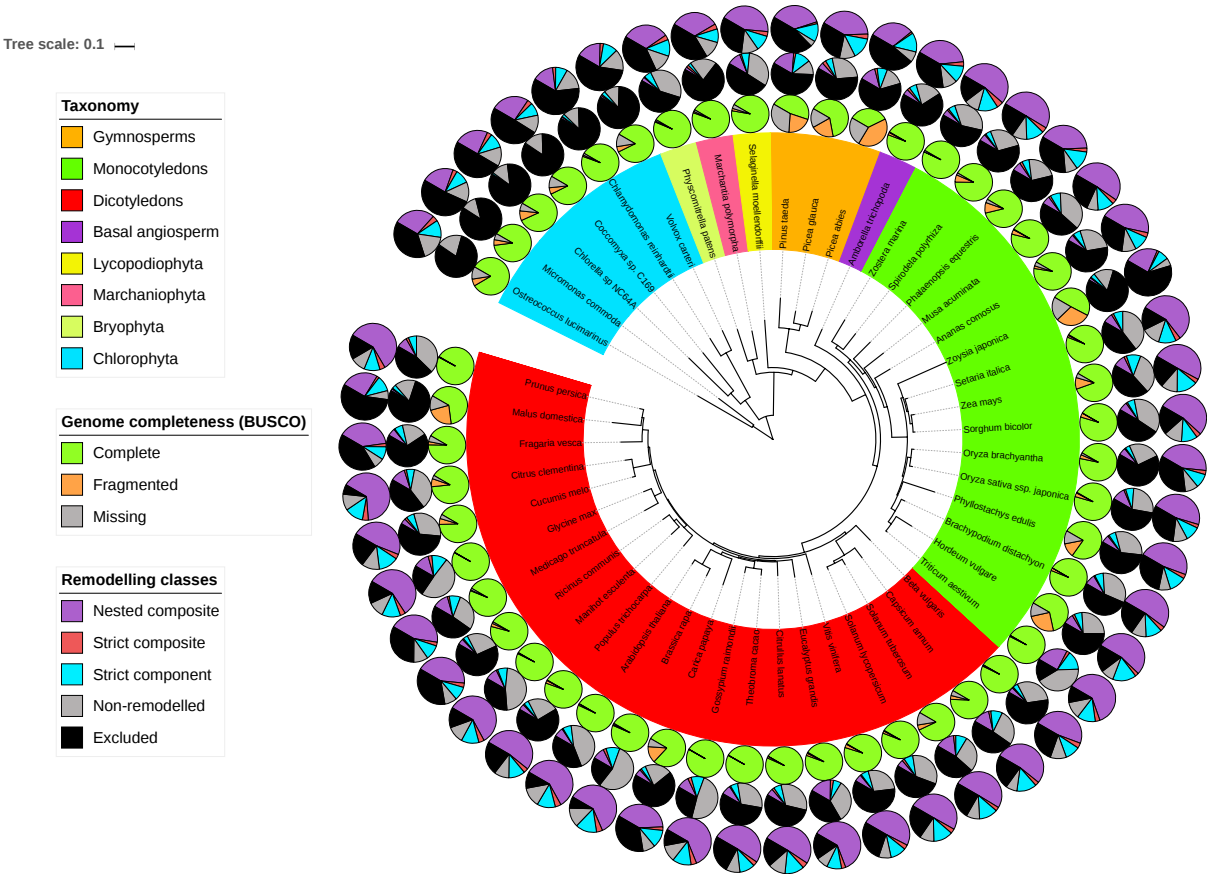

Figure 3.3.6. Representative Viridiplantae phylogeny annotated with genomic characteristics

The Viridiplantae phylogeny (Figure 2.2.3.) was annotated with three genomic datasets, each represented as pie charts. The internal pie chart represents BUSCO completeness, the middle chart represents IRCPs (Table 3.3.19.), and the outer charts represent GRCPs (Table 3.3.18.).

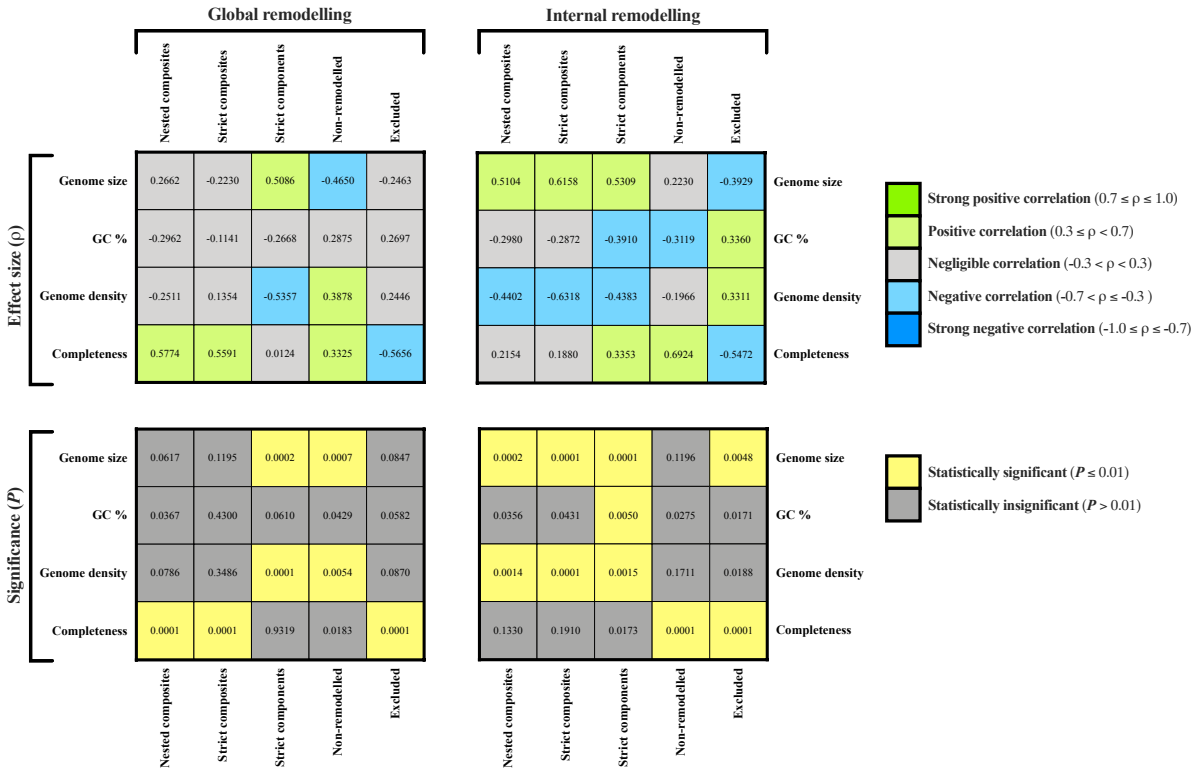

Figure 3.3.7. Correlation matrix between genomic characteristics and gene remodelling extent in Viridiplantae

Illustration of effect sizes (Spearman's ρ) and their significance (*P*) between genomic characteristics for (*a*) GRCPs (Table 3.3.19) on the left and (*b*) IRCPs (Table 3.3.20) on the right.

When relative "internally remodelled" (retained remodelled) RC genomic proportions were compared to genomic characteristics, significant positive correlations were observed between genome size and each remodelled (NC, SC, and SN) proportion ($\rho \geq 0.5104$; $P \leq 0.0002$) and between completeness and the NR proportion ($\rho = 0.6924$; $P < 0.0001$). These results are unsurprising as larger Viridiplantae genomes have typically undergone rounds of polyploidisation which, as discussed previously, promotes remodelling events. It can reasonably be expected that greater genome completeness would positively correlate with the NR proportion due to a higher rate of paralog inclusion when CompositeSearch is constructing families. As a family was required to have a minimum of two members, inadvertent paralog exclusion due to genome incompleteness would result in a non-remodelled gene being classified as excluded. Significant negative correlations were observed between genome density and each remodelled category ($\rho \leq 0.4383$; $P \leq 0.0015$), between GC% and the SN proportions ($\rho = -0.391$; $P = 0.005$), and between each of genome size, and completeness and excluded proportions ($\rho \leq -0.3929$; $P \leq 0.0048$). Again these results are all unsurprising. As a retained remodelling event requires both component families and composite family to be observed within the same genome, genomes replete with polyploidisation events would reasonably be expected to have larger and less dense genomes and greater instances of duplicate genes and a greater frequency of remodelling events overall. In this dataset, much higher GC% were observed in Chlorophyta, which had more dense genomes and lesser remodelled proportions compared to Embryophyta (Table 3.3.1.). Due to these factors, a negative correlation was to be expected between GC% and strict components. Finally, negative correlations were expected between genome size and excluded proportions, and between completeness and excluded proportions. Larger genomes in this dataset were more likely to possess higher frequencies of paralogs (and ohnologs) due to frequent gene and genome duplication (Adams and Wendel, 2005), and as CompositeSearch required a minimum of two

members for each gene family, these results were to be expected; the negative correlation between completeness and excluded proportions were also expected for the same reason.

3.4. Discussion

3.4.1. Gene remodelling is rampant in Viridiplantae

Overall, approximately 60.95% of genes (1,019,409 of 1,672,377 genes) were observed to display a history of remodelling when sampled from the globally remodelled dataset (Tables 3.3.3; 3.3.18). In total, NC, SC, and SN each accounted for an approximately 46.33% (774,886 genes), 0.029% (48,440 genes), and 11.72% (196,083 genes) respectively.

It must be noted that only 70.76% (1,183,398 genes) were sampled from the globally remodelled dataset meaning that approximately 85.69% were observed to have a history of remodelling (Table 3.3.3.). Of sampled genes, NC, SC, and SN accounted for 65.48%, 0.04%, and 16.57% respectively.

Clear distinctions were observed for gene remodelling extent between fungi and plants. With regards to genes (when examined from across the dataset and from just the data sampled by CompositeSearch), plants were observed to have greater proportions of NC and SC than fungi ($P \leq 3.02e^{-45}$; Table 3.3.4.). Comparatively, fungi always presented greater proportions of SN and NR ($P = 0$) than plants. Plants were observed to have greater proportions of all remodelled category (NC, SC, and SN) families ($P \leq 8.44e^{-13}$) and fungi were observed to have a greater proportion of NR families ($P = 0$). This is likely due to the fact that the plant dataset contained considerably more genes than the fungal dataset and that plants were observed to form significantly larger gene families ($P \leq 2.37e^{-208}$; Table 3.3.7). These results suggest that

while both clades are highly plastic, plant genomes are considerably more dynamic and subject to remodelling than their fungal counterparts.

3.4.2 Remodelling mediated evolution is clocklike in Viridiplantae

We could not detect any internal branches with significant ($P \leq 0.0005$ (all data); $P \leq 0.001$ (internal branch subset exclusive)) bursts of birth or decay using a Q-function after the application of a Bonferroni correction. One speciation events were observed to contain bursts (Table 3.3.13). Significant bursts of nested composite and non-remodelled gene birth ($P \leq 0.0005$) was observed during the speciation of *Triticum aestivum*. These sparse results are consistent to what was observed in fungi, suggesting that evolution *via* gene remodelling is relatively clocklike.

With the exception of SC ($P \leq 9.54e^{-04}$), significant differences ($P \leq \alpha_B \leq 0.0125$) were not observed between evolutionary rates (births or decays) when sampled from across plant and fungal phylogenies, and a significant difference was only observed in SC f_d ($P = 1.89e^{-03}$) when sampled from exclusively internal nodes (Table 3.3.13.). Significant differences were observed, however, when leaf node exclusive rates were compared. Significant differences were observed in SC and SN f_b and in SN and NR f_d. These comparisons strengthen the argument that evolution *via* remodelling occurs at a relatively clocklike rate between both clades.

3.4.3 Remodelled genes are highly homoplastic in Viridiplantae

A total of 20,399 NC families, 2,535 SC families, 36,695 SN families, and 24,483 NR families were constructed by CompositeSearch. Any family observed to have appeared more

than once (annotated to have originated at separate branches) was considered to be homoplastic (Table 3.3.8.). A total of 10,444 nested composites (H.P. = 0.497), 1,545 strict composite families (H.P. = 0.609), 15,423 strict component families (H.P. = 0.458), and 8,559 non-remodelled families (H.P. = 0.35) were observed to be homoplastic. A combined total of 84,112 families were plotted, of which, a combined total of 35,971 families were homoplastic (H.P. = 0.428). A Fisher's exact test confirmed that remodelled categories were significantly different to each other ($P \leq 4.78e^{-14}$). These results highlight the dynamic homoplasticity of remodelled gene families compared to non-remodelled families, and between each other, and compared to the dataset background. The observation of homoplasticity in remodelled genes is not so surprising as it is possible that remodelled genes may have arose through epaktology or independent gene fusion events (Nagy and Patthy, 2011; Leonard and Richards, 2012; Avelar *et al.*, 2014; Haggerty *et al.*, 2014). Homoplasticity could also be attributed to a quirk of clustering by CompositeSearch, where large gene families may fall just outside cluster assignment criteria in some groups resulting in a false positive identification due to patchy annotation by the "-apo" function in TNT .

Viridiplantae gene families were also observed to be more homoplastic than fungi in every instance ($P \leq 9.24e^{-05}$) except for NC ($P = 0.044$) when sampled from across the phylogeny and SC ($P = 0.824$) when sampled from the subset of internal nodes.

3.4.4. The role of gene remodelling in the evolution of multicellularity in plants

We observed an enrichment of transcription factor regulatory activity (GO:0140110; GO:0003700; $P_B \leq 4.86e^{-03}$) within both SC and SN (Table 3.3.15.). In our dataset the evolution of multicellularity occurred at two separate occasions, (*i*) the speciation of the alga *Volvox carteri* and (*ii*) during the divergence of embryophytes from chlorophytes. We identified a

single family associated with with transcription factor regulation and multicellular development reported to have been gained at the branch representing embryophyte divergence (F54088), where the ortholog in *Arabidopsis thaliana* was reported to be At2g41980 (SINAT1) an E3 ubiquitin-protein ligase (E.C:2.3.2.27) (Qi *et al.*, 2017).

Regulatory protein ubiquitination is required for a wide variety of plant development and environmental response pathways (Wilkinson, 1999; Furlan *et al*., 2012; Duplan and Rivas, 2014; Miricescu *et al*., 2018). Protein polyubiquitination is dependent on the activity of three enzymes for the activation (E1), conjugation (E2), and ligation (E3) of ubiquitin (Serrano *et al*., 2018). Polyubiquitinated proteins are degraded to constituent amino acids *via* the 26S proteasome for the synthesis of new peptides. E1 and E2 initiate and promote polyubiquitin chain progression, whereas E3 is responsible for target protein selectivity (Sadowski and Sarcevic, 2010). Several E3 families have previously been identified such as U-box proteins (Hatakeyama *et al.*, 2001), Skp-Cullin-F box proteins (Cheng *et al.*, 2011), and anaphase-promoting complexes (Castro *et al.*, 2005) throughout life and all share a RING finger motif or an E6-associated protein carboxyl terminus (HECT) domain which catalyses ubiquitin ligation to lysine residues (Huibregtse *et al.*, 1995; Metzger *et al*., 2012).

Seven in absentia (SINA) proteins, such as SINAT1, are E3 ligases that possess a RING finger motif at their N-termini, neighboured by a conserved SINA domain which functions to bind and dimerise their specific substrate (Hu and Fearon, 1999; Miao *et al.*, 2016). SINA proteins have been observed and well characterised in metazoan (Pepper *et al*., 2017) and plant lineages (M. Wang *et al.*, 2008). In *A. thaliana*, SINAT1 acts with SINAT2 and SINAT6 to regulate AG6 mediated autophagy (Qi *et al.*, 2017). A large variety of hormone mediated plant developmental pathways, including those under the control of cytokinin, auxin, gibberellic acid, jasmontic acid, and DELLA, are at least partially dependant on proteasome degradation, resulting in hormonal defects in E3 knockout studies (Foltz *et al.*, 2006; Ning *et al.*, 2011; Qi

et al., 2017). In light of observations by previous studies, it is not surprising to observe proteasomal mediated regulators during the divergence of embryophytes, mirrored by previous reports of increases in hormonal complexity and transcription factor copy number and class types during this transition to reflect more complex body plans (Bennici, 2008; de Vries *et al.*, 2016; de Vries and Archibald, 2018; Morris *et al.*, 2018).

Of the 26 component families associated with the composition of F54088, two were observed to have occurred either during the emergence or prior to the emergence of Viridiplantae, F401 and F106717. F106717 had a single member in *A. thaliana*, At3g23580 (*RNR2A*), a ribonucleoside-diphosphate reductase (E.C: 1.17.4.1). Ribonucleoside diphosphate reductases catalyses the 2' reduction of ribonucleotides to deoxyribonucleotides, the precursors required for DNA synthesis (Cory, 1983; Guarino *et al.*, 2014). There are two known *RNR2A* paralogs in *A. thaliana* (*RNR2B* and *TSO2*), however they did not meet the clustering requirements to be assigned to the same family by CompositeSearch. Previous studies have demonstrated functional redundancy between *RNR2A, RNR2B*, and *TSO2* with no phenotypic differences observed in either *RNR2A* or *RNR2B* single mutants or *RNR2A*/*RNR2B* double mutants compared to wild type *A. thaliana.* In comparison, *TSO2* mutants displayed reduced deoxyribonucleotide triphosphate (dNTP) accumulation and developmental defects, such as fasciated shoot meristems and callus floral organs (Wang and Liu, 2006). Both TSO2 single mutants and *TSO2*/*RNR2A* double mutants displayed increased DNA damage accumulation, massive apoptosis rates, and decay of transcriptional silencing. *TSO2*/*RNR2A*/*RNR2B* triple mutants were observed to be seedling lethal. Wang and Liu (2006) hypothesised that the functional redundancy between *RNR2A* and *RNR2B* to be a safeguard against mutagenic lethality should one paralog undergo mutation, and illustrates the essentiality of proper RNR function in plant development, and that the observations in *TSO2*/*RNR2A* double mutants illustrated a correlation between increased DNA damage and apoptosis. The finding of a

developmental regulator undergoing a remodelling event to form a new developmental regulator during a major phenotypic transition is interesting, yet unsurprising as previous remodelling studies (specifically gene fusion studies) allude to the likelihood of a fused gene retaining at least one of its components functions.

The other component family predating the divergence of Viridiplantae, F401, is a large family, with 75 members reported in *A. thaliana.* Each member of F401 was found to possess leucine rich repeats (PF00560). The majority of F401 members were observed to be 'receptor like proteins' (eg. AT1G58190, AT2G32660, AT5G27060, and AT1G47890) which are often involved in immune responses and developmental processes (Godiard *et al.*, 2003; Kruijt *et al*., 2005; Wang *et al.*, 2008). Indeed, some F401 member genes (AT5G06860 (*PGIP1*), AT5G06870 (*PGIP2*), and AT3G05360 (*RLP30*)) have been characterised to inhibit phytopathogenic fungal infection (Ferrari *et al.*, 2003, 2006; Zhang *et al.*, 2013). Conversely, four proteins (AT1G65380 (*CLV2*), AT1G17240 (*RLP2*), AT1G71400 (*RLP12*), and AT3G12145 (*FLR1*)) are known to be involved in organ developmental processes in addition to immunity (Wang *et al.*, 2008; Torti *et al.*, 2012). Three of these proteins (*CLV2*, *RLP12,* and *RPL2*)) are *CLAVATA2* and *CLAVATA2*-like RLPs which act as receptors for *CLV3* and *CLV3*-like proteins (Rojo *et al.*, 2002; Wang *et al.*, 2008; Fletcher, 2018). *CLV3* and *CLV3*-like proteins are extracellular hormone precursor signalling molecules that regulate meristem maintenance (Doerner, 2006). *CLV2* and *CLV2*-like RLPs control the sizes of totipotent cell populations in meristematic tissues (Kayes and Clark, 1998; Pan *et al.*, 2016). F401 describes a large cluster of signal receptors involved in diverse pathways. Again, it is interesting yet unsurprising to see developmental regulators being remodelled during emergence of phenotypic complexity. *CLV* genes and *FLR* both interact with another repurposed developmental regulator class, the MADS box genes, to fully orchestrate organogenesis in angiosperms (Kanno *et al.*, 2007).

3.5. Conclusion

In conclusion, Viridiplantae are known to evolve *via* cyclical polyploidization events, where the size of their genetic arsenal doubles, resulting in considerable redundancy, followed by chromosomal rearrangements and decay of redundant, non-selected genes. This model of evolution not only allow for gene remodelling but promote it, as such, it is not surprising to observe such rampant levels of gene remodelling in this clade. Gene remodelling is also clocklike in this clade, with no significant bursts of evolution (composites or otherwise) observed along any internal branch. Considering the rampancy of remodelling, it could be due to the fact that, when successful, composite families comprise a significant proportion of all retained synapomorphic families. The co-option and subfunctionalization of transcription factors could be due to repeated rounds of homoplastic remodelling, thus promoting rapid evolution in the Viridiplantae lineage.

In summation, the models of evolution, propensity for hybridization, and chromosomal architectures provide a nurturing environment for remodelling to occur, so perhaps it is not so surprising to observe such high remodelling rates.

Chapter IV

Development of a

Robust Composite Gene Detection Tool

4.1. Introduction

Gene remodelling is an important and rampant evolutionary process (Pathmanathan *et al.,* 2018). As previously discussed in Chapters I-III, remodelled genes arise from a plethora of different mechanisms including the rapid accumulation of point mutations localised within in a structural or functional domain, shuffling of exons or domains within a gene, or through the fusion and fission of genes and domains (Snel *et al.*, 2000; Braun and Grotewold, 2001; Vogel *et al.,* 2005; Vibranovski *et al.*, 2006; Nagy and Patthy, 2011; Leonard and Richards, 2012). Gene duplication is likely the driving force behind beneficial remodelling events through subfunctionalization and neofunctionalization of non-selected genes, allowing them to accrue mutations or to acquire or lose domains (Causier *et al.*, 2005; He and Zhang, 2005; Des Marais and Rausher, 2008; Freeling, 2009). Due to these events, combined with convergent sequence evolution and epaktologous gene birth (emergence of a gene with the same architecture as another gene but lacking an orthologous or paralogous evolutionary history), it is common to observe partial homology between genes (Haggerty *et al.*, 2014).

Hybridised coding gene composites (and their components), specifically fusion and fission genes, are of considerable interest as they allow for the study and interpretation of evolutionary protein-protein interactions (Enright *et al.*, 1999; Enright and Ouzounis, 2001) or for the rapid development of functional and phenotypic novelty (Avelar *et al.*, 2014; Chen *et al.*, 2014). A successful fusion event (a fusion event that becomes positively selected) is likely to become fixed in a population and have been used to polarise evolutionary relationships (Nakamura *et al.,* 2007; Durrens *et al.*, 2008).

Previous authors have identified subsets of composite genes in *Drosophila* spp, fungi, plants, and bacteria (Wang *et al.*, 2000; Jones, 2005; Pasek *et al.*, 2006; Nakamura *et al.*, 2007; Leonard and Richards, 2012), and the role of composites in metabolic pathway evolution

(Richards *et al.*, 2006; Hagel and Facchini, 2017). Despite these works, it was not possible to perform exhaustive composite gene detection analyses until the publication of CompositeSearch (Pathmanathan *et al.,* 2018).

As discussed in Chapter I, CompositeSearch is a highly sensitive tool, able to detect even slightly shared homologous between a composite and component genes, and is therefore a useful tool for reconstructing the mosaic evolutionary history of a gene family. CompositeSearch, however, is quite unintuitive, and requires considerable postprocessing by the user to analyse any detected composite family in detail. CompositeSearch also makes use of multiple HSPs to detect composites (Pathmanathan *et al.,* 2018). This allows for the detection of composites with a likely shared ancestor between both components (Figure 4.1.1). While these events are still remodelled they are not fusion or fission events. Comparatively, *fdf*BLAST (Leonard and Richards, 2012) is highly selective, requiring not just a high degree of polarised homology between a composite and a component, but shared conserved regions are also required. However, *fdf*BLAST it can be argued that such analyses are too conservative, requiring that a component is differential, where a fused gene is detected in one genome, and two components (but not the fusion) in another genome. In addition to these strict parameters, *fdf*BLAST does not cluster fusion genes into families (fusion events), is dependent on "BLASTall", an outdated version of BLAST and does not use tabular BLAST output, resulting in excessive storage requirements for multi-genomic analyses. We also found *fdf*BLAST to be slow and cumbersome to use.

A need exists for a program which can sensitively and selectively detect fused genes that is also intuitive, has implemented quality control measures, requiring the optional conservation of domain architectures. To fill this need, we have developed compositeBLAST, a Python *v*.3.6 program high confidence composite gene detection and visualisation package.

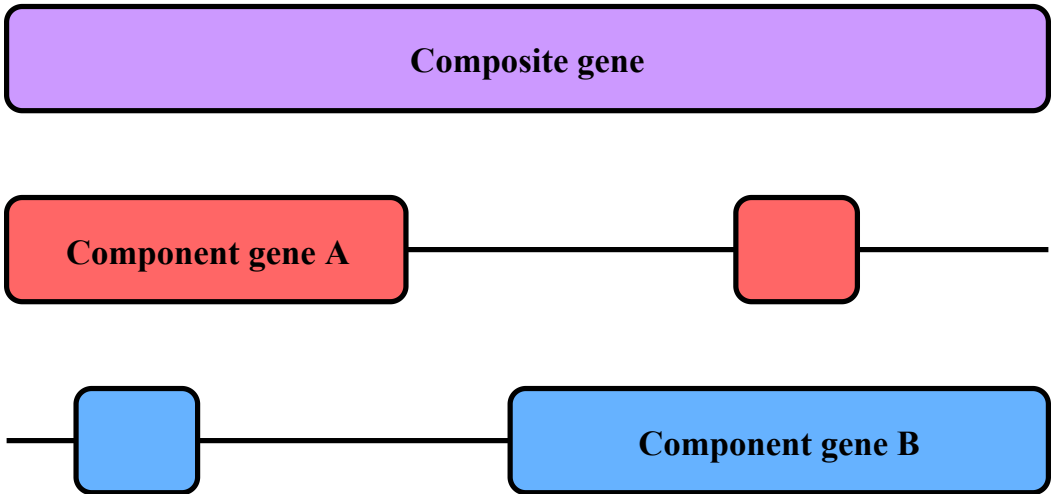

Figure 4.1.1. Multiple HSP processing by CompositeSearch

Representation of how CompositeSearch would report homologies between a composite and components with multiple HSPs. CompositeSearch selects the most significant HSP as the representative alignment and the lesser HSP is ignored. While this is indeed a case of gene remodelling, it is highly unlikely to be a fusion or a fission event. It is possible that this scenario illustrates a subfunctionalization event, however it would not meet the criteria to be classed as a fission in this state.

The following chapter describes its implementation and use on a number of preliminary test datasets.

4.2. Methodology

4.2.1. Gene fusion software development

To our knowledge, no high throughput software exists to parse large genomic datasets for the specific identification of definitive gene fused gene events using BLAST+ (Camacho *et al.*, 2009) and which allows inference from conserved domain architecture. To counter this conundrum we developed "compositeBLAST", a Python *v*3.6 script used to detect distinct fused genes based on sequence similarity, geographic coordinates between alignments, and PFAM domain architectures. The steps of compositeBLAST are described in the next sections.

4.2.1.1. Homolog detection

Homologs were detected using tabular (-outfmt 6) BLASTP (Camacho *et al.,* 2009) with an *e*-value stringency cut-off of $E \leq 1e^{-05}$. Tabular BLAST data was structured in the order required by CompositeSearch (qseqid, sseqid, evalue, pident, bitscore, qstart, qend, qlen, sstart, send, slen). While compositeBLAST can process a reciprocal BLAST file, we were interested in detecting composites where both components could be detected in at least one other species. For analyses such as this, two BLASTP analyses were required, one search between query genome A and subject genome B, and another between query genome B and subject genome A. We did not use SEG filters or limit the amount of target sequences. BLASTP output were loaded into compositeBLAST for further processing.

4.2.1.2. Processing of high scoring pairs

The BLASTP output file is loaded into compositeBLAST. Multiple high scoring pairs (HSPs) arise in cases where a query sequence aligns to more than one segment of a subject sequence. This may arise due to the presence of conserved sequence segments within both sequences. As these sequences could provide a basis for the erroneous report of a fused gene. This can be exemplified in cases where one HSP aligns to the N-terminus of a suspected fusion and another sequence to the C-terminus (Figure 4.1.1.). This would be problematic if no other distinct sequences aligned to the N-terminus thus allowing for a false positive report. To control these type I errors, we removed all alignments where multiple HSPs were detected. This was achieved by counting the amount of times a gene returned a hit for each other gene. Cases where the number of hits between a pair of genes exceeded one were removed. The specifics of this code snippet are explained in Figure 4.2.1.

4.2.1.3. Processing of single HSPs

To further control type I errors in composite reporting we required that a hit between genes was reciprocally reported, sharing a reciprocal HSP. For example, a case where sequence A (*seqA*) returned a hit for *seqB* and *seqB* returned a hit for *seqA* were considered to share an appropriate level of homology for further processing. Conversely, a case where a hit was returned between *seqA* and *seqC* but not between *seqC* and *seqA* was not further processed. The specifics of this code segment are explained in Figure 4.2.2.

```
'''
Remove multiple HSPs
'''

def remove_multiple_hsps(filepath):

    X = filepath
    single_hsps_ids = []
    single_hsps = []

    hsp_count = defaultdict(int)

    for line in filepath:
        term = line.strip().split('\t')
        seq1, seq2, evalue, pident, bitscore, qstart, qend,
        \qlen, sstart, send, slen, length = term[0], term[1],
        \term[2], term[3], term[4], term[5], term[6], term[7],
        \term[8], term[9], term[10], term[11]
        hsp_count[seq1, seq2]+= 1

    filepath = filepath.seek(0)
    already_checked = set()
    for (seq1, seq2), val in hsp_count.items():
        if val == 1:
            already_checked.add((seq1, seq2))

    already_checked = set(already_checked)
    for line in X:
        term = line.strip().split('\t')
        if (term[0],term[1]) in already_checked:
            single_hsps.append(term[0:])

    return single_hsps

Initial_hits = remove_multiple_hsps(input_blast_file)
Reciprocal_hits = remove_multiple_hsps(reciprocal_blast_file)
Single_hsps = Initial_hits + Reciprocal_hits
```

Figure 4.2.1. compositeBLAST multiple HSP removal algorithm

The first step involves opening lists and dictionaries for data storage. BLAST data is then processed through and HSPs between two genes are counted. If a gene pair has just one detected HSP, their associated alignment is extracted from the BLAST file for futher processing. Multiple HSPs are removed from both BLAST files. Full code provided at www.github.io/robleigh/compositeBLAST

```
'''
Identify reciprocal hits
'''

dict = {}

for item in Single_hsps:
    sequences = tuple(item[0:2])
    if sequences[::-1] in dict:
        dict[sequences[::-1]].append((item))
    else:
        dict[sequences] = [item]
for item in Single_hsps:
    rec_sequences = tuple(item[0:2])
    if rec_sequences[0]==rec_sequences[1]:
        x = rec_sequences[0], rec_sequences[1]
        dict[x].append((item))

pairs = {k: v for k, v in dict.items() if len(v) > 1}
```

Figure 4.2.2. compositeBLAST reciprocal HSP retention algorithm

The “sequences = tuple(item[0:2])” function extracts the qseqid and sseqids from Single_hsps (as a single item (“A,B”)). The first step of this snippet is responsible for extracting all “A-B” HSPs from identified single HSPs (“Single_hsps”). The second step retrieves “B-A” HSPs. The final step iterates through both lists and returns hits where “A-B” and “B-A” are observed. Full code provided at www.github.io/robleigh/compositeBLAST

4.2.1.4. Processing of potential components

Potential composites were identified using a "non-transitive triplet" method (Jachiet *et al*., 2012) where one component gene aligns to one region on the composite, and another gene aligns to a separate region. We aimed to detect composites only where both the N- and C-termini were detected amongst potential components. Cases where only the N- or the C-terminus could be detected amongst components may indicate that a fusion or fission event may have occurred, however it is also likely that an erroneous truncation event may also have occurred *via* a stop codon insertion or an error in sequencing or assembly (Des Marais and Rausher, 2008; Pathmanathan *et al.*, 2018). A legitimate model where only an N- or C- terminal component exists in addition to the 'fused' gene could arise through a subfunctionalization event. In such an event (as discussed in Chapter I), a multifunctional gene is duplicated to increase the rate of one function in a particular pathway (Rastogi and Liberles, 2005; Semon and Wolfe, 2008). Regions of the duplicate not involved in its specialised function would no longer be under selective pressure and may be truncated or experience a high level of mutation, thus appearing as a component. The detection of such cases was not our focus as they may produce considerable type I errors due to high levels of sequence similarity.

Erroneous composite reporting may arise due to the query and subject both possessing short homologous regions such as those required in some protein structural motifs. To control for these errors, the aligned region of a potential component ($A_{(c)}$ = (qstart – qend) + 1) was required to be greater than or equal to 30% of its sequence length. We used a 30% cut off as this is what was required for CompositeSearch. Potential components that did not meet this requirement were discarded from further processing.

4.2.1.5. Detection of potential composites by sequence similarity

To detect potential fusions, we treated each alignment as overlapping line segments in a cartesian coordinate system where query sequences were treated as potential fusions and the subject as a potential component. As we aimed to detect composites with distinct component alignments along both their N- and C- termini we bisected a potential composite and each section was considered the N-terminal section or the C-terminal section. To bisect the gene we defined each section based on the length of a potential composite gene, the first 50% of the sequence was the N-terminal section and the final 50% was the C-terminal section. We allowed for some section overlap by extending each section by 10% of the fusion length into the opposing section (Figure 4.3.3.) Therefore, a component was assigned to the N-terminal if its alignment terminated within the first 60% of the composite length ("N-terminal region"), and a component was assigned to a C-terminal if it initiated within the final 60% of the fusion length ("C-terminal region"). The central region was the overlap between the N- and C-terminal regions. As BLAST alignments can slightly overextend, an additional 20 amino acids was added to the N- and C-terminal limits to mirror the methodologies of other remodelling algorithms (Jachiet *et al.,* 2012; Pathmanathan *et al*., 2018). To avoid ambiguous terminal assignment, an N-terminal component was also required to initiate outside of the central region and C-terminal component was required to terminate outside the central region. A potential component with an alignment that either (*a*) crossed the entire central region or (*b*) initiated and terminated within the central region were discarded. Once all N-terminal and C-terminal components have been established, all combinations per potential fusion are generated. The snippet of code pertaining to this section is explained in Figure 4.3.4.

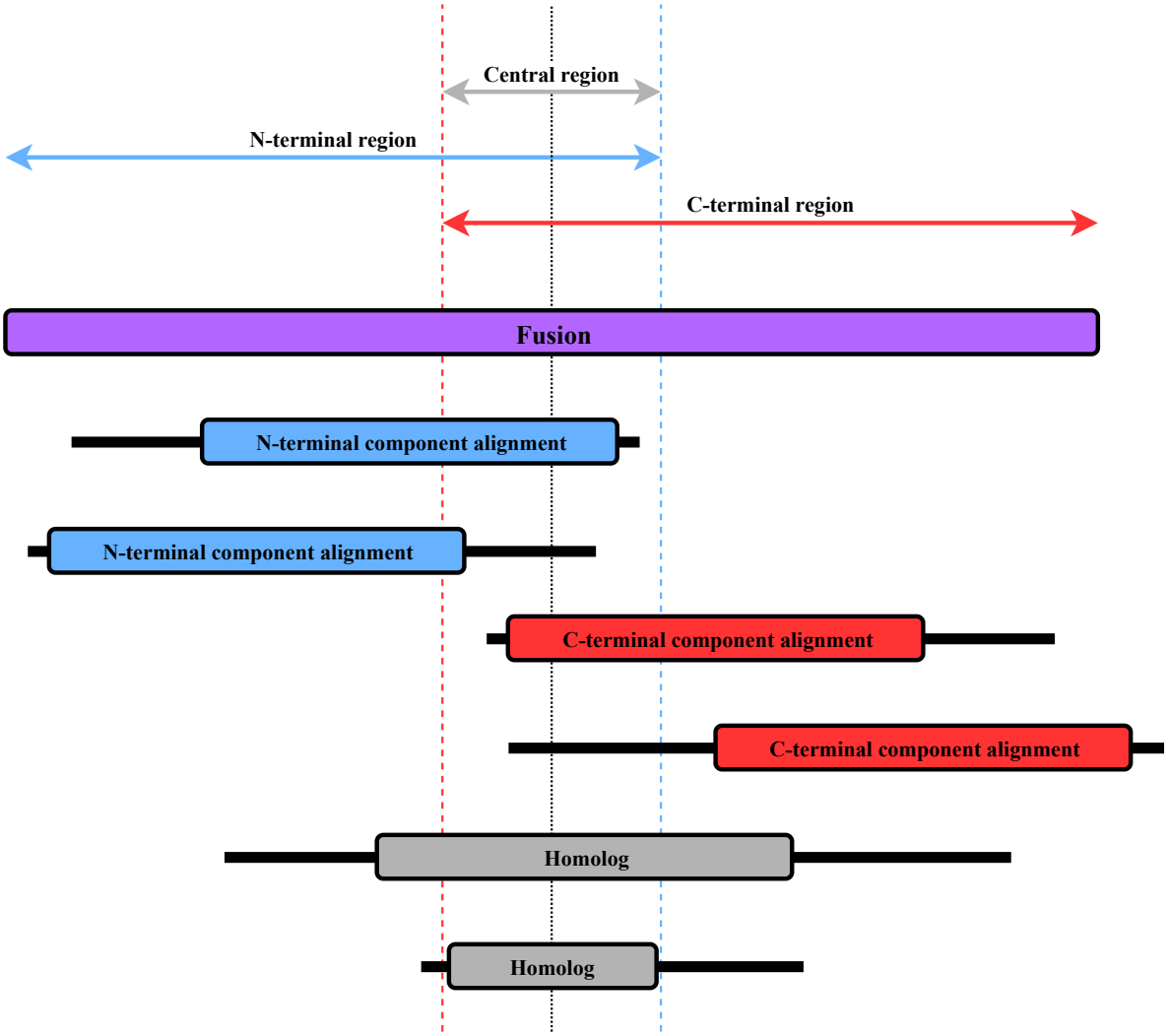

Figure 4.2.3. compositeBLAST gene region assignments

The dotted black line bisects the fusion gene through its midpoint. The blue and red arrows indicate the extent of the N-terminal and C-terminal regions, where red and blue arrows highlight the range of each region past the midpoint. The grey arrow indicates the central region. Thick black lines indicate unaligned regions on the component whereas boxes on genes indicate aligned regions. N-terminal alignments are annotated in blue and C-terminal alignments are aligned in red. The grey "homolog" genes are not identified as components because they either span the length of the central region or initiate and terminate within the central region.

```
'''
Arrange data into a uniform order
'''

arrangement = []
for key, values in pairs.items():
    pairs = values[0] + values[1]
    if int(pairs[7]) > int(pairs[10]):
        x = pairs[0], pairs[1], pairs[2], pairs[14], pairs[3], pairs[15], pairs[4], pairs[16], pairs[5], pairs[6], pairs[7],
        \pairs[8], pairs[9], pairs[10], pairs[17], pairs[18], pairs[19], pairs[20], pairs[21], pairs[22], pairs[23]
        arrangement.append(x)

'''
Quality control and assignment of potential components
'''

C_terminus = []
N_terminus = []

for item in arrangement:
    midpoint = float(item[10])/2
    range = float(item[10])/location_range
    Upper_N_terminus = (midpoint + range) + 20
    Lower_C_terminus = (midpoint - range) - 20
    Comp_align_cov = (int(item[12])-int(item[11])+1)/int(item[13])*100
    Rev_comp_align_cov = (int(item[15])-int(item[14])+1)/int(item[16])*100
    Min_cov = min(Comp_align_cov, Rev_comp_align_cov)
    if (max(float(item[2]), float(item[3])) <= evalue_stringency) and (min(float(item[4]), float(item[5])) >= pident_stringency) and
    \(Min_cov >= minimum_component_alignment_ratio):
        if not ((float(item[8]) >= Lower_C_terminus) and (float(item[9]) <= Upper_N_terminus)) or ((float(item[17]) >= Lower_C_terminus) and
        \(float(item[18]) <= Upper_N_terminus)):
            if (float(item[8]) < Lower_C_terminus) and (float(item[9]) < Upper_N_terminus):
                N_terminus.append(item)
            elif (float(item[8]) > Lower_C_terminus) and (float(item[9]) > Lower_C_terminus):
                C_terminus.append(item)


'''
Combination of potential components
'''

combinations = [(N + C) for C in C_terminus for N in N_terminus if C[0]==N[0] and C[1]!=N[1]]
```

Figure 4.2.4. compositeBLAST fusion detection algorithm

The first step arranges data into an appropriate format, grouping the statistics from both BLAST alignments into one line based on the longest gene in the pair. The second section divides the longest (query) sequence into N-terminal, C-terminal, and central regions and ensures that the alignment covers ≥ 30% of the putative component (subject) sequence. If an alignment initiates and terminates within the C-terminus, it is assigned to the C-terminus. Conversely, if an alignment initiates and terminates within the N-terminus, it is assigned to the N-terminus. If an alignment initiates and terminates within the central region or spans the central region it is excluded. The third step iterates through each potential composite gene and if it possesses a C-terminal component and an N-terminal component, it is considered to be a "putative composite". The combination of all putative *bona fide* C- and N- termini are computed for each putative composite. Full code provided at www.github.io/robleigh/compositeBLAST

4.2.1.6. Confirmation of potential composites by conserved protein architectures

InterProScan *v*5 (Jones *et al.*, 2014) was used to detect conserved protein architectures (PFAMs) for every composite and potential component. The combination of all possible N- and C-terminal components were arranged for each composite (triplet). The set of PFAM domains was compared for each gene in each triplet. A potential fusion was considered to be *bona fide* if (*a*) it shared at least one PFAM domain with each of its N-terminal component and C-terminal component and (*b*) the N-terminal components and C-terminal components did not share any PFAM domains. This ensured that a *bona fide* fusion inherited a conserved domain (and associated structural motif and function) from two separate lineages (Figure 4.2.5.).

4.2.1.7. Clustering composites into events

The sequences of all identified composites were searched against each other using BLASTP ($E \leq 1e^{-05}$). Genes were determined to be in the same family if they possess a mutual alignment overlap ≥ 80% and a pident ≥ 30%. Families were clustered and converted to an edge list using the "Graph.add_edge()" function of NetworkX (nx) python package (Aric *et al.,* 2007). The NetworkX edge list was further converted to a matrix using the "nx.to_scipy_sparse_matrix()" function. Finally, matrices were clustered using the MarkovClustering (mc) python package using the "mc.run_mcl()" and "mc.get_clusters()" function (Allard *et al.,* 2017). Each cluster was considered to be a remodelling event.

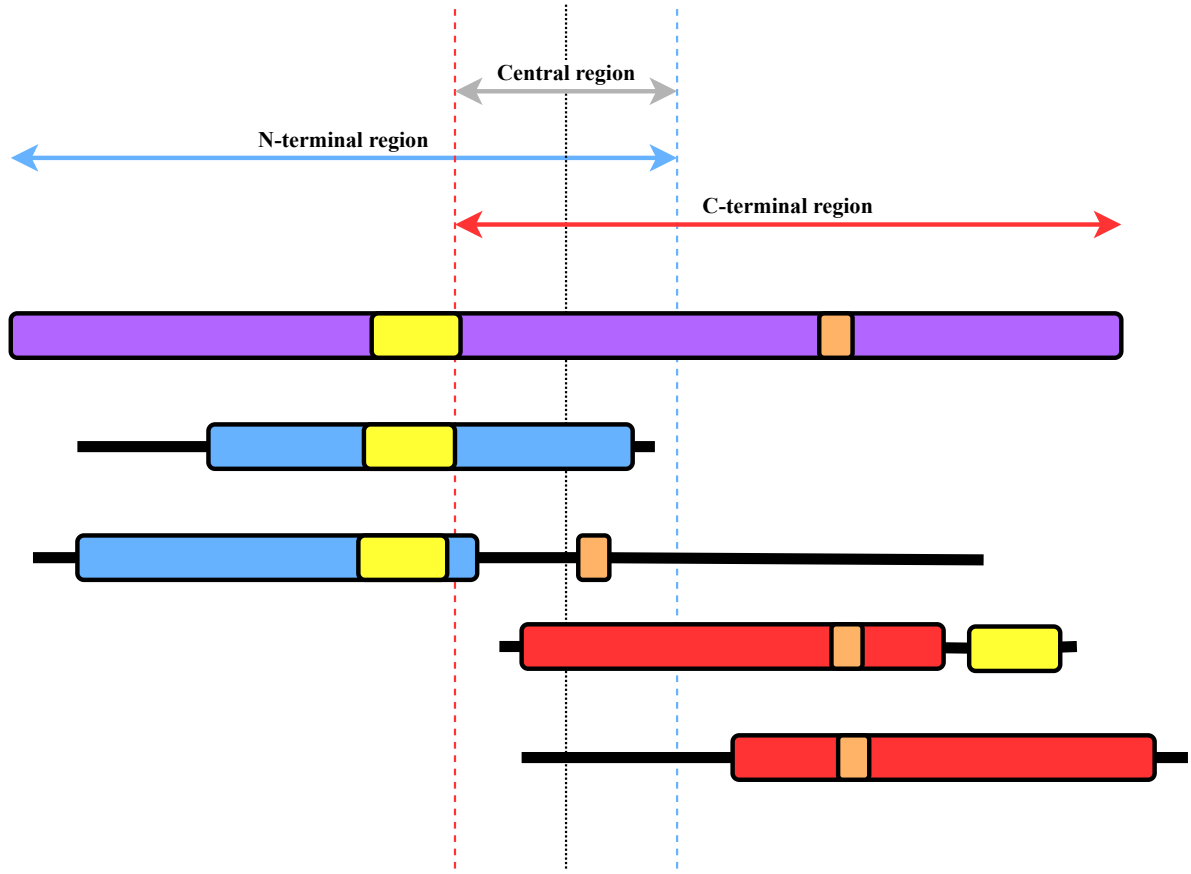

Fig 4.2.5. Comparison of alignments used to determine remodelling by compositeBLAST

Yellow and orange boxes highlight conserved (PFAM) domains on each sequence. In this scenario the only combination of components that is possible the top N-terminal gene and the bottom C-terminal gene. The N- and C-terminal components in the middle are excluded because they share a PFAM domain which is not permitted by compositeBLAST in an effort to maximise divergence between component genes.

4.2.2. Visualisation of composite gene alignments

We designed compositeViewer, a Python *v*3.6 program to view alignments between composites and components when detected by compositeBLAST. The length of the composite gene is drawn as a black line (0,*x*) and all other coordinates are drawn using (0,*x*) as a reference point. Alignments are drawn in colourful blocks on the composite gene, and directly below, the alignment is mapped for the component sequence. Once the alignments have been drawn for all sequences, the rest of the component genes are extended (as a black line) from the alignment using their own coordinates with reference to the alignment to initiate and terminate the extension (Figure 4.2.6.)

4.2.3. Benchmarking compositeBLAST on a fungal dataset

We benchmarked the functionality of compositeBLAST using a 9 fungal genome dataset as used by Leonard and Richards (2012). We used a BLASTP search ($E \leq 1e^{-10}$) to cross reference detected composite genes with the fusions identified by Leonard and Richards in 2012 (Table 4.2.1.). Leonard and Richards were not able to obtain any informational fused genes from two Microsporidia. We replicated this using the same species (*Encephalitozoon cuniculi* and *Antonospora locustae*), however as we were unable to get the accessions for the assemblies Leonard and Richards used, we substituted *Encephalitozoon cuniculi* GB-M1 *v*1.0 and *Antonospora locustae* HM-2013 *v*1.0 obtained from the Joint Genome Institute (https://genome.jgi.doe.gov). Again, we performed a pairwise search (BLASTP; $E \leq 1e^{-10}$) and performed a compositeBLAST search on the dataset using default settings. We used $E \leq 1e^{-10}$ to replicate the parameters used by *fdf*BLAST so results could be accurately compared. We also compared the speed of compositeBLAST and *fdf*BLAST. We used BLASTP (for

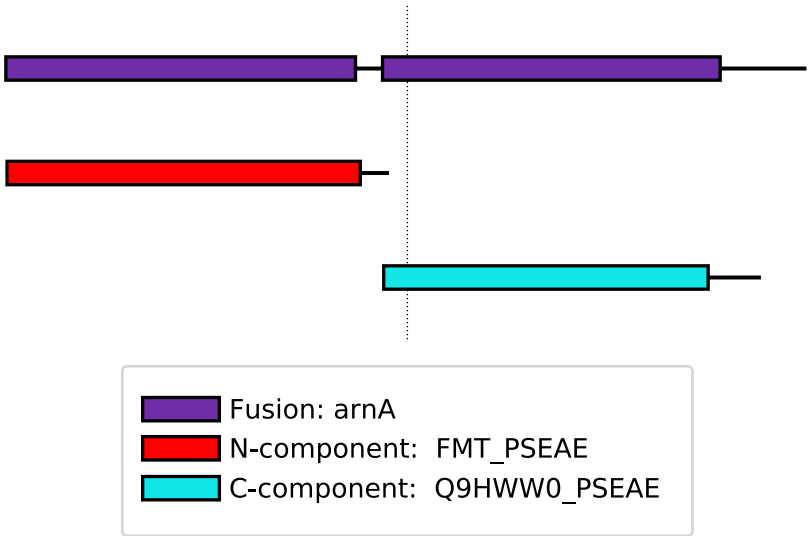

Figure 4.2.6. compositeViewer depiction of fusion gene *arnA*

Example of an alignment using the known fusion gene *arnA* (Williams *et al.*, 2005) illustrated using compositeViewer. The position of the alignments are perpendicular and the extensions of the rest of the component are extended based on the coordinates of their own alignments to the composite gene.

Table 4.2.1. Genomes initially used by Leonard and Richards (2012).

We also replicated their original study by including two Microsporidia species, *Encephalitozoon cuniculi* and *Antonospora locustae* (in bold)

Species	n_{genes}	Phylum
Neurospora crassa OR74A	9908	Ascomycota (Pezizomycotina)
Saccharomyces cerevisiae S288c	5885	Ascomycota (Saccharomycotina)
Schizosaccharomyces pombe 972h	5010	Ascomycota (Taphrinomycotina)
Ustilago maydis 521	6522	Basidiomycota (Ustillagomycotina)
Coprinopsis cinereal FGSC 9003	13394	Basidomycota (Agaricomycotina)
Allomyces macrogynus	17600	Blastoclaidiomycota
Batrachochytrium dendrobatidis JEL423	8732	Chytridiomycota
***Antonospora locustae* HM-2013**	2608	Microsporidia
***Encephalitozoon cuniculi* GB-M1**	1996	Microsporidia
Mucor circinelloides f. lusitanicus	10930	Mucoromucota
Rhizopus oryzae RA 99880	17459	Mucoromycota

compositeBLAST) and BLASTall (for *fdf*BLAST) to detect homologies between a two yeast dataset (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*).

4.2.4. Determination of the evolutionary rate of composite gene formation

In order to decipher the rate of composite generation in fungal and plant lineages, it was important to establish how many remodelling events were ancestral. For fungi, we searched all reported composite genes against a dataset of three genomes ancestral to our dataset. This dataset consisted of two Cryptomycota *Paramicrosporidium saccamoebae* str. KSL3 (GCA_002794465) and *Rozella allomycis* CSF55 (GCA_000442015), and a Holomycota, *Fonticula alba* (GCA_000388065), all of which were downloaded from Ensembl (www.ensembl.org).

We determined a remodelling event to be ancestral if any gene from a cluster returned a hit with pident ≥ 30% and mutual overlap of 80% using BLASTP ($E \leq 1e^{-05}$). Ancestral events cannot be used to determine a rate as it is not known when they first emerged so these clusters were excluded further rate calculations.

Dikarya were reported to have emerged between 392.1-1823 Ma (Betts *et al.,* 2018). We calculated an approximate evolutionary rate (τ) for fungi used in this thesis to be 636.21 Ma when κ = 1 (*subsection 2.2.3.3.2.2.*). The LP from the root of the fungal phylogeny was determined to be 0.88999 (*Pichia membranifaciens*). The divergence time for fungi was approximated to be 566.22 Ma by multiplying τ by LP. The composite birth rate (f_b) was calculated by dividing the sum of non-ancestral remodelling events by 566.22 (as per *subsection 2.2.3.4.*)

4.2.5. Detection of composite antimicrobial resistance genes using compositeBLAST

The complete set of antimicrobial resistance (AMR) genes were downloaded from the Comprehensive Antibiotic Resistance Database (CARD) (Jia *et al.*, 2017) searched against a dataset of 193 reference prokaryote genomes (115 bacteria and 78 archaea) downloaded from UniProtKB (Bateman *et al.*, 2017). A wide breadth of genomes were selected to increase phylogenomic diversity within the dataset (Table 4.2.2.) Model organisms were chosen for each phylum if available. All analyses were conducted using an *e*-value stringency cut off of ($E \leq 1e^{-05}$) and processed through compositeBLAST using default settings ($E \leq 1e^{-05}$; coverage = 30%). We implemented a second quality control step for these analyses so component genes could not display significant homology at $E \leq 1e^{-05}$ using BLASTP. We implemented this step to encourage the most robust composite events.

4.2.6. Assessment of antimicrobial resistance composite distribution

The presence and enumeration of each full length *bona fide* composite homolog in a given genome was determined using the same criteria as the clustering step of compositeBLAST and in the family detection algorithm of CompositeSearch (Pathmanathan *et al.*, 2018). A BLASTP ($E \leq 1e^{-05}$) search was performed between a fusion query and each gene in each genome. If a query returned a hit with (a) ≥ 80% mutual sequence length overlap and (b) a percentage identity (pident) score ≥ 30% it was considered to be present in the genome. For each composite, a dataset was constructed for all identified N-terminal components, and another for all C-terminal components. The presence and numeration for each component type (N- or C-) was also determined using this method. The presence and absence of each fusion and its components in a given genome was compared to each other genome in

Table 4.2.2. Prokaryote genomes used for the detection of composite antimicrobial resistance genes.

All genomes were downloaded from UniProtKB with all associated genomic and assembly statistics provided.

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000245584	*Heimdallarchaeota* archaeon	3521	HEIAB	Archaea, Asgard group, Candidatus Heimdallarchaeota	GCA_003144275.1
UP000185649	*Lokiarchaeota* archaeon	4413	LOKAC	Archaea, Asgard group, Candidatus Lokiarchaeota	GCA_001940655.1
UP000001686	*Korarchaeum cryptofilum*	1602	KORCO	Archaea, Candidatus Korarchaeota, Candidatus Korarchaeum	GCA_000019605.1
UP000000346	*Acidilobus saccharovorans*	1499	ACIS3	Archaea, Crenarchaeota, Thermoprotei	GCA_000144915.1
UP000010469	*Caldisphaera lagunensis*	1477	CALLD	Archaea, Crenarchaeota, Thermoprotei	GCA_000317795.1
UP000002518	*Aeropyrum pernix*	1700	AERPE	Archaea, Crenarchaeota, Thermoprotei	GCA_000011125.1
UP000006903	*Desulfurococcus amylolyticus*	1470	DESA1	Archaea, Crenarchaeota, Thermoprotei	GCA_000020905.1
UP000000262	*Ignicoccus hospitalis*	1434	IGNH4	Archaea, Crenarchaeota, Thermoprotei	GCA_000017945.1
UP000001304	*Ignisphaera aggregans*	1929	IGNAA	Archaea, Crenarchaeota, Thermoprotei	GCA_000145985.1
UP000000254	*Staphylothermus marinus*	1570	STAMF	Archaea, Crenarchaeota, Thermoprotei	GCA_000015945.1
UP000005270	*Thermogladius calderae*	1414	THEC1	Archaea, Crenarchaeota, Thermoprotei	GCA_000264495.1
UP000002376	*Thermosphaera aggregans*	1387	THEAM	Archaea, Crenarchaeota, Thermoprotei	GCA_000092185.1
UP000002593	*Hyperthermus butylicus*	1602	HYPBU	Archaea, Crenarchaeota, Thermoprotei	GCA_000015145.1
UP000001037	*Pyrolobus fumarii*	1967	PYRF1	Archaea, Crenarchaeota, Thermoprotei	GCA_000223395.1
UP000007391	*Fervidicoccus fontis*	1384	FERFK	Archaea, Crenarchaeota, Thermoprotei	GCA_000258425.1
UP000008458	*Acidianus hospitalis*	2329	ACIHW	Archaea, Crenarchaeota, Thermoprotei	GCA_000213215.1
UP000000242	*Metallosphaera sedula*	2256	METS5	Archaea, Crenarchaeota, Thermoprotei	GCA_000016605.1
UP000001974	*Saccharolobus solfataricus*	2938	SACS2	Archaea, Crenarchaeota, Thermoprotei	GCA_000007005.1
UP000001018	*Sulfolobus acidocaldarius*	2221	SULAC	Archaea, Crenarchaeota, Thermoprotei	GCA_000012285.1
UP000001015	*Sulfurisphaera tokodaii*	2805	SULTO	Archaea, Crenarchaeota, Thermoprotei	GCA_000011205.1
UP000000641	*Thermofilum pendens*	1876	THEPD	Archaea, Crenarchaeota, Thermoprotei	GCA_000015225.1
UP000001137	*Caldivirga maquilingensis*	1962	CALMQ	Archaea, Crenarchaeota, Thermoprotei	GCA_000018305.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000002439	*Pyrobaculum aerophilum*	2590	PYRAE	Archaea, Crenarchaeota, Thermoprotei	GCA_000007225.1
UP000002654	*Thermoproteus tenax*	2047	THETK	Archaea, Crenarchaeota, Thermoprotei	GCA_000253055.1
UP000006681	*Vulcanisaeta distributa*	2493	VULDI	Archaea, Crenarchaeota, Thermoprotei	GCA_000148385.1
UP000002199	*Archaeoglobus fulgidus*	2399	ARCFU	Archaea, Euryarchaeota, Archaeoglobi	GCA_000008665.1
UP000002613	*Ferroglobus placidus*	2463	FERPA	Archaea, Euryarchaeota, Archaeoglobi	GCA_000025505.1
UP000001400	*Aciduliprofundum boonei*	1539	ACIB4	Archaea, Euryarchaeota, Diaforarchaea group	GCA_000025665.1
UP000012672	*Methanomethylophilus alvus*	1643	METAX	Archaea, Euryarchaeota, Diaforarchaea group	GCA_000300255.2
UP000014070	*Methanomassiliicoccus intestinalis*	1826	METII	Archaea, Euryarchaeota, Diaforarchaea group	GCA_000404225.1
UP000000438	*Picrophilus torridus*	1535	PICTO	Archaea, Euryarchaeota, Diaforarchaea group	GCA_000008265.1
UP000001024	*Thermoplasma acidophilum*	1482	THEAC	Archaea, Euryarchaeota, Diaforarchaea group	GCA_000195915.1
UP000007490	*Methanobacterium lacus*	2493	METLA	Archaea, Euryarchaeota, Methanomada group	GCA_000191585.1
UP000008680	*Methanobrevibacter ruminantium*	2209	METRM	Archaea, Euryarchaeota, Methanomada group	GCA_000024185.1
UP000001931	*Methanosphaera stadtmanae*	1533	METST	Archaea, Euryarchaeota, Methanomada group	GCA_000012545.1
UP000005223	*Methanothermobacter thermautotrophicus*	1868	METTH	Archaea, Euryarchaeota, Methanomada group	GCA_000008645.1
UP000002315	*Methanothermus fervidus*	1283	METFV	Archaea, Euryarchaeota, Methanomada group	GCA_000166095.1
UP000000805	*Methanocaldococcus jannaschii*	1787	METJA	Archaea, Euryarchaeota, Methanomada group	GCA_000091665.1
UP000002061	*Methanocaldococcus infernus*	1439	METIM	Archaea, Euryarchaeota, Methanomada group	GCA_000092305.1
UP000009227	*Methanotorris igneus*	1753	METIK	Archaea, Euryarchaeota, Methanomada group	GCA_000214415.1
UP000001106	*Methanococcus aeolicus*	1490	META3	Archaea, Euryarchaeota, Methanomada group	GCA_000017185.1
UP000001826	*Methanopyrus kandleri*	1687	METKA	Archaea, Euryarchaeota, Methanopyri	GCA_000007185.1
UP000001169	*Haloarcula marismortui*	4234	HALMA	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000011085.1
UP000001746	*Halomicrobium mukohataei*	3343	HALMD	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000023965.1
UP000011867	*Natronomonas moolapensis*	2723	NATM8	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000591055.1
UP000002698	*Natronomonas pharaonis*	2764	NATPD	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000026045.1
UP000000390	*Halalkalicoccus jeotgali*	3779	HALJB	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000196895.1
UP000000554	*Halobacterium salinarum*	2426	HALSA	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000006805.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000006469	*Haloferax mediterranei*	3826	HALMT	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000306765.2
UP000006663	*Halogeometricum borinquense*	3894	HALBP	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000172995.2
UP000001975	*Haloquadratum walsbyi*	2558	HALWD	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000009185.1
UP000006794	*Halopiger xanaduensis*	4221	HALXS	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000217715.1
UP000001903	*Haloterrigena turkmenica*	5113	HALTV	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000025325.1
UP000010846	*Halovivax ruber*	3099	HALRX	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000328525.1
UP000001879	*Natrialba magadii*	4203	NATMM	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000025625.1
UP000010843	*Natrinema pellirubrum*	4138	NATP1	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000230735.3
UP000010468	*Natronobacterium gregoryi*	3624	NATGS	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000230715.3
UP000000663	*Methanocella arvoryzae*	3071	METAR	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000063445.1
UP000000365	*Methanocorpusculum labreanum*	1739	METLZ	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000015765.1
UP000009007	*Methanoculleus bourgensis*	2575	METBM	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000304355.2
UP000006565	*Methanolacinia petrolearia*	2779	METP4	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000147875.1
UP000002408	*Methanoregula boonei*	2450	METB6	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000017625.1
UP000002457	*Methanosphaerula palustris*	2655	METPE	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000021965.1
UP000001941	*Methanospirillum hungatei*	3087	METHJ	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000013445.1
UP000005877	*Methanosaeta harundinacea*	2358	METH6	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000235565.1
UP000007807	*Methanothrix soehngenii*	2791	METSG	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000204415.1
UP000001979	*Methanococcoides burtonii*	2242	METBU	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000013725.1
UP000000391	*Methanohalobium evestigatum*	2250	METEZ	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000196655.1
UP000001059	*Methanohalophilus mahii*	1986	METMS	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000025865.1
UP000010866	*Methanomethylovorans hollandica*	2551	METHD	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000328665.1
UP000006622	*Methanosalsum zhilinae*	1972	METZD	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000217995.1
UP000002487	*Methanosarcina acetivorans*	4468	METAC	Archaea, Euryarchaeota, Stenosarchaea group	GCA_000007345.1
UP000001013	*Pyrococcus furiosus*	2045	PYRFU	Archaea, Euryarchaeota, Thermococci	GCA_000007305.1
UP000000536	*Thermococcus kodakarensis*	2301	THEKO	Archaea, Euryarchaeota, Thermococci	GCA_000009965.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000000578	*Nanoarchaeum equitans*	536	NANEQ	Archaea, Nanoarchaeota, Nanoarchaeales	GCA_000008085.1
UP000000758	*Cenarchaeum symbiosum*	2022	CENSY	Archaea, Thaumarchaeota, Cenarchaeales	GCA_000200715.1
UP000000792	*Nitrosopumilus maritimus*	1795	NITMS	Archaea, Thaumarchaeota, Nitrosopumilales	GCA_000018465.1
UP000008037	*Nitrososphaera gargensis*	3523	NITGG	Archaea, Thaumarchaeota, Nitrososphaeria	GCA_000303155.1
UP000002207	*Acidobacterium capsulatum*	3363	ACIC5	Bacteria, Acidobacteria, Acidobacteriales	GCA_000022565.1
UP000007113	*Granulicella mallensis*	4804	GRAMM	Bacteria, Acidobacteria, Acidobacteriales	GCA_000178955.2
UP000000343	*Granulicella tundricola*	4514	GRATM	Bacteria, Acidobacteria, Acidobacteriales	GCA_000178975.2
UP000006056	*Terriglobus roseus*	3936	TERRK	Bacteria, Acidobacteria, Acidobacteriales	GCA_000265425.1
UP000006640	*Thermobispora bispora*	3545	THEBD	Bacteria, Actinobacteria, Actinobacteria incertae sedis	GCA_000092645.1
UP000001584	*Mycobacterium tuberculosis*	3993	MYCTU	Bacteria, Actinobacteria, Corynebacteriales	GCA_000195955.2
UP000000738	*Micrococcus luteus*	2207	MICLC	Bacteria, Actinobacteria, Micrococcales	GCA_000023205.1
UP000001973	*Streptomyces coelicolor*	8038	STRCO	Bacteria, Actinobacteria, Streptomycetales	GCA_000203835.1
UP000000798	*Aquifex aeolicus*	1553	AQUAE	Bacteria, Aquificae, Aquificales	GCA_000008625.1
UP000002574	*Hydrogenobacter thermophilus*	1892	HYDTT	Bacteria, Aquificae, Aquificales	GCA_000010785.1
UP000002043	*Thermocrinis albus*	1592	THEAH	Bacteria, Aquificae, Aquificales	GCA_000025605.1
UP000007102	*Desulfurobacterium thermolithotrophum*	1496	DESTD	Bacteria, Aquificae, Desulfurobacteriales	GCA_000191045.1
UP000014227	*Chthonomonas calidirosea*	2809	CHTCT	Bacteria, Armatimonadetes, Chthonomonadetes	GCA_000427095.1
UP000008674	*Salinibacter ruber*	2812	SALRD	Bacteria, Bacteroidetes, Bacteroidetes Order II. Incertae sedis	GCA_000013045.1
UP000001414	*Bacteroides thetaiotaomicron*	4782	BACTN	Bacteria, Bacteroidetes, Bacteroidia	GCA_000011065.1
UP000000723	*Azobacteroides pseudotrichonymphae*	847	AZOPC	Bacteria, Bacteroidetes, Bacteroidia	GCA_000010645.1
UP000006394	*Flavobacterium psychrophilum*	2421	FLAPJ	Bacteria, Bacteroidetes, Flavobacteriia	GCA_000064305.2
UP000002019	*Cloacimonas acidaminovorans*	1813	CLOAI	Bacteria, Candidatus Cloacimonetes, Candidatus Cloacimonas	GCA_000146065.1
UP000000431	*Chlamydia trachomatis*	895	CHLTR	Bacteria, Chlamydiae, Chlamydiales	GCA_000008725.1
UP000000529	*Protochlamydia amoebophila*	1879	PARUW	Bacteria, Chlamydiae, Parachlamydiales	GCA_000011565.1
UP000000495	*Parachlamydia acanthamoebae*	2784	PARAV	Bacteria, Chlamydiae, Parachlamydiales	GCA_000253035.1
UP000001505	*Waddlia chondrophila*	1919	WADCW	Bacteria, Chlamydiae, Parachlamydiales	GCA_000092785.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000001007	*Chlorobaculum tepidum*	2250	CHLTE	Bacteria, Chlorobi, Chlorobia	GCA_000006985.1
UP000007880	*Caldilinea aerophila*	4097	CALAS	Bacteria, Chloroflexi, Caldilineae	GCA_000281175.1
UP000002008	*Chloroflexus aurantiacus*	3850	CHLAA	Bacteria, Chloroflexi, Chloroflexia	GCA_000018865.1
UP000008289	*Dehalococcoides mccartyi*	1502	DEHM1	Bacteria, Chloroflexi, Dehalococcoidia	GCA_000011905.1
UP000002349	*Dehalogenimonas lykanthroporepellens*	1619	DEHLB	Bacteria, Chloroflexi, Dehalococcoidia	GCA_000143165.1
UP000002027	*Sphaerobacter thermophilus*	3471	SPHTD	Bacteria, Chloroflexi, Sphaerobacteridae	GCA_000024985.1
UP000002572	*Desulfurispirillum indicum*	2551	DESIS	Bacteria, Chrysiogenetes, Chrysiogenales	GCA_000177635.2
UP000001732	*Coprothermobacter proteolyticus*	1481	COPPD	Bacteria, Coprothermobacterota, Coprothermobacteria	GCA_000020945.1
UP000000557	*Gloeobacter violaceus*	4406	GLOVI	Bacteria, Cyanobacteria, Gloeobacteria	GCA_000011385.1
UP000092382	*Aphanizomenon flos-aquae*	3783	APHFL	Bacteria, Cyanobacteria, Nostocales	GCA_001672165.1
UP000010480	*Cyanobacterium aponinum*	3415	CYAAP	Bacteria, Cyanobacteria, Oscillatoriophycideae	GCA_000317675.1
UP000001425	*Synechocystis sp.*	3507	SYNY3	Bacteria, Cyanobacteria, Synechococcales	GCA_000009725.1
UP000001420	*Prochlorococcus marinus*	1881	PROMA	Bacteria, Cyanobacteria, Synechococcales	GCA_000007925.1
UP000000440	*Thermosynechococcus elongatus*	2451	THEEB	Bacteria, Cyanobacteria, Synechococcales	GCA_000011345.1
UP000007039	*Calditerrivibrio nitroreducens*	2089	CALNY	Bacteria, Deferribacteres, Deferribacterales	GCA_000183405.1
UP000001520	*Deferribacter desulfuricans*	2338	DEFDS	Bacteria, Deferribacteres, Deferribacterales	GCA_000010985.1
UP000002012	*Denitrovibrio acetiphilus*	2901	DENA2	Bacteria, Deferribacteres, Deferribacterales	GCA_000025725.1
UP000002524	*Deinococcus radiodurans*	3085	DEIRA	Bacteria, Deinococcus-Thermus, Deinococci	GCA_000008565.1
UP000007030	*Marinithermus hydrothermalis*	2194	MARHT	Bacteria, Deinococcus-Thermus, Deinococci	GCA_000195335.1
UP000001916	*Meiothermus silvanus*	3383	MEISD	Bacteria, Deinococcus-Thermus, Deinococci	GCA_000092125.1
UP000008722	*Oceanithermus profundus*	2372	OCEP5	Bacteria, Deinococcus-Thermus, Deinococci	GCA_000183745.1
UP000000532	*Thermus thermophilus*	2227	THET8	Bacteria, Deinococcus-Thermus, Deinococci	GCA_000091545.1
UP000007719	*Dictyoglomus turgidum*	1743	DICTD	Bacteria, Dictyoglomi, Dictyoglomales	GCA_000021645.1
UP000001029	*Elusimicrobium minutum*	1528	ELUMP	Bacteria, Elusimicrobia, Elusimicrobia	GCA_000020145.1
UP000000517	*Fibrobacter succinogenes*	2871	FIBSS	Bacteria, Fibrobacteres, Fibrobacterales	GCA_000146505.1
UP000001570	*Bacillus subtilis*	4260	BACSU	Bacteria, Firmicutes, Bacilli	GCA_000009045.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000001172	*Geobacillus kaustophilus*	3516	GEOKA	Bacteria, Firmicutes, Bacilli	GCA_000009785.1
UP000007397	*Halobacillus halophilus*	4100	HALH3	Bacteria, Firmicutes, Bacilli	GCA_000284515.1
UP000000817	*Listeria monocytogenes*	2844	LISMO	Bacteria, Firmicutes, Bacilli	GCA_000196035.1
UP000008816	*Staphylococcus aureus*	2889	STAA8	Bacteria, Firmicutes, Bacilli	GCA_000013425.1
UP000001415	*Enterococcus faecalis*	3240	ENTFA	Bacteria, Firmicutes, Bacilli	GCA_000007785.1
UP000000586	*Streptococcus pneumoniae*	2030	STRR6	Bacteria, Firmicutes, Bacilli	GCA_000007045.1
UP000001986	*Clostridium botulinum*	3590	CLOBH	Bacteria, Firmicutes, Clostridia	GCA_000063585.1
UP000007053	*Moorella thermoacetica*	2451	MOOTA	Bacteria, Firmicutes, Clostridia	GCA_000013105.1
UP000002521	*Fusobacterium nucleatum*	2046	FUSNN	Bacteria, Fusobacteria, Fusobacteriales	GCA_000007325.1
UP000006875	*Ilyobacter polytropus*	2859	ILYPC	Bacteria, Fusobacteria, Fusobacteriales	GCA_000165505.1
UP000001910	*Leptotrichia buccalis*	2218	LEPBD	Bacteria, Fusobacteria, Fusobacteriales	GCA_000023905.1
UP000000845	*Sebaldella termitidis*	4124	SEBTE	Bacteria, Fusobacteria, Fusobacteriales	GCA_000024405.1
UP000002072	*Streptobacillus moniliformis*	1431	STRM9	Bacteria, Fusobacteria, Fusobacteriales	GCA_000024565.1
UP000002209	*Gemmatimonas aurantiaca*	3932	GEMAT	Bacteria, Gemmatimonadetes, Gemmatimonadales	GCA_000010305.1
UP000007382	*Leptospirillum ferrooxidans*	2413	LEPFC	Bacteria, Nitrospirae, Nitrospirales	GCA_000284315.1
UP000069205	*Nitrospira moscoviensis*	4733	NITMO	Bacteria, Nitrospirae, Nitrospirales	GCA_001273775.1
UP000000718	*Thermodesulfovibrio yellowstonii*	1982	THEYD	Bacteria, Nitrospirae, Nitrospirales	GCA_000020985.1
UP000008631	*Isosphaera pallida*	3721	ISOPI	Bacteria, Planctomycetes, Planctomycetia	GCA_000186345.1
UP000001887	*Pirellula staleyi*	4711	PIRSD	Bacteria, Planctomycetes, Planctomycetia	GCA_000025185.1
UP000002220	*Planctopirus limnophila*	4258	PLAL2	Bacteria, Planctomycetes, Planctomycetia	GCA_000092105.1
UP000001025	*Rhodopirellula baltica*	7271	RHOBA	Bacteria, Planctomycetes, Planctomycetia	GCA_000196115.1
UP000006860	*Rubinisphaera brasiliensis*	4710	RUBBR	Bacteria, Planctomycetes, Planctomycetia	GCA_000165715.3
UP000001364	*Caulobacter vibrioides*	3859	CAUVN	Bacteria, Proteobacteria, Alphaproteobacteria	GCA_000022005.1
UP000002526	*Bradyrhizobium diazoefficiens*	8253	BRADU	Bacteria, Proteobacteria, Alphaproteobacteria	GCA_000011365.1
UP000008320	*Ehrlichia chaffeensis*	1100	EHRCR	Bacteria, Proteobacteria, Alphaproteobacteria	GCA_000013145.1
UP000002480	*Rickettsia prowazekii*	834	RICPR	Bacteria, Proteobacteria, Alphaproteobacteria	GCA_000195735.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000002676	*Bordetella pertussis*	3258	BORPE	Bacteria, Proteobacteria, Betaproteobacteria	GCA_000195715.1
UP000008815	*Burkholderia multivorans*	6040	BURM1	Bacteria, Proteobacteria, Betaproteobacteria	GCA_000010545.1
UP000000425	*Neisseria meningitidis*	2001	NEIMB	Bacteria, Proteobacteria, Betaproteobacteria	GCA_000008805.1
UP000008291	*Thiobacillus denitrificans*	2826	THIDA	Bacteria, Proteobacteria, Betaproteobacteria	GCA_000012745.1
UP000002191	*Pseudodesulfovibrio aespoeensis*	3269	PSEA9	Bacteria, Proteobacteria, Deltaproteobacteria	GCA_000176915.2
UP000002430	*Lawsonia intracellularis*	1342	LAWIP	Bacteria, Proteobacteria, Deltaproteobacteria	GCA_000055945.1
UP000000577	*Geobacter sulfurreducens*	3402	GEOSL	Bacteria, Proteobacteria, Deltaproteobacteria	GCA_000007985.2
UP000001784	*Syntrophobacter fumaroxidans*	4012	SYNFM	Bacteria, Proteobacteria, Deltaproteobacteria	GCA_000014965.1
UP000000799	*Campylobacter jejuni*	1623	CAMJE	Bacteria, Proteobacteria, Epsilonproteobacteria	GCA_000009085.1
UP000000429	*Helicobacter pylori*	1553	HELPY	Bacteria, Proteobacteria, Epsilonproteobacteria	GCA_000008525.1
UP000000422	*Wolinella succinogenes*	2028	WOLSU	Bacteria, Proteobacteria, Epsilonproteobacteria	GCA_000196135.1
UP000000625	*Escherichia coli*	4446	ECOLI	Bacteria, Proteobacteria, Gammaproteobacteria	GCA_000005845.2
UP000001014	*Salmonella typhimurium*	4533	SALTY	Bacteria, Proteobacteria, Gammaproteobacteria	GCA_000006945.2
UP000002716	*Shigella dysenteriae*	3897	SHIDS	Bacteria, Proteobacteria, Gammaproteobacteria	GCA_000012005.1
UP000000815	*Yersinia pestis*	3909	YERPE	Bacteria, Proteobacteria, Gammaproteobacteria	GCA_000009065.1
UP000051497	*Candidatus Berkiella*	3170	CANBE	Bacteria, Proteobacteria, Gammaproteobacteria	GCA_001431295.1
UP000002438	*Pseudomonas aeruginosa*	5564	PSEAE	Bacteria, Proteobacteria, Gammaproteobacteria	GCA_000006765.1
UP000234191	*Brachyspira hyodysenteriae*	2617	BRAHO	Bacteria, Spirochaetes, Brachyspirales	GCA_002850235.1
UP000001408	*Leptospira interrogans*	3676	LEPIN	Bacteria, Spirochaetes, Leptospirales	GCA_000092565.1
UP000006048	*Turneriella parva*	4092	TURPD	Bacteria, Spirochaetes, Leptospirales	GCA_000266885.1
UP000001807	*Borrelia burgdorferi*	1290	BORBU	Bacteria, Spirochaetes, Spirochaetales	GCA_000008685.2
UP000002318	*Sediminispirochaeta smaragdinae*	4211	SEDSS	Bacteria, Spirochaetes, Spirochaetales	GCA_000143985.1
UP000007254	*Spirochaeta thermophila*	2249	SPITZ	Bacteria, Spirochaetes, Spirochaetales	GCA_000184345.2
UP000000811	*Treponema pallidum*	1027	TREPA	Bacteria, Spirochaetes, Spirochaetales	GCA_000008605.1
UP000006061	*Acetomicrobium mobile*	2004	ACEMN	Bacteria, Synergistetes, Synergistia	GCA_000266925.1
UP000002366	*Aminobacterium colombiense*	1872	AMICL	Bacteria, Synergistetes, Synergistia	GCA_000025885.1

Proteome ID	Organism	Gene count	Taxon mnemonic	Taxonomic lineage	Genome assembly ID
UP000002030	*Thermanaerovibrio acidaminovorans*	1737	THEAS	Bacteria, Synergistetes, Synergistia	GCA_000024905.1
UP000005868	*Thermovirga lienii*	1853	THELD	Bacteria, Synergistetes, Synergistia	GCA_000233775.1
UP000008558	*Acholeplasma laidlawii*	1380	ACHLI	Bacteria, Tenericutes, Mollicutes	GCA_000018785.1
UP000002523	Onion yellows *Phytoplasma*	730	ONYPE	Bacteria, Tenericutes, Mollicutes	GCA_000009845.1
UP000006647	*Mesoplasma florum*	683	MESFL	Bacteria, Tenericutes, Mollicutes	GCA_000008305.1
UP000000807	*Mycoplasma genitalium*	483	MYCGE	Bacteria, Tenericutes, Mollicutes	GCA_000027325.1
UP000006793	*Thermodesulfatator indicus*	2184	THEID	Bacteria, Thermodesulfobacteria, Thermodesulfobacteriales	GCA_000217795.1
UP000006583	*Thermodesulfobacterium geofontis*	1594	THEGP	Bacteria, Thermodesulfobacteria, Thermodesulfobacteriales	GCA_000215975.1
UP000002382	*Kosmotoga olearia*	2087	KOSOT	Bacteria, Thermotogae, Kosmotogales	GCA_000023325.1
UP000007161	*Marinitoga piezophila*	2044	MARPK	Bacteria, Thermotogae, Petrotogales	GCA_000255135.1
UP000002415	*Fervidobacterium nodosum*	1725	FERNB	Bacteria, Thermotogae, Thermotogales	GCA_000017545.1
UP000002016	*Pseudothermotoga lettingae*	2040	PSELT	Bacteria, Thermotogae, Thermotogales	GCA_000017865.1
UP000008183	*Thermotoga maritima*	1852	THEMA	Bacteria, Thermotogae, Thermotogales	GCA_000008545.1
UP000009149	*Methylacidiphilum infernorum*	2470	METI4	Bacteria, Verrucomicrobia, Methylacidiphilae	GCA_000019665.1
UP000007013	*Opitutus terrae*	4588	OPITP	Bacteria, Verrucomicrobia, Opitutae	GCA_000019965.1
UP000000925	*Coraliomargarita akajimensis*	3110	CORAD	Bacteria, Verrucomicrobia, Opitutae	GCA_000025905.1
UP000001031	*Akkermansia muciniphila*	2137	AKKM8	Bacteria, Verrucomicrobia, Verrucomicrobiae	GCA_000020225.1

the dataset to determine if a phyletic distribution bias could be observed. Detected fusions were discarded if a phylogenetic pattern could not be observed within N- or C-terminal full length homologs.

4.3. Results

4.3.1. Benchmarking compositeBLAST

The 9 fungal genome dataset used during the debut analysis of *fdf*BLAST (Leonard and Richards, 2012) was used to detect the efficacy of compositeBLAST. We first established if we could determine we could detect the 63 fusions detected by *fdf*BLAST. Our first observation was that only 59 fusions were present in the output files provided by Leonard and Richards. Of these 59 fusion genes, we were not able to detect 12 (Table 4.3.1.). Of these 12 genes, 10 were recoverable by either removing the criteria where components could not span the terminal region or by reducing the minimum amount of coverage required for a component gene. The remaining two components were not recoverable as they had only an N-terminal component in both Leonard and Richards results and in our initial BLAST results. As compositeBLAST requires at least two distinct components, these genes were not recoverable by any means. Differential fusions are observed when a composite is observed in one genome and both components (but not the composite) are observed in a different genome.

4.3.2. Extent of composite genes and fungi

In 9 fungi, a total of 573 composite genes within 300 families were detected using compositeBLAST, of which 219 were singleton genes. A total of 354 genes were dispersed

Table 4.3.1. Fusions identified by Leonard and Richards (2012) not observed during the benchmarking of compositeBLAST.

Of the 12 non-identifiable fusions, 10 could be recovered by removing the requirement that a component does not span the central region of the fusion. The remaining two fusions (highlighted in red) did not have an appropriate C-terminal component in our analyses or in Leonard and Richards (2012) results. As compositeBLAST requires an N- and a C-terminal component these two fusions could not be recovered under any circumstance.

Fusion ID	Reason for non-detection
fusion_10_XP_011394556	Component alignment coverage too low
fusion_21_NP_594836	Alignment spans fusion central region
Fusion_34_XP_001402280	**No associated C-terminal component**
fusion_35_XP_964702	Alignment spans fusion central region
fusion_42_XP_001934738	**No associated C-terminal component**
fusion_43_XP_003303532	Alignment spans fusion central region
fusion_47_XP_001932096	Alignment spans fusion central region
fusion_48_XP_003296895	Component alignment coverage too low
fusion_49_KDQ30737	Alignment spans fusion central region
fusion_6_KNE57841	Alignment spans fusion central region
fusion_7_KNE72089	Alignment spans fusion central region
fusion_8_XP_011388676	Alignment spans fusion central region

amongst 81 multigene families, where an average of 12.6±33.5 genes per composite family was observed. These results are an order of magnitude greater than those reported by Leonard and Richards (2012). We observed compositeBLAST to be much faster than *fdf*BLAST, taking just 118 seconds to process the data in comparison to 8019 seconds.

4.3.3. Rate of composite generation

Using the approximated fungal divergence time of 566.22 Ma, the rate of composite generation in fungi was calculated to be 0.14 events/Ma (excluding singleton families) to 0.53 events/Ma (including singleton families)

There has not been a comparison for fungal genomes, however this figure is far greater than plant composite gene evolutionary rate observed Nakamura *et al.* (2007) who estimated the rate to be $1.0e^{-11}$-$2.0e^{-11}$ events/year ($1.0e^{-05}$-$2.0e^{-05}$ events/Ma).

4.3.4. Detection of composite AMR genes

We detected 13 fused AMR genes using compositeBLAST. Of the 13 detected genes, 9 had been previously reported by other authors (Table 4.3.2.). We detected 4 fusion genes (*mupA*, *mupB*, *rphA*, and *rphB*) from the 2019 CARD gene dataset that, to our knowledge, have not been reported before. The four genes can be broken into two sets of duplicate genes, where *mupA* and *mupB* (*mupAB*) confer resistance to mupirocin (Troeman *et al.*, 2019), *and rphA* and *rphB* (*rphAB*) confer resistance to rifamycin (Sensi, 1983; Baysarowich *et al.*, 2008). The specifics of these resistance mechanisms and fusion genes are explained in the next sections.

Table 4.3.2. 11 AMR composite genes detected by CompositeBLAST

Each composite AMR is represented by its accession ID and gene ID. Genes previously reported as composites are cited. Four detected genes have not been previously reported (emboldened).

Accession ID	Gene	Reported by
ARO:3000444	**rphA**	**This study**
ARO:3000501	rpoB	Zakharova *et al*., 1999
ARO:3000510	**mupB**	**This study**
ARO:3000521	**mupA**	**This study**
ARO:3000535	macB	Coleman *et al*., 2015
ARO:3002970	vanTC	Meziane-Cherif *et al*., 2015
ARO:3002971	vanTE	Meziane-Cherif *et al*., 2015
ARO:3002972	vanTG	Meziane-Cherif *et al*., 2015
ARO:3002975	vanTN	Meziane-Cherif *et al*., 2015
ARO:3002985	arnA	Williams *et al*., 2005
ARO:3003324	mprF(A)	Maloney *et al*., 2009
ARO:3003770	mprF(B)	Maloney *et al*., 2009
ARO:3003992	**rphB**	**This study**

4.3.4.1 Rifamycin resistance

Two rifamycin resistance genes, *rphA* (ARO:3000444) and *rphB* (ARO:3003992) were reported to be composites by compositeBLAST. Both *rphA* and *rphB* belong to the rifamycin phosphotransferase (E.C: 2.7.9.6) class of AMR resistance cassettes (Boehme *et al.*, 2010).

Rifamycins are a class of ansamycin bacterial polyketides or artificially manufactured antibiotics used primarily to combat mycobacterial infections in clinical settings (Sensi, 1983). Rifamycins bind the β subunit of DNA-dependant RNA polymerase (*rpoB*) and physically block RNA elongation (Feklistov *et al.*, 2008). As rifamycins act through steric occlusion, resistance may arise through point mutation in *rpoB* that reduce rifamycin binding affinity (Campbell *et al.*, 2001). A 531 Ser→Leu point mutation (S531L) in *rpoB* is the most common rifamycin resistance conferring point mutation in *Mycobacterium tuberculosis* (Lemus *et al.*, 2004). In *Norcardia farcinica*, an *rpoB* duplicate (*rpoB2*) has undergone sufficient cumulative mutation to confer resistance, where 88% sequence similarity is observed between both sequences (Ishikawa *et al.*, 2006). Ishikawa and company (2006) also observed differential *rpoB* and *rpoB2* expression is observed depending on rifamycin presence.

Resistance may also arise through rifamycin inactivation through position 21 phosphorylation, as has been observed in both *rphA* and *rphB* (Campbell *et al.*, 2001; Goldstein, 2014). Phosphorylation mediated rifamycin resistance is observed throughout a range of environments as a competitive response between bacterial communities in addition to clinical settings (Baysarowich *et al.*, 2008).

rphA and *rphB* are full length homologs (FLHs) of each other (*rphAB*), and display an overlapping set of *bona fide* N- and C-terminal components and phylogenetic distributions with the exception of an *rphB* FLH presenting in *Halogeometricum borinquense* (Euryarcheota) where an *rphA* FLO is not observed (Table 4.3.3.). *rphAB* FLOs observed in three

Table 4.3.3. Distribution of FLHs of each composite AMR gene.

For each genome, presence and absence of a composite (F), N-terminal component (N), and C-terminal component are highlighted in a coloured box. In each box a "1" represents presence and a "0" represents absence. The presence of a composite, N-terminal component and C-terminal components are indicated by purple, red and blue colouration, respectively.

		rphA			rphB			mupA			mupB		
Species	**Clade**	**F**	**N**	**C**	**F**	**N**	**C**	**F**	**N**	**C**	**F**	**N**	**C**
Heimdallarchaeota archaeon	Archaea, Asgard group	0	1	0	0	1	0	1	0	1	1	0	1
Lokiarchaeota archaeon	Archaea, Asgard group	0	0	0	0	0	0	0	0	0	0	0	0
Korarchaeum cryptofilum	Archaea, Candidatus Korarchaeota	0	1	0	0	1	0	1	1	0	1	0	0
Acidilobus saccharovorans	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	0	1	0	1
Caldisphaera lagunensis	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	0	0	1	0	0
Aeropyrum pernix	Archaea, Crenarchaeota, Thermoprotei	0	1	1	0	1	1	1	1	0	1	0	0
Desulfurococcus amylolyticus	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	0	0	1	0	1
Ignicoccus hospitalis	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	0	1	0	1
Ignisphaera aggregans	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	0	1	0	1
Staphylothermus marinus	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	0	0	1	0	1
Thermogladius calderae	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	1	1	0	1
Thermosphaera aggregans	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	0	1	1	0	1
Hyperthermus butylicus	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	0	1	0	1
Pyrolobus fumarii	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	1	1	0	1
Fervidicoccus fontis	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	0	0	1	0	0
Acidianus hospitalis	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	1	1	0	1
Metallosphaera sedula	Archaea, Crenarchaeota, Thermoprotei	0	1	0	0	1	0	1	1	0	1	0	1
Saccharolobus solfataricus	Archaea, Crenarchaeota, Thermoprotei	0	1	0	0	1	0	1	0	1	1	0	1
Sulfolobus acidocaldarius	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	1	1	0	1
Sulfurisphaera tokodaii	Archaea, Crenarchaeota, Thermoprotei	0	1	0	0	1	0	1	1	1	1	0	1
Thermofilum pendens	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	0	1	0	0
Caldivirga maquilingensis	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	0	0	1	0	0
Pyrobaculum aerophilum	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	1	1	0	0
Thermoproteus tenax	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	0	1	0	1	0	0
Vulcanisaeta distributa	Archaea, Crenarchaeota, Thermoprotei	0	0	0	0	0	0	1	1	0	0	0	0
Archaeoglobus fulgidus	Archaea, Euryarchaeota, Archaeoglobi	0	0	0	0	0	0	1	1	1	1	0	1
Ferroglobus placidus	Archaea, Euryarchaeota, Archaeoglobi	0	1	1	0	1	1	1	1	1	1	0	1

Species	Clade	rphA			rphB			mupA			mupB		
		F	N	C	F	N	C	F	N	C	F	N	C
Aciduliprofundum boonei	Archaea, Euryarchaeota, Diaforarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanomethylophilus alvus	Archaea, Euryarchaeota, Diaforarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanomassiliicoccus intestinalis	Archaea, Euryarchaeota, Diaforarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Picrophilus torridus	Archaea, Euryarchaeota, Diaforarchaea	0	0	0	0	0	0	1	0	0	1	0	0
Thermoplasma acidophilum	Archaea, Euryarchaeota, Diaforarchaea	0	0	0	0	0	0	1	0	0	1	0	0
Methanobacterium lacus	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	0	1	1	0	1
Methanobrevibacter ruminantium	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	0	1	1	0	1
Methanosphaera stadtmanae	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	0	1	1	0	1
Methanothermobacter thermautotrophicus	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	1	1	1	0	1
Methanothermus fervidus	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	1	1	1	0	1
Methanocaldococcus infernus	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	1	1	1	0	1
Methanocaldococcus jannaschii	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	0	1	1	0	1
Methanotorris igneus	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	0	1	1	0	1
Methanococcus aeolicus	Archaea, Euryarchaeota, Methanomada	0	0	0	0	0	0	1	0	1	1	0	1
Methanopyrus kandleri	Archaea, Euryarchaeota, Methanopyri	0	0	0	0	0	0	1	0	1	1	0	1
Haloarcula marismortui	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	0	1	0	0	0	0
Halomicrobium mukohataei	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	0	0	0
Natronomonas moolapensis	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	0	0	0
Natronomonas pharaonis	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	1	0	0
Halalkalicoccus jeotgali	Archaea, Euryarchaeota, Stenosarchaea	1	0	0	1	0	0	0	1	0	1	0	0
Halobacterium salinarum	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	0	1	0	0	0	0
Haloferax mediterranei	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	0	0	0	0	0	0
Halogeometricum borinquense	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	1	0	0	1	1	0	0	0	1
Haloquadratum walsbyi	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	0	0	0
Halopiger xanaduensis	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	1	0	0
Haloterrigena turkmenica	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	1	0	0
Halovivax ruber	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	0	1	0	0	0	0
Natrialba magadii	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	1	0	0
Natrinema pellirubrum	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	1	0	0
Natronobacterium gregoryi	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	0	1	0	0
Methanocella arvoryzae	Archaea, Euryarchaeota, Stenosarchaea	1	0	0	1	0	0	1	0	1	1	0	1
Methanocorpusculum labreanum	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	1	1	0	1
Methanoculleus bourgensis	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanolacinia petrolearia	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	1	1	0	1
Methanoregula boonei	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	1	1	0	1
Methanosphaerula palustris	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanospirillum hungatei	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	1	1	1	0	1
Methanosaeta harundinacea	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanothrix soehngenii	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1

Species	Clade	rphA			rphB			mupA			mupB		
		F	N	C	F	N	C	F	N	C	F	N	C
Methanococcoides burtonii	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanohalobium evestigatum	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanohalophilus mahii	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanomethylovorans hollandica	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanosalsum zhilinae	Archaea, Euryarchaeota, Stenosarchaea	0	0	0	0	0	0	1	0	1	1	0	1
Methanosarcina acetivorans	Archaea, Euryarchaeota, Stenosarchaea	1	1	0	1	1	0	1	1	1	1	0	1
Pyrococcus furiosus	Archaea, Euryarchaeota, Thermococci	0	0	0	0	0	0	1	1	1	1	0	1
Thermococcus kodakarensis	Archaea, Euryarchaeota, Thermococci	0	0	0	0	0	0	1	1	0	1	0	1
Nanoarchaeum equitans	Archaea, Nanoarchaeota, Nanoarchaeales	0	0	0	0	0	0	0	0	0	0	0	0
Cenarchaeum symbiosum	Archaea, Thaumarchaeota, Cenarchaeales	0	0	0	0	0	0	1	1	0	1	1	0
Nitrosopumilus maritimus	Archaea, Thaumarchaeota, Nitrosopumilales	0	0	0	0	0	0	1	1	0	1	0	0
Nitrososphaera gargensis	Archaea, Thaumarchaeota, Nitrososphaeria	0	0	0	0	0	0	1	0	0	0	0	0
Acidobacterium capsulatum	Bacteria, Acidobacteria, Acidobacteriales	0	0	0	0	0	0	0	1	0	0	0	0
Granulicella mallensis	Bacteria, Acidobacteria, Acidobacteriales	0	0	0	0	0	0	0	1	1	0	0	1
Granulicella tundricola	Bacteria, Acidobacteria, Acidobacteriales	0	0	0	0	0	0	0	1	0	0	0	0
Terriglobus roseus	Bacteria, Acidobacteria, Acidobacteriales	0	0	0	0	0	0	0	1	1	0	0	1
Thermobispora bispora	Bacteria, Actinobacteria, Actinobacteria	0	0	0	0	0	0	1	1	1	1	0	1
Mycobacterium tuberculosis	Bacteria, Actinobacteria, Corynebacteriales	0	0	0	0	0	0	1	1	1	1	0	1
Micrococcus luteus	Bacteria, Actinobacteria, Micrococcales	0	0	0	0	0	0	1	1	1	1	0	1
Streptomyces coelicolor	Bacteria, Actinobacteria, Streptomycetales	0	0	0	0	0	0	1	1	1	1	0	1
Aquifex aeolicus	Bacteria, Aquificae, Aquificales	0	0	0	0	0	0	1	1	1	0	1	1
Hydrogenobacter thermophilus	Bacteria, Aquificae, Aquificales	0	0	0	0	0	0	1	1	0	1	1	1
Thermocrinis albus	Bacteria, Aquificae, Aquificales	0	0	0	0	0	0	1	1	1	0	1	1
Desulfurobacterium thermolithotrophum	Bacteria, Aquificae, Desulfurobacteriales	0	0	0	0	0	0	1	0	1	0	0	1
Chthonomonas calidirosea	Bacteria, Armatimonadetes, Chthonomonadetes	0	0	0	0	0	0	0	1	1	0	0	1
Salinibacter ruber	Bacteria, Bacteroidetes, Bacteroidetes	0	0	0	0	0	0	1	1	1	1	0	1
Bacteroides thetaiotaomicron	Bacteria, Bacteroidetes, Bacteroidia	0	0	0	0	0	0	1	0	1	1	0	1
Azobacteroides pseudotrichonymphae	Bacteria, Bacteroidetes, Bacteroidia	0	0	0	0	0	0	1	0	0	1	0	0
Flavobacterium psychrophilum	Bacteria, Bacteroidetes, Flavobacteriia	0	0	0	0	0	0	1	0	1	1	0	1
Cloacimonas acidaminovorans	Bacteria, Candidatus Cloacimonetes	0	0	0	0	0	0	1	1	1	1	0	1
Chlamydia trachomatis	Bacteria, Chlamydiae, Chlamydiales	0	0	0	0	0	0	1	1	0	1	0	0
Protochlamydia amoebophila	Bacteria, Chlamydiae, Parachlamydiales	0	0	0	0	0	0	1	1	1	1	0	1
Parachlamydia acanthamoebae	Bacteria, Chlamydiae, Parachlamydiales	0	0	0	0	0	0	1	1	1	1	0	1
Waddlia chondrophila	Bacteria, Chlamydiae, Parachlamydiales	0	0	0	0	0	0	1	1	1	1	0	1
Chlorobaculum tepidum	Bacteria, Chlorobi, Chlorobia	0	0	0	0	0	0	1	1	1	1	0	1
Caldilinea aerophila	Bacteria, Chloroflexi, Caldilineae	0	0	0	0	0	0	1	1	0	1	0	0
Chloroflexus aurantiacus	Bacteria, Chloroflexi, Chloroflexia	0	0	0	0	0	0	1	0	0	1	0	0
Dehalococcoides mccartyi	Bacteria, Chloroflexi, Dehalococcoidia	0	0	0	0	0	0	1	1	0	1	0	0
Dehalogenimonas lykanthroporepellens	Bacteria, Chloroflexi, Dehalococcoidia	0	0	0	0	0	0	1	1	0	1	0	1

Species	Clade	rphA			rphB			mupA			mupB		
		F	N	C	F	N	C	F	N	C	F	N	C
Sphaerobacter thermophilus	Bacteria, Chloroflexi, Sphaerobacteridae	0	1	1	0	1	1	1	1	1	1	0	1
Desulfurispirillum indicum	Bacteria, Chrysiogenetes, Chrysiogenales	0	0	0	0	0	0	1	1	1	0	0	1
Coprothermobacter proteolyticus	Bacteria, Coprothermobacterota	0	0	0	0	0	0	0	1	1	0	0	1
Gloeobacter violaceus	Bacteria, Cyanobacteria, Gloeobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Cyanobacterium aponinum	Bacteria, Cyanobacteria, Oscillatoriophycideae	0	0	0	0	0	0	0	1	1	0	0	1
Synechocystis sp.	Bacteria, Cyanobacteria, Synechococcales	0	0	0	0	0	0	0	1	1	0	0	1
Prochlorococcus marinus	Bacteria, Cyanobacteria, Synechococcales	0	0	0	0	0	0	0	1	0	0	0	1
Thermosynechococcus elongatus	Bacteria, Cyanobacteria, Synechococcales	0	0	0	0	0	0	0	1	1	0	0	1
Calditerrivibrio nitroreducens	Bacteria, Deferribacteres, Deferribacterales	0	0	0	0	0	0	0	1	1	0	0	1
Deferribacter desulfuricans	Bacteria, Deferribacteres, Deferribacterales	0	0	0	0	0	0	0	1	1	0	0	1
Denitrovibrio acetiphilus	Bacteria, Deferribacteres, Deferribacterales	0	0	0	0	0	0	0	1	1	0	0	1
Deinococcus radiodurans	Bacteria, Deinococcus-Thermus, Deinococci	0	1	1	0	1	1	1	1	1	1	0	1
Marinithermus hydrothermalis	Bacteria, Deinococcus-Thermus, Deinococci	0	0	0	0	0	0	1	1	1	1	0	1
Meiothermus silvanus	Bacteria, Deinococcus-Thermus, Deinococci	0	0	0	0	0	0	1	1	1	1	0	1
Oceanithermus profundus	Bacteria, Deinococcus-Thermus, Deinococci	0	0	0	0	0	0	1	1	1	1	0	1
Thermus thermophilus	Bacteria, Deinococcus-Thermus, Deinococci	0	0	0	0	0	0	1	1	1	1	0	1
Dictyoglomus turgidum	Bacteria, Dictyoglomi, Dictyoglomales	0	0	0	0	0	0	1	1	1	1	0	1
Elusimicrobium minutum	Bacteria, Elusimicrobia, Elusimicrobia	0	0	0	0	0	0	0	1	1	0	0	1
Fibrobacter succinogenes	Bacteria, Fibrobacteres, Fibrobacterales	0	0	0	0	0	0	1	1	1	1	0	1
Bacillus subtilis	Bacteria, Firmicutes, Bacilli	1	0	0	1	0	0	0	1	1	0	0	1
Geobacillus kaustophilus	Bacteria, Firmicutes, Bacilli	0	0	0	0	0	0	0	1	1	0	0	1
Halobacillus halophilus	Bacteria, Firmicutes, Bacilli	0	0	0	0	0	0	1	1	1	1	0	1
Listeria monocytogenes	Bacteria, Firmicutes, Bacilli	0	1	0	0	1	0	0	1	1	0	0	1
Staphylococcus aureus	Bacteria, Firmicutes, Bacilli	0	0	0	0	0	0	1	1	1	1	0	1
Enterococcus faecalis	Bacteria, Firmicutes, Bacilli	0	0	0	0	0	0	0	1	1	0	0	1
Streptococcus pneumoniae	Bacteria, Firmicutes, Bacilli	0	0	0	0	0	0	0	1	1	0	0	1
Clostridium botulinum	Bacteria, Firmicutes, Clostridia	0	0	0	0	0	0	1	1	1	1	0	1
Moorella thermoacetica	Bacteria, Firmicutes, Clostridia	0	0	0	0	0	0	1	1	1	0	0	1
Fusobacterium nucleatum	Bacteria, Fusobacteria, Fusobacteriales	0	0	0	0	0	0	1	1	1	0	0	1
Ilyobacter polytropus	Bacteria, Fusobacteria, Fusobacteriales	0	0	0	0	0	0	0	1	1	0	0	1
Leptotrichia buccalis	Bacteria, Fusobacteria, Fusobacteriales	0	0	0	0	0	0	0	1	1	1	0	1
Sebaldella termitidis	Bacteria, Fusobacteria, Fusobacteriales	1	0	0	1	0	0	0	1	1	0	0	1
Streptobacillus moniliformis	Bacteria, Fusobacteria, Fusobacteriales	0	0	0	0	0	0	0	1	1	0	0	1
Gemmatimonas aurantiaca	Bacteria, Gemmatimonadetes, Gemmatimonadales	0	0	0	0	0	0	1	1	1	1	0	1
Leptospirillum ferrooxidans	Bacteria, Nitrospirae, Nitrospirales	0	0	0	0	0	0	0	1	1	0	0	1
Nitrospira moscoviensis	Bacteria, Nitrospirae, Nitrospirales	0	0	0	0	0	0	0	1	1	0	0	1
Thermodesulfovibrio yellowstonii	Bacteria, Nitrospirae, Nitrospirales	0	0	0	0	0	0	1	1	1	0	0	1
Isosphaera pallida	Bacteria, Planctomycetes, Planctomycetia	0	0	0	0	0	0	0	1	1	0	0	1
Pirellula staleyi	Bacteria, Planctomycetes, Planctomycetia	0	0	0	0	0	0	1	1	1	1	0	1

Species	Clade	rphA			rphB			mupA			mupB		
		F	N	C	F	N	C	F	N	C	F	N	C
Planctopirus limnophila	Bacteria, Planctomycetes, Planctomycetia	0	0	0	0	0	0	0	1	0	0	0	0
Rhodopirellula baltica	Bacteria, Planctomycetes, Planctomycetia	0	0	0	0	0	0	1	1	0	1	0	0
Rubinisphaera brasiliensis	Bacteria, Planctomycetes, Planctomycetia	0	0	0	0	0	0	0	0	0	0	0	0
Caulobacter vibrioides	Bacteria, Proteobacteria, Alphaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Bradyrhizobium diazoefficiens	Bacteria, Proteobacteria, Alphaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Ehrlichia chaffeensis	Bacteria, Proteobacteria, Alphaproteobacteria	0	0	0	0	0	0	1	1	0	1	0	1
Rickettsia prowazekii	Bacteria, Proteobacteria, Alphaproteobacteria	0	0	0	0	0	0	1	1	1	1	0	1
Bordetella pertussis	Bacteria, Proteobacteria, Betaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Burkholderia multivorans	Bacteria, Proteobacteria, Betaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Neisseria meningitidis	Bacteria, Proteobacteria, Betaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Thiobacillus denitrificans	Bacteria, Proteobacteria, Betaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	1
Pseudodesulfovibrio aespoeensis	Bacteria, Proteobacteria, Deltaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Lawsonia intracellularis	Bacteria, Proteobacteria, Deltaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Geobacter sulfurreducens	Bacteria, Proteobacteria, Deltaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Syntrophobacter fumaroxidans	Bacteria, Proteobacteria, Deltaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Campylobacter jejuni	Bacteria, Proteobacteria, Epsilonproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Helicobacter pylori	Bacteria, Proteobacteria, Epsilonproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Wolinella succinogenes	Bacteria, Proteobacteria, Epsilonproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Escherichia coli	Bacteria, Proteobacteria, Gammaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Salmonella typhimurium	Bacteria, Proteobacteria, Gammaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Shigella dysenteriae	Bacteria, Proteobacteria, Gammaproteobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Yersinia pestis	Bacteria, Proteobacteria, Gammaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Pseudomonas aeruginosa	Bacteria, Proteobacteria, Gammaproteobacteria	0	0	0	0	0	0	0	1	0	0	0	0
Brachyspira hyodysenteriae	Bacteria, Spirochaetes, Brachyspirales	0	0	0	0	0	0	1	1	1	1	0	1
Leptospira interrogans	Bacteria, Spirochaetes, Leptospirales	0	0	0	0	0	0	0	1	0	0	0	0
Turneriella parva	Bacteria, Spirochaetes, Leptospirales	0	0	0	0	0	0	0	1	1	0	0	1
Borrelia burgdorferi	Bacteria, Spirochaetes, Spirochaetales	0	0	0	0	0	0	1	1	0	1	0	0
Sediminispirochaeta smaragdinae	Bacteria, Spirochaetes, Spirochaetales	0	0	0	0	0	0	1	1	0	1	0	1
Spirochaeta thermophila	Bacteria, Spirochaetes, Spirochaetales	0	0	0	0	0	0	1	1	0	1	0	0
Treponema pallidum	Bacteria, Spirochaetes, Spirochaetales	0	0	0	0	0	0	1	1	0	1	0	0
Acetomicrobium mobile	Bacteria, Synergistetes, Synergistia	0	0	0	0	0	0	0	1	1	0	0	1
Aminobacterium colombiense	Bacteria, Synergistetes, Synergistia	0	0	0	0	0	0	0	1	1	0	0	1
Thermanaerovibrio acidaminovorans	Bacteria, Synergistetes, Synergistia	0	0	0	0	0	0	0	1	1	0	0	1
Thermovirga lienii	Bacteria, Synergistetes, Synergistia	0	0	0	0	0	0	0	1	1	0	0	1
Acholeplasma laidlawii	Bacteria, Tenericutes, Mollicutes	0	0	0	0	0	0	0	1	1	0	0	1
Onion yellows	Bacteria, Tenericutes, Mollicutes	0	0	0	0	0	0	0	1	1	0	0	1
Mesoplasma florum	Bacteria, Tenericutes, Mollicutes	0	0	0	0	0	0	0	1	1	0	0	1
Mycoplasma genitalium	Bacteria, Tenericutes, Mollicutes	0	0	0	0	0	0	0	1	1	0	0	1

Species	Clade	rphA			rphB			mupA			mupB		
		F	N	C	F	N	C	F	N	C	F	N	C
Thermodesulfatator indicus	Bacteria, Thermodesulfobacteria	0	0	0	0	0	0	1	1	1	0	0	1
Thermodesulfobacterium geofontis	Bacteria, Thermodesulfobacteria	0	0	0	0	0	0	0	1	1	0	0	1
Kosmotoga olearia	Bacteria, Thermotogae, Kosmotogales	0	0	0	0	0	0	1	1	1	0	0	1
Marinitoga piezophila	Bacteria, Thermotogae, Petrotogales	0	0	0	0	0	0	0	1	1	0	0	1
Fervidobacterium nodosum	Bacteria, Thermotogae, Thermotogales	0	0	0	0	0	0	1	1	1	1	0	1
Pseudothermotoga lettingae	Bacteria, Thermotogae, Thermotogales	0	0	0	0	0	0	1	1	1	1	0	1
Thermotoga maritima	Bacteria, Thermotogae, Thermotogales	0	0	0	0	0	0	1	1	1	0	0	1
Methylacidiphilum infernorum	Bacteria, Verrucomicrobia, Methylacidiphilae	0	0	0	0	0	0	0	1	0	0	0	1
Opitutus terrae	Bacteria, Verrucomicrobia, Opitutae	0	0	0	0	0	0	0	1	1	0	0	1
Coraliomargarita akajimensis	Bacteria, Verrucomicrobia, Opitutae	0	0	0	0	0	0	0	1	1	0	0	1
Akkermansia muciniphila	Bacteria, Verrucomicrobia, Verrucomicrobiae	0	0	0	0	0	0	0	1	0	0	0	1

Euryarchaeota (*Halalkalicoccus jeotgali*, *Methanocella arvoryzae*, and *Methanosarcina acetivorans*), in *Bacillus subtilis* (Firmicutes), and in *Sebaldella termitidis* (Fusobacteria).

Both N-terminal and C-terminal component FLHs are observed in four distantly related species, *Aeropyrum pernix* (Crenarchaeota), *Ferroglobus placidus* (Euryarchaeota), *Sphaerobacter thermophilus* (Chloroflexi), and *Deinococcus radiodurans* (Deinococcus-Thermus). N-terminal component FLOs were also observed in the Heimdallarchaeota archaeon (Asgard clade), in *Korarchaeum cryptofilum* (Korarcheaota), in three Crenarchaeota (*Metallosphaera sedula*, *Saccharolobus solfataricus*, and *Sulfurisphaera tokodaii*), in two *Methanosarcina acetivorans* (Euryarchaeota) and in *Listeria monocytogenes* (Firmicutes). C-terminal FLOs were only observed in conjunction with N-terminal FLOs.

Representative C- and N-terminal *bona fide* components (D1C8T8_SPHTD and D1C8T7_SPHTD for *rphA* and *rphB* were selected from *Sphaerobacter thermophilus* (Figures 4.3.1.-2.). D1C8T8_SPHTD was reported to be a phosphoenolpyruvate (PEP) utilizing protein mobile region in UniProt and as ENOG4106850 (phosphoenolpyruvate synthase) in EggNOG v5. Comparatively, D1C8T7_SPHTD was identified as a pyruvate phosphate dikinase PEP/pyruvate-binding protein in UniProt and as ENOG4107R95 (pyruvate kinase (by similarity) in EggNOG v5 (Huerta-Cepas *et al.*, 2018). D1C8T8_SPHTD shared a PEP utilizing mobile domain (PF00391) with the N-termini or *rphAB* and a pyruvate phosphate dikinase, PEP/pyruvate binding domain (PF01326) with the *rphAB* C-termini.

4.3.4.2. Mupirocin resistance

Two mupirocin resistance genes (ARO:3000510 (*mupA*) and ARO:3000521 (*mupB*)) were reported to be composites. Mupirocin is administered for topical Firmicute infections, especially MRSA infections (Troeman *et al.*, 2019). Mupirocin inhibits protein synthesis

Figure 4.3.1. Representative composite-component alignments for *rphA*

Both homologous domains are located towards the termini, with no observed homology within the centre. *rphA* shares a pyruvate phosphate dikinase domain (PF01326) with D1C8T8 along their N-termini. *rphA* and D1C8T7 share a PEP utilizing mobile domain (F00391) along their C-termini.

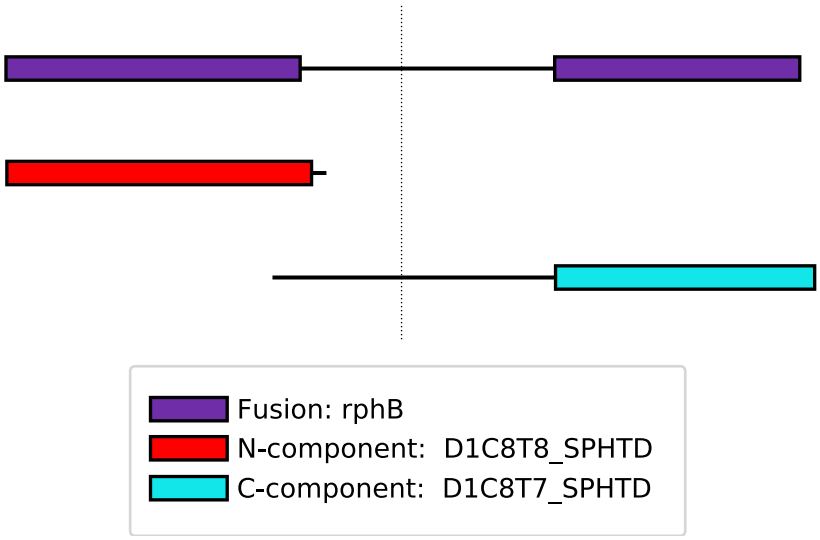

Figure 4.3.2. Representative alignments for an *rphB* composite gene.

rphB is almost idistinguishable from *rphA*. Both homologous domains are located towards the termini, with no observed homology within the centre. *rphB* shares a pyruvate phosphate dikinase domain (PF01326) with D1C8T8 along their N-termini. *rphB* and D1C8T7 share a PEP utilizing mobile domain (F00391) along their C-termini.

through reversible t-RNA synthetase binding (Martin and Simpson, 1989). Both low-level and high-level mupirocin resistance types have been reported. Low-level resistance is hypothesised to arise through cumulative point mutations in wild type *ileS*, whereas level resistance is conferred by two distinct loci, *mupA* and *mupB*, which are reported to display considerable sequence differences to wild type *ileS* (Gilbart *et al.*, 1993; Cookson, 1998)

Both *mupA* and *mupB* share *bona fide* C-terminal FLOs (SYM_LISMO was selected as a representative), however no mutual *bona fide* N-terminal FLH was observed so SYL_HELPY was selected as a representative for *mupB* and D3DI52_HYDTT for *mupA*. An overlapping alignment was observed between both components in both cases (Figures 4.3.3.-4.).

SYL_HELPY (*leuS*) was identified as *Helicobacter pylori* leucine--tRNA ligase (E.C: 6.1.1.4) using UniProtKB (Bateman *et al.*, 2017). D3DI52_HYDTT was also identified as leucine--tRNA ligase (*leuS*) from *Hydrogenobacter thermophilus.* Conversely, SYM_LISMO (*metG*) was identified as *Listeria monocytogenes* methionine--tRNA ligase (E.C: 6.1.1.10). Both *mupA* and *mupB* (*mupAB*) share a tRNA synthetases class I (I, L, M and V) domain (PF00133) with *leuS* and a tRNA anticodon binding domain (PF08264) with *metG*. *mupAB* were reported to contain just these two PFAM domains. Interestingly, two PFAM domains were observed in *leuS* and in *metG* that are not present in the other component or in *mupAB*. *leuS* was reported to contain an additional tRNA synthetases class I (M) domain (PF09334) and a leucyl-tRNA synthetase, domain 2 (PF13603). Comparatively, *metG* was observed to also contain a tRNA binding domain (PF01588) and a tRNA synthetases class I (M) domain (PF09334).

mupA FLHs were distributed across the dataset (Table 4.3.3.). Interestingly, no *mupA* FLOs were observed in the phyla Acidobacteria, Cyanobacteria, Deferribactres, Fusobacteria, Tenericutes, Proteobacteria, or Verrucomicrobia. *mupA* N-terminal component FLHs were

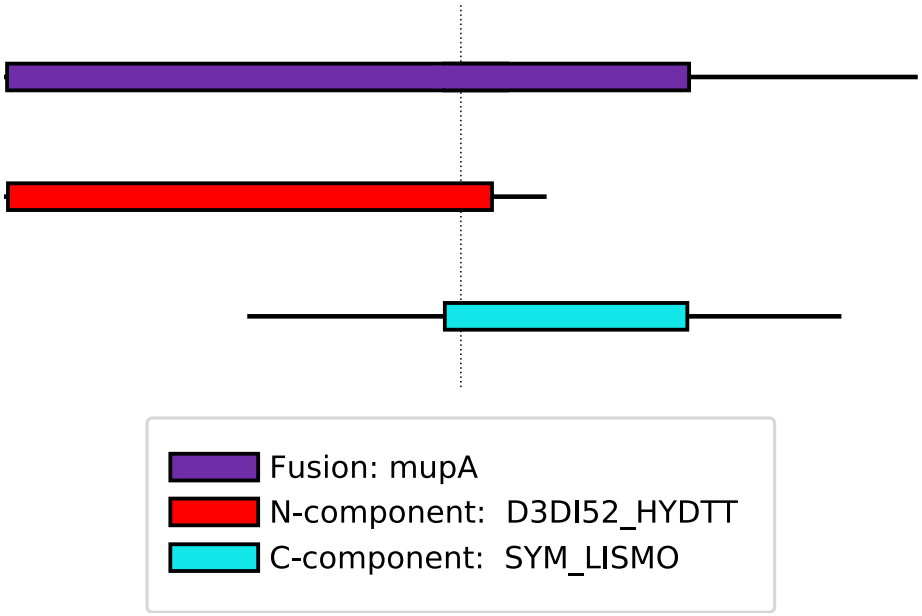

Figure 4.3.3. Representative composite-component alignments for *mupA*

Both homologous domains are located towards the centre. The N-terminal component (D3D152) and *mupA* share a tRNA synthetases class I (I, L, M and V) domain (PF00133) and the C-terminal component (SYM) share a tRNA anticodon binding domain (PF0268). *mupA* displays a "fused alignment" where both the alignments of both components partially overlap against the composite, resulting in a long alignment, unlike the "polarised" alignments observed in *rphAB*.

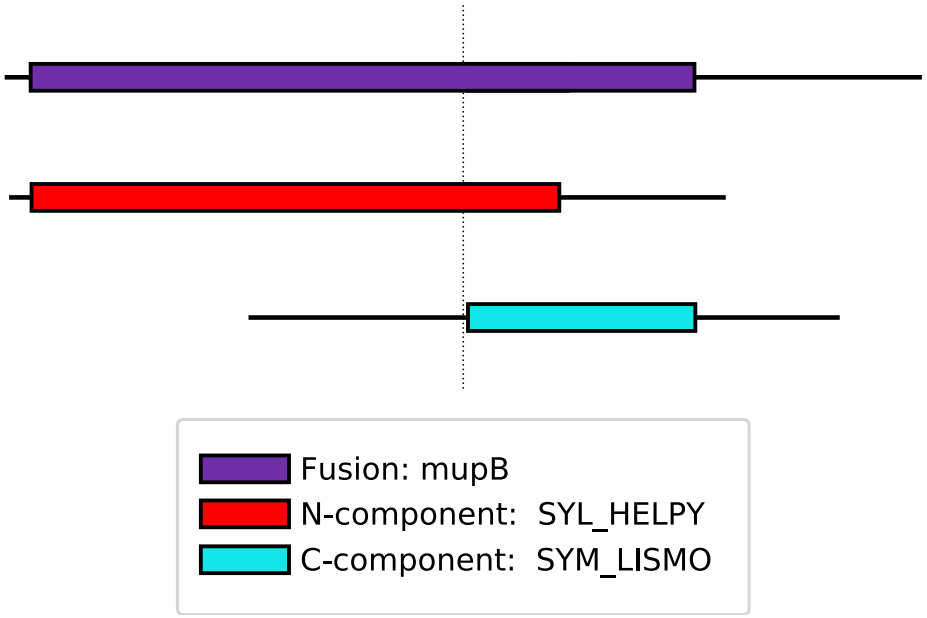

Figure 4.3.4. Representative composite-component alignments for *mupB*

Both homologous domains are located towards the centre. The N-terminal component (D3D152) and mupA share a tRNA synthetases class I (I, L, M and V) domain (PF00133) and the C-terminal component (SYM) share a tRNA anticodon binding domain (PF0268). mupB displays a "fused alignment" where both the alignments of both components partially overlap against the composite, resulting in a long alignment, unlike the "polarised" alignments observed in *rphAB*.

observed in a patchy distribution across the archaea but were almost ubiquitous across the bacteria. *mupA* C-terminal component FLHs were also relatively evenly distributed across the bacteria, however no FLHs were observed within the Spirocheates or Betaproteobacteria.

4.4. Discussion

4.4.1. compositeBLAST is a useful tool for composite detection

In this chapter we have presented compositeBLAST to be a valuable tool for detecting and visualising high quality composite genes. We have demonstrated that compositeBLAST to be more sensitive than *fdf*BLAST (recovering an order of magnitude more genes than those found by Leonard and Richards (2012)).

The parameters for composite detection *via* homology statistics alone, while good for determining relationships between two sequences, may not be adequate for composite detection through the lens of "fusion" and "fission". The restriction of multiple HSPs aids in the reduction of Type I errors that may be encountered when using tools such as CompositeSearch (Pathmanathan *et. al.,* 2018). As compositeBLAST requires a degree of domain conservation between the composite and each component and for there to be no domain conservation between both components we believe this tool to be highly selective, depending on input data quality. Composite genes have the potential for resolving trifurcations in phylogenies and evolutionary histories (Leonard and Richards, 2012). As compositeBLAST provides a "middle ground" between CompositeSearch and *fdf*BLAST (in terms of sensitivity and selectivity), it may be used to detect previously unknown evolutionary traits throughout the tree of life. The sensitivity and selectivity of compositeBLAST is further exemplified by the rate of composite

emergence in plants between those detected by compositeBLAST and those detected by Nakamura and company (2007).

4.4.2. Efficacy of detecting clinically relevant composite genes

Composite genes, specifically fusion genes, are aetiologies of many cancers due to the disruption of signalling cascades (Mitelman *et al.* , 2007; Liu *et al.*, 2009; Stransky *et al.*, 2014). Differential gene fusion events between hosts and pathogens are also potential targets for increased success in the treatment of infectious diseases (Trimpalis *et al.*, 2013). By detecting genes from metabolically important pathways that are fused in the host and unfused in the pathogen, or *vice versa*, it is possible to use these as pathway disruptive targets. We identified four AMR genes that confer resistance to mupirocin and ripamycin to be composites.

While composite genes have been implicated in secondary metabolite and AMR evolution (Coleman *et al.* 2014), the extent of such processes on the evolution of pathogenicity is yet to be elucidated. Therefore, compositeBLAST could be used to decipher such relationships and aid in the development of new treatment regimens for emerging and persistent infectious microorganisms.

4.5. Conclusion

We have demonstrated the efficacy of compositeBLAST in detecting *bona fide* composite genes and by using compositeBLAST we were able to identify two classes of AMR genes as composite genes which have not been previously reported. Considering the wealth of high quality genomes available, compositeBLAST provides a mechanism to detect previously unknown high-confidence gene remodelling events which may highlight previously unknown evolutionary relationships, and may be used to help resolve trifurcations in phylogenies.

Chapter V

Concluding Remarks and Future Work

5.1 Concluding remarks

It is evident that the evolution of genes and of genomes is not dependant on a vertical model of inheritance (Haggerty *et al.*, 2014). It has been long established that prokaryote genomes frequently evolve through horizontal gene transfer (Kurland *et al.*, 2003; Groussin *et al.*, 2016), and more recently, evidence of horizontal gene transfer (HGT) between eukaryotes (Fukatsu, 2010) and interdomain HGT (McCarthy and Fitzpatrick, 2016). These genes can lead to incongruent phylogenies and, as such, are quite often cause misrepresentation and incorrect placement of taxa during the construction of prokaryote superalignments (Puigbò *et al.*, 2009, 2010). Comparatively gene evolution while once thought to be mostly under a model of vertical inheritance, displays striking evidence of horizontal evolution through the transfer and rearrangement of domains (Yanai *et al.*, 2002; Leonard and Richards, 2012; Jachiet *et al.*, 2013; Haggerty *et al.*, 2014; McLysaght and Guerzoni, 2015; Pathmanathan *et al.*, 2018). The extent of sequence rearrangement and the sharing and merging of gene domains was not fully appreciated until the development of CompositeSearch due to computational restrictions (Pathmanathan *et al.*, 2018). Reconstructing evolutionary histories of genes through the lens of gene remodelling uncovers previously unseen or underappreciated trends, relying more on a network model as opposed to a simple tree model.

In Chapter 2, the methodology and network based detection mechanism of CompositeSearch were introduced. We benchmarked the functionality of CompositeSearch on a dataset of fungal genomes previously used to detect remodelling events (Leonard and Richards, 2012) where it was deemed to be a highly sensitive detection tool. Alignments between composite and component genes usually display fragmented homologies (Pathmanathan *et al.,* 2018), Type I errors in composite detection could easily arise due to poor gene calls or incorrect genome assemblages. As a precaution, we designed and statistically

modelled a quality control procedure to reduce the Type I error rate. It was observed that restricting a composite family to contain 2 or more *bona fide* composite genes and restricting component families to contain 2 or more members sufficed to remove additional Type I errors in 100 repeated controlled experiments. We performed a CompositeSearch analysis on 107 fungal genomes (1,150,918 genes). We found that 9.71% (111,768) genes could not be analysed due to low complexity or being identified as a singleton gene. It was observed that 49.94% of all genes (55.31% of all sampled genes) displayed a history of remodelling, of which 376,968 (32.76% of all genes and 36.28% of sampled genes) were identified as composites. We did not anticipate so many genes to display a history of remodelling based on previous studies (Nakamura *et al.*, 2007; Leonard and Richards, 2012). We plotted the presence and absence of remodelled gene families to an independently derived phylogeny and did not observe any bursts of evolution attributed to any branch for any remodelling category. These results suggest that successful remodelling events, remodelling events that persist post-speciation events, are clocklike in fungi, evolving at a stable pace. The observation of an profound increase in remodelling within nodes representing speciation events when compared to internal nodes further lend evidence this hypothesis. We observed that between 43.9% and 47.7% of families from remodelled categories (nested composites, strict composites, and strict component families) were homoplastic, which was in stark contrast to the 30.5% of homoplastic non-remodelled families. This rate of homoplasy suggests that convergent remodelling occurs at a relatively high rate, possibly as a rapid response to a stressor. Again, this was compounded when we observed composite genes to be functionally overrepresented for the production of secondary metabolites and small molecule transport. We also observed that remodelled gene families were enriched for a bacterial or undefined prokaryotic origin, whereas eukaryote specific genes families were more likely to be non-remodelled.

In chapter 3, we investigated the extent of gene remodelling in 50 Viridiplantae genomes (1,672,377 genes) using CompositeSearch. We observed a much higher rate of remodelling in plants when compared to fungi. In total, we observed nested composite, strict composite, strict component, and non-remodelled genes to account for 46.33%, 2.9%, 11.72% and 29.24% of the entire dataset, and 65.48%, 4.1%, 16.57%, and 13.86% of sampled genes. Like fungi, we observed plants to display a stable rate of remodelled gene acquisition and decay, where evolutionary bursts were only observed during speciation events where organisms have underwent significant chromosomal and genomic restructuring (for example, the speciation of the hexaploid *Triticum aestivum*). Again, with respect to the phylogeny, considerable differences were observed in the evolutionary rate of leaf nodes and internal branches, further compounding the hypothesis that remodelling events occur much more frequently during speciation events. We found the same categories of remodelled genes to be more homoplastic than in fungi. We observed remodelled categories to be 42-60.9% homoplastic (compared to 43.9-47.7%) and non-remodelled categories to be 35% homoplastic. We observed remodelled genes to be enriched for stress responses and metabolism as in fungi, but also for photosynthesis and multicellular development (*via* transcription factor remodelling). A considerable proportion of genes enriched for transcriptional regulation were found to be involved in multicellularity, and were found to emerge during major phenotypic transitions in plant evolutionary history. This highlights the evolutionary importance of gene remodelling in the evolutionary history of the green tree of life. We observed remodelled plant gene families to be enriched for a prokaryote origin, and eukaryote specific gene families to be non-remodelled, mirroring what we observed with fungi.

In Chapter 4, we developed compositeBLAST, a sensitive and selective tool for the detection of gene remodelling events. We found compositeBLAST to be a fast and intuitive tool. From a functional perspective, we found compositeBLAST to be more sensitive than

*fdf*BLAST, and were able to recover 10 times more composite genes than those reported by Leonard and Richards (2012) when we replicated our procedure on the dataset they used for composite detection in fungi. We report compositeBLAST to be more selective than CompositeSearch for the detection of polarised remodelling events, such as gene fusions and fissions due to the exclusion of multiple HSP derived homology and due to the requisite for the conservation of domains between composites and components which lend further robustness in our results. We detected two unreported classes of antimicrobial resistance genes as composites conferring resistance to mupirocin and rifamycin using CompositeBLAST. These results illustrate the importance of composite gene detection not just in macroevolutionary biology studies but also for observing trends in microevolution such as in the rapid emergence of AMR.

5.2. Future work

Evolutionary analyses have become more nuanced and informative as the number of high quality genome assemblies continues to expand. In our analyses of 107 fungi and 50 plants we observed some highly interesting results, however we feel that we just scratched the surface of the extent of gene remodelling. With a greater sample size of genomes that cover a broader range of taxa, more interesting and previously unreported phylogenomic trends may come to light. It would be interesting to observe the rate of remodelling in fungal and plant species that were previously unavailable to us either due to the genomes not being available or due to low quality assemblages. In particular, it would be interesting to observe the rate of remodelling in non-Dikarya species. Dikarya accounted for 101 of 107 of our species, so undoubtedly, results were biased due to the lack of available, reliable non-Dikarya genomes during the time of analysis. Similarly, in plants 38 of 50 were angiosperms, due to the lack of reliable, non-

angiosperm genomes. There are a wealth of genomes that we could have used for the plant analyses to help correct these biases, however, these were constructed from transcriptomes (Van Bel *et al.,* 2018) so they would not have been useful for the detection of remodelling genes due to Type I errors arising from alternative transcripts appearing as composites and components of each other. It would be of considerable interest to observe the rate of remodelling in other clades such as the Metazoa or 'protists', and within the Prokaryotes. Ultimately, a large study encompassing all domains of life would likely uncover considerable previously unseen evolutionary trends. Further development of compositeBLAST to encompass DNA alignments (BLASTN), and transcription alignments (BLASTX and tBLASTn) in addition to protein alignments would provide a tool for the detection of polarised composites in transcriptomes, untranslated genomes, and in entities that do not have a protein complement, for example viroids. We aim to implement GFF2/GFF3 parsing to decipher the location of genes within a genome to flag adjacent component genes for manual curation in an attempt to further reduce Type I errors. We aim to perform compositeBLAST on larger, more diverse datasets in future studies to uncover trends in polarised composite evolution throughout the domains of life.

5.3. Final remark

Gene remodelling is a relatively unexplored evolutionary concept. As the rate of genome sequencing and curation rises, the material to uncover remodelled gene evolutionary trends also grows. We must stop observing gene evolution through the lens of a vertically-exclusive model if we are to begin to understand the vast complexities observed throughout the Network of Life.

Bibliography

Adams, K. L. and Wendel, J. F. (2005) 'Polyploidy and genome evolution in plants', *Current Opinion in Plant Biology*. Elsevier Ltd, pp. 135–141. doi: 10.1016/j.pbi.2005.01.001.

Avelar, G. M. *et al.* (2014) 'A rhodopsin-guanylyl cyclase gene fusion functions in visual perception in a fungus.', *Current biology : CB*. Elsevier, 24(11), pp. 1234–40. doi: 10.1016/j.cub.2014.04.009.

Baldwin, S. J. and Husband, B. C. (2011) 'Genome duplication and the evolution of conspecific pollen precedence.', *Proceedings. Biological sciences*. The Royal Society, 278(1714), pp. 2011–7. doi: 10.1098/rspb.2010.2208.

Bateman, A. *et al.* (2017) 'UniProt: the universal protein knowledgebase', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Baysarowich, J. *et al.* (2008) 'Rifamycin antibiotic resistance by ADP-ribosylation: Structure and diversity of Arr.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(12), pp. 4886–91. doi: 10.1073/pnas.0711939105.

Beltrao, P. *et al.* (2013) 'Evolution and functional cross-talk of protein post-translational modifications', *Molecular Systems Biology*, 9(1), p. 714. doi: 10.1002/msb.201304521.

Bennici, A. (2008) 'Origin and early evolution of land plants: Problems and considerations.', *Communicative & integrative biology*. Taylor & Francis, 1(2), pp. 212–8.

Bidartondo, M. I. *et al.* (2011) 'The dawn of symbiosis between plants and fungi.', *Biology letters*. The Royal Society, 7(4), pp. 574–7. doi: 10.1098/rsbl.2010.1203.

Boehme, C. C. *et al.* (2010) 'Rapid Molecular Detection of Tuberculosis and Rifampin Resistance', *New England Journal of Medicine*, 363(11), pp. 1005–1015. doi: 10.1056/NEJMoa0907847.

BONFERRONI and C. (1936) 'Teoria statistica delle classi e calcolo delle probabilita',

Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8, pp. 3–62.

Box, G. E. P. and Cox, D. R. (1985) 'An Analysis of Transformations ({P}kg: P463-504)', in *The Collected Works of George E. P. Box: Volume II*. WileyRoyal Statistical Society, pp. 463–495. doi: 10.2307/2984418.

Braun, E. L. and Grotewold, E. (2001) 'Fungal Zuotin proteins evolved from MIDA1-like factors by lineage-specific loss of MYB domains.', *Molecular biology and evolution*, 18(7), pp. 1401–12.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Campbell, E. A. *et al.* (2001) 'Structural mechanism for rifampicin inhibition of bacterial rna polymerase.', *Cell*, 104(6), pp. 901–12. doi: 10.1016/S0092-8674(01)00286-0.

Castro, A. *et al.* (2005) 'The anaphase-promoting complex: a key factor in the regulation of cell cycle', *Oncogene*. Nature Publishing Group, 24(3), pp. 314–325. doi: 10.1038/sj.onc.1207973.

Causier, B. *et al.* (2005) *Evolution in Action: Following Function in Duplicated Floral Homeotic Genes*, *Current Biology*. doi: 10.1016/j.cub.2005.07.063.

Chen, Z. *et al.* (2014) 'Identification of two forms of the Eso1 protein in *Schizosaccharomyces pombe*', *Cell Biology International*, 38(5), pp. 682–688. doi: 10.1002/cbin.10230.

Cheng, Y. T. *et al.* (2011) 'Stability of plant immune-receptor resistance proteins is controlled by SKP1-CULLIN1-F-box (SCF)-mediated protein degradation', *Proceedings of the National Academy of Sciences*, 108(35), pp. 14694–14699. doi: 10.1073/pnas.1105685108.

Cookson, B. D. (1998) 'The emergence of mupirocin resistance: a challenge to infection

control and antibiotic prescribing practice.', *The Journal of antimicrobial chemotherapy*, 41(1), pp. 11–8. doi: 10.1093/jac/41.1.11.

Cory, J. G. (1983) 'Role of ribonucleotide reductase in cell division', *Pharmacology & Therapeutics*. Pergamon, 21(2), pp. 265–276. doi: 10.1016/0163-7258(83)90076-1.

Doerner, P. (2006) 'Plant Meristems: What You See Is What You Get?', *Current Biology*. Cell Press, 16(2), pp. R56–R58. doi: 10.1016/J.CUB.2006.01.001.

Dunn, J., Dunn, J. and Dunn, O. J. (1961) 'Multiple Comparisons Among Means', *AMERICAN STATISTICAL ASSOCIATION*, pp. 52--64. Available at: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.309.1277 (Accessed: 9 December 2019).

Dunn, O. J. (1959) 'Estimation of the Medians for Dependent Variables', *The Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 30(1), pp. 192–197. doi: 10.1214/aoms/1177706374.

Duplan, V. and Rivas, S. (2014) 'E3 ubiquitin-ligases and their target proteins during the regulation of plant innate immunity', *Frontiers in Plant Science*, 5, p. 42. doi: 10.3389/fpls.2014.00042.

Durrens, P., Nikolski, M. and Sherman, D. (2008) 'Fusion and Fission of Genes Define a Metric between Fungal Genomes', *PLoS Computational Biology*. Edited by A. J. Enright. Public Library of Science, 4(10), p. e1000200. doi: 10.1371/journal.pcbi.1000200.

Enright, A. J. *et al.* (1999) 'Protein interaction maps for complete genomes based on gene fusion events', *Nature*, 402(6757), pp. 86–90. doi: 10.1038/47056.

Enright, A. J. and Ouzounis, C. A. (2001) 'Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions.', *Genome biology*, 2(9), p. RESEARCH0034.

Feklistov, A. *et al.* (2008) 'Rifamycins do not function by allosteric modulation of binding of

Mg2+ to the RNA polymerase active center.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(39), pp. 14820–5. doi: 10.1073/pnas.0802822105.

Ferrari, S. *et al.* (2003) 'Tandemly duplicated Arabidopsis genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection.', *The Plant cell*, 15(1), pp. 93–106.

Ferrari, S. *et al.* (2006) 'Antisense Expression of the *Arabidopsis thaliana AtPGIP1* Gene Reduces Polygalacturonase-Inhibiting Protein Accumulation and Enhances Susceptibility to *Botrytis cinerea*', *Molecular Plant-Microbe Interactions*, 19(8), pp. 931–936. doi: 10.1094/MPMI-19-0931.

Fisher, R. A. (1922) 'On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P', *Journal of the Royal Statistical Society*, 85(1), p. 87. doi: 10.2307/2340521.

Fletcher, J. (2018) 'The CLV-WUS Stem Cell Signaling Pathway: A Roadmap to Crop Yield Optimization', *Plants*, 7(4), p. 87. doi: 10.3390/plants7040087.

Foltz, D. R. *et al.* (2006) 'The human CENP-A centromeric nucleosome-associated complex.', *Nature cell biology*, 8(5), pp. 458–69. doi: 10.1038/ncb1397.

Freeling, M. (2009) 'Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition.', *Annual review of plant biology*, 60(1), pp. 433–53. doi: 10.1146/annurev.arplant.043008.092122.

Furlan, G., Klinkenberg, J. and Trujillo, M. (2012) 'Regulation of plant immune receptors by ubiquitination.', *Frontiers in plant science*. Frontiers Media SA, 3, p. 238. doi: 10.3389/fpls.2012.00238.

Gilbart, J., Perry, C. R. and Slocombe, B. (1993) 'High-level mupirocin resistance in Staphylococcus aureus: evidence for two distinct isoleucyl-tRNA synthetases.',

Antimicrobial agents and chemotherapy, 37(1), pp. 32–8. doi: 10.1128/aac.37.1.32.

Godiard, L. *et al.* (2003) 'ERECTA, an LRR receptor-like kinase protein controlling development pleiotropically affects resistance to bacterial wilt.', *The Plant journal : for cell and molecular biology*, 36(3), pp. 353–65. doi: 10.1046/j.1365-313x.2003.01877.x.

Goldstein, B. P. (2014) 'Resistance to rifampicin: a review.', *The Journal of antibiotics*, 67(9), pp. 625–30. doi: 10.1038/ja.2014.107.

Groussin, M. *et al.* (2016) 'Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msv249.

Guarino, E., Salguero, I. and Kearsey, S. E. (2014) 'Cellular regulation of ribonucleotide reductase in eukaryotes', *Seminars in Cell & Developmental Biology*. Academic Press, 30, pp. 97–103. doi: 10.1016/J.SEMCDB.2014.03.030.

Hagel, J. M. and Facchini, P. J. (2017) 'Tying the knot: occurrence and possible significance of gene fusions in plant metabolism and beyond', *Journal of Experimental Botany*. Oxford University Press, 68(15), pp. 4029–4043. doi: 10.1093/jxb/erx152.

Haggerty, L. S. *et al.* (2014) 'A pluralistic account of homology: adapting the models to the data.', *Molecular biology and evolution*. Oxford University Press, 31(3), pp. 501–16. doi: 10.1093/molbev/mst228.

Harris, L. W. and Davies, T. J. (2016) 'A Complete Fossil-Calibrated Phylogeny of Seed Plant Families as a Tool for Comparative Analyses: Testing the "Time for Speciation" Hypothesis', *PLOS ONE*. Edited by S. Ho, 11(10), p. e0162907. doi: 10.1371/journal.pone.0162907.

Hartmann, T. (2004) 'Plant-derived secondary metabolites as defensive chemicals in herbivorous insects: A case study in chemical ecology', *Planta*, 219(1), pp. 1–4. doi: 10.1007/s00425-004-1249-y.

Hatakeyama, S. *et al.* (2001) 'U Box Proteins as a New Family of Ubiquitin-Protein Ligases', *Journal of Biological Chemistry*, 276(35), pp. 33111–33120. doi: 10.1074/jbc.M102755200.

He, X. and Zhang, J. (2005) 'Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution', *Genetics*, 169(2).

Hu, G. and Fearon, E. R. (1999) 'Siah-1 N-terminal RING domain is required for proteolysis function, and C-terminal sequences regulate oligomerization and binding to target proteins.', *Molecular and cellular biology*. American Society for Microbiology (ASM), 19(1), pp. 724–32.

Huibregtse, J. M. *et al.* (1995) 'A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 92(7), pp. 2563–7.

Ishikawa, J. *et al.* (2006) 'Contribution of rpoB2 RNA Polymerase Subunit Gene to Rifampin Resistance in Nocardia Species', *Antimicrobial Agents and Chemotherapy*, 50(4), pp. 1342–1346. doi: 10.1128/AAC.50.4.1342-1346.2006.

Jachiet, P.-A. *et al.* (2013) 'MosaicFinder: identification of fused gene families in sequence similarity networks', *Bioinformatics*, 29(7), pp. 837–844. doi: 10.1093/bioinformatics/btt049.

Jia, B. *et al.* (2017) 'CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database', *Nucleic Acids Research*, 45(D1), pp. D566–D573. doi: 10.1093/nar/gkw1004.

Jones, C. D. (2005) 'Origin and Evolution of a Chimeric Fusion Gene in Drosophila subobscura, D. madeirensis and D. guanche', *Genetics*, 170(1), pp. 207–219. doi: 10.1534/genetics.104.037283.

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.

Kanno, A. *et al.* (2007) 'Class B Gene Expression and the Modified ABC Model in Nongrass Monocots', *The Scientific World JOURNAL*, 7, pp. 268–279. doi: 10.1100/tsw.2007.86.

Kasarskis, A., Manova, K. and Anderson, K. V. (1998) 'A phenotype-based screen for embryonic lethal mutations in the mouse', *Proceedings of the National Academy of Sciences of the United States of America*, 95(13), pp. 7485–7490. doi: 10.1073/pnas.95.13.7485.

Kayes, J. M. and Clark, S. E. (1998) 'CLAVATA2, a regulator of meristem and organ development in Arabidopsis.', *Development (Cambridge, England)*, 125(19), pp. 3843–51. doi: 10.1242/dev.00998.

Kellis, M., Birren, B. W. and Lander, E. S. (2004) *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*. Available at: http://www.broad.mit.edu/seq/YeastDuplication (Accessed: 9 December 2019).

KRUIJT, M., DE KOCK, M. J. D. and DE WIT, P. J. G. M. (2005) 'Receptor-like proteins involved in plant disease resistance', *Molecular Plant Pathology*, 6(1), pp. 85–97. doi: 10.1111/j.1364-3703.2004.00264.x.

Kurland, C. G., Canback, B. and Berg, O. G. (2003) 'Horizontal gene transfer: a critical view.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(17), pp. 9658–62. doi: 10.1073/pnas.1632870100.

Lemus, D. *et al.* (no date) 'Rapid alternative methods for detection of rifampicin resistance in Mycobacterium tuberculosis', *academic.oup.com*.

Leonard, G. and Richards, T. A. (2012) 'Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life', *Proceedings of the National Academy of Sciences*, 109(52), pp. 21402–21407. doi: 10.1073/pnas.1210909110.

Lilliefors, H. W. (1967) 'On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown', *Journal of the American Statistical Association*, 62(318), pp. 399–402. doi: 10.1080/01621459.1967.10482916.

Liu, H., Cheng, E. H. Y. and Hsieh, J. J. D. (2009) 'MLL fusions: Pathways to leukemia', *Cancer Biology & Therapy*, 8(13), pp. 1204–1211. doi: 10.4161/cbt.8.13.8924.

Madlung, A. (2013) 'Polyploidy and its effect on evolutionary success: Old questions revisited with new tools', *Heredity*, pp. 99–104. doi: 10.1038/hdy.2012.79.

Mann, H. B. and Whitney, D. R. (1947) 'On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other', *The Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 18(1), pp. 50–60. doi: 10.1214/aoms/1177730491.

Des Marais, D. L. and Rausher, M. D. (2008) 'Escape from adaptive conflict after duplication in an anthocyanin pathway gene.', *Nature*, 454(7205), pp. 762–5. doi: 10.1038/nature07092.

Martin, F. M. and Simpson, T. J. (1989) 'Biosynthetic studies on pseudomonic acid (mupirocin), a novel antibiotic metabolite of Pseudomonas fluorescens', *Journal of the Chemical Society, Perkin Transactions 1*, (1), p. 207. doi: 10.1039/p19890000207.

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. American Society for Microbiology Journals, 1(5), pp. e00195-16. doi: 10.1128/mSphere.00195-16.

McLysaght, A. and Guerzoni, D. (2015) 'New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), p. 20140332. doi: 10.1098/rstb.2014.0332.

Metzger, M. B., Hristova, V. A. and Weissman, A. M. (2012) 'HECT and RING finger families of E3 ubiquitin ligases at a glance.', *Journal of cell science*. Company of Biologists, 125(Pt 3), pp. 531–7. doi: 10.1242/jcs.091777.

Miao, M. *et al.* (2016) 'The ubiquitin ligase SEVEN IN ABSENTIA (SINA) ubiquitinates a defense-related NAC transcription factor and is involved in defense signaling', *New*

Phytologist. John Wiley & Sons, Ltd (10.1111), 211(1), pp. 138–148. doi: 10.1111/nph.13890.

Miricescu, A., Goslin, K. and Graciet, E. (2018) 'Ubiquitylation in plants: signaling hub for the integration of environmental signals', *Journal of Experimental Botany*, 69(19), pp. 4511–4527. doi: 10.1093/jxb/ery165.

Mitelman, F., Johansson, B. and Mertens, F. (2007) 'The impact of translocations and gene fusions on cancer causation', *Nature Reviews Cancer*, 7(4), pp. 233–245. doi: 10.1038/nrc2091.

Morris, J. L. *et al.* (2018) 'The timescale of early land plant evolution.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(10), pp. E2274–E2283. doi: 10.1073/pnas.1719588115.

Nagy, A. and Patthy, L. (2011) 'Reassessing Domain Architecture Evolution of Metazoan Proteins: The Contribution of Different Evolutionary Mechanisms', *Genes*, 2(4), pp. 578–598. doi: 10.3390/genes2030578.

Nakamura, Y., Itoh, T. and Martin, W. (2007) 'Rate and Polarity of Gene Fusion and Fission in Oryza sativa and Arabidopsis thaliana', *Molecular Biology and Evolution*, 24(1), pp. 110–121. doi: 10.1093/molbev/msl138.

Nikolić, J., Perić, Z. and Marković, A. (2017) 'Proposal of Simple and Accurate Two-Parametric Approximation for the Q -Function', *Mathematical Problems in Engineering*. Hindawi Limited, 2017. doi: 10.1155/2017/8140487.

Ning, Y. *et al.* (2011) 'The SINA E3 Ligase OsDIS1 Negatively Regulates Drought Response in Rice', *Plant Physiology*. American Society of Plant Biologists, 157(1), pp. 242–255. doi: 10.1104/PP.111.180893.

PabÃ³n-Mora, N., Wong, G. K.-S. and Ambrose, B. A. (2014) 'Evolution of fruit development genes in flowering plants', *Frontiers in Plant Science*. Frontiers, 5, p. 300. doi:

10.3389/fpls.2014.00300.

Pan, L. *et al.* (2016) 'The Multifunction of CLAVATA2 in Plant Development and Immunity', *Frontiers in Plant Science*, 7, p. 1573. doi: 10.3389/fpls.2016.01573.

Pasek, S., Risler, J.-L. and Brezellec, P. (2006) 'Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins', *Bioinformatics*, 22(12), pp. 1418–1423. doi: 10.1093/bioinformatics/btl135.

Pathmanathan, J. S. *et al.* (2018) 'CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection', *Molecular Biology and Evolution*, 35(1), pp. 252–255. doi: 10.1093/molbev/msx283.

Pepper, I. J., Van Sciver, R. E. and Tang, A. H. (2017) 'Phylogenetic analysis of the SINA/SIAH ubiquitin E3 ligase family in Metazoa', *BMC Evolutionary Biology*, 17(1), p. 182. doi: 10.1186/s12862-017-1024-x.

Pola, E. *et al.* (2012) 'Medical and surgical treatment of pyogenic spondylodiscitis.', *European review for medical and pharmacological sciences*, 16 Suppl 2, pp. 35–49.

Prabakaran, S. *et al.* (2012) 'Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, pp. 565–583. doi: 10.1002/wsbm.1185.

Puigbò, P., Wolf, Y. I. and Koonin, E. V. (2010) 'The Tree and Net Components of Prokaryote Evolution', *Genome Biology and Evolution*, 2, pp. 745–756. doi: 10.1093/gbe/evq062.

Puigbò, P., Wolf, Y. I. and Koonin, E. V (2009) 'Search for a "Tree of Life" in the thicket of the phylogenetic forest', *Journal of Biology*, 8(6), p. 59. doi: 10.1186/jbiol159.

Qi, H. *et al.* (2017) 'TRAF Family Proteins Regulate Autophagy Dynamics by Modulating AUTOPHAGY PROTEIN6 Stability in Arabidopsis.', *The Plant cell*. American Society of Plant Biologists, 29(4), pp. 890–911. doi: 10.1105/tpc.17.00056.

Rastogi, S. and Liberles, D. A. (2005) ‘Subfunctionalization of duplicated genes as a transition state to neofunctionalization’, *BMC Evolutionary Biology*. BioMed Central, 5(1), p. 28. doi: 10.1186/1471-2148-5-28.

Retallack, G. J. and Landing, E. (2014) ‘Affinities and architecture of Devonian trunks of Prototaxites loganii’, *Mycologia*. Allen Press Inc., 106(6), pp. 1143–1158. doi: 10.3852/13-390.

Richards, T. A. *et al.* (2006) ‘Evolutionary Origins of the Eukaryotic Shikimate Pathway: Gene Fusions, Horizontal Gene Transfer, and Endosymbiotic Replacements’, *Eukaryotic Cell*, 5(9), pp. 1517–1531. doi: 10.1128/EC.00106-06.

Rojo, E. *et al.* (2002) ‘CLV3 is localized to the extracellular space, where it activates the Arabidopsis CLAVATA stem cell signaling pathway.’, *The Plant cell*, 14(5), pp. 969–77. doi: 10.1105/tpc.002196.

Sadowski, M. and Sarcevic, B. (2010) ‘Mechanisms of mono- and poly-ubiquitination: Ubiquitination specificity depends on compatibility between the E2 catalytic core and amino acid residues proximal to the lysine’, *Cell Division*. BioMed Central, 5(1), p. 19. doi: 10.1186/1747-1028-5-19.

Semon, M. and Wolfe, K. H. (2008) ‘Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis’, *Proceedings of the National Academy of Sciences*, 105(24), pp. 8333–8338. doi: 10.1073/pnas.0708705105.

Sensi, P. (1983) ‘History of the development of rifampin.’, *Reviews of infectious diseases*, 5 Suppl 3, pp. S402-6.

Serrano, I., Campos, L. and Rivas, S. (2018) ‘Roles of E3 Ubiquitin-Ligases in Nuclear Protein Homeostasis during Plant Stress Responses.’, *Frontiers in plant science*. Frontiers Media SA, 9, p. 139. doi: 10.3389/fpls.2018.00139.

Snel, B., Bork, P. and Huynen, M. (2000) ‘Genome evolution. Gene fusion versus gene

fission.', *Trends in genetics : TIG*, 16(1), pp. 9–11.

Spanu, P. D. *et al.* (2010) 'Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 330(6010), pp. 1543–6. doi: 10.1126/science.1194573.

Sparkes, I. A. *et al.* (2003) 'An Arabidopsis pex10 Null Mutant Is Embryo Lethal, Implicating Peroxisomes in an Essential Role during Plant Embryogenesis', *Plant Physiology*, 133(4), pp. 1809–1819. doi: 10.1104/pp.103.031252.

Spearman, C. (1904) 'The Proof and Measurement of Association between Two Things', *The American Journal of Psychology*, 15(1), p. 72. doi: 10.2307/1412159.

Stransky, N. *et al.* (2014) 'The landscape of kinase fusions in cancer', *Nature Communications*, 5, p. 4846. doi: 10.1038/ncomms5846.

Student (1908) 'The Probable Error of a Mean', *Biometrika*, 6(1), p. 1. doi: 10.2307/2331554.

T, F. (2010) 'Evolution. A fungal past to insect color', *Science*, 328(5978), pp. 574–5. doi: 10.1126/science.1190417.

Torti, S. *et al.* (2012) 'Analysis of the Arabidopsis Shoot Meristem Transcriptome during Floral Transition Identifies Distinct Regulatory Patterns and a Leucine-Rich Repeat Protein That Promotes Flowering', *THE PLANT CELL ONLINE*, 24(2), pp. 444–462. doi: 10.1105/tpc.111.092791.

Trimpalis, P. *et al.* (2013) 'Gene fusion analysis in the battle against the African endemic sleeping sickness.', *PloS one*. Public Library of Science, 8(7), p. e68854. doi: 10.1371/journal.pone.0068854.

Troeman, D. P. R., Van Hout, D. and Kluytmans, J. A. J. W. (2019) 'Antimicrobial approaches in the prevention of Staphylococcus aureus infections: a review.', *The Journal of antimicrobial chemotherapy*, 74(2), pp. 281–294. doi: 10.1093/jac/dky421.

Vibranovski, M. D., Sakabe, N. J. and de Souza, S. J. (2006) ‘A possible role of exon-shuffling in the evolution of signal peptides of human proteins’, *FEBS Letters*, 580(6), pp. 1621–1624. doi: 10.1016/j.febslet.2006.01.094.

Vogel, C., Teichmann, S. A. and Pereira-Leal, J. (2005) ‘The Relationship Between Domain Duplication and Recombination’, *Journal of Molecular Biology*, 346(1), pp. 355–365. doi: 10.1016/j.jmb.2004.11.050.

de Vries, J. *et al.* (2016) ‘Streptophyte Terrestrialization in Light of Plastid Evolution’, *Trends in Plant Science*, 21(6), pp. 467–476. doi: 10.1016/j.tplants.2016.01.021.

de Vries, J. and Archibald, J. M. (2018) ‘Plant evolution: landmarks on the path to terrestrial life’, *New Phytologist*, 217(4), pp. 1428–1434. doi: 10.1111/nph.14975.

Wan, T. *et al.* (2018) ‘A genome for gnetophytes and early evolution of seed plants’, *Nature Plants*. Palgrave Macmillan Ltd., 4(2), pp. 82–89. doi: 10.1038/s41477-017-0097-2.

Wang, C. and Liu, Z. (2006) ‘Arabidopsis ribonucleotide reductases are critical for cell cycle progression, DNA damage repair, and plant development.’, *The Plant cell*. American Society of Plant Biologists, 18(2), pp. 350–65. doi: 10.1105/tpc.105.037044.

Wang, G. *et al.* (2008) ‘A genome-wide functional investigation into the roles of receptor-like proteins in Arabidopsis.’, *Plant physiology*. American Society of Plant Biologists, 147(2), pp. 503–17. doi: 10.1104/pp.108.119487.

Wang, M. *et al.* (2008) ‘Genome-wide analysis of SINA family in plants and their phylogenetic relationships †’, *DNA Sequence*. Taylor & Francis, 19(3), pp. 206–216. doi: 10.1080/10425170701517317.

Wang, W. *et al.* (2000) ‘The Origin of the Jingwei Gene and the Complex Modular Structure of Its Parental Gene, Yellow Emperor, in Drosophila melanogaster’, *Molecular Biology and Evolution*, 17(9), pp. 1294–1301. doi: 10.1093/oxfordjournals.molbev.a026413.

WELCH, B. L. (1947) ‘THE GENERALIZATION OF “STUDENT’S” PROBLEM WHEN

SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED', *Biometrika*, 34(1–2), pp. 28–35. doi: 10.1093/biomet/34.1-2.28.

Wilkinson, K. D. (1999) 'Ubiquitin-Dependent Signaling: The Role of Ubiquitination in the Response of Cells to Their Environment.', *The Journal of Nutrition*, 129(11), pp. 1933–1936. doi: 10.1093/jn/129.11.1933.

Williams, G. J. *et al.* (2005) 'Structure and function of both domains of ArnA, a dual function decarboxylase and a formyltransferase, involved in 4-amino-4-deoxy-L-arabinose biosynthesis.', *The Journal of biological chemistry*. Europe PMC Funders, 280(24), pp. 23000–8. doi: 10.1074/jbc.M501534200.

Wink, M. (2003) 'Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective', *Phytochemistry*. Elsevier Ltd, pp. 3–19. doi: 10.1016/S0031-9422(03)00300-5.

Wink, M. (2018) 'Plant Secondary Metabolites Modulate Insect Behavior-Steps Toward Addiction?', *Frontiers in Physiology*, 9. doi: 10.3389/fphys.2018.00364.

Yanai, I., Wolf, Y. I. and Koonin, E. V (2002) 'Evolution of gene fusions: horizontal transfer versus independent events.', *Genome biology*, 3(5), p. research0024.

Zhang, W. *et al.* (2013) 'Arabidopsis receptor-like protein30 and receptor-like kinase suppressor of BIR1-1/EVERSHED mediate innate immunity to necrotrophic fungi.', *The Plant cell*, 25(10), pp. 4227–41. doi: 10.1105/tpc.113.117010.

Adams, K. L. and Wendel, J. F. (2005) 'Polyploidy and genome evolution in plants', *Current Opinion in Plant Biology*. Elsevier Ltd, pp. 135–141. doi: 10.1016/j.pbi.2005.01.001.

Avelar, G. M. *et al.* (2014) 'A rhodopsin-guanylyl cyclase gene fusion functions in visual perception in a fungus.', *Current biology : CB*. Elsevier, 24(11), pp. 1234–40. doi:

10.1016/j.cub.2014.04.009.

Baldwin, S. J. and Husband, B. C. (2011) 'Genome duplication and the evolution of conspecific pollen precedence.', *Proceedings. Biological sciences*. The Royal Society, 278(1714), pp. 2011–7. doi: 10.1098/rspb.2010.2208.

Bateman, A. *et al.* (2017) 'UniProt: the universal protein knowledgebase', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Baysarowich, J. *et al.* (2008) 'Rifamycin antibiotic resistance by ADP-ribosylation: Structure and diversity of Arr.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(12), pp. 4886–91. doi: 10.1073/pnas.0711939105.

Beltrao, P. *et al.* (2013) 'Evolution and functional cross-talk of protein post-translational modifications', *Molecular Systems Biology*, 9(1), p. 714. doi: 10.1002/msb.201304521.

Bennici, A. (2008) 'Origin and early evolution of land plants: Problems and considerations.', *Communicative & integrative biology*. Taylor & Francis, 1(2), pp. 212–8.

Bidartondo, M. I. *et al.* (2011) 'The dawn of symbiosis between plants and fungi.', *Biology letters*. The Royal Society, 7(4), pp. 574–7. doi: 10.1098/rsbl.2010.1203.

Boehme, C. C. *et al.* (2010) 'Rapid Molecular Detection of Tuberculosis and Rifampin Resistance', *New England Journal of Medicine*, 363(11), pp. 1005–1015. doi: 10.1056/NEJMoa0907847.

BONFERRONI and C. (1936) 'Teoria statistica delle classi e calcolo delle probabilita', *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8, pp. 3–62.

Box, G. E. P. and Cox, D. R. (1985) 'An Analysis of Transformations ({P}kg: P463-504)', in *The Collected Works of George E. P. Box: Volume II*. WileyRoyal Statistical Society, pp. 463–495. doi: 10.2307/2984418.

Braun, E. L. and Grotewold, E. (2001) 'Fungal Zuotin proteins evolved from MIDA1-like

factors by lineage-specific loss of MYB domains.', *Molecular biology and evolution*, 18(7), pp. 1401–12.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Campbell, E. A. *et al.* (2001) 'Structural mechanism for rifampicin inhibition of bacterial rna polymerase.', *Cell*, 104(6), pp. 901–12. doi: 10.1016/S0092-8674(01)00286-0.

Castro, A. *et al.* (2005) 'The anaphase-promoting complex: a key factor in the regulation of cell cycle', *Oncogene*. Nature Publishing Group, 24(3), pp. 314–325. doi: 10.1038/sj.onc.1207973.

Causier, B. *et al.* (2005) *Evolution in Action: Following Function in Duplicated Floral Homeotic Genes*, *Current Biology*. doi: 10.1016/j.cub.2005.07.063.

Chen, Z. *et al.* (2014) 'Identification of two forms of the Eso1 protein in *Schizosaccharomyces pombe*', *Cell Biology International*, 38(5), pp. 682–688. doi: 10.1002/cbin.10230.

Cheng, Y. T. *et al.* (2011) 'Stability of plant immune-receptor resistance proteins is controlled by SKP1-CULLIN1-F-box (SCF)-mediated protein degradation', *Proceedings of the National Academy of Sciences*, 108(35), pp. 14694–14699. doi: 10.1073/pnas.1105685108.

Cookson, B. D. (1998) 'The emergence of mupirocin resistance: a challenge to infection control and antibiotic prescribing practice.', *The Journal of antimicrobial chemotherapy*, 41(1), pp. 11–8. doi: 10.1093/jac/41.1.11.

Cory, J. G. (1983) 'Role of ribonucleotide reductase in cell division', *Pharmacology & Therapeutics*. Pergamon, 21(2), pp. 265–276. doi: 10.1016/0163-7258(83)90076-1.

Doerner, P. (2006) 'Plant Meristems: What You See Is What You Get?', *Current Biology*. Cell Press, 16(2), pp. R56–R58. doi: 10.1016/J.CUB.2006.01.001.

Dunn, J., Dunn, J. and Dunn, O. J. (1961) ‘Multiple Comparisons Among Means’, *AMERICAN STATISTICAL ASSOCIATION*, pp. 52--64. Available at: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.309.1277 (Accessed: 9 December 2019).

Dunn, O. J. (1959) ‘Estimation of the Medians for Dependent Variables’, *The Annals of Mathematical Statistics*. Institute of Mathematical Statistics, 30(1), pp. 192–197. doi: 10.1214/aoms/1177706374.

Duplan, V. and Rivas, S. (2014) ‘E3 ubiquitin-ligases and their target proteins during the regulation of plant innate immunity’, *Frontiers in Plant Science*, 5, p. 42. doi: 10.3389/fpls.2014.00042.

Durrens, P., Nikolski, M. and Sherman, D. (2008) ‘Fusion and Fission of Genes Define a Metric between Fungal Genomes’, *PLoS Computational Biology*. Edited by A. J. Enright. Public Library of Science, 4(10), p. e1000200. doi: 10.1371/journal.pcbi.1000200.

Enright, A. J. *et al.* (1999) ‘Protein interaction maps for complete genomes based on gene fusion events’, *Nature*, 402(6757), pp. 86–90. doi: 10.1038/47056.

Enright, A. J. and Ouzounis, C. A. (2001) ‘Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions.’, *Genome biology*, 2(9), p. RESEARCH0034.

Feklistov, A. *et al.* (2008) ‘Rifamycins do not function by allosteric modulation of binding of Mg2+ to the RNA polymerase active center.’, *Proceedings of the National Academy of Sciences of the United States of America*, 105(39), pp. 14820–5. doi: 10.1073/pnas.0802822105.

Ferrari, S. *et al.* (2003) ‘Tandemly duplicated Arabidopsis genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection.’, *The Plant cell*, 15(1), pp. 93–106.

Ferrari, S. *et al.* (2006) 'Antisense Expression of the *Arabidopsis thaliana AtPGIP1* Gene Reduces Polygalacturonase-Inhibiting Protein Accumulation and Enhances Susceptibility to *Botrytis cinerea*', *Molecular Plant-Microbe Interactions*, 19(8), pp. 931–936. doi: 10.1094/MPMI-19-0931.

Fisher, R. A. (1922) 'On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P', *Journal of the Royal Statistical Society*, 85(1), p. 87. doi: 10.2307/2340521.

Fletcher, J. (2018) 'The CLV-WUS Stem Cell Signaling Pathway: A Roadmap to Crop Yield Optimization', *Plants*, 7(4), p. 87. doi: 10.3390/plants7040087.

Foltz, D. R. *et al.* (2006) 'The human CENP-A centromeric nucleosome-associated complex.', *Nature cell biology*, 8(5), pp. 458–69. doi: 10.1038/ncb1397.

Freeling, M. (2009) 'Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition.', *Annual review of plant biology*, 60(1), pp. 433–53. doi: 10.1146/annurev.arplant.043008.092122.

Furlan, G., Klinkenberg, J. and Trujillo, M. (2012) 'Regulation of plant immune receptors by ubiquitination.', *Frontiers in plant science*. Frontiers Media SA, 3, p. 238. doi: 10.3389/fpls.2012.00238.

Gilbart, J., Perry, C. R. and Slocombe, B. (1993) 'High-level mupirocin resistance in Staphylococcus aureus: evidence for two distinct isoleucyl-tRNA synthetases.', *Antimicrobial agents and chemotherapy*, 37(1), pp. 32–8. doi: 10.1128/aac.37.1.32.

Godiard, L. *et al.* (2003) 'ERECTA, an LRR receptor-like kinase protein controlling development pleiotropically affects resistance to bacterial wilt.', *The Plant journal : for cell and molecular biology*, 36(3), pp. 353–65. doi: 10.1046/j.1365-313x.2003.01877.x.

Goldstein, B. P. (2014) 'Resistance to rifampicin: a review.', *The Journal of antibiotics*, 67(9), pp. 625–30. doi: 10.1038/ja.2014.107.

Groussin, M. *et al.* (2016) 'Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msv249.

Guarino, E., Salguero, I. and Kearsey, S. E. (2014) 'Cellular regulation of ribonucleotide reductase in eukaryotes', *Seminars in Cell & Developmental Biology*. Academic Press, 30, pp. 97–103. doi: 10.1016/J.SEMCDB.2014.03.030.

Hagel, J. M. and Facchini, P. J. (2017) 'Tying the knot: occurrence and possible significance of gene fusions in plant metabolism and beyond', *Journal of Experimental Botany*. Oxford University Press, 68(15), pp. 4029–4043. doi: 10.1093/jxb/erx152.

Haggerty, L. S. *et al.* (2014) 'A pluralistic account of homology: adapting the models to the data.', *Molecular biology and evolution*. Oxford University Press, 31(3), pp. 501–16. doi: 10.1093/molbev/mst228.

Harris, L. W. and Davies, T. J. (2016) 'A Complete Fossil-Calibrated Phylogeny of Seed Plant Families as a Tool for Comparative Analyses: Testing the "Time for Speciation" Hypothesis', *PLOS ONE*. Edited by S. Ho, 11(10), p. e0162907. doi: 10.1371/journal.pone.0162907.

Hartmann, T. (2004) 'Plant-derived secondary metabolites as defensive chemicals in herbivorous insects: A case study in chemical ecology', *Planta*, 219(1), pp. 1–4. doi: 10.1007/s00425-004-1249-y.

Hatakeyama, S. *et al.* (2001) 'U Box Proteins as a New Family of Ubiquitin-Protein Ligases', *Journal of Biological Chemistry*, 276(35), pp. 33111–33120. doi: 10.1074/jbc.M102755200.

He, X. and Zhang, J. (2005) 'Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution', *Genetics*, 169(2).

Hu, G. and Fearon, E. R. (1999) 'Siah-1 N-terminal RING domain is required for proteolysis function, and C-terminal sequences regulate oligomerization and binding to target proteins.',

Molecular and cellular biology. American Society for Microbiology (ASM), 19(1), pp. 724–32.

Huibregtse, J. M. *et al.* (1995) 'A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 92(7), pp. 2563–7.

Ishikawa, J. *et al.* (2006) 'Contribution of rpoB2 RNA Polymerase Subunit Gene to Rifampin Resistance in Nocardia Species', *Antimicrobial Agents and Chemotherapy*, 50(4), pp. 1342–1346. doi: 10.1128/AAC.50.4.1342-1346.2006.

Jachiet, P.-A. *et al.* (2013) 'MosaicFinder: identification of fused gene families in sequence similarity networks', *Bioinformatics*, 29(7), pp. 837–844. doi: 10.1093/bioinformatics/btt049.

Jia, B. *et al.* (2017) 'CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database', *Nucleic Acids Research*, 45(D1), pp. D566–D573. doi: 10.1093/nar/gkw1004.

Jones, C. D. (2005) 'Origin and Evolution of a Chimeric Fusion Gene in Drosophila subobscura, D. madeirensis and D. guanche', *Genetics*, 170(1), pp. 207–219. doi: 10.1534/genetics.104.037283.

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.

Kanno, A. *et al.* (2007) 'Class B Gene Expression and the Modified ABC Model in Nongrass Monocots', *The Scientific World JOURNAL*, 7, pp. 268–279. doi: 10.1100/tsw.2007.86.

Kasarskis, A., Manova, K. and Anderson, K. V. (1998) 'A phenotype-based screen for embryonic lethal mutations in the mouse', *Proceedings of the National Academy of Sciences of the United States of America*, 95(13), pp. 7485–7490. doi: 10.1073/pnas.95.13.7485.

Kayes, J. M. and Clark, S. E. (1998) 'CLAVATA2, a regulator of meristem and organ

development in Arabidopsis.', *Development (Cambridge, England)*, 125(19), pp. 3843–51. doi: 10.1242/dev.00998.

Kellis, M., Birren, B. W. and Lander, E. S. (2004) *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*. Available at: http://www.broad.mit.edu/seq/YeastDuplication (Accessed: 9 December 2019).

KRUIJT, M., DE KOCK, M. J. D. and DE WIT, P. J. G. M. (2005) 'Receptor-like proteins involved in plant disease resistance', *Molecular Plant Pathology*, 6(1), pp. 85–97. doi: 10.1111/j.1364-3703.2004.00264.x.

Kurland, C. G., Canback, B. and Berg, O. G. (2003) 'Horizontal gene transfer: a critical view.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 100(17), pp. 9658–62. doi: 10.1073/pnas.1632870100.

Lemus, D. *et al.* (no date) 'Rapid alternative methods for detection of rifampicin resistance in Mycobacterium tuberculosis', *academic.oup.com*.

Leonard, G. and Richards, T. A. (2012) 'Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life', *Proceedings of the National Academy of Sciences*, 109(52), pp. 21402–21407. doi: 10.1073/pnas.1210909110.

Lilliefors, H. W. (1967) 'On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown', *Journal of the American Statistical Association*, 62(318), pp. 399–402. doi: 10.1080/01621459.1967.10482916.

Liu, H., Cheng, E. H. Y. and Hsieh, J. J. D. (2009) 'MLL fusions: Pathways to leukemia', *Cancer Biology & Therapy*, 8(13), pp. 1204–1211. doi: 10.4161/cbt.8.13.8924.

Madlung, A. (2013) 'Polyploidy and its effect on evolutionary success: Old questions revisited with new tools', *Heredity*, pp. 99–104. doi: 10.1038/hdy.2012.79.

Mann, H. B. and Whitney, D. R. (1947) 'On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other', *The Annals of Mathematical Statistics*.

Institute of Mathematical Statistics, 18(1), pp. 50–60. doi: 10.1214/aoms/1177730491.

Des Marais, D. L. and Rausher, M. D. (2008) 'Escape from adaptive conflict after duplication in an anthocyanin pathway gene.', *Nature*, 454(7205), pp. 762–5. doi: 10.1038/nature07092.

Martin, F. M. and Simpson, T. J. (1989) 'Biosynthetic studies on pseudomonic acid (mupirocin), a novel antibiotic metabolite of Pseudomonas fluorescens', *Journal of the Chemical Society, Perkin Transactions 1*, (1), p. 207. doi: 10.1039/p19890000207.

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. American Society for Microbiology Journals, 1(5), pp. e00195-16. doi: 10.1128/mSphere.00195-16.

McLysaght, A. and Guerzoni, D. (2015) 'New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), p. 20140332. doi: 10.1098/rstb.2014.0332.

Metzger, M. B., Hristova, V. A. and Weissman, A. M. (2012) 'HECT and RING finger families of E3 ubiquitin ligases at a glance.', *Journal of cell science*. Company of Biologists, 125(Pt 3), pp. 531–7. doi: 10.1242/jcs.091777.

Miao, M. *et al.* (2016) 'The ubiquitin ligase SEVEN IN ABSENTIA (SINA) ubiquitinates a defense-related NAC transcription factor and is involved in defense signaling', *New Phytologist*. John Wiley & Sons, Ltd (10.1111), 211(1), pp. 138–148. doi: 10.1111/nph.13890.

Miricescu, A., Goslin, K. and Graciet, E. (2018) 'Ubiquitylation in plants: signaling hub for the integration of environmental signals', *Journal of Experimental Botany*, 69(19), pp. 4511–4527. doi: 10.1093/jxb/ery165.

Mitelman, F., Johansson, B. and Mertens, F. (2007) 'The impact of translocations and gene

fusions on cancer causation', *Nature Reviews Cancer*, 7(4), pp. 233–245. doi: 10.1038/nrc2091.

Morris, J. L. *et al.* (2018) 'The timescale of early land plant evolution.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(10), pp. E2274–E2283. doi: 10.1073/pnas.1719588115.

Nagy, A. and Patthy, L. (2011) 'Reassessing Domain Architecture Evolution of Metazoan Proteins: The Contribution of Different Evolutionary Mechanisms', *Genes*, 2(4), pp. 578–598. doi: 10.3390/genes2030578.

Nakamura, Y., Itoh, T. and Martin, W. (2007) 'Rate and Polarity of Gene Fusion and Fission in Oryza sativa and Arabidopsis thaliana', *Molecular Biology and Evolution*, 24(1), pp. 110–121. doi: 10.1093/molbev/msl138.

Nikolić, J., Perić, Z. and Marković, A. (2017) 'Proposal of Simple and Accurate Two-Parametric Approximation for the Q -Function', *Mathematical Problems in Engineering*. Hindawi Limited, 2017. doi: 10.1155/2017/8140487.

Ning, Y. *et al.* (2011) 'The SINA E3 Ligase OsDIS1 Negatively Regulates Drought Response in Rice', *Plant Physiology*. American Society of Plant Biologists, 157(1), pp. 242–255. doi: 10.1104/PP.111.180893.

PabÃ³n-Mora, N., Wong, G. K.-S. and Ambrose, B. A. (2014) 'Evolution of fruit development genes in flowering plants', *Frontiers in Plant Science*. Frontiers, 5, p. 300. doi: 10.3389/fpls.2014.00300.

Pan, L. *et al.* (2016) 'The Multifunction of CLAVATA2 in Plant Development and Immunity', *Frontiers in Plant Science*, 7, p. 1573. doi: 10.3389/fpls.2016.01573.

Pasek, S., Risler, J.-L. and Brezellec, P. (2006) 'Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins', *Bioinformatics*, 22(12), pp. 1418–1423. doi: 10.1093/bioinformatics/btl135.

Pathmanathan, J. S. *et al.* (2018) 'CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection', *Molecular Biology and Evolution*, 35(1), pp. 252–255. doi: 10.1093/molbev/msx283.

Pepper, I. J., Van Sciver, R. E. and Tang, A. H. (2017) 'Phylogenetic analysis of the SINA/SIAH ubiquitin E3 ligase family in Metazoa', *BMC Evolutionary Biology*, 17(1), p. 182. doi: 10.1186/s12862-017-1024-x.

Pola, E. *et al.* (2012) 'Medical and surgical treatment of pyogenic spondylodiscitis.', *European review for medical and pharmacological sciences*, 16 Suppl 2, pp. 35–49.

Prabakaran, S. *et al.* (2012) 'Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, pp. 565–583. doi: 10.1002/wsbm.1185.

Puigbò, P., Wolf, Y. I. and Koonin, E. V. (2010) 'The Tree and Net Components of Prokaryote Evolution', *Genome Biology and Evolution*, 2, pp. 745–756. doi: 10.1093/gbe/evq062.

Puigbò, P., Wolf, Y. I. and Koonin, E. V (2009) 'Search for a "Tree of Life" in the thicket of the phylogenetic forest', *Journal of Biology*, 8(6), p. 59. doi: 10.1186/jbiol159.

Qi, H. *et al.* (2017) 'TRAF Family Proteins Regulate Autophagy Dynamics by Modulating AUTOPHAGY PROTEIN6 Stability in Arabidopsis.', *The Plant cell*. American Society of Plant Biologists, 29(4), pp. 890–911. doi: 10.1105/tpc.17.00056.

Rastogi, S. and Liberles, D. A. (2005) 'Subfunctionalization of duplicated genes as a transition state to neofunctionalization', *BMC Evolutionary Biology*. BioMed Central, 5(1), p. 28. doi: 10.1186/1471-2148-5-28.

Retallack, G. J. and Landing, E. (2014) 'Affinities and architecture of Devonian trunks of Prototaxites loganii', *Mycologia*. Allen Press Inc., 106(6), pp. 1143–1158. doi: 10.3852/13-390.

Richards, T. A. *et al.* (2006) 'Evolutionary Origins of the Eukaryotic Shikimate Pathway: Gene Fusions, Horizontal Gene Transfer, and Endosymbiotic Replacements', *Eukaryotic Cell*, 5(9), pp. 1517–1531. doi: 10.1128/EC.00106-06.

Rojo, E. *et al.* (2002) 'CLV3 is localized to the extracellular space, where it activates the Arabidopsis CLAVATA stem cell signaling pathway.', *The Plant cell*, 14(5), pp. 969–77. doi: 10.1105/tpc.002196.

Sadowski, M. and Sarcevic, B. (2010) 'Mechanisms of mono- and poly-ubiquitination: Ubiquitination specificity depends on compatibility between the E2 catalytic core and amino acid residues proximal to the lysine', *Cell Division*. BioMed Central, 5(1), p. 19. doi: 10.1186/1747-1028-5-19.

Semon, M. and Wolfe, K. H. (2008) 'Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis', *Proceedings of the National Academy of Sciences*, 105(24), pp. 8333–8338. doi: 10.1073/pnas.0708705105.

Sensi, P. (1983) 'History of the development of rifampin.', *Reviews of infectious diseases*, 5 Suppl 3, pp. S402-6.

Serrano, I., Campos, L. and Rivas, S. (2018) 'Roles of E3 Ubiquitin-Ligases in Nuclear Protein Homeostasis during Plant Stress Responses.', *Frontiers in plant science*. Frontiers Media SA, 9, p. 139. doi: 10.3389/fpls.2018.00139.

Snel, B., Bork, P. and Huynen, M. (2000) 'Genome evolution. Gene fusion versus gene fission.', *Trends in genetics : TIG*, 16(1), pp. 9–11.

Spanu, P. D. *et al.* (2010) 'Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 330(6010), pp. 1543–6. doi: 10.1126/science.1194573.

Sparkes, I. A. *et al.* (2003) 'An Arabidopsis pex10 Null Mutant Is Embryo Lethal, Implicating Peroxisomes in an Essential Role during Plant Embryogenesis', *Plant*

Physiology, 133(4), pp. 1809–1819. doi: 10.1104/pp.103.031252.

Spearman, C. (1904) 'The Proof and Measurement of Association between Two Things', *The American Journal of Psychology*, 15(1), p. 72. doi: 10.2307/1412159.

Stransky, N. *et al.* (2014) 'The landscape of kinase fusions in cancer', *Nature Communications*, 5, p. 4846. doi: 10.1038/ncomms5846.

Student (1908) 'The Probable Error of a Mean', *Biometrika*, 6(1), p. 1. doi: 10.2307/2331554.

T, F. (2010) 'Evolution. A fungal past to insect color', *Science*, 328(5978), pp. 574–5. doi: 10.1126/science.1190417.

Torti, S. *et al.* (2012) 'Analysis of the Arabidopsis Shoot Meristem Transcriptome during Floral Transition Identifies Distinct Regulatory Patterns and a Leucine-Rich Repeat Protein That Promotes Flowering', *THE PLANT CELL ONLINE*, 24(2), pp. 444–462. doi: 10.1105/tpc.111.092791.

Trimpalis, P. *et al.* (2013) 'Gene fusion analysis in the battle against the African endemic sleeping sickness.', *PloS one*. Public Library of Science, 8(7), p. e68854. doi: 10.1371/journal.pone.0068854.

Troeman, D. P. R., Van Hout, D. and Kluytmans, J. A. J. W. (2019) 'Antimicrobial approaches in the prevention of Staphylococcus aureus infections: a review.', *The Journal of antimicrobial chemotherapy*, 74(2), pp. 281–294. doi: 10.1093/jac/dky421.

Vibranovski, M. D., Sakabe, N. J. and de Souza, S. J. (2006) 'A possible role of exon-shuffling in the evolution of signal peptides of human proteins', *FEBS Letters*, 580(6), pp. 1621–1624. doi: 10.1016/j.febslet.2006.01.094.

Vogel, C., Teichmann, S. A. and Pereira-Leal, J. (2005) 'The Relationship Between Domain Duplication and Recombination', *Journal of Molecular Biology*, 346(1), pp. 355–365. doi: 10.1016/j.jmb.2004.11.050.

de Vries, J. *et al.* (2016) 'Streptophyte Terrestrialization in Light of Plastid Evolution', *Trends in Plant Science*, 21(6), pp. 467–476. doi: 10.1016/j.tplants.2016.01.021.

de Vries, J. and Archibald, J. M. (2018) 'Plant evolution: landmarks on the path to terrestrial life', *New Phytologist*, 217(4), pp. 1428–1434. doi: 10.1111/nph.14975.

Wan, T. *et al.* (2018) 'A genome for gnetophytes and early evolution of seed plants', *Nature Plants*. Palgrave Macmillan Ltd., 4(2), pp. 82–89. doi: 10.1038/s41477-017-0097-2.

Wang, C. and Liu, Z. (2006) 'Arabidopsis ribonucleotide reductases are critical for cell cycle progression, DNA damage repair, and plant development.', *The Plant cell*. American Society of Plant Biologists, 18(2), pp. 350–65. doi: 10.1105/tpc.105.037044.

Wang, G. *et al.* (2008) 'A genome-wide functional investigation into the roles of receptor-like proteins in Arabidopsis.', *Plant physiology*. American Society of Plant Biologists, 147(2), pp. 503–17. doi: 10.1104/pp.108.119487.

Wang, M. *et al.* (2008) 'Genome-wide analysis of SINA family in plants and their phylogenetic relationships †', *DNA Sequence*. Taylor & Francis, 19(3), pp. 206–216. doi: 10.1080/10425170701517317.

Wang, W. *et al.* (2000) 'The Origin of the Jingwei Gene and the Complex Modular Structure of Its Parental Gene, Yellow Emperor, in Drosophila melanogaster', *Molecular Biology and Evolution*, 17(9), pp. 1294–1301. doi: 10.1093/oxfordjournals.molbev.a026413.

WELCH, B. L. (1947) 'THE GENERALIZATION OF "STUDENT'S" PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED', *Biometrika*, 34(1–2), pp. 28–35. doi: 10.1093/biomet/34.1-2.28.

Wilkinson, K. D. (1999) 'Ubiquitin-Dependent Signaling: The Role of Ubiquitination in the Response of Cells to Their Environment.', *The Journal of Nutrition*, 129(11), pp. 1933–1936. doi: 10.1093/jn/129.11.1933.

Williams, G. J. *et al.* (2005) 'Structure and function of both domains of ArnA, a dual function

decarboxylase and a formyltransferase, involved in 4-amino-4-deoxy-L-arabinose biosynthesis.', *The Journal of biological chemistry*. Europe PMC Funders, 280(24), pp. 23000–8. doi: 10.1074/jbc.M501534200.

Wink, M. (2003) 'Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective', *Phytochemistry*. Elsevier Ltd, pp. 3–19. doi: 10.1016/S0031-9422(03)00300-5.

Wink, M. (2018) 'Plant Secondary Metabolites Modulate Insect Behavior-Steps Toward Addiction?', *Frontiers in Physiology*, 9. doi: 10.3389/fphys.2018.00364.

Yanai, I., Wolf, Y. I. and Koonin, E. V (2002) 'Evolution of gene fusions: horizontal transfer versus independent events.', *Genome biology*, 3(5), p. research0024.

Zhang, W. *et al.* (2013) 'Arabidopsis receptor-like protein30 and receptor-like kinase suppressor of BIR1-1/EVERSHED mediate innate immunity to necrotrophic fungi.', *The Plant cell*, 25(10), pp. 4227–41. doi: 10.1105/tpc.113.117010.