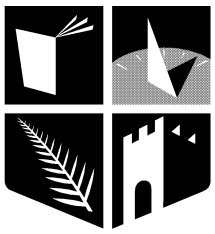
De novo sequencing, assembly and analysis of the genome and transcriptome of the nematode *Panagrolaimus superbus*

by

Georgina O'Mahony Zamora, BSc.



NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Dissertation submitted in partial fulfillment of the requirements
for candidate for the degree of

Doctor of Philosophy

Department of Biology,
National University of Ireland Maynooth, Co. Kildare, Ireland.

July, 2013

Head of Department: Prof. Paul Moynagh
Supervisors: Prof. Ann Burnell and Dr. Simon Wong

This thesis is dedicated to Mum and Dad

For their endless love, support and encouragement

Contents

Page

1 Introduction **1**

1.1 Nematodes 1

1.1.1 The *Panagrolaimus superbus* Nematode 4

1.1.2 Environmental Stress Tolerance in Nematodes 11

1.1.3 Freezing Tolerance 11

1.1.4 Anhydrobiosis and Desiccation 12

1.2 Eukaryote Genomes 16

1.2.1 Introduction and Overview 16

1.2.2 Prokaryote and Eukaryote Genome Organisation 19

1.2.3 Eukaryote Gene Structure 22

1.2.4 Content Comparisons of Selected Model Eukaryote Genomes 24

1.2.5 Nematode Nuclear Genomes 27

1.3 Mitochondrial Genomes of Nematodes and Other Animals 30

1.4 Genome Sequencing 33

1.5 Transcriptome Sequencing 38

1.5.1 High Throughput Transcriptome Sequencing 38

1.5.2 Transcriptome Assembly 39

1.6 Aims and Objectives of this Project 43

2 Expressed Sequence Tags **44**

2.1 Introduction 44

2.2 Methods & Materials 46

2.2.1 Nematode Culture 46

2.2.2 cDNA Library Construction and EST Generation 46

2.2.3 Clustering and Sequence Analysis 47

2.2.4 Translation and Primary Structure Analysis of Novel ESTs . 49

2.2.5 Real-time Relative qPCR Analysis of Gene Expression . . . 50

2.3 Results 53

2.3.1 Functional Annotation using BLAST2GO 53

2.3.2 Most Abundant Contigs 54

2.3.3 Assignments to Metabolic Pathways using KEGG 61

2.3.4 Gene Ontology Assignments 63

2.3.5 Putative Anhdyrobiotic and Stress Response Genes 67

2.3.6 Signal Transduction, Protein Kinases and Transcription Factors 67

2.3.7 Anti-oxidant Activity 69

2.3.8 Late Embryogenesis Abundant Proteins 70

2.3.9 Molecular Chaperones & Unfolded Protein Response 73

2.3.10 Removal of Damaged Proteins - the Ubiquitin - Proteasome (UPS) and Autophagy Systems . 74

2.3.11 A Comparison of the *P. superbus* EST Unigene Dataset with EST Datasets from other Anhydrobiotic Nematodes 74

2.3.12 Analysis of Novel Transcripts 78

2.3.13 Expression of Putative Stress Related Genes upon Desiccation 82

2.4 Discussion 86

3 Transcriptome Assembly **89**

3.1 Introduction 89

3.2 Materials & Methods 93

3.2.1 Growth Conditions, Stress Treatments and RNA Extraction 93

3.2.2 cDNA Synthesis and Normalisation 95

3.2.3 454 Titanium Pyrosequencing 96

3.2.4 Sequence Assembly 96

3.2.5 BLAST Homology Searches 101

3.3 Results 102

3.3.1 General Assembly Statistics 102

3.3.2 Similarity Searches to Establish Quality of the *P. superbus* Datasets 113

3.3.3 Homologues to other Nematode Genes and Transcripts . . . 116

3.3.4 Quality Evaluation of *P. superbus* Assemblies by Assessing 5' to 3' Coverage of Individual CEG Genes 120

3.3.5 Quality Evaluation of *P. superbus* Assemblies by Identifying Under-assembled Contigs and Chimeras 126

3.3.6 Quality Evaluation of *P. superbus* Assemblies by Assessing Length Coverage of Individual *C. elegans* Genes 131

3.3.7 Ranking *P. superbus* Transcriptome Assemblies using Unweighted Quality Criteria 137

3.4 Discussion 140

4 Transcriptome Annotation **144**

4.1 Introduction 144

4.2 Methods & Materials 145

4.2.1 Functional Annotation using the BLAST2GO Tool 145

4.3 Results 147

4.3.1 Summary . . . 147
4.3.2 BLAST Results . . . 148
4.3.3 InterProScan . . . 157
4.3.4 Gene Ontology Annotations . . . 158
4.3.5 EC and KEGG Annotations . . . 173
4.4 Discussion . . . 178
4.4.1 Heat Shock Proteins . . . 178
4.4.2 Removal of Damaged Proteins - The Ubiquitin-proteasome (UPS) and Autophagy Systems . . . 179
4.4.3 DNA Damage Response Proteins . . . 179
4.4.4 Signal Transduction, Protein Kinases and Transcription Factors . . . 180
4.4.5 Other Putative Anhydrobiotic Genes . . . 180

5 Assembling the Nuclear Genome of *P. superbus* **182**
5.1 Introduction . . . 182
5.2 Materials & Methods . . . 190
5.2.1 Chemicals . . . 190
5.2.2 Nematode Collection and Care . . . 190
5.2.3 gDNA Extractions . . . 191
5.2.4 Karyotyping . . . 193
5.3 Results . . . 194
5.3.1 Preparation of *P. superbus* gDNA for High-Throughput Sequencing . . . 194
5.3.2 Genome Assembly and Statistics . . . 200
5.3.3 Further Analysis of 454 gDNA . . . 206
5.3.4 The Karyotypes of *P. superbus* and *P. davidi* . . . 211
5.4 Discussion . . . 220

5.4.1 Genome 220
5.4.2 Karyotyping 223

6 The Mitochondrial Genome of *Panagrolaimus superbus* 224
6.1 Introduction 224
6.1.1 Structure and Functions of the Mitochondria 225
6.1.2 Mitochondrial Genome Structure 226
6.1.3 Why have Mitochondria Retained a mtDNA Genome? 229
6.1.4 Nematode Mitochondrial Genomes 230
6.2 Materials & Methods 232
6.2.1 Traditional Approach (PCR Amplification of mtDNA Gene Fragments) 232
6.2.2 Agarose Gel Electrophoresis 236
6.2.3 Computational Approach 236
6.3 Results 240
6.3.1 PCR Amplification of mtDNA Fragments 240
6.4 Discussion 249

7 General Discussion 252
7.1 Discussion 252
7.2 Future Work 257

Declaration

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

Signed _____

Georgina O'Mahony Zamora

Abbreviations

18S	18 Svedberg
28S	28 Svedberg
3'	3 Prime
5'	5 Prime
60S	60 Svedberg
A	Adenine
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
Bps	Basic pairs
°C	Degree Celcius
C	Cytosine
cDNA	Complementary DNA
CEG	Core Eukaryotic Genes
CEGMA	Core Eukaryotic Genes Mapping Approach
CK2	Casein kinase 2
cm	Centimetre
COB	Cytochrome b
COI	Cytochrome c oxidase subunit 1
COX	Cytochrome c oxidase
Cys	Cytosine
DAG	Diacetylglycerol
dbEST	Database od Expressed Sequence Tags
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
ds	Double stranded
DSBs	Double Strand DNA Breaks

DSN	Duplex-specific nuclease
e-value	Expectation value
EC	Enzyme Commission
EMBL	European molecular biology laboratory
ER	Endoplasmic Reticulum
ERM	Ezrin, Radixin and Muesin
EST	Expressed sequence tag
G	Guanine
GC	Guanine Cytosine
GRX	Glutaredoxin
GSH	Glutathione
GST	Glutathione S-transferase
gDNA	Genomic deoxyribonucleic acid
GO	Gene Ontology
GPX	Glutathione Peroxidase
GRAVY	Grand average of hydropathicity
HGT	Horizontal Gene Transfer
HMG	High mobility group
HSP	Heat shock protein
hsp	Minimal alignment length
ICDH	Isocitrate dehydrogenase
ID	Identification number
IDP	Intrinsically disordered proteins
IGF	Insulin-like growth factor
IUBMB	International Union of Biochemistry and Molecular Biology
IUP	Intrinsically unfolded protein
IUPred	Intrinsically unstructured/disordered proteins

JmjC	Jumonji domain-containing
KAAS	Kgg Automatic Annotation Server
Kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
kms	Kilometres
KOG	Eukaryotic orthologous groups
L	Litre
LDPCR	Long-distance Polymerase chain reaction
LEA	Late embryogenesis abundant
LSU	Large subunit
MAP	Mitogen-activated protein
MAPK	Mitogen-activated protein kinase
Mb	Mega bases
Mbp	Mega base pairs
MCL	Monte Carlo Localisation
ml	Millilitre
ML	Maximum likelihood
MP	Maximum parsimony
MSP	Major sperm protein
mtDNA	Mitochondrial Deoxyribonucleic acid
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate-oxidase
NCBI	National Center for Biotechnology Information
ng	Nanogram
NGM	Next Generation Multiplex
NGM	Nutrient Growth Media
NJ	Neighbour joining

NR	Non redundant
NT	Nucleotide
ORF	Open Reading Frame
PCR	Polymerase chain reaction
PDI	Protein Disulfide Isomerases
Prx	Peroxi Redoxins
QPCR	Quantitative Polymerase chain reaction
rDNA	Recombinant DNA
RET	Rearranged during transfection
ROS	Reactive oxygen species
RH	Relative Humidity
RNA	Reoxyribonucleic acid
RNAi	RNA interference
RNase	Ribonuclease
RPM	Revolutions per minute
rRNA	Ribosomal reoxyribonucleic acid
rrNL	Large ribonucleic acid subunit
rrNS	Small ribonucleic acid subunit
SAPKs	Stress-activated protein kinases
SEM	Scanning electron microscope
SPF	Sun protection factor
T	Thymine
TAE	Tris-acetate-EDTA
TGS	Third generation sequencing
Tris	Tris (hydroxy-methyl) amino methane
tRNA	Transfer Reoxyribonucleic acid
ttr	transthyretin-related

UBS	Ubiquitin binding surface
UPR	Unfolded Protein Response
UPS	Ubquitin Proteasome System
UTRs	Untranslated regions
μ**g**	Microgram
μ**L**	Microlitre

List of Figures

Page

1.1 A phylogeny of the Phylum *Nematoda* based on maximum parsimony analysis of the small subunit ribosomal DNA sequence 3

1.2 A *P. superbus* female. 5

1.3 *P. superbus* head region with six distinct lips used for morphological characterisation. 7

1.4 The nuclear rRNA gene phylogeny for *Panagrolaimus*. 9

1.5 Possible steps involved in the detection and expression of anhydrobiotic protection mechanisms in nematodes. 15

1.6 Known genome size ranges for extant life forms on Earth. 18

1.7 A *C. elegans* operon. 20

1.8 Prokaryotic and eukaryotic gene structures. 23

1.9 The number of Open Reading Frames in each genome versus genome size for the three extant domains of life. 25

1.10 Expression profile of operon genes during the life cycle of *C. elegans* 29

1.11 *C.elegans* mtDNA genome structure and gene order. 31

1.12 General structure of a tRNA. 32

1.13 Steps in completion of the annotation of genome data. 36

1.14 Trends in generation of genomic data. 37

2.1 Sequence counts vs sequence length in base pairs. 47
2.2 Overview of the BLAST2GO results distribution for the *P. superbus* EST sequences. 54
2.3 Species distribution from BLAST hits to the *P. superbus* EST dataset following a BLASTX to NR with a cutoff of $1e^{-10}$. 55
2.4 The distribution of e-value hits to the *P. superbus* EST dataset shown by number of hits for that e-value. 56
2.5 Biological process categories for the *P. superbus* unigenes having at least 100 EST sequences in each category. 64
2.6 Molecular function categories for the *P. superbus* unigenes having at least 100 EST sequences in each category. 65
2.7 Cellular component categories for the *P. superbus* unigenes having at least 100 EST sequences in each category. 66
2.8 Number of unigene families that contain representatives from one or more anhydrobiotic nematodes. 77
2.9 A plot of the hydropathy value (GRAVY Index). 80
2.10 Plot of the putative glycine content and the hydropathy of the protein sequences. 81
2.11 Real-Time Relative qPCR analysis. 84

3.1 Steps used in the generation of hybrid pipeline showing CAP3 and Phrap assemblies. 100
3.2 The mean size of the contigs generated 108
3.3 The total size of the *P. superbus* transcriptome generated by different assembly programs 110
3.4 Transcriptome Assemblers compared on the basis of the N50, N90 and N95 lengths 111
3.5 The percentage coverage across all the *C. elegans* CEG set 127

3.6 Under-assembled contigs 129
3.7 Chimeric contigs 130
3.8 Variable width box plots showing gene-coverage 133
3.9 Venn diagram showing the relationships between *C. elegans* genes which had homologous sequences in up to five *P. superbus* transcriptome assemblies 134
3.10 Variable width boxplots showing the percentage coverage of *C. elegans* by *P. superbus* homologues based on the number of *P. superbus* transcriptome assemblies that that contain the gene. 136
3.11 Boxplots showing the percentage coverage of *C. elegans* by *P. superbus* homologs when the *P. superbus* genes are found by only one assembler 137

4.1 BLAST2GO annotation pipeline beginning with sequences in FASTA format and resulting with an annotated dataset. 146
4.2 Number of sequences in the *P. superbus* transcriptome assembly in proportion to length of sequence. 147
4.3 Data distribution post BLAST2GO annotation 148
4.4 e-value distribution of the *P. superbus* transcriptome sequences which returned BLAST hits following BLASTX against NR with a cut off of 1e-3. 149
4.5 Species distribution of all BLAST hits following BLASTX against the NR database with a cut off of 1e-3. 150
4.6 Species distribution of the top BLAST hits returned following BLASTX against the NR database with a cut off of 1e-3. 152
4.7 Sequence similarity distribution of the *P. superbus* transcriptome sequences to the sequences of top BLAST hits following BLASTN against the NCBI NR database with a cutoff of 1e-3. 153

4.8 Evidence code distribution for the *P. superbus* transcriptome BLAST hits. 155
4.9 Databases accessed to find annotations for *P. superbus* transcriptome. 156
4.10 Summary of the InterProScan hits for the *P. superbus* transcriptome. 158
4.11 The GO annotation distribution for the *P. superbus* transcriptome showing the number of annotations found versus the count of sequences for that number. 160
4.12 Percentage of *P. superbus* sequences versus sequence length that were annotated using GO. 161
4.13 *P. superbus* transcriptome dataset distribution showing the abundance of terms for each vocabulary at each level. 162
4.14 Overview of the GO molecular function annotation for the *P. superbus* transcriptome dataset. 163
4.15 Molecular function categories identified for the *P. superbus* transcriptome dataset at level 3. 164
4.16 The 10 most common molecular functions identified 166
4.17 Overview of the cellular component annotations for the *P. superbus* transcriptome dataset in a combined graph 167
4.18 Cellular component categories identified at level 6 for the *P. superbus* transcriptome. 168
4.19 The 10 most common cellular components identified in the GO analysis of the *P. superbus* transcriptome. 169
4.20 Overview of the Gene ontology biological processes annotations for the *P. superbus* transcriptome 171
4.21 The 10 most common biological processes identified in the GO analysis of the *P. superbus* transcriptome. 172

4.22 Purine metabolism pathway with EC numbers found in the annotation of the *P. superbus* dataset shown in colour. 175
4.23 Oxidative Phosphorylation pathway with EC numbers found in the annotation of the *P. superbus* dataset shown in colour. 176

5.1 Suggested exploration avenues in next generation genomic sequencing (Edwards et al., 2013). 183
5.2 454 Titanium workflow for genomic sequencing focusing on crop genomes (DNA Sequencing Core, 2013). 185
5.3 Solexa Illumina workflow for genomic sequencing 186
5.4 *P. superbus* gDNA sample for 454 sequencing (pre RNAse treatment). 195
5.5 *P. superbus* gDNA sample for 454 sequencing (post RNAse treatment). 195
5.6 *P. superbus* gDNA sample for 454 sequencing (post pooling). Letters indicate a $\frac{1}{10}$ dilution of the previously labelled sample. 196
5.7 An absorbance spectrum for a *P. superbus* gDNA sample obtained using a Qubit 2.0 Fluorometer 197
5.8 *P. superbus* gDNA sample for 50bp Illumina sequencing (pre RNAse treatment). 199
5.9 *P. superbus* gDNA sample for 100bp Illumina sequencing (post RNAse treatment). 199
5.10 454 gDNA assembly showing biological process categories breakdown. 208
5.11 454 gDNA assembly showing molecular function categories breakdown. 209
5.12 454 gDNA assembly using cellular component categories breakdown. 210
5.13 *C. elegans* egg cell following DAPI staining as seen thorough a confocal microscope 212
5.14 *P. davidi* oocyte following DAPI staining as seen through a confocal microscope 213

5.15 *P. davidi* egg cell following DAPI staining as seen through a confocal microscope 214
5.16 *P. superbus* sperm cell following DAPI staining as seen through a confocal microscope 216
5.17 *P. superbus* egg cell following DAPI staining as seen through a confocal microscope 217
5.18 *P. superbus* egg cell following DAPI staining as seen thorough a confocal microscope 218
5.19 *P. superbus* oocyte cell following DAPI staining as seen through a confocal microscope 219

6.1 The sizes of mtDNA genomes compared with an α-Proteobacterial (*Rickettsia*) genome 227
6.2 Universal nematode primers shown to amplify the whole genome in two fragments (Hu, 2002). 233
6.3 PCR product (∼1,300bp) obtained from *P. superbus* using the primer pairs AATP6F/ACOBR and confirming the ATP6/COB border . . 241
6.4 PCR product (∼600bp) obtained from *P. superbus* using the primer pairs CX2F/RLR and confirming COXII/rrnL gene border. 242
6.5 PCR product (∼500bp) obtained from *P. superbus* using the primer pairs ACX3F/AND4R and confirming ND4/COXIII gene border. . 242
6.6 PCR product (∼800bp) obtained from *P. superbus* using the primer pairs BND1F/BATP6R and confirming ND1/ATP6 gene border. . . 243
6.7 PCR product (∼850bp) obtained from *P. superbus* using the primer pairs CYTB2F/CX3R and confirming COB and COXIII gene border.243
6.8 PCR product (∼200bp) obtained from *P. superbus* using the primer pairs CYTBF/CYTBR and confirming the COB gene presence. . . 244

6.9 PCR product (∼800bp) obtained from *P. superbus* using the primer pairs CX1F/CX2R2 and confirming the COXI and COXII gene border. 244
6.10 PCR product (∼1,000bp) obtained from *P. superbus* using the primer pairs ND2F2/CY2R and confirming the NADH2 and COB gene border. 245
6.11 Nematode mtDNA molecule showing the locations of the successful PCR amplifications for the *P. superbus* mtDNA genome. 246
6.12 *P. superbus* mtDNA fragment showing the overlapping sequences from various datasets used in its assembly 247
6.13 The mtDNA genome of *P. superbus* 248
6.14 The breakdown of gene orders from nematode mitochondrial genomes 250

List of Tables

Page

1.1 Nematode Nuclear Genome statistics 21
1.2 Eukaryotic genome statistics. 24

2.1 The primer sequences used for real time qPCR. 52
2.2 Summary of the analysis of EST sequences. 53
2.3 Most abundantly represented transcripts. 59
2.4 Putative anhdyrobiotic and stress response genes. 60
2.5 Summary of KEGG orthology assignments. 62
2.6 BLASTX similarity searches of *P. superbus*. 71

3.1 Stress states and their corresponding environmental values 94
3.2 Assemblers used and parameters executed. 98
3.3 Post-filtering read statistics for the two *P. superbus* cDNA libraries obtained by 454 Titanium FLX pyrosequencing. 99
3.4 Databases used for homology searching. 101
3.5 Assembly statistics (a). 104
3.6 Assembly statistics (b) Newbler assemblies 105
3.7 Assembly statistics (c) Newbler assemblies continued. w indicates with URT wo indicates without URT. 106

3.8 Assembly statistics (d). Latest Newbler release (Feb 2013). w indicates with URT wo indicates without URT. 107
3.9 Transcriptome assemblies compared on the basis of N50 values. . . . 112
3.10 BLAST searches carried out using the CEG genes from six organisms 115
3.11 BLAST searches carried out using NemBase4+ sequences 118
3.12 TBLASTN searches against other nematode transcriptomes (a) . . 121
3.13 BLASTN searches against other nematode transcriptomes (b) . . . 122
3.14 BLASTX searches against other nematode transcriptomes (c) . . . 123
3.15 BLASTX searches against other nematode transcriptomes (d) . . . 124
3.16 The number of putative under-assembled and chimeric contigs in different *P. superbus* transcriptome assemblies detected in reciprocal best BLAST analyses of the *C. elegans* transcriptome. 132
3.17 Assembly metrics, quality assessment and ranked scores for five assemblies 139

4.1 Stress response genes found in the transcriptome of *P. superbus*. . . 177

5.1 Genome sample 454 gDNA which was sent for sequencing by 454 Roche Titanium platform and assembled using the Newbler 2.3 assembler. 200
5.2 First Velvet assembly of 50bp Solexa Illumina genome sequences. . . 200
5.3 Second Velvet assembly of 50bp Solexa Illumina genome sequences following removal of putative bacterial sequences from the Solexa reads. 201
5.4 Velvet assembly of the 100bp Solexa Illumina genome sequences. . . 202
5.5 CLCBio assembly of the 50 and 100bp Solexa Illumina genome sequences. 202

5.6 Hybrid assembly, using Velvet, of the 454 genome reads, 50bps Solexa Reads and 100bp Solexa Illumina reads post-filtering to remove for bacterial contaminants. 203
5.7 CLCBio assembly of the 50 and 100bp Solexa Illumina reads(using the sequences as non paired end reads) and the 454 Titanium gDNA reads. 203
5.8 CLCBio assembly of the 50 and 100bp Solexa Illumina (using the sequences as paired end reads) and the 454 Titanium gDNA reads. 204
5.9 Comparison of nuclear genome assemblies on the basis of number of contigs, number of base pairs and N50 205
5.10 UniRef90 BLASTX matches for the 20 longest contigs. 207
5.11 Published nematode genomes. 222

6.1 Nematode mitochondrial 'universal primers' tested in this study. . . 232
6.2 mtDNA primers designed using EST sequences. 234
6.3 Reagent quantities required when using GoTaq (Promega) and Platinum Taq (Invitrogen) to amplify *P. superbus* mtDNA gene fragments using PCR. 235
6.4 PCR cycling conditions for GoTaq (Promega) and Platinum Taq (Invitrogen) when amplifying *P. superbus* mtDNA gene fragments. . 236
6.5 mtDNA gene ID's. 237
6.6 Nematode mitochondrial genomes used in this study 239
6.7 The PCR primer pair combinations which were successfully used to amplify *P. superbus* mtDNA fragements. 240

Abstract

The nematode *Panagrolaimus superbus* can survive for extended periods of time in a desiccated state (anhydrobiosis) and is also freezing tolerant (cryobiotic). These adaptations make it an interesting candidate for genome and transcriptome sequencing using second generation high throughput methods. In this project the transcriptome of *P. superbus* was sequenced using the 454 (Roche) platform. To enrich for stress-related genes, nematodes were exposed to one of the following stresses (desiccation, cold, heat or oxidation). Equal numbers of nematodes from each stress treatment were combined with unstressed control nematodes prior to RNA extraction. Normalised and unnormalised cDNA libraries were prepared from this mixed population. A *de novo* assembly of the transcriptome was generated using a variety of assembly programs and strategies. A Sanger sequenced expressed sequence dataset comprising 3,982 unigenes was fully annotated and integrated into the *de novo* transcriptome assembly. The *de novo* assembly has also been annotated and putative stress response genes were identified. The haploid karyotype of *P. superbus* was determined to be n=4. *P. superbus* genomic DNA was sequenced using 454 (Roche) methods along with 50 bp and 100 bp paired end Illumina reads. Eight different gDNA assemblies were prepared, generating predicted genome sizes ranging from 87.9 kilobases to 159.7 kilobases. The longest contigs were obtained from the 454 genomic DNA assembly and the assemblies of the Illumina reads generated shorter contigs. The gene order of the *P. superbus* mitochondrial DNA genome was obtained and a draft assembly of the mitochondrial genome is presented. The current transcriptome assembly is a resource suitable for use as a reference for aligning high throughput RNA Seq reads. Both the transcriptome and genome assemblies can be used to generate a protein reference database for the mass spectrometry based identification of the proteome of control and desiccated *P. superbus* for future studies.

Chapter 1

Introduction

1.1 Nematodes

Krogh's Principle states that "For a large number of problems there will be some animal of choice, or a few such animals, on which it can be most conveniently studied" (Krogh, 1929). This is indeed true of the nematode *Caenorhabditis elegans*, a cosmopolitan animal that lives in soil in virtually every part of the world. The nematode has been used to study neurological disorders, congenital heart disease, kidney disease and diabetes due to the large number of genes it has with functional counterparts in humans. Nematodes are even believed to provide insights into mechanisms for counteracting the effects of ageing (Kaletta & Hengartner, 2006).

The word "Nematoda" means "the thread-like ones", and comes from the Ancient Greek words *nema* ("thread") and *-ode*, ("like") (Chitwood, 1957). The group was originally established by Karl Rudolphi under the name *Nematoidea* (Rudolphi, 1810), but reclassified as family Nematodes by Burmeister in 1837 (Burmeister, 1837). They were eventually renamed *ordo Nematoda* by K. M. Diesing (Diesing, 1860). Nathan Cobb later argued that they should be called

Nemata or *Nemates*, or the English version "nemas " rather than "nematodes " (Cobb, 1919), but Diesing's revision had been established.

The phylum *Nematoda* is an exceptionally diverse ancient phylum with over a million species, many of which have not been classified (Wasmuth et al., 2008). They are said to be the most species-rich multicellular phylum on the planet. Using phylogenetic techniques the phylum has been separated into three major classes: (*Dorylaimia, Enoplia*, and *Chromadoria*) and various different clades (Clade A: *Plectida* and *Rhabditida*, Clade B: Clade A plus *Axonolaimidae, Desmolaimus zeelandicus* and *Isolaimium sp.*, and Clade C: Clade B plus *Desmodoridae* and *Monhysterida* (including *Comesomatidae*)) (Meldal et al., 2007). The nematode phylogenetic tree shows that parasitic nematodes have evolved independently on many occasions. They occur in various clades, sharing these clades with a multitude of parasitic nematodes (Blaxter et al., 1998; Dorris et al., 1999) as illustrated in Figure 1.1.

Nematodes can be categorised into two feeding types:

- free-living, by which it is meant that the nematode feeds on bacteria, fungi or is carnivorous,
- parasitic nematodes, which infect plants and animals.

Given the importance of parasitic nematodes in agriculture and human health, much of the research done on nematodes has gone into investigating parasitic nematodes. Free-living nematodes have a place in the earth's ecosystem as decomposers and predators and are typically smaller than parasitic worms, with some just a few millimeters long. It is the aim of this project to show how the free-living nematode *Panagrolaimus superbus* could be used as a suitable model organism in the study of anhydrobiotic and cold tolerance genes.

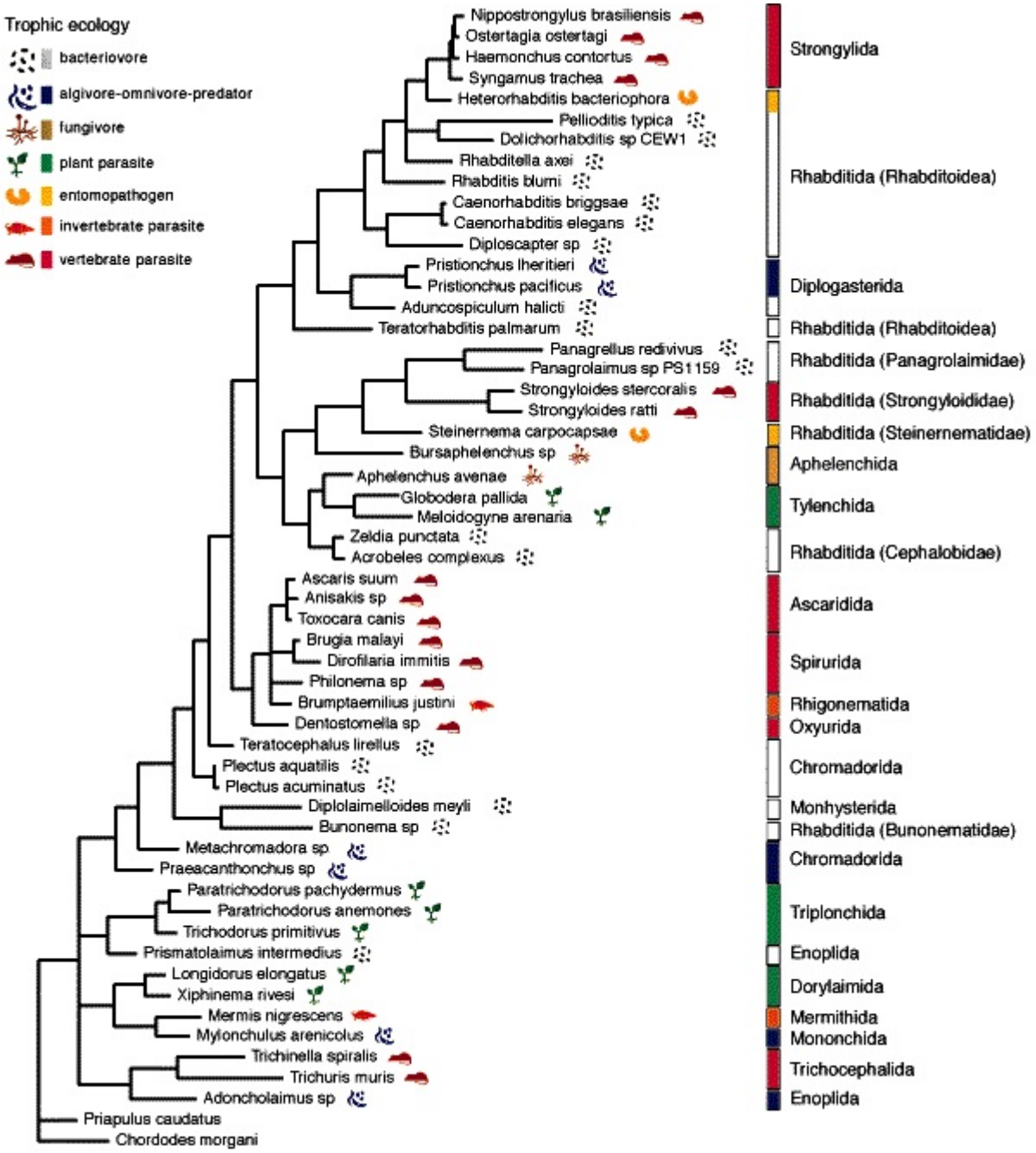


Figure 1.1: A phylogeny of the Phylum *Nematoda* based on maximum parsimony analysis of the small subunit ribosomal DNA sequence Reproduced with permission from (Blaxter et al., 1998).

1.1.1 The *Panagrolaimus superbus* Nematode

P. superbus (strain DF5050) was first isolated on the volcanic island of Surtsey, 30 kilometers off the coast of Iceland (Sohlenius, 1972). Surtsey was formed from a volcanic eruption in 1963 and has been studied extensively over the past 50 years by scientists interested in primary succession. Emil Olaisson isolated *P. superbus* on the island in July 1981, where the organism was found in a small nest of the hybrid gulls *Larus fuscus* and *Larus argentatus*. The nest was found in a small crevice in the lava and filled with *Rhacomitrium sp.* moss. It is presumed that the gulls arrived on the island from mainland Iceland. Lava rock would not be thought to contain very substantial growing conditions, but the gull waste produces nitrogen which would make the environment suitable for protozoa and nematodes.

After five years it was discovered that two independent species were growing along side each other on the agar culture plates and from this the species were separated and new cultures were started from a single gravid female, as depicted in Figure 1.2, by Bjorn Sohlenius (Bostrom, 1988).

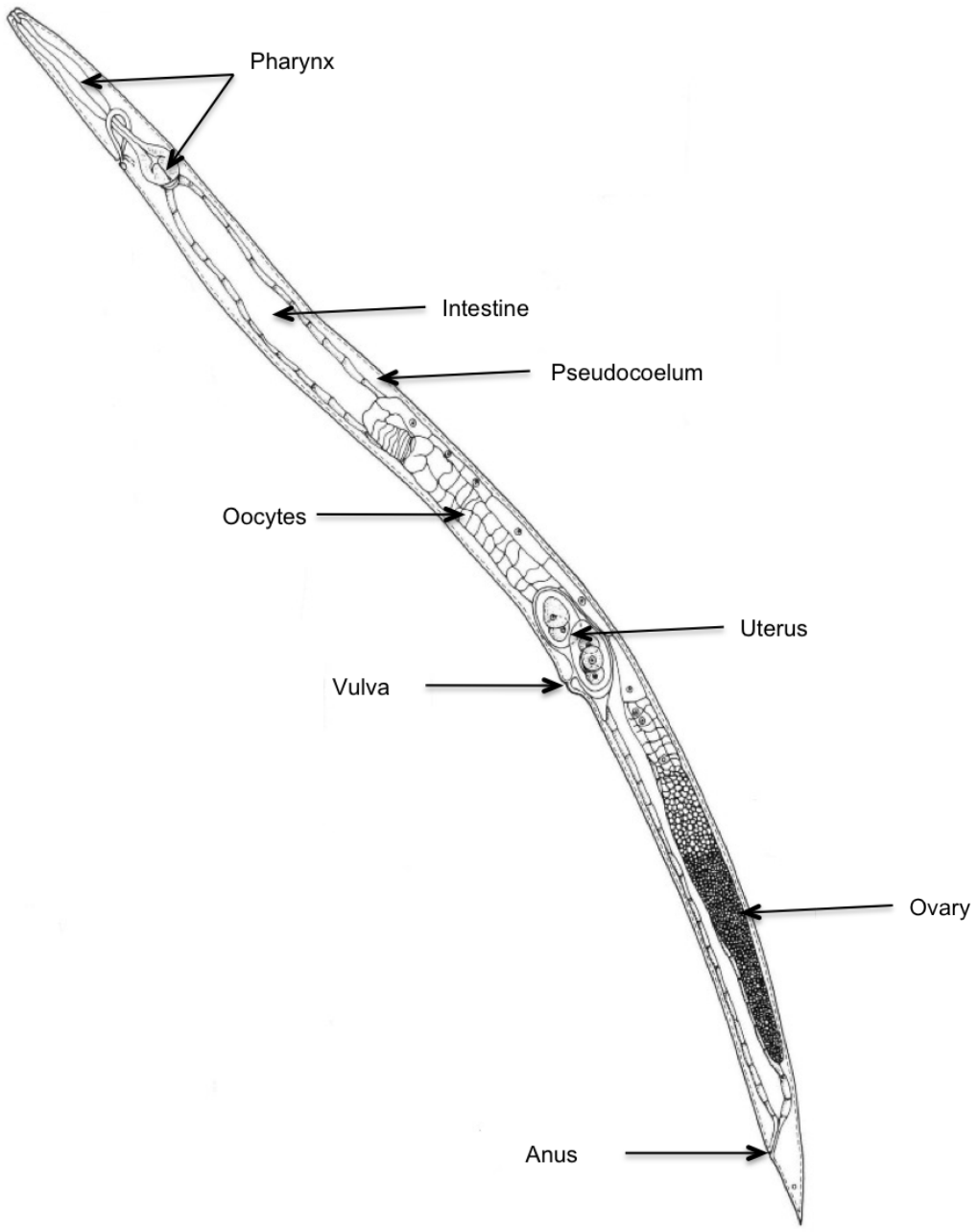


Figure 1.2: A *P. superbus* female (from Bostrom (1988)).

General Biology

P. superbus is a gonochoristic (i.e., sexually reproducing with separate males and females) free-living nematode that feeds on bacteria (Shannon et al., 2005). In the laboratory it is cultured on nutrient growth agar with a lawn of *Escherichia coli*. It has been shown to be widespread in soil and is the most common species in the United Kingdom and United States amongst the soil-dwelling nematodes (Williams, 1986). A *Panagrolaimus sp.*, isolated in Surtsey in 1995, was described thus: "corpus:isthmus ratio was 2.0 - 2.1 for the females, and 2.3 - 2.5 for the males. The length of the postvulval sac was 75 - 80% of the body width, and the length of the gubernaculum was 11μm. The lateral field of the female had four lines and was 5μm wide." (Frederiksen et al., 2001).

Panagrolaimus have been classified based on many different morphological features since the genus was first presented by Fuchs (1930). These include deMan's ratios, positions of plasmids, papillae arrangement, spicule shape and stomatal rhabdions. This has lead to confusion and misclassification due to inconsistency with measurements and observations. Andrássy (1984) compiled a list of 35 members of the genus but, as with previous reviews, this was also done on the basis of morphological characteristics. Given that few members of the genus have distinct characteristics, Williams (1986) used scanning electron microscopy (SEM) on 32 identified strains of *Panagrolaimus*. This allowed for the definitive characterisation of the *Panagrolaimus* species on the basis of lip separation and shape. This work separated these strains into four groups, which were named as four separate species based on original species descriptions considered by the authors to be most similar to these groups. Members of group 1 are described as having six lips with distinct separations between each one. Fuchs had previously described *P. superbus* in 1930 as having "six, slightly convex small papillae bearing lips" (see Figure 1.3). This was in accordance with Williams' observations and thus Williams placed members

of group 1 in the species *P. superbus*. Bostrom used morphological, morphometric and SEM data to compare the two distinct *Panagrolaimus* species isolated from a gulls nest on Surtsey Island with a third population of *Panagrolaimus* isolated from agricultural soil in Sweden (Bostrom, 1988). He concluded that the Swedish strain belonged to *P. rigidus* and the two species from Surtsey corresponded to *P. superbus* and *P. detritophagus*.

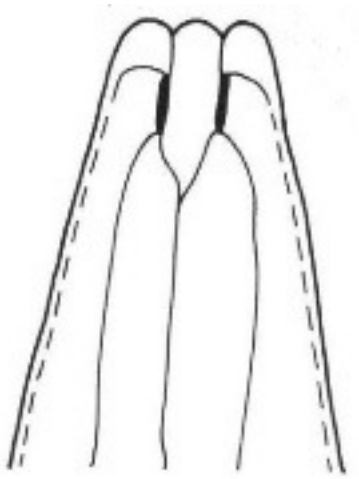


Figure 1.3: *P. superbus* head region with six distinct lips used for morphological characterisation (from Bostrom (1988)).

Given the subjective nature of differentiating species on the basis of morphology alone, it was vital that a molecular study be done to verify positioning of the genus on both the nematode and the *Panagrolaimus* phylogenetic tree. Lewis et al. (2009) showed the position of *P. superbus* in the genus *Pangrolaimus* in clade IV (*Tylenchida*) of the phylum *Nematoda*. This study was done using the 18S and 28S rRNA regions of each species (Lewis et al., 2009). Figure 1.4 shows the phylogenetic tree for *Panagrolaimus* species based on the neighbour joining (NJ) algorithm. The NJ phylogeny was highly convergent with the Maximum Parsimony (MP) phylogeny. The authors acknowledged that, to redefine the genus, closely related nematodes from other genera would need to be included, but for now this

remains the most complete phylogeny of the genus.

Geographical Distribution

Members of *Panagrolaimoidea* are found in all moist environments - marine, freshwater and soil (Lewis et al., 2009) *Panagrolaimus* is the second-most common genus found in dry soil in the Kinchega National Park in Australia (Nicholas & Stewart, 1985). Some have been found as far north as the Arctic (*P. superbus*), as far south as the Antarctic (*P. davidi*) and as close by as the roof of the Callan building, National University of Ireland Maynooth (*Panagrolaimus sp. AS02*). Despite the nematode *P. magnivulvatus n. sp.* being an Antarctic nematode, it also shows many morphological similarities to *P. superbus* (Bostrom, 1995). Considering the similarity in morphology this may be a closely related species, but as no DNA sequencing has been done, this hypothesis is speculative.

Panagrolaimus is one of five nematode genera that dominate bacterial mats in caves. A *Panagrolaimus* species has been found in the Bakwena Cave, Guateng Province in South Africa and in the Movile cave in Romania (Poinar & Sarbu, 1994). The worm was found only in guano deposits so it is assumed that the nematode's source of entry was the cave's resident bats, much the same as its appearance on Surtsey. It is suggested that *Panagrolaimus* have a phoretic relationship with insects. If the bats fed on beetles and the worms associate with the beetles then this would lead to entry to the cave. Beetle remains found in the guano deposits support this theory. An assumption could be made that *Panagrolaimus* could have environmentally adapted to this habitat and its sporadic food resources, by developing genes to allow it to enter a dormant state. The appearance of *Panagrolaimus* species in such harsh environments such as caves, the Antarctic and the Arctic also lends support to this theory (Rensburg et al., 2010).

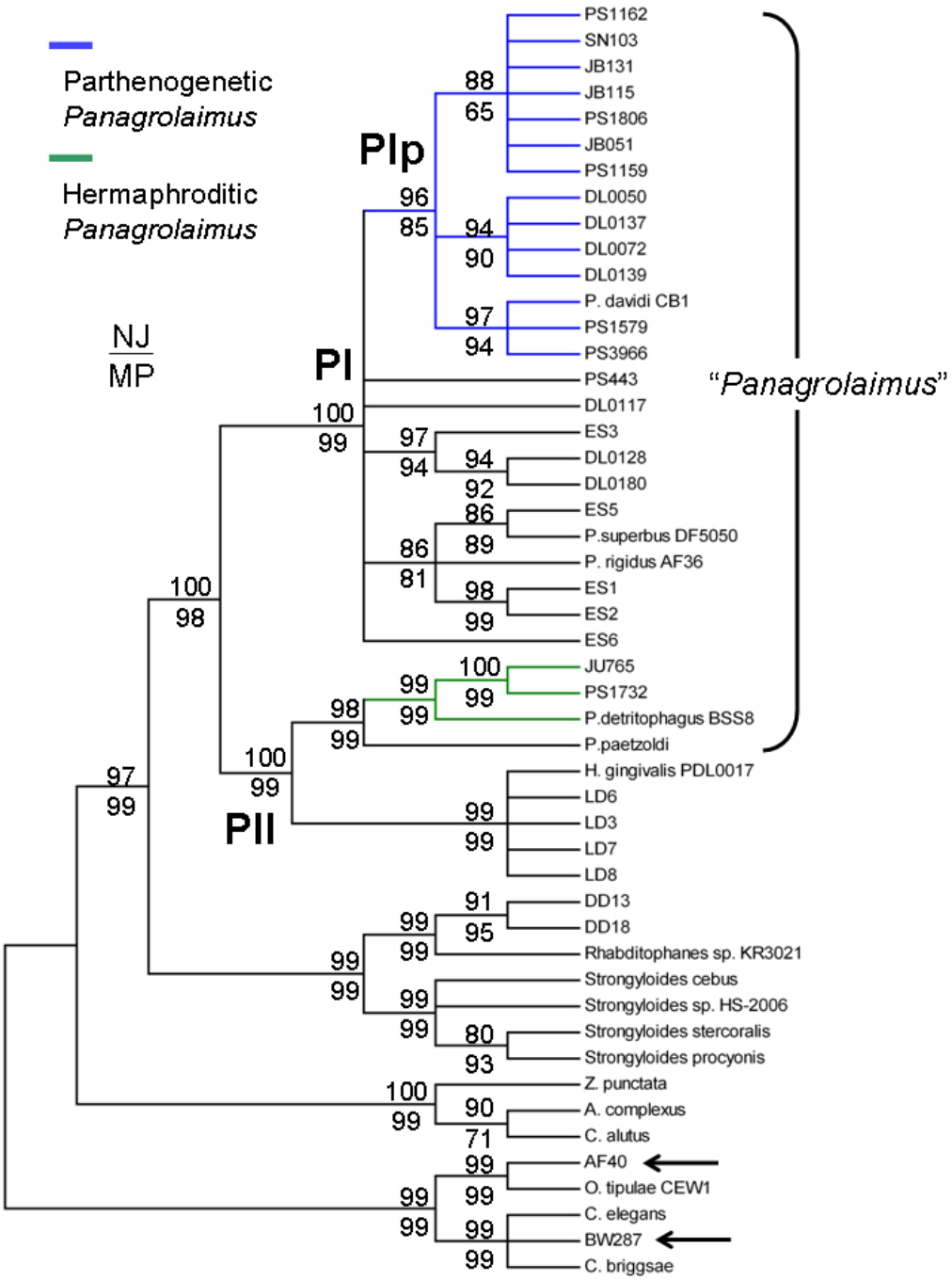


Figure 1.4: The nuclear rRNA gene phylogeny for *Panagrolaimus* (from Lewis et al. (2009)). Blue lines indicate parthenogenetic strains with hermaphrodite strains shown in green. The arrows denote *rhabditid* strains initially misclassified as belonging to the *Panagrolaimus* species. Node-specific bootstrap values are shown with NJ values over MP.

Recently, a member of the *Panagrolaimidae* was isolated from fracture water from a gold mine 1.3km below the surface (Borgonie et al., 2011). These nematodes correspond to a new species named *Halicephalobus mephisto*. *H. mephisto* is tolerant to high temperatures as its maximum growth temperature is 41°C. It is a bacterial feeding parthenogenetic nematode.

Panagrolaimus species have also been used in studies of comparative development. The genus is quite unique in that gonochoristic, hermaphroditic and parthenogenetic species are found. Studies done on the phylogenetics of the genus have shown a single origin of parthenogenesis from what is presumed to be a gonochoristic ancestor (Lewis et al., 2009).

Life Cycle

Sohlenius (1988) studied the life cycle of *P. superbus* and *P. detritophagus* on agar plates at 20°C. He found that *P. superbus* had a generation time of about eight days from egg to egg. Total life span was found to be 16-17 days. The numbers of offspring produced range from 7-39 eggs daily depending on environmental conditions. Exponential growth occurs 14-22 days after sub culturing. Mean body length and frequency of laying eggs decreases as time passes, with longest females visible during exponential growth and most females containing eggs at that time. After 23 days, almost no females contain eggs due to the depletion of food source which can be identified though the disappearance of visible bacterial growth. Due to the abundance of food during the period of exponential growth, there is a resulting population of small juveniles with a small food supply towards the end of the culturing cycle (Sohlenius, 1988). Similarly *P. detritophagus* has a life span of 7-16 days.

1.1.2 Environmental Stress Tolerance in Nematodes

Nematodes are aquatic animals that require a film of water over their body for normal activity (Wharton, 1996), nevertheless, they are probably the most species-rich multicellular phylum on the planet. Free-living nematodes occupy a great diversity of niches including in marine and freshwater sediments, soil and moist terrestrial habitats. These terrestrial habitats range from the Antarctic, temperate and semi-arid soils to terrestrial mosses. Nematodes are also successful endoparasites of animals and plants. Due to the wide range of habitats that they occupy, nematodes are often exposed to periods of environmental stress. The occurrence and duration of a change in optimal conditions can be unpredictable, so several nematodes have evolved a number of physiological and biochemical adaptations that allow their survival in response to severe stress such as high temperatures, cold and freezing conditions, desiccation, oxidative stress and starvation. Given the natural habitat of *P. superbus*, exposure to desiccation and/or freezing, could be an issue at certain times of year and, therefore, the nematode has adapted to survive both of these stresses. The following sections will focus on these stresses.

1.1.3 Freezing Tolerance

Some nematodes have adapted to survive extreme cold, e.g., the Antarctic nematode *P. davidi* (Wharton, 1996; Lewis et al., 2009). Cryobiosis or the ability to survive freezing with little or no preconditioning is an interesting characteristic also displayed by *P. superbus* (Shannon et al., 2005). Several species in the genus are also freezing tolerant, that is, they have the ability to survive intracellular freezing using a cryoprotective dehydration strategy (Wharton, 1996). Much work has been carried out in *P. davidi* and an interesting study would be a genome comparison between sister species from the Arctic and Antarctic. Freezing tolerant

nematodes have evolved to survive by allowing their extracellular fluids to freeze or by supercooling. Due to the sub-zero climates in the habitats of the Antarctic and Arctic nematodes, this cold tolerance is vital for survival. Freezing tolerance has also been identified in *Trichostrongylus colubriformis* (Wharton, 1996), *Aphelenchoides ritzemabosi* (Asahina, 1959), *Coomanus gerlachei* (Pickup, 1990*a*) and *Tetracephalus tilbrooki* (Pickup, 1990*b*). These organisms amongst others provide a good basis to study extreme habitat survival skills and some suggest that they could be used to study climate change (Wharton & Marshall, 2009).

1.1.4 Anhydrobiosis and Desiccation

Dehydration is a severe stress for organisms where most animals die if they lose more than 15-20% of their body water. However, some organisms are able to survive conditions in which all the free water is removed from their cells, and where the hydration shell of their biomolecules is lost. They do this by entering into a state of suspended animation known as anhydrobiosis (life without water). In this state organisms, including nematodes, can survive without water for extended periods of time (Crowe et al., 1992). This is believed to be possible through a series of adaptations, including the production of the carbohydrates trehalose and glycerol which prevent damage to the cell membranes. There are many types of organisms that can enter a state of anhydrobiosis in different stages of their life cycles. These include rotifers, nematodes and tardigrades in the animal kingdom. Many other organisms can undergo anhydrobiosis and the most common example is plant seeds, which can be stored for many years and will germinate on the application of water.

This distribution demonstrates that anhydrobiotic phenotypes are likely to have evolved independently on multiple occasions and provides support for the concept of anhydrobiotic engineering (Tyson et al., 2012). Invertebrate anhydrobiotes

include members of the *Nematoda*, *Rotifera*, *Tardigrada*, *Crustacea* and *Insecta*. These anhydrobiotes typically occupy aquatic or terrestrial habitats that are prone to temporary water loss. Free-living nematodes, rotifers and tardigrades contain representatives which are capable of entering anhydrobiosis at all stages of their life cycle. Crustacean anhydrobiotic stages are confined to the embryonic cysts of aquatic brine shrimps and other microcrustaceans. This advantageous characteristic is being researched in conjugation with the use of dry vaccines, which will eliminate the need for their cool storage. An understanding of the molecular mechanisms responsible for anhydrobiotic survival will provide insights which may ultimately lead to the ability to confer desiccation tolerance on desiccation sensitive organisms (Tyson et al., 2012).

The free-living nematodes *Acrobeloides nanus* and *Aphelenchus avenae* have been shown to recover after being desiccated for 6.5 years and 18 months respectively. The longest accounts of nematodes surviving anhydrobiosis are the parasitic nematodes *Anguina tritici* and *Filenchus polyhypnus* which survived in seed galls for 28 years and in a herbarium for 38 years respectively (Aroian et al., 1993). A large number of plant and animal parasitic nematodes have anhydrobiotic eggs and infective larval stages (Perry, 1999).

Womersley recognised two broad categories of anhydrobiotic nematodes: fast dehydration and slow dehydration strategists (Womersley, 1987). Fast dehydration strategists can survive rapid dehydration; slow dehydration strategists need to firstly be preconditioned by exposure to a slow reduction in relative humidity, before they can survive a severe loss of water. *P. superbus* has been shown to be desiccation tolerant (Shannon et al., 2005). While some work has been carried out on the physiological characteristics of anhydrobiosis in *Panagrolaimus*, DNA sequencing has been limited to individual genes of interest and the D3 region of the 28S rRNA gene for phylogenetic studies.

When cells suffer severe dehydration, metabolism ceases, macromolecules denature, membranes undergo phase changes and fuse with other normally separate membranes. Unlike desiccation sensitive taxa, anhydrobiotes have evolved mechanisms which maintain the structure and integrity of macromolecules and membranes in the absence of water and also during rehydration and revival. Comparative studies of the desiccation tolerance phenotypes of anhydrobiotes show lineage specific differences in the response patterns and biochemical adaptations, which implies that anhydrobiotic phenotypes can be achieved in different taxa by the expression of functionally equivalent molecules. Based on currently available data from nematodes and other anhydrobiotic animals, Tyson et al. (2012) have presented a model showing the possible steps involved in the detection and expression of anhydrobiotic protection mechanisms in nematodes (Figure 1.5). This diagram lists some of the main effector proteins and biosynthetic enzymes which have been shown to have a role in anhydrobiotic protection. The application of high throughput transcriptome sequencing methods to anhydrobiotic nematodes and other anhydrobiotic organisms will greatly extend our knowledge of the biochemical and genetic mechanisms responsible for anhydrobiotic survival.

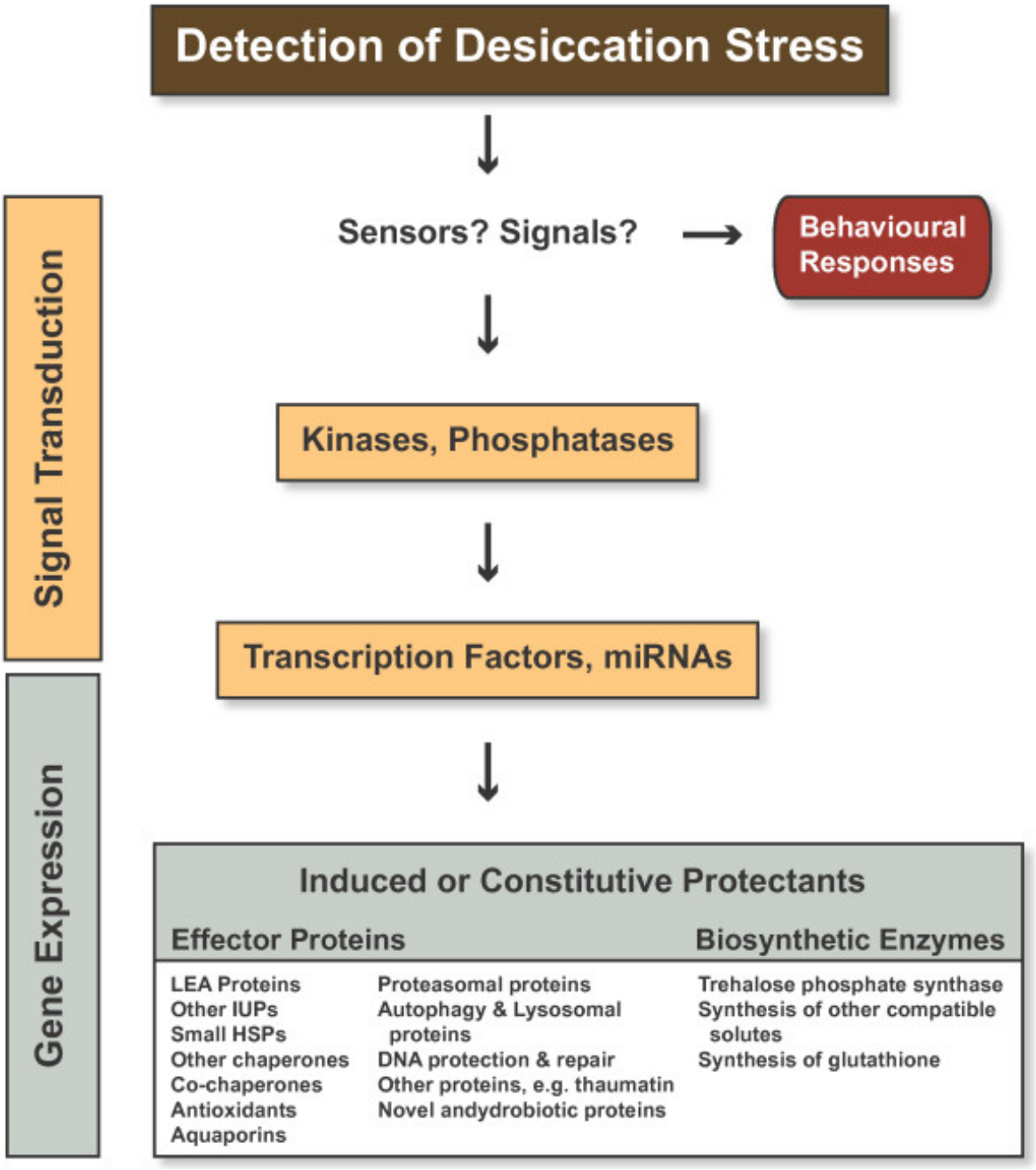


Figure 1.5: Possible steps involved in the detection and expression of anhydrobiotic protection mechanisms in nematodes (HSP = heat shock protein; LEA = late embryogenesis abundant protein; IUP = intrinsically unfolded protein)(From Tyson et al. (2012)).

1.2 Eukaryote Genomes

1.2.1 Introduction and Overview

All of an organism's hereditary information is encoded in its genome. In eukaryotes the genome comprises the haploid set of chromosomes. In bacteria, the genome is usually contained in one single chromosome, which is usually circular, but may be linear, as for example, in *Streptomyces* and *Borrelia* (Casjens, 1998). In viruses, the genome may be composed of DNA or RNA (or both) and be circular, linear or segmented in structure (Dimmock et al., 2007). Extrachromosomal DNA elements (plasmids) occur in many species of bacteria. Genome sizes range from (<3kb) for single-stranded DNA viruses (Rosario et al., 2009) to as much as 150,000Mbp for the angiosperm plant *Paris japonica* which has the largest genome described to date (Pellicer et al., 2010). The genome size ranges for extant life forms on Earth are presented in Figure 1.6. Genome sequences are made up of protein coding genes, RNA genes, functional intergenic and nongenic regions and 'non-functional' DNA. The functional non-coding DNA includes promoters, telomeres and regulators while the 'non-functional' portion includes repeats and transposable elements.

The first bacterial genome sequences were published in 1995: these were from *Haemophilus influenzae* (Fleischmann & Adams, 1995) and *Mycoplasma genitalium* (Fraser et al., 1995). The genome sequence of the *E. coli K12* strain was published in 1997 (Blattner, 1997). The genome sequence of the unicellular eukaryote *Saccharomyces cerevisiae* was published in 1996 (Goffeau, 1996). The first published genome sequence of a multicellular organism was that of the nematode *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998). Shortly after, the genomes of the fruit fly *Drosophila melanogaster* (Adams, 2000) and the plant *Arabidopsis thaliana* (Kaul, 2000), human (Lander & International, 2001;

Venter et al., 2001), mouse (Waterston, 2002) and rice *Oryza sativa* (Goff et al., 2002) were published. The genome of the first tree, the black cotton wood, *Populus trichocarpa*, was published in 2006 (Tuskan, 2006). In the past 10 years, the number of genome sequencing projects has increased dramatically and currently the complete genomes of 4,126 organisms are curated in Genomes Online Database [1]. These comprise Archaeal: 181; Bacterial: 3,762 and Eukaryal: 183 (data accessed on 28 January, 2013). Studies of these genomes have demonstrated that significant differences in genome organisation exist between prokaryotes and eukaryotes, but many aspects of gene and genome structure are conserved across living phyla and domains.

[1] http://www.genomesonline.org/cgi-bin/GOLD/index.cgi

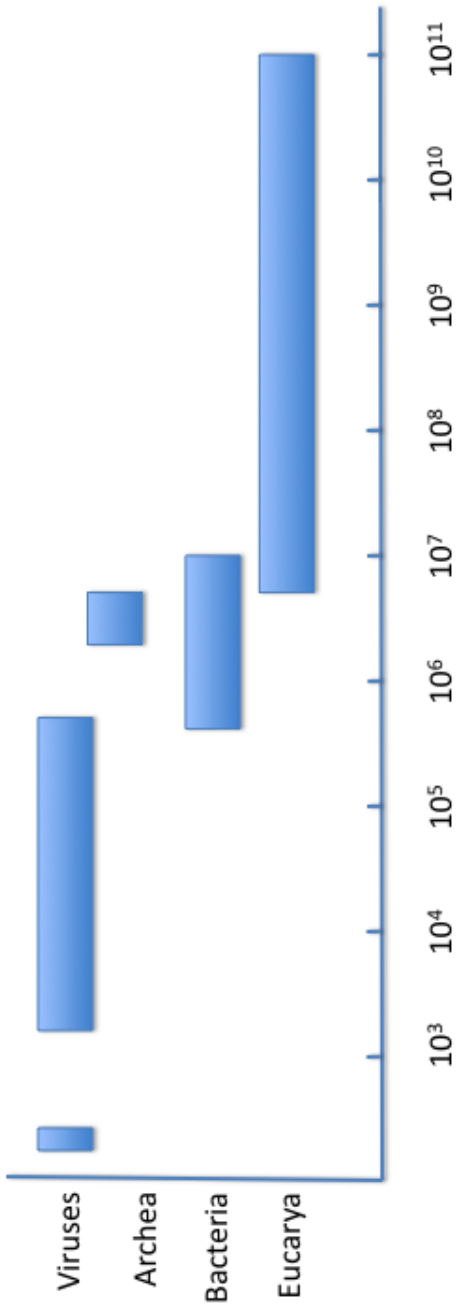


Figure 1.6: Known genome size ranges for extant life forms on Earth. The smaller outliers in the virus and Eucarya groups are viroids and green algal endosymbionts respectively (from Casjens (1998)).

1.2.2 Prokaryote and Eukaryote Genome Organisation

Most bacterial genomes are small, ranging in size from ∼0.5 Mb to 10 Mb (Casjens, 1998) and bacterial gene numbers range from ∼500 to 9,000 (Friar et al., 2012). The genome of *E. coli* K-12 is 4.1 Mb and contains 4,281 open reading frames (ORFs), which occupy ∼88% of the genome; regulatory sequences comprise ∼11% and repetitive sequences ∼0.7% (Blattner, 1997). Gene density is high in bacteria, averaging one gene per Kb of DNA. Bacterial genomes contain little non-coding DNA and bacterial genome size shows a strong positive relationship with gene number. Many bacterial genes, whose protein products form part of a common biochemical pathway, are organised into operons (reviewed by Rocha (2008)). Operons are clusters of coregulated genes encoded on a single polycistronic unit (see Figure 1.7). In addition to being physically close in the genome, these genes are regulated such that they are all turned on or off together. In *E. coli* 27% of all genes are grouped into 600 operons. Grouping related genes under a common control mechanism allows bacteria to rapidly adapt to changes in the environment while also minimising genome size. Prokaryotic genes can be found by looking in genome sequences for long ORFs.

Eukaryotic genome sizes range from 10Kbp for some fungi (Gregory et al., 2007) to 150Mbp for the angiosperm plant *Paris japonica* (Pellicer et al., 2010). Haploid eukaryotic chromosome numbers range from one in male ants of the genus *Myrmica pilulosa* (Crosland & Crozier, 1986) to several hundred in polyploid ferns of the genus *Ophioglossum* (Khandelwal, 1990). However, the number of protein coding genes in eukaryote genomes varies less dramatically than either genome size or chromosome number. Surprisingly, the nematode *C. elegans*, the plant *A. thaliana* and human *H. sapiens* all have similar gene numbers. Eukaryotic genomes have several features not found in prokaryotes. These include the presence of introns and mRNA splicing, an apparent lack of constraint on genome size, which has led to the

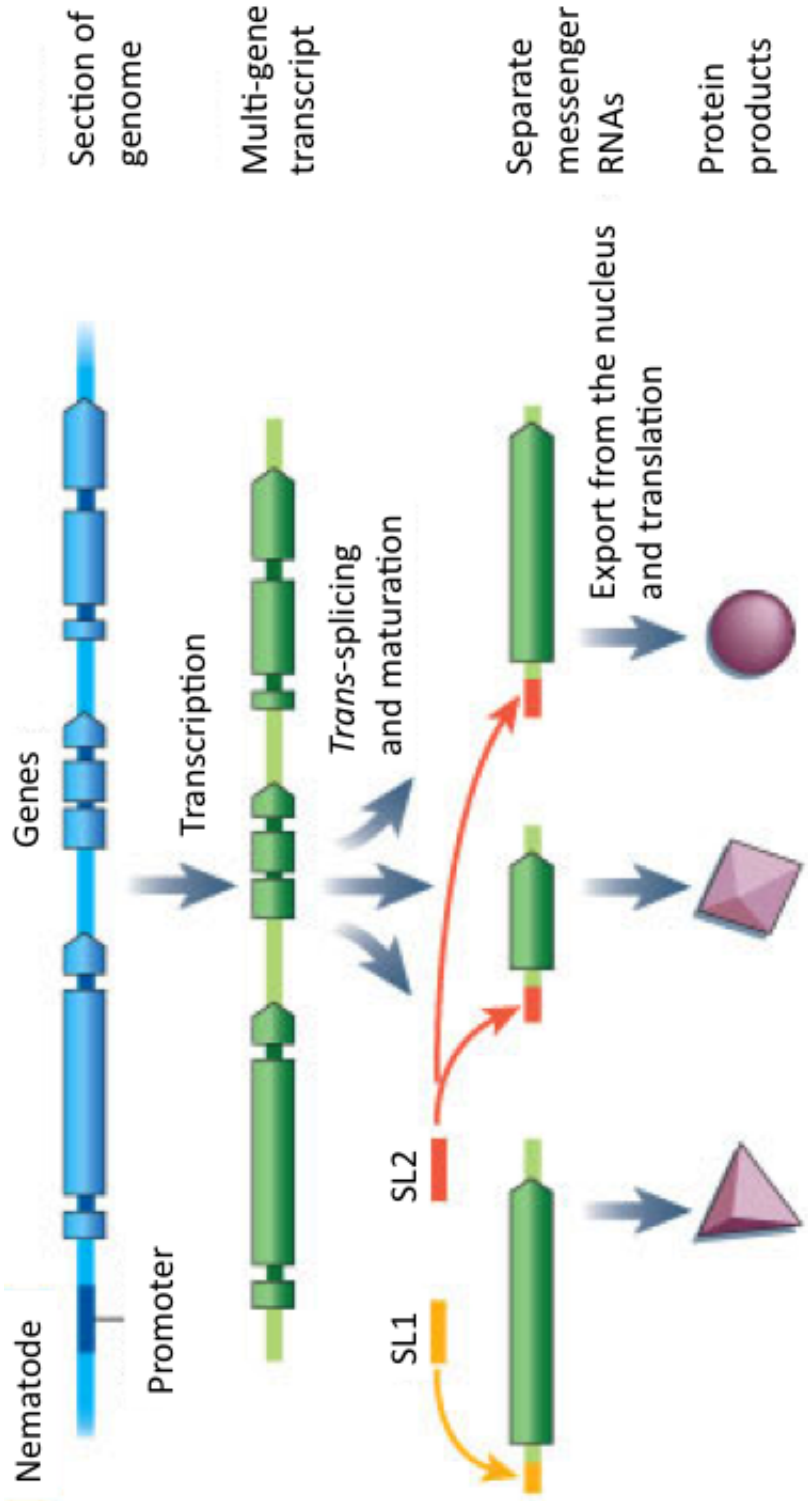


Figure 1.7: A *C. elegans* operon (von Mering et al., 2002).

accumulation of transposons and repetitive sequences in eukaryote genomes and a substantially lower gene density than that found in prokaryotes. Additionally, eukaryotic genes have complex regulatory regions and in multicellular species such regulatory regions have a modular structure that facilitates tissue specific gene expression. A review of the sequence statistics for some of the published nematode genomes is presented in Table 1.1.

Table 1.1: Nematode Nuclear Genome statistics (data from Coghlan (2005)).

Organism	**Genome size (Mbp)**	**Chromosome Number**	**Number of Genes**
C. elegans	100	6	20,621
C. briggsae	104	6	19,507
Ascaris suum	272	12	18,500
Brugia malayi	90	5	11,453
Meloidogyne hapla	62	14	14,200
Pristionchus pacificus	169	6	23,500
Trichinella spiralis	240	3	15,808

1.2.3 Eukaryote Gene Structure

A typical eukaryotic gene is organised into protein-coding regions called exons, separated by non-protein-coding regions called introns. Introns are transcribed, but not translated, being spliced out of the mRNA prior to translation by an RNA-protein complex called the spliceosome, producing the mature mRNA. In the typical process of eukaryotic gene expression, a gene is transcribed from DNA to pre-mRNA. mRNA is then produced by RNA processing, which includes the capping, splicing and polyadenylation of the transcript. It is then transported from the nucleus to the cytoplasm for translation. For a given transcript, there may be alternate splice patterns, each of which produces a different mature mRNA and may give rise to different protein isoforms (see Figure 1.8, panel B), which often have tissue specific expression. In humans it is estimated that alternative splicing occurs in 95% of the multi-exon genes (Pan et al., 2008) and that 68% of all alternative splicing events show tissue specific regulation (Wang, 2008).

Introns are thought to play a regulatory role in cells and the splicing process itself might help regulate mRNA passing from the nucleus to the cytoplasm. During RNA processing, splicing can give rise to different proteins by splicing various combinations of exons together. Introns play a role in the evolution of new and useful proteins. Genetic recombination could modify the function of a protein by changing the domain structure of that protein. The more introns a gene has the higher the frequency of recombination. Theoretically, exons could be exchanged between different genes and thus give rise to new novel genes with the potential to create advantageous functions (Roy & Gilbert, 2006).

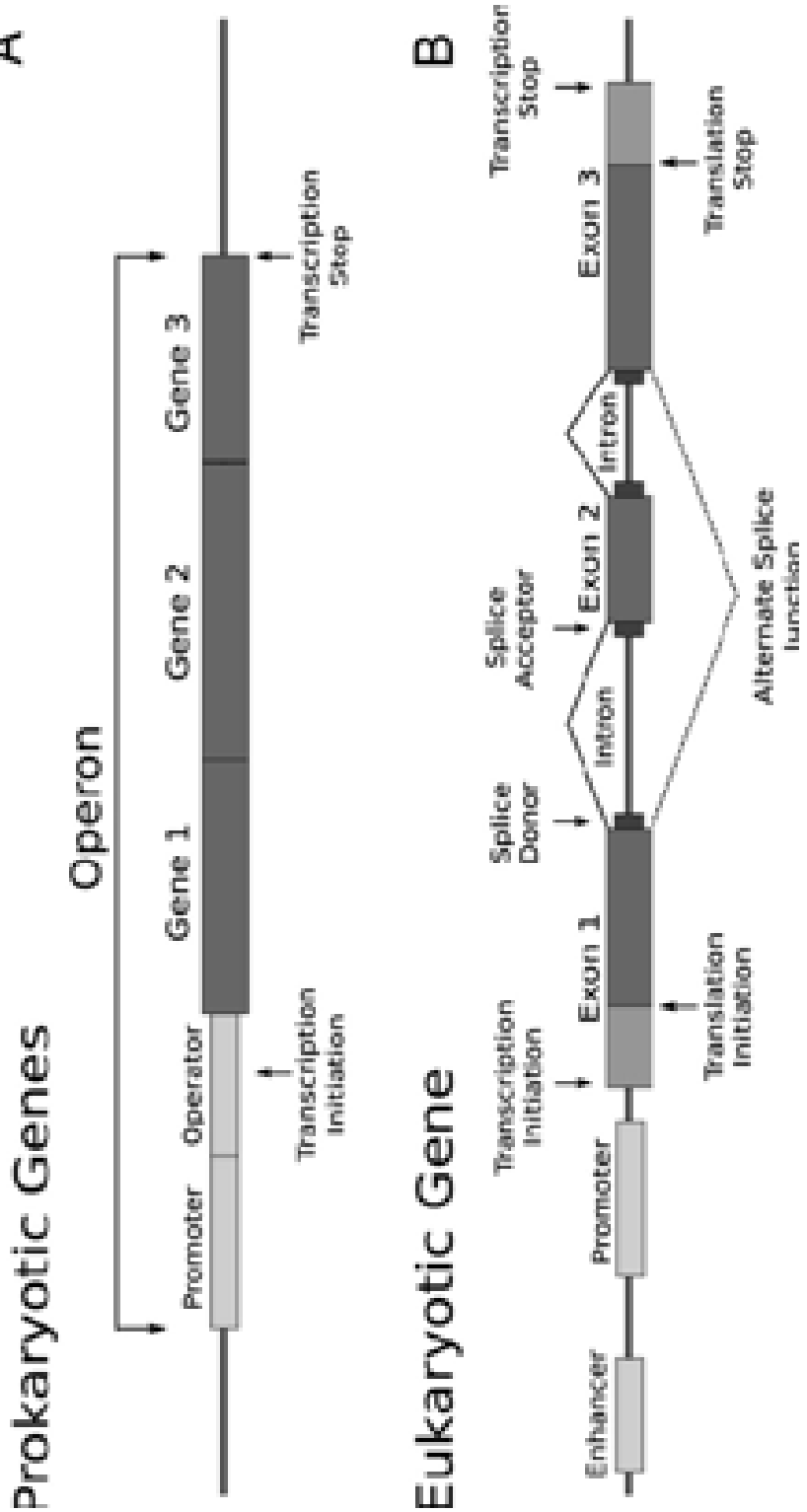


Figure 1.8: Prokaryotic and eukaryotic gene structures (Figure from Castellana & Bafna (2010)).

1.2.4 Content Comparisons of Selected Model Eukaryote Genomes

The genome content data for selected model eukaryotes is presented in Table 1.2.

Table 1.2: Eukaryotic genome statistics.

Organism	**Genome size (Mbp)**	**Chromosome Number**	**Number of Genes**
Saccharomyces cerevisiae	12.5	16	5,777
Caenorhabditis. elegans	100	6	20,553
Drosophila melanogaster	180	4	13,600
Homo sapiens	3,200	23	23,000
Mus musculus	250	20	30,000
Arabidopsis thaliana	125	5	25,498
Oryza sativa	420	12	55,986
Polulus trichocarpa	410	19	45,000

There has been a big increase in gene number between *S. cerevisiae* and the multicellular eukaryotes. By comparing genome sizes of more than 1,000 publicly available genomes from the three extant domains of life (Friar et al., 2012) it was observed that the number of ORFs varies with overall genome size for each domain. They found that for prokaryotes the number of ORFs increases linearly with genome size, up to a limit of 10,000Kb but, for eukaryotes the relationship between number of ORFs and genome size is non-linear (see Figure 1.9). These authors conclude that arising from an increase in genome size and organismal complexity there is a requirement to have additional non-coding DNA to control and efficiently regulate gene expression in eukaryotes.

Plant genomes also tell an interesting story. Even though we can observe a similarity in gene order and a high level of sequence similarity in genes, (Hu et al., 2011) there is a reported difference of approximately 80Mb (over 200Mb compared with 125Mb) in genome size between *A. thaliana* and *A. lyrata*. This is surprising as genome size shift in *A. thaliana* is consistent with genome loss and there are clear reductions in size due to chromosome rearrangements, TE copy number, small and large deletions, and even gene number. Furthermore, with the exception of single

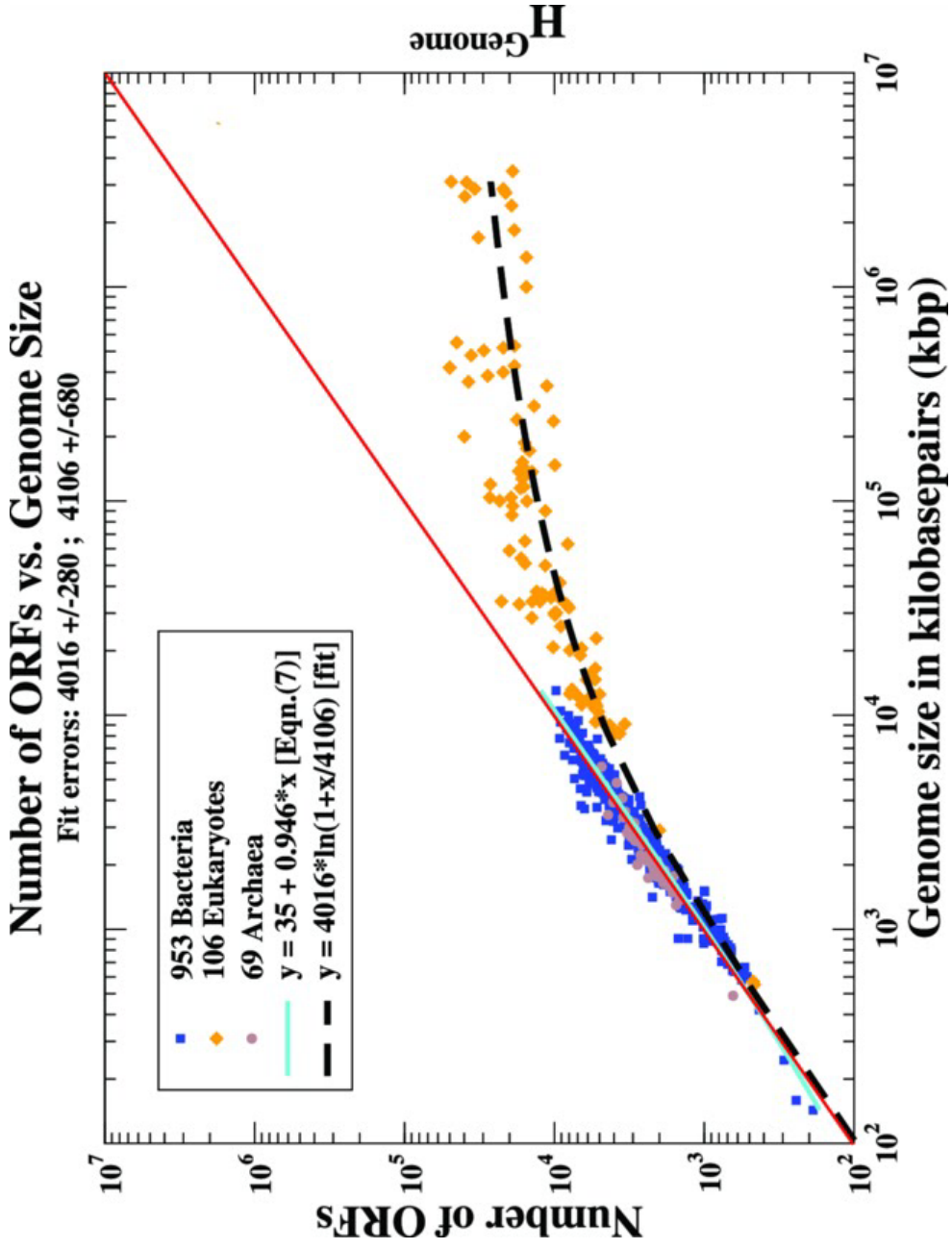


Figure 1.9: The number of Open Reading Frames in each genome versus genome size for the three extant domains of life (from Friar et al. (2012)).

base pair deletions, the DNA size change between these two species is apparent for deletion/insertion events at all size ranges, although it is especially exaggerated at the larger size range.

1.2.5 Nematode Nuclear Genomes

Free-living nematodes are under-represented in the research literature, with the exception of *C. elegans* (Brenner, 1974) which was the first model organism and first multi-cellular eukaryotic organism to have its genome fully sequenced. The genome sequencing of *C. briggsae* followed this and several other *Caenorhabditis* genomes have been or are currently being sequenced. In recent years the genome sequences of seven other nematodes were published, most significantly: the animal parasite *Brugi malayi* (Ghedin et al., 2007), a free-living nematode *P. pacificus* (Dieterich et al., 2008) and the plant parasites *M. incognita* (Abad et al., 2008) and *M. hapla* (Opperman et al., 2008).

Nematode genome sizes range from 19.56Mb for the plant parasitic nematode *Pratylenchus coffeae*, to 2,445Mb for *Parascaris univalens*, a roundworm parasite of horses (Animal Genome Size Database. http://www.genomesize.com). The mean size of all nematodes in the database is 146.7Mb.

The genomes of nematodes, ascidians (sea squirts, Phylum *Chordata*), and trypanosomes are unusual among eukaryotes in that they contain operons. Approximately 15% of genes in the genome of *C. elegans* occur in operons (Blumenthal et al., 2002) with operons being prevalent among nematodes inside and outside this genus as well (Guiliano & Blaxter, 2006). The ascidian *Ciona intestinalis* also harbours >20% of its genes in operon structures (Satou et al., 2006).

Many of the genes in *C. elegans* operons encode proteins required for basic cellular processes such as metabolism, transcription and RNA processing. Zaslaver et al. (2011) found that 88% of the genes in *C. elegans* operons are growth related and that the average expression level of operon genes is about 2-fold higher compared to the average expression level of nonoperon genes (Figure 1.10) They also tested the hypothesis that the need for rapid activation of multiple genes during recovery from growth-arrested states may explain operon evolution in metazoans.

They obtained transcriptional evidence that operon genes are upregulated in *C. elegans* during recovery from the growth arrested L1 and dauer larval stages (Figure 1.10 panels B and C). Similarly they found that in the sea squirt *Ciona intestinalis* expression of operon genes increased dramatically immediately following metamorphosis. Metamorphosis of *C. intestinalis* is characterised by the transformation of a non-feeding mobile larva into a filter-feeding adult. Thus Zaslaver et al. conclude that during metabolic arrest *C. elegans* and *C. intestinalis* maintain low levels of transcriptional resources when they are not needed while also ensuring that these low levels will support a fast and efficient transition from arrest into growth (Zaslaver et al., 2011).

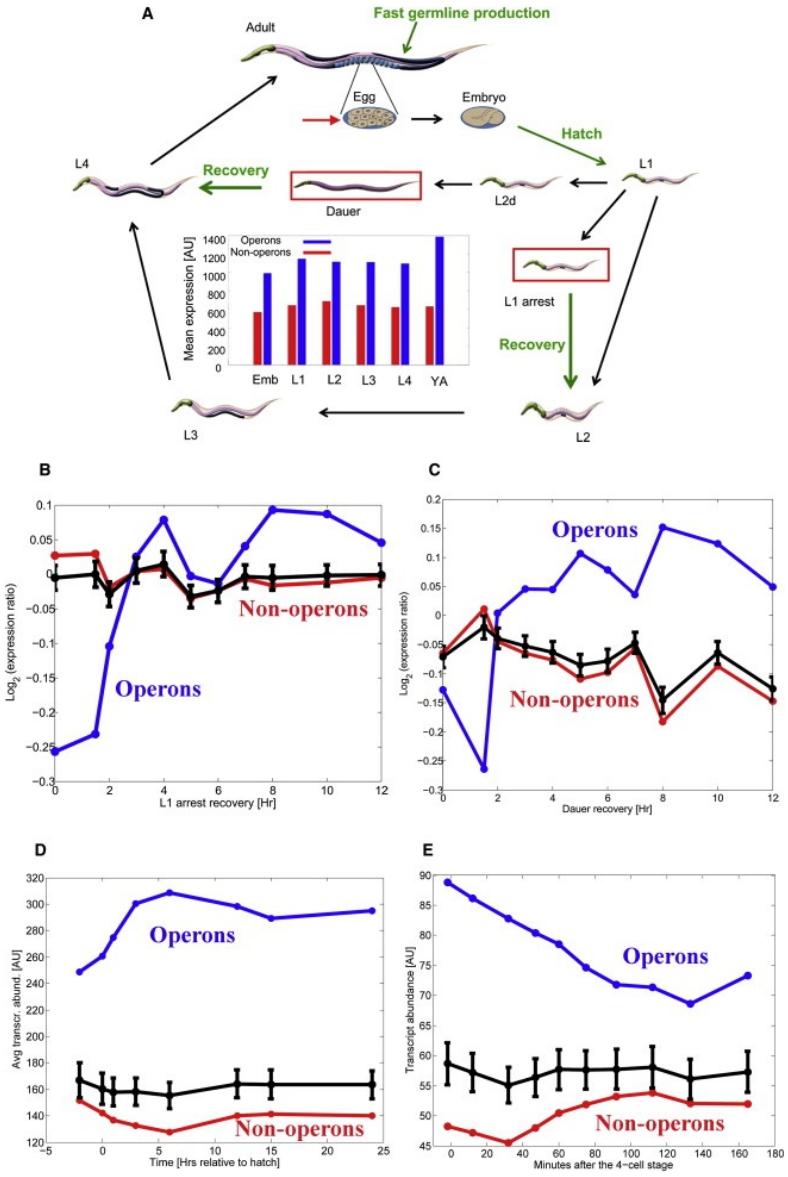


Figure 1.10: Expression profile of operon genes during the life cycle of *C. elegans* (A) and during recovery from growth arrested states (B,C). The lifecycle of *C. elegans* consists of four larval stages followed by an adult stage. If unfavorable conditions arise during larval development, worms stop growing and arrest at the L1 state or as dauer larvae, a highly resistant and long-lived state. When conditions improve, the worms recover and resume normal development. (B,C) The expression dynamics of operon genes upon recovery from both L1 and dauer arrest based on cDNA microarray data are presented in panels B and C. Figure taken from Zaslaver et al. (2011).

1.3 Mitochondrial Genomes of Nematodes and Other Animals

Mitochondria are the site of oxidative phosphorylation, which is essential for the production of ATP. Thus mitochondria have a central role in cellular metabolism. Within the mitochondria, there is a genome that is separate from the nuclear genome and referred to as mitochondrial DNA (mtDNA). In animals mtDNA is generally a small genome (12 to 20Kb). Although much larger animal mtDNA genomes have occasionally been found, these are the product of duplications of the mtDNA and are not due to the presence of additional genes (Boore, 1999). In human cells, 2-10 mtDNA molecules per mitochondrion have been reported (Griffiths, 2000).

Unlike nuclear DNA, mitochondrial DNA is a circular molecule. Nematode mtDNA genomes typically contain 12-14Kb and are composed of 12 protein coding genes, 22 transfer RNA, 2 ribosomal RNA genes (rrnL and rrnS) and a non-coding hyper-variable region which initiates transcription and replication (See Figure 1.11).

While the mtDNA genes have a relatively conserved sequence amongst members of the phylum, the order of these genes in the circular genome varies between different nematode genera. This makes each mitochondrial genome sequencing project a challenge. The 12 mtDNA protein coding genes in nematodes are involved in ATP synthesis which takes place in the matrix of the mitochondria. These genes are cytochrome oxidase subunits I-III (COXI, COXII and COXIII), cytochrome B apoenzyme (COB), NADH dehydrogenase subunits 1-6 (NADH1, NADH2, NADH3, NADH4, NADH4L, NADH5 and NADH6) and ATP synthase subunit 6 (ATP6). The tRNAs have a 7bp amino - acyl stem, a 4bp DHU stem with a 4-8bp loop, a 5bp anticodon stem with a loop of 7bps and a TV replacement

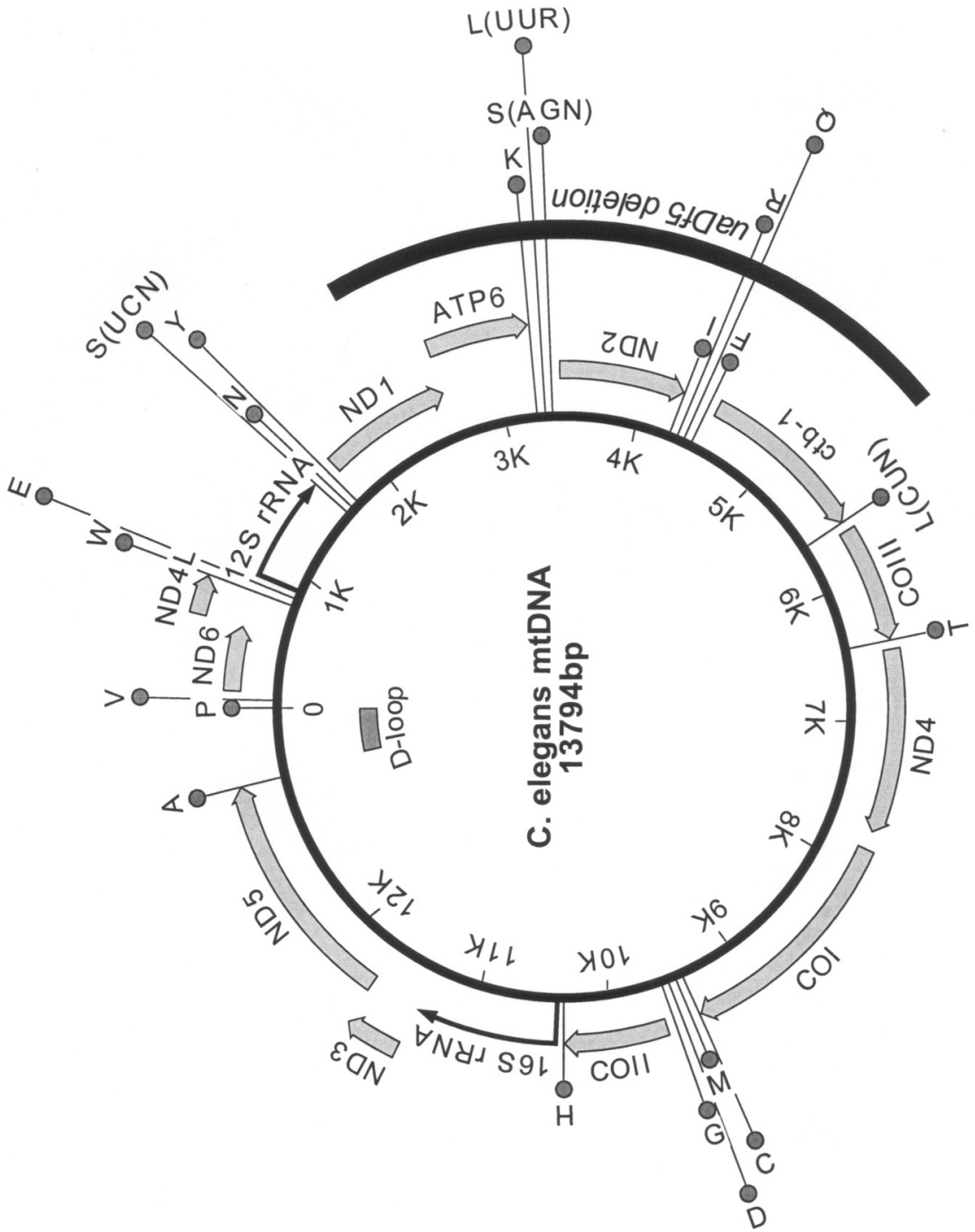


Figure 1.11: *C.elegans* mtDNA genome structure and gene order. Figure from Lemire (2005).

loop of 6-12bps, as show in Figure 1.12. Due to the variation in mitochondrial DNA genomes, PCR amplification requires the use of primers in a range of combinations to identify those primer pairs which amplify contiguous DNA sequences.

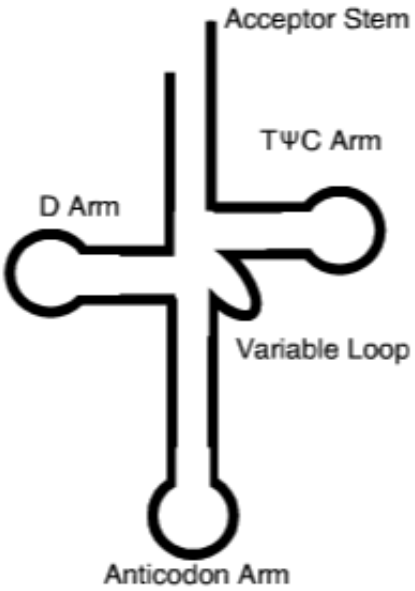


Figure 1.12: General structure of a tRNA (McNulty et al., 2012).

Previous authors have used a 454 sequencer from Roche to generate the sequence of their previously isolated clones and genes following long PCR (Jex et al., 2008). An interesting study was undertaken which compared sequencing mitochondrial genomes by traditional methods, versus re-sequencing them using next-generation technology (Jex et al., 2010). It was found that there were just 2 errors in 14,055bp of sequence, generated using the Roche high-throughput sequencing system. This establishes high-throughput sequencing as a reliable and fast method of generating sequences even from AT rich genomes.

1.4 Genome Sequencing

DNA sequencing began in 1977 (Sanger & Nicklen, 1977). The first step in Sanger sequencing involves separating the double stranded DNA into single strands by denaturation. A primer binds to the template strand and DNA polymerase is used to synthesise a complementary strand. A dye terminator nucleotide is included in the reaction mix and when incorporated into the new DNA strand it stops its growth, as well as leaving a marker for further identification. By labeling the fragments of DNA, they can be separated by size on electrophoresis genes and visualised. By giving each dye terminator molecule a different colour, a pattern, or sequence, can be established. The discovery of capillary electrophoresis meant that sequencing speed was increased and about 1Mb of DNA could be sequenced daily by a single machine.

The speed of sequencing greatly advanced with the introduction of shotgun sequencing. With this technique, a large fragment of DNA, or indeed a genome, is fragmented into smaller random pieces. Each piece is cloned into an *E. coli* plasmid vector, sequenced and assembled. Due to the large number of copies of each fragment, the coverage of the genome sequenced is increased (Weber & Myers, 1997). This method ensures that each base is sequenced at least twice. The quality of each base is measured and the 'Bermuda agreement', which states that there must be less than one error in 10,000 bases sequenced is applied. If this quality standard is not achieved the genome must be referred to as a draft. The assembly stage comes with its own set of challenges. Regions of low coverage or that have had just a single strand sequenced and gaps in the sequence coverage mean that a new technology had to be developed to deal with this. This technology is referred to as paired end reads or mate pairs (Roach et al., 1995), where two stretches of DNA, from each end of a single DNA molecule, are sequenced. They also have a recorded distance between them.

High-throughput or next generation sequencing is a relatively new and rapidly evolving technology. The predominant technologies on offer currently are 454 Roche (www.454.com) and Solexa Illumina (www.illumina.com). 454 was developed by Roche in 2005 and is based on pyrosequencing technology. One million reads with an average of ∼700bps per read and with greater than 99.5% accuracy can be achieved. Illumina operates on a sequencing by synthesis technology which uses reversible terminators and clonal single molecule array technology to produce 180 million reads of up to 100bps. Six Gbps per run are produced with a greater than 98.5% accuracy. Another emerging technology is LifeTech SOLiD (www.appliedbiosystems.com). This technique is based on sequencing by ligation which generates one hundred and eighty billion mappable bases per run of 75+35 bases in length.

These technologies can be used for *de novo* sequencing of transcriptomes and genomes as well as being used in re-sequencing projects. Both technologies possess advantages and disadvantages. Illumina is cheaper per base and generates more data per run, but 454 affords longer reads which are invaluable downstream at an assembly stage. The relatively cheap cost of sequencing a genome has lead to various people having their entire genome sequenced. This will lead to future development of personalised drug treatments, etc. Genomes of bacteria are much smaller and sequencing of these provides indications as to what makes them virulent and to the establishment of new and more effective control measures. In fact, *Haemophilus influenzae* was the first organism to have its genome fully sequenced (Fleischmann & Adams, 1995). Comparison of genomes of different organisms allows us to make evolutionary relationship predictions as to common ancestors amongst species. Using these fully sequenced and somewhat annotated genomes, we can identify similar proteins in newly sequenced genomes. However, it should be noted that depending on the organism sequenced, many of the genes discovered

may have to be labelled as novel as no known gene with similar sequence may have been identified previously (Blaxter, 2012).

Once a genome has been sequenced it needs to be annotated, that is, biological sense needs to be made of the masses of DNA sequence data returned from the sequencer. This is done by using *ab initio* gene finder programs such as GLIMMER (Gene Locator and Interpolated Markov ModelER) (Salzberg et al., 1998), or GENESCAN (Burge & Karlin, 1997), identifying the RNA genes, promoter and other regulatory regions. Different levels of completion of genome sequencing have been established which outline the characteristics a data set must achieve to be defined as a particular point on the completion scale (Chain et al., 2009) (Figure 1.13). The authors also hypothesise about the growth curve that will develop in new sequencing data sets generated as technology becomes more commercially viable for the average researcher and, indeed, individuals wishing to sequence their own genomes. In fact, Chain et al. (2009) underestimated these figures and, according to the US Department of Energy's Genome Institution, as of September 2011, 11,472 genome sequencing projects were listed in a publication by the group (Figure 1.14). As of July 2012 their website lists over 16,000 projects (Pagani et al., 2012).

Figure 1.13: Steps in completion of the annotation of genome data. Modified from Chain et al. (2009)

.

Figure 1.14: Trends in generation of genomic data with a conservative estimate of future projects shaded in light blue (from Chain et al. (2009).)

1.5 Transcriptome Sequencing

Expressed Sequence Tags (ESTs) are short nucleotide sequences of about 200 to 500bps in length. ESTs are a portion of a gene or whole gene in a cDNA clone which corresponds to an mRNA. Firstly, RNA is isolated and converted to cDNA using reverse transcriptase. RNAase H is used in conjunction with Polymerase Chain Reaction (PCR) and the resulting PCR products are cloned into a bacterial plasmid that is then isolated and sequenced. They can be sequenced from both the 3' and the 5' end of an expressed gene. They can be generated at relatively low cost from cDNA expressed at various different stages or under different stresses. They can be used in gene finding, mapping of genomes and in identifying coding regions of genes. ESTs were used extensively during the human genome project for gene discovery (Adams et al., 1991).

1.5.1 High Throughput Transcriptome Sequencing

An mRNA encodes the amino acid sequence of a peptide. The set of mRNA transcripts in a cell is referred to as the transcriptome. As mRNAs encode the proteins responsible for cellular function, the transcriptome can be used as an indicator of phenotype and function at a cellular level. The set of transcripts in a cell are not a stable entity. Depending on the environment, the set of transcripts expressed will change. Given a particular stressed state, a gene expression profile can be established and thus important genes for survival in that state will be highly expressed. This expression profile is regulated at all stages in the pathway from DNA to RNA to protein.

Many high-throughput transcriptome projects have been carried out. These projects offer insights into methods for sequencing technologies to establish a data-set, resulting in functional analysis pipelines and the answer to various biological

questions. Some examples of high-throughput transcriptome projects for nematodes are the *A. suum* transcriptome project (Wang et al., 2011) and the *Ancylostoma caninum* project (Wang et al., 2010*b*).

1.5.2 Transcriptome Assembly

Assembly algorithms are used to piece together the large amounts of short reads that are produced from sequencers. Up until the end of the last century, this was seen as an intractable problem. There existed some algorithmic solution to assembling reads but the amount of time it took to do this was not deemed reasonable. Computer hardware has advanced considerably since then and the accessibility of cloud computing has meant that these questions can now be answered. Reference mapping is performed by assemblers when there is a previously sequenced reference genome dataset that can be used to map the reads. *De novo* assemblers have no reference data set and the assembler must perform the assembly based on overlaps in the reads. If a transcriptome, rather than a genome, is being assembled this becomes an obvious difficulty. In a genome, an area of high sequence coverage may indicate repeat sequences but with cDNA sequencing this may also be indicative of a highly expressed gene.

Thus, the selection of an assembler is paramount in the pipeline of transcriptome analysis. The wrong choice could affect downstream analysis and as many authors have previously described, the choice is not easy and should not be trivial. The assembler chosen should be one that is most appropriate for the sequence data set to be assembled and will need to be optimised. A complete transcriptome can be defined as one which has long enough sequences to be deemed full length transcripts but not so long that the possibility of chimeric sequences should need to be considered. It should also have limited redundancy without excluding too many reads. Using a hybrid assembler such as CAP3 (Huang, 1999) also should be

considered as it has been shown that these assemblers can generate larger contigs and preserve better metrics, but redundancy and formation of chimeric contigs may be an issue, as found in this project (See Chapter 3). Thus they need to be considered carefully.

Martin & Wang (2011) established a set of metrics for evaluating the quality of an assembly. These were:

- Accuracy: Percentage of correctly assembled bases using reference transcripts.
- Completeness: Percentage of reference transcripts covered by all assembled transcripts.
- Contiguity: Percentage of reference transcripts covered by a single longest-assembled transcript.
- Chimerism: Percentage of chimeras that occur due to misassembly among all assembled transcripts.
- Variant resolution: Percentage of transcripts assembled.

Martin and Wang (2011) also propose that various factors should be considered when choosing an assembly strategy. These factors are:

- The existence of a complete reference genome: Given that the *P.superbus* genome has not yet been completed this made our choice of a *de novo* assembler straightforward.
- The availability of computer resources: We were fortunate enough to have access to Stoney, a Bull Novascale R422-E2 cluster with 64 compute nodes. Each compute node has two 2.8GHz Intel (Nehalem EP) Xeon X5560 quad-core processors and 48GB of RAM. This results in a total of 496 cores and

2976GB of RAM available for jobs. We also had access to Stokes, an SGI Altix ICE 8200EX cluster with 320 compute nodes. Each compute node has two Intel (Westmere) Xeon E5650 hex-core processors and 24GB of RAM. This results in a total of 3840 cores and 7680GB of RAM available for jobs (ICHEC, 2013). This meant that computer resource availability was not a critical factor on which to base choosing an assembler.

- The type of data generated: The sequencing was carried out on a Roche 454 FLX Titanium machine. As Newbler is the assembler designed for use with 454 data, it can correct for long stretches of homopolymers of unknown length, which are caused by ambiguities in the signal intensity. This is an interesting point and should be considered. It also means that the use of the *de novo* assemblers such as Velvet (Zerbino & Birney, 2008), Oasis (Majoros et al., 2005), etc., which are designed for shorter Illumina sequences, were not considered when it came to choosing an assembler. Overlap-Layout-Consensus assemblers such as MIRA (Chevreux et al., 2004), CAP3 (Huang, 1999) and Newbler are usually chosen for 454 data sets.

- The overall goal of the project: The goal of this study was to establish a list of transcripts that could be involved in stress pathways. Taking this into consideration, larger gene size transcripts would be preferable, as shorter sequences would mean transcripts could be fragmented across various sequences. While chimeras are not ideal, blasts hits to known stress transcripts should be feasible if two sequences were falsely aligned, as long as they truly belong in the same transcript.

Many authors including Kumar & Blaxter (2010) have used a variety of different parameters to evaluate the quality of an assembly. These include assembly metrics, resource usage, number of reads assembled, similarity/coverage to a reference

genome, gene coverage, integrity of assembled transcripts, sensitivity, specificity and percentage coverage to a proteome. This was also the approach that was taken in this study. Assembly metrics include N50 (number at which half the sequences of the assembly are that length or greater), max sequence length, total number of all bases used in assembly, number of large contigs (>1Kb), average sequence length, and number of contigs generated. Specificity can be examined by looking at the number of sequences in the assembly that had hits to CEG (Core Eukaryotic Genes) genes with >70, 80 and 90% coverage. CEG genes are a set of 248 highly conserved core eukaryotic genes present in low copy number in higher eukaryotes (Parra et al., 2007). Ambiguity can be checked by taking three CEG genes of different sizes and aligning them against each of the assemblies. Though N50 is predominately used to choose a genome assembly, it is important to note that it should not necessarily be weighted as the most important metric when choosing a transcriptome assembly as reads from separate transcripts might have been forced to assemble together.

1.6 Aims and Objectives of this Project

It is the aim of this project to suggest *P. superbus* as an alternative model organism for the use in stress studies.

The objectives of this project are to present:

- An EST data-set of just over 4,000 unigenes as a snapshot of genes of *P. superbus*, with an emphasis on those involved in stress,
- A transcriptome data set with a wide range of transcripts, particularly stress-related transcripts, pooled to be used as a baseline for stress,
- A complete mitochondrial genome for the nematode *P. superbus* including protein coding genes, transfer RNA and ribosomal RNA genes,
- A nuclear genome of *P. superbus* to an automated/directed improvement level of completion, for future use in gene comparisons and pathway reconstruction (Chain et al., 2009).

The data-set generated from this project will be made freely available and should greatly aid in focusing wet-bench studies to genes of interest and thus the characterisation and evolutionary history of stress genes in nematodes.

Chapter 2

Expressed Sequence Tags

2.1 Introduction

An EST is a short fragment of DNA traditionally generated by Sanger dideoxy terminator sequencing. In Sanger sequencing the fragment of DNA grows from the 3' end as deoxynucleotides are attached to the fragment by a phosphodiester bridge. Each of the deoxynucleotides is labelled with a different colour dye and thus the sequence of the DNA fragment can be identified. Sanger sequencing has been used in a multitude of sequencing projects to identify expressed genes of interest. By inducing an environmental state such as stress, the gene expression patterns characteristic of this stress state can be identified using EST sequencing. NCBI defines the dbEST as "a collection of short single-read transcript sequences from GenBank. These sequences provide a resource to evaluate gene expression, find potential variation, and annotate genes." As of February 2013, dbEST release No. 130101 has 74,186,692 EST sequences. Total numbers of nematode EST sequences are listed as 1,252,785 from 73 species. Of the nematodes, *C. elegans* has the most sequences with a count of 396,687 (Wheeler et al., 2005).

To identify constitutively expressed candidate anhydrobiotic genes, 9,216 ESTs

were obtained from an unstressed mixed stage population of *P. superbus*. 4,009 unigenes were derived from these ESTs. A set of 187 constitutively expressed candidate anhydrobiotic genes were manually annotated. Notable among those is a putative lineage expansion of the LEA (late embryogenesis abundant) gene family. The most abundantly expressed sequence was a member of the nematode specific sxp/ral-2 family that is highly expressed in parasitic nematodes and secreted onto the surface of the nematodes' cuticles.

There were 2,059 novel unigenes (51.7% of the total), 149 of which are predicted to encode intrinsically disordered proteins lacking a fixed tertiary structure. One unigene may encode an exo-β-1,3-glucanase (GHF5 family), most similar to a sequence from *Phytophthora infestans*. GHF5 enzymes have been reported from several species of plant parasitic nematodes, with horizontal gene transfer (HGT) from bacteria proposed to explain their evolutionary origin. This *P. superbus* sequence represents another possible HGT event within the Nematoda. The expression of five of the 19 putative stress response genes tested was upregulated in response to desiccation. These were the antioxidants glutathione peroxidase, dj-1 and 1-Cys peroxiredoxin, an sHSP sequence and an LEA gene.

In addition to providing cDNA clones and sequence data for candidate anhydrobiotic genes, the dataset presented here has also provided anchor sequences important for the assembly of the genome and transcriptome of *P. superbus*.

2.2 Methods & Materials

2.2.1 Nematode Culture

P. superbus (strain DF5050) was obtained from Prof. Bjorn Sohlenius, Swedish Museum of Natural History, Stockholm. The nematodes were cultured at 20°C in the dark on nematode growth medium (NGM) plates containing a lawn of streptomycin resistant *E. coli* strain HB101 obtained from the *Caenorhabditis* Genetics Center, University of Minnesota, USA. The NGM was supplemented with streptomycin sulfate (30μg ml^{-1}) as described in Section 5.2.2.

2.2.2 cDNA Library Construction and EST Generation

Total RNA was extracted from mixed stage unstressed worms using TRIzol reagent (Invitrogen, Carlsbad, USA). The cDNA library was prepared by Ms Mairin Skelton at the Scottish Crops Research Institute, Dundee, using the SMART cDNA Library Construction Kit Long-Distance (LD) PCR protocol (Clontech, Mountain View, CA 94043, USA). Fifty ng of total RNA was used for the SMART cDNA synthesis and there were 25 PCR cycles in the LD PCR amplification step. The cDNAs were cloned into the pDNR-Lib vector (Clontech) and transformed into *E. coli* DH10B cells. A total of 15,360 recombinant *E. coli* were picked using a Q-Bot robot (Genetix, Hampshire BH25 5NN, UK) and transferred to 384 well microtitre plates containing freezing media (Sambrook & Russell, 2011) and chloramphenicol (30μg ml^{-1}) and the plates were stored at -80°C. The cDNA inserts from individual transformants (n = 9,216) from the cDNA library were sequenced by the Sanger method at the Scottish Crop Research Institute, Dundee (4,224 clones) and at The GenePool, University of Edinburgh (4,992 clones). As can be seen from Figure 2.1, the majority of the sequences are between 50-800bps in length.

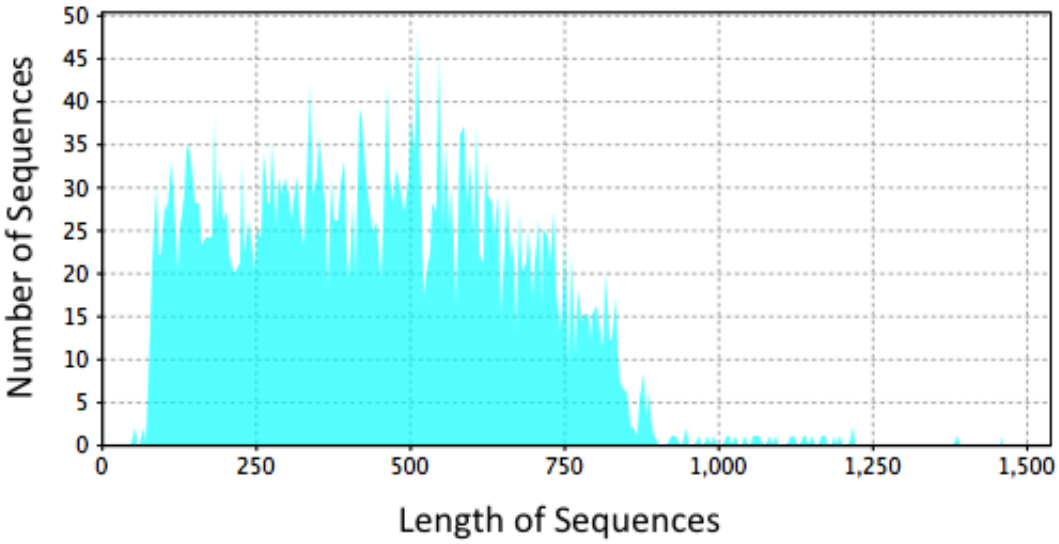


Figure 2.1: Sequence counts vs sequence length in base pairs.

2.2.3 Clustering and Sequence Analysis

The raw EST sequences were processed through the PartiGene pipeline at the Genepool, Edinburgh (Parkinson et al., 2004), first using trace2dbEST, which removes vector-derived sequences, poor quality sequences and ESTs shorter than 150bp, followed by CLOBB (Parkinson et al., 2002), an iterative program which groups the sequences on the basis of BLAST similarity into clusters that are putatively derived from the same gene. The Partigene pipeline clustered these ESTs into 1,079 consensus sequences (contigs) and 2,958 singletons. Removal of putative bacterial sequences and rRNA genes yielded a total of 3,982 putative protein-coding transcripts (unigenes). Clusters containing more than one sequence were then assembled into consensus sequences using Phrap (Green, 2012). The partial transcriptome consists of these consensus sequences, along with those clusters that contain only one sequence (singletons). Potential bacterial contaminant sequences (28 contigs) and nematode rRNA genes (27 contigs) were identified using

a BLASTN search of the *P. superbus* consensus sequences against the GenBank nucleotide (nt) database, with an e-value cut-off of $1e^{-50}$ to identify significant matches. These *P. superbus* EST sequences have been deposited in dbEST with the accession numbers GW405912-GW413517. The unigene sequences and annotations along with their constituent ESTs can be downloaded from the NEMBASE4 database (Elsworth et al., 2011) online and the unigene annotations can also be subjected to keyword queries using the NEMBASE4 database.

Genes encoded on the mitochondrial genome, as discussed in Chapter 6, were identified by BLASTX using 496 mitochondrial genes from 41 nematode mtDNA genomes from GenBank. These were used as queries against the *P. superbus* consensus sequences with an e-value cut-off of $1e^{-10}$. This resulted in 293 unique hits with a BLAST score of greater than 60 hits.

The consensus sequence for each unigene was translated using prot4EST (Wasmuth & Blaxter, 2004). Each unigene was then subjected to a BLASTP search (e-value cut-off of $1e^{-4}$) against a non-redundant custom database containing sequences from a variety of sources: the GenBank NR database, Wormpep (version 224) and an extended version of the NEMBASE4 database which will be referred as NEMBASE4^{+}. NEMBASE4^{+} has been supplemented with sequences from the following nematodes: *Plectus murrayi*, *Ditylenchus africanus*, *Aphelenchus avenae*, *Trihinella spiralis*, *Wuchereria bancrofti*, *Loa loa* and *Pristionchus pacificus*. Putative genes were annotated using the annot8r algorithm (Schmid & Blaxter, 2008). This software tool assigns Gene Ontology (GO) terms, Enzyme Commission (EC) numbers (Bairoch, 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway data (Kanehisa & Goto, 2000) to EST sequences based on BLAST searches against annotated subsets of the EMBL UniProt database (Jain et al., 2009). All BLAST results were parsed and the corresponding annotations were saved in a relational postgreSQL database. A web interface where the annot8r

annotations can be subjected to keyword queries and where output clusters can be retrieved is available online. Supplemental ortholog assignment and pathway mapping were carried out using the KAAS-KEGG Automatic Annotation Server (Moriya et al., 2007).

To identify putative unigene families among four anhydrobiotic nematodes, EST consensus sequences were kindly provided by various groups: 1,387 *Plectus murrayi* sequences (Adhikari et al., 2009) from Dr. B. Adhikari; 2,596 *Ditylenchus africanus* sequences (Haegeman et al., 2009) from Dr. A. Haegeman and 2,700 *A. avenae* sequences (Karim et al., 2009) from Dr. N. Karim. All ESTs were translated into peptide sequences using prot4EST and they were subjected to an all-vs-all BLASTP analysis to identify pairwise similarities. A graph representation of the homologous relationships among the unigenes was constructed, where each node is a unigene and an edge is drawn between any two nodes that have a BLASTP match. Each edge is weighted by -logE where E is the e-value of the alignment between two similar unigenes. e-values of 0 are transformed into $1e^{-200}$, i.e., an edge weight of 200. This graph is then used by the MCL algorithm (Enright et al., 2002) as input, with an inflation parameter of 2.1, to classify the unigenes into putative families.

2.2.4 Translation and Primary Structure Analysis of Novel ESTs

Novel ESTs were translated into putative peptide sequences using prot4EST which incorporates the ESTScan2.0 (Lottaz et al., 2003) and DECODER (Fukunishi & Hayashizaki, 2012) programs, using *de novo* prediction methods for predicting the amino acid sequence of cDNA sequences with putative sequencing errors (particularly insertions and deletions). The glycine content and the Grand Average of Hydropathy (GRAVY) values of these putative peptides were determined using

the ProtParam tool (Gasteiger, 2003). The GRAVY value is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence (Kyte & Doolittle, 1982). Predictions of the extent of intrinsically disordered regions within each putative peptide were determined using the IUPred program (using the long disorder prediction algorithm) (Dosztányi et al., 2005) kindly provided by Dr. Zsuzsanna Dosztnyi.

2.2.5 Real-time Relative qPCR Analysis of Gene Expression

A mixed population of nematodes was vacuum filtered onto 25mm Super Membrane Disc Filters at a concentration of 2,000 nematodes per filter. Five replicate filters were prepared for each treatment. The filters were placed in an 8L glass desiccation chamber over a saturated solution of potassium dichromate for 12 hours at 20°C in the dark, to generate an RH of 98% (Winston & Bates, 1960). Nematodes were then washed off the filters with distilled water and the nematodes from the five filters were pooled together. RNA was extracted using the TRIsure (BIO-38033, Bioline) method followed by treatment with DNAse I (Promega, M6101). Control nematodes were placed directly in to TRIsure and flash frozen with liquid Nitrogen without vacuum filtration.

Total RNA (1μg per reaction) was converted to cDNA using the Roche Transcriptor First Strand cDNA Synthesis Kit (04 379 012 001). One μl of cDNA from the above reaction was used for each real time qPCR reaction. These reactions were carried out on a Roche LightCycler 480 thermocycler using Roche SYBR I Master 1 kit (04 707 516 001). Each qPCR reaction also contained 5μl SYBR Master Mix, 0.002pmole of each primer and 2μl water. Primers were designed to produce an amplicon of approximately 125bp for each gene tested. These primer sequences are presented in Table 2.1. Relative expression data were calculated

with the LightCycler 480 Efficiency Method analysis software using the second derivative maximum option. The *P. superbus ama-1* and *rpl32* genes were used as a reference. Having established that the crossing point data were normally distributed and that the variance of the controls and treatment data were equal, two sample Student's t-tests were carried out to identify statistically significant differences in expression levels between the controls and the experimental treatments.

Table 2.1: The primer sequences used for real time qPCR analysis of gene expression in *P. superbus* in response to desiccation stress.

P. superbus **Cluster ID**	**Description**	**Primer**	**Sequence 5'-> 3'**
PSC00673	HSP70 family	PShsp-70F	ACGTGCAATGACCAAAGACA
		PShsp-70R	ACCATTGGCATCAACATCAA
PSC02842	HSP40/DNAJ family member	PShsp-40F	AAACAAGCCGTTGAAGCACT
		PShsp-40R	GCAGGCGATACTCCAAGAAC
PSC03116	sHSP family member	PShsp-12F	ACTCCAACATGGACGGAAAA
		PShsp-12R	ACGGTTGCCAATTTGCTATT
PSC01018	sHSP 21 Bursaphelenchus	PShsp-21F	GTTCATTCCTTCGTCGGGTA
		PShsp-21R	AGGCTTTGGAGCAAAGATGA
PSC00782	LEA Protein	PSlea-2F	TGGAATCCTCATCTCCAACA
		PSlea-2R	GCAGCATCATAGGCATCAGA
PSC01853	LEA Protein	PSlea-5F	GGAGCTGCAAAGGTTAAAGC
		PSlea-5R	ATGGCATCTTGTTGTTCACG
PSC00514	LEA Protein	PSlea-8F	GCTGGTAAAGCTAAGGATGTTATG
		PSlea-8R	GAACATTATCCCATGTTTCTTCAGC
PSC02695	Cyclophilin family member	PScyp-3F	TATCTGCACTGCCGTTACCA
		PScyp-3R	TCGGCAGAAGTTTTTCCACT
PSC00740	Protein disulfide isomerase	PS00740F	GCAAAACTGGAGCTGGTCTC
		PS00740R	AAACAGGCAATTTGCGTACA
PSC01029	Aquaporin	PS001029F	TTAGGAAATGCCCTCATTGG
		PS001029R	CAAGAACAAGGAAGGCAAGG
PSC03895	Peroxiredoxin 1	PScys-2F	TGGGGCTTAAACTTGGTGAC
		PScys-2R	GTTGTGCAGACAGGCGTAAA
PSC01468	RIC1 (Putative stress responsive protein)	Psric-1F	CCCCGATTATGTTGCTCTGT
		Psric-1R	ATCCGGGGATATAACCCAAA
PSC02304	DJ-1 family protein	PSdj-1F	AGCGCCAGTTATTTTTGCAC
		PSdj-1R	CCTGGAGCTCGACTCGTTAC
PSC02494	Glutathione peroxidise	PSC02494-F	TGATGATGCAGCACCACTTT
		PSC02494-R	TGGAGCGAAACGTTTAACAA
PSC04819	Glutathione peroxidise	PSC04819-F	TCAAGAACCTGCGGAAAATC
		PSC04819-R	GCCGTTGACTTCAAGCTTTC
PSC02624	Glutathione S-transferase	PSC02624-F	CCCCAAGAATGATTTTGCAT
		PSC02624-R	TTTGCCATCAACTTCAAGGA
PSC04040	Glutathione S-transferase	PSC04040-F	GGAGCTCCATGGTTTGTCAT
		PSC04040-R	ATGGGCTCCAACAAAATCAA
PSC01063	Aldehyde dehydrogenase	PSC01063-F	GTTGCACGTCGAATTGTTTG
		PSC01063-R	CAAGTTCATCACGCTTTGGA
PSC01095	Aldehyde dehydrogenase	PSC01095-F	TGATTTCGCTGTAGGCCTTT
		PSC01095-R	AACCCCAACAACACCAAGAG
PSC01944	RNA polymerase II	PSRNAPOLIIF	GATGACTTTATGGAAGAAGATGAGG
		PSRNAPOLIIR	CTATGATCACAATTTCGGCAAG
PSC00238	60S ribosomal protein L32	PS60SL32-F	GTTCGTAGACGTTTCAAGGGTACT
		PS60SL32-R	TCGAGATCTCTGACATTATTGACG

2.3 Results

2.3.1 Functional Annotation using BLAST2GO

The working number of 3,982 contigs post filtering were subjected to the BLAST2GO (Conesa et al., 2005) pipeline as well as all further steps discussed below. A BLASTX search of the *P. superbus* consensus sequences against the GenBank nucleotide database (NR), with an e-value cut-off of $1e^{-10}$ was completed. The overall statistics generated from this data set can be seen in Table 2.2.

Table 2.2: Summary of the analysis of EST sequences from a cDNA library prepared from a mixed stage unstressed culture of *P. superbus*. The number of putative unigenes excludes bacterial contaminants and rRNA genes but includes mtDNA genes.

	Number of EST sequences
Number of raw sequences	9,216
Number of high quality sequences	7,606
Average length of high quality sequence	425
Total number of contigs	1,079
Total number of singletons	2,958
Number of putative bacterial contaminant sequences	28
Number of rRNA gene consensus sequences	27
Number of mtDNA consensus sequences	10
Number of putative unigenes	3,982
Number of unigenes with signigicant hits to NEMBASE4, WormPep and NR	1,923
Number of unigenes with significant hits to NR	1,544 (39%)
Number of unigenes with a match to GO	1,373 (35%)
Number of unigenes with a match to EC	1,121 (28%)
Number of unigenes with a match to InterProScan	2,979 (75%)

Just over half the sequences are shown as having no BLAST hits or being novel sequences. Over $1,000$ sequences have been fully annotated with less than 500 shown as having a BLAST hit but no GO mapping or having a BLAST hit, but no annotation. The distribution of annotation identified from this BLAST step can be seen in Figure 2.2. Of all hits found, 82% were to nematodes. The breakdown of species is shown in Figure 2.3. Figure 2.4 shows the e-value distribution versus the number of hits. While the majority of the hits fall between 0 to $1e^{-25}$, there

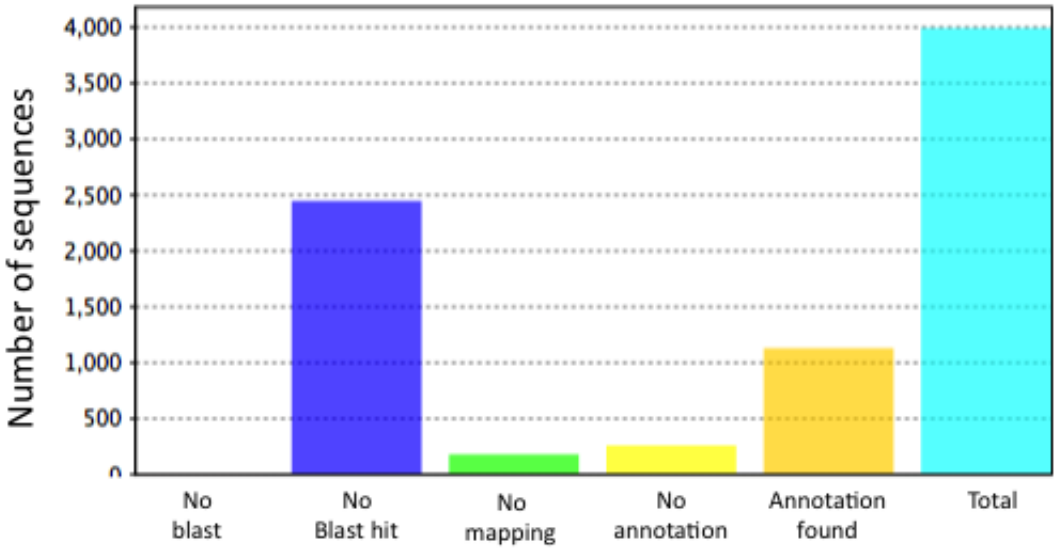


Figure 2.2: Overview of the BLAST2GO results distribution for the *P. superbus* EST sequences.

are some hits with a significant e-value below that.

2.3.2 Most Abundant Contigs

The 44 most abundantly expressed *P. superbus* protein-coding sequences comprise 1,200 ESTs and represent 15.7% of the total EST dataset (Table 2.3). The most abundant sequence, containing 79 ESTs, encodes a member of the nematode specific family of SXP/RAL-2 proteins (Gallin et al., 1989; Rao et al., 2000). These proteins have been detected in the pharyngeal glands and as surface associated antigens in diverse animal parasitic nematodes. Immunisation with recombinant antigens derived from SXP/RAL-2 has been effective in protecting treated animal hosts against filarial worm (Wang et al., 1997), roundworm (Tsuji et al., 2003) and hookworm (Fujiwara et al., 2007) infections. SXP/RAL-2 proteins have also been described in plant parasitic nematodes (Jones et al., 2000; Tytgat et al.,

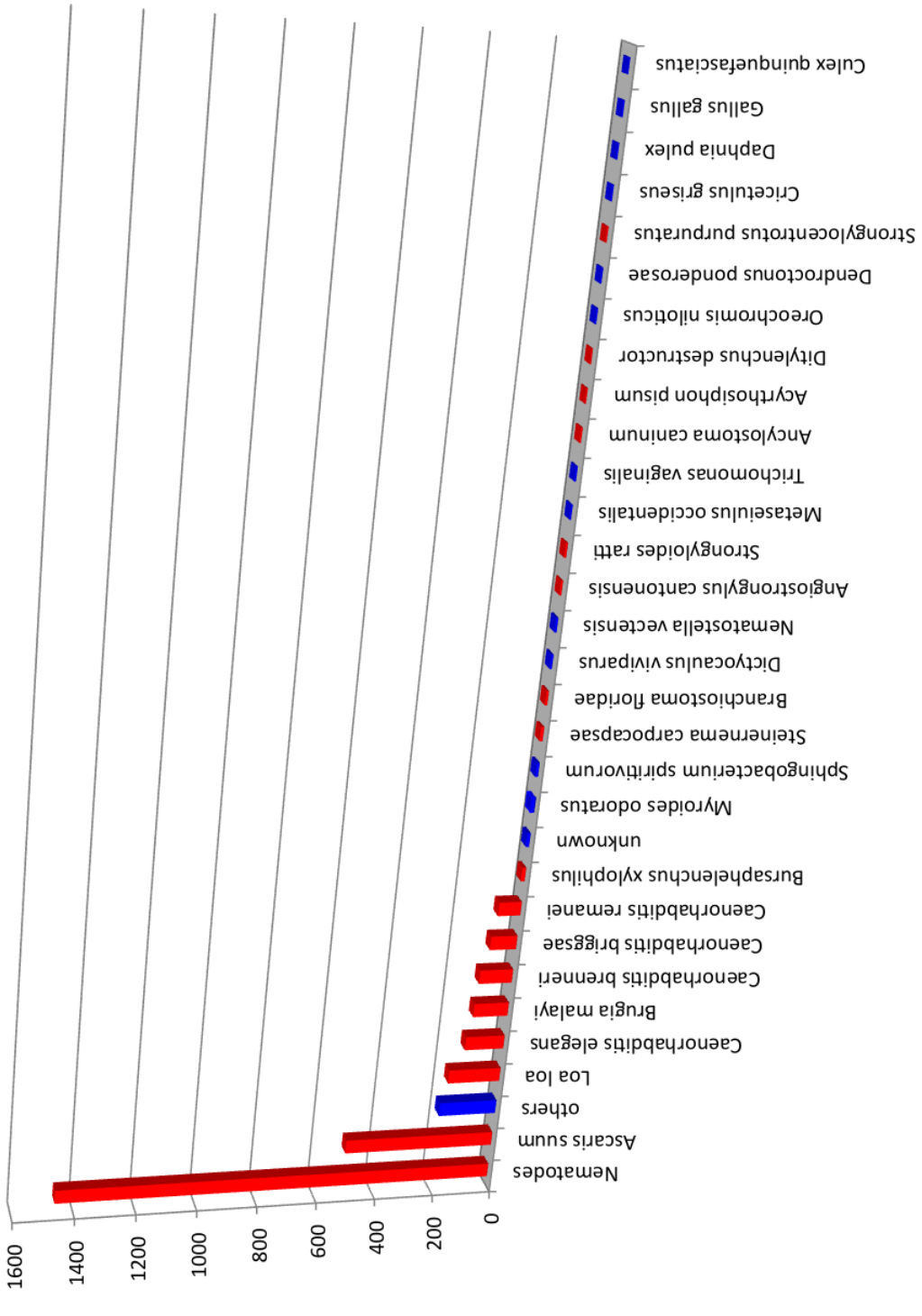


Figure 2.3: Species distribution from BLAST hits to the *P. superbus* EST dataset following a BLASTX of NR with a cutoff of $1e^{-10}$. Nematode species are shown in red.

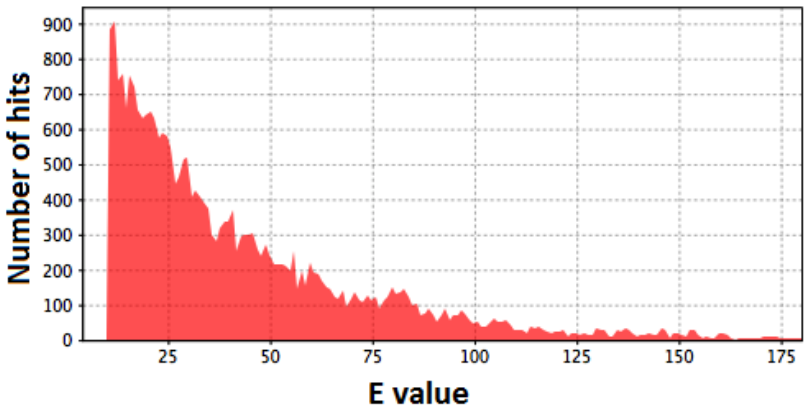


Figure 2.4: The distribution of e-value hits to the *P. superbus* EST dataset shown by number of hits for that e-value.

2005) and SXP/RAL-2 sequences from 11 species of plant parasitic nematodes and are represented in NEMBASE4. GenBank searches show that SXP/RAL-2 homologs also occur in other free-living nematodes in addition to *P. superbus*. Nematode SXP/RAL-2 sequences are likely to be encoded by a small multigene family (Jones et al., 2000; Tytgat et al., 2005). Five SXP/RAL-2 unigenes were detected in *P. superbus*, comprising of 95 ESTs and representing 2.2% of the total EST dataset. With the exception of *Ascaris lumbricoides*, the level of SXP/RAL-2 expression in *P. superbus* was higher than that observed for any of the 19 species of parasites with SXP/RAL-2 homologs in NEMBASE4. SXP/RAL-2 are small (16-21 kDa) basic proteins which share a common domain of unknown function (DUF148, PF02520) (Finn et al., 2008). No RNAi phenotypes have been detected for SXP/RAL-2 homologs in *C. elegans*, but one homolog was among 14 genes which were upregulated in *C. elegans* in response to fungal infection (Pujol et al., 2008). All SXP/RAL-2 sequences characterised to date, including PSC00077, have a signal peptide indicative of a secreted protein. Parasitic nematode studies suggest that these nematode-specific proteins are most likely secreted from the pharyngeal glands onto the surface of the cuticle, where they appear to carry out a structural

or protective function. The very high level of expression of SXP/RAL-2 sequences in *P. superbus* suggests that this cuticular protein may have an important role in anhydrobiotic protection in this nematode.

Ten abundantly expressed sequences were associated with reproductive function, eight corresponding to major sperm protein genes (MSPs) and two to vitellogenin genes. In total, 32 *P. superbus* MSP unigenes and 7 vitellogenin unigenes were detected. In *C. elegans*, MSPs are encoded by a multigene family comprising more than 50 genes (Ward et al., 1988; Tarr & Scott, 2005). Nematode MSPs are small, basic proteins required for the amoeboid movement of sperm. A family of six genes *vit-1* to *vit-6* encode *C. elegans* vitellogenin (Blumenthal et al., 1984; Spieth & Blumenthal, 1985), a major yolk component which is expressed exclusively in the adult hermaphrodite intestine from which it is secreted into the pseudocoelomic space and taken up by oocytes (Kimble & W.J., 1983). Four structural genes were abundantly expressed in *P. superbus*: an actin family member (homolog of *C. elegans act-2*; a gene encoding a core histone protein required for chromatin assembly and chromosome function (Jorcano & Ruizcarrillo, 1979) and genes encoding two proteins associated with the 60S ribosomal subunit. Two abundantly expressed contigs were associated with lipid metabolism. PSC00187 encodes a homolog of *C. elegans* HEH-1 and human NPC2/He1, a cholesterol-binding protein whose deficiency in humans causes Niemann-Pick type C2 disease involving retention of cholesterol in lysosome (Storch & Xu, 2009; Vanier & Millat, 2004). Transcripts for the mitochondrially encoded cytochrome c oxidase subunit 1, essential for oxidative phosphorylation and ATP synthesis, were also highly expressed in this mixed stage *P. superbus* library.

Twenty one of the abundant sequences listed in Table 2.3 are novel. These novel unigenes correspond to 641 ESTs, representing 8.4% of the total EST dataset. Data on the predicted physico-chemical parameters of the putative proteins encoded

by these 21 unigenes are presented in Table 2.4. Thirteen (65%) of these novel unigenes encode a signal peptide indicative of a secreted protein. The association between sequence novelty and likely secretion has been noted previously in the parasitic nematode *Nippostrongylus brasiliensis* (Harcus et al., 2004). Three of the putative novel proteins are predicted to be natively unfolded over 80-100% of their primary sequence. The *P. superbus* dataset contains a total of 2,059 novel unigenes. Further analysis of these novel sequences is presented in a later section.

Table 2.3: Most abundantly represented transcripts in the dataset of 7,606 ESTs prepared from a mixed stage unstressed culture of *P. superbus.*

Cluster ID	Best BLAST hit [Organism]	Accession Number	Blast Score	E value	Number of ESTs (% of Total Dataset)
PSC00077	Immunodominant antigen (SXP/RAL-2 protein) [*A. caninum*]	ABD98404.1	3.04E+02	3.00E-26	79 (1.04)
PSC00006	No significant similarity found	-	-	-	71 (0.93)
PSC00009	No significant similarity found	-	-	-	70 (0.92)
PSC00076	No significant similarity found	-	-	-	54 (0.71)
PSC00137	No significant similarity found	-	-	-	49 (0.64)
PSC00511	Major sperm protein [*C. elegans*]	NP_494858.1	635	1.00E-64	40 (0.53)
PSC00155	Hypothetical protein (Y105C5A.8) [*C. elegans*]	NP_001041003	430	1.00E-40	39 (0.51)
PSC00051	No significant similarity found	-	-	-	35 (0.46)
PSC01915	Major sperm protein [*C. elegans*]	NP_494858.1	640	3.00E-65	30 (0.39)
PSC00043	Yolk protein (vitellogenin) (CEW1-vit-6) [*Oscheius sp.*]	U35449	280	6.00E-23	29 (0.38)
PSC00163	No significant similarity found	-	-	-	29
PSC00182	No significant similarity found	-	-	-	29
PSC00025	No significant similarity found	-	-	-	28 (0.37)
PSC00165	Major sperm protein [*C. elegans*]	NP_494858.1	640	3.00E-65	27 (0.35)
PSC00883	No significant similarity found	-	-	-	27
PSC00633	No significant similarity found	-	-	-	26 (0.34)
PSC00187	Lysosomal protein (heh-1) [*C. elegans*]	NP_497671.2	295	3.00E-25	25 (0.33)
PSC00252	Expressed sequence tag [*M. chitwoodi*]	MCP06382	358	2.00E-32	25
PSC00203	No significant similarity found	-	-	-	25
PSC00316	No significant similarity found	-	-	-	25
PSC00610	No significant similarity found	-	-	-	24 (0.31)
PSC00167	Vitellogenin (vit-5) [*C. elegans*]	NP_508589	238	2.00E-18	23 (0.30)
PSC00241	No significant similarity found	-	-	-	22 (0.29)
PSC00429	No significant similarity found	-	-	-	22
PSC00179	Major sperm protein [*C. elegans*]	NP_501781.1	634	1.00E-64	21 (0.28)
PSC00876	Lipid binding protein (lbp-3) [*C. elegans*]	NP_001041249	360	9.00E-33	21
PSC00004	No significant similarity found	-	-	-	21
PSC00010	No significant similarity found	-	-	-	21
PSC00566	Major sperm protein [*C. elegans*]	NP_494898	633	2.00E-64	20 (0.26)
PSC00265	60S ribosomal protein L7a [*Loa loa*]	XP_003139379.1	8.83E+02	e-93	19 (0.25)
PSC00457	Actin family member [*Panagrellus redivivus*]	AAM47606.1	1.19E+03	3.00E-148	19
PSC00015	Histone H4d [*Xenopus laevis*]	NP_001128541.1	4.69E+02	2.00E-45	17 (0.22)
PSC00128	Cytochrome c oxidase subunit 1 [*Chabertia ovina*]	YP_003434131.1	1300	3.00E-141	17
PSC00326	Major sperm protein [*B. malayi*]	XP_001902608.1	2.10E+02	2.00E-15	17
PSC00122	No significant similarity found	-	-	-	17
PSC00047	60S ribosomal protein L37a [*B. malayi*]	XP_001902009.1	416	3.00E-39	16 (0.21)
PSC00764	Expressed sequence tag [*M. chitwoodi*]	MCP06382_1	5.22E+02	4.00E-51	16
PSC00097	No significant similarity found	-	-	-	16
PSC00962	Expressed sequence tag[*Angiostrongylus cantonensis*]	AAC00593_1	198	6.00E-14	15 (0.19)
PSC00184	Major sperm protein [*Dictyocaulus viviparus*]	ABW37697.1	624	2.00E-63	15
PSC00064	Eukaryotic elongation factor 1A [*B. xylophilus*]	ACZ13348.1	1,348	8.00E-147	15
PSC00486	Major sperm protein [*C. elegans*]	NP_494858.1	640	3.00E-65	15
PSC00127	No significant similarity found	-	-	-	15
PSC00725	No significant similarity found	-	-	-	15

Table 2.4: Putative anhdyrobiotic and stress response genes constitutively expressed by unstressed *P. superbus*.

Description	Number of Clusters	Number of ESTs
Signal Transduction		
Mitogen-activated protein kinases	3	4
Serine/threonine protein kinases	12	15
Casein kinases	10	15
Protein tyrosine kinases	6	7
Other protein kinases	4	4
Transcription factors/activators	6	9
Total	**41**	**54**
Antioxidant Activity		
Manganese superoxide dismutase (sod-2)	1	2
Glutathione peroxidase	3	8
Peroxiredoxin	2	4
Glutathione S-transferase	7	11
Glutaredoxin	2	2
Thioredoxin	1	1
Aldehyde dehydrogenase	2	4
Aldo/keto reductase	2	2
NADP Isocitrate dehydrogenase	1	1
Total	**21**	**35**
Late Embryogenesis Abundant Proteins		
Total	**13**	**34**
Heat Shock Proteins (HSP)		
HSP90 family	3	10
HSP70 family	6	13
HSP60 family	1	1
HSP40/DNaJ family	9	14
Small heat shock protein/α-crystallin family	4	6
HSP90 co-chaperone Cdc37	1	1
HSP70 co-chaperone BAG1	1	2
Tetratricopeptide repeat containing protein	1	1
Total	**26**	**48**
Other Chaperone/Chaperonin Proteins		
Mitochondrial chaperone BCS1 family member	1	1
Mitochondrial prohibitin complex protein 2	1	1
Protein disulfide isomerase	3	9
Cyclophylin family member	5	7
Derlin-2	1	1
DJ-1 family protein	1	2
Prefoldin subunit	2	3
Cytosolic T-complex protein 1	2	3
Putative α-tubulin folding cofactor B	1	1
Total	**17**	**28**
The Ubiquitin System		
Ubiquitin family protein	8	16
Ubiquitin-conjugating enzyme E1	2	2
Ubiquitin-conjugating enzyme E2	5	6
E3 Ubiquitin ligase	5	9
Ubiquitin elongating factor E4	1	1
Ubiquitin carboxyl-terminal hydrolase	2	2
Ubiquitin fusion degradation protein UFD1	1	1
Total	**25**	**38**
The Proteasome		
Proteasome subunit alpha family	4	6
Proteasome subunit beta family	4	6
Proteasome regulatory subunit family	15	24
Total	**23**	**36**
Autophagy		
Autophagy-related protein 2-like (atg2)	1	1
LC3, GABARAP AND GATE-16 family member (lgg-1)	1	1
Total	**2**	**2**
DNA Damage Response Proteins		
Total	**12**	**12**
Others		
Aquaporin related family member	2	2
Ezrin/Radixin/Moesin family member (erm-1)	2	2
Thaumatin family member (thn-3)	1	1
AN1-like Zinc finger family protein	1	1
RIC1 Putative stress responsive protein	1	1
Mitochondrial Lon protease	1	1
Total	**8**	**8**
Total	**187**	**294**

2.3.3 Assignments to Metabolic Pathways using KEGG

One thousand six hundred and eighty four KEGG orthology assignments were inferred by searching for *P. superbus* unigenes that have homologs among the default set of manually curated eukaryotic genes in the KEGG database (which contains 26 genomes); similarly, 1,412 KEGG assignments specific to the *C. elegans* genome were also inferred (Table 2.5).

KEGG pathways associated with metabolism had the highest representation, with a large number of the *P. superbus* sequences associated with 'carbohydrate metabolism', 'energy metabolism', 'lipid metabolism' and 'amino acid metabolism' pathways. In the environmental information processing category, 'signal transduction' was highly represented. Other highly represented pathways were found in the genetic information processing category including 'translation' and 'folding, sorting and degradation' and a large number of sequences had KEGG assignments to the human neurodegenerative disease sub-category. Many neurodegenerative diseases are associated with the dysfunction or overload of the protection systems responsible for repairing or degrading damaged proteins and macromolecules (Martínez et al., 2010; Irvine et al., 2008; Hegde & Upadhya, 2007). Cells exposed to severe water stress experience serious damage to their macromolecules and membranes; proteins lose their structures, become unfolded and aggregate. Thus anhydrobiotic organisms are adapted to survive cellular dehydration by deploying efficient cellular protection and repair systems and it is likely that some gene products that have roles in anhydrobiotic protection in nematodes may also have human homologs which are required for neural survival. For example the molecular chaperone DJ-1, which is associated with familial Parkinson's disease (Bonifati et al., 2003; Shendelman et al., 2004), is also upregulated in response to desiccation stress in the anhydrobiotic nematode *A. avenae* (Reardon et al., 2010); and AAvLEA1, a natively unfolded LEA protein which is upregulated in

Table 2.5: Summary of KEGG orthology assignments of *P. superbus* unigenes to biochemical pathways.

KEGG Pathway Category	**Eukaryotes**	**C. elegans**
1. Metabolism	**474**	**438**
1.1 Carbohydrate Metabolism	113	102
1.2 Energy Metabolism	70	68
1.3 Lipid Metabolism	61	47
1.4 Nucleotide Metabolism	36	33
1.5 Amino Acid Metabolism	76	71
1.6 Metabolism of Other Amino Acids	26	24
1.7 Glycan Biosynthesis and Metabolism	21	19
1.8 Metabolism of Cofactors and Vitamins	22	18
1.9 Biosynthesis of Polyketides and Terpenoids	7	9
1.10 Biosynthesis of Secondary Metabolites	13	14
1.11 Xenobiotics Biodegradation and Metabolism	29	33
2. Genetic Information Processing	**294**	**272**
2.1 Transcription	50	43
2.2 Translation	132	124
2.3 Folding, Sorting and Degradation	91	87
2.4 Replication and Repair	21	18
3. Environmental Information Processing	**88**	**62**
3.1 Membrane Transport	4	3
3.2 Signal Transduction	73	53
3.3 Signalling Molecules and Interaction	11	6
4. Cellular Processes	**236**	**163**
4.1 Transport and catabolism	91	73
4.2 Cell Motility	19	10
4.3 Cell Growth and Death	63	39
4.4. Cell Communication	63	41
5. Organismal Systems	**250**	**181**
5.1 Immune System	52	38
5.2 Endocrine System	60	49
5.3 Circulatory System	21	15
5.4 Digestive System	44	31
5.4 Excretory System	15	14
5.5 Nervous System	20	10
5.6 Sensory System	19	11
5.7 Development	10	5
5.8 Environmental Adaptation	9	8
6. Human Diseases	**342**	**296**
6.1 Cancers	63	39
6.2 Immune System Diseases	22	21
6.3 Neurodegenerative Diseases	134	130
6.4 Cardiovascular Diseases	36	33
6.5 Metabolic Diseases	2	3
6.6 Infectious Diseases	85	70
Total (Unigenes)	**1,684 (854)**	**1,412 (714)**

response to desiccation stress in *A. avenae* (Browne et al., 2002), has been shown *in vitro* to protect complex mixtures of proteins from aggregation (Chakrabortee et al., 2007).

2.3.4 Gene Ontology Assignments

The Gene Ontology consortium has developed a vocabulary of defined terms that describe gene products in the context of three domains: biological process, molecular function and cellular component in a species-independent manner (Consortium, 2000). The representation of GO terms as found by BLAST searches of the *P. superbus* unigenes against genes in the GO database are presented using a filter of at least 100 sequences. A pie chart was generated for each category; biological process (Figure 2.5), molecular function (Figure 2.6) and cellular component (Figure 2.7). In summary, these representations consist of 3,148, 1,671 and 1,245 hits respectively. Several of these hits were to gene products whose descriptions indicate roles in anhydrobiotic protection.

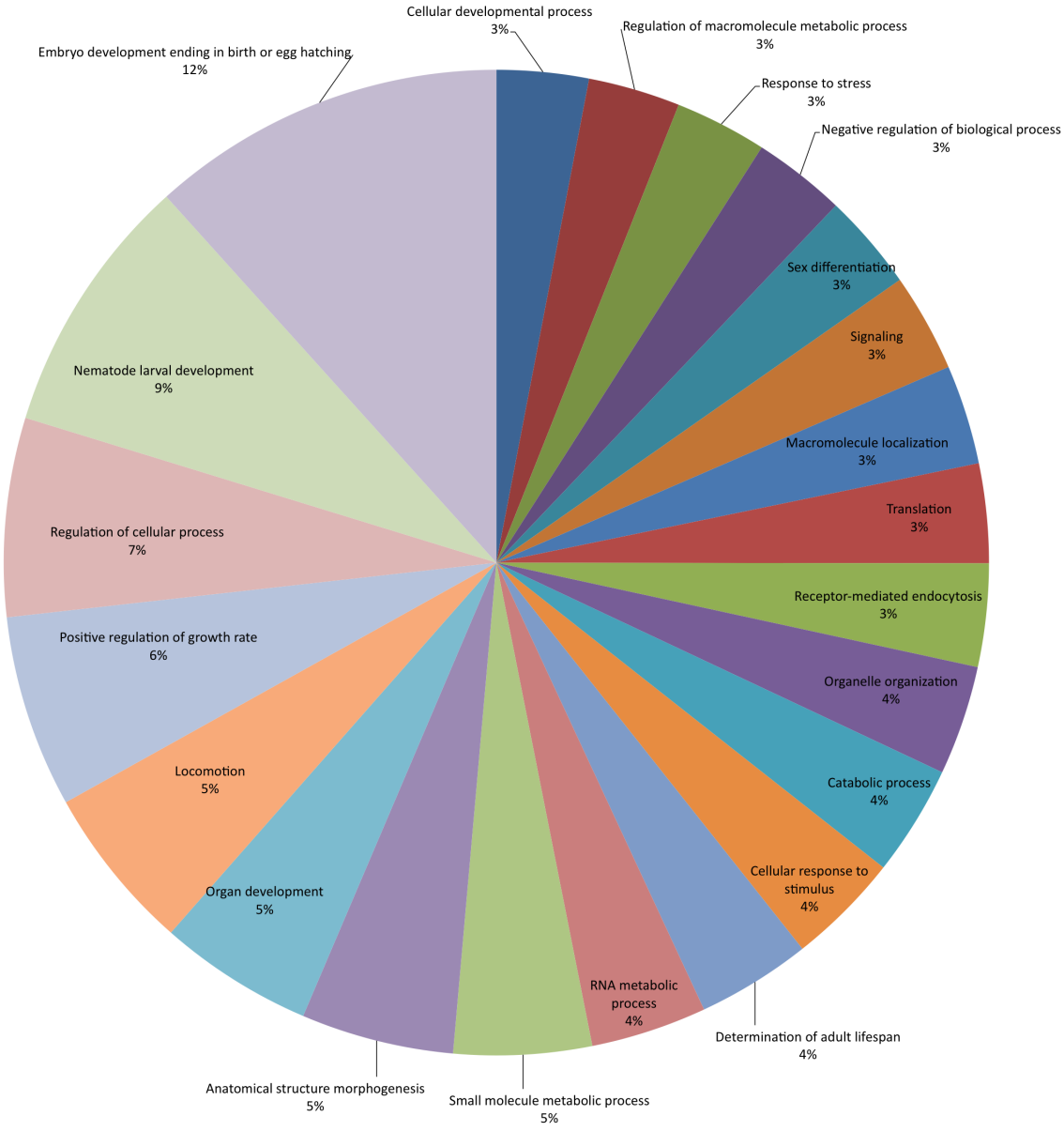


Figure 2.5: Biological process categories for the *P. superbus* unigenes having at least 100 EST sequences in each category. 3,148 hits are represented.

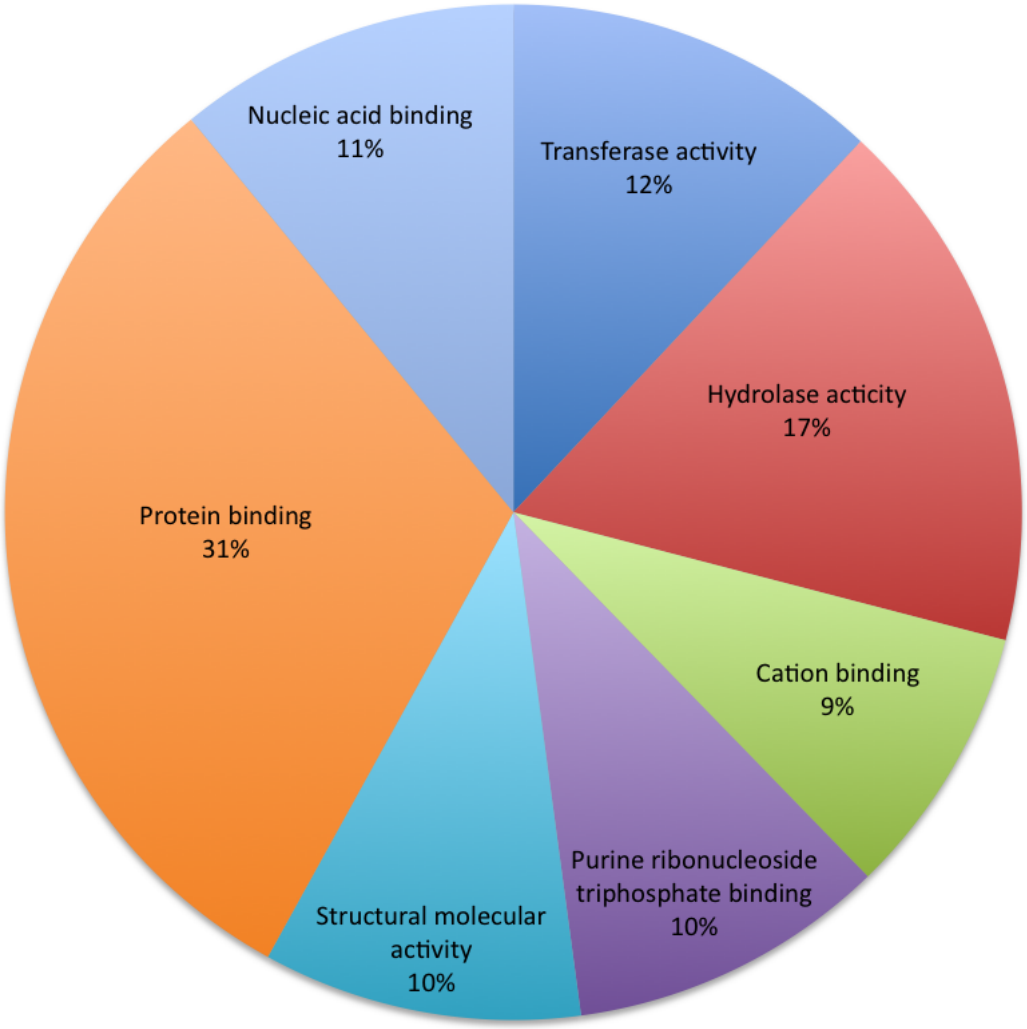


Figure 2.6: Molecular function categories for the *P. superbus* unigenes having at least 100 EST sequences in each category. 1,671 hits are represented.

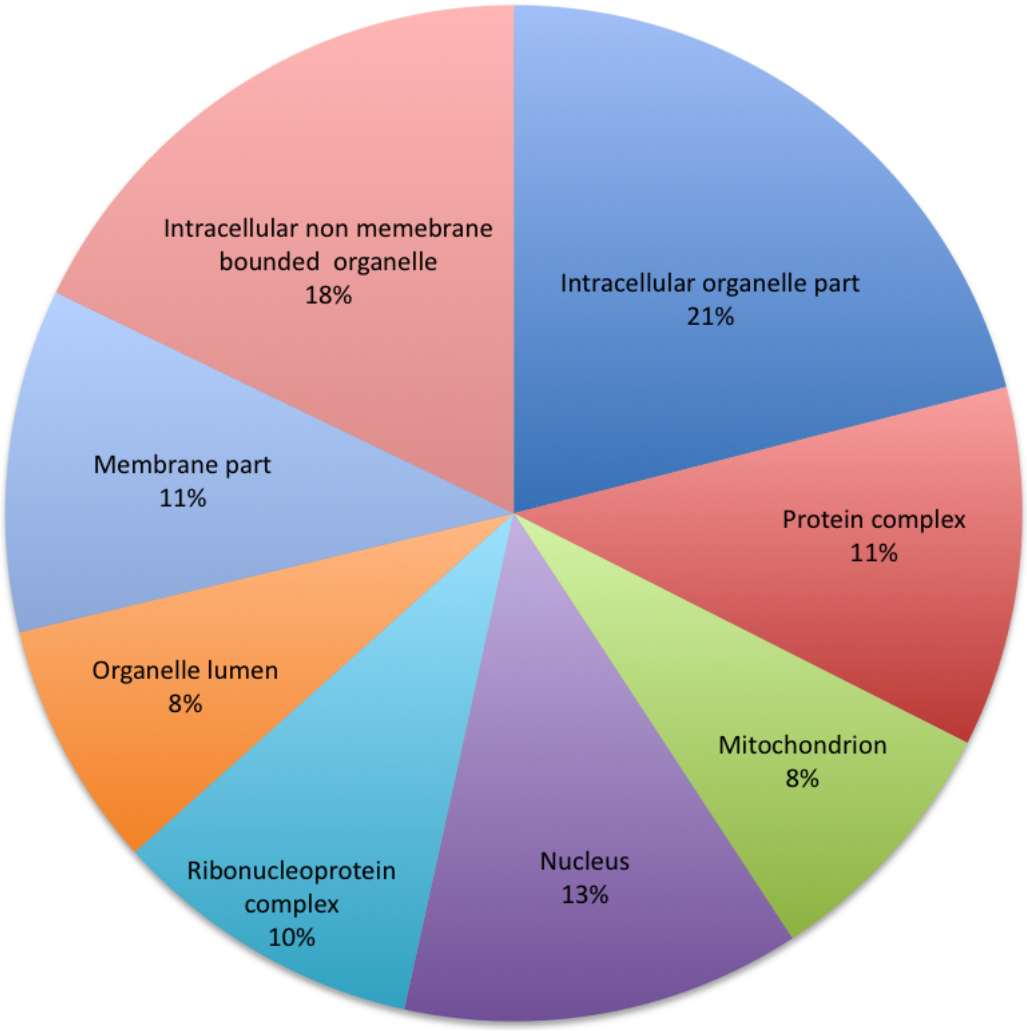


Figure 2.7: Cellular component categories for the *P. superbus* unigenes having at least 100 EST sequences in each category. $1,245$ hits are represented .

2.3.5 Putative Anhdyrobiotic and Stress Response Genes

When cells suffer severe dehydration, metabolism ceases, macromolecules denature, membranes undergo phase changes and fuse with other normally separate membranes. Unlike desiccation sensitive taxa, anhydrobiotes have evolved mechanisms which maintain the structure and integrity of macromolecules and membranes in the absence of water and also during rehydration and revival. Comparative studies of the desiccation tolerance phenotypes of anhydrobiotes show lineage specific differences in the response patterns and biochemical adaptations, which implies that anhydrobiotic phenotypes can be achieved in different taxa by the expression of functionally equivalent molecules. Based on currently available data from nematodes and other anhydrobiotic animals, a model showing the possible steps involved in the detection and expression of anhydrobiotic protection mechanisms in nematodes is presented and shown in Chapter 1. Using GO, KEGG and BLAST description data, the components of this model in the *P. superbus* unigene dataset have been assembled and manually annotated and a set of 187 candidate genes whose products may be involved in the anhydrobiotic response of *P. superbus* are presented. This dataset is summarised in Table 2.4.

2.3.6 Signal Transduction, Protein Kinases and Transcription Factors

The *P. superbus* unigene data set contains 35 protein kinases. Among these were three contigs encoding MAP kinases. One of these, PSC00478, is a homolog of MAPKAP kinase-2 which is responsible for the phosphorylation of the small heat-shock proteins HSP27 (Stokoe et al., 1992) and α-B-crystallin (Ito et al., 1997). The phosphorylation of these small HSPs in response to stresses such as heat shock and oxidation is proposed to regulate actin filament dynamics and to stabilise

microfilaments (Dalle-Donne et al., 2001). Two unigenes encode members of the STE20 family of serine/threonine kinases (Strange et al., 2006). One of these, PSC03670, has high BLAST identity (3e-51) to the *gck-3* gene whose function is required by *C. elegans* for volume recovery and survival after exposure to extreme hypertonic stress (Choe & Strange, 2007). Nine unigenes encode putative casein kinases, important regulatory molecules in cell division and differentiation and in DNA damage repair (Cheung et al., 2005; Knippschild et al., 2005). Casein kinase 2 (CK2) was found to be upregulated in response to desiccation stress in the nematode *Steinernema feltiae* (Gal et al., 2003). Other *P. superbus* stress-related kinases include a homolog of *akt-1*, a regulatory component within the insulin/IGF-1 signaling pathway (Paradis & Ruvkun, 1998; Padmanabhan et al., 2009), which plays an important role in regulating nematode life span, dauer formation and stress tolerance and diacetylglycerol (DAG) kinase, which modulates DAG levels in the cell membrane, regulating intracellular signalling proteins that have evolved the ability to bind this lipid (Merida et al., 2008). DAG kinase is also activated in plants during cold and osmotic stress (Arisz et al., 2009).

Several *P. superbus* unigenes are predicted to encode transcription factors. Among these are a forkhead protein, a member of a conserved family of transcriptional regulators of cellular processes including metabolism, ageing, apoptosis, cell cycle progression and stress resistance (Sedding, 2008; Smith & Shanley, 2010). Two unigenes encode putative jumonji (JmjC) domain-containing proteins. The *C. elegans jmjc-1* gene functions as a transcriptional activator of stress related genes in response to multiple stimuli, including heat-shock and osmotic and oxidative stress (Kirienko & Fay, 2010). Two unigenes encode putative high mobility group (HMG) proteins, which are important in modulating chromatin structure and gene expression. HMG transcripts are upregulated in response to desiccation, osmotic, heat and cold stresses in the anhydrobiotic nematode *A. avenae* (Reardon

et al., 2010) and HMG function is required for the transcription of stress-responsive genes in *Arabidopsis thaliana* (Lildballe et al., 2008).

2.3.7 Anti-oxidant Activity

Reactive oxygen species (ROS) accumulate in cells as a result of cellular dehydration (Kranner & Birtic, 2005; Franca et al., 2007). ROS cause oxidative damage to proteins, lipids, DNA and other macromolecules. Therefore, proteins with antioxidant properties are required to rapidly neutralise ROS immediately after they are formed. Twenty-one unigenes encoding proteins that fall into this category were characterised: PSC00113 encodes a manganese superoxide dismutase responsible for converting oxygen radicals into hydrogen peroxide; three unigenes encode glutathione peroxidases which function to reduce hydrogen peroxide; and sequences encoding both the 1-Cys and 2-Cys class of peroxiredoxin enzymes (whose main function is the reduction of peroxides (Wood et al., 2003)) were also identified.

The tripeptide glutathione (GSH) functions as a co-factor for the antioxidant enzymes glutaredoxin (Grx) and glutathione S-transferase (GST). GSTs catalyse the conjugation of glutathione to reactive electrophilic compounds from endogenous and xenobiotic sources and are thus capable of detoxifying a large variety of cytotoxic molecules (Hayes et al., 2005). The *P. superbus* dataset contains seven GST unigenes, four from the sigma class of cytosolic GSTs and two from the kappa class of mitochondrial GSTs. The GST sigma and kappa classes are considered to be involved in protection against endogenously produced ROS (Corona & Robinson, 2006). PSC02300 encodes a Grx enzyme. Protein deglutathionylation is carried out by Grx (Gallogly & Mieyal, 2007), and this enzyme can also reduce the disulphide bridges of oxidised proteins (Meyer et al., 2009). Such disulphide bridges can also be reduced by thioredoxin (PSC00712), which shares a similar structure and overlapping function with Grx (Holmgren, 1989). Other *P.*

superbus unigenes, whose gene products are likely to be involved in antioxidant activity and redox regulation, include: aldehyde dehydrogenase which deactivates malondialdehyde, an important end product of lipid peroxidation; two aldo-keto reductase sequences (Chang & Petrash, 2008; Malik & Storey, 2009) and a putative cytosolic NADP-isocitrate dehydrogenase (NADP-ICDH). NADP-ICDH catalyses the production of NADPH and, by supplying NADPH to the antioxidant systems, NADP-ICDH is an important component in the control of redox balance and the modulation of oxidative damage in the cytosol (Lee et al., 2002; Leterrier et al., 2007).

2.3.8 Late Embryogenesis Abundant Proteins

Thirteen *P. superbus* unigenes encode predicted LEA proteins (Table 2.6). Although LEA proteins have been shown to accumulate during the onset of desiccation in anhydrobiotic animals, including nematodes (Browne et al., 2002; Gal et al., 2003; Adhikari et al., 2009), genes encoding LEA proteins are particularly numerous and heterogeneous in plant genomes. For example, the *Arabidopsis thaliana* genome contains 51 LEA genes placed into 9 different Pfam groups (Hundertmark & Hincha, 2008; Bies-Ethève et al., 2008). In animal genomes LEA genes are less abundant and predominantly belong to the Group 3 LEA family (Pfam F02987) (Tunnacliffe & Wise, 2007), with members of LEA Group 1 (PF00477) being described to date only in the brine shrimp *Artemia franciscana* (Sharon et al., 2009) and in an unspecified tardigrade species (Forster et al., 2009). Group 3 LEA proteins are highly hydrophilic and largely lacking in secondary structure when fully hydrated (Tunnacliffe & Wise, 2007). Group 3 proteins also contain blocks of tandemly repeated 11-mer amino acid motifs (Dure, 1993), the number of repeats per LEA protein typically ranging in number from 5-24 (Brown et al., 2004).

Table 2.6: BLASTX similarity searches of *P. superbus* unigenes predicted to encode LEA proteins against the NCBI NR and LEAP

Contig ID and (No of ESTs)	Best BLAST hit and (Organism)	Database	Accession No	BLAST Bit Score	e-value	Disordered AAs (%)	GRAVY Index No
PSC00061 (10)	LEA protein (*C. briggsae*)	NCBI nr	CAP25449	82.8	4.00E-14	86%	0.765
PSC00061	LEA protein (*A. thaliana*)	LEAPdb	AAL59922	74.3	8.00E-16		
PSC00416 (6)	LEA protein K08H10.1e (*C. elegans*)	NCBI nr	CCA65580	60.8	1.00E-07	18%	-0.618
PSC00489	LEA protein (*C. elegans*)	LEAPdb	AAB69446	80.5	9.00E-18		
PSC00489 (2)	LEA3 protein (*Glycine max*)	NCBI nr	CAA80491.1	68.6	2.00E-10	10%	-0.81
PSC00416	LEA protein (*C. elegans*)	LEAPdb	AAB69446	74.7	3.00E-16		
PSC00514 (2)	LEA3 protein (*G. max*)	NCBI nr	CAA80491.1	64.7	5.00E-12	9%	-0.854
PSC00514	LEA protein (*Pisum sativum*)	LEAPdb	CAF32327	94	9.00E-22		
PSC00782 (2)	Hypothetical protein (*C. briggsae*)	NCBI nr	CAP25465	66.2	9.00E-12	100%	-1.367
PSC00782	LEA-like protein (*A. thaliana*)	LEAPdb	BAB10116	67.4	4.00E-22		
PSC01414 (1)	LEA protein (*C. briggsae*)	NCBI nr	XP_002637990.1	84.7	5.00E-15	100%	-1.424
PSC01414	LEA-like protein (*A. thaliana*)	LEAPdb	BAD43695	76.6	1.00E-16		
PSC01455 (2)	Hypothetical protein NCU01912 (*Neurospora crassa*)	NCBI nr	XP_965543.1	62	7.00E-08	99%	-1.034
PSC01455	LEA-like protein (*A. thaliana*)	LEAPdb	NP_193834	61.2	6.00E-12		
PSC01720 (1)	Predicted protein Gls24 (*Gemella haemolysans*)	NCBI nr	ZP_04776234.1	78.2	8.00E-13	40%	-0.623
PSC01720	LEA-like protein (*A. thaliana*)	LEAPdb	BAD43695	90.5	9.00E-21		
PSC01853 (1)	Hypothetical protein (*H. influenzae*)	NCBI nr	ZP_01786547	67	1.00E-09	100%	-0.988
PSC01853	Hypothetical protein (*H. influenzae*)	LEAPdb	ZP_01786547	67	9.00E-14		
PSC03871 (2)	Hypothetical protein K08H10.1f (*C. elegans*)	NCBI nr	CCA65610	42.4	3.00E-07	77%	-0.656
PSC03871	LEA group 3 protein (*Lindernia brevidens*)	LEAPdb	ACA49509	54.7	5.00E-10		
PSC04142 (1)	Hypothetical protein (*Toxoplasma gondii*)	NCBI nr	EEE21041	55.8	2.00E-06	100%	-1.34
PSC04142	LEA-like protein (*A. thaliana*)	LEAPdb	BAD43695	50.4	4.00E-09		
PSC04118 (1)	Protein At5g44310 (*A. thaliana*)	NCBI nr	AAS49101	47	1.00E-05	100%	-1.405
PSC04118	LEA-like protein (*A. thaliana*)	LEAPdB	BAD43695	47	2.00E-09		
PSC04695 (2)	Hypothetical protein K08H10.1f (*C. elegans*)	NCBI nr	CCA65610	73.6	1.00E-11	100%	-1.297
PSC04695	LEA-like protein (*A. thaliana*)	LEAPdb	BAD43695	67.4	7.00E-14		

A database of LEA proteins was established recently (Hunault & Jaspard, 2010). Although 89% of the sequences in this database are from land plants, the LEAPdb database includes LEA sequences from animal taxa. Among BLAST searches against the LEAPdb (Table 2.6), ten *P. superbus* unigenes had best hits to LEA3 proteins from plant species, one unigene was most similar to a putative LEA protein from *H. influenzae* (Hogg et al., 207) and two to an LEA3 protein from *C. elegans*, the abundance of *P. superbus* sequences which had hits to plant genes (using a default E value) may be a consequence of the large number of plant LEA sequences represented in the LEAP database. The LEA sequences encoded by six of the *P. superbus* unigenes are predicted to be 100% natively unfolded. Three of the 13 *P. superbus* LEA sequences are predicted to lack substantial regions of unfolded structure, however, all 13 LEA sequences had negative GRAVY (Grand Average of Hydropathy) indices characteristic of hydrophilic proteins. All the predicted sequences showed evidence of tandemly repeated 11-mer amino acid motifs. The *C. elegans* genome has been reported to contain three LEA genes (Brown et al., 2004) and four LEA genes have been detected in the *C. briggsae* genome (Brown et al., 2004). The best characterised nematode LEA protein is AAv1 which is upregulated in the nematode *A. avenae* in response to desiccation. AAv1 has been shown to protect complex mixtures of proteins from aggregation in vitro and in vivo (Chakrabortee et al., 2007). It is possible that some of the *P. superbus* unigenes reported here may represent alternatively spliced forms of a single LEA gene. However, the relative abundance of LEA genes in *P. superbus* as compared to *C. elegans*, along with their constitutive expression (34 LEA-encoding are ESTs), suggest that LEA Group 3 proteins are an important component of the anhydrobiotic protection repertoire of *P. superbus*.

2.3.9 Molecular Chaperones & Unfolded Protein Response

Heat shock proteins (HSPs) are essential for the correct folding and maturation of a great diversity of client proteins and for protecting proteins from stress induced unfolding and aggregation (Morimoto, 2008; Richter et al., 2010). Eukaryotic HSP families contain multiple genes, which may be either constitutively expressed or stress inducible and targeted to specific cellular compartments (Kabani & Martineau, 2008; Vos et al., 2008). The HSP expression repertoire of an anhydrobiotic organism may thus be very important in maintaining the integrity of the proteome during the dehydration and recovery phases of anhydrobiosis (Sales et al., 2000; Jonsson & Schill, 2007; Cho & Choi, 2009; Hu et al., 2003). The *P. superbus* dataset contains representatives of all the HSP classes characteristic of nematodes, including four distinct small heat shock proteins (sHSP). sHSP are the major "holding" chaperones, retaining unfolding proteins in a conformation suitable for subsequent refolding, thus preventing their irreversible aggregation (Stengel et al., 2010; Eyles & Gierasch, 2010). Anhydrobiotic encysted larvae of the brine shrimp *Artemia franciscana* accumulate large quantities of a sHSP known as p26 which constitutes 15% of the non-yolk protein in these larvae (Liang et al., 1997). *A. franciscana* cysts are resistant to desiccation, high temperature, gamma-irradiation and anoxia, and the chaperoning activity of p26 is likely to be a very significant component of this remarkable stress resistance (Sun et al., 2006).

The accumulation of unfolded proteins in the endoplasmic reticulum (ER) arising from physiological or abiotic stress and leads to the expression of several protein folding chaperones, including members of the HSP90 and HSP70 families and their co-chaperones (Schroder, 2008). Unfolded protein response (UPR) chaperones from the *P. superbus* dataset include three protein disulfide isomerases (PDI), which catalyse the formation and isomerisation (rearrangement) of cysteine bonds during protein folding (Winter et al., 2007; Karala et al., 2007); five cyclophilin-

type peptidyl-prolyl cis-trans isomerases which catalyse the isomerisation of the peptide bonds preceeding proline residues, and a homolog of Derlin-2 which is required for the degradation of misfolded glycoproteins in the ER (Oda et al., 2006). Five *P. superbus* unigenes encode proteins required for the facilitated folding of actin and tubulin to form microtubules: two prefolding subunits, two subunits of cytosolic T-complex protein 1 and alpha-tubulin folding cofactor B. Changes in microtubule dynamics have been shown to occur during osmotic stress in *Zea mays* (Lü et al., 2007) and during desiccation in *Brassica napus* (Bagniewska-Zadworna, 2008); it is also possible that adjustments to the stability of the microtubule cytoskeleton are also required by *P. superbus* for successful entry into anhydrobiosis.

2.3.10 Removal of Damaged Proteins - the Ubiquitin - Proteasome (UPS) and Autophagy Systems

The importance of the proteasomal system to unstressed nematodes is also apparent from its abundant representation in the *P. superbus* EST dataset, which contains 44 UPS unigenes comprising 68 ESTs. In contrast, autophagy genes are not well represented in unstressed *P. superbus*. Only two *P. superbus* homologs of the 19 core *C. elegans* autophagy genes were detected (Kovacs & Zhang, 2010).

2.3.11 A Comparison of the *P. superbus* EST Unigene Dataset with EST Datasets from other Anhydrobiotic Nematodes

The 3,982 *P. superbus* unigenes were compared to EST unigene datasets from three other species of anhydrobiotic nematodes (Karim et al., 2009; Adhikari et al., 2009) and (Haegeman et al., 2009) to identify putative homologous protein families which

may reveal some of the core anhydrobiotic processes shared by these nematodes. *Plectus murrayi* is an Antarctic soil nematode adapted to survive desiccation and freezing (Adhikari et al., 2009). *A. avenae* is a slow desiccation strategist soil dwelling fungiverous nematode. *D. africanus* is an endoparasite of plants with peanut as its primary host. It migrates to the pods and seeds of the ground nut and can survive in an anhydrobiotic state in the seeds (Adhikari et al., 2009). The phylogenetic relationships of these nematodes are indicated in Chapter 1.

The combined dataset from the four anhydrobiotic nematodes comprised 10,791 unigenes. All against all BLASTP analyses of the predicted peptide sequences for these unigenes, followed by their classification into putative homologous groups using the TRIBE Markov clustering algorithm as implemented in the MCL software package, has identified 7,063 unigene families, where 6,308 consist of singletons. The distribution of these unigene families across the four nematode species is summarised in Figure 2.8. A total of 67 unigene families contain transcripts from all four anhydrobiotic nematodes. While the analysis is based on an incomplete coverage of the transcriptomes of all four nematodes, these 67 families provide a first indication of subsets of genes common to the four species, some of which may be involved in anhydrobiotic processes. These families include representatives of several of the anhydrobiotic and stress response proteins discussed in the previous section. Among these are protein kinases and HMG proteins; glutathione S-transferase; sHSP, HSP70; HSP90; peptidyl-prolyl cis-trans isomerase; several components of the UPS system and RIC1, a poorly characterised family which encodes plasma membrane proteins that are expressed in response to high salt or low temperature conditions in plants (Navarre & Goffeau, 2000). Members of the nematode specific transthyretin-related *ttr* family (Jacob et al., 2007) are also included among the 67 unigenes. The function of *ttr* genes remains elusive; so far, the function of just one nematode *ttr* gene product has been discovered (TTR-52

mediates the recognition and engulfment of apoptotic cells in *C. elegans* (Wang et al., 2010*a*)). Since this analysis is based on partial transcriptomes of the four nematodes the results need to be interpreted conservatively; however, the data shows that these four anhydrobiotic nematodes express a great diversity of stress responsive genes. Surprisingly, none of the 13 LEA unigenes were common to all four nematode datsets, and 8 LEA sequences were found only in *P. superbus*. This may indicate that constitutive expression of LEA transcripts is higher in *P. superbus* than in the three other anhydrobiotic nematodes. When more complete coverage of the transcriptomes of anhydrobiotic nematodes and other anhydrobiotic animals becomes available, comparative transcriptomic analyses will be a powerful tool for the identification of candidate genes and processes required for successful anhydrobiotic survival.

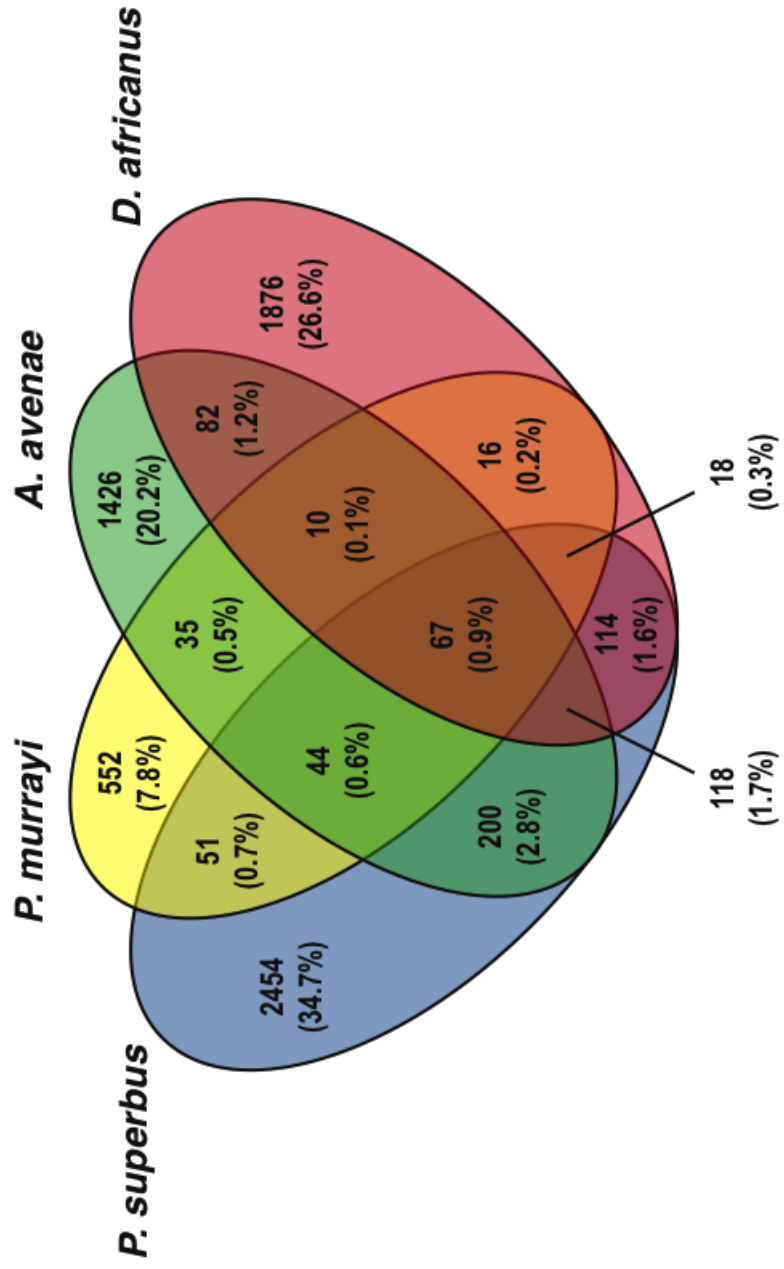


Figure 2.8: Venn diagram indicating the number of unigene families that contain representatives from one or more anhydrobiotic nematodes: *P. superbus*, *P. murrayi*, *D. africanus*, *A. avenae*.

2.3.12 Analysis of Novel Transcripts

Of the 3,982 unigenes in the dataset, 2,059 (51.7%) have no significant similarity to any sequences in the Genbank or NEMBASE4 databases. The Prot4EST algorithm was used to translate these novel unigenes into putative peptides. Analysis of the physical properties of these putative peptides reveals that 149 of them are predicted to lack a fixed tertiary structure (100% intrinsically disordered), while an additional 296 peptides are predicted to be 50-99% disordered. Intrinsically disordered proteins (IDPs) are hydrophilic, being characterised by a high proportion of polar and charged amino acids and low sequence complexity; they also have a low content of the hydrophobic amino acids which would normally form the core of a folded globular protein (Dyson & Wright, 2005). These physical features also occur in LEA proteins. A plot of the hydropathy (Kyte & Doolittle, 1982) of *P. superbus* putative novel peptides and the 13 predicted *P. superbus* LEA proteins against their predicted degree of disorder (determined using the IUPred program (Dosztányi et al., 2005)) shows that there are 225 *P. superbus* peptides with a GRAVY (Grand Average of Hydropathy) value of $\preceq$ -1 and $\succ$ 50% disordered (Figure 2.9).

Garay-Arroyo et al. (Garay-Arroyo et al., 2000) proposed that LEA proteins are contained within a larger group of proteins called 'hydrophilins' that accumulate in response to osmotic stress in prokaryotes and eukaryotes. The characteristics that define this group are a glycine content of greater than 6% and hydropathy index of less than -1. This dataset contains 170 novel putative peptides that meet these criteria (Figure 2.10). The *P. superbus* unigenes predicted to encode LEA proteins were identified on the basis of BLAST searches. Analysis of their physical properties reveals that all of these putative LEA proteins are hydrophilic, having GRAVY values ranging from -0.62 to -1.42; six are predicted to be 100% unstructured, a further three are largely (77-99%) unstructured, but three putative LEA

unigenes: PSC01853, PSC00514 and PSC00416 are predicted to be only partly disordered (9, 10 and 18% disorder, respectively). Eleven of the 13 putative LEA sequences also have a glycine content of greater than 6%.

Many IDP proteins function by molecular recognition: either by transient, or permanent binding to a structured partner molecule (Tompa, 2005). However, the functions of some IDPs depend directly on the extended random coil conformation of the disordered state - the so-called entropic chain effect (Dunker et al., 2002). Entropic chain effects are likely to be central to many of the functions of LEA proteins. The elongated, natively unfolded conformation of LEA proteins may help to form a "molecular shield" (Chakrabortee et al., 2007), preventing protein aggregation and denaturation. These hydrophilic, proteins also have the capacity to bind and retain water molecules and, at later stages of the dehydration process, an abundance of charged amino acids may enable some LEA proteins to replace water at the hydrogen bonding sites of dehydrated proteins. Although LEA proteins are natively unfolded when fully hydrated, some LEA proteins, including AavLEA1 (Goyal et al., 2003), have been shown to develop a secondary structure as they become desiccated (Hand et al., 2011), leading to the suggestion that some LEA proteins might function as intracellular space-filler molecules which prevent the collapse of cells as they become desiccated (Tunnacliffe & Wise, 2007). The combined group of putative hydrophilic proteins identified in Figure 2.10 contains 294 individual novel sequences. These sequences represent an important group of candidate anhydrobiotic genes, meriting further investigation.

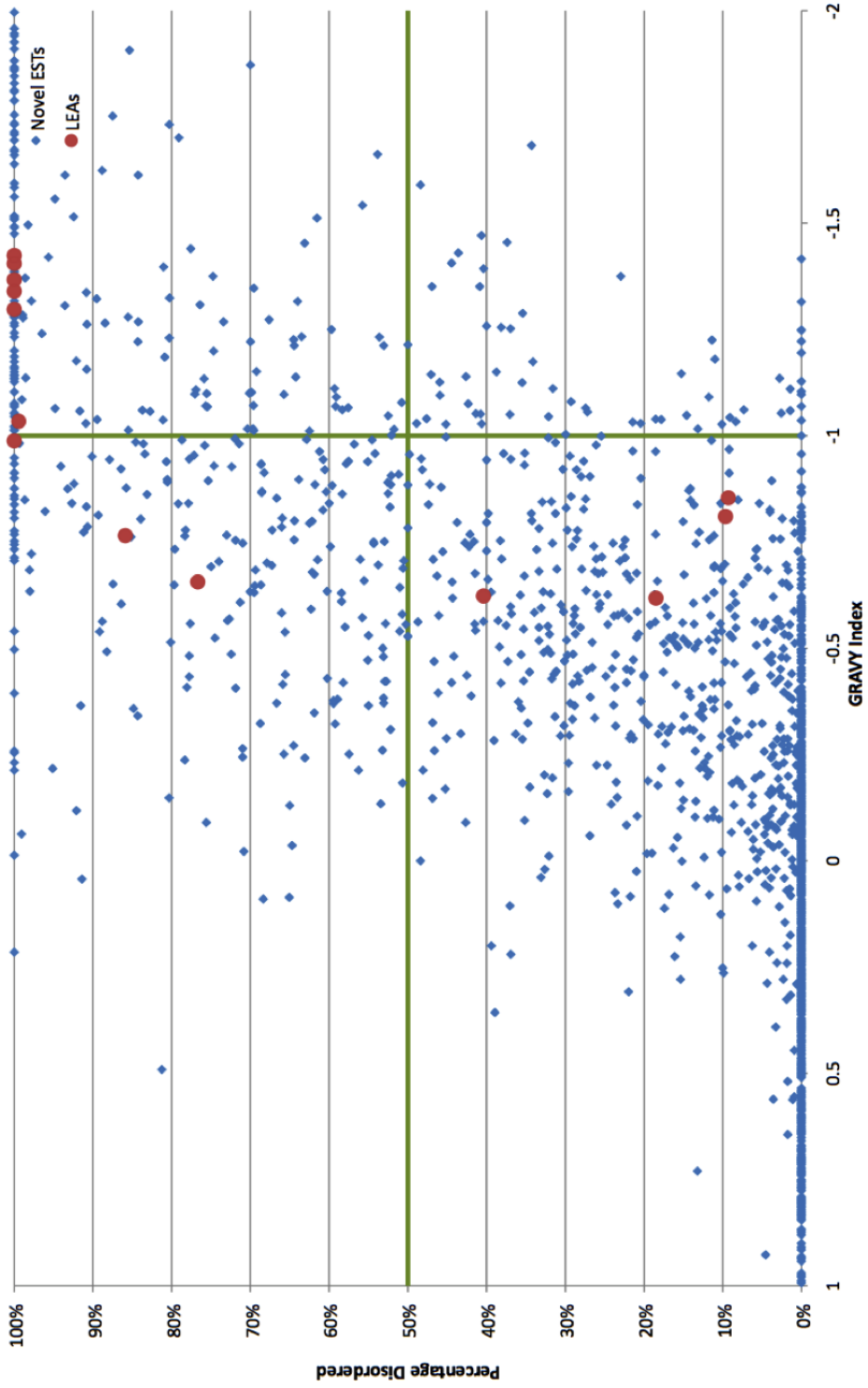


Figure 2.9: A plot of the hydropathy value (GRAVY Index) (Kyte & Doolittle, 1982) of *P. superbus* putative novel peptides and the 13 predicted *P. superbus* LEA proteins against their predicted degree of disorder, as determined by the IUPred program (Dosztányi et al., 2005). Hydrophilic proteins typically have hydropathy values ≺-1 (Garay-Arroyo et al., 2000). Green lines represent the boundaries that delimit a group of novel hydrophilic peptides that are predominantly disordered.

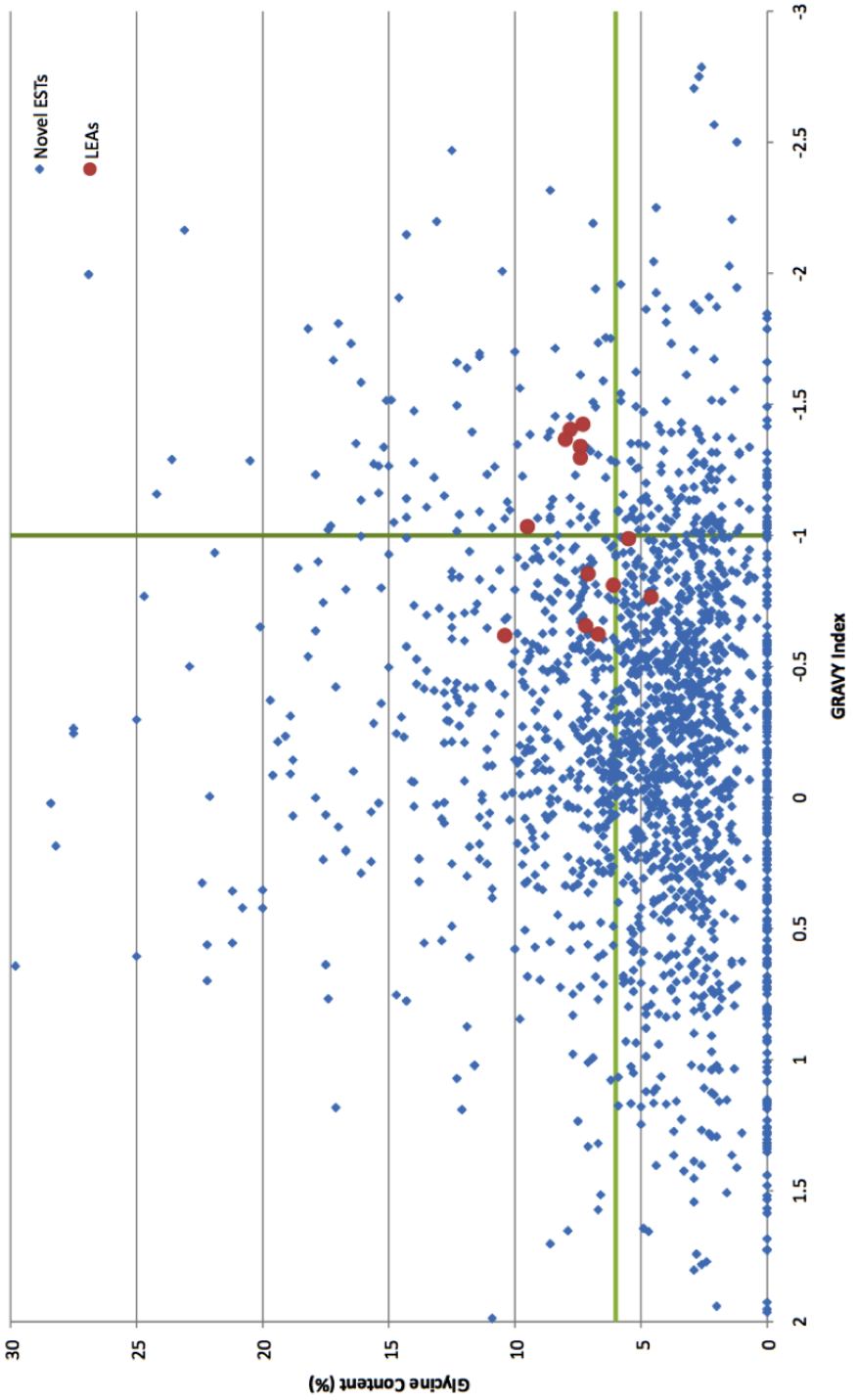


Figure 2.10: Plot of the putative glycine content and the hydropathy of the protein sequences encoded by the novel ESTs and putative LEA proteins in the *P. superbus* dataset. Green lines represent the boundaries of the properties that define hydrophilins; glycine content $\succ$ 6% and a hydropathicity index of $\prec$ -1.

2.3.13 Expression of Putative Stress Related Genes upon Desiccation

P. superbus is capable of surviving exposure to a dry atmosphere in desiccation chambers containing silica gel without the need for prior pre-incubation to mild desiccation stress (Shannon et al., 2005). However, it is likely that in its natural habitat *P. superbus* would experience more gradual change from a condition in which its cells and tissues are fully hydrated to one of extreme dehydration. In addition, intrinsic behavioural (coiling/clumping) responses or morphological adaptations (such as surface lipids (Wharton & Marshall, 2009) or possibly SXP/RAL-2 cuticular proteins) may slow the rate of water loss in *P. superbus* and allow time for inducible molecular protection mechanisms to be put in place. Dr Trevor Tyson used qPCR to investigate the inducible expression of several unigenes that represent homologues of stress related genes in other organisms. The expression of five of the 19 genes tested was upregulated in *P. superbus* following exposure to 98% RH for 12 hours.

Three antioxidant genes *gpx* (glutathione peroxidase), *dj-1* and *prx* (which encodes a 1-Cys peroxiredoxin) were upregulated in response to desiccation stress in *P. superbus*. ROS accumulation is triggered by cellular dehydration and the qPCR data show the importance of enzymatic antioxidant defense systems during the induction of anhydrobiosis. Glutathione peroxidases (GPx) catalyse the reduction of Hydrogen Peroxide, and GPx have been previously found to be upregulated in *A. avenae* and in *P. murrayi*, in response to desiccation (Reardon et al., 2010; Adhikari et al., 2010). DJ-1 is a multifunctional protein associated with familial Parkinson's disease (Bonifati et al., 2003; Shendelman et al., 2004). One of its proposed functions is a redox-dependent molecular chaperone activity (Shendelman et al., 2004) and a role for DJ-1 as an atypical peroxiredoxin-like peroxidase in inactivating mitochondrial hydrogen peroxide has also been proposed (Andres-

Mateos et al., 2007). From Figure 2.11 it can be seen that the expression of *dj-1* is upregulated 9-fold in *P. superbus* in response to desiccation stress. This gene is also upregulated in response to desiccation stress in the anhydrobiotic nematode *Aphelenchus avenae* (Reardon et al., 2010). Peroxiredoxins (Prx) comprise two classes: 1-Cys Prx and 2-Cys Prx, based on the number of cysteinyl residues directly involved in catalysis (Wood et al., 2003). Animal Prx sequences comprise 3 clades (Dubreuil et al., 2011): clades A and B contain 2-Cys Prx, while 1-Cys Prx occur in clade C, which also contains plant 1-Cys Prx sequences (Dubreuil et al., 2011). In plants 1-Cys Prx are seed-specific (Aalen, 1999): they accumulate during seed maturation and their expression declines during germination, an expression pattern is also characteristic of many LEA genes. A seed-specific 1-Cys Prx, was found to be abundantly expressed during desiccation in the leaves of the resurrection plant *Xerophyta viscosa* (Mowla et al., 2002) and transcripts encoding a 1-Cys Prx are also upregulated during rehydration of the anhydrobiotic moss *Tortula ruralis* (Oliver, 1996). A 1-Cys Prx is upregulated in response to desiccation in *P. superbus*, revealing a further parallel between the desiccation tolerance mechanisms of anhydrobiotic nematodes and plants.

Only one of the three LEA sequences tested was upregulated, but this sequence (PSC01853) was upregulated 9.8 fold in response to desiccation stress. Of the four *P. superbus* HSP sequences assayed only one, an sHSP sequence, was upregulated in response to desiccation. The genes encoding HSP70 and HSP90 are constitutively expressed in the Antarctic nematode *P. murrayi* and are not upregulated further by desiccation (Adhikari et al., 2009). Similarly only 2 of 6 HSP70 paralogues show higher expression levels in diapausing eggs of the rotifer *Brachionus plicatilis* than in other metabolically active life stages (Denekamp et al., 2011). sHSP are the major holding chaperones which prevent the irreversible aggregation of unfolding proteins (Stengel et al., 2010; Eyles & Gierasch, 2010). They have been

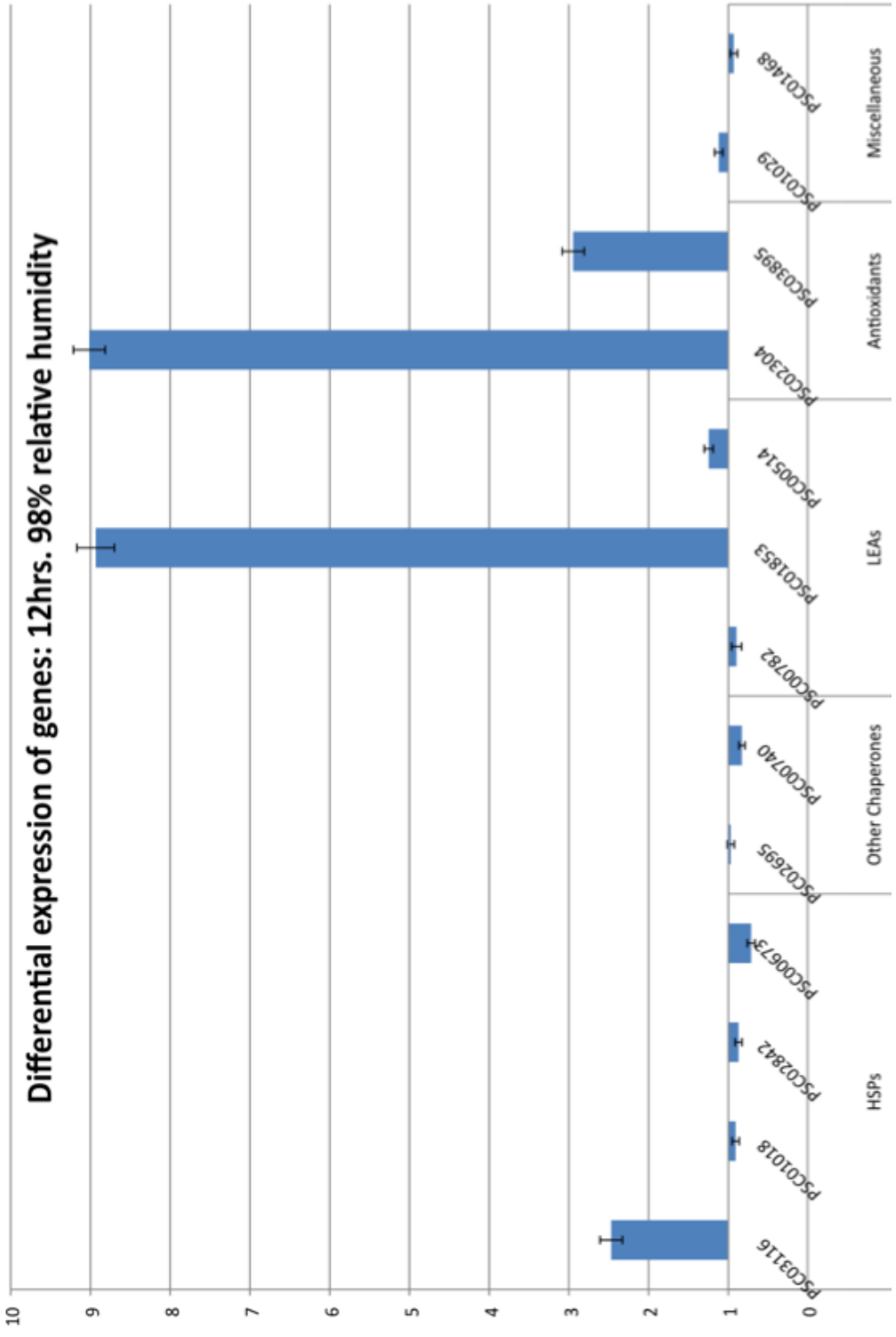


Figure 2.11: Real-Time Relative qPCR analysis of the expression of some putative stress-response genes in *P. superbus* following 12 hours of desiccation at 98% RH.

shown to accumulate in anhydrobiotic encysted larvae of the brine shrimp *A. franciscana* (Liang et al., 1997) and they are also abundantly expressed in diapausing eggs of *B. plicatilis* (Denekamp et al., 2011). These expression profiles for representatives of different HSP classes in anhydrobiotic animals from three different phyla suggest that sHSP proteins, in particular, have an important role in maintaining the integrity of the proteome during anhydrobiosis. The following transcripts were tested: PSC03116: sHSP family member; PSC01018 sHSP family member; PSC02842 HSP40/DNaJ protein family member; PSC00673 HSP70 family member; PSC02695 cyclophilin family member; PSC00740 protein disulfide isomerase; PSC00782 LEA3 protein; PSC01853 putative LEA3 protein; PSC00514 LEA3 protein; PSC02304 DJ-1; PSC03895 1-Cys peroxiredoxin; PSC02494 gluthatione peroxidase; PSC04819 gluthatione peroxidase; PSC02624 glutathione S-transferase (sigma class); PSC04040 glutathione S-transferase (kappa class); PSC01063 aldehyde dehydrogenase; PSC01095 aldehyde dehydrogenase; PSC01029 aquaporin; PSC01468; RIC1 putative stress responsive protein. The reference genes were the *P. superbus* 60S ribosomal protein L32 and *ama-1* genes. Statistically significant differences (Student's t test) are indicated, $^{**}p < 0.001$.

2.4 Discussion

P. superbus appears to utilise a strategy of combined constitutive and inducible gene expression in preparation for entry into anhydrobiosis. The apparent lineage expansion of LEA genes, together with their constitutive and inducible expression, suggests that LEA3 proteins are important components of the anhydrobiotic protection repertoire of *P. superbus*. The finding that 51% of these unigenes correspond to novel sequences is consistent with previous metagenomic analyses of nematode EST datasets, and is a reflection of the diversity of nematode gene space. Functional annotation of the *P. superbus* unigenes has identified 187 constitutively expressed consensus sequences encoding putative stress-related genes that may have a role in anhydrobiosis. Among these were: MAP-kinases; members of the jumonji family of transcription activators; antioxidant enzymes; molecular chaperones; components of the ubiquitin-proteasome system; DNA damage response proteins and LEA proteins.

Thirteen *P. superbus* unigenes encode predicted LEA proteins, all members of Group 3, as is typical of animal LEA proteins. The relative abundance of LEA genes in *P. superbus*, as compared to *C. elegans*, along with their constitutive expression (34 LEA-encoding ESTs were detected), suggest that LEA proteins are an important component of the anhydrobiotic protection repertoire of *P. superbus* and that the LEA gene family may have undergone lineage-specific expansion in this species. Five of the 19 putative *P. superbus* stress response genes tested were upregulated in response to desiccation. Three of the upregulated genes encoded antioxidant enzymes, an indication of the importance of enzymatic antioxidant defense systems during the induction of anhydrobiosis. One of the upregulated genes encoded a 1-Cys Prx, revealing a parallel between the desiccation tolerance mechanisms of plant seeds and resurrection plants with those of anhydrobiotic nematodes. Of the four *P. superbus* HSP sequences assayed, only one, an sHSP

sequence, was upregulated in response to desiccation. This is consistent with the expression profiles for representatives of different HSP classes in anhydrobiotic animals from other phyla and suggests that sHSP proteins have an important role in maintaining the integrity of the proteome during the dehydration phases of anhydrobiosis.

A large number of *P. superbus* unigenes are homologous to human disease genes, particularly those implicated in neurodegenerative diseases. Many neurodegenerative diseases are associated with the dysfunction of the protection systems responsible for repairing or degrading damaged proteins and macromolecules, thus some gene products that have roles in anhydrobiotic protection in *P. superbus* may have human homologs that are required for neural survival. Therefore, in addition to providing candidate genes for use in anhydrobiotic engineering experiments, knowledge of the molecular mechanisms responsible for anhydrobiotic protection of macromolecules may also provide insights into some of the gene products required for the integrity of neural tissues.

Analysis of the physical properties of the putative peptides encoded by the 2,059 novel *P. superbus* unigenes reveals that 149 of them are predicted to be 100% IDPs and that 170 novel sequences meet the criteria used to define 'hydrophilin' molecules which accumulate in response to osmotic stress in prokaryotes and eukaryotes. These IDPs and putative hydrophilins represent a key group of potential stress-related genes. The most highly expressed *P. superbus* sequence belongs to the nematode specific family of SXP/RAL-2 proteins, which had previously been identified as a class of secreted and surface associated antigens in diverse animal parasitic nematodes. The abundant representation of the SXP/RAL-2 in *P. superbus* may be indicative of a role for this protein in stabilising the nematodes' integument and slowing the rate of water loss during evaporative desiccation.

Panagrolaimus is an excellent model system for the study of anhydrobiosis and

cryobiosis. *Panagrolaimid* nematodes can be readily cultured in the laboratory and have a short generation time. The anhydrobiotic and cryobiotic species and strains of *Panagrolaimus* described to date belong to a single clade, which will facilitate comparative transcriptomic analyses of the molecular basis of anhydrobiosis in a single genus. This study is the first investigation of the putative molecular mechanisms involved in anhydrobiosis in *Panagrolaimus*. In addition to providing cDNA clones and sequence data for candidate anhydrobiotic genes, the dataset presented here has provided anchor sequences important for the assembly of the genome and transcriptome of *P. superbus* from high-throughput sequence data as described in Chapters 3, 4, 5 and 6.

Chapter 3

Transcriptome Assembly

3.1 Introduction

Currently, technology has caught up with, and indeed surpassed, the thirst for sequence data to be produced. Therefore, the focus must now be on meaningful interpretation of the data through functional annotation. The launch of the first next generation sequencing platform by Roche in 2005 led the way for advanced sequencing technologies, which now dominate the field, ahead of more traditional methods. The cost of sequencing has dropped significantly, meaning it has now become accessible to most research groups. Indeed, some researchers have found that in order to study a mutant of interest it is simply easier and more efficient to sequence the whole genome than to isolate and sequence the gene itself (Shen et al., 2008).

The human genome project took upwards of $10 million to complete (Lander & International, 2001) and now, with the advent of new cheaper sequencing technologies, the goal of a $1,000 genome is not too far in the future. In fact, some companies (as of July 2012) are offering human exome sequencing at a coverage of 30X for $698 for new customers with 1st sample only (www.otogenetics.com).

This could mean huge advances in the field of personalised medicine and the area of pharmacogenomics (where a person's genes could predict how their body will respond to a drug treatment) is quickly gaining mass approval, not only from those in the research domain. The true power of pharmacogenomics is in the area of prediction. For example, if sequencing their genome indicates that a person may have a predestined risk of developing a skin cancer, such as melanoma, then the application of a high SPF sunscreen should feature in their day to day routine.

Transcriptome sequencing, where only messenger RNAs are sequenced, has been used extensively in research since the first next-generation transcriptome was published shortly after the technology was launched in 2005 (Margulies et al., 2005). This technology has begun to replace microarray technology for gene expression studies. The speed and cost efficiency of next-generation technologies are continuously being updated to produce longer, more accurate and cheaper reads in less time. Once these reads are produced, they undergo quality control and are then assembled into longer contiguous sequences, so-called contigs. The assembly of the transcriptome shall be the focus of this chapter, while functional annotation of the resulting dataset produced post-assembly will be discussed in Chapter 4.

The Newbler assembler, or GS De Novo Assembler as it is referred to in the literature, is a *de novo* assembler developed by Roche for use specifically with 454 reads. Both FLX and longer Titanium reads can be assembled alongside traditional Sanger reads. In the most simplistic view, the assembly process works by taking all the reads and aligning them to one another. Overlapping reads can then be merged, and there may be multiple iterations of this process before contigs are produced. Ideally, these contigs represent the complete repertoire of full-length transcripts from expressed genes.

There are problems associated with assembly. For example, with *P. superbus*,

as it is grown on a lawn of *Esherichia coli*, there is the potential that the contigs generated from the sequencing of *P. superbus* could actually originate from *E. coli* or other bacteria in the food source. This can be remedied by removing any sequences with high similarity to bacterial genes. However, it is possible that some of these bacterial sequences may be *bona fide* components of the *P. superbus* genome, such as remnants from horizontal gene transfer events. Primers and adapters need to be removed before any analysis is done, as otherwise these sequences would cause a large number of reads from different transcripts to overlap. As with any sequencing technology, there may be a certain degree of sequencing error. This is remedied by using quality scores and filtering at a pre-defined quality cutoff point. Homopolymer errors may occur when using the 454 pyrosequencing technology (Balzer et al., 2010). These errors have been detected when the target sequence contains runs of five (or more) of the same base. Challenges associated with sequencing also include repeat regions and AT-rich regions. The possibility of chimera formation needs to be considered and abundantly expressed ribosomal RNA reads need to be identified.

The aim of this transcriptome analysis was to assemble the best representation of all expressed transcripts with the expectation that, by stressing the worms prior to sequencing, the resulting transcriptome would be enriched for stress response genes. To enrich for stress-related genes, a mixed population of *P. superbus* was exposed to one of the following stresses: desiccation, cold, heat or oxidation. Both normalised and unnormalised cDNA libraries were prepared from this population. The reason for including a normalised cDNA library in the sequencing step was to reduce the over-representation of abundantly expressed genes (normally associated with housekeeping functions) in the sequencing library and hence low abundance sequences can be identified. The aim of the work presented in this chapter was to identify the best assembly pipeline for 454 Roche Titanium reads with a view

to performing downstream functional annotation. This resulted in establishing a set of metrics by which to evaluate the performance and merits of each assembler. It should be noted at this point, as was discussed in Chapter 1, that establishing the 'best' assembly is not simply a matter of choosing which performs best given a series of test. It is necessary also to examine what is being questioned of the sequences resulting from the assembly.

3.2 Materials & Methods

3.2.1 Growth Conditions, Stress Treatments and RNA Extraction

P. superbus was grown at 20°C on 9cm NGM agar plates (Brenner, 1974) supplemented with streptomycin (30μg/ml) and containing a lawn of *Escherichia coli* strain HB101. Twenty five NGM plates, containing a mixed population of adults and larvae, were flooded with sterile S Basal buffer (Brenner, 1974). The plates were left shaking for 30 minutes, then the supernatant was transferred to a sterile 1L beaker and the nematodes were allowed to settle for 30 minutes at 4°C. The supernatant containing *E. coli* was then removed and the nematodes were re-suspended in sterile S buffer. This washing process was repeated three times. The nematodes were then transferred to 50ml Falcon tubes and centrifuged at 1,000rpm for 5 minutes. The supernatant was removed and the pellet re-suspended in sterile S buffer. This process was also repeated three times.

The nematodes were pooled and their final concentration was adjusted to 3,000 nematodes/ml. Nine replicates were set up for each of the four stress conditions (as can be seen in Table 3.1), each replicate containing 3,000 nematodes in 1ml of S buffer. For the controls, 36 replicates were set up (i.e., the number of control nematodes equalled that used in all the stress treatments combined).

With the exception of the desiccation treatments, all nematodes were incubated with shaking at 50rpm on a Braun Centromat-R shaking platform for 24 hours. For the desiccation treatment, individual 1ml suspensions of nematodes were vacuum filtered onto a 2.5cm SuporR-450 filter (45μM, Pall Life Sciences). Each filter was then transferred to a 3cm Petri dish and these were placed, uncovered, in a 1L desiccation chamber containing a 200ml saturated solution of potassium dichromate (K_2Cr_2O7) (Winston & Bates, 1960). Following exposure to 98% RH for 24 hours

Table 3.1: Stress states and their corresponding environmental values

Stress	**Environmental setup**	**Time in stress state**	**No. of nematodes**
Heat	32°C	24hrs	25,000
Cold	4°C	24hrs	25,000
Oxidation	38μM paraquat	24hrs	25,000
Desiccation	98% RH at 20°C	60hrs	25,000
Control	20°C	24hrs	108,000

the desiccated worms were washed off the filters using 1ml of sterile water. The nematodes corresponding to each individual stress were pooled in a 50ml Falcon tube, they were allowed to settle by gravity and were then transferred to a 1.5ml Eppendorf tube, centrifuged at 5,000rpm for 5 minutes and re-suspended in 1ml TRIzol Reagent (Invitrogen). The tubes were snap frozen in liquid nitrogen and then stored at -80°C until required. In the case of the paraquat treatment, an additional washing step was included to remove the paraquat prior to the TRIzol step.

The worms were homogenised in a mortar and pestle using liquid nitrogen. Worms from the four stress treatments were pooled prior to homogenisation and the control worms were homogenised separately. RNA was isolated from the homogenates following the protocol provided by the TRIzol supplier. The RNA concentration was measured using a Qubit fluorometer (Invitrogen) and RNA quality was monitored by Tris-acetate-EDTA (TAE) agarose gel electrophoresis. TAE is a buffer solution containing a mixture of Tris base, acetic acid and EDTA (Ethylenediaminetetraacetic acid).

3.2.2 cDNA Synthesis and Normalisation

cDNA was prepared using a MINT-Universal cDNA synthesise kit SK002 (Evrogen), designed to syntheses full-length enriched ds cDNA. First strand cDNA was generated following the manufacturer's protocol using 1μg of RNA for each reaction (the completed reaction volume was 15μl). The first strand product was used as a template for ds cDNA synthesis by PCR amplification for 20 cycles using the M1 PCR primer in a 10μl reaction volume. The cDNA template comprised 0.5μl first strand cDNA product from stressed nematodes and 0.5μl first strand product from the control nematodes (thus the resulting ds cDNA library contains transcripts from both the stressed and control nematodes). The optimum number

of PCR cycles was determined by evaluative PCR following the manufacturer's protocol. The ds cDNA from eight PCR amplifications was pooled and purified using a Qiagen PCR purification kit, yielding 17.4μg of cDNA in total. One half of this cDNA was retained for sequencing as an unnormalised cDNA library (library PS1) and the remainder was used to prepare a normalised library using a TRIMMER-DIRECT cDNA normalisation kit NK002 (Evrogen). In the TRIMMER normalisation protocol, ds cDNA is denatured and allowed to re-anneal; the ds-fraction formed by abundant transcripts during re-annealing is degraded using Kamchatka crab duplex-specific nuclease (DSN) enzyme as described by Zhulidov *et al.* (Zhulidov et al., 2004) and the normalised cDNA fraction is PCR amplified. The optimal concentration of DSN (0.21 Units) and optimum number of PCR cycles (19) were determined experimentally following the Evrogen protocol using 300ng of non-normalised cDNA. This process yielded 5.9μg of normalised cDNA (library PS2) for sequencing. cDNA concentrations were measured using a Qubit fluorometer (Invitrogen) and cDNA quality was monitored by TAE agarose gel electrophoresis.

3.2.3 454 Titanium Pyrosequencing

The samples were nebulised, adapter ligated and pyrosequenced using the GS-FLX Titanium single end reads platform (Roche) at the GenePool DNA Sequencing Centre, University of Edinburgh. MINT SMART adapters, low complexity and low quality sequences were trimmed and filtered, also removing the poly-A tails.

3.2.4 Sequence Assembly

Following filtering, these sequences were used as input into a variety of commercial and freely available assembly programs using the Stokes and Stoney computing clusters at the Irish Centre for High-End Computing (ICHEC, 2013). Independent

assembler software packages used were Newbler (Roche, 2012) (www.454.com), MIRA (Chevreux et al., 2004) (*www.chevreux.org/projects_mira.html*), CLCBio (Knudsen et al., 2010) (www.clcbio.org), Celera (Myers, 2000) (www.jcvi.org) and iAssembler (Zheng et al., 2011) (www.bioinfo.bti.cornell.edu/tool/iAssembler/). Their command line execution parameters are shown in Table 3.2. Several different versions of Newbler were examined as newer ones were released during the investigation. The sample referred to as PS1 is the unnormalised pooled sample of control and stress worms whereas PS2 represents the normalised sample of pooled control and stress worms. PS1 read count was 444,453 and PS2 read count was 414,248 for assembly. Read statistics can be seen in Table 3.3. As discussed in Chapter 2, 7,606 *P. superbus* ESTs that were sequenced using conventional Sanger sequencing were also annotated. These sequences were also included in the assembly steps to maximise the amount of data for assembly.

Table 3.2: Assemblers used and parameters executed. -vt is used to trim primers, adapters or polyA tails from the start or end of reads, -cdna identifies the sequences as transcriptome sequences and -URT (use read tips) which means that contigs can be generated even if found in a region of low coverage.

Assembler	**Parameters**
Newbler 2.3 without URT Contigs	-cdna -vt
Newbler 2.3 without URT Isotigs	-cdna -vt
Newbler 2.5 pre release with URT Contigs	-cdna -vt
Newbler 2.5 pre release with URT Isotigs	-cdna -vt
Newbler 2.5 without URT Contigs	-cdna -vt
Newbler 2.5 without URT Isotigs	-cdna -vt
Newbler 2.5 with URT Contigs	-cdna -vt -URT
Newbler 2.5 with URT Isotigs	-cdna -vt -URT
Newbler 2.6 without URT Contigs	-cdna -vt
Newbler 2.6 without URT Isotigs	-cdna -vt
Newbler 2.6 with URT Contigs	-cdna -vt -URT
Newbler 2.6 with URT Isotigs	-cdna -vt -URT
CLCBio Assembly Cell version 3.10.47055	default
Celera version wgs-6.1-Linux-amd64	default
Mira version 3.2.0rc1	default
Mira version 3.4.0.1	default
CAP3 version 2010	-o 50 -p 98
CAP3 version 2012	-o 50 -p 98
Phrap	-o 50 -p 98
iAssembler version v1.3.2	-y 30 -p 97 -o 40 -s 251 -f 6

Table 3.3: Post-filtering read statistics for the two *P. superbus* cDNA libraries obtained by 454 Titanium FLX pyrosequencing.

	PS1	**PS2**
Number of reads	444,453	414,248
Average length	320.52	284.14
Standard deviation	130.13	133.66
GC%	36	33
Number of bps (Mb)	142.5	117.7
Longest read length(bps)	668	693

Also investigated was the use of hybrid assemblers using output from independent assemblers (see Figure 3.1). The CAP3 software package [1] was used on the output from MIRA and Newbler to produce the CAP3 assembly. Given this dataset, further processing was carried out according to the Partigene pipeline (Parkinson et al., 2004), which involved the use of two extra steps: CLOBB [2] and Phrap [3], producing what will now be referred to as the Phrap assembly. The command line execution parameters used for these hybrid assemblers are shown in Table 3.2.

[1] www.seq.cs.iastate.edu
[2] www.nematodes.org/bioinformatics/Clobb2
[3] www.phrap.org

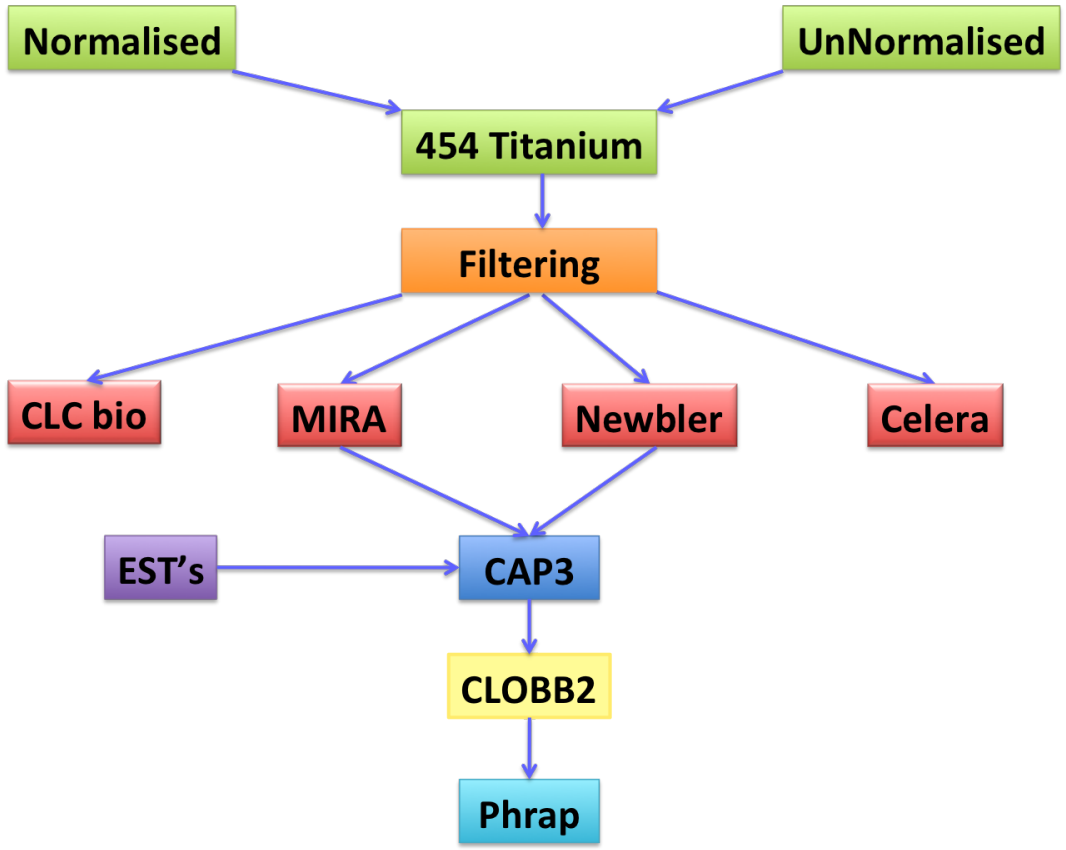


Figure 3.1: Steps used in the generation of hybrid pipeline showing CAP3 and Phrap assemblies.

3.2.5 BLAST Homology Searches

The databases used for the homology searches are shown in Table 3.4.

Table 3.4: Databases used for homology searching.

Database	**URL**
BLAST (NR)	www.blast.ncbi.nlm.nih.gov
CEG version 2.4	www.korflab.ucdavis.edu/datasets/cegma
Nembase 4	www.nematodes.org/nembase4
UniRef100	www.ebi.ac.uk/uniref/

3.3 Results

3.3.1 General Assembly Statistics

General information on the sequences from each assembly was generated using a custom script. Initially, contigs which were less than 150bp were filtered out from further analyses. The percentage of assembly removed at this filtering stage varied drastically depending on the assembler. The assemblers which had most sequences removed were iAssembler and Celera along with the Newbler contigs datasets. The other assemblies lost relatively few sequences at this filtering stage, as can be seen in Tables 3.5 - 3.8.

Newbler assembles reads into contigs, isotigs and isogroups. Respectively, these correspond to exons, splice variants and gene sequences. However, there are potential anomalies in this classification. For example, while the number of isogroups should correspond to the number of expressed genes identified, some large isogroups may cluster together transcripts from different genes due to a highly conserved domain. Another potential issue is that different parts of a transcript may be classified into two un-related isotigs due to low sequencing coverage. It is also important to note that some level of untranslated regions (UTRs) and introns (in case of primary transcripts) may be represented among the reads.

Various metrics were used to select the most appropriate assembly for downstream annotation. These metrics will be discussed in the following sections. The assemblies discussed will be divided into Newbler assemblers and other assemblers. The objective of all assembly programs is to reassemble sequencing reads into longer contigs which accurately reflect the underlying transcriptome from which the reads were derived. The data in Tables 3.5 - 3.8 present the key metrics obtained for all the assembly programs and hybrid assemblies used in this project with the 454 Titanium FLX cDNA sequences. These metrics describe the num-

ber of sequences (contigs and singletons) generated by each assembly program, the average sequence length, the number of contigs ≥1,000bp and the size of the assembled transcriptome in Mb.

The mean contig length ranges from 450bp (Celera) and 481bp (iAssembler) (Table 3.5) to 1,032bp (Newbler version 2.6 without URT isotigs) (Table 3.8). Figure 3.2 shows that different assembly programs and strategies generate a wide diversity of mean contig lengths. The optimum assembler will ideally assemble all the reads corresponding to a transcript into a single contig, thus the mean contig length is an important criterion. However, it is important that maximisation of contig length does not lead to the introduction of assembly errors, such as the generation of chimeric sequences.

As expected, the mean contig length tends to be inversely proportional to the number of contigs generated by the assembly. The highest number of sequences in any assembly was found in the iAssembler dataset, 105,156 sequences, while the CAP3 v2012, Phrap and the Newbler assemblies had fewer sequences (range 17,934 - 27,470 for Newbler - see Table 3.8), with 31,836 sequences for Phrap and 31,836 for CAP3 v2012.

The size of the assembled transcriptome provides an indication of redundancy. Although animal genome sizes vary greatly in magnitude, transcriptome sizes fall into a relatively narrow size range. Thus, for *P. superbus* a transcriptome which is greater than ∼25Mb (size of the *C. elegans* transcriptome) is likely to contain redundant contigs. To a certain extent some redundancy is to be expected as splice variants need to be taken into account and some assemblers, most notably Newbler Isotigs, will include variations on the same sequences. This can be seen in Figure 3.3 which shows the total size of the assemblies in relation to the *C. elegans* transcriptome (He et al., 2007). It shows that the large transcriptome sizes generated by the iAssembler, CAP3 v2010 and Mira assemblies are likely to

Table 3.5: Assembly statistics (a).

	CLCBio	**Celera**	**Mira v3.2.0rc1**	**Mira v3.4.0.1**	**CAP3 v2010**	**CAP3 v2012**	**Phrap**	**iAssembler**
Seq. no. post filtering	27,976	38,981	41,422	42,161	46,540	31,836	26,360	105,156
Seq. no. pre filtering	27,976	54,973	42,568	43,275	48,413	33,282	26,905	161,177
% removed through filtering	0	29	3	3	4	4	2	35
% GC	32	34	33	33	32	33	33	32
Average sequence length	694	450	725	696	805	830	898	481
Average sequence length s.d.	444	393	428	407	531	523	590	324
Total size (Mb)	19	18	30	29	37	26	24	51
Large contigs (1000bp+)	5,469	4,399	8,418	7,617	12,582	9,350	8,842	7,881
Maximum length (bp)	4,434	4,386	6,764	5,482	6,842	6,203	6,842	6,303

Table 3.6: Assembly statistics (b) Newbler assemblies. w indicates with URT wo indicates without URT.

	Newbler 2.3 woURT Contigs	**Newbler 2.3 woURT Isotigs**	**Newbler pre2.5 wURT Contigs**	**Newbler pre2.5 wURT Isotigs**
Seq. no. post filtering	20,074	18,183	17,922	18,690
Seq. no. pre filtering	25,505	18,196	23,961	18,706
% removed through filtering	21	0	25	0
% GC	33	33	33	33
Average sequence length	894	1,041	758	1,038
Average sequence length s.d.	505	529	475	521
Total size (Mb)	18	19	14	19
Large contigs (1000bp+)	7,043	8,283	4,663	8,557
Maximum length (bp)	5,469	5,472	4,368	5,471

Table 3.7: Assembly statistics (c) Newbler assemblies continued. w indicates with URT wo indicates without URT.

	Newbler 2.5 woURT Contigs	**Newbler 2.5 woURT Isotigs**	**Newbler 2.5 wURT Contigs**	**Newbler 2.5 wURT Isotigs**
Seq. no. post filtering	17,934	18,617	26,797	27,470
Seq. no. pre filtering	24,176	18,630	34,934	28,616
% removed through filtering	26	0	23	4
% GC	33	33	32	32
Average sequence length	758	1,032	685	934
Average sequence length s.d.	475	514	484	591
Total size (Mb)	14	19	18	26
Large contigs (1000bp+)	4,661	8,409	5,673	10,506
Maximum length (bp)	4,368	5,470	4,389	6,888

Table 3.8: Assembly statistics (d). Latest Newbler release (Feb 2013). w indicates with URT wo indicates without URT.

	Newbler 2.6 woURT Contigs	**Newbler 2.6 woURT Isotigs**	**Newbler 2.6 wURT Contigs**	**Newbler 2.6 wURT Isotigs**
Seq. no. post filtering	15,900	14,960	26,595	25,402
Seq. no. pre filtering	18,630	14,980	31,491	26,824
% removed through filtering	15	0	16	5
% GC	32	33	32	32
Average sequence length	813	1,031	698	871
Average sequence length s.d.	492	507	489	562
Total size (Mb)	13	15	19	22
Large contigs (1000bp+)	4,857	6,866	5,852	8,612
Maximum length (bp)	4,367	5,475	4,389	5,545

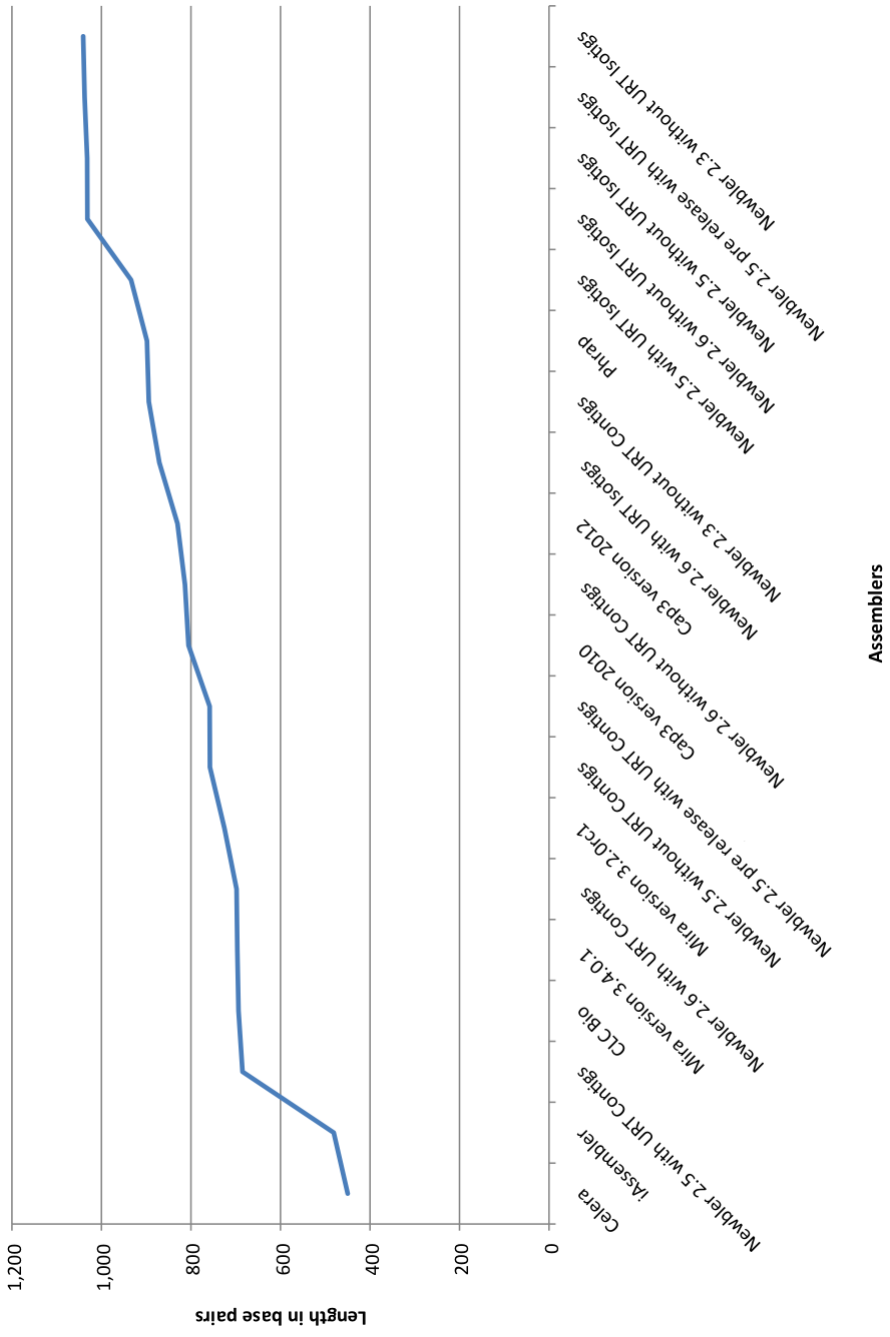


Figure 3.2: Average length of the *P. superbus* transcriptome generated by the different assembly programs evaluated during this project.

contain redundant sequences, whereas some of the early versions of the Newbler software and the Celera program are liable to generate transcriptome assemblies that could under-represent the full repertoire of the *P. superbus* transcriptome.

A common metric used to establish the quality of an assembly is the N50 statistic. It is the length of the smallest contig in the set that contains the fewest (i.e., largest) contigs whose combined length represents just over half the size of the assembly. Genes in *C. elegans* were found to be between 1 and 1.5Kb long. Therefore, an assembly with an N50 in this range, would indicate that full length genes have been sequenced. The CAP3 assemblies, Phrap, and all of the Newbler Isotigs assemblies fall in this range, as seen in Figure 3.4, and the corresponding information in Table 3.9. While it is often tempting to use the N50 statistic as a determinant of the quality of an assembly, i.e., higher N50 values indicate better representation by long, potentially full-length, sequences in the assembly, it is also important to examine other criteria to ensure that the contigs are not over-assembled, i.e., chimeric sequences.

Figures 3.2 - 3.4 present some of the key metrics for the *P. superbus* transcriptome obtained with the different assembler programs.

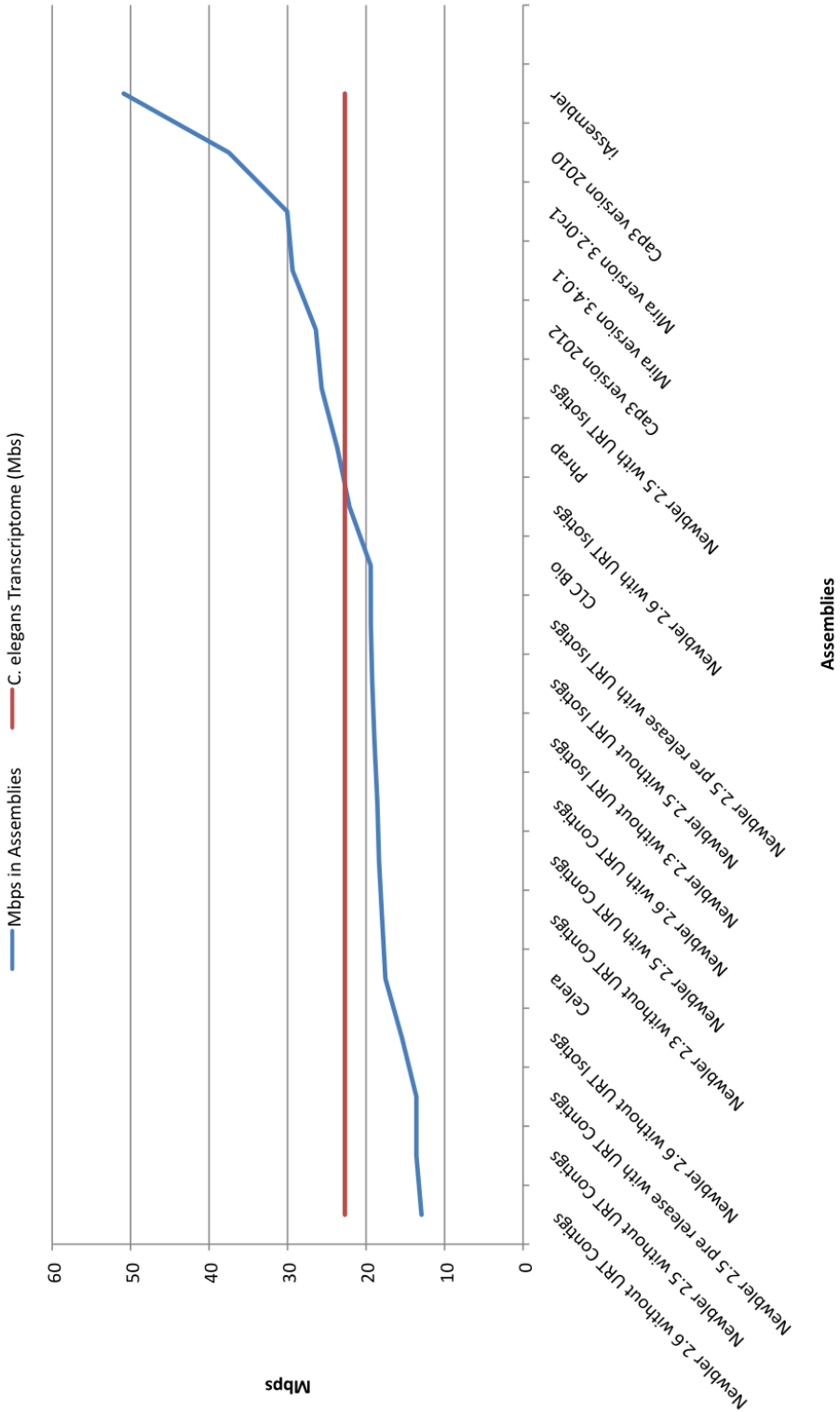


Figure 3.3: The total size (Mb) of the *P. superbus* transcriptome generated by the different assembly programs evaluated in this project. The horizontal red line corresponds to the size of the transcriptome of *C. elegans*.

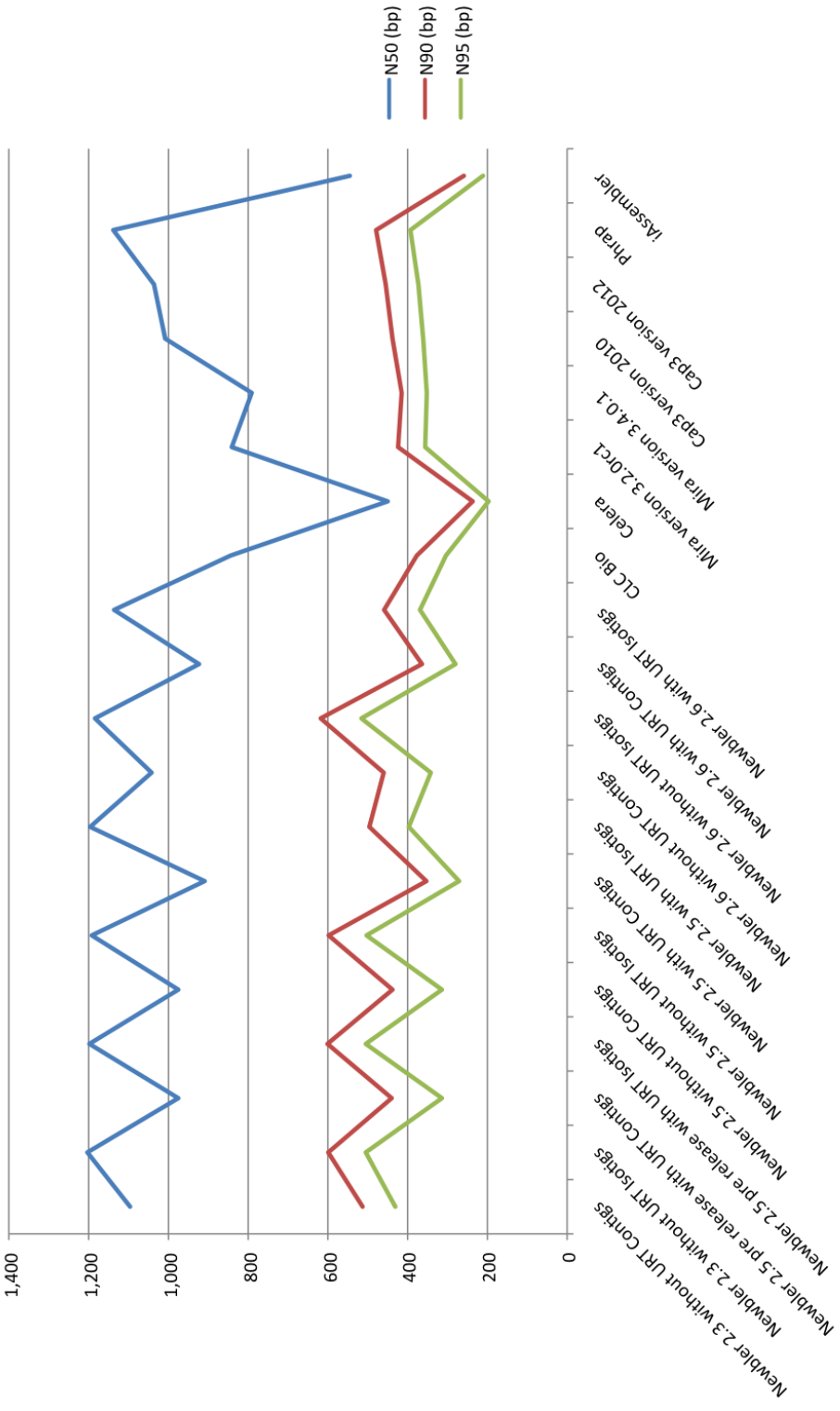


Figure 3.4: Transcriptome Assemblers compared on the basis of the N50, N90 and N95 lengths (bp) of the assembled *P. superbus* sequences.

Table 3.9: Transcriptome assemblies compared on the basis of N50 values.

Assembler	**N50 (bps)**
Newbler 2.3 without URT Contigs	1,096
Newbler 2.3 without URT Isotigs	1,203
Newbler 2.5 pre release with URT Contigs	975
Newbler 2.5 pre release with URT Isotigs	1,198
Newbler 2.5 without URT Contigs	975
Newbler 2.5 without URT Isotigs	1,192
Newbler 2.5 with URT Contigs	909
Newbler 2.5 with URT Isotigs	1,196
Newbler 2.6 without URT Contigs	1,042
Newbler 2.6 without URT Isotigs	1,184
Newbler 2.6 with URT Contigs	923
Newbler 2.6 with URT Isotigs	1,136
CLCBio	844
Celera	450
Mira version 3.2.0rc1	841
Mira version 3.4.0.1	791
Cap3 version 2010	1,008
Cap3 version 2012	1,036
Phrap	1,138
iAssembler	545

3.3.2 Similarity Searches to Establish Quality of the *P. superbus* Datasets

Other important criteria which can be used to evaluate the quality of a genome, or transcriptome assembly, from a eukaryotic organism are whether the assembly in question contains transcripts similar to gene/transcript sequences from other organisms, i.e., the extent to which one is able to recover homologs from each assembly. BLAST analyses were carried out, comparing the *P. superbus* transcriptomes generated by the different assembly programs to different databases containing sequences that represent (a) highly conserved eukaryotic genes, and (b) genes from closely related nematode species.

Assembly Quality Control using the Core Eukaryotic Genes (CEGs) Dataset

The CEG genes are a set of 248 highly conserved genes found in a wide range of organisms (Parra et al., 2007, 2009). This gene set has previously been used as a measure of the quality of *de novo* genome assembly data (Parra et al., 2009). It is derived from a set of low copy number conserved genes from six model organisms: *Saccharomyces cerevisiae*; *Schizosaccharomyces pombe*; *Arabidopsis thaliana*; *C. elegans*; *Drosophila melanogaster* and *Homo sapiens*. In selecting this gene set, the authors chose conserved genes which were present in the KOG (euKaryotic Orthologous Groups) database (Tatusov et al., 2003) in single copy for at least four of the six model organisms (Parra et al., 2009).

Due to their conservation among the eukaryotes, the CEG genes should be well represented in the *P. superbus* transcriptome. Hence the ability to retrieve CEG homologs can be used as a basic 'litmus' test for the output of an assembly. Each assembly was searched using BLAST for the presence of CEG homologs. The

results are shown in Table 3.10, where it can be seen that, apart from CLCBio, 97-98% of all CEG genes have homologs in each assembly, i.e., all assemblies passed this first quality control test, apart from CLCBio.

Table 3.10: BLAST searches carried out using the CEG genes from six model organisms against the different *P. superbus* transcriptome assemblies. A BLAST cut-off of > 60 bits was used. *: TBLASTN search using CEG as the query. **: BLASTX search using CEG genes as the database.

Transcriptome Assembly	**CEG(query) ***	%	**CEG(db) ****	%
CLCBio	898	60	366	1
Newbler 2.5 with URT Isotigs	1,444	97	834	3
Phrap	1,445	97	843	5
Newbler 2.5 pre release with URT Contigs	1,447	97	645	4
Newbler 2.6 with URT Isotigs	1,449	97	748	3
MIRA version 3.2.0rc1	1,450	97	665	4
Cap3 version 2010	1,450	97	659	4
Newbler 2.5 pre release with URT Isotigs	1,451	98	696	5
Celera	1,451	98	860	5
MIRA version 3.4.0.1	1,451	98	835	4
CAP3 version 2012	1,451	98	817	4
Newbler 2.6 without URT Isotigs	1,453	98	1,363	3
Newbler 2.6 with URT Contigs	1,455	98	1,404	3
iAssembler	1,455	98	1,882	2
Newbler 2.5 without URT Isotigs	1,456	98	720	3
Newbler 2.6 without URT Contigs	1,457	98	1,127	4
Newbler 2.5 without URT Contigs	1,459	98	788	3
Newbler 2.3 without URT Isotigs	1,460	98	714	3
Newbler 2.3 without URT Contigs	1,461	98	950	3
Newbler 2.5 with URT Contigs	1,463	98	1,366	3

3.3.3 Homologues to other Nematode Genes and Transcripts

NemBase4 is a set of Sanger sequenced nematode EST sequences curated by the GenePool at the University of Edinburgh (Elsworth et al., 2011). As of July 2012, it contained 62 species of nematodes including the EST sequences from *P. superbus* discussed in Chapter 2. NemBase4+ dataset was created by removing the *P. superbus* sequences (to avoid hits to these sequences) and adding EST cDNA sequences for the following species: *Plectus murrayi*, *Ditylenchus africanus*, *Aphelenchus avenae*, *Trihinella spiralis*, *Wuchereria bancrofti*, *Loa loa* and *Pristionchus pacificus*. This resulted in a total of 213,570 sequences which were compared against the assemblies using TBLASTN. The results are shown in Table3.11, where it can be seen that 39-41% of Nembase4+ sequences are consistently represented by the assemblies. The exception to this is the Celera assembly where only 25% of Nembase4+ sequences are represented, indicating potential issues with this assembly.

Complementary BLASTX analyses were also carried out (using Nembase4+ as the database and different assemblies as queries) to examine what proportion of each assembly was similar to Nembase4+ sequences (results shown in Table 3.11). Assuming that the ideal transcriptome assembly would have a fixed proportion of transcripts that are similar to the Nembase4+ database, significantly higher/lower proportions may indicate over- and under- assembly issues respectively. Congruent with the previous finding of under-representation by Nembase4+ sequences, the Celera assembly has a low proportion of its transcripts (35%) with significant similarity to the Nembase4+ dataset. Other assemblies have 41-59% of their respective transcripts with similarity to 39-41% of Nembase4+ sequences. The exception to this is the iAssembler assembly where it has the lowest proportion of its transcripts having similarity to Nembase4+ sequences. This is despite having the highest number of transcripts overall (28,060), 86,773 or 41% of these have a hit

to Nembase4+ sequences (consistent with most other assemblers). One obvious explanation lies in the fact that iAssembler produced the highest total number of transcripts. However, these sequences are shortest when compared to others, i.e., it may be subjected to under-assembly issues where transcripts are overly fragmented.

Table 3.11: BLAST searches carried out using NemBase4+ sequences against the different *P. superbus* transcriptome assemblies. A BLAST cut-off > 60 bits was used. *: TBLASTN algorithm. **: BLASTX algorithm.

Transcriptome Assembly	**NemBase4+ (query) ***	%	**NemBase4+ (db) ****	%)
Celera	86,155	25	13,524	35
Newbler 2.6 without URT Isotigs	87,501	38	8,860	59
Newbler 2.5 without URT Contigs	83,099	39	8,881	50
Newbler 2.3 without URT Isotigs	82,903	39	10,716	59
Newbler 2.5 pre release with URT Contigs	82,540	39	8,870	49
Newbler 2.5 without URT Isotigs	82,536	39	10,981	59
Newbler 2.5 pre release with URT Isotigs	82,566	39	10,954	59
CAP3 version 2012	88,467	40	15,303	48
CLCBio	87,501	40	11,357	41
Mira version 3.4.0.1	86,475	40	19,335	46
Mira version 3.2.0rc1	53,875	40	19,128	46
Phrap	84,447	41	12,398	47
Newbler 2.6 with URT Contigs	81,837	41	10,987	41
Newbler 2.5 with URT Contigs	83,093	41	11,019	41
Newbler 2.6 with URT Isotigs	87,044	41	11,860	47
Newbler 2.6 without URT Contigs	87,715	41	8,242	52
Newbler 2.5 with URT Isotigs	87,046	41	13,494	49
iAssembler	86,773	41	28,060	27
CAP3 version 2010	86,241	41	20,280	44

While the Nembase4+ database is a rich resource for gene sequences representing the diversity of nematode gene space, many of the 62 nematode species represented in the database have less than 100 ESTs. The sequences in Nembase4 are predominantly from parasitic nematodes and the database does not contain EST data from either *C. elegans* or *C. briggsae*. Thus, analogous BLAST searches were carried out using transcriptome data from nematodes whose genomes are fully sequenced and annotated. The nematode species selected for this experiment were the free living nematodes *C. elegans*, *C. briggsae* and *P. pacificus* and the parasitic nematodes *B. malayi* and *M. hapla*. Although it is a parasite with a small transcriptome (Ghedin et al., 2007), the *B. malayi* dataset was included because of its phylogenetic relatedness to *P. superbus* (see Figure 1.1). Again, the different assemblies were used as queries and databases compared against each of the nematode transcriptomes. The results are shown in Tables 3.12 to 3.15. They show that when the transcriptome datasets from the five nematodes are used as BLAST queries against the *P. superbus* assemblies, *C. elegans* is most significant in similarity to *P. superbus* (45-51%). Here the CLCBio assembly is again an outlier with only 39% of the *P. superbus* transcriptome represented. The trend can also be seen where the iAssembler has the largest number of contigs but the lowest proportion of its transcripts with significant similarity to other nematodes (Tables 3.14 and 3.15), indicative of under-assembly issues.

3.3.4 Quality Evaluation of *P. superbus* Assemblies by Assessing 5' to 3' Coverage of Individual CEG Genes

The analyses in Sections 3.3.4 to 3.3.7 were carried out in collaboration with Dr. Chris Creevey, Teagasc, Grange, Dunsany, Co. Meath. The data in Tables 3.5 to 3.8 show that different assembler programs generate transcriptome assemblies which differ substantially in numbers of contigs, mean contig length and transcriptome size. Very high numbers of contigs and low mean contig size in an assembly are possible indicators of under-assembled contigs, where most likely the assembler has failed to merge overlapping reads from the same transcript. Under-assembly of reads results in contigs which may not contain the full sequences of individual genes and leads to inflated estimates of transcriptome size. From Table 3.5 it can be seen that the *P. superbus* transcriptome assemblies generated by iAssembler are likely to have generated under-assembled contigs.

The data in Section 3.3.2 shows that, with the exception of the CLCBio assembly, all of the other assemblies had greater than 60 bit hits to 97-98% of the CEG (Core Eukaryotic Genes) genes. Thus these conserved core genes are very well represented in the *P. superbus* transcriptome. This is an indication that the high-throughput 454 Titanium FLX sequencing does provide a good representation of genes expressed in the *P. superbus* transcriptome. Although the different assemblies generally provide a good representation of the gene space in the *P. superbus* transcriptome, another very important quality criterion is whether full length transcripts are represented. Where a closely related reference species is available, sequence alignment strategies can be used to identify the assembly whose genes have the most complete coverage of individual genes in the reference transcriptome. Lacking access to a closely related reference transcriptome for *P. superbus*, the strategy used here was to compare the alignment of the *P. super-*

Table 3.12: TBLASTN using complete transcriptome datasets from *C. elegans*, *P. pacificus* and *C. briggsae* as queries against the different *P. superbus* transcriptome assemblies as databases. A BLAST cutoff of > 60 bits was used.

Transcriptome Assembly	***C. elegans*(query)**	**%**	***P. pacificus*(query)**	**%**	***C. briggsae*(query)**	**%**
MIRA version 3.4.0.1	11,769	47	8,442	35	8,824	40
Newbler 2.3 without URT Isotigs	12,635	50	9,026	37	9,421	43
Newbler 2.5 pre release with URT Contigs	11,353	45	8,222	34	8,551	39
Newbler 2.5 pre release with URT Isotigs	11,423	45	8,256	34	8,604	39
Newbler 2.5 without URT Isotigs	12,644	50	9,029	37	9423	43
CAP3 version 2010	11,690	46	8,350	34	8,761	40
CAP3 version 2012	11,761	47	8,436	35	8,815	40
Celera	11,767	47	8,442	35	8,825	40
CLCBio	9,862	39	6,795	28	7,210	33
iAssembler	12,881	51	9,226	38	9,514	43
MIRA version 3.2.0rc1	11,693	46	8,346	34	8,761	40
Newbler 2.3 without URT Contigs	12,722	50	9,125	38	9,494	43
Newbler 2.5 with URT Contigs	12,810	51	9,211	38	9,556	43
Newbler 2.5 with URT Isotigs	12,639	50	9,059	37	9,419	43
Newbler 2.5 without URT Contigs	12,716	50	9,086	38	9,489	43
Newbler 2.6 with URT Contigs	12,385	49	8,922	37	9,266	42
Newbler 2.6 with URT Isotigs	12,481	49	8,906	37	9,304	42
Newbler 2.6 without URT Contigs	12,034	48	8,624	36	9,021	41
Newbler 2.6 without URT Isotigs	12,362	49	8,886	37	9,249	42
Phrap	11,754	47	8,415	35	8,800	40

Table 3.13: TBLASTN using complete transcriptome datasets from *B. malayi* and *M.hapla* against the different *P. superbus* transcriptome assemblies. A BLAST cutoff of > 60 bits was used.

Transcriptome Assembly	***B. malayi*(query)**	**%**	***M. hapla*(query)**	**%**
MIRA version 3.4.0.1	8,777	41	4,834	37
Newbler 2.3 without URT Isotigs	10,007	47	5,287	40
Newbler 2.5 pre release with URT Contigs	8,700	41	4,674	36
Newbler 2.5 pre release with URT Isotigs	8,724	41	4,710	36
Newbler 2.5 without URT Isotigs	9,853	46	5,301	41
CAP3 version 2010	8,892	42	4,790	37
CAP3 version 2012	8,844	41	4,830	37
Celera	8,757	41	4,837	37
CLCBio	7,679	36	3,680	28
iAssembler	10,441	49	5,352	41
MIRA version 3.2.0rc1	9,007	42	4,793	37
Newbler 2.3 without URT Contigs	9,833	46	5,340	41
Newbler 2.5 with URT Contigs	10,290	48	5,400	41
Newbler 2.5 with URT Isotigs	10,050	47	5,309	41
Newbler 2.5 without URT Contigs	9,892	46	5,342	41
Newbler 2.6 with URT Contigs	10,046	47	5,208	40
Newbler 2.6 with URT Isotigs	10,078	47	5,203	40
Newbler 2.6 without URT Contigs	9,615	45	4,978	38
Newbler 2.6 without URT Isotigs	9,977	47	5,195	40
Phrap	8,724	41	4,824	37

Table 3.14: BLASTX using complete transcriptome datasets from *C. elegans*, *P. pacificus* and *C. briggsae* as databases against the different *P. superbus* transcriptome assemblies as queries. A BLAST cutoff of > 60 bits was used.

Transcriptome Assembly	***C. elegans*(db)**	**%**	***P.pacificus*(db)**	**%**	***C. briggsae*(db)**	**%**
MIRA version 3.4.0.1	10,136	54	9,193	49	10,127	54
Newbler 2.3 without URT Isotigs	10,352	39	9,157	34	10,264	38
Newbler 2.5 pre release with URT Contigs	7,586	48	6,776	43	7,547	47
Newbler 2.5 pre release with URT Isotigs	8,188	55	7,395	49	8,159	55
Newbler 2.5 without URT Isotigs	10,341	39	9101	34	10,283	39
CAP3 version 2010	8,202	46	7,305	41	8,133	45
CAP3 version 2012	9,922	49	8,957	45	9,905	49
Celera	10,127	54	9,161	49	10,134	54
CLCBio	9,205	24	9,153	23	9,049	23
iAssembler	25,290	24	22,141	21	24,841	24
MIRA version 3.2.0rc1	8,202	46	7,306	41	8,140	45
Newbler 2.3 without URT Contigs	12,636	46	11,385	41	12,633	46
Newbler 2.5 with URT Contigs	18,287	39	16,377	35	18,164	39
Newbler 2.5 with URT Isotigs	11,417	43	10,175	39	11,312	43
Newbler 2.5 without URT Contigs	11,196	44	9,972	39	11,146	44
Newbler 2.6 with URT Contigs	16,987	41	15,283	37	16,787	41
Newbler 2.6 with URT Isotigs	10,613	38	9,307	33	10,486	37
Newbler 2.6 without URT Contigs	13,564	43	12,226	38	13,446	42
Newbler 2.6 without URT Isotigs	16,932	40	15,286	36	16,714	40
Phrap	9,915	55	8,935	49	9,932	55

Table 3.15: BLASTX using complete transcriptome datasets from *B. malayi* and *M. hapla* as databases and individual *P. superbus* transcriptome assemblies as queries. A BLAST cut-off of 60 bits was used.

Transcriptome Assembly	***B. malayi*(db)**	**%**	***M. hapla*(db)**	**%**
MIRA version 3.4.0.1	9,002	48	8,188	44
Newbler 2.3 without URT Isotigs	9,031	34	7,936	30
Newbler 2.5 pre release with URT Contigs	6,714	42	5,997	38
Newbler 2.5 pre release with URT Isotigs	7,255	48	6,557	44
Newbler 2.5 without URT Isotigs	9,050	34	7,930	30
CAP3 version 2010	7,164	40	6,408	36
CAP3 version 2012	8,749	44	7,874	39
Celera	8,962	48	8,148	44
CLCBio	12,941	33	6,538	17
iAssembler	22,458	21	19,972	44
MIRA version 3.2.0rc1	7,167	40	6,411	36
Newbler 2.3 without URT Contigs	11,240	41	10,055	37
Newbler 2.5 with URT Contigs	16,663	36	14,066	30
Newbler 2.5 with URT Isotigs	10,383	39	8,822	33
Newbler 2.5 without URT Contigs	9,848	39	8,714	34
Newbler 2.6 with URT Contigs	15,834	38	13,081	32
Newbler 2.6 with URT Isotigs	9,272	33	8,060	29
Newbler 2.6 without URT Contigs	12,626	40	10,504	33
Newbler 2.6 without URT Isotigs	16,330	39	12,957	31
Phrap	8,754	48	7,983	44

bus CEG homologues against their full length counterparts from *C. elegans*. The results obtained are presented in Figure 3.5 and they show that for all the *P. superbus* assemblies there is a lower transcript coverage of the CEG genes at the 5' end of the *P. superbus* contigs. This analysis also shows that the genes from the CAP3 v2010 hybrid assembly and Newbler 2.6 without URT isotigs and contigs have a higher percentage coverage at each point along the matched *C. elegans* CEG genes than those from any of the other *P. superbus* assemblies. (At the time this analysis was carried out the CAP3 v2012 hybrid assembly was not available for inclusion in this study).

3.3.5 Quality Evaluation of *P. superbus* Assemblies by Identifying Under-assembled Contigs and Chimeras

In Section 3.3.3 it was found that when full transcriptome datasets from four nematodes (*C. elegans*, *C. briggsae*, *P. pacifcus* and *B. malayi*) were used as a BLAST query against the *P. superbus* transcriptome assemblies, a higher proportion of hits was recovered for the *C. elegans* sequences than for the other nematode datasets tested. So even when only 50% of the *C. elegans* genes had homologues in the *P. superbus* assemblies (Table 3.13), the *C. elegans* transcriptome was selected as the reference to investigate the number of under-assembled contigs and chimeric contigs in the different *P. superbus* transcriptome assemblies.

To investigate the number of under-assembled contigs in different *P. superbus* assemblies, the number of *P. superbus* contigs that had reciprocal best BLAST hits to the same non-overlapping region of a *C. elegans* gene was obtained (see Figure 3.6). The number of under-assembled contigs is then expressed as a percentage of the total number of contigs with significant similarity to any *C. elegans* gene.

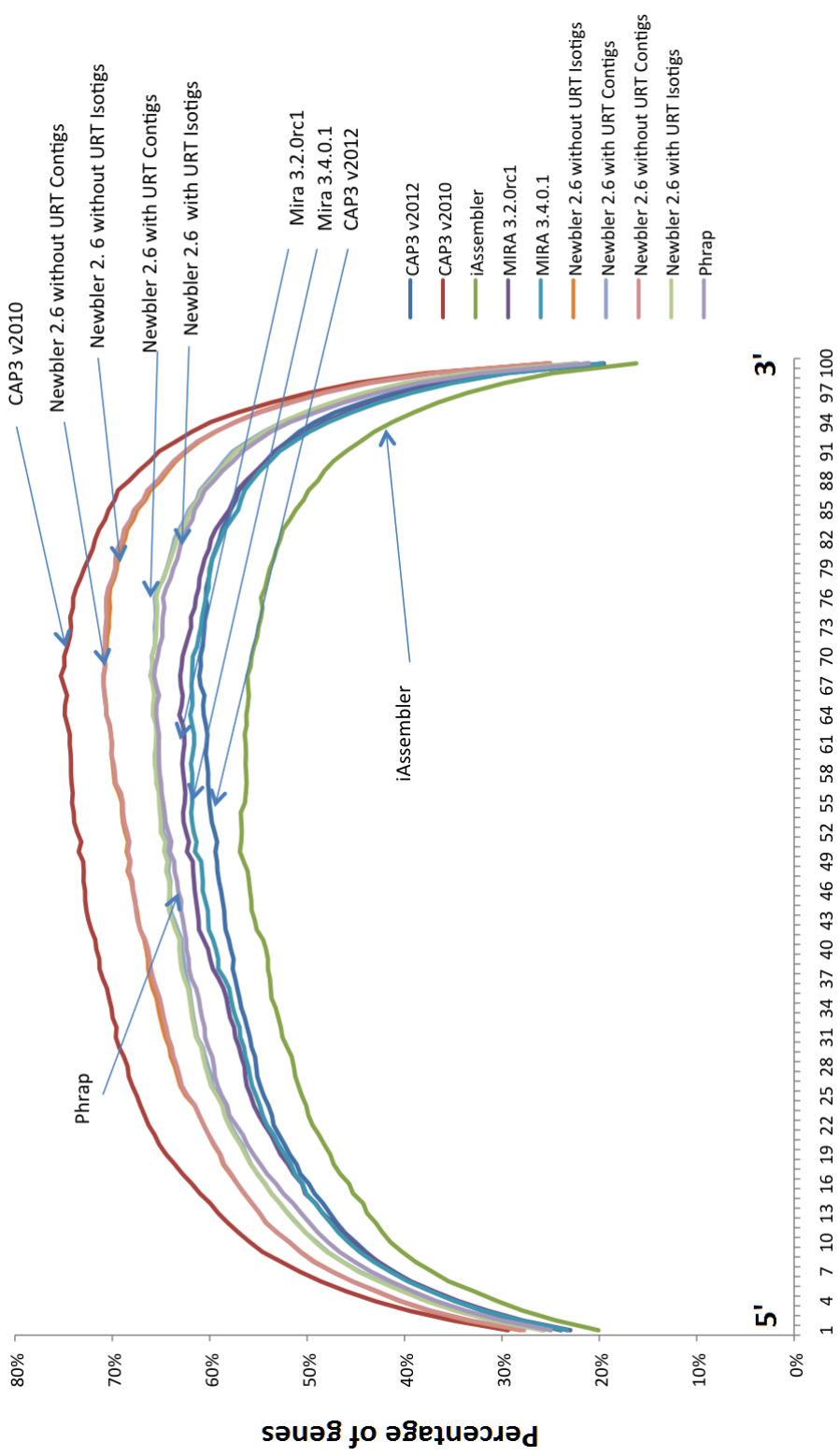


Figure 3.5: The percentage coverage across all the *C. elegans* CEG (Core Eukaryotic Genes) set (Parra et al., 2009) by different *P. superbus* transcriptome assemblies. The x-axis represents each CEG gene from its 5' end to the 3' end and y-axis shows the percentage of genes from each transcriptome assembly which cover each stretch of the matched *C. elegans* CEG gene.

Chimeras are mis-assembled contigs where the contig in question is made up of what should be two independent contigs (see Figure 3.7). These are identified by looking at the reciprocal best BLAST hits to *C. elegans* genes for each contig, and where two non-overlapping regions in the same contig are similar to different *C. elegans* genes. The number of chimeric contigs is then expressed as a percentage of the total number of contigs with significant similarity to any *C. elegans* gene.

The percentage of under-assembled and chimeric contigs found in each assembly are presented in Table 3.16. The results obtained show that Newbler 2.6 without URT Isotigs had the least number of putative under-assembled contigs (5.85%), followed by CAP3 version 2012 (7.16%), while the iAssembler transcriptome had the highest number of putative under-assembled contigs (18.52%). Conversely, the iAssembler dataset had the lowest number of putative chimeras (1.67%). Overall, it seems that for most assemblers the generation of chimeras occurs less frequently than the under-assembly of contigs. The highest number of putative chimeras observed was 5.43% for the Phrap assembly.

Figure 3.6: When more than one contig has the same gene as the reciprocal top-BLAST hit and the contigs align to non-overlapping regions of the *C. elegans* gene they can be defined as under-assembled contigs.

Figure 3.7: Contigs that had reciprocal top BLAST hits with *C. elegans*, which had different non-overlapping regions of the contig that matched (as a top BLAST hit) to different genes.

3.3.6 Quality Evaluation of *P. superbus* Assemblies by Assessing Length Coverage of Individual *C. elegans* Genes

Figure 3.8 uses variable width box plots to show the distribution of the percentage gene coverage of *C. elegans* genes matched by transcripts from each *P. superbus* assembly. Ten *P. superbus* assemblies were investigated in this experiment and the results obtained show that the Newbler 2.6 assemblies and both CAP3 hybrid assemblies perform better than the other assemblies in terms of their percentage coverage of individual *C. elegans* genes. The assemblies generated with the earlier versions of the Newbler assembly software were not included in this analysis.

Table 3.16: The number of putative under-assembled and chimeric contigs in different *P. superbus* transcriptome assemblies detected in reciprocal best BLAST analyses of the *C. elegans* transcriptome.

Assembly	**No. of matched *C. elegans* genes**	**No. of matched *P. superbus* contigs**	**Unassembled Contigs (%)**	**Chimeric Contigs (%)**
CAP3 v2010	7,461	8,216	10.78	3.42
CAP3 v2012	6,761	7,049	7.16	3.84
iAssembler	8,854	11,109	18.52	1.67
MIRA version 3.2.0rc1	7,479	8,330	11.61	3.19
MIRA version 3.4.0.1	7,465	8,357	11.82	2.99
Newbler 2.6 without URT Contigs	6,048	6,332	8.7	4.75
Newbler 2.6 with URT Isotigs	7,290	7,970	10.53	3.93
Newbler 2.6 without URT Isotigs	5,899	5,984	5.85	4.73
Newbler 2.6 with URT Contigs	7,455	8,435	13.39	3.89
Phrap	7,364	7,912	10.55	5.43

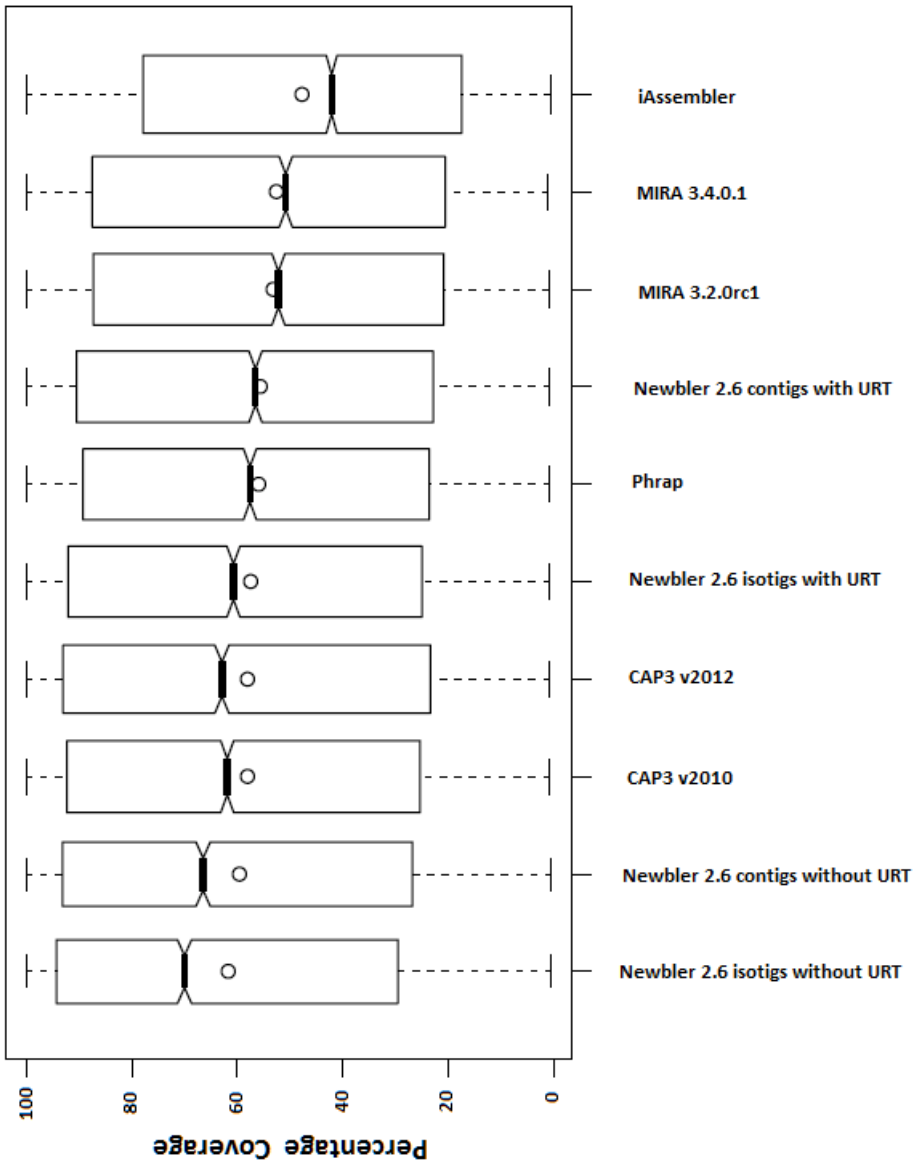


Figure 3.8: Variable width box plots showing the gene-coverage for all the *C. elegans* genes which had hits in individual *P. superbus* transcriptome assemblies. The y-axis shows the % coverage, the circles are the means and the notches are the medians. The width of each box is proportional to the number of contigs in each transcriptome.

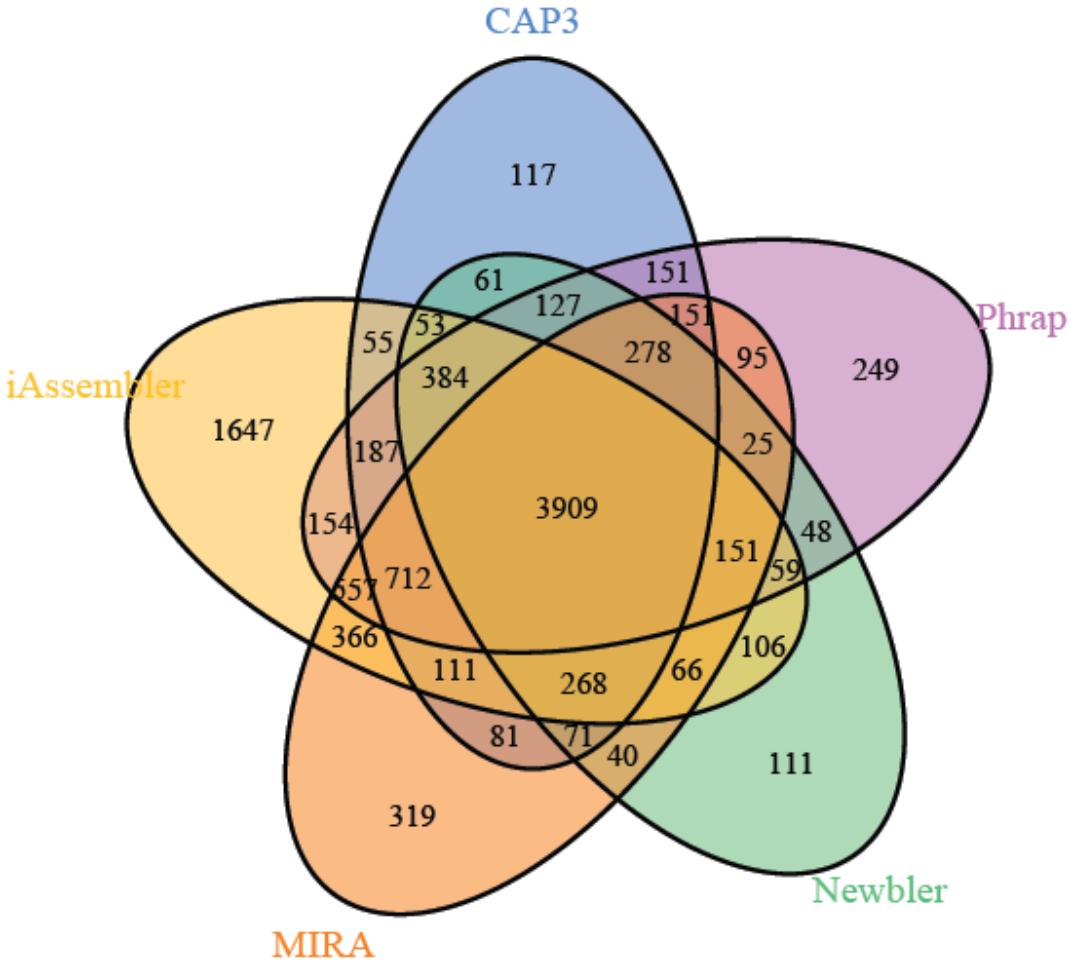


Figure 3.9: Venn diagram showing the break-down of the *C. elegans* genes where its homologues have been identified in up to five different *P. superbus* assemblies. The assemblies were as follows: Newbler without URT Isotigs; CAP3 v2012; MIRA v3.4.0.1 (7,465); Phrap and iAssembler.

Figure 3.9 is a Venn diagram that shows the break-down of the *C. elegans* genes where its homologs have been identified in up to five different *P. superbus* assemblies. The assemblies selected and the number of homologous *C. elegans* sequences contained in each assembly were as follows: Newbler 2.6 without URT Isotigs (5,899); CAP3 v2012 (6,761); MIRA v3.4.0.1 (7,465); Phrap (7,364) and iAssembler (8,854). This diagram shows that there are 3,909 *C. elegans* genes

which had homologous sequences in all five *P. superbus* transcriptome assemblies. It can also be seen that each *P. superbus* assembly had hits to *C. elegans* genes which were unique to that assembly alone: Newbler 2.6 without URT Isotigs (111); CAP3 v2012 (117); MIRA v3.4.0.1 (319); Phrap (249) and iAssembler (1,647). Interestingly, the iAssembler transcriptome, which has the largest number of contigs with the smallest mean contig size, had the largest number of homologous *C. elegans* sequences uniquely present in that assembly. This raises the possibility that the iAssembler may have recovered or assembled transcriptome *P. superbus* sequences that were not assembled by the other assemblers.

From Figure 3.10 it can be seen that for the dataset from the Venn diagram (Figure 3.9), the *P. superbus* genes which are assembled by all five assemblers have the highest average transcript coverage in length. The lowest average length coverage is detected for genes that are present only in one assembly. Figure 3.11 breaks it down further, showing length coverage for *C. elegans* genes where its homologs are found only in one assembler. The results show a clear trend where uniquely identified homologs (from each assembly) are generally not full length transcripts but only represent partial gene sequences. So, while the iAssembler dataset contain the largest number of *C. elegans* homologues, the data in Figure 3.11 show that its contigs that uniquely match a *C. elegans* gene cover on average less than 20% of that gene. This observation also extends to other assemblies (Figure 3.11) where unique *C. elegans* gene coverage is typically less than 20% of that *C. elegans* gene.

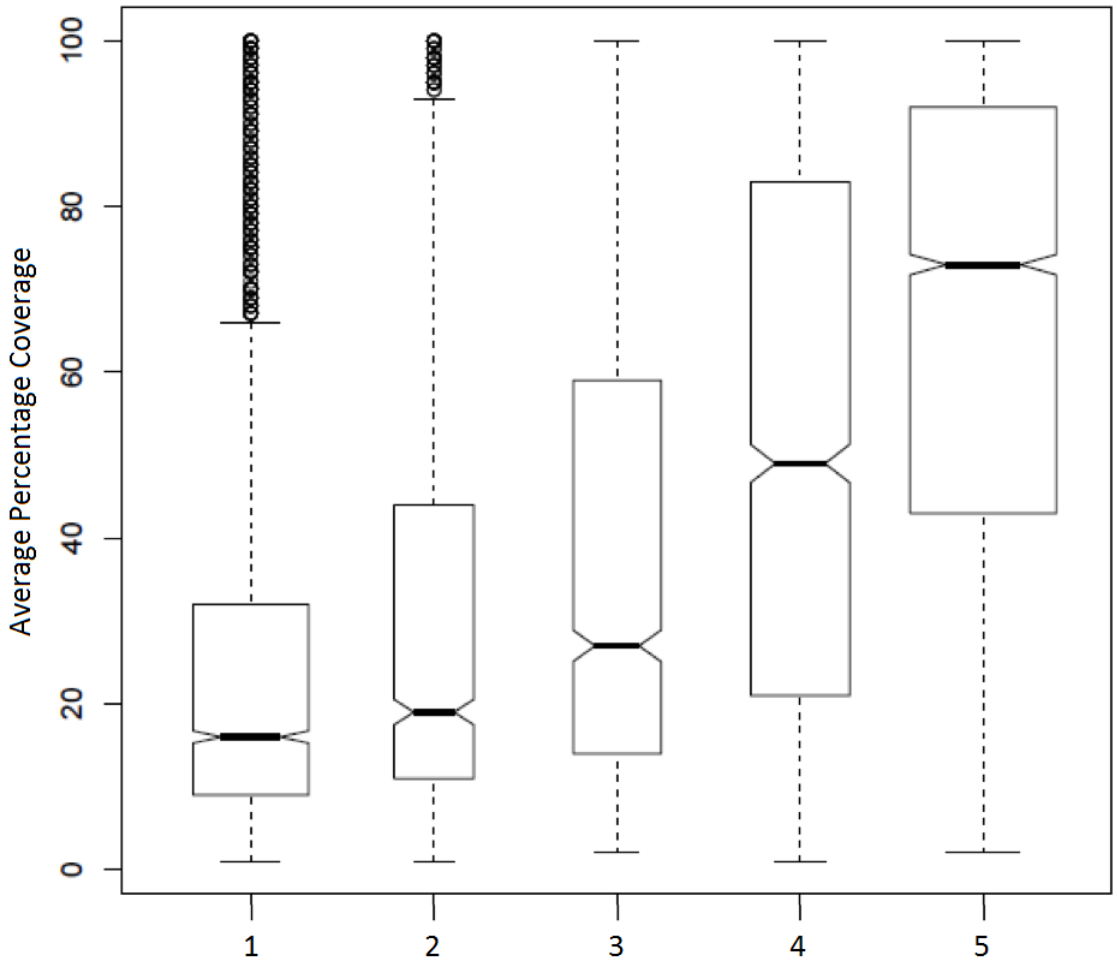


Figure 3.10: Variable width boxplots showing the percentage coverage of *C. elegans* by *P. superbus* homologs based on the number of *P. superbus* transcriptome assemblies that contain the gene. The circles at the top and bottom are outliers to the main distribution, and the dark bars are the medians.

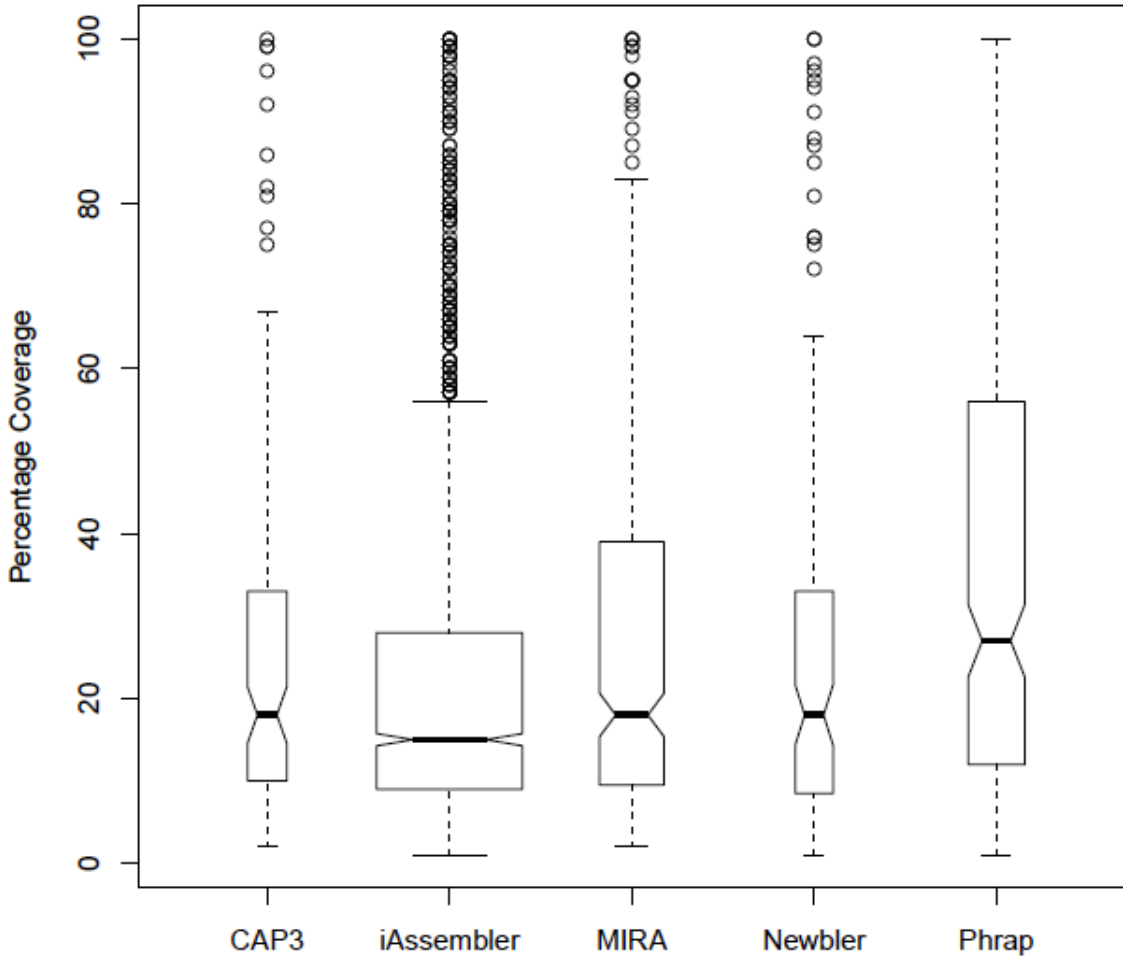


Figure 3.11: Variable width boxplots showing the percentage coverage of *C. elegans* by *P. superbus* homologs when the *P. superbus* genes are found by only one assembler. The circles at the top and bottom are outliers to the main distribution, and the dark bars are the medians.

3.3.7 Ranking *P. superbus* Transcriptome Assemblies using Unweighted Quality Criteria

To select the assembly for downstream analyses, each assembly was scored based on the unweighted sum of its rankings for a number of quality criteria, mentioned above, i.e., the lower the total score the better. Some of these criteria are likely

not to be independent and may be positively correlated (e.g., number of contigs, and percentage of under-assembled contigs; coverage of *C. elegans* genes, *C. elegans* gene length coverage ≥80% and coverage of the 5' end of *C. elegans* genes; number of *C. elegans* BLAST hits) or negatively correlated (e.g., number of contigs and mean contig length; N50 length and number of contigs in N50). The assembly metrics and ranked scores are presented in Table 3.17. Transcriptome size is included in Table 3.16, but was not used as a ranking criterion. It appears to be inflated in the iAssembler, and possibly also in MIRA, but it is difficult to decide among the other assemblies because the target transcriptome size is not known. The results obtained show that two assemblies, CAP3 v2012 and Newbler 2.6 without URT isotigs, achieved a substantially better rank score than the other four assemblies which were evaluated. Thus, either of these transcriptomes could be recommended/selected for downstream annotation of the *P. superbus* transcriptome. Analysis of their contig metrics show that the main difference between these two assemblies is the larger number of contigs in the CAP3 assembly (31,836) as compared to the Newbler isotigs assembly (14,960); the mean contig length is longer for the Newbler isotigs assembly, but the number of contigs > 1kb is larger for the CAP3 assembly (9,350) than for the Newbler isotigs assembly (6,866) and the coverage of the 5' end of the *C. elegans* genes is slightly higher for the CAP3 assembly (56%) than for the Newbler isotigs assembly (52%).

Table 3.17: Assembly metrics, quality assessment and ranked scores for five assemblies of the *P. superbus* transcriptome obtained using different assembly programs and strategies. Ranked scores (in parentheses) are presented for key quality metrics.

	iAssembler	**MIRA v 3.4.0.1**	**Newbler 2.6 without URT Isotigs**	**Newbler 2.6 with URT isotigs**	**CAP3 v2012**	**Phrap**
Number of contigs	105,156 (6)	42,161 (5)	14,960 (1)	25,402 (2)	31,836 (4)	26,360 (3)
Mean contig length (bp)	481 (6)	696 (5)	1,031(1)	871 (3)	830 (4)	898 (2)
Number of contigs > 1kb	7,881(4)	7,617 (5)	6,866 (6)	8,612 (3)	9,350(1)	8,842 (2)
Contig N50 (bp)	545 (6)	791 (5)	1,184(1)	1,136 (3)	1,036 (4)	1,138 (2)
Number of contigs in N50	6,165 (5)	12,183(1)	4,805(6)	6,848 (4)	8,732 (2)	7,050 (3)
Number of contigs > 1kb	7,881(4)	7,617 (5)	6,866 (6)	8,612 (3)	9,350 (1)	8,842 (2)
Transcriptome Size (Mb)	51	29	15	22	26	24
Mergable contigs %	18.52(6)	11.82 (5)	5.85(1)	10.53 (3)	7.16 (2)	10.55 (4)
Chimeras %	1.67(1)	2.99(2)	4.73 (5)	3.93(4)	3.84(3)	5.43 (6)
CEG orthologs	231 (5)	231 (5)	239 (1)	237 (2)	237 (2)	235 (4)
***C. elegans* BLAST hits**	8,854 (1)	7,465 (2)	5,899(6)	7,290 (4)	6,761 (5)	7,364 (3)
NemBase 4+BLAST hits	86,773 (4)	86,475 (5)	87,501(2)	87,044 (3)	88,467 (1)	84,447 (6)
Mean *C. elegans* gene coverage (%)	47.68 (6)	52.74 (5)	61.76 (1)	57.67 (3)	58.01 (2)	55.9 (4)
***C. elegans* gene coverage > 80%**	23.76(6)	30.17(5)	42.27 (1)	36.77 (3)	38.66 (2)	33.78 (4)
Coverage of *C. elegans* 5' end (%)	41(6)	46(5)	52(2)	49 (3)	56 (1)	48 (4)
Sum of Ranks	61	55	34	40	33	47

3.4 Discussion

This chapter showed the transcriptome experiment from the early stages of sample preparation and normalisation through to the sequencing and assembly. The functional annotation is a significant step which could not be undertaken without a good deal of effort being put into choosing the best assembly for the dataset and the type of biological questions being asked. It was decided to establish a set of tests or metrics by which to evaluate the performance of the assemblers and that way a 'best' or set of 'best' assemblies could be chosen for the downstream functional annotation analysis discussed in the next chapter. These metrics were chosen on the basis that they had been used previously in other publications, but this evolved into choosing them based on discussion and analysis of previous metric results. What had previously been thought to be an initial routine stage in the annotation process, became a study in its own right.

As expected, there was a greater number of reads generated from the unnormalised rather than the normalised sample. This is expected, because the normalisation step functions to reduce the amount of housekeeping genes. It can be assumed that these housekeeping genes would also be present in a high copy number. Default parameters were used with the assemblies, as it was felt that there were so many different assemblies having to also identify the different parameters on offer, it would place unsustainable time constraints on the project.

An interesting observation to note was that different assemblers generated contigs of different lengths and thus, post-filtering by length (>150bp) meant more data was lost at this stage for some assemblies than for others. It was deemed that the Newbler Isotigs performed best in this way, as very few sequences were removed through size filtering and thus resulted in minimal data loss. Furthermore, an examination of the number of base pairs in the finished assembly showed that Newbler contigs and Celera are perhaps too strict when choosing which reads

are deemed acceptable for assembly.

Comparing assemblies based on N50 length is an interesting metric. This was the popular metric of choice to be optimised in the early stages of next-generation sequencing. This metric does give a good estimate of whether or not full-length transcripts have been achieved. If the gene size in *C. elegans* is used as a comparison, any assemblies with an N50 of 1,000-1,500bps could be deemed reasonable. These are the CAP3s, Phrap and the newer Newbler contigs and isotigs. Completing the pattern, Celera, CLCBio, Mira and iAssembler don't generate comparable N50 values. An argument could be made here that this is due to the stringency of their alignments and a further investigation into the parameters on offer for these assemblers could lead to better results.

P. superbus homologs to other organisms from different assemblies were identified by similarity searches using BLAST (see Section 3.3.2). While this is primarily done for the purpose of annotation, i.e., functional inference, it also proved useful for assessment of assembly quality. For example, the recovery of CEG genes acted as a basic 'litmus' test for the assemblies (Section 3.3.2); all assemblies passed this test, apart from CLCBio, by virtue of having almost all CEG genes represented. The extent to which genes from other nematodes are represented in each assembly (by homologs) also provide clues to quality (Section 3.3.3), e.g., the Celera assembly has significantly lower hits to other nematode sequences relative to others. In the case of iAssembler, nematode homologs tend to be as well represented as in the other datasets, but this is achieved through exceptionally high numbers of iAssembler transcripts compared to others, i.e., indication that the iAssembler dataset may be under-assembled.

Other information has also been gleaned from the BLAST analyses involving other nematode genomes/transcriptomes. For all assemblies, *C. elegans* came out consistently as the closest to *P. superbus* in terms of homologous sequences, with

about 50% of its genes having homologs in the *P. superbus* assemblies (Section 3.3.3). *P. pacificus* was found to be least similar to *P. superbus*, where around 35% of its genome have *P. superbus* homologs. This is surprising as *P. pacificus* is a free-living nematode and it is more closely related phylogenetically to *P. superbus* than is *C. elegans* as shown in Figure 1.1. One likely explanation for this discrepancy is that the *P. pacificus* dataset is probably not as extensively curated as that of the model organism *C. elegans*.

Finally, evaluation of the quality of the different *P. superbus* assemblies was performed by assessing contig coverage of individual CEG genes and *C. elegans* genes and by estimating the proportion of chimeras and under-assembled contigs in each assembly as shown in Sections 3.3.4, 3.3.5 and 3.3.6. This analysis also shows that the genes from the CAP3 v2010 hybrid assembly and Newbler 2.6 without URT Isotigs and contigs have a higher percentage coverage at each point along the matched *C. elegans* and CEG genes than those from any of the other *P. superbus* assemblies. It was also found that Newbler 2.6 without URT Isotigs had the least number of putative under-assembled contigs (5.85%), followed by CAP3 v2012 (7.16%), while the iAssembler dataset had the highest number of putative under-assembled contigs (18.52%). Overall, it seems that for most assemblers the generation of chimeras occurs less frequently than the under assembly of contigs. The highest number of putative chimeras observed was 5.43% for the Phrap assembly.

To select the assembly for downstream analyses, each assembly was scored based on the unweighted sum of its rankings for a number of quality criteria (Section 3.3.7). The results obtained show that two assemblies, CAP3 v2012 and Newbler 2.6 without URT Isotigs, have a substantially better rank score than the other four assemblies which were evaluated. Thus, either of these transcriptomes could be recommended/selected for analysis an annotation of the *P. superbus* tran-

scriptome. Analysis of their contig metrics show that the main difference between these two assemblies is the larger number of contigs in the CAP3 assembly (31,836) as compared to the Newbler isotigs assembly (14,960); the mean contig length is longer for the Newbler isotigs assembly, but the number of contigs >1kb is larger for the CAP3 assembly (9,350) than for the Newbler Isotigs assembly (6,866). Transcript coverage at the 5' end of *C. elegans* genes is also higher for the CAP3 assembly (56%) than for the Newbler Isotigs assembly (52%).

Taking into account all of the metrics used and the aim of this transcriptome study, it was deemed that it would be wise to immediately exclude iAssembler, CLCBio, Celera and early Newbler versions from inclusion in any further studies. Since Phrap, CAP3, and MIRA neither shone nor performed exceptionally badly it was clear that the best performance on all metrics were the Newbler without URT Isotigs and CAP3 v2012 assemblies. Newbler is a specific assembler developed by 454(Roche) to accommodate the types of reads coming from their own in-house sequencer and it is clear that the most recent version of the Newbler software performs very well relative to the earlier versions of Newbler.

Chapter 4

Transcriptome Annotation

4.1 Introduction

Assembly and generation of a dataset is crucial to the success of any transcriptome project. Further downstream, functional annotation of the dataset allows for a comprehensive picture of the quantification of genes and their isoforms. This is very computationally intense and leads to many challenges, both of time and resources. It is, however, a necessary step in the establishment of expression levels, function and novel components yet to be explored. Using transcripts of organisms stressed in different ways, we can reach a deeper understanding of how that organism is affected by stress versus how it normally functions. While we don't have a clear picture as to the function of most genes, those that are known can be used to annotate transcriptome datasets. This is done using comparison methods, and can be used to generate a 'best guess' as to the function of that gene. All cells contain the same genomic sequence, but what differentiates them is their expressed genes (Nicol et al., 2012), thus the transcriptome provides valuable insight into gene expression and was used to investigate gene expression in response to environmental stress (desiccation, cold, heat and oxidation) in this study.

4.2 Methods & Materials

4.2.1 Functional Annotation using the BLAST2GO Tool

BLAST2GO is a functional annotation tool and pipeline that was used on the transcriptome assembly (Conesa et al., 2005). BLAST2GO allows homologous mapping using BLAST. It also integrates the Gene Ontology database, Enzyme Commission database, InterPro database and KEGG database, and allows an extensive annotation of novel datasets. BLAST2GO PRO was used on an iMac 3.33 GHz Intel Core 2 Duo with 8Gb of RAM. BLAST2GO runs took approximately 30 days per run of 20,000 sequences and the software was found to need constant monitoring as it had a tendency to crash when it became overloaded (Conesa et al., 2005). The pipeline used is outlined in Figure 4.1. The BLAST analysis took about seven days to complete using BLAST2GO. GO, KEGG and EC were quick to run in the PRO mode and the remaining time was spent running InterProScan.

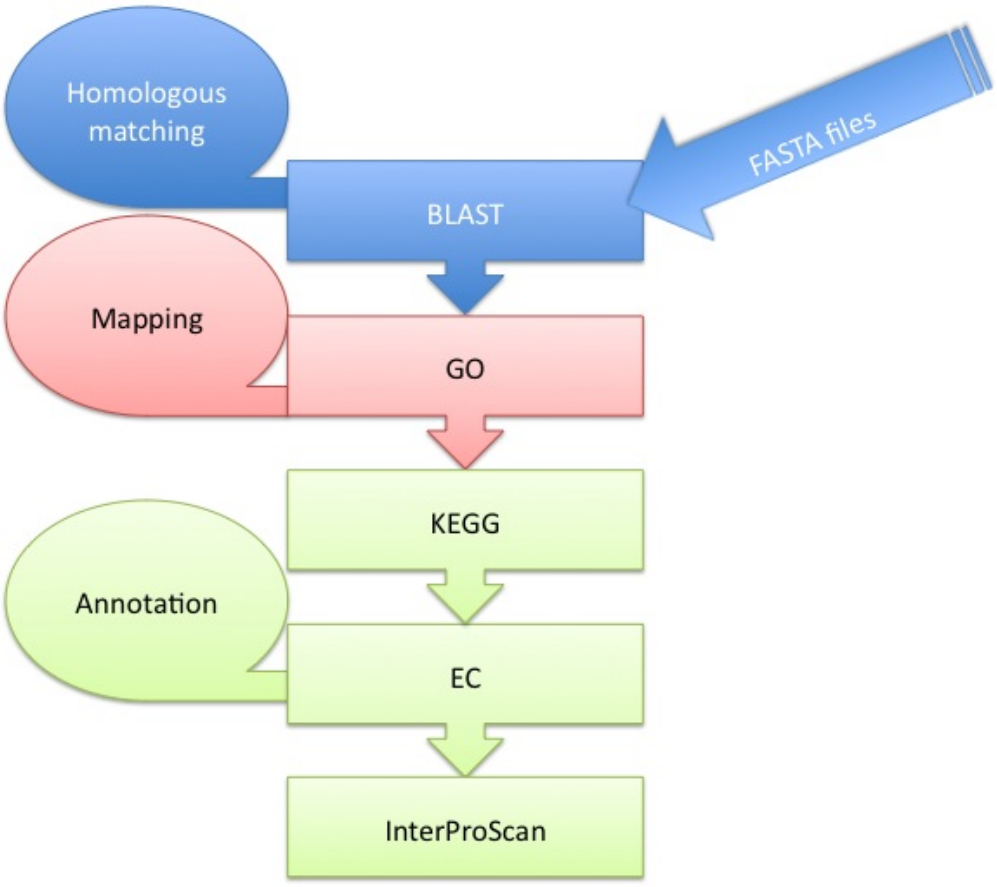


Figure 4.1: BLAST2GO annotation pipeline beginning with sequences in FASTA format and resulting with an annotated dataset.

4.3 Results

4.3.1 Summary

Version 2.6.3 of BLAST2GO was used to functionally annotate the transcriptome. BLAST2GO is an extensive tool for annotating and data mining sequences using Gene Ontology (GO). It integrates visualisation and statistical software, including InterPro, enzyme codes, KEGG pathways, GO direct acyclic graphs (DAGs) and GOSlim (Conesa et al., 2005). The annotation step is based on homology transfer. As has been outlined previously, the Newbler 2.6 Isotigs without URT was deemed one of the most suitable assemblies and was chosen for annotation. Firstly, a BLASTX against the NCBI NR database was performed. This was followed by GO mapping and downstream annotation. The number of sequences found at each length is shown in Figure 4.2. The majority of sequences in the *P. superbus* assembly were found to be of length 750-1,500bps, which is the average length of a nematode gene.

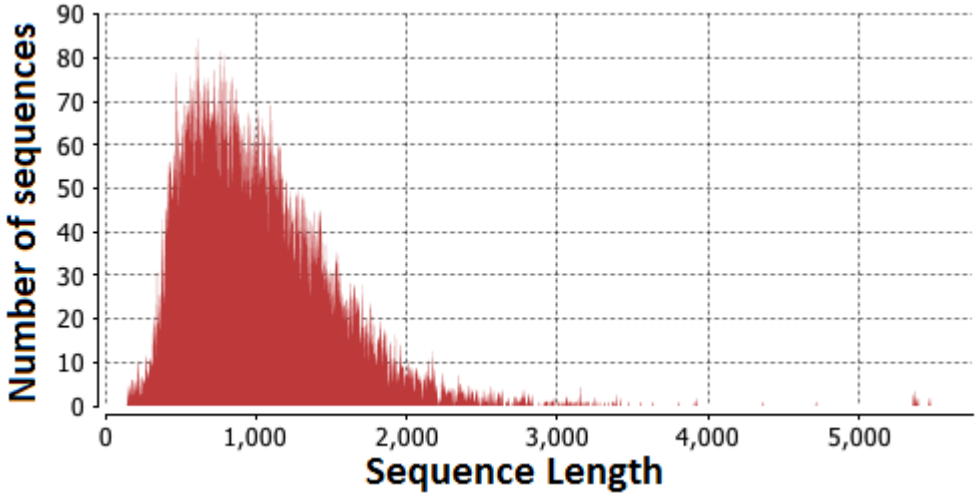


Figure 4.2: Number of sequences in the *P. superbus* transcriptome assembly in proportion to length of sequence.

Following the removal of sequences shorter than 150bps, 14,960 post filtering sequences were put through the pipeline. Of these sequences, 6,140 (41%) did not generate a BLAST hit. 712 (4%) generated a BLAST result, but no further downstream annotation, and the remaining 8,170 (55%) were annotated. These results can be seen in Figure 4.3.

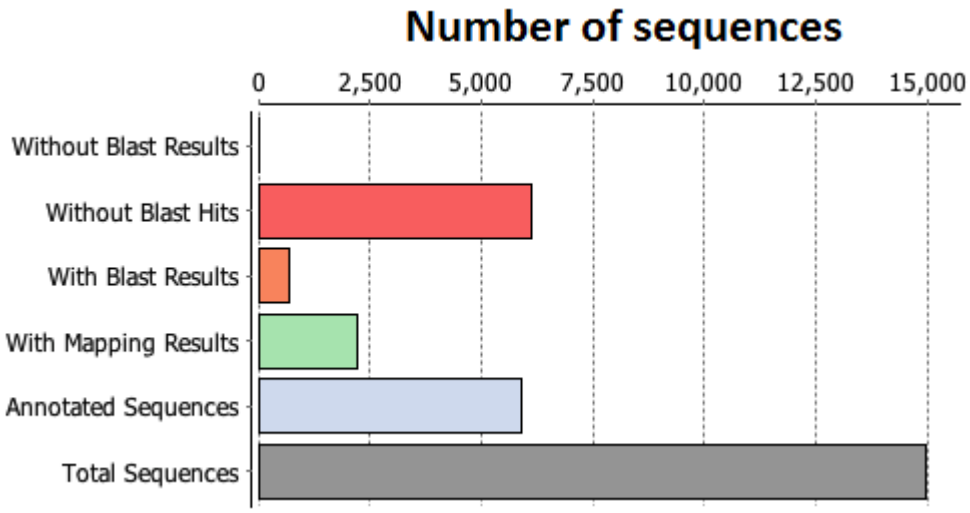


Figure 4.3: Data distribution of the BLAST analysis of the *P. superbus* transcriptome dataset post BLAST2GO annotation. Graph is colour coded to match BLAST2GO output as can be seen on the accompanying CD. (File: BLAST2GO_output_transcriptome.dat).

4.3.2 BLAST Results

BLASTX using NR as a database with a cutoff of 1e-3 was used. The top 20 resulting BLAST hits were retained by BLAST2GO and used for annotation. The e-value depends on three parameters: the alignment itself, the length (and composition) of the query sequence, and the total length (and composition) of the sequences in the database (Altschul et al., 1990). The e-value distribution can be seen in Figure 4.4. An e-value of 0.0 indicates a perfect or self-hit. As the e-value

decreases, the probability of a hit by chance also decreases; as a result, there are less hits.

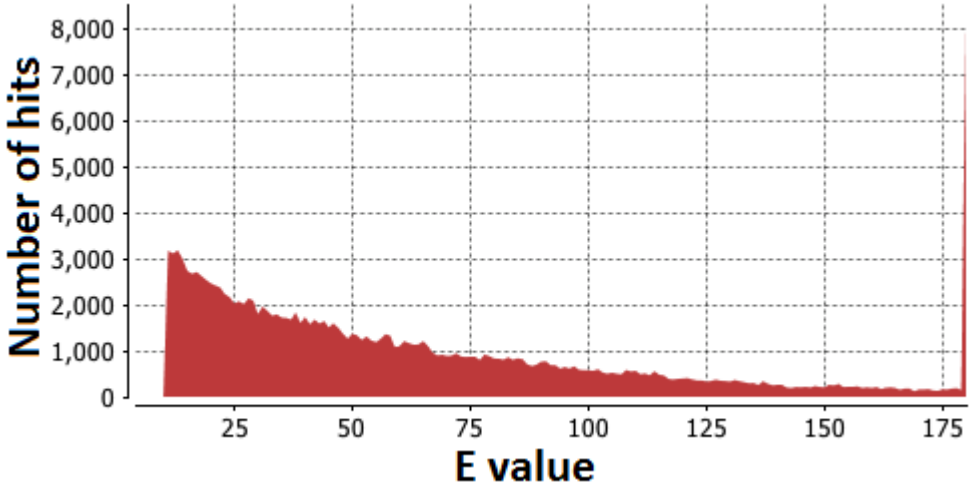


Figure 4.4: e-value distribution of the *P. superbus* transcriptome sequences which returned BLAST hits following BLASTX against NR with a cut off of 1e-3.

An analysis of the species distribution of the BLAST hits was carried out, and the majority of sequences had hits to the *Caenorhabditis*, *A. suum*, *L. loa* and *B. malayi* species. Some other species include the model organisms (*H. sapiens, M. musculus* and *D. rerio*) and insects, including the *D. melanogaster* and *Apis mellifera*. The distribution by species can be seen in Figure 4.5.

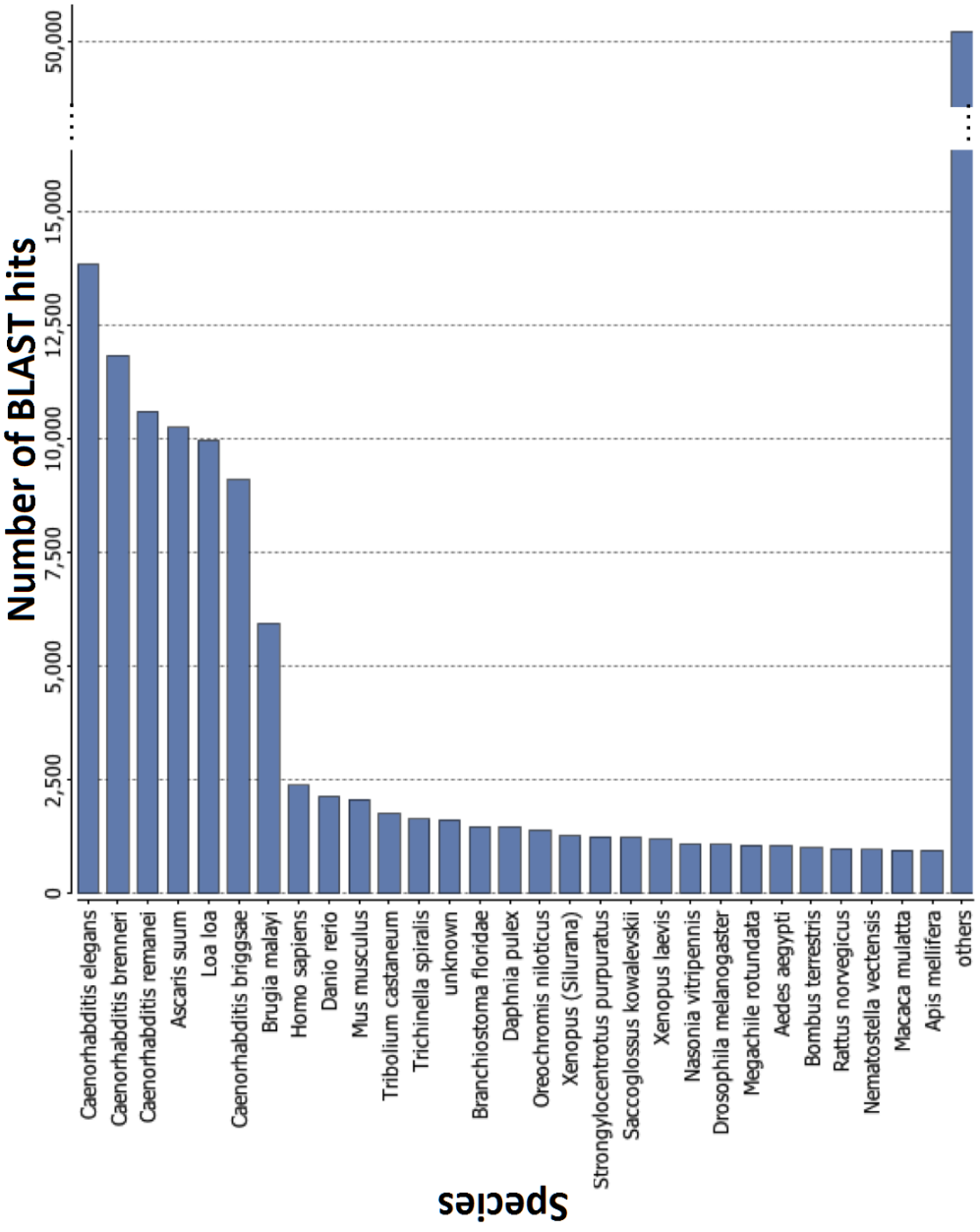


Figure 4.5: Species distribution of all BLAST hits following BLASTX against the NR database with a cut off of 1e-3.

The top hits were predominantly to *A. suum* followed by the eye worm *L. loa*. The *Caenorhabditis* species and *B. malayi* were also represented in the top 7 species. This is not surprising, as although *A. suum* is a Clade III nematode, it is in the class *Rhabditia*, as are *P. superbus* and the *Caenorhabditis* species. The top BLASTX hit distribution by species can be seen in Figure 4.6.

The sequence similarity goes from 40% to 60%, before decreasing. Any hits of 100% could be "self-hits" or sequence pattern of 100% similarity. Figure 4.7 shows the similarity of the query set (transcriptome dataset) and the selected database (NR).

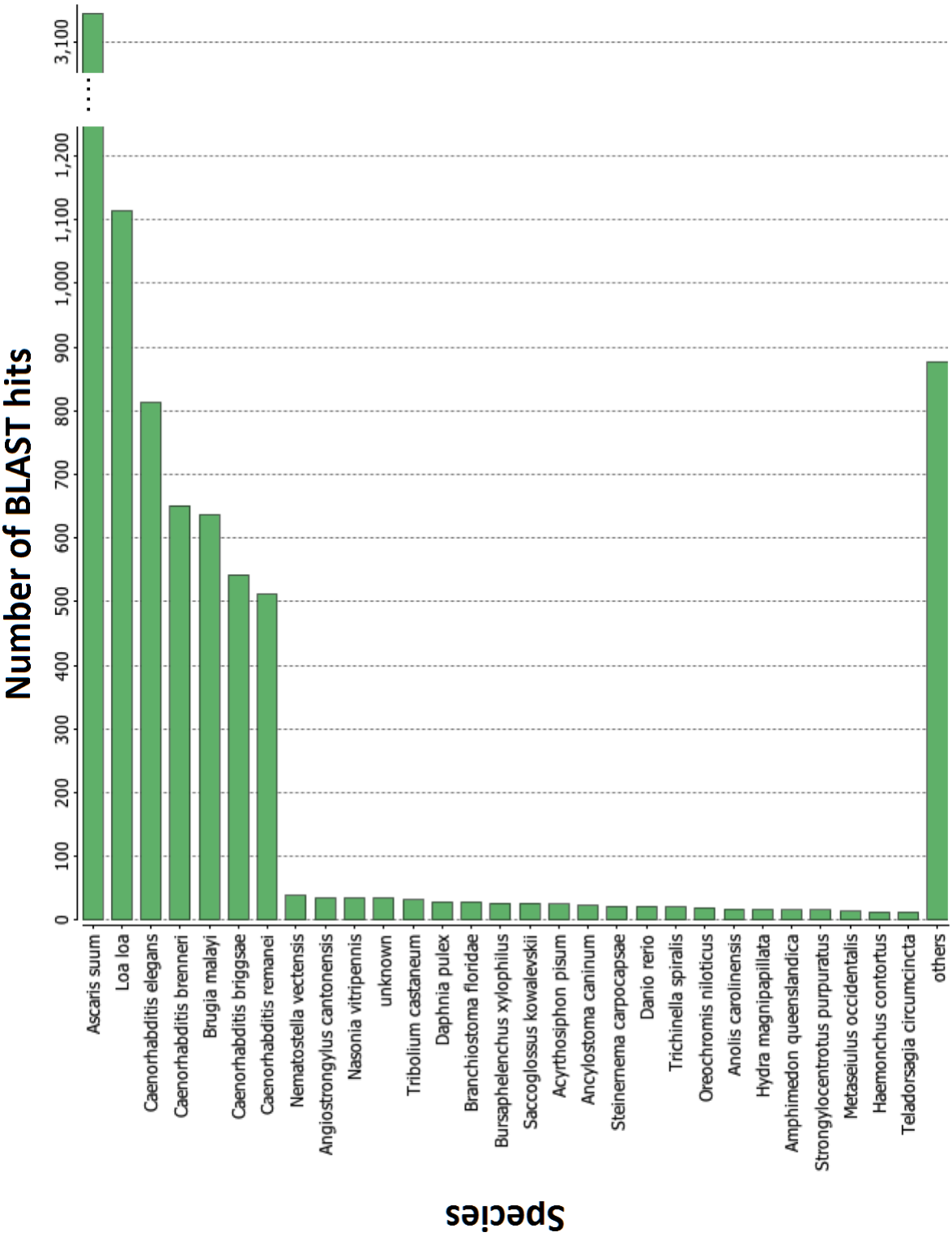


Figure 4.6: Species distribution of the top BLAST hits returned following BLASTX against the NR database with a cut off of 1e-3.

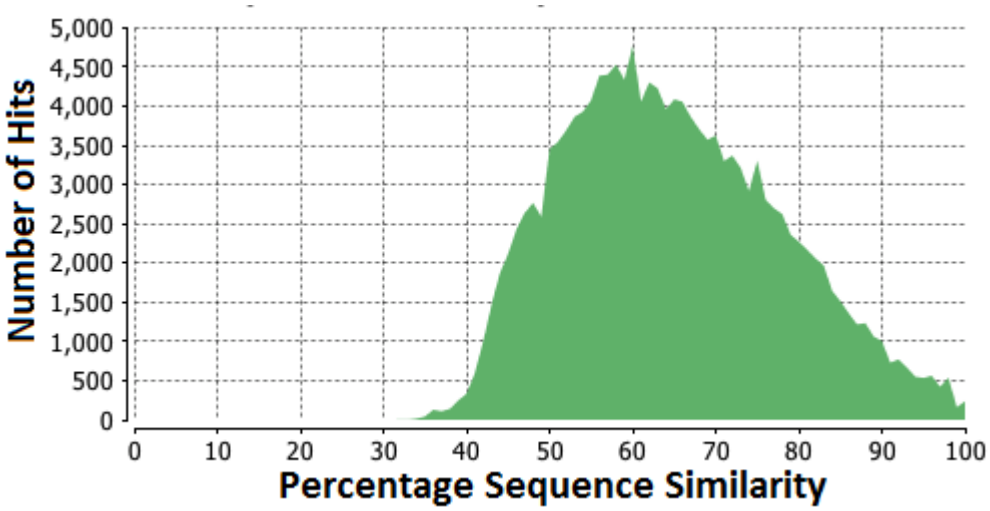


Figure 4.7: Sequence similarity distribution of the *P. superbus* transcriptome sequences to the sequences of top BLAST hits following BLASTN against the NCBI NR database with a cutoff of 1e-3.

Evidence codes are used to distinguish the method used by the GO curator to associate a GO term with a specific reference. However, while evidence codes reflect the type of experiment which supports the GO term to gene product association, they are not a classification of types of experiments/analyses (www.geneontology.org). Evidence codes should be used alongside GO terms. The evidence codes found in this dataset are as follows:

- Experimental Evidence Codes:

  IMP: Inferred from Mutant Phenotype,

  IDA: Inferred from Direct Assay,

  IPI: Inferred from Physical Interaction,

  EXP: Inferred from Experiment,

  IGI: Inferred from Genetic Interaction,

  IEP: Inferred from Expression Pattern,

- Computational Analysis Evidence Codes:

  ISO: Inferred from Sequence Orthology,

  ISS: Inferred from Sequence or Structural Similarity,

  ISA: Inferred from Sequence Alignment,

  RCA: inferred from Reviewed Computational Analysis,

- Author Statement Evidence Codes:

  TAS: Traceable Author Statement,

  NAS: Non-traceable Author Statement,

- Curator Statement Evidence Codes:

  ND: No biological Data available,

  IC: Inferred by Curator,

- Automatically-assigned Evidence Codes:

  IEA: Inferred from Electronic Annotation,

Experimental evidence codes indicate that a physical characterisation of the gene or gene product was carried out and subsequently published. Computational analysis evidence codes indicate that the experimentation was done *in silico* on the gene sequences and then published. Author Statement Evidence Codes indicate that the annotation was done based on a piece of text in a published reference. Curator Statement Evidence Codes indicate that a GO curator made the judgement when it did not fit into any of the other evidence code categories. Automatically-assigned Evidence Codes indicate that annotation has been given based on sequence similarity, database records and keyword mapping files. This is the only evidence code where a curator is not involved. This, for example, will be used when a dataset has not yet been published. As can be seen in Figure 4.8, the

majority of evidence codes in this dataset fell into the "inferred from electronic annotation" category.

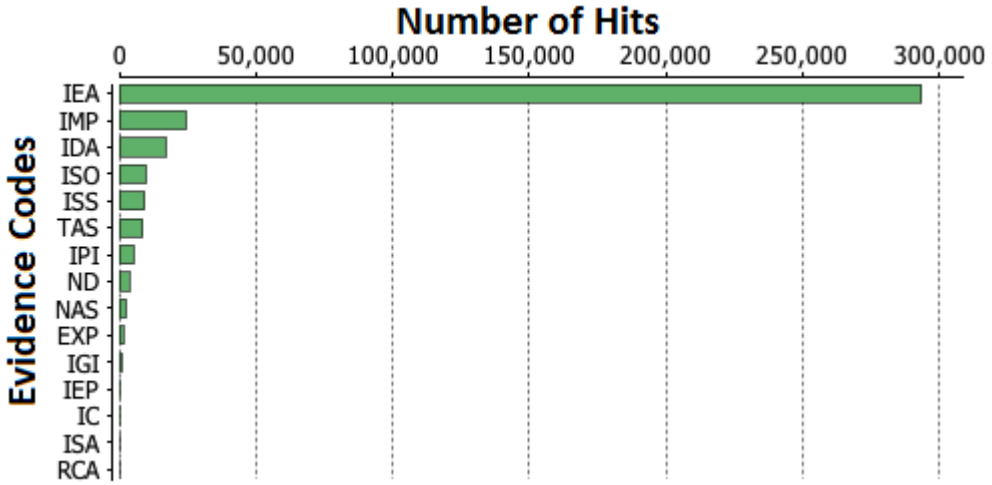


Figure 4.8: Evidence code distribution for the *P. superbus* transcriptome BLAST hits.

Annotations came predominantly from the UniProt Knowledgebase (UniProtKB) which is a collection of accurate and consistent functional annotation of proteins. This can be seen in Figure 4.9.

Figure 4.9: Databases accessed to find annotations for *P. superbus* transcriptome.

4.3.3 InterProScan

More than half the sequences (8,000) had an InterProScan hit, while just over 3,000 sequences were annotated using GO terms. Each InterProScan hit is given a category to help identify what the annotation infers (www.ebi.ac.uk/interpro). These categories are:

- Family:

  The sequences belong to a group of proteins with a common evolutionary origin. This is identified by their functions, similarities in sequence, or similarity in primary, secondary or tertiary structure.

- Domain:

  Domains are distinct functional, structural or sequence units that may exist in a variety of biological contexts.

- Repeat:

  A short sequence that is typically repeated within a protein.

- Site:

  A short sequence that contains one or more conserved residues. These could include active sites, binding sites, post-translational modification sites and conserved sites.

An overview of the breakdown of InterProScan results found can be seen in Figure 4.10.

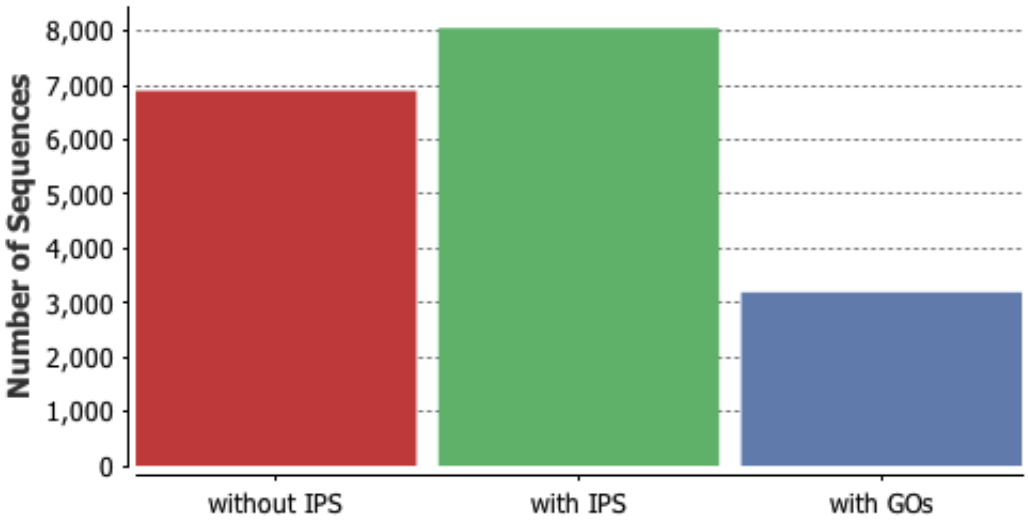


Figure 4.10: Summary of the InterProScan hits for the *P. superbus* transcriptome.

4.3.4 Gene Ontology Annotations

GO is comprised of three vocabularies used in the annotation of genes and gene products (www.geneontology.org). These vocabularies are:

- Molecular Function:

  Molecular function describes activities that occur at a molecular level. Terms in this vocabulary include activities rather than entities and molecules and complexes that carry out these activities. No location for these activities is given. These activities are usually performed by individual gene products, but some may be in complexes.

- Biological Process:

  Biological processes are those in which the molecular functions are involved. To be part of this vocabulary, a process must have more than one distinct step or series of events, but it is not the same as a pathway.

- Cellular component:

Cellular component is the area or place that the molecular function acts in. It may be some part of a larger cellular compartment or a gene product group.

Figure 4.11 shows the distribution GO terms in the annotations. Just over 9,000 sequences had no annotation, while there were 1 - 5 annotations for at least 2,500 sequences. As the number of annotations found increases, the number of sequences decreases.

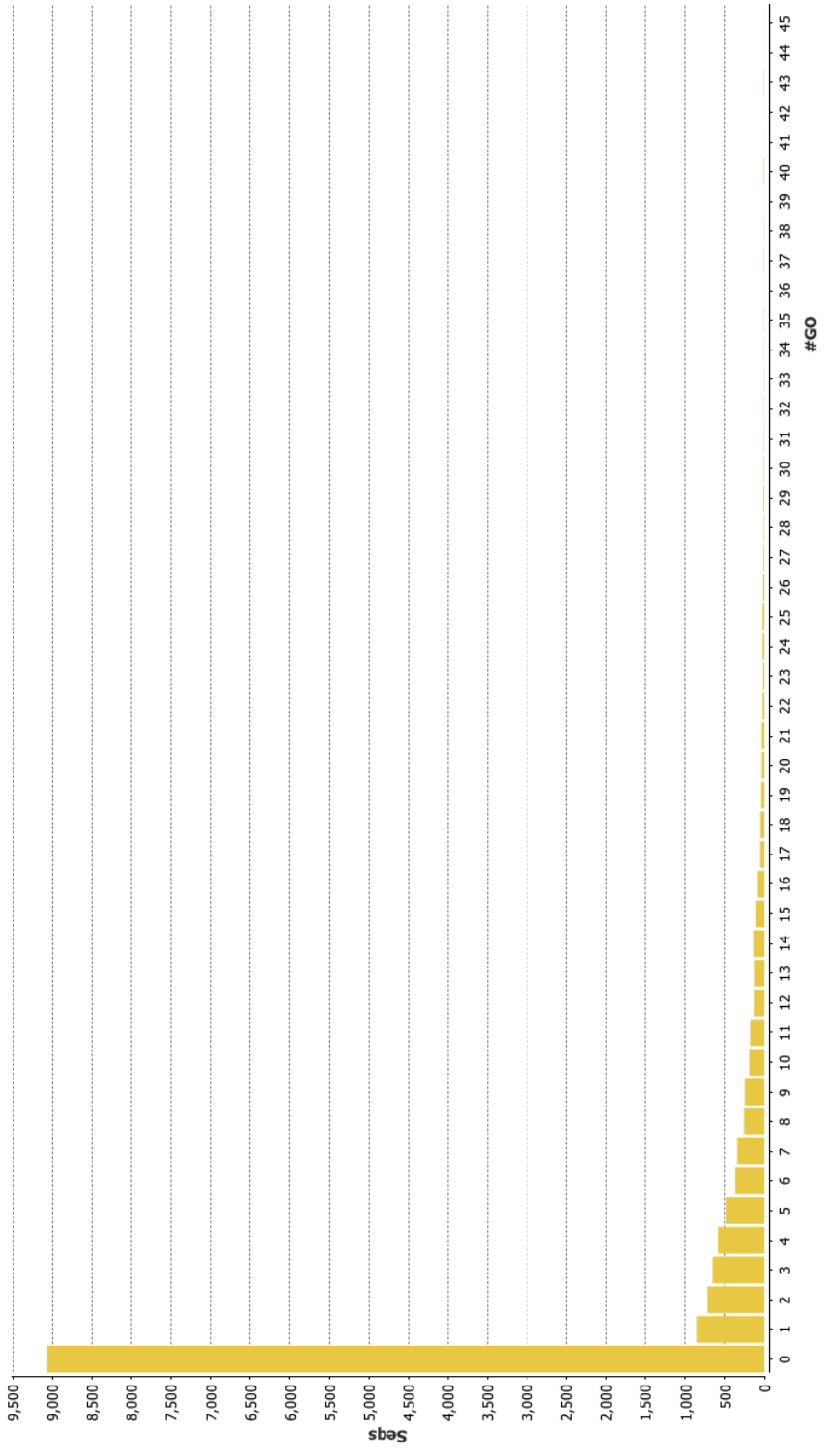


Figure 4.11: The GO annotation distribution for the *P. superbus* transcriptome showing the number of annotations found versus the count of sequences for that number.

Annotation is most likely found for sequences of between 2,200-3,200bp, as shown in Figure 4.12. Longer sequences, in the 5,000bp plus range, were also substantially annotated. Sequences from 150-2,200bps became increasingly more frequently annotated as length increased, suggesting longer sequences (or those closer to full length genes) are easier to annotate. The top 30 longest contigs were annotated with both GO and InterProScan. There is a dip in annotation success between lengths 4,000-5,400bps as there were no sequences of that length in the dataset.

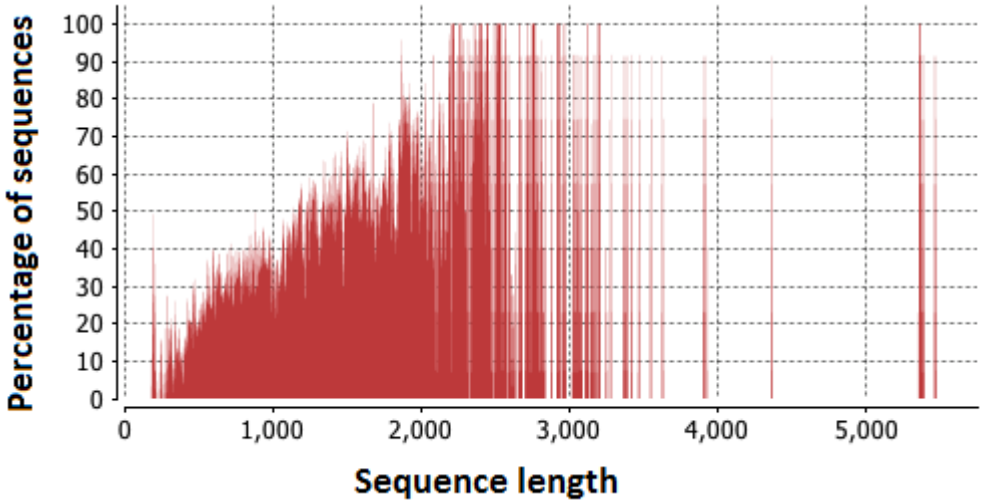


Figure 4.12: Percentage of *P. superbus* sequences versus sequence length that were annotated using GO.

Combined graphs can be generated showing the far-reaching scope for annotation. Different levels of annotation can be identified from the graph, and these levels of annotation, as well as their abundancy of terms for that level, are recorded in Figure 4.13. GO follows the true path rule, meaning annotation at a term implies annotation to all its parents' terms. GO has a Directed Acyclic Graph (DAG) structure where some categories have more than one parent category (Li et al., 2005).

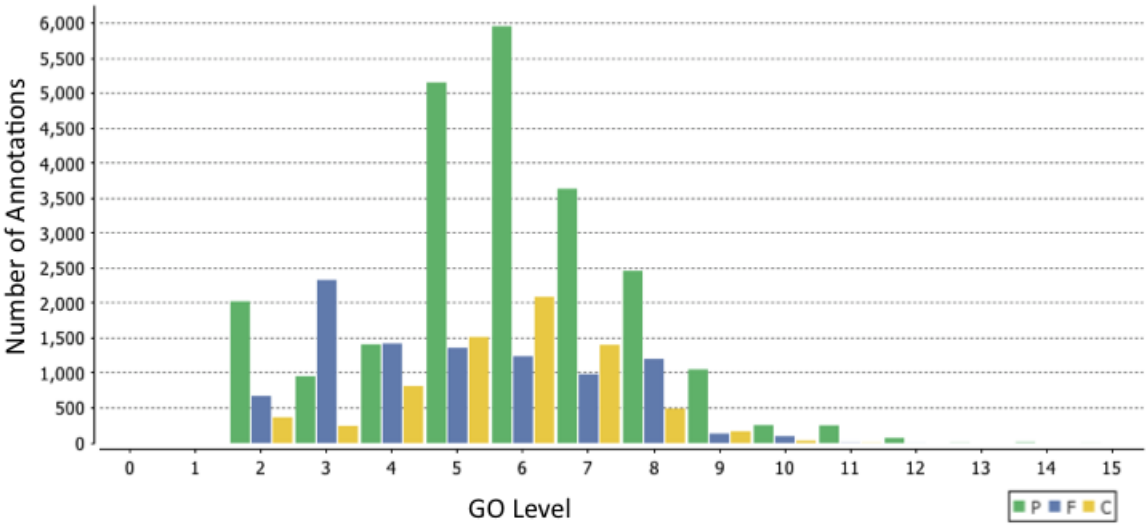


Figure 4.13: *P. superbus* transcriptome dataset distribution showing the abundance of terms for each vocabulary at each level.

Molecular Function

Molecular functions describe the action of a gene at a molecular level. Each gene may have one or more functions. The combined graph for molecular function shows that in this dataset there are 9 levels of annotation (see Figure 4.14).

For the molecular function vocabulary, the majority of the annotations fall in level 3. These are summarised in Figure 4.15. Protein binding, organic cyclic compound binding and hydrolase activity make up just over 50% of the GO terms associated with this level of molecular function annotation. The top 10 GO categories for overall molecular function are shown in Figure 4.16.

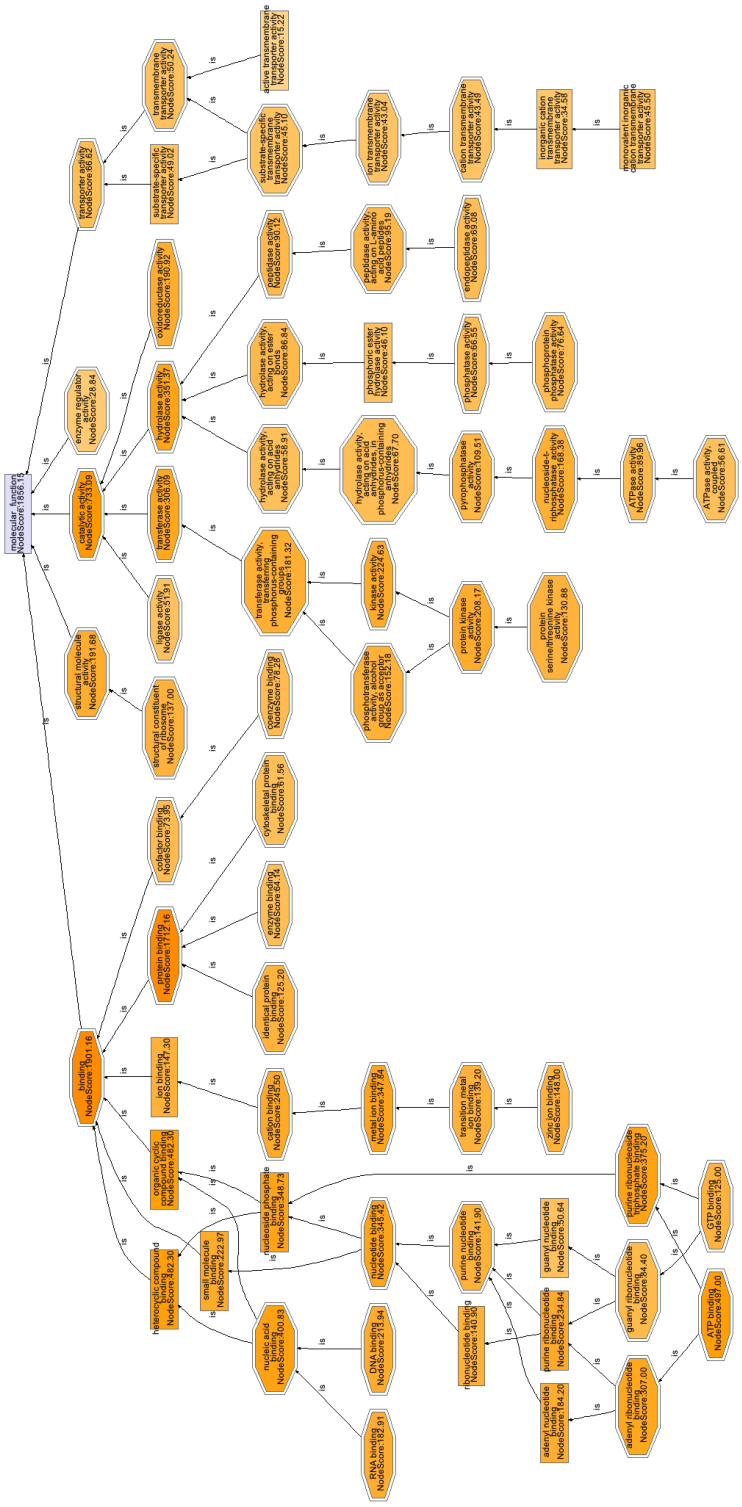


Figure 4.14: Overview of the GO molecular function annotation for the *P. superbus* transcriptome dataset. (See Figure on attached CD with filename of Molecular_function_graph_transcriptome.png).

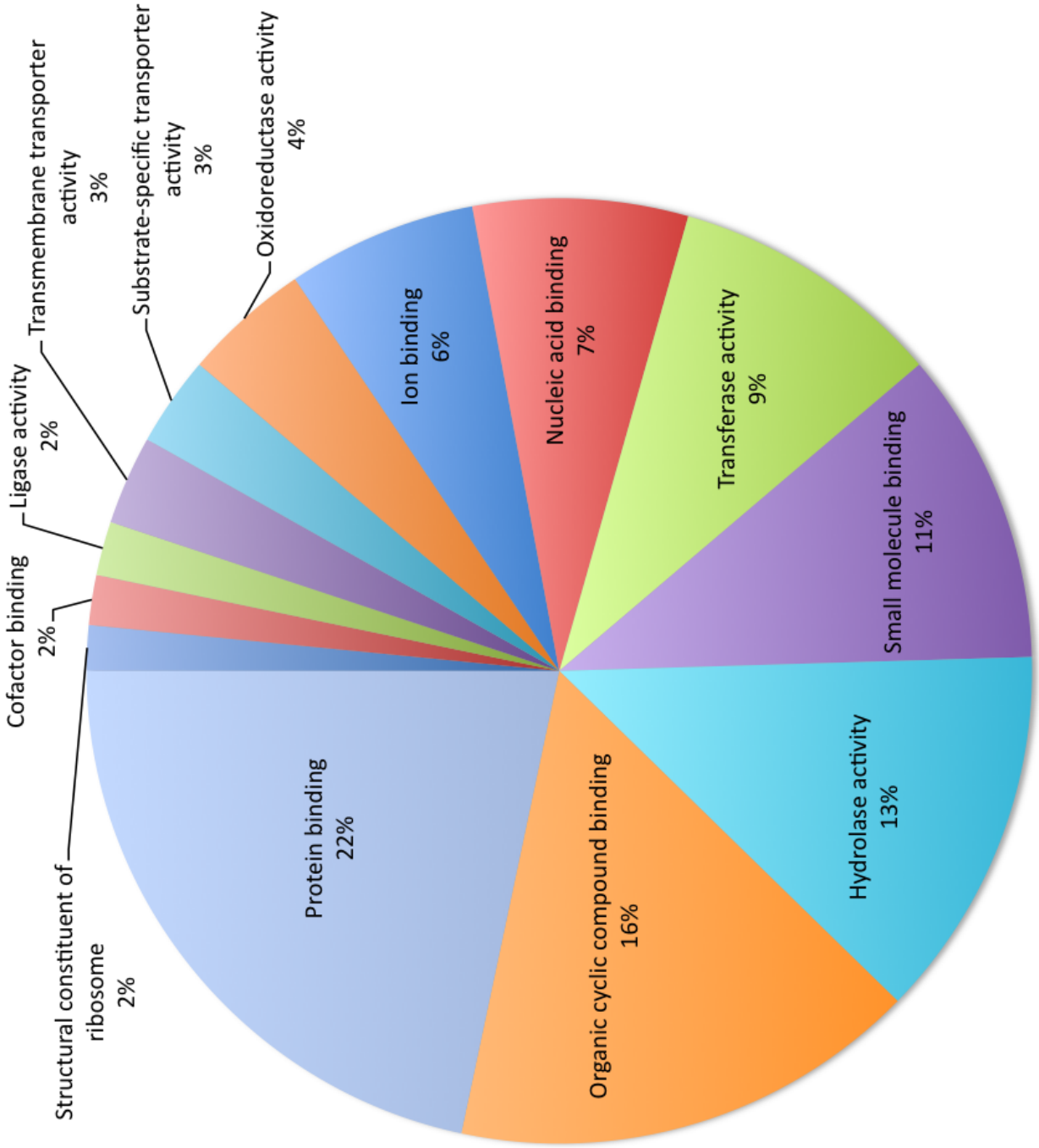


Figure 4.15: Molecular function categories identified for the *P. superbus* transcriptome dataset at level 3.

Cellular Component

The combined graph for cellular components found in this dataset shows that there are 15 levels of annotation, as shown in Figure 4.17. Level 6 in the cellular component category has most annotations, and these are summarised in Figure 4.18. Intracellular non-membrane and membrane bound organelles make up 39% of the annotations found at this level. The top 10 GO categories overall for cellular component are shown in Figure 4.19.

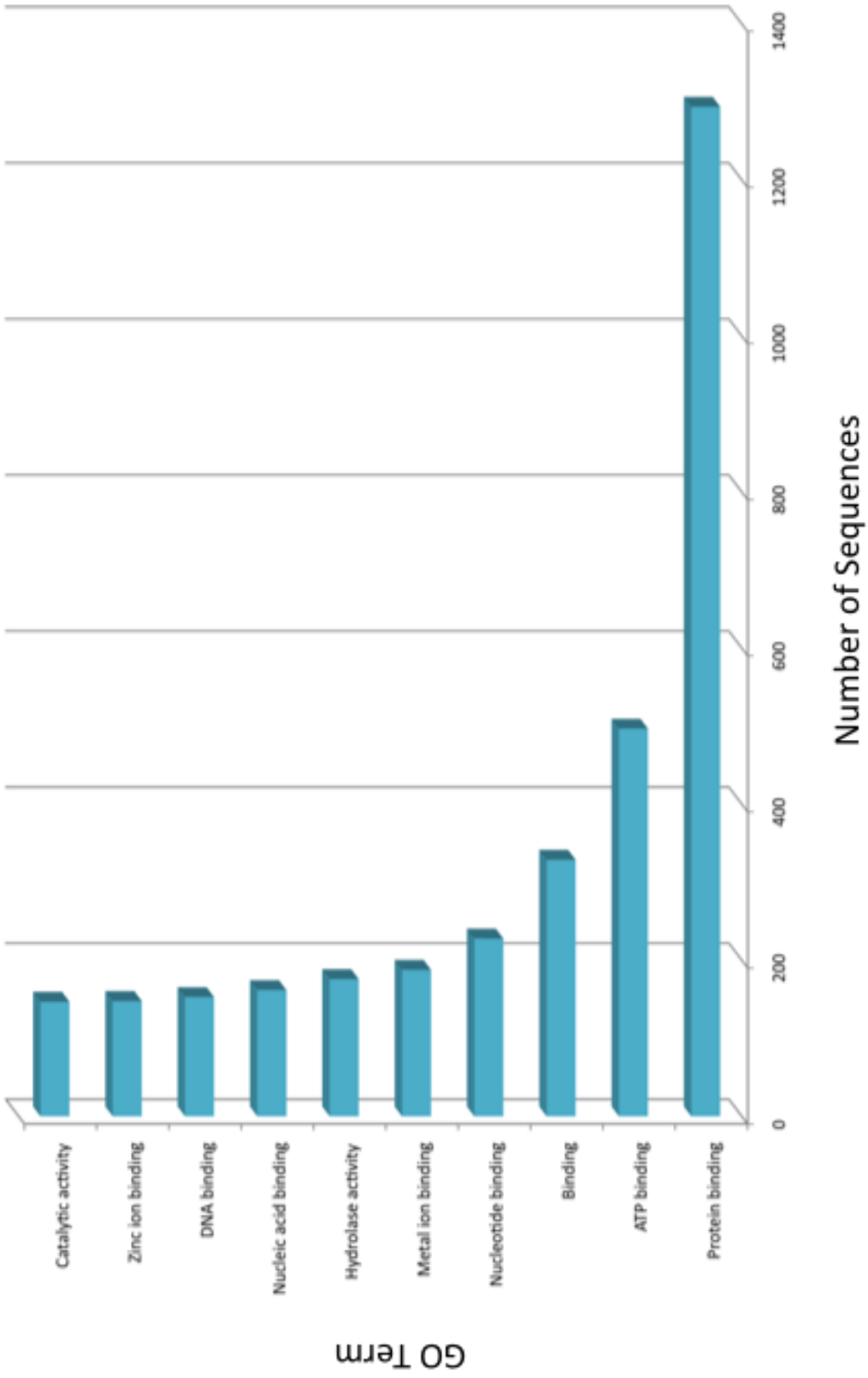


Figure 4.16: The 10 most common molecular functions identified in the GO analysis of the *P. superbus* transcriptome.

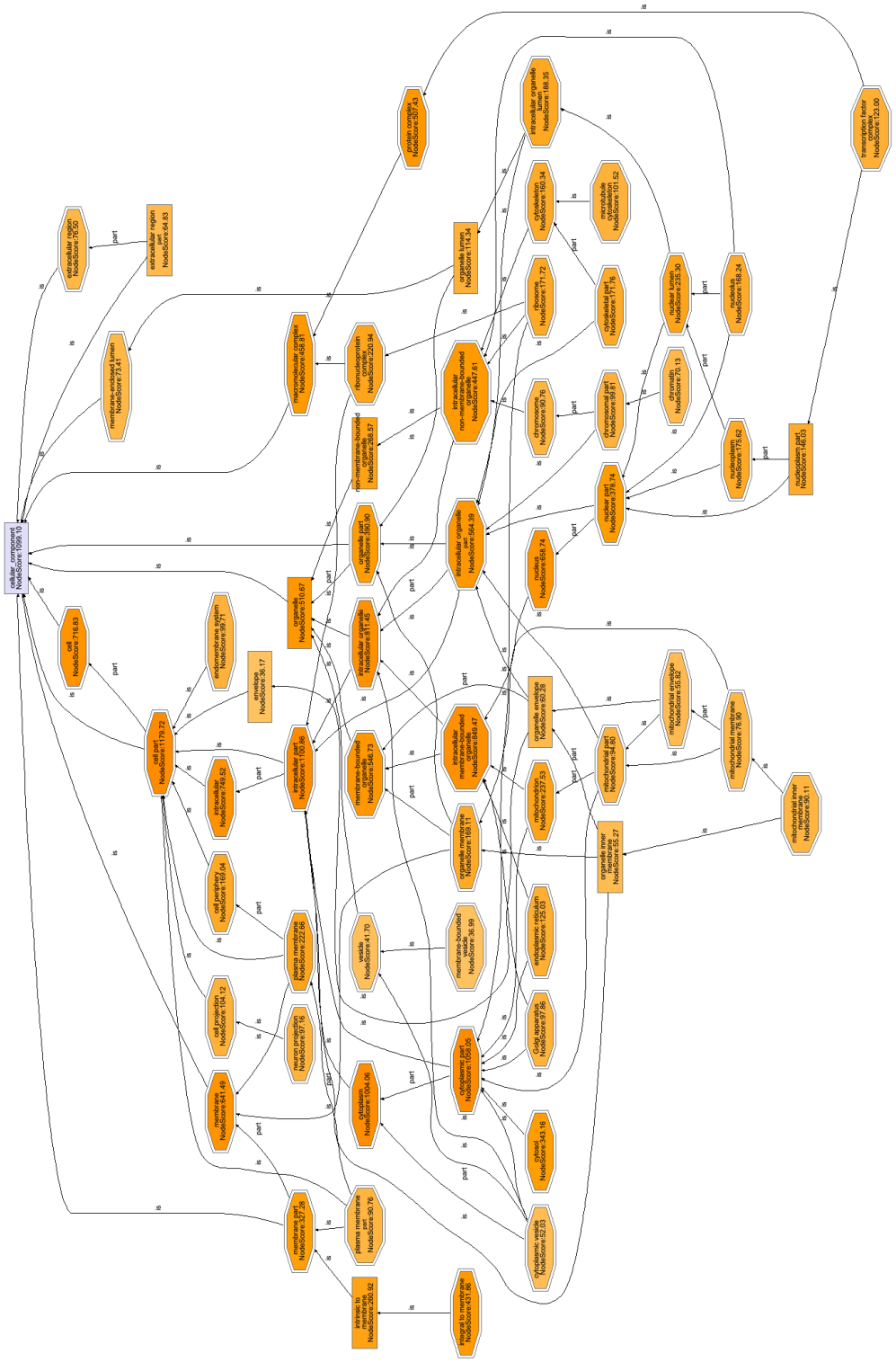


Figure 4.17: Overview of the cellular component annotations for the *P. superbus* transcriptome dataset in a combined graph. (See figure on attached CD with a filename of Cellular_component_graph_transcriptome.png).

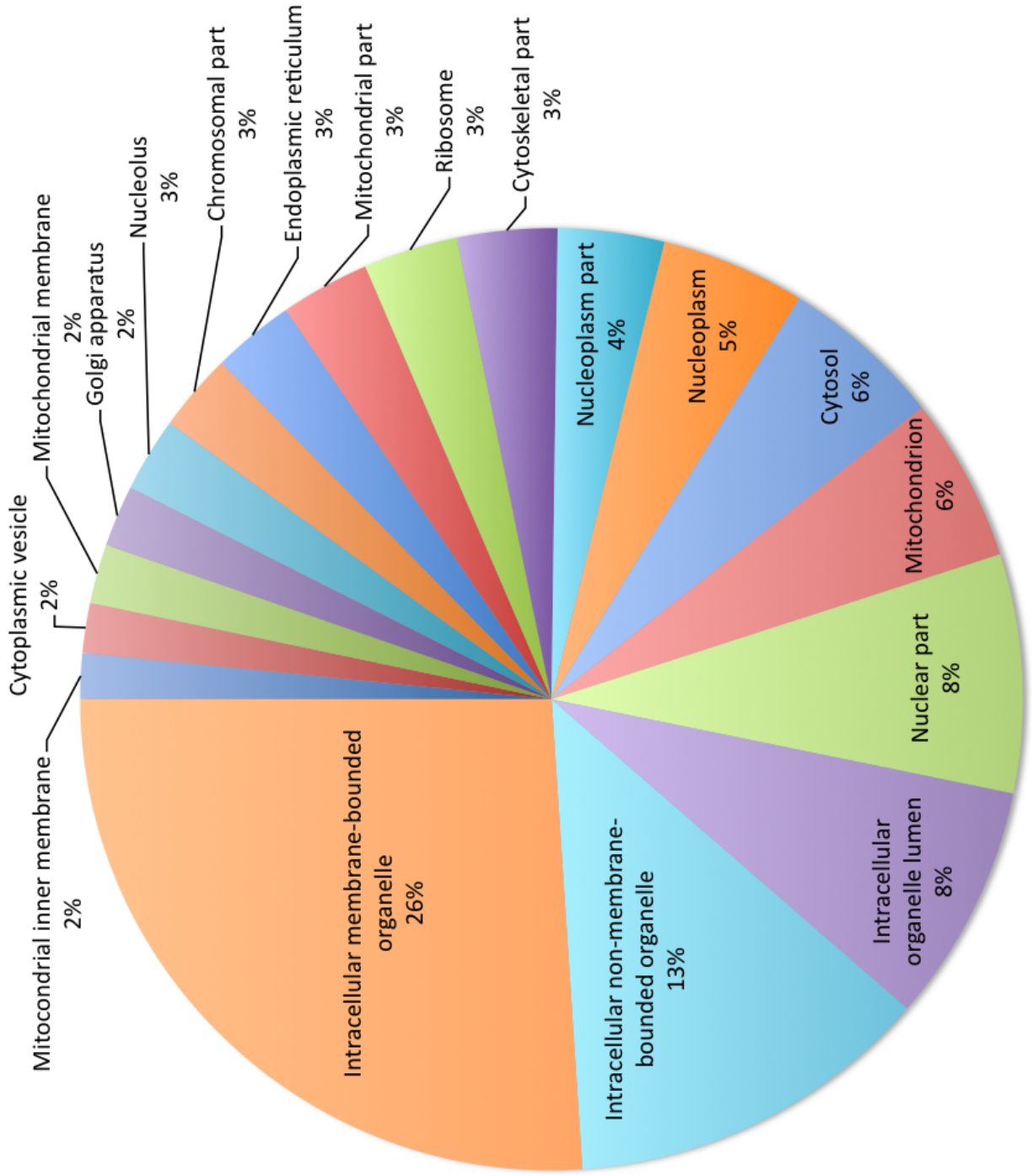


Figure 4.18: Cellular component categories identified at level 6 for the *P. superbus* transcriptome.

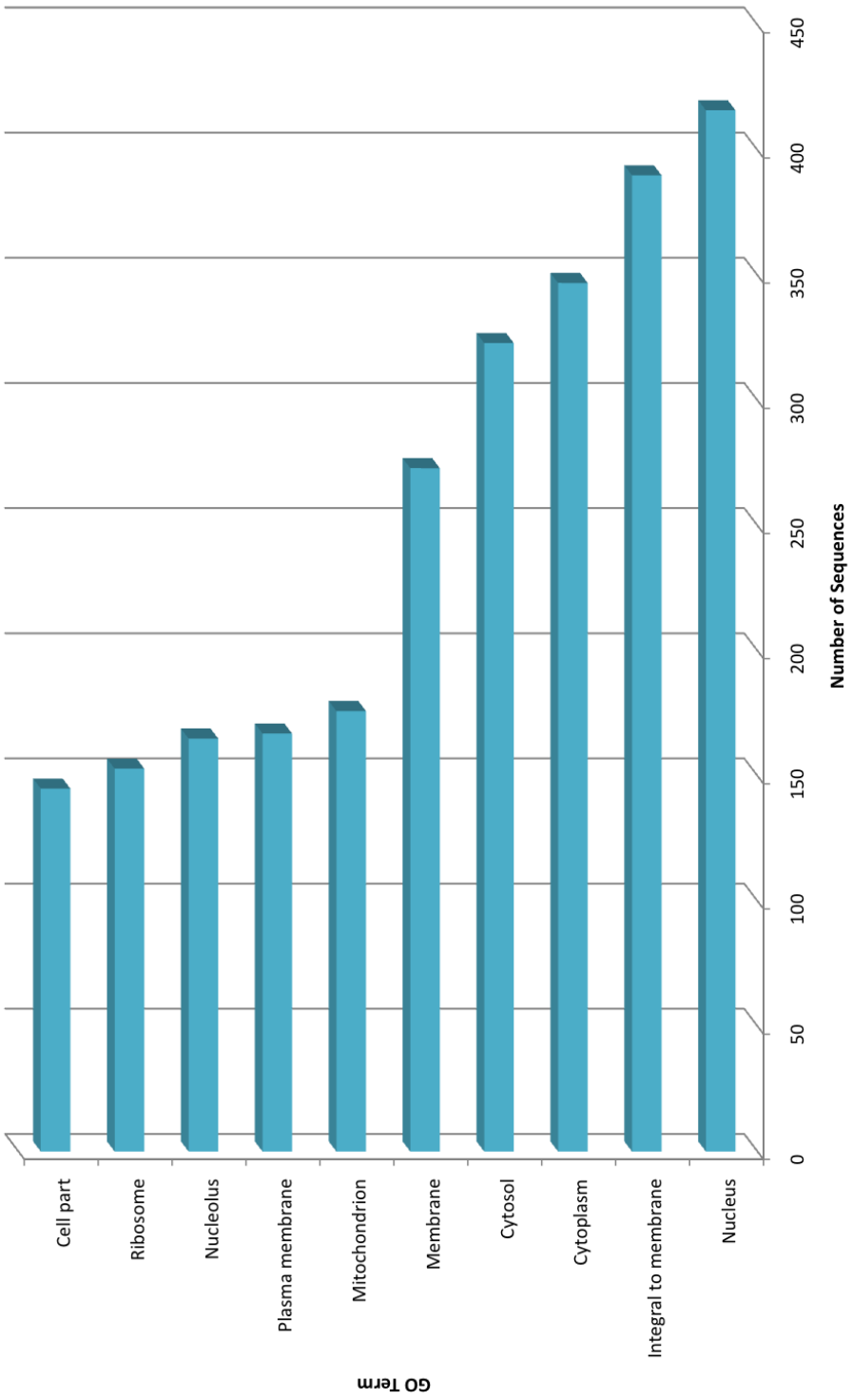


Figure 4.19: The 10 most common cellular components identified in the GO analysis of the *P. superbus* transcriptome.

Biological Process

The combined graph for Biological Process is extensive and shows how integral the DAG structure is to the algorithm. There are 17 levels of annotation, as shown in Figure 4.20.

The top 10 GO categories for biological processes are identified in Figure 4.21.

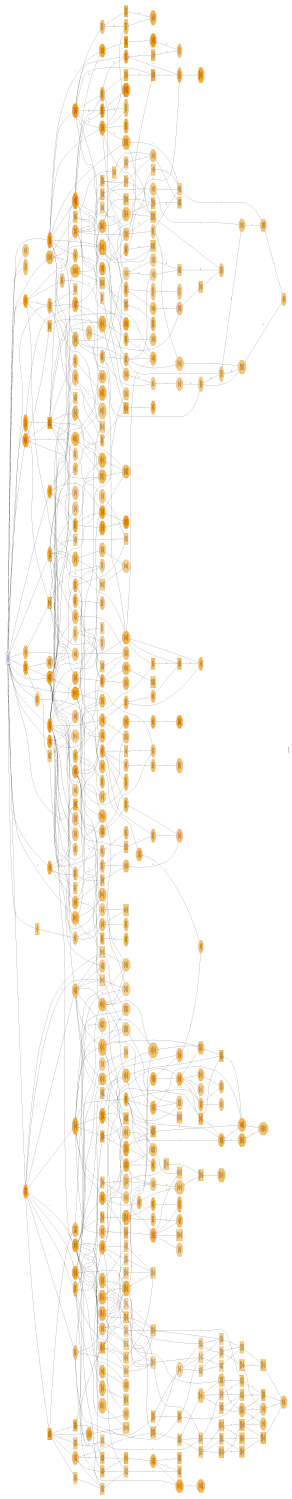


Figure 4.20: Overview of the Gene ontology biological processes annotations for the *P. superbus* transcriptome. (See figure on attached CD with filename biological_processes_combined_graph_transcriptome.png).

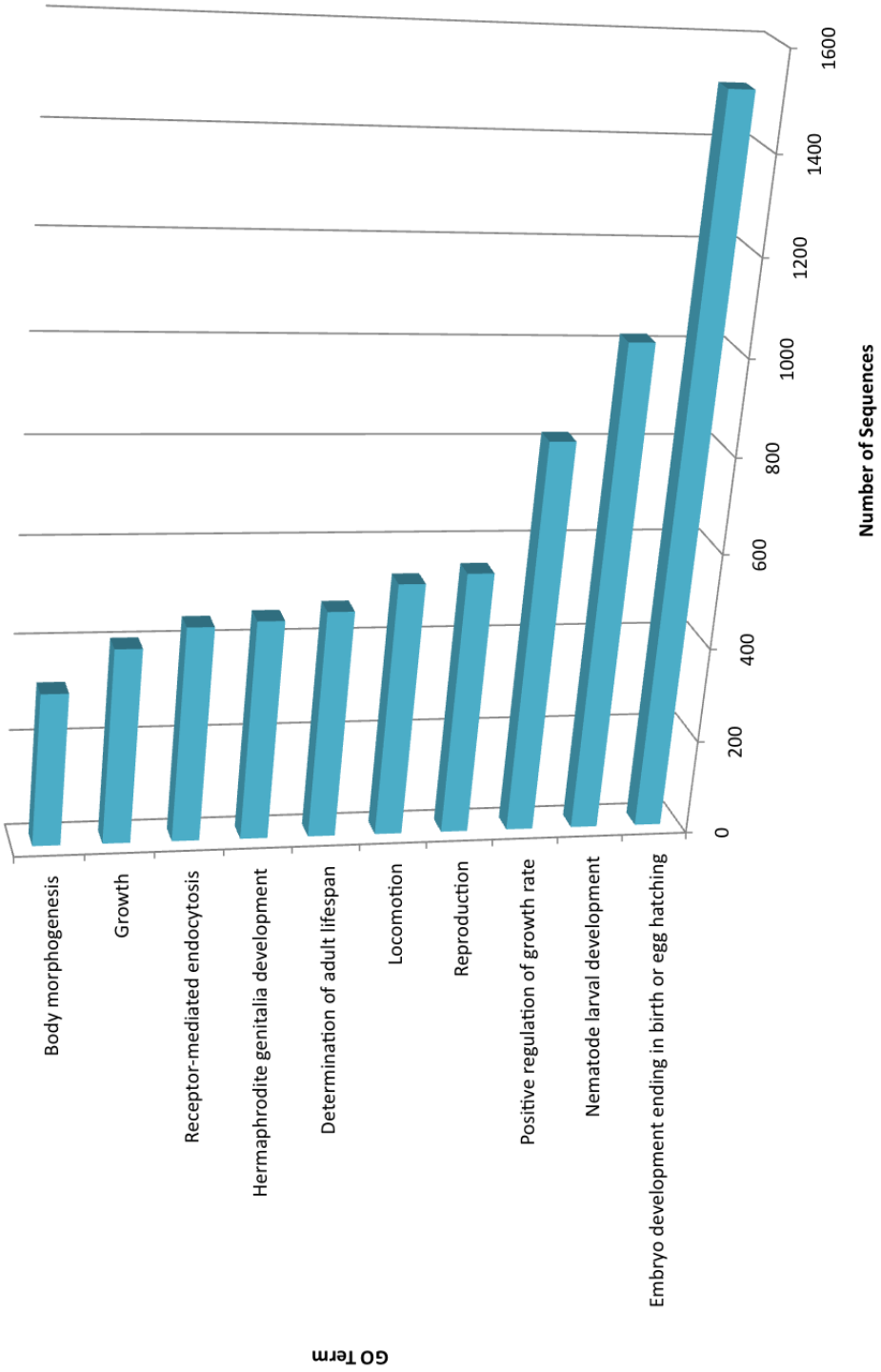


Figure 4.21: The 10 most common biological processes identified in the GO analysis of the *P. superbus* transcriptome.

4.3.5 EC and KEGG Annotations

Enzyme Commission (EC) numbers are based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) (Bairoch, 2000). Each enzyme is described using four figures. The first figure represents one of six classes that the enzyme belongs to: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases. The second figure indicates the subclass, the third figure indicates the sub-subclass, and the fourth the serial number of the enzyme in its sub-subclass. Part of the BLAST2GO pipeline allows for a matching of the sequences to EC numbers to further the annotation. This is beneficial, as on occasion some *P. superbus* sequences were not identified through BLAST but did result in an EC hit.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) is a collection of pathways representing all of the major interactions and reactions in molecules. BLAST2GO uses the EC numbers identified during the EC step of the pipeline to map against KEGG, and identify which components of various pathways are located in the dataset. KEGG maps are used in conjunction with EC numbers to return which sequence IDs correspond to that pathway. Those found are identified and highlighted in colour in the various pathways. Showing the full scope of this analysis is beyond a hard copy version of the thesis, so all data collected are located on the accompanying CD. To give an overview of the analysis, the two pathways whose components are most represented in the dataset are discussed.

The pathway showing purine metabolism is shown in Figure 4.22. Purines, including Adenine and Guanine are essential six sided rings which are an integral part of both DNA and RNA. Purine nucleotides give humans their energy and repair cell membranes (Zollner, 1982). The purine metabolism pathways includes 32 enzymes, which are present in this dataset. These are represented in colour

in Figure 4.22. These enzymes are represented by 141 sequences in the dataset. In this pathway, most *P. superbus* sequences matched EC number 3.6.1.3, which is Adenosine triphosphatase. This functions in the conversion of ATP to ADP in metabolism. When ATP is broken down to ADP by Adenosine triphosphatase, energy is released (Mader, 2001).

The oxidative phosphorylation pathway is shown in Figure 4.23. Oxidative phosphorylation refers to the generation of ATP due to energy release during the electron transport chain. ATP production is vital for the cell as it is later converted to ADP which, as previously described, gives cells their energy. In this pathway, 8 of the enzymes involved are represented in the dataset by 53 sequences, the most common being EC number 3.6.3.6 or H+− exporting ATPase.

GO:0006950 response to stress is described as "Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a disturbance in organismal or cellular homeostasis, usually, but not necessarily, exogenous (e.g., temperature, humidity, ionising radiation)" (www.geneontology.org). This is a high level biological process term with multiple child terms such as GO:0006979 (response to oxidative stress), GO:0009409 (response to cold), GO:0009414 (response to water deprivation) and GO:0009408 (response to heat). In this dataset, 465 sequences were identified as having a hit to GO:0006950 or one of its child terms. This is just over 5% of all annotated sequences and is subsequently more than the 187 unigenes identified in Chapter 2 in the *P. superbus* EST dataset (Table 2.4). Functional analysis resulted in just over 250 sequences with unique descriptors. These could be genes induced by the stress the nematodes were subjected to prior to sequencing. Among these were: MAP-kinases; members of the jumonji family of transcription activators; antioxidant enzymes; molecular chaperones; components of the ubiquitin-proteasome system; DNA damage response

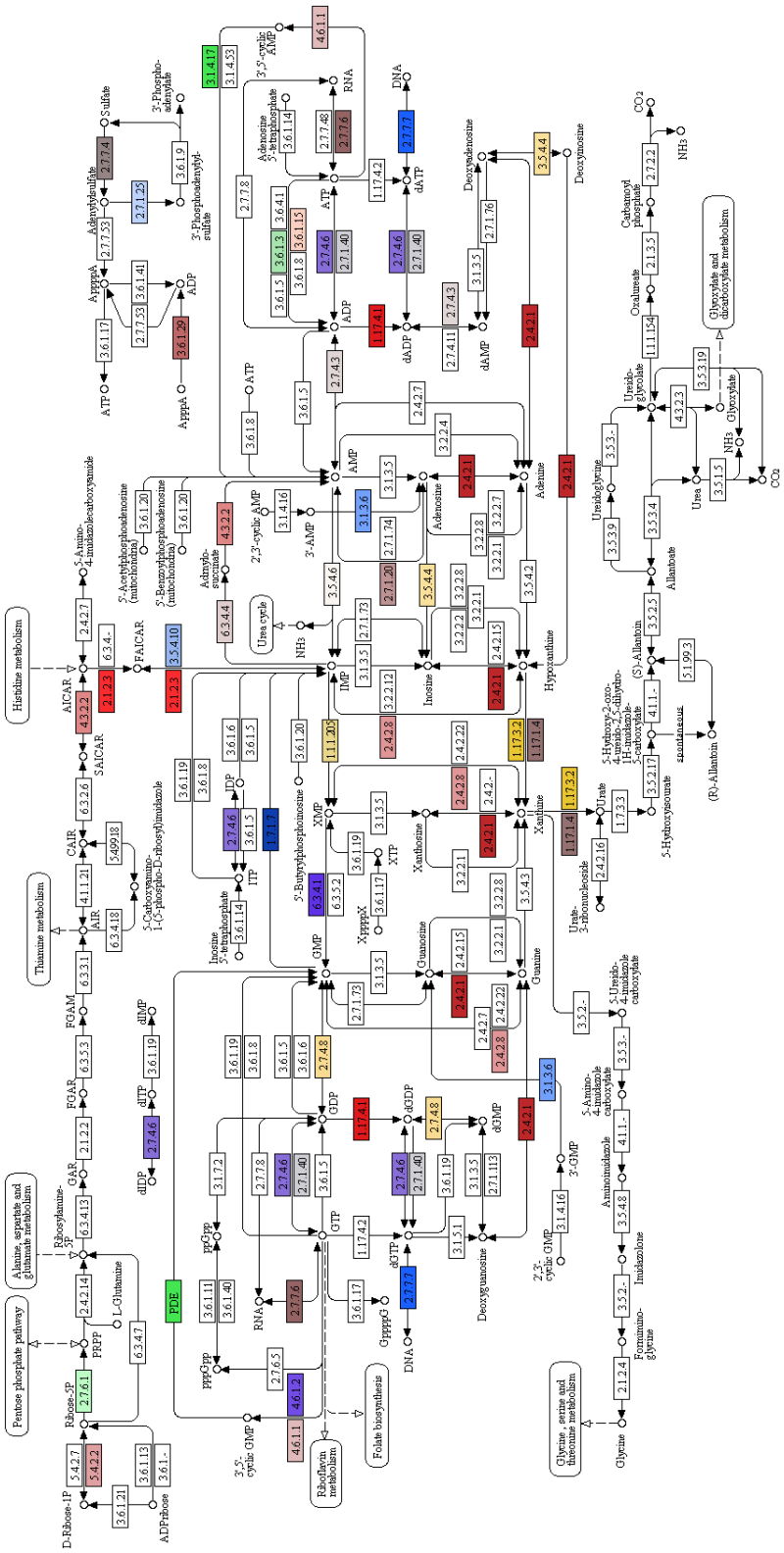


Figure 4.22: Purine metabolism pathway with EC numbers found in the annotation of the *P. superbus* dataset shown in colour.

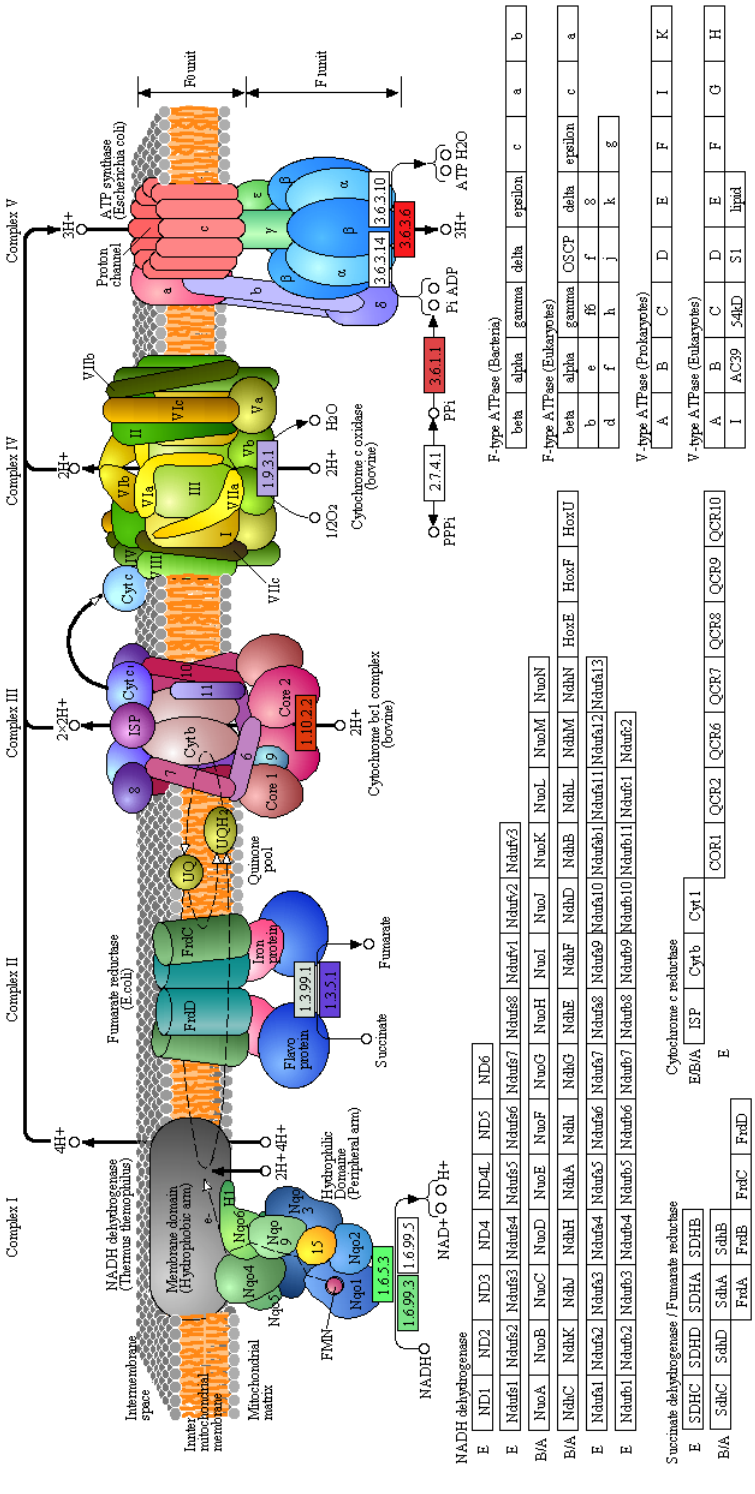


Figure 4.23: Oxidative Phosphorylation pathway with EC numbers found in the annotation of the *P. superbus* dataset shown in colour.

proteins and LEA proteins. Some of these are summarised in Table 4.1.

Table 4.1: Stress response genes found in the transcriptome of *P. superbus*

Description	**Number of Sequences**
Ubiquitin family protein	17
DNA Damage Response Proteins	15
HSP70 family	11
Serine/threonine protein kinases	10
Glutathione peroxidase	7
HSP90 family	7
Proteasome subunit alpha family	5
NADP Isocitrate dehydrogenase	4
Transcription factors/activators	4
Ubiquitin-conjugating enzyme E2	4
Aldehyde dehydrogenase	3
Mitogen-activated protein kinases	3
Protein tyrosine kinases	3
Thioredoxin	3
Derlin-1	2
Casein kinases	1
HSP60 family	1
Late Embryogenesis Abundant Proteins (LEA)	1
Mitochondrial prohibitin complex protein 1	1
Protein disulfide isomerase	1
Small heat shock protein/α-crystallin family	1

4.4 Discussion

Annotation is an extensive task involving many factors. First, a comprehensive search for the best assembler must be performed to ensure the dataset is as accurate and representative as it can be. In this dataset the Newbler assembler with the following parameters was used: Version 2.6 without URT Isotigs. At 14,900 contigs, it is believed that this assembly allows for a conservative estimate of gene numbers without the doubt of over-assembly and thus false positives. This assembly was chosen given various metrics identified in Chapter 3. Transcriptome sequencing, using next generation techniques, can be employed to address questions that were difficult when using previous techniques. Transcriptome sequencing allows a glimpse of expressed genes during a particular point in the life cycle of an organism. In this study, an investigation of genes expressed during stress are presented. The following sections describe some of the identified sequences found in the *P. superbus* transcriptome believed to be involved in stress.

4.4.1 Heat Shock Proteins

Heat shock proteins (HSPs) are some of the most highly conserved genes in existence. HSPs can be induced by heat as well as other stresses (Lindquist, 1988). They are proteins essential for the correct folding and maturation of a great diversity of client proteins, and for protecting proteins from stress-induced unfolding and aggregation (Morimoto, 2008; Richter et al., 2010). Eukaryotic HSP families contain multiple genes, which may be either constitutively expressed or stress-inducible and targeted to specific cellular compartments (Kabani & Martineau, 2008; Vos et al., 2008). The HSP expression repertoire of an anhydrobiotic organism may thus be very important in maintaining the integrity of the proteome during the dehydration and recovery phases of anhydrobiosis (Sales et al., 2000;

Jonsson & Schill, 2007; Cho & Choi, 2009; Hu et al., 2003). The generated dataset contains all the heat shock protein classes including HSP20, which acts as a chaperone protecting other proteins against heat induced denaturation and aggregation (Groenen et al., 1994) and HSP70, which aids in oxidative stress by preventing partial proteins from becoming aggregated and thus non-functional (Tavaria et al., 1996).

4.4.2 Removal of Damaged Proteins - The Ubiquitin-proteasome (UPS) and Autophagy Systems

When the HSP chaperone system fails to correctly fold a denatured protein, the misfolded protein is polyubiquitinated. The 26S proteasome, a large multiprotein complex, then translocates polyubiquitinated proteins into the inner proteolytic chamber where they are hydrolysed (Kubota, 2009). If the generation of misfolded proteins exceeds the proteolytic capacity of the ubiquitin proteasomal system (UPS), misfolded proteins accumulate into aggregates which are degraded by autophagy (Salomons et al., 2009; Lamark & Johansen, 2010). A whole genome RNA interference (RNAi) screen in *C. elegans* identified 40 genes that are essential for survival during acute hypertonic stress (Choe & Strange, 2008). Half of these genes encode proteins that function to detect, transport, and degrade damaged proteins. Ubiquitin is well represented in this dataset with 21 sequences identified with this description.

4.4.3 DNA Damage Response Proteins

DNA extracted from anhydrobiotic stages of the plant parasitic nematode *D. dipsaci* was intact, showing no increase in the frequency of double-strand DNA breaks (DSBs) as compared with hydrated worms (Barrett & Butterworth, 1985). Data

from the anhydrobiotic chironomid *P. vanderplanki* (Gusev et al., 2010) and anhydrobiotic tardigrades (Neumann et al., 2009; Rebecchi et al., 2009) show that DSBs accumulate with time in the dry state in these organisms. DSBs also accumulate during desiccation in the anhydrobiotic and radiation resistant bacterium *Deinococcus radiodurans*. Similar to *P. vanderplanki* (Gusev et al., 2010) and anhydrobiotic tardigrades (Neumann et al., 2009; Rebecchi et al., 2009), *D. radiodurans* has acquired the ability to rapidly repair DNA damage when rehydrated (Mattimore & Battista, 1996). DNA repair and DNA polymerase are represented in the dataset by 25 sequences.

4.4.4 Signal Transduction, Protein Kinases and Transcription Factors

Transduction of environmental stress signals is achieved in eukaryotes through a conserved cascade of sequentially acting stress activated protein kinases (SAPKs) which form a branch of the mitogen-activated kinase (MAP-kinase) system (Jonak et al., 1996; de Nadal & Alepuz, 2002; Cuenda & Eousseau, 2007; Whitmarsh, 2010). In *S. cerevisiae* the SAPK pathway is activated by osmostress and the terminal kinase Hog1 (Brewster et al., 1993), when phosphorylated, translocates to the nucleus (Ferrigno et al., 1998). Here it phosphorylates several transcription factors, and associates at stress-responsive promoters through such transcription factors (Ferrigno et al., 1998), resulting in the expression of osmotic response genes. These were also represented in the dataset.

4.4.5 Other Putative Anhydrobiotic Genes

Other transcripts whose products may play a role in the anhydrobiotic response of *P. superbus* include two putative aquaporins; an ERM (ezrin, radixin, and moesin)

family member; an an1-like Zinc finger sequence; a thaumatin-like transcript; two copies of *lon-1* which encodes a protease that selectively degrades oxidised mitochondrial proteins (Ngo & Davies, 2009) and a homolog of the *Ric1* family (van West et al., 1999), which encodes plasma membrane proteins that are expressed in response to high salt or low temperature conditions in plants (Navarre & Goffeau, 2000). ERM proteins are activated by osmotic shrinkage (Rasmussen et al., 2008) and they are thought to function as cross-linkers between plasma membranes and actin-based cytoskeletons (Sato et al., 1992). Thaumatin-like proteins are induced in plants in response to pathogens, cold, drought and osmotic stress (Liu et al., 2010). The an1-like multigene family is involved in plant abiotic stress responses and in inflammation responses in mammals (Jin et al., 2007).

The analysis described was carried out on a *P. superbus* transcriptome assembly derived from a combination of an unnormalised (PS1) and a normalised (PS2) cDNA library and a Sanger sequenced EST dataset. It would be very informative to map the individual sequencing reads from the PS1 and PS2 libraries onto the combined assembly as this would give an insight into the stress-response genes of the *P. superbus*. The cDNA libraries were derived from pooled cDNA's extracted from control nematodes and nematodes which have been exposed to one of the following stresses: heat, cold oxidation or desiccation.

Chapter 5

Assembling the Nuclear Genome of *P. superbus*

5.1 Introduction

Genome annotation is the process by which we extract meaningful information from sequences of genomic DNA (gDNA). This can be done using laboratory or computational techniques. The vast advances being made in the field of generation of genomic data have meant that when it comes to their bioinformatic analysis it is a challenging task; however, the number of tools available for this task are rapidly increasing in number. The suggested pipelines and exploration avenues available are shown in Figure 5.1.

The generation of good quality DNA in a large enough quantity poses the first challenge in genomic sequencing. Once the gDNA has been extracted, the next potential stumbling block is the generation of a library at the sequencing centre, but, by far, one of the greatest challenge lies in the assembly step of the pipeline. Genomes can be difficult to assemble because they contain biological anomalies such as heterozygosity and repeat regions. It is also difficult to be accurate with

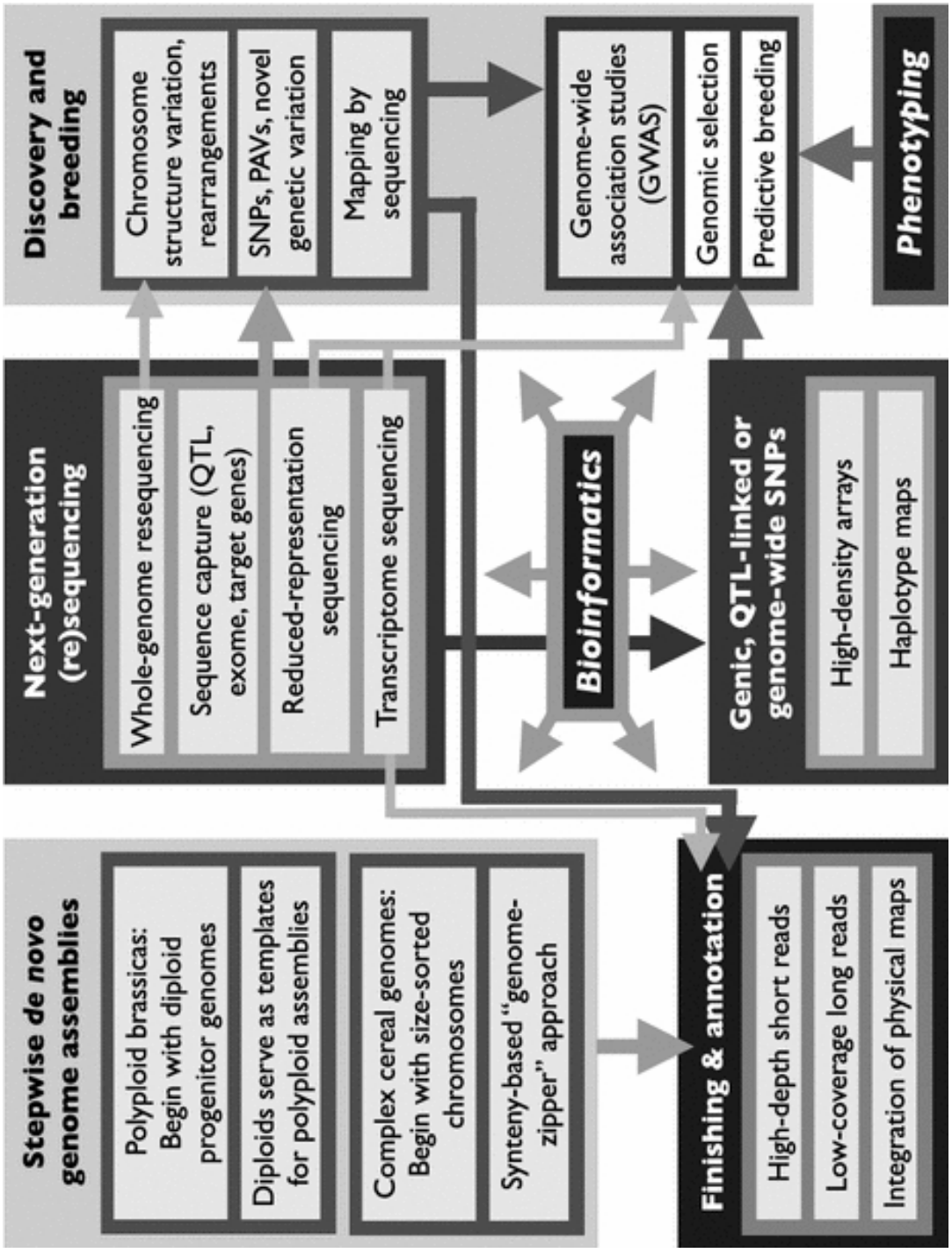


Figure 5.1: Suggested exploration avenues in next generation genomic sequencing (Edwards et al., 2013).

the assembly when it is a *de novo* assembly of a genome without a closely related reference available. Three main focus points should be considered when trying to create a good assembly. These are:

- Coverage: a large oversampling of the genome needs to be carried out to ensure there are long overlaps between the reads and that all parts of the genome are represented. From a mathematical point of view, having low coverage is not practical for a genome annotation project;
- Read length: read length and mate pairs must be longer than the repeat regions to avoid having false overlaps, the longer the reads the easier to assemble contigs;
- Quality: The better the quality score of each base the more accurate the assembly will be (Schatz et al., 2012).

Assembly difficulties can include incomplete coverage due to gaps of unknown size between contigs. Sequencing errors increase in proportion to the length of the read generated and the reads are sequenced in an unknown orientation so these may need to be reverse-complemented in order to establish the correct reading frame. All of these difficulties add to the challenge of generating a draft genome, let alone a completely sequenced genome. To try to minimise these challenges as much as possible, a hybrid approach is often employed. In this research project two sequencing platforms were used: Roche 454 Titanium and Solexa Illumina. In addition, the Solexa Illumina platform was used three times to maximise the technology available at that time. A work flow diagram of the 454 platform can be seen in Figure 5.2, and a workflow diagram of the Solexa Illumina platform can be seen in Figure 5.3.

Once the genome assembly has been established to a satisfactory level, gene finding can be done in several ways. Sequence similarity to known genes from

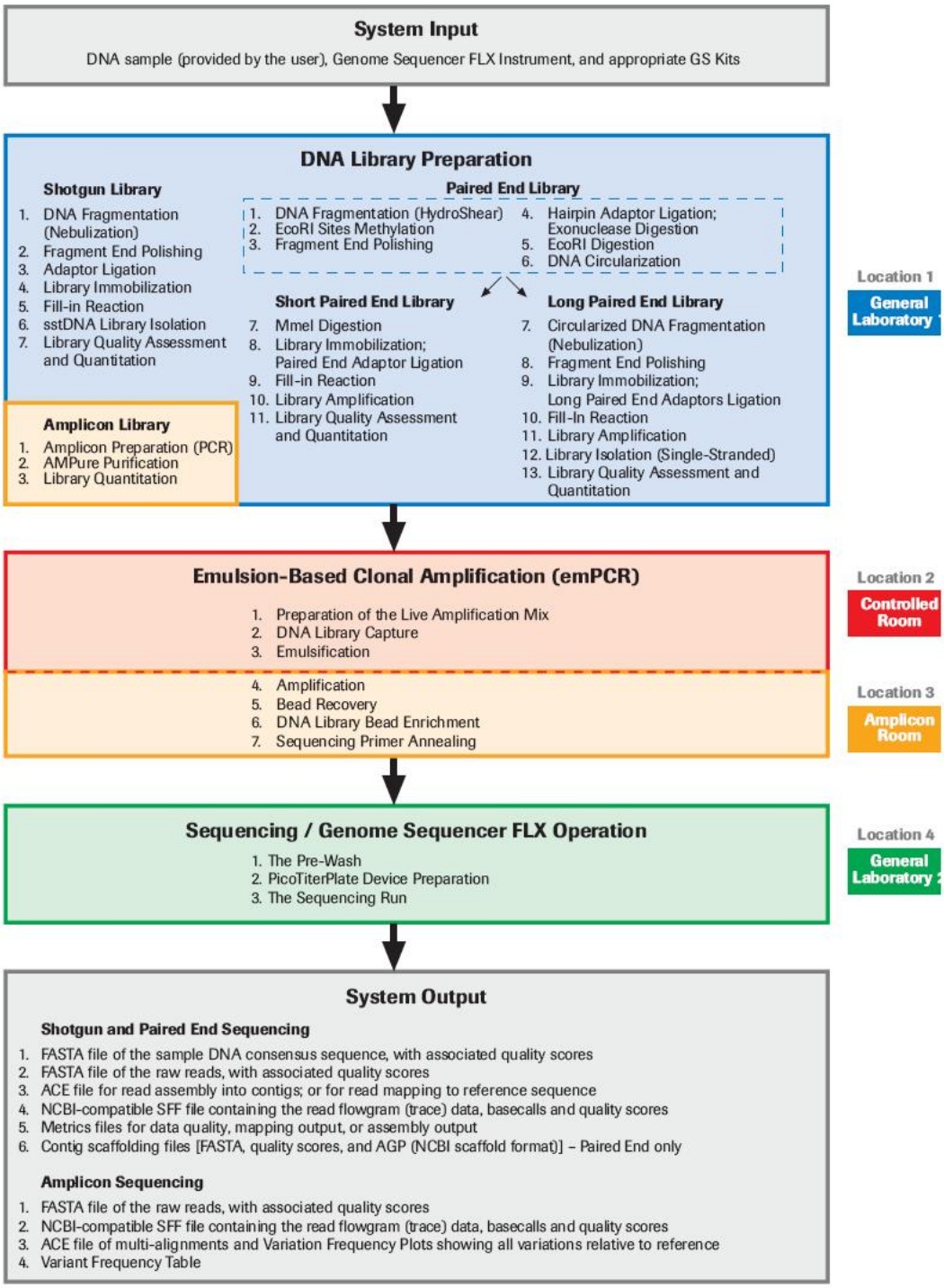


Figure 5.2: 454 Titanium workflow for genomic sequencing focusing on crop genomes (DNA Sequencing Core, 2013).

Figure 5.3: Solexa Illumina workflow for genomic sequencing at the University of Michigan (*www.seqcore.brcf.med.umich.edu/*) (Illumina, 2013).

other organisms is one technique that can be used. This is known as an extrinsic approach. More commonly, this would integrate some form of the BLAST algorithm. The method allows identification of approximately 50% of the genes in the genome as shown in the EST dataset discussed in Chapter 3. The use of ESTs in this way can be useful for the identification of exons. Transcriptome data can be used in this way to identify exons and alternative splicing. Comparison of a genome with itself can identify repeat regions, transposons and retrotransposons as well as abundantly expressed gene families.

Alternatively, protein prediction could be used to identify potential protein coding regions. This *ab initio* approach can be used to identify these potential gene regions. Codon usage frequency can be used as synonymous codons are not utilised equally. Hexamer frequency can be used in the same way and is perhaps a more sensitive way of measuring codon frequency. Another method that could be used is a sliding window of the GC%. GC count in a gene will differ from that outside of a gene. Two main algorithms are used in gene finding. These are:

- Statistical-based methods:
    1. Hidden Markov Models,
    2. Neural Networks,
    3. Integration of various different approaches.
- Homology-based methods:
    1. Local Alignment methods,
    2. Pattern-based Alignments.

There are difficulties with each method. Statistical-based methods tend to over-predict and so, lead to false positives. They also cannot predict alternative splicing. Homology-based methods are time-consuming and require report parsing.

A pattern must also be chosen and maintained, which is challenging. Gene finders can run into difficulties if the exon regions are small and the introns are big. They can fuse neighbouring genes or split genes. They can miss genes or overlap genes. Pseudogenes, alternative splicing and non-canonical splice sites can also cause difficulty.

Subsequently, annotation of the genome begins. Annotation involves identifying the function of a gene given its sequence. Biological questions such as:

- When and where is this gene expressed?
- What genes does this gene interact with?
- Does it contain functional domains?

can then be asked. These questions and the one previously discussed are extensive and generally require a genome sequence analysis workbench to assist in answering them (The Blaxter Lab, 2013).

There are three categories of genome annotation. These are:

- Nucleotide level Annotation:

  Mapping to known genes, genetic markers, tRNAs, rRNAs, repeat elements, ancient duplications, ORFs. etc.,

  Gene Finding,

  Non coding RNAs and regulatory regions,

  Identifying repetitive elements,

  Mapping segmental duplications,

  Mapping variations.

- Protein level Annotation:

  Define the category of proteins, name them and assign a function,

Classify on the basis of domains, folds and motifs.

- Process level Annotation:

  Molecular function,

  Biological process,

  Cellular component.

In this chapter extraction of gDNA and sequencing of various samples is discussed. Assembly and initial annotation is reported and future work is identified.

5.2 Materials & Methods

5.2.1 Chemicals

Chemicals were obtained from Sigma-Aldrich Co. Ltd. (Gillingham, UK), Thermo Fisher Scientific Ltd. (Massachusetts, US), Novagen Division of Merck/EMD (Wisconsin, US), Invitrogen Ltd. (Paisley, UK), Promega UK Ltd. (Southampton, UK), Fermentas (Maryland, US), and Pierce Division of Thermo Fisher Scientific Ltd. (Cramlington, UK). Enzymes were purchased from New England Biolabs (NEB) (Beverly, MA, US) or Roche (Clarecastle, IRE), Promega or Novagen. Oligonucleotide primers were ordered from Eurofins MWG Operon (Edersberg, DE). Sterile plasticware was purchased from Sartorius AG (Goettingen, DE).

5.2.2 Nematode Collection and Care

P. superbus was first isolated in 1970 on Surtsey Island, Iceland from a gulls nest as described in Chapter 1. *P. superbus* DF5050 was used in this study. Nematode growth medium (NGM) plates were prepared by dissolving 3g Sodium Chloride, 2.5g peptone, 17g agar and made up to 1L with distilled water. The solution was autoclaved, and once cooled, 1ml cholesterol in ethanol (5mg/ml), 25ml 1M potassium phosphate buffer pH6, 1ml 1M Calcium Chloride, 1ml 1M magnesium sulphate was added to the solution and supplemented with the antibiotic Streptomycin to a final concentration of 30μg/ml. The media was poured into 9cm plates and left to set. Upon setting, 40μL of *E. coli* HB101 culture was spread on the plates and left to grow overnight at 37°C. A 1cm^2 piece of an established NGM nematode culture was transferred using a sterile scalpel onto the new plates. The plates were cultured in the dark at 20°C until the nematodes reached a large mixed population ($\sim$14 days depending on strain). These methods were developed for the model nematode *C. elegans* (Brenner, 1974). To confirm the *Panagrolaimus*

strain identity, the rDNA D3 expansion region was PCR amplified and cloned into the TOPO pCR 2.1 vector and transformed into *E. coli* TOP10 cells. The DNA sequences obtained for the cloned DNA fragment were identical to the *P. superbus* rDNA D3 sequence in GenBank (Accession No. AY878376.1) thus confirming the strain.

Nematodes were washed off the NGM plates using sterile water and left to gently agitate on a shaker for 20 minutes in order to digest any bacteria present in the nematode gut. The liquid was poured off the plates into sterile 50ml Falcon tubes and left to settle at 4°C for 20 minutes. The supernatant was removed and sterile water was added. This process was repeated a total of three times in order to obtain bacteria free nematodes. The supernatant was removed and the resulting pellet was snap-frozen in liquid nitrogen prior to storage at -80°C. The nematodes for gDNA sequencing were re-suspended in Nematode Lysis buffer (20mM Tris pH 7.5, 50mM EDTA, 200mM NaCl, 0.5% SDS (w/v)) before freezing. For every 100μL of nematode pellet, 500μL of Lysis buffer was used.

5.2.3 gDNA Extractions

Nematodes were cultured and harvested as described previously. The tubes were defrosted at room temperature and ground into a fine powder (at least three times) using an autoclaved pestle and mortar and liquid nitrogen. The supernatant was then placed into a 1.5ml Eppendorf tube and proteinase K (10mg/ml in water) was added to a final concentration of 2mg/ml. The mixture was then incubated at 56° C for 1 hour 30 minutes (with inversion every 20 minutes) or until no nematode carcasses could be identified on examination under a microscope. The solution was cooled to room temperature, RNase A was added to a final concentration of 1.2mg/ml and the tube was incubated for 15 minutes. The solution was then extracted with 2.5 volume of phenol:chloroform:isoamyl alcohol (24:24:1) at room

temperature for 10 minutes. The tubes were spun in a pre-cooled microcentrifuge (Eppendorf) at 4° C at 12,000 x g for 10 minutes . The aqueous layer was transferred to a new tube and a second phenol:chloroform:isoamyl alcohol extraction was performed. The solution was spun as previously described and the supernatant transferred to a new tube. A final 10 minute chloroform:isoamyl alcohol (24:1) extraction was carried out to remove any residual phenol and the sample was centrifuged as previously described. The aqueous layer was then transferred to a new tube and 1/30th volume of 3M sodium acetate and 2.5 volume of ice-cold 100% ethanol were added. The solution was stored at -20° C for 1 hour. The DNA was pelleted by centrifugation at 12000 x g for 20 minutes at 4° C. The supernatant was removed and the pellet washed with 1ml of 70% Ethanol and spun for 5 minutes at 7,500 x g. The supernatant was removed and the pellet air-dried. The pellet was then resuspended in 30μL sterile water and the DNA concentration and integrity was determined by using a Qubit DNA assay kit with the Qubit 2.0 Fluorometer. DNA was also visually compared to standards of known concentration on ethidium bromide stained gels.

Electrophoresis was carried out using 0.7% agarose gels in 1X Tris Acetate EDTA buffer (TAE): 40mM Tris, 20mM acetic acid and 2mM EDTA, pH 8.1. Agarose powder was dissolved by heating in 1X TAE buffer. Upon cooling to 60°C ethidium bromide was added (10mg/ml). The solution was then poured into a casting tray and allowed to solidify. Samples were mixed with loading buffer (5mg/ml bromophenol blue, 5mg/ml xylene cyanol, 50% glycerol) at a ratio of 5:1. 5μL of 1kb bench top DNA ladders (Promega) was loaded on each gel. Gels were typically run at 100V using BioRad electrophoresis equipment. The ethidium bromide stained DNA gels were visualised under UV light using a UV transilluminator at 365nm. Gels were photographed using an AlphaDigiDoc gel documentation system (Alpha Innotech).

5.2.4 Karyotyping

Karyotyping using DAPI stain was performed on *C. elegans*, *P. davidi* and *P. superbus* to establish chromosomal number. The karyotypes were determined using unfertilised oocytes as these contain the haploid chromosome number. Firstly, young adult males and females were isolated from a six-day-old culture grown on NGM media. A $1cm^2$ agar section containing a high concentration of nematodes was transferred to a 3cm Petri dish containing M9 buffer: 6g Na_2HPO_4, 3g KH_2PO_4, 5g NaCl, 0.25g $MgSO_4$ and distilled water to a volume of 1 litre. Using an aspirator, six individual nematodes were removed from the buffer and placed in 6cm Petri dish 4', 6-diamidino-2-phenylindole (DAPI) to a final concentration of 2μg/ml (Catalog No. S4651). Using a 0.2mm insect needle mounted on a 10cm wooden stick, the nematodes were dissected below the pharynx to release the gonads. The gonads were transferred to a silane-prep microscope slide (Sigma catalog No. S4651) containing 40μL DAPI and left to incubate at room temperature for 15 minutes in a moisture chamber.

Embryos were also examined. NGM plates were flooded with M9 buffer and stirred gently for two minutes. The supernatant was transferred to a new 9cm Petri dish and stirred gently to concentrate the embryos in the middle. Using the aspirator, some eggs were removed to a clean 3cm Petri dish and washed with 40μL M9 buffer. These eggs were then removed into DAPI and left to incubate at room temperature for 15 minutes in a moisture chamber.

At this point, for both adults and embryo samples, a cover slip was then mounted on the sample and gently pressed to spread the cells to make the visualistion of the cellular contents easier. The slides were then examined using a confocal microscope with the assistance of Dr. Ilora Dix, Confocal Microscopy Unit, NUI Maynooth.

5.3 Results

5.3.1 Preparation of *P. superbus* gDNA for High-Throughput Sequencing

A large number of samples preparations were carried out. It took some time to establish the best procedure for procurement of the nematodes and extraction of gDNA. Various factors hindered this process including contamination of plates, which meant the culture had to be restarted from an egg stage, significantly slowing up the progress. The successful sample preparations sent for sequencing are outlined below.

The first sample preparation will be referred to as 454 gDNA. The gel electrophoresis images for this sample are shown in Figure 5.4 (pre RNAse step) and Figure 5.5 (post RNAse step). Samples 1, 2 and 9 were pooled (now referred to as sample 1). Samples 5, 6, 7 and 8 were pooled (now referred to as sample 2). Samples 3, 4 and 10 were pooled (now referred to as sample 3). These samples were re-run on an electrophoresis gel using ethidium bromide staining and are shown in Figure 5.6. Concentration of sample 1 was calculated to be 13μg/30μL. Concentration of sample 2 was calculated to be 13μg/12μL as shown in Figure 5.7. The purity of the samples was 1.78 for sample 1 and 1.58 for sample 2. The purity of the third sample was not deemed to be of a satisfactory standard for sequencing and was excluded. Samples 1 and 2 were sent to The Gene Pool, University of Edinburgh, Scotland, for sequencing using Roche 454 Titanium platform. 709.5Mb of data were generated representing an estimated 7.9X coverage of the *P. superbus* genome.

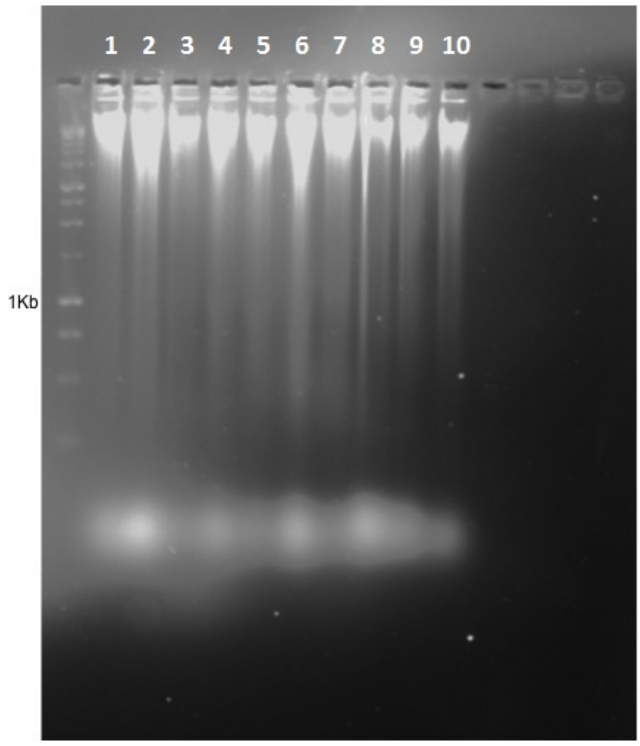


Figure 5.4: *P. superbus* gDNA sample for 454 sequencing (pre RNAse treatment).

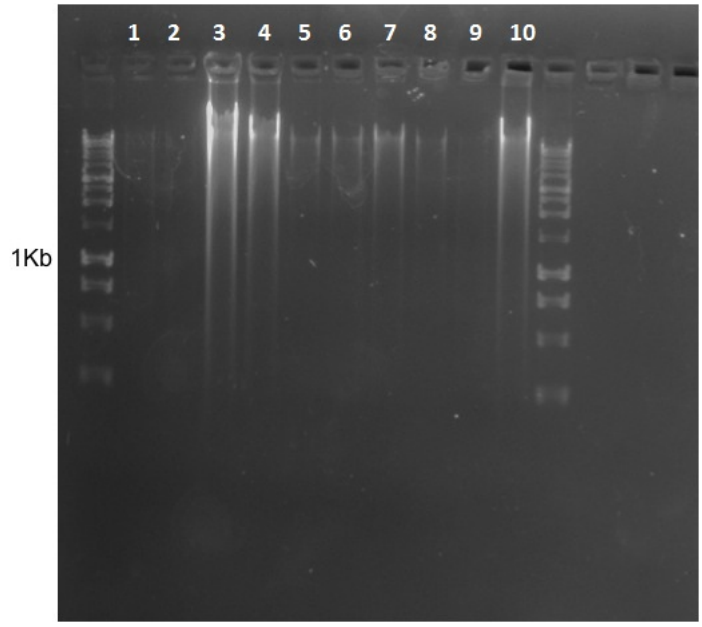


Figure 5.5: *P. superbus* gDNA sample for 454 sequencing (post RNAse treatment).

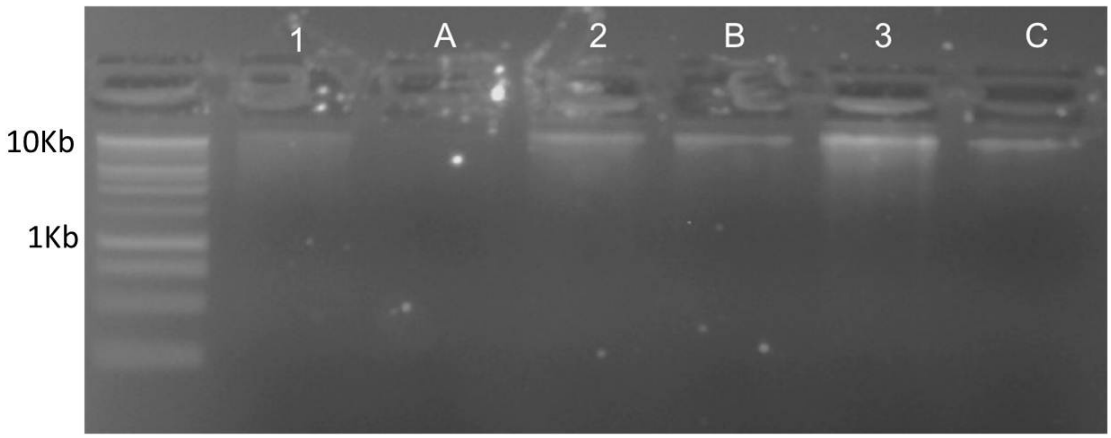


Figure 5.6: *P. superbus* gDNA sample for 454 sequencing (post pooling). Letters indicate a $\frac{1}{10}$ dilution of the previously labelled sample.

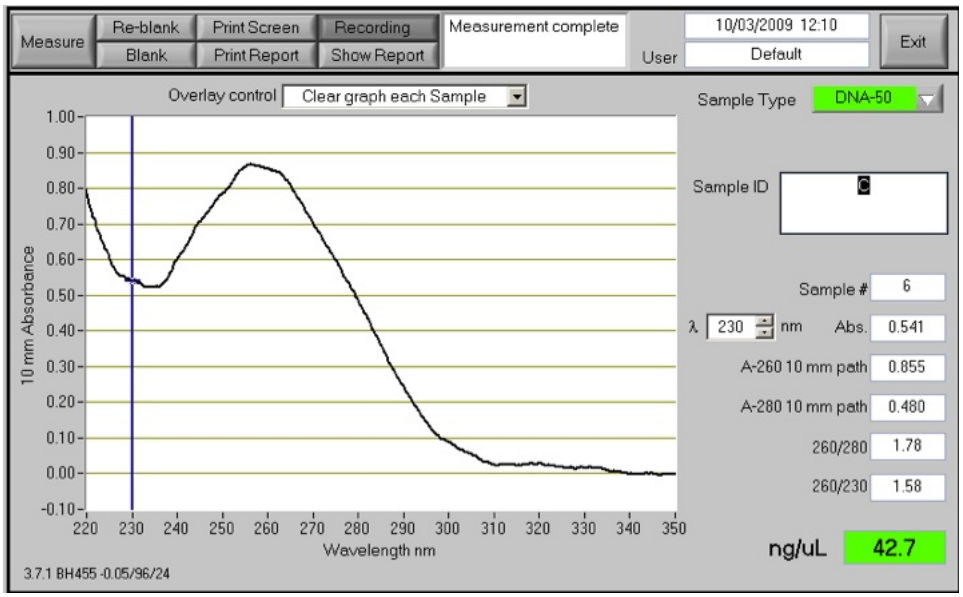


Figure 5.7: An absorbance spectrum for a *P. superbus* gDNA sample obtained using a Qubit 2.0 Fluorometer, indicating the A260/A280 ratios used to assess DNA quality.

A second sample preparation was carried out (Figure 5.8). This will be referred to as 50bp Illumina. Samples 2, 6, 7, 8, 9, 12 were pooled and became sample 1 and samples 3, 4, 5, 8, 10, 11, 12 were pooled and became sample 2. Concentration of sample 1 was calculated at 351ng/20μL and 254ng/20μL. Purity was calculated at 1.83 for sample 1 and 1.89 for sample 2. These samples were sent for sequencing by Illumina Solexa 50bp paired end read sequencing. 3,539.6Mb of data or an estimated 39X coverage of the genome was achieved.

A third sample prep as shown in Figure 5.9 was performed and will be referred to as 100bp Illumina. Concentration was calculated at 11.2μg/135μL and sample purity was estimated at 1.8. This was sent for 100bp Illumina Solexa paired end sequencing and 3,600Mb were generated with an estimated 40X coverage of the genome. The technological advances in high-throughput sequencing techniques are ever expanding and thus the project changed over time to accommodate this.

A fourth and final set of sample preps were attempted to procure enough gDNA

to send for mate paired sequencing using the Illumina Solexa platform. This was attempted by traditional phenol chloroform methods, using commercial kits, in large batches and in pooled smaller batches. The gDNA concentration was estimated at a high enough level to be sent for sequencing but the various different batches either failed quality control in Scotland or passed quality control but failed to generate enough usable data. This whole process took a significant amount of time. As a hybrid assembly was to be attempted when all data was returned and it was believed that the mate-paired library would add significantly to the project, the progress on the assembly of the already generated gDNA data was put on hold. The protocol for generating mate libraries for *P. superbus* was never successfully accomplished, so the genomic work on the project was not focused on for long periods while the transcriptome work was prioritised. Some suggestions for future work would be to use gene finding and alignment tools to known annotated genomes to establish homologs as well as further downstream annotation. A substantial review of the available software was performed and this is present on the accompanying CD with a filename of *software.xls*.

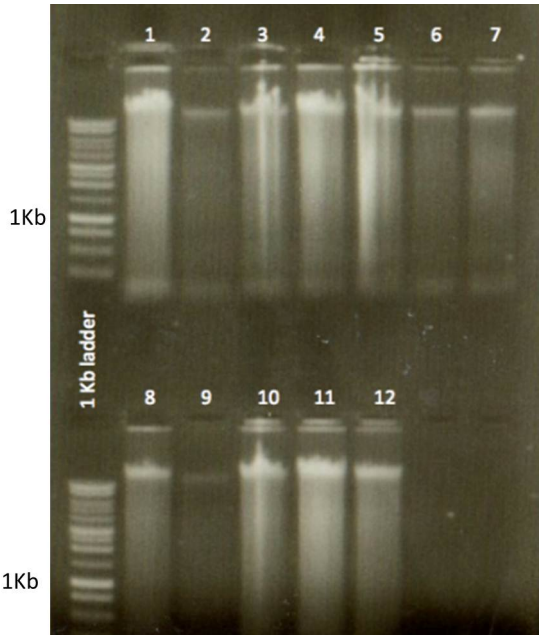


Figure 5.8: *P. superbus* gDNA sample for 50bp Illumina sequencing (pre RNAse treatment).

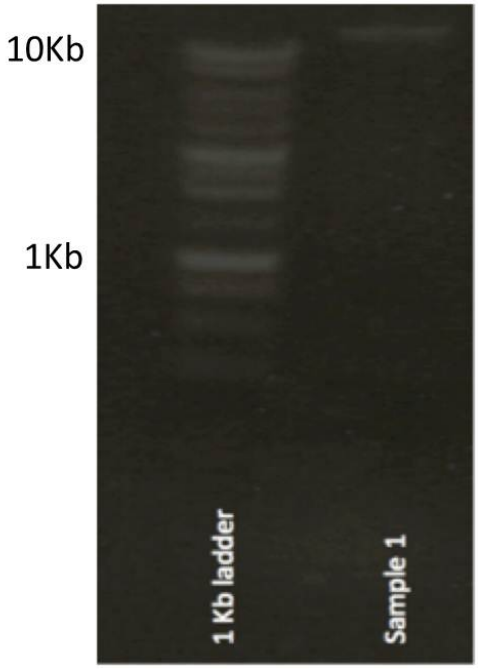


Figure 5.9: *P. superbus* gDNA sample for 100bp Illumina sequencing (post RNAse treatment).

5.3.2 Genome Assembly and Statistics

The mean read length from 454 gDNA was 332bps and over two million reads were generated. These were assembled using the 454 Newbler assembler, and statistics and can be seen in Table 5.1. These contigs were filtered to remove bacterial contaminants. This was done by BLASTing against NR and removing any reads contained in contigs that returned a hit to a bacterial sequence.

Table 5.1: Genome sample 454 gDNA which was sent for sequencing by 454 Roche Titanium platform and assembled using the Newbler 2.3 assembler.

Total number of base pairs generated (Mbps)	707
Total number of base pairs assembled (Mbps)	87
N50 (bps)	2,275
Max contig length (bps)	33,510
Min contig length (bps)	94
Mean contig length (bps)	1,005
Number of bps in all contigs	87,977,093
Total number of contigs	87,354
Number of contigs longer than 1kb	22,638

The 50bp Solexa Illumina sequences generated from six lanes of sequencing were assembled using the Velvet assembler. This is presented in Table 5.2.

Table 5.2: First Velvet assembly of 50bp Solexa Illumina genome sequences.

N50	1,373
No. of contigs in N50	19,971
No. of contigs greater than 1Kb	28,648
Max contig length (bps)	97,629
Average contig length (bps)	409
Number of bps in all contigs	121,236,604
Total number of contigs	295,912

The sequences were searched, using the BLAST algorithm against a bacterial

subset of NR with e-value of 1e-4 and only including top hits using BLAST parameters of b=1 and v=1. The contigs that showed any bacterial hit were taken out and raw reads were aligned against these contigs. The reads that didn't match the contigs were filtered out. The total number of filtered reads remaining was 53,247,304. The longest contig still showed a bacterial hit when BLASTed against NR so the BLAST searches were redone by lowering the e-value to a default e-value. 45,226 contigs were shown to have hits to bacteria. To avoid the possibility that these results might be sequences that could be common to both nematodes and bacteria, the contigs were BLAST against a subset of NR, including the nematodes. The unique contigs hitting bacteria and nematodes were compared. If the bit score of nematode BLAST hit was greater than the score of bacterial BLAST hit, the contig was considered to be coming from nematode and bacteria. As a result, 33,855 contigs were deemed to be bacterial. The reads that aligned to generate these contigs were filtered out and the remaining reads (69,405,012) were reassembled using Velvet. A summary of this is shown in Table 5.3. The 100bp Solexa Illumina was also assembled independently using Velvet. A summary of this is shown in Table 5.4.

Table 5.3: Second Velvet assembly of 50bp Solexa Illumina genome sequences following removal of putative bacterial sequences from the Solexa reads.

N50	1,433
No. of contigs in N50	18,091
No. of contigs greater than 1Kb	27,527
Max contig length (bps)	65,323
Average contig length (bps)	446
Number of bps in all contigs	118,718,696
Total number of contigs	266,030

The two sets of Solexa Illumina reads were assembled together using the commercially available hybrid assembler ClcBio. The resulting summary is shown in

Table 5.4: Velvet assembly of the 100bp Solexa Illumina genome sequences.

N50	365
No. of contigs in N50	79,577
No. of contigs greater than 1Kb	20,612
Max contig length (bps)	17,790
No. of bps in all contigs	140,886,927
No. of contigs	748,365

Table 5.5.

Table 5.5: CLCBio assembly of the 50 and 100bp Solexa Illumina genome sequences.

N50	1,058
No. of contigs in N50	18,994
No. of contigs greater than 1Kb	20,745
Max contig length (bps)	85,445
Average contig length (bps)	769
Number of bps in all contigs	95,487,185
Total number of contigs	124,097

Following post-filtering of the independent datasets for bacterial sequences, it was decided to reassemble all post-filtered reads from 454 genome, 50bps Illumina and 100bps Illumina in a series of hybrid assemblers. Velvet had previously succeeded with the volumes of data involved, so this was attempted. A summary of the results are shown in Table 5.6.

CLCBio, the commercial assembler was also attempted for the hybrid cross platform assembly. This was approached in two ways. Initially, the assembler was fed the reads as 'non paired end' reads. A summary of this is shown in Table 5.7.

Subsequently, the CLCBio assembler was given the reads as 'paired end' reads to see if this would improve the performance as the N50 recorded was quite low. Summary results are seen in Table 5.8. Table 5.9 shows that the longest contigs

Table 5.6: Hybrid assembly, using Velvet, of the 454 genome reads, 50bps Solexa Reads and 100bp Solexa Illumina reads post-filtering to remove for bacterial contaminants.

N50	1,686
No. of contigs in N50	20,484
No. of contigs greater than 1Kb	37,106
Max contig length (bps)	78,636
Average contig length (bps)	660
Number of bps in all contigs	159,666,933
Total number of contigs	241,903

Table 5.7: CLCBio assembly of the 50 and 100bp Solexa Illumina reads(using the sequences as non paired end reads) and the 454 Titanium gDNA reads.

N50	987
No. of contigs in N50	32,908
No. of contigs greater than 1Kb	32,318
Max contig length (bps)	69,511
Average contig length (bps)	709
Number of bps in all contigs	142,513,835
Total number of contigs	200,973

were obtained from the assembly of the 454 gDNA sequences and as expected the assemblies of the shorter Solexa Illumina reads generated assemblies with lower contig sizes. But, surprisingly, the hybrid assembly of the Solexa Illumina and 454 reads returned an assembly with a smaller N50 than either of the single platform assemblies derived for these sequences. It was hoped that the larger insert paired reads would resolve this problem but, due to technical and time constraints, these reads were not obtained.

Table 5.8: CLCBio assembly of the 50 and 100bp Solexa Illumina (using the sequences as paired end reads) and the 454 Titanium gDNA reads.

N50	945
No. of contigs in N50	34,211
No. of contigs greater than 1Kb	31,544
Max contig length (bps)	64,318
Average contig length (bps)	692
Number of bps in all contigs	142,873,836
Total number of contigs	206,215

Table 5.9: Comparison of nuclear genome assemblies on the basis of number of contigs, number of base pairs and N50. 1: 454 gDNA reads assembled using Newbler 2.3. 2: 50bp Solexa Illumina reads assembled using Velvet (first assembly). 3: 50bp Solexa Illumina using Velvet (second assembly). 4: 100bp Solexa Illumina reads assembled using Velvet. 5: 50 and 100bp Solexa Illumina reads assembled using CLCBio. 6: 50 and 100bp Solexa Illumina reads and 454 gDNA reads assembled using Velvet. 7: 50 and 100bp Solexa Illumina reads and 454 gDNA reads assembled using CLCBio (using the sequences as non paired end reads). 8: 50 and 100bp Solexa Illumina reads and 454 gDNA reads assembled using CLCBio (using the sequences as paired end reads).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
No. of base pairs in all contigs (Mbps)	87.9	121.2	118.7	140.9	95.5	159.7	142.5	142.9
No. of contigs	87,354	295,912	266,030	748,365	124,097	241,903	200,973	206,215
N50	2,275	1,373	1,433	365	1,058	1,686	987	945

5.3.3 Further Analysis of 454 gDNA

A BLASTX of the top 20 longest contigs from the 454 assembly was carried out against UniRef and these results are shown in Table 5.10. All of the longest contigs, except one had hits to nematodes. Eleven of the twenty longest contigs had hits to *B. malayi* with five to *C. elegans* and four to *C. briggsae*. The lack of bacterial hits would infer that the filtering step was sufficiently thorough as bacterial genomes would be expected to assemble more readily and yield longer contigs than their nematode counterparts.

These contigs were annotated using BLAST2GO, as previously described, and the breakdown of biological processes can be seen in Figure 5.10. Under this category, almost 40% of the contigs gave a hit to regulation of biological process or multicellular organismal development with significant number of hits to metabolic process also.

Molecular functions are outlined in Figure 5.11. A significant portion of the hits (56%) were to binding, with a large portion to catalytic activity (11%).

In terms of cellular component, the majority of hits were to the intracellular region, with almost one third of the additional hits to the membrane as can be seen in Figure 5.12.

Table 5.10: UniRef90 BLASTX matches for the 20 longest contigs.

EValue	**Query ID.**	**Subject ID.**	**Subject description**	**Species**
0	contig37668	UniRef90_A8NW38	DNA replication licensing factor MCM3	*Brugia malayi*
0	contig37224	UniRef90_UPI0000223743	Hypothetical protein CBG06205	*Caenorhabditis briggsae* AF16
0	contig21182	UniRef90_P34369	Pre-mRNA-splicing factor 8 homolog	*Caenorhabditis elegans*
0	contig20524	UniRef90_Q86NF8	VAB-10A protein	*Caenorhabditis elegans*
0	contig42814	UniRef90_A8WFI8	Filamin (Actin binding protein) alpha protein 1, isoform b	*Caenorhabditis elegans*
0	contig33482	UniRef90_UPI0000122F7C	Hypothetical protein CBG05936	*Caenorhabditis briggsae* AF16
9.00E-180	contig30266	UniRef90_A8XWJ4	CBR-GLF-1 protein	*Caenorhabditis briggsae*
7.00E-172	contig03993	UniRef90_A8NFV4	Cation efflux family protein	*Brugia malayi*
1.00E-161	contig04202	UniRef90_Q9XWE1	Protein Y47H9C.5a, partially confirmed by transcript evidence	*Caenorhabditis elegans*
8.00E-154	contig22133	UniRef90_A8PYW9	RhoGAP domain containing protein	*Brugia malayi*
8.00E-140	contig37127	UniRef90_A8Q011	Laminin-like protein K08C7.3, putative	*Brugia malayi*
2.00E-139	contig13381	UniRef90_A8PDY6	Putative uncharacterized protein	*Brugia malayi*
4.00E-131	contig27536	UniRef90_A7LPD6	Immunoglobulin-like cell adhesion molecule family protein 3, isoform b	*Caenorhabditis elegans*
1.00E-130	contig55087	UniRef90_A8QGU5	Zinc finger, C2H2 type family protein	*Brugia malayi*
2.00E-128	contig01446	UniRef90_A8PUJ4	RuvB-like 2, putative	*Brugia malayi*
2.00E-124	contig64304	UniRef90_A8PZ19	Zinc finger, C2H2 type family protein	*Brugia malayi*
2.00E-99	contig07431	UniRef90_A8PVA2	DOMON domain containing protein	*Brugia malayi*
6.00E-59	contig12787	UniRef90_UPI000180D3B6	PREDICTED: similar to novel EGF domain containing protein	*Ciona intestinalis*
8.00E-39	contig37561	UniRef90_A8XKV6	Putative uncharacterized protein	*Caenorhabditis briggsae*
4.00E-33	contig57911	UniRef90_A8PIA7	Rhomboid family protein	*Brugia malayi*

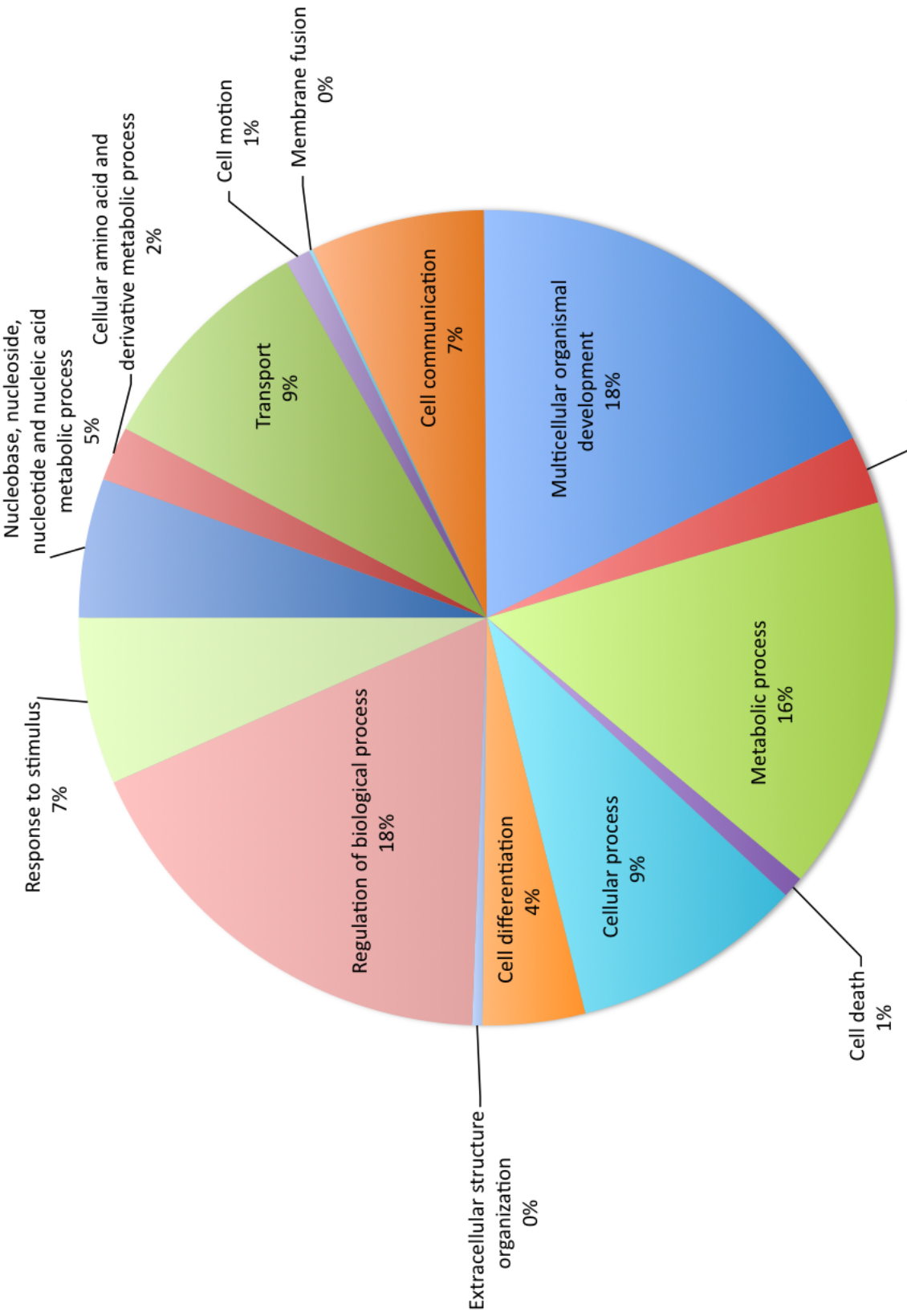


Figure 5.10: 454 gDNA assembly showing biological process categories breakdown.

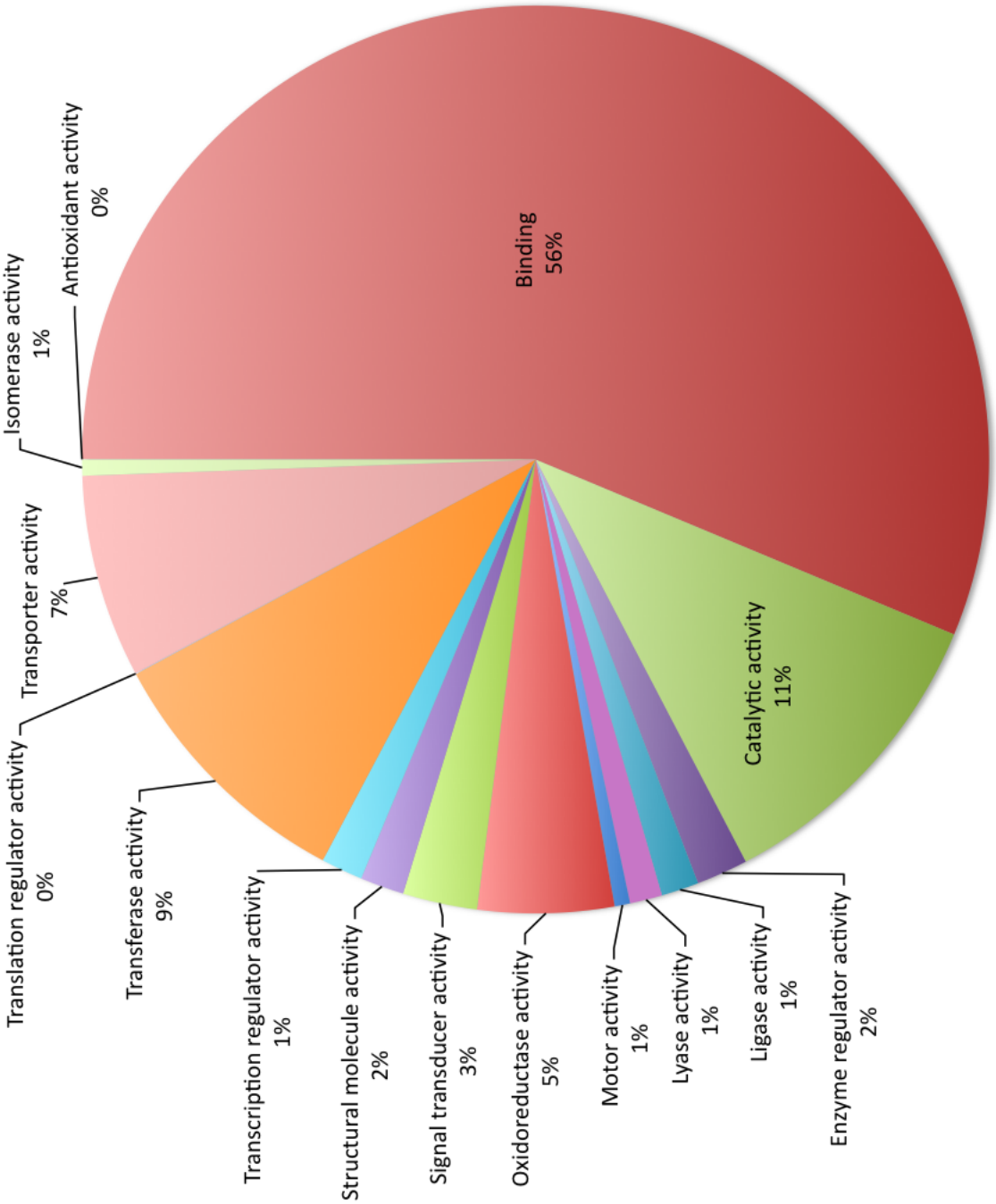


Figure 5.11: 454 gDNA assembly showing molecular function categories breakdown.

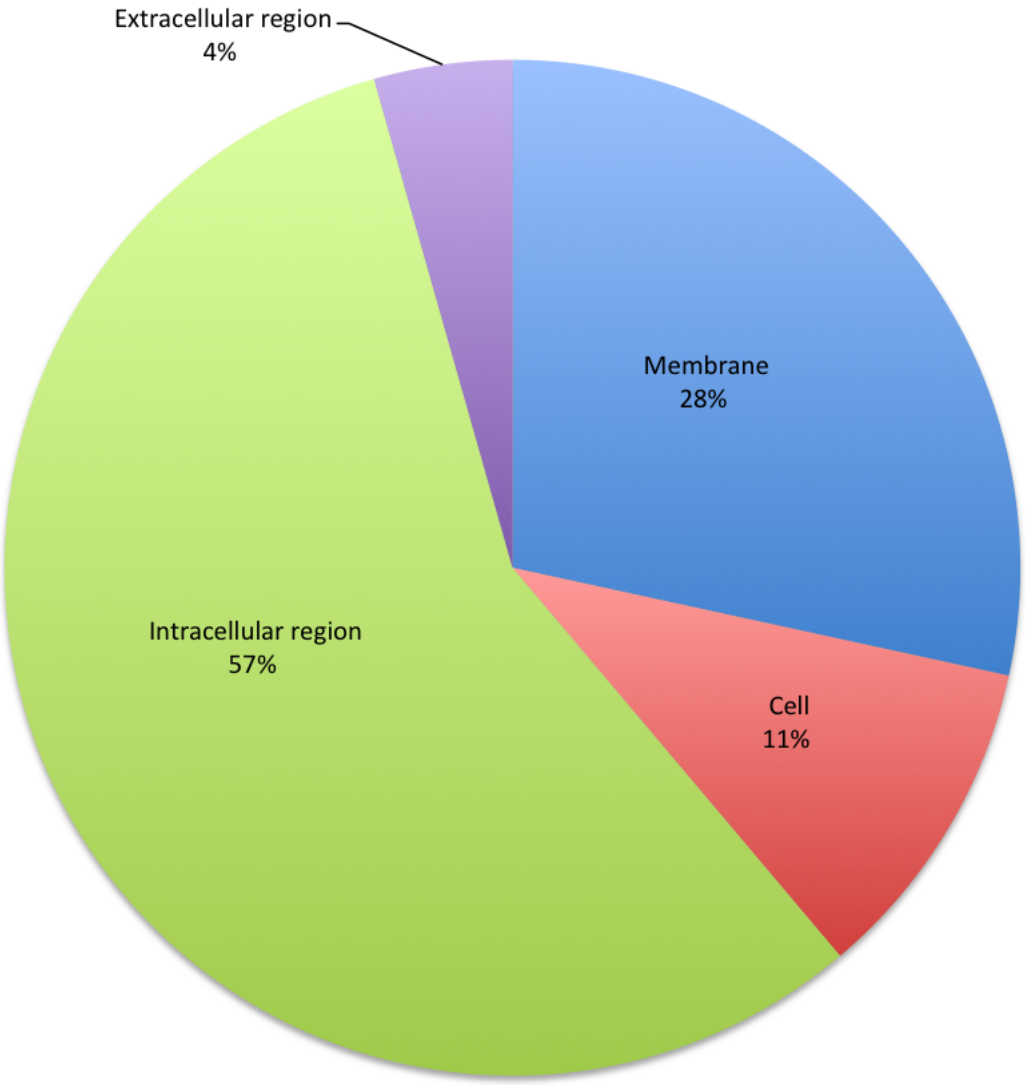


Figure 5.12: 454 gDNA assembly using cellular component categories breakdown.

5.3.4 The Karyotypes of *P. superbus* and *P. davidi*

C. elegans and *P. davidi* were used as controls in this experiment as the chromosomal number of each was known. *C. elegans* was shown to have $n = 6$ chromosomes as was expected (Hillier et al., 2005).

Confocal images of the *C.elegans* chromosomes can be seen in Figure 5.13. The karyotype of *P. davidi* was investigated as it is also a freezing tolerant nematode and a sister species of *P. superbus*. It had previously been shown that the chromosome number of *P. davidi* was $n = 7$ (Goldstein & Wharton, 1996). It is clearly shown in Figure 5.14 of the *P. davidi* oocyte and in particular in Figure 5.15 of the *P. davidi* egg that the chromosome number is actually $n = 12$.

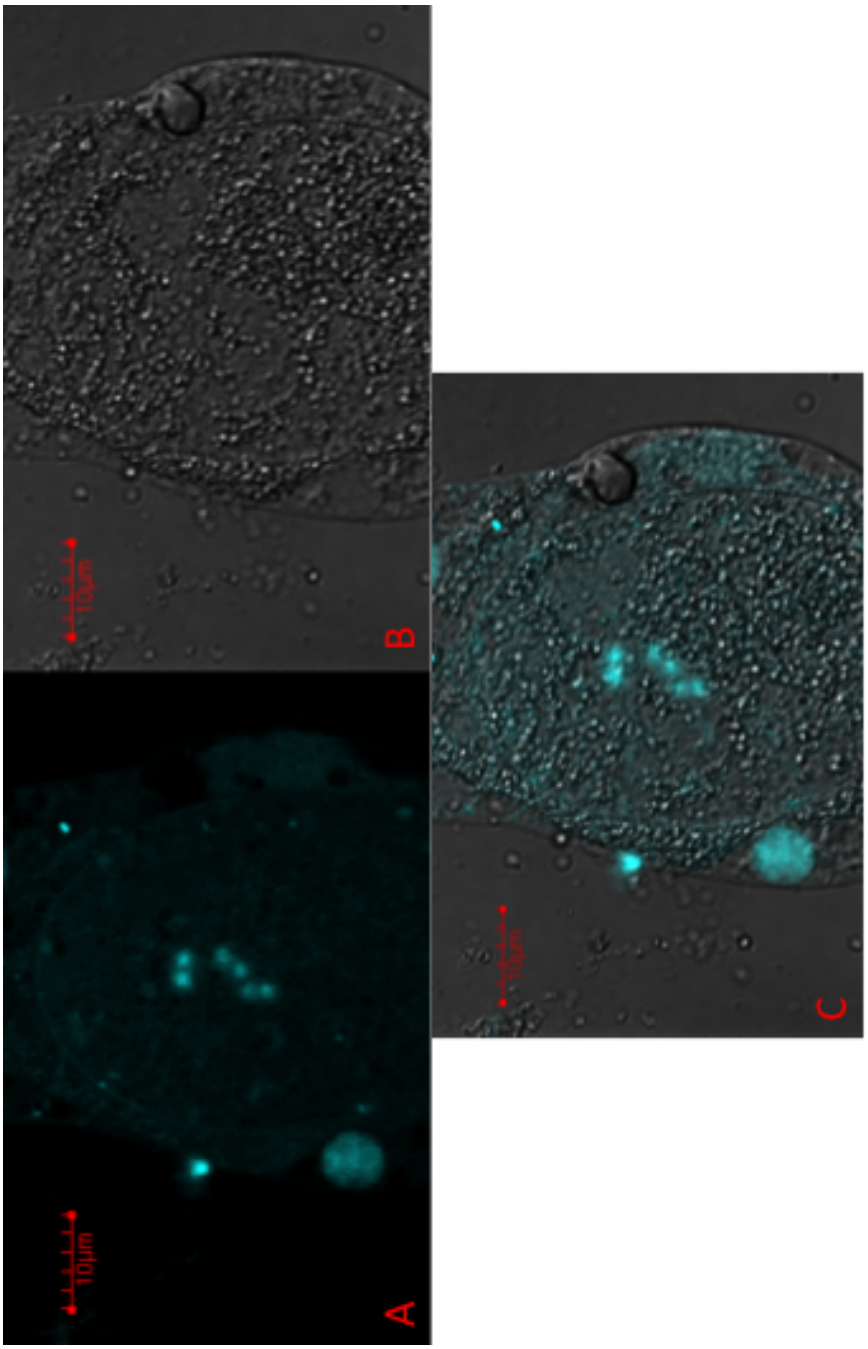


Figure 5.13: *C. elegans* egg cell following DAPI staining as seen thorough a confocal microscope. Panel C shows the overlap of Panel A and Panel B to show that the 6 chromosomes lie within the egg cell.

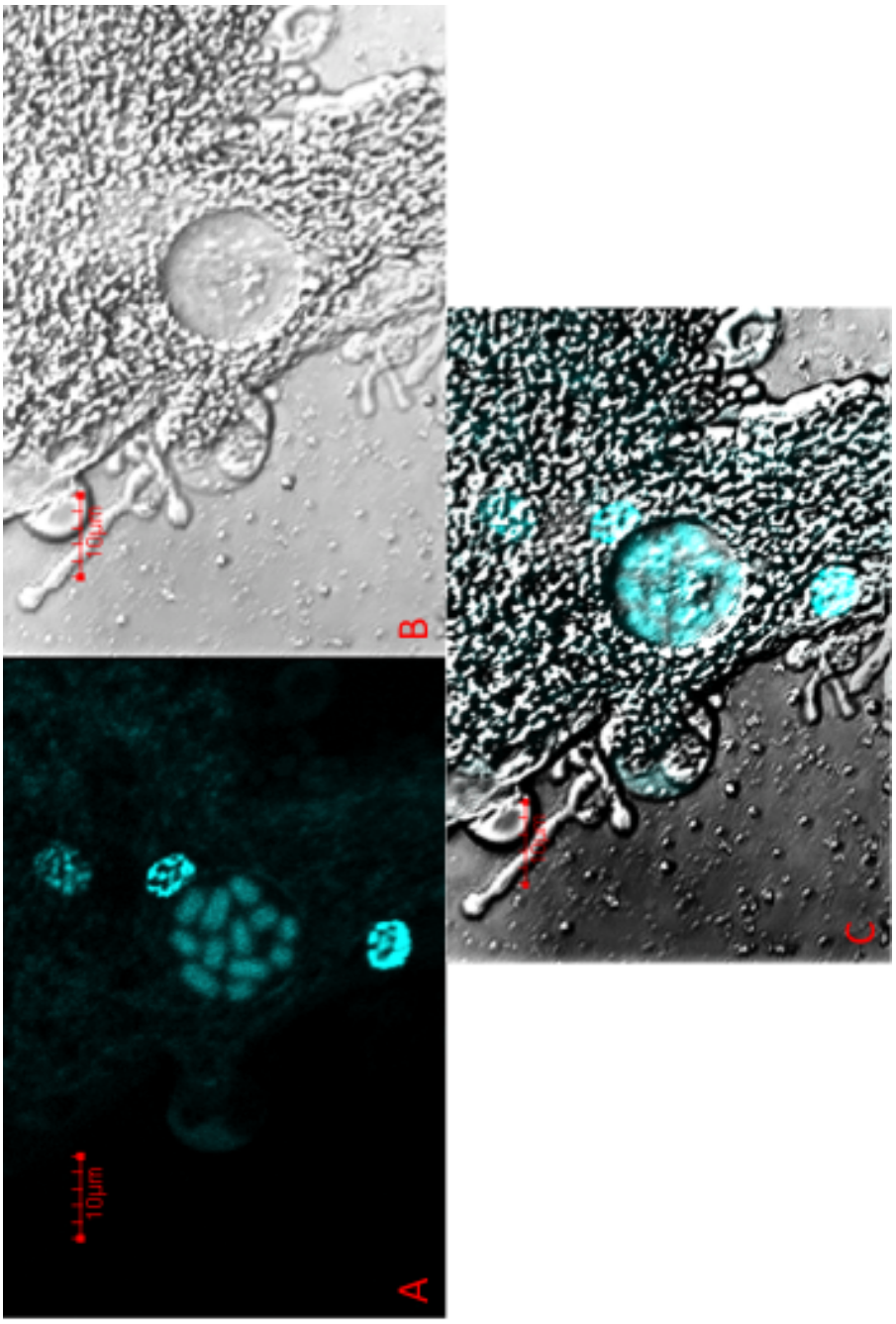


Figure 5.14: *P. davidi* oocyte following DAPI staining as seen through a confocal microscope. Panel C shows the overlap of Panel A and Panel B to show that the 12 chromosomes lie within the egg cell.

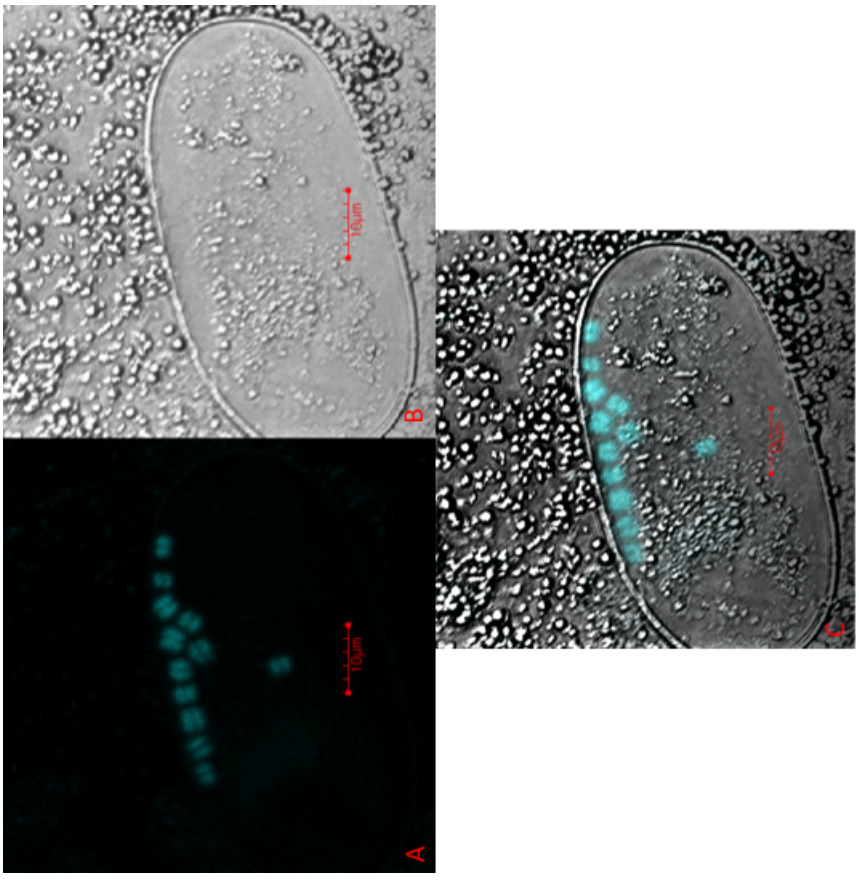


Figure 5.15: *P. davidi* egg cell following DAPI staining as seen through a confocal microscope. Panel C shows the overlap of Panel A and Panel B to show that the 12 chromosomes lie within the egg cell.

The chromosome number in *P. superbus* was investigated as thoroughly as possible with testis, egg, oocyte, ovary and egg cells all examined. Each confirmed the chromosome number of the other with a total of 4 chromosomes. This is shown in Figures 5.16, 5.17, 5.18 and 5.19.

Figure 5.16: *P. superbus* sperm cell following DAPI staining as seen through a confocal microscope. Panel C shows the overlap of Panel A and Panel B to show that the 4 chromosomes lie within the sperm cell.

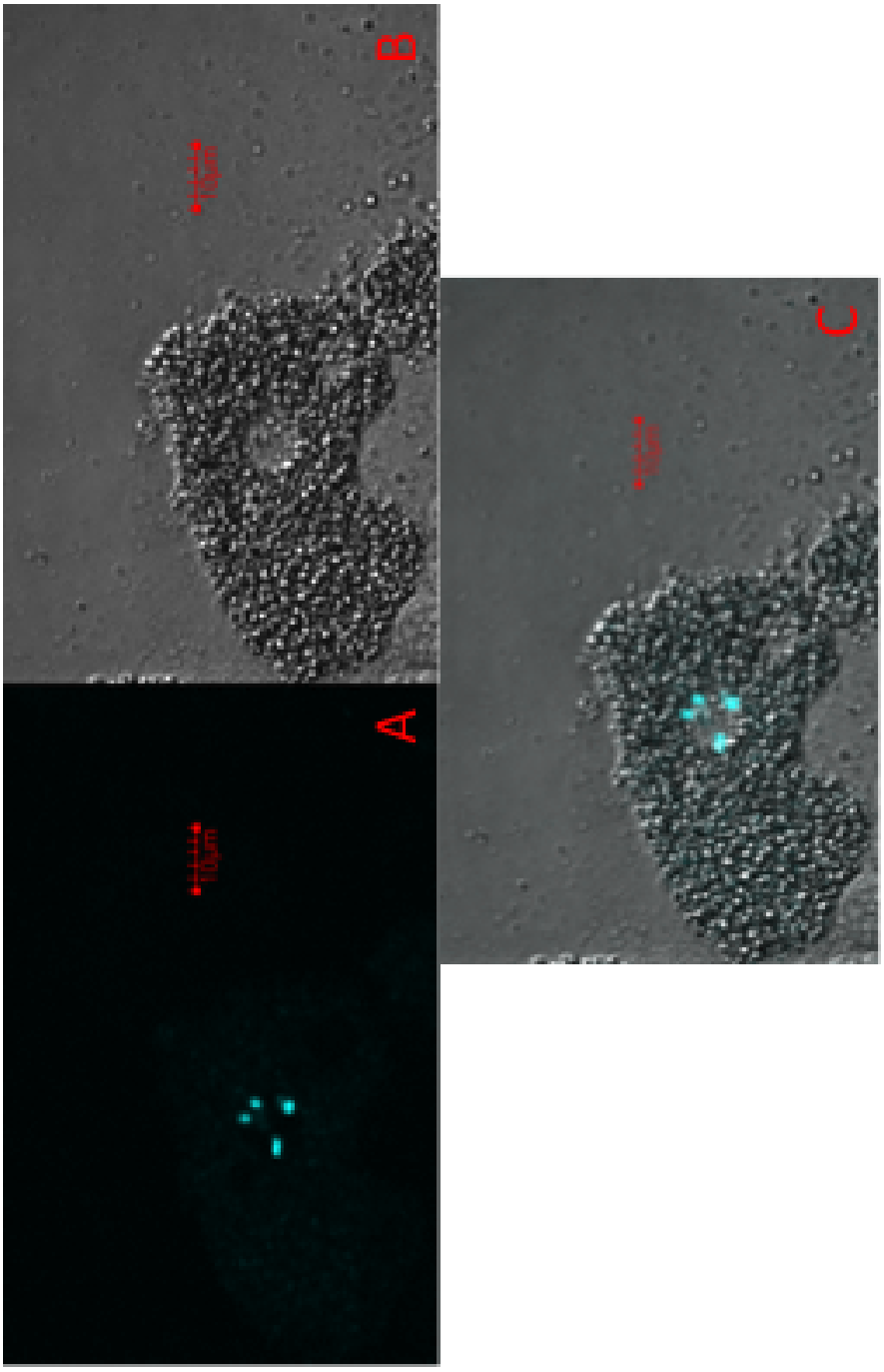


Figure 5.17: *P. superbus* egg cell following DAPI staining as seen through a confocal microscope. Panel C shows the overlap of Panel A and Panel B to show that the 4 chromosomes lie within the egg cell.

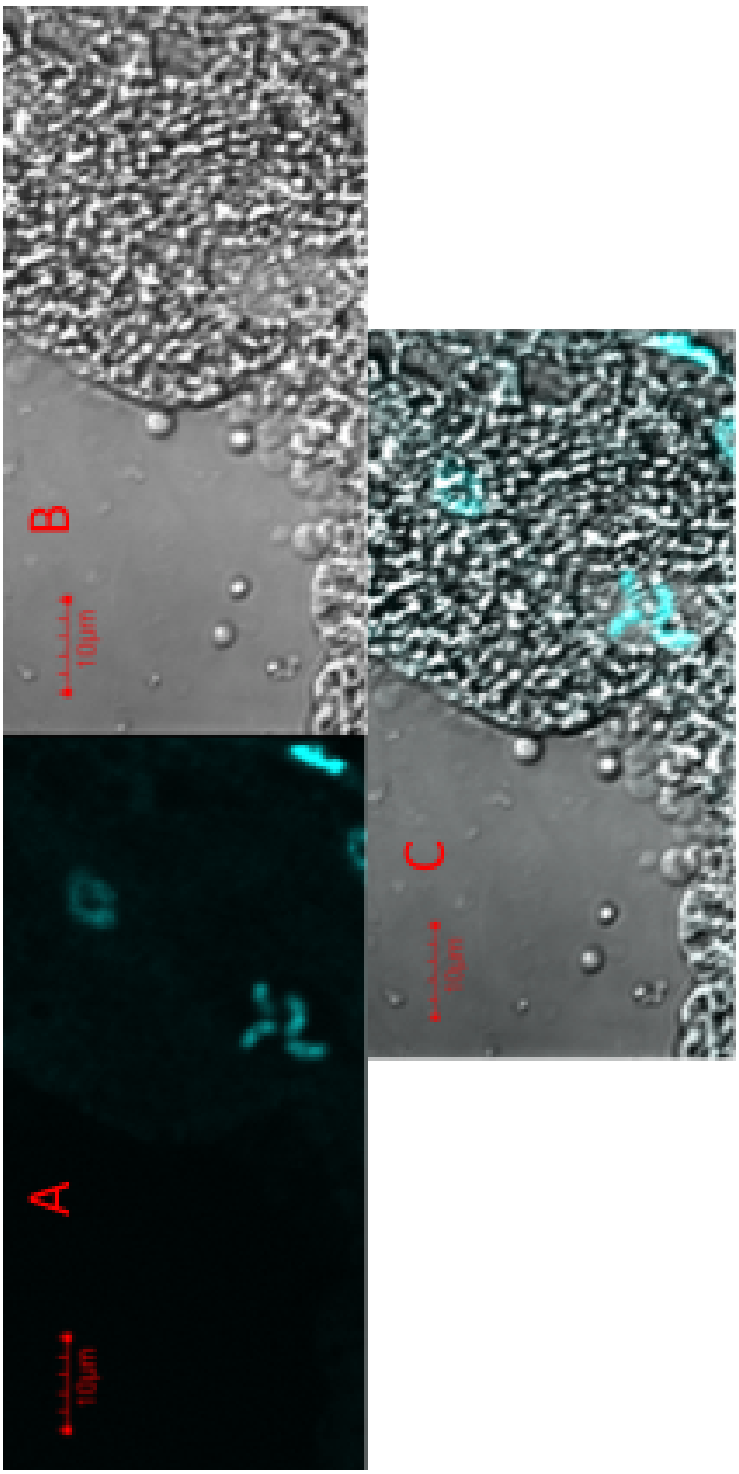


Figure 5.18: *P. superbus* egg cell following DAPI staining as seen thorough a confocal microscope. Panel C shows the overlap of Panel A and Panel B to show that the 4 chromosomes lie within the egg cell.

Figure 5.19: *P. superbus* oocyte cell following DAPI staining as seen through a confocal microscope. Panel C shows the overlap of Panel A and Panel B and the 4 chromosomes lie within the egg cell.

5.4 Discussion

5.4.1 Genome

Genome statistics pre- and post-filtering for bacterial contamination show the importance of a filtering step in the assembly process. As the nematodes are free-living and consume bacteria, the washing stage is not enough to ensure they have fully digested the bacteria in their gut. The stringent checks used to ensure removal of all bacterial sequences may appear over-zealous, but they ensure that no bacterial sequences are misassembled as nematodes sequences. The Newbler assembler has been shown in previous chapters to work well; this is also noted in the genome assembly of the 454 dataset. The longest N50 was recorded using this assembler at almost double what was achieved with either Velvet or ClcBio. The length of the reads generated by the 454 Titanium platform are significantly longer in length than those generated with Solexa Illumina. Given the difficulties in both generating a library and assembling the mass of data generated by Solexa Illumina the 454 Titanium platform holds its weight in the argument. The custom-made assembly software would also be an important metric to consider, as shown by the statistics generated. The examination of the top 20 hits produced and annotated using UniRef give confidence to the dataset as almost all hits were to nematode species.

The previously published nematode genomes, as of February 2013, are shown in Table 5.11. The free living nematodes have genomes over 100Mb while the plant parasitic nematodes have smaller genomes. This gives confidence that the *P. superbus* will be over 100Mb. Using the assemblies, an estimate could be made that the *P. superbus* genome is between 100-150Mb which fits well with the estimates from other data available. An N50 size of around 1-1.5Kb, as was generated, means that potentially full length genes have been assembled. Future work could include

gene finding and use of a scaffolding technique to link the sequences together.

Table 5.11: Published nematode genomes.

Strain	**Species**	**Strain bclade**	**Strain genome url**
Ascaris suum	*Ascaris suum*	Bclade III	*ftp://ftp.wormbase.org/pub/wormbase/species/a_suum/*
Brugia malayi TRS	*Brugia malayi*	Bclade III	*http://www.wormbase.org/db/gb2/gbrowse/b_malayi*
Bursaphelenchus xylophilus Ka4C1	*Bursaphelenchus xylophilus*	Bclade IV	*http://www.genedb.org/Homepage/Bxylophilus*
Caenorhabditis angaria PS1010	*Caenorhabditis angaria*	Bclade V	*http://www.wormbase.org/db/gb2/gbrowse/c_angaria/*
Caenorhabditis briggsae AF16	*Caenorhabditis briggsae*	Bclade V	*http://www.wormbase.org/db/gb2/gbrowse/c_briggsae/*
Caenorhabditis elegans N2	*Caenorhabditis elegans*	Bclade V	*http://www.wormbase.org/db/gb2/gbrowse/c_elegans/*
Dirofilaria immitis Edinburgh/TRS/Basel	*Dirofilaria immitis*	Bclade III	*http://dirofilaria.org*
Meloidogyne hapla VW9	*Meloidogyne hapla*	Bclade IV	*http://www.hapla.org/*
Meloidogyne incognita Morelos	*Meloidogyne incognita*	Bclade IV	*http://www.inra.fr/meloidogyne_incognita*
Pristionchus pacificus California	*Pristionchus pacificus*	Bclade V	*http://www.pristionchus.org/*
Trichinella spiralis Not specified	*Trichinella spiralis*	Bclade I	*http://www.nematode.net/*

5.4.2 Karyotyping

Most nematodes have a haploid chromosome number of between n=4-12, but karyotype variation has been shown across the species studied. The smallest chromosome number was found in *Parascaris univalens* with just n=1, and the largest to date is *M. hapla* where n=12. Most of the Rhabditida have n=5 or 6, as was confirmed with the study of the *C. elegans* chromosomes (Coghlan, 2005). *P. davidi* was previously reported as having n=7 chromosomes. As the study on *P. davidi* was published in 1996 the method of visualising these chromosomes may not have been as effective as if performed using a confocal microscope (Goldstein & Wharton, 1996). It was evident from all images taken that *P. davidi* has n=12 chromosomes. As it was found that *P. superbus* has just 4 chromosomes, it could be suggested that there has been a substantial amount of gene duplication in *P. davidi* and that *P. davidi* may, therefore, have a longer genome size.

Chapter 6

The Mitochondrial Genome of *Panagrolaimus superbus*

6.1 Introduction

Mitochondrial research began in the 19th century where the term symbiosis was first used in 1879 (De Bary, 1879). In Latin, symbiosis means "living together" and draws on a theory that two organisms can coexist together and, even, perhaps create a single organism which combines both original organisms. This theory is incorporated into the endosymbiotic hypothesis for the origin of mitochondria. Mereschkowski (1903) published a very influential paper on the theory that there was a double origin for living organisms. Wallin (1927) later suggested that mitochondria were descendants of endosymbiotic bacteria. The endosymbiotic theory lost popularity until the mid 1960's (Sagan, 1967). Advances in technology, such as the electron microscope in the 1950's (Scheffler, 2000), the identification of the central dogma of biology in 1953 (Watson, 1953), together with progress in biochemistry, lead to a better understanding of the structure and functions of mitochondria and to the discovery of mtDNA in the 1960's (reviewed by Mounolou

& Lacroute (2005)). In 1970 Lynn Margulis published Origin of Eukaryotic Cells, an influential book which revitalised the endosymbiotic hypothesis for the origin of mitochondria and chloroplasts.

The belief that mitochondria originated from bacterial cells stems from the many similarities between them (Margulis, 1981). Mitochondrial and bacterial DNA is circular, while eukaryotic nuclear DNA is linear; mitochondrial and bacterial ribosomes are of similar size and structure (70S), whereas eukaryote ribosomes are larger (80S). Mitochondrial protein synthesis starts with N-formyl methionine, as in bacteria whereas in eukaryotes protein synthesis is initiated by methionine (Kozak, 1983). Members of the rickettsial subdivision of the α-Proteobacteria are considered to be the closest eubacterial relatives of mitochondria (Yang et al., 1985). The past few years have seen advances in sequencing technology and reduced costs and as a result many mitochondrial genomes have been sequenced. This led to an interesting study on the mitochondrial genomes of protists which revealed compelling evidence towards a single protomitochondrial ancestor (Gray, 1999).

6.1.1 Structure and Functions of the Mitochondria

Mitochondria are bounded by a double-membrane system, consisting of inner and outer membranes. Folds of the inner membrane (cristae) extend into the matrix. The matrix contains the mitochondrial genetic system as well as the enzymes responsible for the central reactions of oxidative metabolism. Mitochondria play a central role in metabolism and bioenergetics, with over one thousand mitochondria found in each cell. The mitochondria are the site of oxidative phosphorylation which synthesises ATP from the oxidative breakdown of glucose and fatty acids. Mitochondria have also been identified as performing a role in programmed cell death, or apoptosis (Scheffler, 2000).

6.1.2 Mitochondrial Genome Structure

In 1981, complete mtDNA genome sequences of humans (Anderson et al., 1981) and mice (Bibb et al., 1981) were published. These were followed by full sequence data for the mtDNA genomes of several other vertebrates. In 1985 the mtDNA sequence of *Drosophila yakuba* was completed by Clary & Wolstenholme (1985) and Okimoto et al. (1992) published the mtDNA genomes of the nematodes *C. elegans* and *A. suum*. These early mtDNA datasets revealed that animal mtDNAs were small molecules (∼14-16Kb) which encoded relatively few protein coding genes (13 in mammals), along with 22 transfer RNA (tRNA) and 2 ribosomal RNA (rRNA) genes. The structure of these animal mtDNA genomes, being circular and compact, was very different from nuclear DNA. The mtDNA genes were very densely packed, with their coding sequences running directly into each other, leaving very little room for regulatory DNA sequences, except for a single large non-coding region thought to contain elements that regulate replication and transcription (Shadel & Clayton, 1997). There are only 22 tRNA species to translate the mitochondrial mRNAs, whereas at least 30 tRNA species are required to translate the universal genetic code (Crick et al., 1961). Translation of the mtDNA genetic code is accomplished by an extreme form of wobble in which U in the anticodon of the tRNA can pair with any of the four bases in the third codon position of mRNA, allowing four codons to be recognised by a single tRNA (Barrell et al., 1980). In addition, some codons specify different amino acids in mtDNA than in the universal code.

As more mtDNA molecules were sequenced from a greater diversity of eukaryotes, it was found that many mtDNA molecules were considerably larger than those of animals (Figure 6.1).

For example, mtDNA from the protozoan *Acanthamoeba castellanii* is 41.5Kb (Burger et al., 1995); that of the yeast *Saccharymyces cerevisiae* is 85.8Kb (Foury

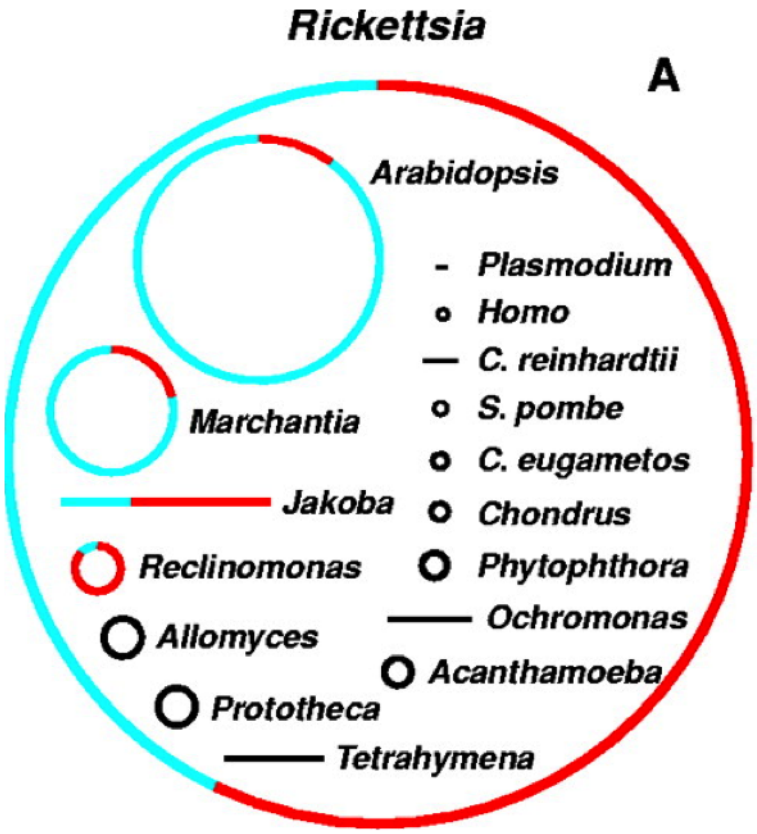


Figure 6.1: The sizes of mtDNA genomes compared with an α-Proteobacterial (*Rickettsia*) genome. Circles and lines represent circular and linear genome shapes, respectively. For genomes >60Kb, the DNA coding for genes with known function (red) is distinguished from that coding for unidentified ORFs and intergenic sequences (blue) (Gray, 1999).

et al., 1998), while the mtDNA genome of the plant *A. thaliana* is 336.9Kb (Unseld et al., 1997). Comparisons of mtDNA genome structures and sequences have shown that there are two basic types of mtDNAs, designated ancestral and derived (Gray, 1999). Ancestral mtDNA genomes retain clear evidence of their eubacterial ancestry and are characterised by:

- The presence of many extra genes, as compared with animal mtDNA, including additional *nad*, *atp*, and especially ribosomal protein genes (*rps* and *rpl*),
- rRNA genes that encode eubacteria-like LSU (23S), SSU (16S), and 5S rRNAs,
- A complete or almost complete set of tRNA genes,
- A tight packing of genetic information in a genome that consists mostly of coding sequences, with no or few introns
- A standard "universal" genetic code.

The small derived mtDNA genomes which are found in animals and fungi are characterised by:

- Extensive loss of both protein-coding and tRNA genes,
- A highly modified rRNA structure ; there is a severe truncation of rRNA sequences and modified secondary structure,
- Adoption of a highly biased codon usage strategy in protein genes, including the elimination of certain codons,
- Introduction of nonstandard codon assignments such as the use of alternative start codons and abbreviated stop codons.

6.1.3 Why have Mitochondria Retained a mtDNA Genome?

The difference in gene content between all mtDNA genomes and present day α-Proteobacteria, both ancestral and derived, is considered to be the result of gene loss in the transition from bacterial symbiont to cellular organelle, combined with horizontal gene transfer of the symbiont genes to the host nucleus (Blanchard & Lynch, 2000; Gray, 1999). Similarly plant chloroplasts have retained some of their ancestral bacterial genome, but the majority of genes required for chloroplast function have been transferred to the nuclear genome. The products of these organellar genes, which now reside on the nuclear genome, are translated on nuclear ribosomes and are imported in to the organelles as completed polypeptide chains. Studies in the mouse estimate the mitochondrial proteome comprises 940 proteins, with the vast majority of these proteins encoded by the nuclear DNA (Zhang et al., 2008). This raises two questions:

- Why have the organelle genes been translocated to the nucleus? (Blanchard & Lynch, 2000), and
- If most genes have been transferred to the nuclear genome during evolution, why have some genes and their translation apparatus (rRNAs) and tRNAs been retained by the organelles? (Race et al., 1999; Wright et al., 2009).

Concerning the translocation of mitochondria genes to the nucleus, it is hypothesised that mutations accumulate more rapidly in asexually propagated genomes than in sexually propagated genomes because mutations can't be removed through recombination (Lynch & Blanchard, 1998). In addition, respiratory electron transport in mitochondria generates high concentrations of mutagenic reactive oxygen species (ROS). Thus, since mitochondria are maternally inherited, a genetic load

should build up rapidly in mtDNA genes, providing a very strong selective pressure for these genes to be transferred to the nucleus (Lynch & Blanchard, 1998). Nevertheless, mitochondria (and chloroplasts) do retain a small proportion of their ancestral genes, which implies that the location of these genes on organellar genomes is likely to carry a selective advantage. mtDNA genes encode proteins with key roles in electron transport and energy coupling. Mitochondrial electron transport chains are extremely harmful when short circuits occur, leading to enhanced production of ROS. Allen (2003) proposed that organelle genomes have persisted because the structural proteins of bioenergetic membranes must be synthesised rapidly when and where they are needed, to minimise the unavoidable side effects of electron transport.

6.1.4 Nematode Mitochondrial Genomes

To date, 63 complete mtDNA genomes from nematodes have been published (Sultana et al., 2013). Nematode mtDNA genomes typically comprise 12 -14 Kb and contain 12 protein coding genes, 22 transfer RNA and 2 ribosomal RNA genes (Jex et al., 2008). The proteins encoded by nematode mtDNA are as follows: cytochrome oxidase subunits I-III (COXI, COXII and COXIII), cytochrome B apoenzyme (COB), NADH dehydrogenase subunits 1-6 and NADH4L (NADH1, NADH2, NADH3, NADH4, NADH4L, NADH5 and NADH6) and ATP synthase subunit 6 (ATP6). Unlike most other animal mtDNA, nematodes lack ATP synthase 8 (Boore, 1999). While mtDNA genes have a relatively conserved sequence amongst members of the phylum, the order of these genes in the circular genome varies between different nematode genera, even among closely related species (Hyman et al., 2011). This makes each nematode mtDNA genome sequencing project a challenge. mtDNA is very AT rich, with only 30.7% GC content. This adds to the challenge of finding suitable primers and achieving successful PCRs.

Apart from the intrinsic interest in the evolution of mtDNA genes and genomes, mtDNA genes have also been used extensively for phylogenetic studies and as molecular markers for systematic and population genetic studies across a broad range of animal groups. In this research project the *P. superbus* mtDNA genome sequence was assembled and the gene order will now be discussed.

6.2 Materials & Methods

6.2.1 Traditional Approach (PCR Amplification of mtDNA Gene Fragments)

A set of universal PCR primers have been developed for the amplification, in two fragments, of complete mtDNA genomes from diverse nematode species as shown in Figure 6.2 (Hu, 2002). This technique which utilises the Expand 20Kb PCR system (Roche) was attempted with the *P. superbus* mtDNA genome using the universal primers shown in Table 6.1.

Table 6.1: Nematode mitochondrial 'universal primers' tested in this study.

Primer Name	**Primer Sequence**
39F	TAAATGGCAGTCTTAGCGTGA
42R	CCCAATAAATGACGCTCATA
38R	AGAAAAAGCAATCTCATAAGAA
5F	TATGAGCGTCATTTATTGGG
40R	GAATTAAACTAATATCACGT
44R	TCACGCTAAGACTGCCATTTA
58R	CTATAATTACGGCCATCTTGTTG

However, after many attempts including trouble-shooting to changing reagents, length of PCR steps, fresh gDNA, and other long PCR kits, a long mtDNA genome fragment was not achieved. It was decided to amplify smaller fragments of the mtDNA genome. This was done by designing custom-made primers using the sequences that had been identified in the EST study as being of mtDNA origin (Chapter 2). The primers were designed to be strictly between 18-22bps and have a primer melting temperature of between 52-58°C. GC content was kept to a level of 40-60%. Repeats and runs of any one particular base were avoided and primers were checked for cross-homology. Length, melting temperature and optimal annealing temperature of primer pairs were kept as constant as possible. Fifty-eight primers were designed and PCRs were attempted in all likely combinations.

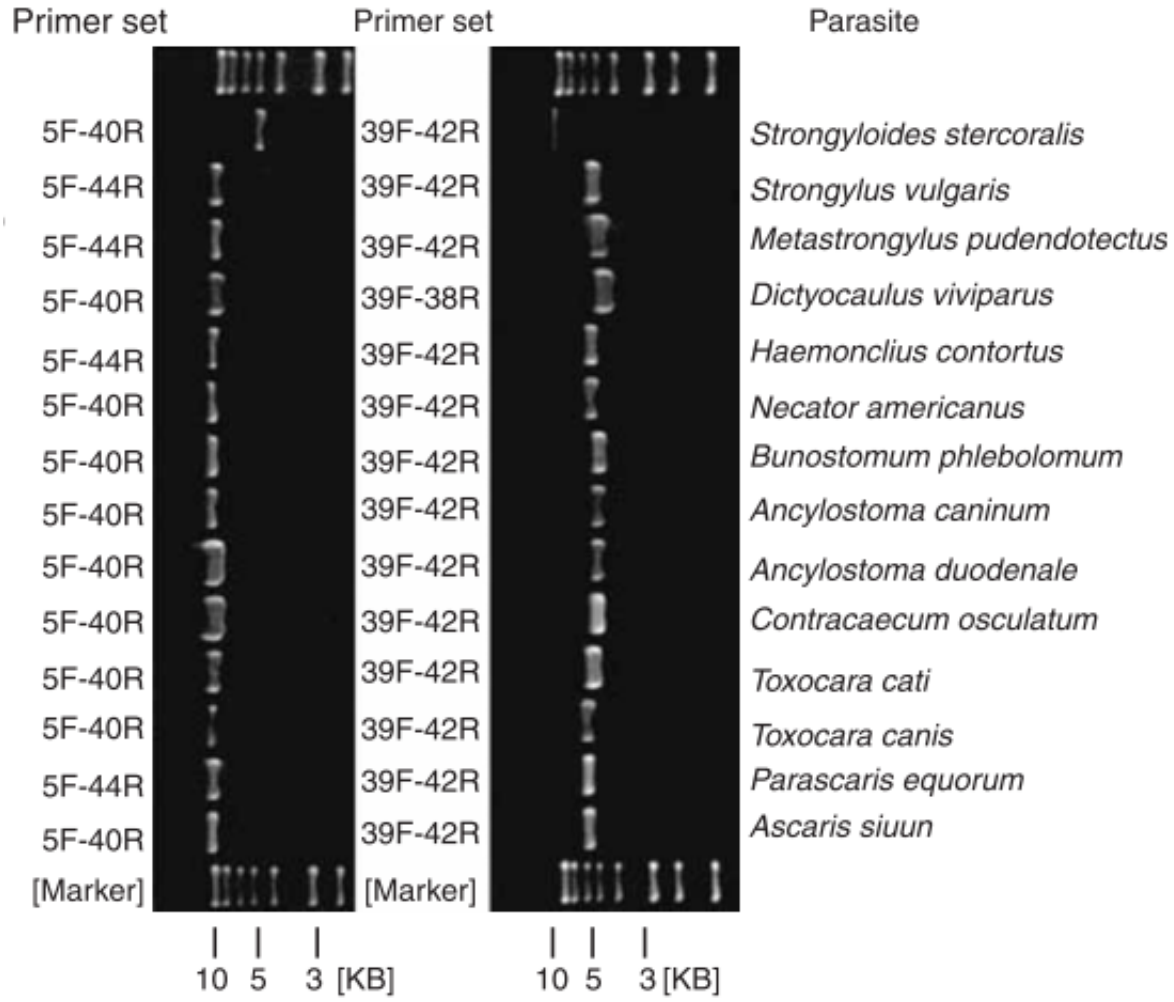


Figure 6.2: Universal nematode primers shown to amplify the whole genome in two fragments (Hu, 2002).

As the gene order varies in nematode mtDNA genomes, PCR amplification requires the use of primers in a range of combinations to identify those primer pairs which amplify contiguous DNA sequences. The successful primers identified are shown in Table 6.2. The naming structure was used for identification purposes. Sequence names containing an F indicates a forward primer and those containing an R indicate a reverse primer. These primers were designed to annotate the following genes: COXI, COXII, COXIII, COB, ND1, ND2, ND4, ATP6 and rnL.

Table 6.2: mtDNA primers designed using EST sequences.

Primer Name	**Primer Sequence**	**Target mtDNA gene**
AATP6F	GTATAGGTCTTCTAAACTTATCTAG	ATP6
ACX2F	ATTTACAAAAGTGCTGAAACATCG	COXII
ACX3F	TGGATGTTGTTTGATTATTTTTATTTG	COXIII
AND4R	GAGCCTTAGGTAATCATAAATG	ND4
ArrnLR	CATGGATAACATCTGCAGAAG	rrnL
CX1F	GTATTAAGATTAGGTGCTGTTTTTGG	COXI
CX2R	GTTACTTCAAGAGCAATAGG	COXII
CX2R2	CTACATAAAAGTTCTCCAAACTGATATTC	COXII
CX3R	GTCAATAAATAATAGCAAATTCTAAGCC	COXIII
CY2R	CAGCTTCAATAAATATTTCTGGATCTC	COB
CYTB2F	GATTAGGACAATGTTTAGTTGAAGATC	COB
CYTBF	TTGCATATGCTATTCTACGTGCT	COB
CYTBR	AAGGATCTTCAACTAAACATTGTCCT	COB
ND1F	GTGCACCTTTTGATTTTTCAGAAGG	ND1
ND2F2	AACATTTACACAAAAAAAATTTGAAGATC	ND2

PCRs were carried out in 25μl volumes. GoTaq from Promega was used for general PCRs, while Platinum Taq from Invitrogen was used when high fidelity was required. PCR conditions can be seen in Table 6.3 with various cycling parameters shown in Table 6.4. Denaturing, annealing and extension steps were repeated for 25-40 cycles. The annealing temperature was estimated as 5°C less than the melting temperature (T_m) of the primer pairs. The extension times were estimated as 1min/Kb of DNA to be synthesised. PCRs were done in an Eppen-

dorf PCR Thermal Cycler or a G-Strorm GS1 Thermal Cycler. PCRs were loaded on ethidium bromide stained DNA gels as described in Section 5.2.3. This method resulted in a fragmented mtDNA genome which presented some initial ideas about gene order.

Table 6.3: Reagent quantities required when using GoTaq (Promega) and Platinum Taq (Invitrogen) to amplify *P. superbus* mtDNA gene fragments using PCR.

Reagent	**Volume per reaction**
GoTaq	
5X Green reaction buffer	5μl
PCR Nucleotide Mix (10mM each)	1μl
$MgCl_2$	2.5μl
Forward primer (10pmol)	1μl
Reverse primer (10pmol)	1μl
GoTaq Polymerase	0.2μl
Template gDNA	100ng - 0.5μg / 50μl
PCR grade water	Bring total volume to 25μl
Platinum Taq	
10X High Fidelity buffer	2.5μl
PCR Nucleotide Mix (10mM each)	0.5μl
$MgCl_2$	1μl
Forward primer (10pmol)	1μl
Reverse primer (10pmol)	1μl
Platinum Taq	0.25μl
Template gDNA	100ng - 0.5μg / 50μl
PCR grade water	Bring total volume to 25μl

Table 6.4: PCR cycling conditions for GoTaq (Promega) and Platinum Taq (Invitrogen) when amplifying *P. superbus* mtDNA gene fragments.

PCR Step	**GoTaq (temp/time)**	**Platinum Taq(temp/time)**
Denaturation/Activation	95°C/2min	94°C/30secs
Denaturation	95°C/0.5-1min	94°C/30secs
Annealing	5°C below the T_m/0.5 − 1min	5°C below the T_m/30sec
Extension	72°C/1min	68°C/1min/Kb
Final extension	72°C/10min	68°C/10min

6.2.2 Agarose Gel Electrophoresis

0.7% agarose gels were made by dissolving pH 8.1 agarose in 1X Tris Acetate EDTA buffer (TAE): 40 mM Tris, 20 mM acetic acid and 2 mM EDTA, pH 8.1 and heating. Once cooled to 60°C ethidium bromide was added (10mg/ml). The solution was then poured into a casting tray and allowed to solidify. 1Kb bench top DNA ladders (Promega) were loaded on the gel. Gels were run at 100V using Sigma-Aldrich or BioRad electrophoresis equipment. The gels were visualised under UV light using a UV transilluminator at 365nm. Gels were photographed using an Eagle-Eye gel documentation system (Stratagene), or an AlphaDigiDoc gel documentation system (Alpha Innotech).

6.2.3 Computational Approach

PCRs were unsuccessful in bridging all gaps in the *P. superbus* mtDNA molecule so it was decided to assemble the genome using computational approaches. This was done using the high-throughput cDNA and gDNA sequences which had been generated for *P. superbus*. Forty one mitochondrial nematode genomes were downloaded from NCBI, which are presented in Table 6.6. These nematodes were predominately sequenced using traditional methods by first designing primers from closely related species, then using PCR and cloning. Long PCR and primer walking strategies were regularly employed. Genomes from closely related species such

as *Strongyloides spp.* and *Steinernema carpocapsae* were included along with other known nematode mtDNA genomes.

The BLASTN algorithm was used to identify the most significant hits to *P. superbus* datasets. These included the EST dataset as described in Chapter 2, the Newbler 2.6 without URT Isotigs, as described in Chapter 3 and Chapter 4 and the various genome assemblies as described in Chapter 5. Using the nematodes listed in Table 6.6, each gene within each of the genomes was individually BLASTed against the EST (as shown in Table 6.5), gDNA and transcriptome datasets using BLASTN and tBLASTX. The hits returned were filtered to remove any lower than 60 bits.

Table 6.5: mtDNA gene names and their corresponding sequence IDs in the EST dataset.

EST ID Number	**mtDNA Gene identified**
PSC00036	ND1
PSC00067	ND4L
PSC00128	CX1
PSC00166	ND5
PSC00814	CX3
PSC00957	ND4
PSC01122	CX2
PSC01234	ATP6
PSC01888	COB
PSC01968_2	ND2
PSC03880	ND4L
PSC00194	rrnL
PSC00065_1	rrnS

The hits found most commonly across all the queries to a particular gene were aligned and the longest sequence was used as a match for that gene. On occasion, for example, with COXI and COXII the same sequence from a gDNA contig was identified as being the best hit for both COXI and COXII. This method was used

to establish that COXI and COXII can be found next to each other in the genome. When COXI and COXII from *C. elegans* aligned to this contig it was found that COXI aligned to one part of the contig and COXII aligned to another. In this way the largest contigs found in the dataset were identified and when aligned it was found that they overlapped each other, eventually leading to a complete mtDNA molecule.

Once a complete mtDNA sequence had been established, individual genes were found by doing alignments with known genes from other nematodes and thus the gene order was established.

Table 6.6: Nematode mitochondrial genomes used in this study

Species Name	**NCBI ID.**	**Length**	**Genes**	**tRNAs**
Agamermis species BH2006	NC_008231	16,561 nt	12	28
Ancylostoma caninum / dog hookworm	NC_012309	13,717 nt	12	24
Ancylostoma duodenale	NC_003415	13,721 nt	12	24
Anisakis simplex / herring worm	NC_007934	13,916 nt	12	24
Ascaris suum / pig roundworm	NC_001327	14,284 nt	12	24
Brugia malayi	NC_004298	13,657 nt	12	24
Bunostomum phlebotomum / cattle hookworm	NC_012308	13,790 nt	12	24
Caenorhabditis briggsae	NC_009885	14,420 nt	12	24
Caenorhabditis elegans	NC_001328	13,794 nt	12	24
Chabertia ovina / large-mouth bowel worm	NC_013831	13,682 nt	12	24
Cooperia oncophora	NC_004806	13,636 nt	12	24
Cucullanus robustus	NC_016128	13,972 nt	12	24
Dirofilaria immitis / dog heartworm nematode	NC_005305	13,814 nt	12	24
Enterobius vermicularis / human pinworm	NC_011300	14,010 nt	12	24
Haemonchus contortus / barber pole worm	NC_010383	14,055 nt	12	24
Heliconema longissimum	NC_016127	13,610 nt	12	24
Heterorhabditis bacteriophora	NC_008534	18,128 nt	12	24
Hexamermis agrotis	NC_008828	24,606 nt	14	26
Metastrongylus pudendotectus	NC_013813	13,793 nt	12	24
Necator americanus	NC_003416	13,605 nt	12	24
Oesophagostomum dentatum	NC_013817	13,869 nt	12	24
Onchocerca volvulus	NC_001861	13,747 nt	12	24
Romanomermis culicivorax	NC_008640	26,194 nt	14	34
Romanomermis iyengari	NC_008693	18,919 nt	13	27
Romanomermis nielseni	NC_008692	15,546 nt	12	27
Setaria digitata	NC_014282	13,839 nt	12	24
Steinernema carpocapsae	NC_005941	13,925 nt	12	24
Strelkovimermis spiculatus	NC_008047	18,030 nt	12	29
Strongyloides stercoralis	NC_005143	13,758 nt	12	24
Strongylus vulgaris	NC_013818	14,301 nt	12	24
Syngamus trachea	NC_013821	14,647 nt	12	24
Teladorsagia circumcincta	NC_013827	14,066 nt	12	24
Thaumamermis cosgrovei	NC_008046	20,013 nt	12	28
Toxocara canis / dog roundworm	NC_010690	14,322 nt	12	24
Toxocara cati / cat roundworm	NC_010773	14,029 nt	12	24
Toxocara malaysiensis	NC_010527	14,266 nt	12	24
Trichlinella spiralis	NC_002681	16,706 nt	13	24
Trichostrongylus axei	NC_013824	13,653 nt	12	24
Trichostrongylus vitrinus	NC_013807	13,800 nt	12	24
Wellcomia siamensis	NC_016129	14,128 nt	12	24
Xiphlinema americanum	NC_005928	12,626 nt	12	21

6.3 Results

6.3.1 PCR Amplification of mtDNA Fragments

The PCR primers which were successfully used to amplify *P. superbus* mtDNA fragments were presented in Table 6.2. Figures 6.3 to 6.10 show images of the successfully amplified PCR products when visualised on a 0.7% agarose gel. The successful PCR primer pairs and the estimated fragment sizes are presented in Table 6.7. For each of these PCR reactions the primer pairs were selected such that the forward and reverse primers annealed to different mtDNA genes and in the case of COB, the two ends of the gene. Thus, a successful amplification indicated that the two genes corresponding to these primers were located at contiguous positions on the circular mtDNA molecule of *P. superbus*.

Table 6.7: The PCR primer pair combinations which were successfully used to amplify *P. superbus* mtDNA fragements.

Forward Primer	**Reverse Primer**	**Band size**	**Figure No.**
AATP6F	ACOBR	$\sim$ 1300bps	6.3
CX2F	RLR	$\sim$ 600bps	6.4
ACX3F	AND4R	$\sim$ 500bps	6.5
BND1F	BATP6R	$\sim$ 800bps	6.6
CYTB2F	CX3R	$\sim$ 850bps	6.7
CYTBF	CYTBR	$\sim$ 200bps	6.8
CX1F	CX2R2	$\sim$ 800bps	6.9
ND2F2	CY2R	$\sim$ 1000bps	6.10

Figure 6.11 is a diagram of the mtDNA molecule with typical gene order as found in many species including *C. elegans* showing the locations of the successful PCR amplifications for the *P. superbus* mtDNA genome. This figure shows that the following genes are located in sequence on the *P. superbus* mtDNA molecule: ND1, ATP6, ND2, COB, COXIII, ND4 and that the following genes are located

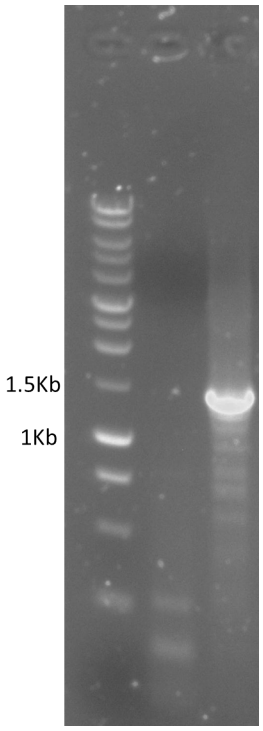


Figure 6.3: PCR product (∼1,300bp) obtained from *P. superbus* using the primer pairs AATP6F/ACOBR and confirming the ATP6/COB border. The marker DNA shown in the left lane is the Promega 1Kb marker.

in sequence, COXI, COXII, rrnL failed to generate a band. However, PCR amplifications using primer pairs in various combinations for ND3, ND4L, ND5, ND6 and rrnL. Thus a computational approach utilising the high-throughput cDNA and gDNA sequences was attempted as previously described.

A diagrammatic representation of a 9,661bp fragment of the genome can be seen in Figure 6.12. This shows an alignment of three sequences, isotig04908 from the Newbler 2.6 without Isotigs transcriptome dataset, contig1203 from the Velvet assembly of the 100bp Solexa Illumina genome dataset and contig08802 from the 454 gDNA dataset. The relevant ESTs used to find the overlapping sequences are shown in green. The identified tRNAs can be seen in pink.

Using this method the genome was assembled and found to be 13,970bps long

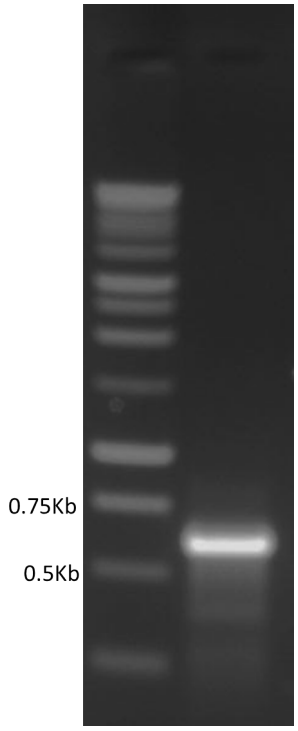


Figure 6.4: PCR product (∼600bp) obtained from *P. superbus* using the primer pairs CX2F/RLR and confirming COXII/rrnL gene border.

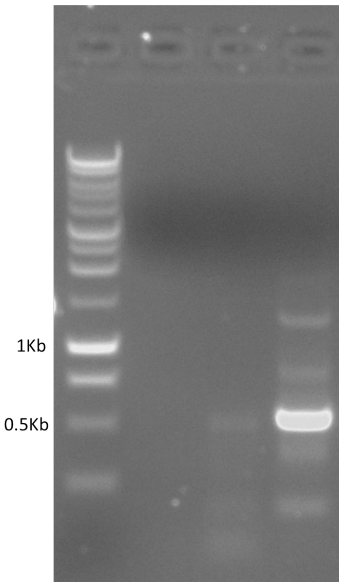


Figure 6.5: PCR product (∼500bp) obtained from *P. superbus* using the primer pairs ACX3F/AND4R and confirming ND4/COXIII gene border.

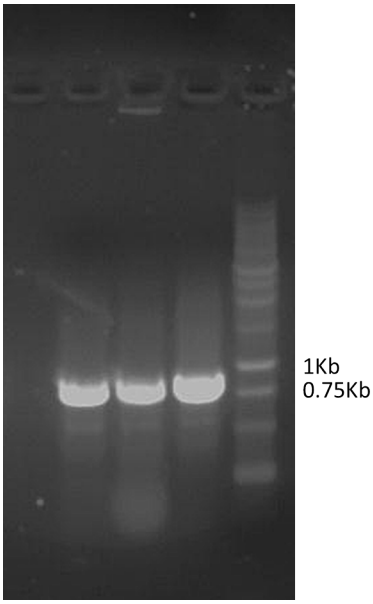


Figure 6.6: PCR product (∼800bp) obtained from *P. superbus* using the primer pairs BND1F/BATP6R and confirming ND1/ATP6 gene border.

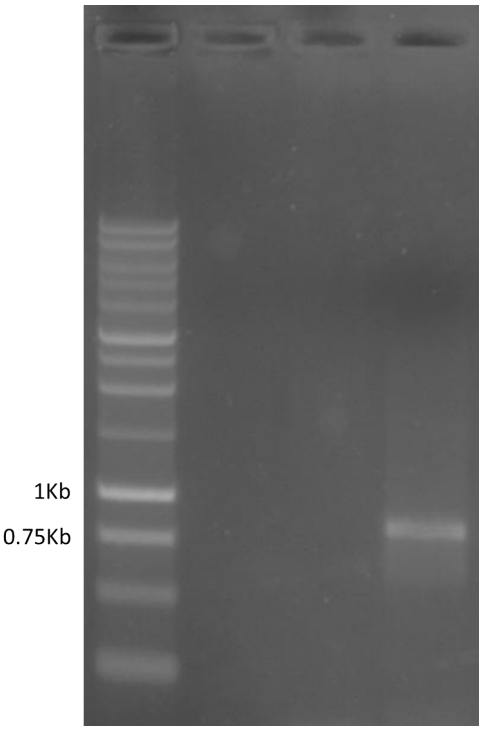


Figure 6.7: PCR product (∼850bp) obtained from *P. superbus* using the primer pairs CYTB2F/CX3R and confirming COB and COXIII gene border.

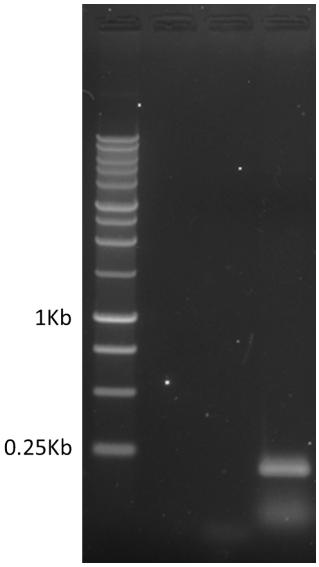


Figure 6.8: PCR product (∼200bp) obtained from *P. superbus* using the primer pairs CYTBF/CYTBR and confirming the COB gene presence.

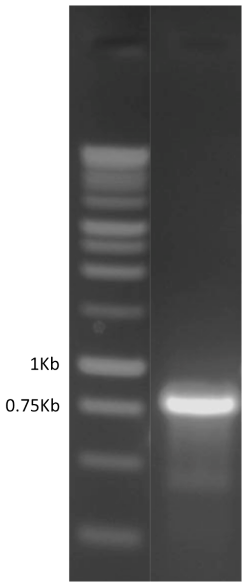


Figure 6.9: PCR product (∼800bp) obtained from *P. superbus* using the primer pairs CX1F/CX2R2 and confirming the COXI and COXII gene border.

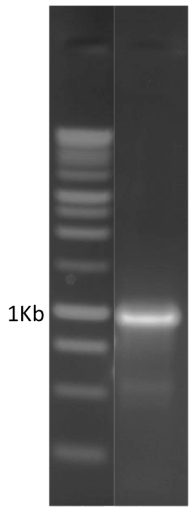


Figure 6.10: PCR product (∼1,000bp) obtained from *P. superbus* using the primer pairs ND2F2/CY2R and confirming the NADH2 and COB gene border.

(see file on the attached CD called mtDNA.fas) and the following genes were identified in contigs from the EST, transcriptome or genomic datasets, or by running tRNAscan:

- Protein coding genes:

  COXI, COXII and COXIII, COB, NADH1, NADH3, NADH4, NADH4L, NADH5, and ATP6.

- Ribosomal RNA genes:

  rrnL and rrnS.

- tRNAs:

  Cys, Met, Asp, Gly, His, Ala, Pro, Val, Trp, Glu, Tyr, Lys, Leu, Ile, Arg, Gln, Phe, Leu2 and Asn.

No homologous sequence was found to either NADH2 or NADH6. The resulting genome was found to have the same gene order as *S. carpocapsae*. It was thus inferred that NADH2 and NADH6 were in the same position in the genome as in

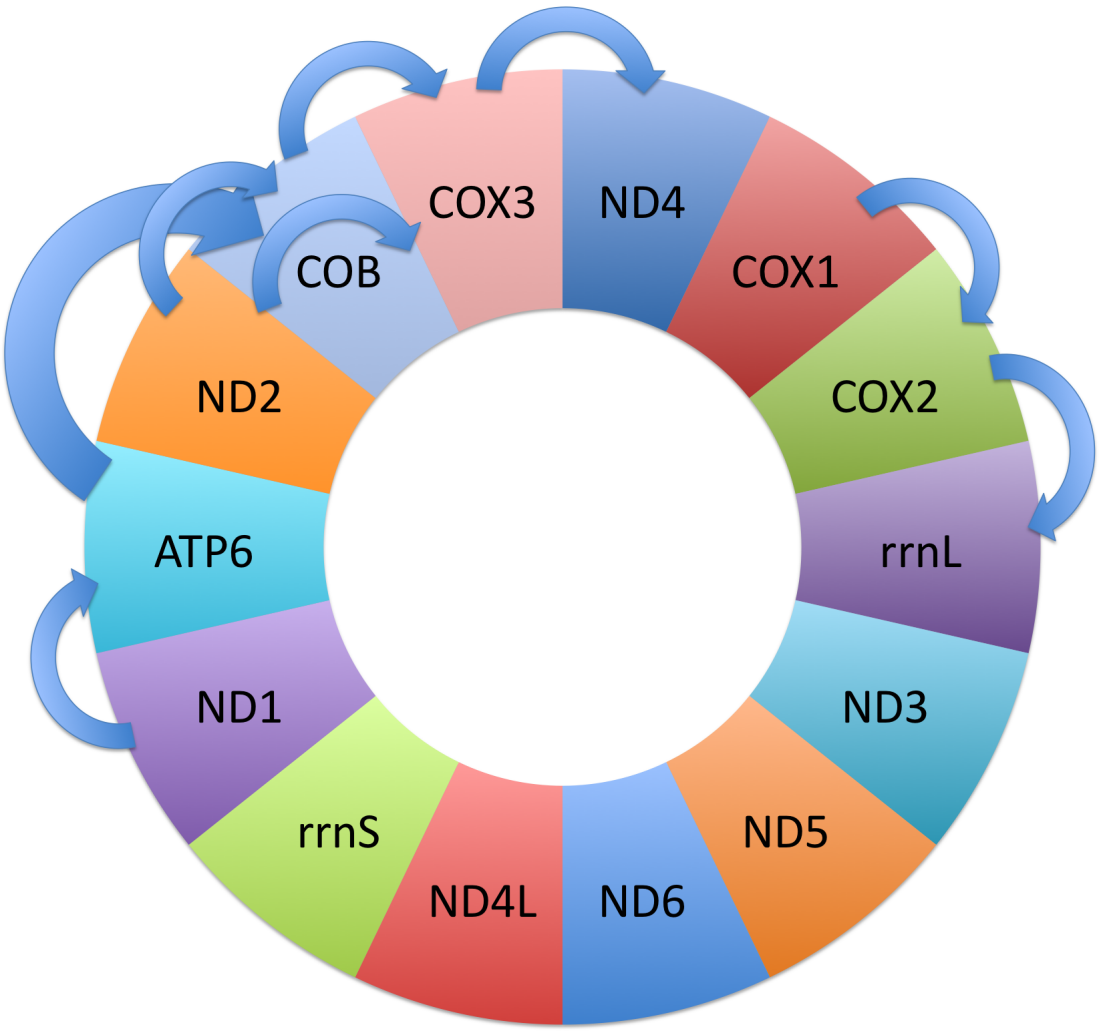


Figure 6.11: Nematode mtDNA molecule showing the locations of the successful PCR amplifications for the *P. superbus* mtDNA genome.

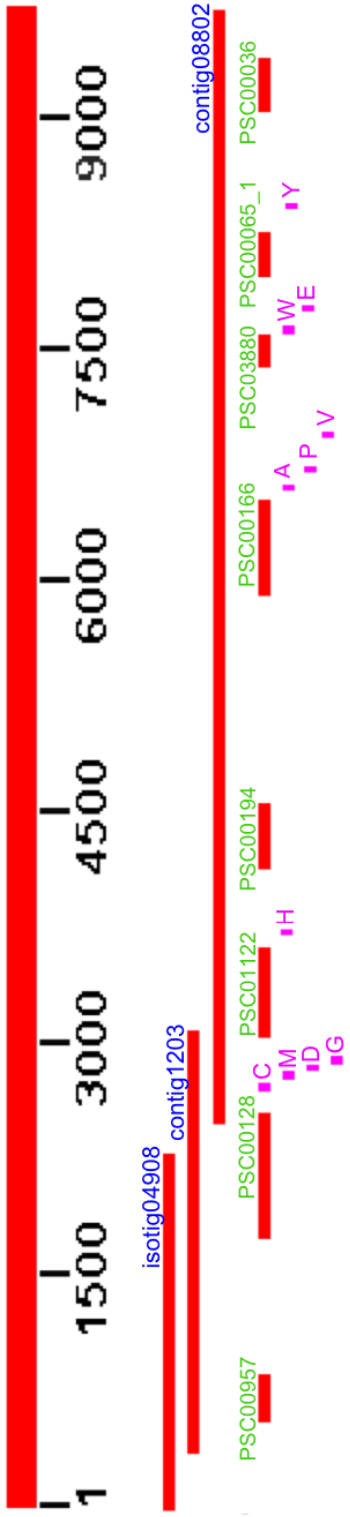


Figure 6.12: *P. superbus* mtDNA fragment showing the overlapping sequences from various datasets used in its assembly labelled in blue. Also shown are the ESTs used to identify these sequences as mitochondrial labelled in green and the relevant tRNAs found post assembly labelled in pink.

S. carpocapsae but were significantly different in gene sequence, thus, no homologous hits found. A diagrammatic representation of the genome can be seen in Figure 6.13.

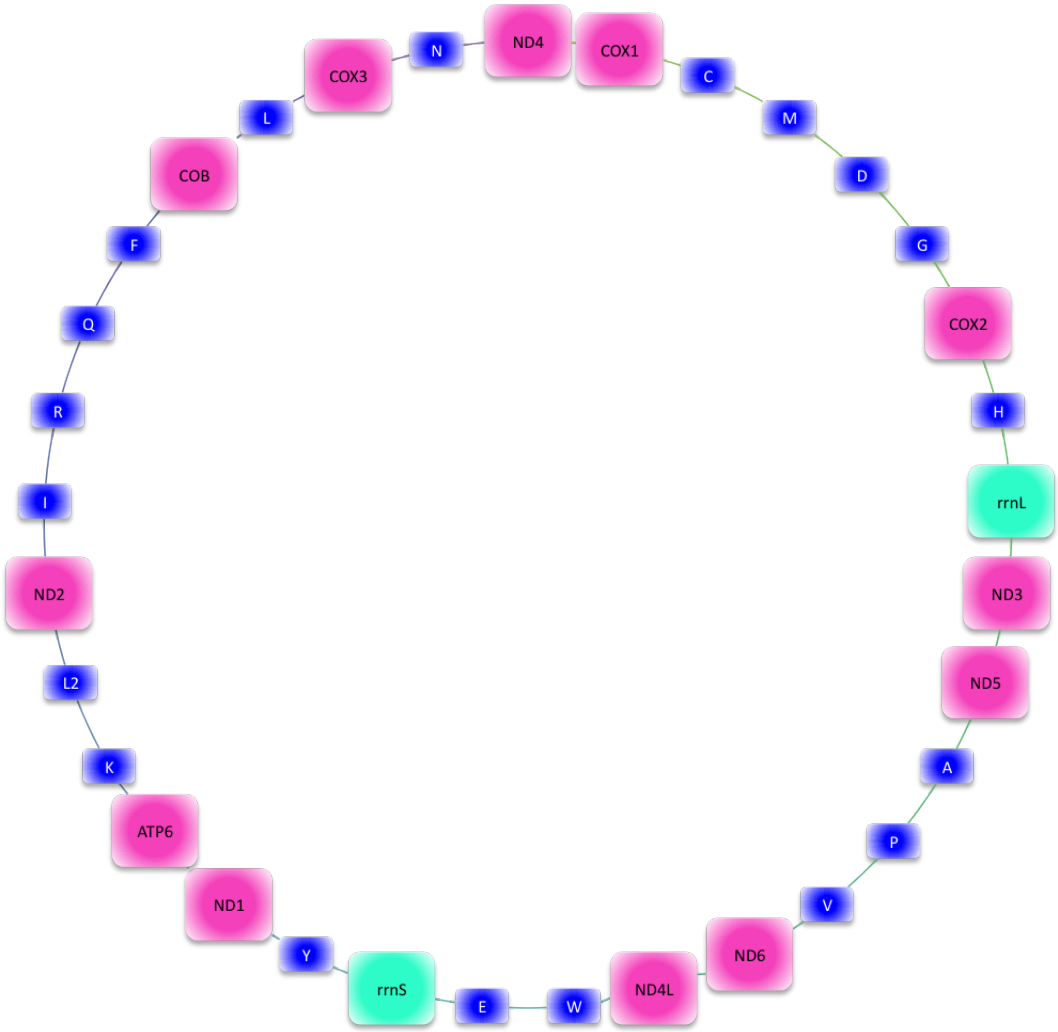


Figure 6.13: The mtDNA genome of *P. superbus* with protein coding genes shown in pink, rRNA genes shown in green and tRNAs shown in blue.

6.4 Discussion

The similarity in the gene order of *P. superbus* and *S. carpocapsae* fits with the evolutionary proximities of these two nematodes on the phylogenetic tree, as can be seen in Figure 1.1. It is interesting that *Strongyloides* species are placed on the same branch as *Panagrolaimus*, but yet, in terms of gene order, they differ significantly from that of the other two nematode species. While *Panagrolaimus* is free-living, *Strongyloides* is a vertebrate parasite and *Steinernema* is an entomopathogen. This could be a suggested reason for the changes in gene order. Enoplean nematodes usually have unpatterned gene orders while chromadorea nematodes like *Panagrolaimus*, usually have fixed gene orders (Montiel et al., 2006). The various gene orders can be seen in Figure 6.14. *P. superbus* has a G6/G7 formation while *S. stercoralis* has a G3 formation. A suggested reason for this is that the phylogeny was constructed using partial genomes or singular genes so the order in which the phylogeny is presented may need further analysis to confirm true positioning (Hyman et al., 2011). With advances in high-throughput next-generation sequencing, and the copy number of mitochondria in the cell, full mitochondrial genomes are being assembled amongst genomic datasets, and so a more thorough and complete phylogeny may soon be available.

When the completed genome was assembled, the primer sequences were aligned and the bases between primers were counted. Fifty percent of the genome had been verified by PCR, which thus establishes the computational approach as a more thorough, money-saving and time-saving method to establish the DNA sequence of mtDNA genomes. In this chapter, the complete mtDNA genome of *P. superbus*, which has a gene order most similar to that of *S. carpocapsae*, was presented. Future work could include phylogenetic studies of individual genes and whole genome alignment with other nematodes to establish a more complete phylogenetic tree of mitochondrial nematode genomes. Additional PCRs should be attempted to

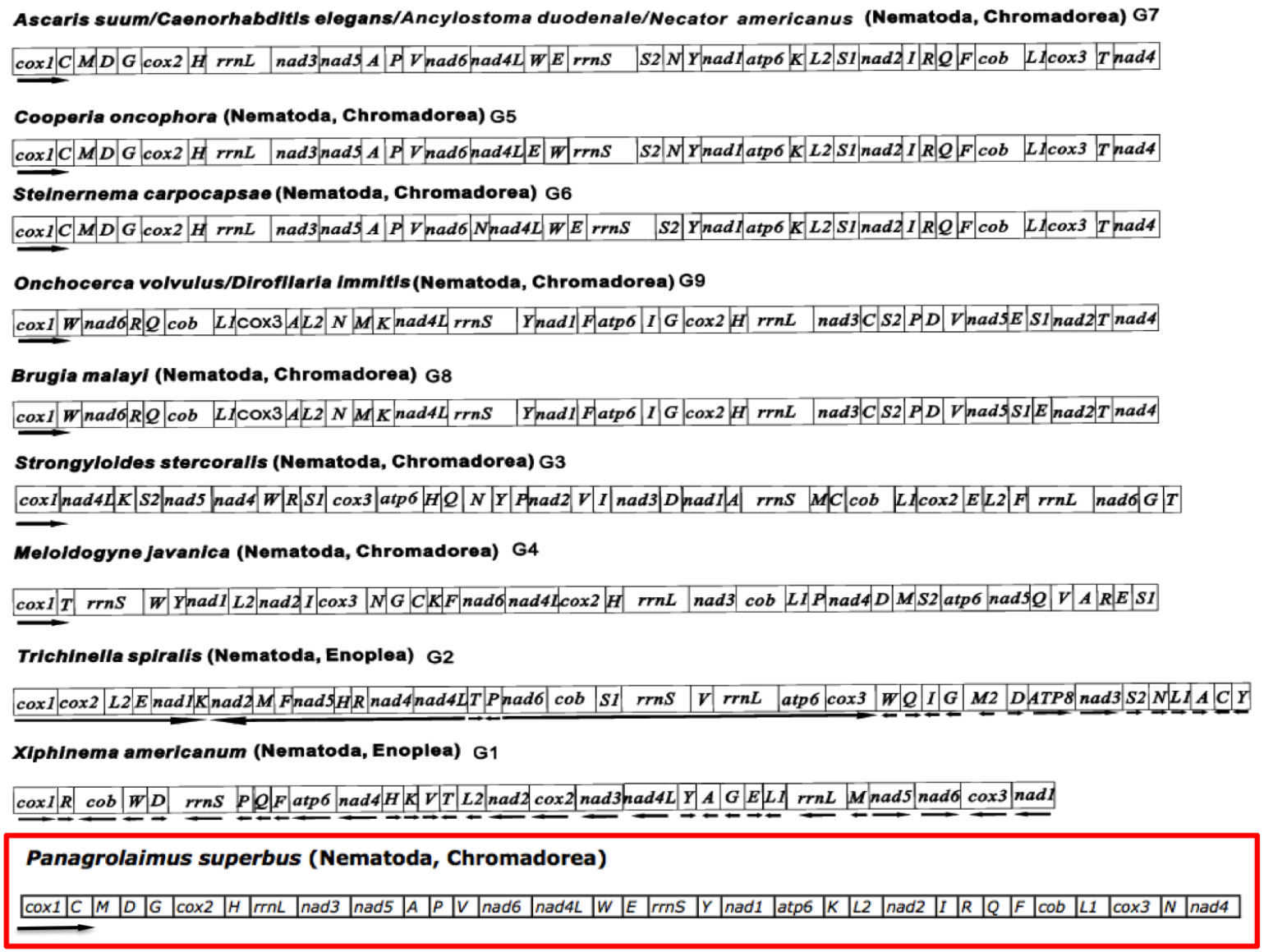


Figure 6.14: The breakdown of gene orders from nematode mitochondrial genomes with *P.superbus* shown in a red box. Modified from He et al. (2005)

resequence the NADH6 and NADH2 genes to confirm the hypothesis about their position with the genome.

This study is unique in that a predominantly bioinformatics based approach was used to find the genes of interest and, using alignments, these sequences were mapped to other previously sequenced nematodes and a gene order was thus established.

Chapter 7

General Discussion

7.1 Discussion

Next generation high-throughput sequencing has revolutionised several areas of biological research with particular advances being made in genomics and transcriptomics, cell biology and molecular medicine. Beginning with the sequencing of model organisms (Lander & International, 2001; The C. elegans Sequencing Consortium, 1998), the advances that have been made in the subsequent 15 years are substantial: new ways of generating samples for sequencing, the sequencing technologies themselves and, of course, the computational methods developed to investigate the sequences generated.

The high-throughput sequencing technology offered by 454(Roche) was first released in 2005, followed by the Illumina sequencing platform, released in 2006. Initially the Illumina reads were short, but accurate. However the current Illumina technology can generate reads up to 200 bp [1]. The demand for high throughput, low cost "second generation" sequencing machines has led to several other biotech companies entering the market. The SOLiD short read system (released 2006)

[1] (*http://www.illumina.com/Documents/systems/hiseq/datasheet_hiseq_systems.pdf*)

and Ion Torrent mid read system (released 2010) are both from Applied Biosystems/Life Technologies (reviewed by Liu et al. (2012)). Recently, a TGS technology capable of much longer reads has become available. The PacBio RS system can yield reads of average length over 2,500 bp and some longer reads can reach 10,000bp (Mason & Elemento, 2012). However, these continuous long reads tend to have a high error rate of up to 15% (Au et al., 2012). According to Eisenstein (2012) the current generation of 454 FLX machines are still being used in many sequencing centers, generating individual sequence reads that "routinely exceed those generated by most other sequencers". However, Eisenstein estimates that Illumina "now controls roughly 60% of the multibillion dollar sequencing technology sector".

With high-throughput technologies now capable of sequencing billions of bases in a single run, these technologies present computational challenges. Reliable computational tools and infrastructures are required to accurately assemble genomes and transcriptomes from these high throughput reads and to interpret the assembled sequence data. The relatively low cost and ready access to high throughput sequencing have greatly increased the applications and scope of DNA sequencing enabling researchers to investigate a great diversity of phenomena in human disease, organismal and cellular biology and environmental science (reviewed by Shendure et al. (2004)). Although basic research is still the major market for high-throughput sequencing, it is generally expected that the clinical diagnostics market is soon likely to be the biggest consumer of this technology. For example, sequencing based characterisation of cancer genomes and transcriptomes can aid in the identification of specific cancer types (Slack & Gascoyne, 2013) and this can be used to guide patient treatment. Transcriptome profiling of an individual adenocarcinoma found that the pattern of gene expression suggested the tumor cells were driven by the RET oncogene. This led to the successful clinical decision

to treat the patient with RET inhibitors (Jones et al., 2010).

An important application of high-throughput sequencing technologies is in the area of *de novo* sequencing of the transcriptomes and genomes of non model organisms, the subject of this thesis. The cost of the technology is gradually decreasing, making it possible to sequence the genome or transcriptome of any organism of choice. While model organisms, and previously sequenced species have their own set of challenges, non-model *de novo* sequencing can be particularly difficult. Without a completed reference genome to align high throughput reads, it can be challenging to know how much of the complete sequence has been generated. High-throughput sequencing allows for mass over sequencing of a dataset to ensure a good chance of sequencing all bases at least once. A large number of previously unstudied genomes and transcriptomes have been published in the last three years and there is an increasing emphasis on development of new technologies and algorithms to deal with the ever expanding databases of sequencing information being generated.

The advances in next generation technology now allow nematologists to investigate a variety of topics such as nematode diversity, parasitism and evolution. The data being generated for the Phylum is expanding rapidly, with a strong emphasis on data from parasitic nematodes. Recent publications include the genome sequences of the pinewood nematode *Bursaphelenchus xylophilus* (Kikuchi et al., 2011), the heart-worm *Dirofilaria immitis* (Godel et al., 2012) and the free-living nematode *Panagrellus redivivus* (Srinivasan et al., 2013). Recently published nematode transcriptomes include those from the following parasitic nematodes: *Trichostrongylus colubriformis* (Cantacessi et al., 2010), *Pratylenchus coffeae* (Haegeman et al., 2011), *Trichuris suis* (Cantacessi et al., 2011), *Ascaris suum* (Ma et al., 2011), *Dirofilaria immitis* (Fu et al., 2012), *Trichinella spiralis* (Liu et al., 2012) and *Strongyloides venezuelensis* (Nagayasu et al., 2013). Thus far, there have been

no publications of *de novo* assembled transcriptomes of free living nematodes.

In this study four datasets for the free living nematode *Panagrolaimus superbus* are presented. In Chapter 2, a Sanger sequenced EST dataset was discussed. This dataset offers a highly annotated summary of some *P. superbus* genes with potential roles in anhydrobiosis. Chapter 3 gives an account of the high throughput sequencing of the *P. superbus* transcriptome following exposure of the nematodes to various environmental stresses (heat, cold, oxidation and desiccation). A great deal of time and effort was put into choosing the best assembly for the data set and it was found that different assembly programs generated widely varying assembly and "quality" metrics. During the time these transcriptome sequences were being assembled, new assembly algorithms were being developed and the frequency with which new versions of assembly software (particularly Newbler software) were released delayed the generation of the final version of the *P. superbus* transcriptome assembly. Two assemblies, CAP3 v2012 and Newbler 2.6 without URT Isotigs, achieved a substantially better rank score than the other assemblies which were evaluated. Thus, either of these transcriptomes could be recommended/selected for downstream annotation of the *P. superbus* transcriptome. Analysis of their contig metrics show that the main difference between these two assemblies is the larger number of contigs in the CAP3 assembly (31,836) as compared to the Newbler isotigs assembly (14,960); the mean contig length is longer for the Newbler isotigs assembly, but the number of contigs >1kb is larger for the CAP3 assembly (9,350) than for the Newbler isotigs assembly (6,866) and the coverage of the 5' end of the *C. elegans* genes is slightly higher for the CAP3 assembly (56%) than for the Newbler isotigs assembly (52%).

Transcriptome sequencing allows a glimpse of expressed genes during a particular point in the life cycle or adaptive state of an organism. Chapter 4 presents an overview of the stress transcriptome of *P. superbus*. The Newbler assembly

comprising 14,960 isotigs was selected for this analysis because it was considered that this assembly allows for a conservative estimate of gene numbers without the doubt of over assembly and thus false positives. In this dataset, 465 sequences were identified as having a hit to the Gene Ontology term GO:0006950 ("response to stress"). This is slightly over 5% of all annotated sequences and is subsequently more than the 187 unigenes identified in Chapter 2 in the *P. superbus* EST dataset. In Chapter 5, the nuclear genome following sequencing by 454 (Roche) and Illumina platforms was put forward as a dataset to allow future assembly and annotation. From the genome assemblies generated in this project, an estimate could be made that the *P. superbus* genome is between 100 - 150Mb, which fits well with the other data available for other nematode genomes. The haploid chromosome number of *P. superbus* was determined as n=4. Chapter 6 describes the work done to obtain the complete mtDNA genome of *P. superbus*. A combination of conventional PCR experiments was used together with computational approaches employing high-throughput cDNA and gDNA sequences. While nematode mtDNA genes have a relatively conserved sequence amongst members of the phylum, the order of these genes in the circular genome varies between different nematode genera, even among closely related species (Hyman et al., 2011). This makes each nematode mtDNA genome sequencing project a challenge. The mtDNA genome is very AT rich, with only 30.7% GC content and this adds to the challenge of finding suitable primers and achieving successful PCRs. The mtDNA genome of *P. superbus*, when assembled was found to have a gene order most similar to that of *Steinernema carpocapsae*. This fits with the evolutionary proximities of these two nematodes on the phylogenetic tree, as can be seen in Figure 1.1.

The datasets presented here can be used as a stepping stone towards generating a fully annotated and complete genome sequence for *P. superbus*. These current draft assemblies are already suitable as references in gene discovery studies using

RNA Seq and high throughput proteomics as described below.

7.2 Future Work

The EST dataset described in Chapter 2 has been published Tyson et al. (2012) and these sequences are freely available to download as a complete dataset, or to search having particular queries in mind [2]. This dataset has been used to select target genes for further downstream RNAi work in the laboratory.

The transcriptome dataset described in Chapter 4 is currently being prepared for publication and the associated sequences and assemblies will also be made freely available to the research community. The generation of an unnormalised and normalised transcriptome dataset can potentially answer the question as to whether or not it is beneficial to normalise the dataset when studying a particular treatment, condition or stress. We intend to map the individual reads from the PS1 (unnormalised) and PS2 (normalised) libraries onto the main assemblies to identify up-regulated putative stress response sequences within the PS1 dataset.

The transcriptome dataset is being used in a separate project in the laboratory in an RNA-Seq experiment (using Illumina/Solexa reads) to identify genes up regulated in response to cold and desiccation stress. In this experiment the *P. superbus de novo* transcriptome was used as the reference to which the Illumina short reads were aligned and the statistical tests using the DESeq program (Anders & Huber, 2010) were used to identify up-regulated genes. The transcriptome assembly is currently being used to generate a protein reference database for the mass spectrometry based identification of the proteome of control and desiccated *P. superbus*. The possibility that the current draft genome assembly could also be used to generate a *P. superbus* protein reference database is also being investigated.

[2] ($www.nematodes.org/nembase4/species\_info.php?species = PSC$)

The mtDNA genome is not yet ready for submission to the NCBI database. Further PCR experiments are required to verify the computational work. Long PCR could again be attempted, this time with primers designed from *P. superbus* sequences. Completion of a mtDNA genome, from a free living nematode, from clade IV will greatly add to phylogenetic studies within the phylum and would also have applications in phylogenetic population genetics analyses within the genus *Panagrolaimus*.

Due to technical difficulties in generating the mate pair libraries at the sequencing centre and the large size of the Illumina datasets, the genome assemblies need to be reviewed and possibly reassembled. The computational challenges faced due to large quantities of data mean that software needed for assembly will have to be thoroughly investigated and refined to give optimal performance. It is hoped that the EST, transcriptome and RNA-Seq datasets can be used to aid in scaffolding and generation of a high N50 value for the genome contigs. Downstream annotation for gene finding will result in a valuable nematode genome resource for study of stress biology and comparative genomics.

Cited References

Aalen, R. (1999), 'Peroxiredoxin antioxidants in seed physiology.', *Seed Science Research* **9**, 285–295.

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V. C., Caillaud, M.-C., Coutinho, P. M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J. V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T. R., Markov, G. V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Ségurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T. J., Blaxter, M., Bleve-Zacheo, T., Davis, E. L., Ewbank, J. J., Favery, B., Grenier, E., Henrissat, B., Jones, J. T., Laudet, V., Maule, A. G., Quesneville, H., Rosso, M.-N., Schiex, T., Smant, G., Weissenbach, J. & Wincker, P. (2008), 'Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita.', *Nature Biotechnology* **26**, 909–915.

Adams, M. (2000), 'The genome sequence of Drosophila melanogaster', *Science* **287**, 2185–2195.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. & Moreno, R. F. (1991),

'Complementary DNA sequencing: expressed sequence tags and human genome project.', *Science* **252**, 1651–1656.

Adhikari, B. N., Wall, D. H. & Adams, B. J. (2009), 'Desiccation survival in an Antarctic nematode: molecular analysis using expressed sequenced tags.', *BMC Genomics* **10**, 69.

Adhikari, B., Wall, D. & Adams, B. (2010), 'Effect of slow desiccation and freezing on gene transcription and stress survival of an Antarctic nematode.', *Journal of Experimental Biology* **213**, 1803–1812.

Allen, J. F. (2003), 'The function of genomes in bioenergetic organelles.', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **358**(1429), 19–37; discussion 37–8.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990), 'Basic local alignment search tool.', *Journal of Molecular Biology* **215**, 403–410.

Anders, S. & Huber, W. (2010), 'Differential expression analysis for sequence count data.', *Genome Biology* **11**, R106.

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. & Young, I. G. (1981), 'Sequence and organization of the human mitochondrial genome.', *Nature* **290**, 457–465.

Andrássy, I. (1984), *Klasse Nematoda*, Bestimmungsbücher zur Bodenfauna Europas, Berlin-Stuttgart.

Andres-Mateos, E., Perier, C., Zhang, L., Blanchard-Fillion, B., Greco, T. M., Thomas, B., Ko, H. S., Sasaki, M., Ischiropoulos, H., Przedborski, S., Dawson, T. M. & Dawson, V. L. (2007), 'DJ-1 gene deletion reveals that DJ-1 is an

atypical peroxiredoxin-like peroxidase.', *Proceedings of the National Academy of Sciences of the United States of America* **104**, 14807–14812.

Arisz, S., Testerink, C. & Munnik, T. (2009), 'Plant PA signaling via diacylglycerol kinase.', *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biololgy of Lipids* **1791**, 869–875.

Aroian, R. V., Carta, L., Kaloshian, I. & Sternberg, P. W. (1993), 'A Free-living Panagrolaimus sp. from Armenia Can Survive in Anhydrobiosis for 8.7 Years.', *Journal of Nematology* **25**, 500–502.

Asahina, E. (1959), 'Frost-resistance in a nematode, Aphelenchoides ritzemabosi', *Low Temperature Science* **17**, 51–62.

Au, K. F., Underwood, J. G., Lee, L. & Wong, W. H. (2012), 'Improving PacBio long read accuracy by short read alignment.', *PLoS ONE* **7**(10), e46679.

Bagniewska-Zadworna, A. (2008), 'The root microtubule cytoskeleton and cell cycle analysis through desiccation of Brassica napus seedlings.', *Protoplasma.* **233**, 177–185.

Bairoch, A. (2000), 'The ENZYME database in 2000.', *Nucleic Acids Research* **28**, 304–305.

Balzer, S., Malde, K., Lanzén, A., Sharma, A. & Jonassen, I. (2010), 'Characteristics of 454 pyrosequencing data–enabling realistic simulation with flowsim.', *Bioinformatics (Oxford, England)* **26**(18), i420–425.

Barrell, B. G., Anderson, S., Bankier, a. T., de Bruijn, M. H., Chen, E., Coulson, a. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. a., Sanger, F., Schreier, P. H., Smith, a. J., Staden, R. & Young, I. G. (1980), 'Different pattern of codon

recognition by mammalian mitochondrial tRNAs.', *Proceedings of the National Academy of Sciences of the United States of America* **77**(6), 3164–3166.

Barrett, J. & Butterworth, P. (1985), 'DNA stability in the anabiotic fourth-stage juveniles of Ditylenchus dipsaci (Nematoda)', *Annals of Applied Biology* **106**, 121–124.

Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. & Clayton, D. A. (1981), 'Sequence and gene organization of mouse mitochondrial DNA.', *Cell* **26**(2 Pt 2), 167–180.

Bies-Ethève, N., Gaubier-Comella, P., Debures, A., Lasserre, E., Jobet, E., Raynal, M., Cooke, R. & Delseny, M. (2008), 'Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in Arabidopsis thaliana.', *Plant Molecular Biology* **67**, 107–124.

Blanchard, J. & Lynch, M. (2000), 'Organellar genes: why do they end up in the nucleus?', *Trends in Genetics* **16**, 315 – 320.

Blattner, F. (1997), 'The complete genome sequence of Escherichia coli K-12.', *Science* **277**, 1453–1462.

Blaxter, M. L. (2012), 'Genomes and Genomics 3 2011 - 2012'.

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T. & Thomas, W. K. (1998), 'A molecular evolutionary framework for the phylum Nematoda.', *Nature* **392**, 71–75.

Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thirry-Mieg, D., Wei, L. C., Duke, K., Kiraly, M. & Kim, S. K. (2002), 'A

global analysis of Caenorhabditis elegans operons.', *Letters to Nature* **417**, 851–854.

Blumenthal, T., Squire, M., Kirtland, S., Cane, J., Donegan, M., Spieth, J. & Sharrock, W. (1984), 'Cloning of a yolk protein gene family from Caenorhabditis elegans.', *Journal of Molecular Biology* **174**, 1–18.

Bonifati, V., Rizzu, P., Squitieri, F., Krieger, E., Vanacore, N., van Swieten, J. C., Brice, A., van Duijn, C. M., Oostra, B., Meco, G. & Heutink, P. (2003), 'DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism.', *Neurological Sciences : Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* **24**, 159–160.

Boore, J. L. (1999), 'Animal mitochondrial genomes.', *Nucleic Acids Research* **27**(8), 1767–1780.

Borgonie, G., García-Moyano, A., Litthauer, D., Bert, W., Bester, A., van Heerden, E., Möller, C., Erasmus, M. & Onstott, T. C. (2011), 'Nematoda from the terrestrial deep subsurface of South Africa.', *Nature* **474**(7349), 79–82.

Bostrom, S. (1988), 'Descriptions and morphological variability of three populations of Panagrolaimus fuchs, 1930 (Nematoda: Panagrolaimidae)', *Nema* **34**, 144–155.

Bostrom, S. (1995), 'Populations of Plectus acuminatus Bastian , 1865 and Panagrolaimus magnivulvatus n . sp . (Nematoda) from nunatakks in Dronning Maud Land , East Antarctica', *Fundamental and Applied Nematology* **18**, 25–34.

Brenner, S. (1974), 'Caenorhabditis elegans', *Genetics* pp. 71–94.

Brewster, J. J., de. Valoir, T., Dwyer, N., Winter, E., Gustin, M. & Devaloir, T. (1993), 'An osmosensing signal transduction pathway in yeast.', *Science* **259**, 1760–1763.

Brown, I. M., Wharton, D. a. & Millar, R. B. (2004), 'The influence of temperature on the life history of the Antarctic nematode Panagrolaimus davidi', *Nematology* **6**, 883–890.

Browne, J., Tunnacliffe, A. & Burnell, A. (2002), 'Plant desiccation gene found in a nematode', *Nature* **416**, 38.

Burge, C. & Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA.', *Journal of Molecular Biology* **268**, 78–94.

Burger, G., Plante, I., Lonergan, K. M. & Gray, M. W. (1995), 'The mitochondrial DNA of the amoeboid protozoon, Acanthamoeba castellanii: complete sequence, gene content and genome organization.', *Journal of Molecular Biology* **245**, 522–537.

Burmeister, H. (1837), *Handbuch der Naturgeschicht. Zum Gebrauch bei Vorlesungenentworfen.*, Zoologie, Berlin.

Cantacessi, C., Mitreva, M., Campbell, B. E., Hall, R. S., Young, N. D., Jex, A. R., Ranganathan, S. & Gasser, R. B. (2010), 'First transcriptomic analysis of the economically important parasitic nematode, Trichostrongylus colubriformis, using a next-generation sequencing approach.', *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* **10**, 1199–1207.

Cantacessi, C., Young, N. D., Nejsum, P., Jex, A. R., Campbell, B. E., Hall, R. S., Thamsborg, S. M., Scheerlinck, J.-P. & Gasser, R. B. (2011), 'The Transcrip-

tome of Trichuris suis First Molecular Insights into a Parasite with Curative Properties for Key Immune Diseases of Humans', *PLoS ONE* **6**, 10.

Casjens, S. (1998), 'The diverse and dynamic structure of bacterial genomes', *Annual Review of Genetics* **1998**, 339–377.

Castellana, N. & Bafna, V. (2010), 'Proteogenomics to discover the full coding content of genomes: A computational perspective', *Journal of Proteomics* **73**, 2124–2135.

Chain, P., Field, D., Snape, J., Tiwari, B., Service, S. D., Sansone, S. A., Quackenbush, J., Giles, J. & Lau, F. (2009), 'Genome Project Standards in a New Era of Sequencing', *Science* **326**, 4–5.

Chakrabortee, S., Boschetti, C., Walton, L. J., Sarkar, S., Rubinsztein, D. C. & Tunnacliffe, A. (2007), 'Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function.', *Proceedings of the National Academy of Sciences of the United States of America* **104**, 18073–18078.

Chang, Q. & Petrash, J. (2008), 'Disruption of aldo-keto reductase genes leads to elevated markers of oxidative stress and inositol auxotrophy in Saccharomyces cerevisiae.', *Biochimica et Biophysica Acta - Molecular and Cell Biology Research* **1783**, 237–245.

Cheung, W. L., Turner, F. B., Krishnamoorthy, T., Wolner, B., Ahn, S.-H., Foley, M., Dorsey, J. a., Peterson, C. L., Berger, S. L. & Allis, C. D. (2005), 'Phosphorylation of histone H4 serine 1 during DNA damage requires casein kinase II in S. cerevisiae.', *Current Biology* **15**, 656–660.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E. G., Wetter, T. & Suhai, S. (2004), 'Using the miraEST assembler for reliable and automated

mRNA transcript assembly and SNP detection in sequenced ESTs.', *Genome Research* **14**, 1147–1159.

Chitwood, B. G. (1957), 'The English Word "Nema" Revised', *Systematic Zoology* **6**, 184.

Cho, E. & Choi, Y. (2009), 'A nuclear-localized HSP70 confers thermoprotective activity and drought-stress tolerance on plants.', *Biotechnology Letters* **31**, 597–606.

Choe, K. & Strange, K. (2007), 'Evolutionarily conserved WNK and Ste20 kinases are essential for acute volume recovery and survival after hypertonic shrinkage in Caenorhabditis elegans.', *American Journal of Physiology - Cell Physiology* **293**, C915–C927.

Choe, K. & Strange, K. (2008), 'Genome-wide RNAi screen and in viv protein aggregation reporters identify degradation of damaged proteins as an essential hypertonic stress response.', *American Journal of Physiology - Cell Physiology* **295**, C1488–C1498.

Clary, D. O. & Wolstenholme, D. R. (1985), 'The mitochondrial DNA molecular of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code.', *Journal of Molecular Evolution* **22**, 252–271.

Cobb, N. A. (1919), 'The orders and classes of nemas', *Ibid.* **8**, 213–216.

Coghlan, A. (2005), 'Nematode genome evolution.', *WormBook : the Online Review of C. elegans Biology* pp. 1–15.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J. & Talón, M. (2005), 'Blast2GO: A universal tool for annotation and visualization in functional genomics research', *Bioinformatics* **21**, 3674–3676.

Consortium, T. G. O. (2000), 'Gene Ontology : tool for the unification of biology.', *Gene Expression* **25**, 25–29.

Corona, M. & Robinson, G. (2006), 'Genes of the antioxidant system of the honey bee: annotation and phylogeny.', *Insect Molecular Biology* **15**, 687–701.

Crick, F., Brenner, S., Watstobi, R. & Barnett, L. (1961), 'General nature of the genetic code for proteins', *Nature* **192**, 1227 – 1232.

Crosland, M. W. J. & Crozier, R. H. (1986), 'Myrmecia pilosula, an ant with only one pair of chromosomes', *Science* **231**, 1278.

Crowe, H., Hoekstra, F. A. & Crowe, L. A. (1992), 'Anhydrobiosis', *Annual Review of Physiology* **54**, 579 – 599.

Cuenda, A. & Eousseau, S. (2007), 'MAP-Kinases pathway regulation, function and role in human diseases.', *Biochimica et Biophysica Acta - Molecular and Cell Biology Research* **1773**, 1358–1375.

Dalle-Donne, I., Rossi, R., Milzani, A., Di Simplicio, P. & Colombo, R. (2001), 'The actin cytoskeleton response to oxidants: From small heat shock protein phosphorylation to changes in the redox state of actin itself.', *Free Radical Biology and Medicine* **31**, 1624–1632.

De Bary, A. (1879), 'Die erscheinung der symbiose', *Verlag VonKarl J.Trubner*.

de Nadal, E. & Alepuz, P. (2002), 'Dealing with osmostress through MAP kinase activation.', *EMBO Reports* **3**, 735–740.

Denekamp, N., Reinhardt, R., Albrecht, M., Drungowski, M., Kube, M. & E., L. (2011), 'The expression pattern of dormancy-associated genes in multiple life-history stages in the rotifer Brachionus plicatilis.', *Hydrobiologia* **662**, 51–63.

Diesing, K. (1860), 'Revision der Nematoden.', *Akad. d. Wissensch., Wien, math.-naturw. Cl.* **42**, 595–736.

Dieterich, C., Clifton, S. W., Schuster, L. N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte, H., Yang, S. P., Wilson, R. K. & Sommer, R. J. (2008), 'The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism', *Nature Genetics* **40**, 1193–1198.

Dimmock, N., Easton, A. & Leppard, K. (2007), *Introduction to Modern Virology*, 6th edn, Wiley-Blackwell Paperback.

DNA Sequencing Core (2013), 'University of Michigan DNA Sequencing Core'.

Dorris, M., Ley, P. D. & Blaxter, M. L. (1999), 'Molecular Analysis of Nematode Diversity and the Evolution of Parasitism', *Parasitology* **15**, 188–193.

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005), 'IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.', *Bioinformatics (Oxford, England)* **21**(16), 3433–4.

Dubreuil, G., Deleury, E., Magliano, M., Jaouannet, M., Abad, P. & Rosso, M. (2011), 'Peroxiredoxins from the plant parasitic root-knot nematode. Meloidogyne incognita. are required for successful development within the host.', *International Journal of Parasitology* **41**, 385–396.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. (2002), 'Intrinsic Disorder and Protein Function', *Biochemistry* **41**, 6573–6582.

Dure, L. (1993), 'A repeating 11-mer amino acid motif and plant desiccation.', *The Plant Journal : for Cell and Molecular Biology* **3**, 363–369.

Dyson, H. & Wright, P. (2005), 'Intrinsically unstructured proteins and their functions.', *Nature Reviews Molecularal and Cellular Biology* **6**, 197–208.

Edwards, D., Batley, J. & Snowdon, R. J. (2013), 'Accessing complex crop genomes with next-generation sequencing', *Theoretical Applied Genetics* **126**, 1–11.

Eisenstein, M. (2012), 'The battle for sequencing supremacy.', *Nature Biotechnology* **30**, 1023–1026.

Elsworth, B., Wasmuth, J. & Blaxter, M. (2011), 'NEMBASE4: the nematode transcriptome resource.', *International Journal of Parasitology* **41**, 881–94.

Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002), 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Research* **30**, 1575–1584.

Eyles, S. J. & Gierasch, L. M. (2010), 'Nature's molecular sponges: small heat shock proteins grow into their chaperone roles.', *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2727–2728.

Ferrigno, P., Posas, F., Koepp, D., Saito, H. & Silver, P. (1998), 'Regulated nucleo/cytoplasmic exchange of HOG1 MAPK requires the importin beta homologs NMD5 and XPO1.', *EMBO Journal* **17**, 5606–5614.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L. & Bateman, A. (2008), 'The Pfam protein families database.', *Nucleic Acids Research* **36**, D281–288.

Fleischmann, R. D. & Adams, M. D. (1995), 'Whole-Genome Random Sequencing and Assembly of Haemophilus-Influenzae', *Science* **269**, 496–512.

Forster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., Schnölzer, M., Frohme, M., Müller, T., Schill, R. O. & Dandekar, T. (2009), 'Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades.', *BMC Genomics* **10**, 469.

Foury, F., Roganti, T., Lecrenier, N. & Purnelle, B. (1998), 'The complete sequence of the mitochondrial genome of Saccharomyces cerevisiae.', *FEBS letters* **440**, 325–331.

Franca, M., Panek, A. & Eleutherio, E. (2007), 'Oxidative stress and its effects during dehydration.', *Comparative Biochemistry and Physiology - Part A: Molecular and Integrative Physiology* **146**, 621–631.

Fraser, C. M., Gocayne, J. D. & White, O. (1995), 'The Minimal Gene Complement of Mycoplasma-Genitalium', *Science* **270**, 397–403.

Frederiksen, H. B., Kraglund, H.-O. & Ekelund, F. (2001), 'Microfaunal primary succession on the volcanic island of Surtsey, Iceland', *Polar Research.*

Friar, J. L., Goldman, T. & Perez-Mercader, J. (2012), 'Genome Sizes and the Benford Distribution', *PLoS ONE* **7**, e36624.

Fu, Y., Lan, J., Zhang, Z., Hou, R., Wu, X., Yang, D., Zhang, R., Zheng, W., Nie, H., Xie, Y., Yan, N., Yang, Z., Wang, C., Luo, L., Liu, L., Gu, X., Wang, S., Peng, X. & Yang, G. (2012), 'Novel insights into the transcriptome of Dirofilaria immitis.', *PLoS ONE* **7**, e41639.

Fuchs, G. (1930), 'Neue an Borkenkafer und Russelkafer Gebundene Nematoden, Halbparasitische und Wohnuungseinmieter', *Zoologische Jahrbucher* **59**, 586–608.

Fujiwara, R. T., Zhan, B., Mendez, S., Loukas, A., Bueno, L. L., Wang, Y., Plieskatt, J., Oksov, Y., Lustigman, S., Bottazzi, M. E., Hotez, P. & Bethony, J. M. (2007), 'Reduction of worm fecundity and canine host blood loss mediates protection against hookworm infection elicited by vaccination with recombinant Ac-16.', *Clinical and Vaccine Immunology : CVI* **14**, 281–287.

Fukunishi, Y. & Hayashizaki, Y. (2012), 'Amino acid translation program for full-length cDNA sequences with frameshift errors Amino acid translation program for full-length cDNA sequences with frameshift errors', *Physiological Genomics* pp. 81–87.

Gal, T., Glazer, I. & Koltai, H. (2003), 'Differential gene expression during desiccation stress in the insect-killing nematode Steinernema feltiae IS-6.', *Parasitology* **89**, 761–766.

Gallin, M. Y., Tan, M., Kron, M. A., Rechnitzer, D., Bruce, M., Newland, H. S., White, A. T., Taylor, H. R. & Unnasch, T. R. (1989), 'Onchocerca volvulus Recombinant Antigen : Physical Characterization with Serum Clinical Correlates Reactivity', *Journal of Infectious Diseases* **160**, 521–529.

Gallogly, M. & Mieyal, J. (2007), 'Mechanisms of reversible protein glutathionylation in redox signaling and oxidative stress.', *Current opinion in Pharmacology* **7**, 381–391.

Garay-Arroyo, A., Colmenero-Flores, J., Garciarrubio, A. & Covarrubias, A. (2000), 'Highly hydrophilic proteins in prokaryotes and eukaryotes are common during conditions of water deficit.', *Journal of Biological Chemistry* **275**, 5668–5674.

Gasteiger, E. (2003), 'ExPASy: the proteomics server for in-depth protein knowledge and analysis', *Nucleic Acids Research* **31**, 3784–3788.

Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L., Guiliano, D. B., Miranda-Saavedra, D., Angiuoli, S. V., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N. M., Wortman, J. R., Feldblyum, T., Tallon, L., Schatz, M., Shumway, M., Koo, H., Salzberg, S. L., Schobel, S., Pertea, M., Pop, M., White, O., Barton, G. J., Carlow, C. K. S., Crawford, M. J., Daub, J., Dimmic, M. W., Estes, C. F., Foster, J. M., Ganatra, M., Gregory, W. F., Johnson, N. M., Jin, J., Komuniecki, R., Korf, I., Kumar, S., Laney, S., Li, B.-W., Li, W., Lindblom, T. H., Lustigman, S., Ma, D., Maina, C. V., Martin, D. M. a., McCarter, J. P., McReynolds, L., Mitreva, M., Nutman, T. B., Parkinson, J., Peregrín-Alvarez, J. M., Poole, C., Ren, Q., Saunders, L., Sluder, A. E., Smith, K., Stanke, M., Unnasch, T. R., Ware, J., Wei, A. D., Weil, G., Williams, D. J., Zhang, Y., Williams, S. a., Fraser-Liggett, C., Slatko, B., Blaxter, M. L. & Scott, A. L. (2007), 'Draft genome of the filarial nematode parasite Brugia malayi.', *Science (New York, N.Y.)* **317**, 1756–1760.

Godel, C., Kumar, S., Koutsovoulos, G., Ludin, P., Nilsson, D., Comandatore, F., Wrobel, N., Thompson, M., Schmid, C. D., Goto, S., Bringaud, F., Wolstenholme, A., Bandi, C., Epe, C., Kaminsky, R., Blaxter, M. & Mäser, P. (2012), 'The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets.', *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* **26**, 4650–4661.

Goff, S., Ricke, D., Lan, T. & Presting, G. (2002), 'A draft sequence of the rice genome (Oryza sativa L. ssp japonica)', *Science* **296**, 92–100.

Goffeau, A. (1996), 'Life with 6000 genes', *Science* **274**, 563–567.

Goldstein, P. & Wharton, D. (1996), 'The synaptonemal complexes of the meiotic parthenogenetic Antarctic nematode Panagrolaimus davidi: karyotype analysis and three-dimensional reconstruction of pachytene nuclei.', *Cytobios* **85**, 81–90.

Goyal, K., Tisi, L., Basran, A., Browne, J., Burnell, A., Zurdo, J. & Tunnacliffe, A. (2003), 'Transition from natively unfolded to folded state induced by desiccation in an anhydrobiotic nematode protein.', *The Journal of Biological Chemistry* **278**, 12977–12984.

Gray, M. W. (1999), 'Mitochondrial Evolution', *Science* **283**, 1476–1481.

Green, P. (2012), 'Phrap Assembler'.

Gregory, T. R., Nicol, J., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., Murray, B., Kapraun, D. F., Greilhuber, J. & Bennett, M. (2007), 'Eukaryotic genome size databases', *Nucleic Acids Research* **35**, D332–D338.

Griffiths, E. J. (2000), 'Mitochondria–potential role in cell life and death.', *Cardiovascular Research* **46**, 24–27.

Groenen, P. J., Merck, K. B., DeJon, W. W. & Bloemendal, H. (1994), 'Structure and modifications of the junior chaperone alpha-crystallin. From lens transparency to molecular pathology', *European Journal of Biochemistry* **225**, 1 – 19.

Guiliano, D. B. & Blaxter, M. L. (2006), 'Operon conservation and the evolution of trans-splicing in the phylum', *Nematoda. PLoS Genet* **2**, 198.

Gusev, O., Nakahara, Y., Vanyagina, V., Malutina, L., Cornette, R., Sakashita, T., Hamada, N., Kikawada, T., Kobayashi, Y. & Okuda, T. (2010), 'Anhydrobiosis-associated nuclear DNA damage and repair in the sleeping chironomid: linkage with radioresistance.', *PLoS ONE* **5**, e14008.

Haegeman, A., Jacob, J., Vanholme, B., Kyndt, T., Mitreva, M. & Gheysen, G. (2009), 'Expressed sequence tags of the peanut pod nematode Ditylenchus

africanus: the first transcriptome analysis of an Anguinid nematode.', *Molecular and Biochemical Parasitology* **167**, 32–40.

Haegeman, A., Joseph, S. & Gheysen, G. (2011), 'Analysis of the transcriptome of the root lesion nematode Pratylenchus coffeae generated by 454 sequencing technology.', *Molecular and Biochemical Parasitology* **178**, 7–14.

Hand, S. C., Menze, M. a., Toner, M., Boswell, L. & Moore, D. (2011), 'LEA proteins during water stress: not just for plants anymore.', *Annual Review of Physiology* **73**, 115–34.

Harcus, Y. M., Parkinson, J., Fernández, C., Daub, J., Selkirk, M. E., Blaxter, M. L. & Maizels, R. M. (2004), 'Signal sequence analysis of expressed sequence tags from the nematode Nippostrongylus brasiliensis and the evolution of secreted proteins in parasites.', *Genome Biology* **5**, R39.

Hayes, J., Flanagan, J. & Jowsey, I. (2005), 'Glutathione transferases.', *Annual Review of Pharmacology Toxicol* **45**, 51–88.

He, H., Wang, J., Liu, T., Liu, S., Li, T., Wang, Y., Qian, Z., Zheng, H., Zhu, X., Wu, T., Shi, B., Deng, W., Zhou, W., Skogerbo, G. & Chen, R. (2007), 'Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray', *Genome Research* **17**, 1471 –1477.

He, Y., Jones, J., Armstrong, M., Lamberti, F. & Moens, M. (2005), 'The mitochondrial genome of Xiphinema americanum sensu stricto (Nematoda: Enoplea): considerable economization in the length and structural features of encoded genes.', *Journal of Molecular Evolution* **61**, 819–833.

Hegde, A. & Upadhya, S. (2007), 'The ubiquitin-proteasome pathway in health and disease of the nervous system.', *Trends in Neuroscience* pp. 587–595.

Hillier, L. W., Coulson, A., Murray, J. I., Bao, Z., Sulston, J. E. & Waterston, R. H. (2005), 'Genomics in C . elegans : So many genes, such a little worm', *Genome Research* pp. 1651–1660.

Hogg, J. S., Hu, F. F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C. & Ehrlich, G. G. D. (207), 'Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains.', *Genome biology* **8**(6), R103.

Holmgren, A. (1989), 'Electron Transport to Reductive Enzymes', *Biochemistry* **264**, 13963–13966.

Hu, M. (2002), 'Long PCR-based amplification of the entire mitochondrial genome from single parasitic nematodes', *Molecular and Cellular Probes* **16**, 261–267.

Hu, M., Chilton, N. B. & Gasser, R. B. (2003), 'The mitochondrial genome of Strongyloides stercoralis (Nematoda) idiosyncratic gene order and evolutionary implications q', *International Journal for Parasitology* **33**, 1393–1408.

Hu, T. T., Pattyn, P. & Bakker, E. G. (2011), 'The Arabidopsis lyrata genome sequence and the basis of rapid genome size change', *Nature Genetics* **43**, 476.

Huang, X. (1999), 'CAP3: A DNA Sequence Assembly Program', *Genome Research* **9**, 868–877.

Hunault, G. & Jaspard, E. (2010), 'LEAPdb: a database for the late embryogenesis abundant proteins.', *BMC Genomics* **11**, 221.

Hundertmark, M. & Hincha, D. (2008), 'LEA (Late Embryogenesis Abundant) proteins and their encoding genes in Arabidopsis thaliana.', *BMC Genomics* **9**, 22.

Hyman, B. C., Lewis, S. C., Tang, S. & Wu, Z. (2011), 'Rampant gene rearrangement and haplotype hypervariation among nematode mitochondrial genomes.', *Genetica* **139**, 611–5.

ICHEC (2013), 'Irish Centre for High-End Computing'.

Illumina (2013), 'www.illumina.com'.

Irvine, G. B., El-Agnaf, O. M., Shankar, G. M. & Walsh, D. M. (2008), 'Protein aggregation in the brain: the molecular basis for Alzheimer's and Parkinson's diseases.', *Molecular Medicine (Cambridge, Mass.)* **14**, 451–464.

Ito, H., Okamoto, K., Nakayama, H., Isobe, T. & Kato, K. (1997), 'Phosphorylation of alphaB-crystallin in response to various types of stress.', *The Journal of Biological Chemistry* **272**, 29934–29941.

Jacob, J., Vanholme, B., Haegeman, A. & Gheysen, G. (2007), 'Jacob J, Vanholme B, Haegeman A, Gheysen G: Four transthyretin-like genes of the migratory plant-parasitic nematode Radopholus similis: members of an extensive nematode-specific family.', *Gene* **402**, 9–19.

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P. & Gasteiger, E. (2009), 'Infrastructure for the life sciences: design and implementation of the UniProt website.', *BMC Bioinformatics* **10**, 136.

Jex, A. R., Hall, R. S., Littlewood, D. T. J. & Gasser, R. B. (2010), 'An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes.', *Nucleic Acids Research* **38**, 522–533.

Jex, A. R., Hu, M., Littlewood, D. T. J., Waeschenbach, A. & Gasser, R. B. (2008), 'Using 454 technology for long-PCR based sequencing of the complete

mitochondrial genome from single Haemonchus contortus (Nematoda).', *BMC Genomics* **9**, 11.

Jin, Y., Wang, M., Fu, J., Xuan, N., Zhu, Y., Lian, Y., Jia, Z., Zheng, J. & Wang, G. (2007), 'Phylogenetic and expression analysis of ZnF-AN1 genes in plants.', *Genomics* **90**, 265–275.

Jonak, C., Kiegerl, S., Ligterink, W., Barker, P. J., Huskisson, N. S. & Hirt, H. (1996), 'Stress signaling in plants: a mitogen-activated protein kinase pathway is activated by cold and drought.', *Proceedings of the National Academy of Sciences of the United States of America* **93**, 11274–11279.

Jones, J., Smant, G. & Blok, V. C. (2000), 'SXP/RAL-2 proteins of the potato cyst nematode Globodera rostochiensis: secreted proteins of the hypodermis and amphids', *Nematology* **2**, 887–893.

Jones, S. J., Laskin, J., Li, Y. Y., Griffith, O. L., An, J., Bilenky, M., Butterfield, Y. S., Cezard, T., Chuah, E., Corbett, R., Fejes, A. P., Griffith, M., Yee, J., Martin, M., Mayo, M., Melnyk, N., Morin, R. D., Pugh, T. J., Severson, T., Shah, S. P., Sutcliffe, M., Tam, A., Terry, J., Thiessen, N., Thomson, T., Varhol, R., Zeng, T., Zhao, Y., Moore, R. A., Huntsman, D. G., Birol, I., Hirst, M., Holt, R. A. & Marra, M. A. (2010), 'Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors', *Genome Biology* **11**, R82.

Jonsson, K. & Schill, R. (2007), 'Induction of Hsp70 by desiccation. ionising radiation and heat-shock in the eutardigrade Richtersius coronifer.', *Comparative Biochemistry and Physiology - Part B: Molecular and Integrative Physiology* **146**, 456–460.

Jorcano, J. & Ruizcarrillo, A. (1979), 'H3.H4 tetramer directs DNA and core his-

tone octamer assembly in the nucleosome core particle.', *Biochemistry* **18**, 768–774.

Kabani, M. & Martineau, C. (2008), 'Multiple Hsp70 isoforms in the eukaryotic cytosol: Mere redundancy or functional specificity?', *Current Genomics* **9**, 338–348.

Kaletta, T. & Hengartner, M. O. (2006), 'Finding function in novel targets : C . elegans as a model organism', *Nature Reviews Drug Discovery* **AOP**(April), 1–12.

Kanehisa, M. & Goto, S. (2000), 'KEGG: kyoto encyclopedia of genes and genomes.', *Nucleic Acids Research* **28**, 27–30.

Karala, A., Psarrakos, P., Ruddock, L. & Klappa, P. (2007), 'Protein disulfide isomerases from C. elegans are equally efficient at thiol-disulfide exchange in simple peptide-based systems but show differences in reactivity towards protein substrates.', *Antioxid Redox Signal* **9**, 1815–1823.

Karim, N., Jones, J. T., Okada, H. & Kikuchi, T. (2009), 'Analysis of expressed sequence tags and identification of genes encoding cell-wall-degrading enzymes from the fungivorous nematode Aphelenchus avenae.', *BMC Genomics* **10**, 525.

Kaul, S. (2000), 'Analysis of the genome sequence of the flowering plant Arabidopsis thaliana', *Nature* **408**, 796 – 815.

Khandelwal, S. (1990), 'Chromosome evolution in the genus Ophioglossum L.', *Botanical Journal of the Linnean Society* **102**, 205–217.

Kikuchi, T., Cotton, J. A. & Dalzell, J. J. (2011), 'Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen Bursaphelenchus xylophilus', *PLOS Pathogens* **7**, e1002219.

Kimble, J. & W.J., S. (1983), 'Tissue-specific synthesis of yolk proteins in Caenorhabditis elegans.', *Developmental Biology* **96**, 189 – 196.

Kirienko, N. V. & Fay, D. S. (2010), 'SLR-2 and JMJC-1 regulate an evolutionarily conserved stress-response network.', *The EMBO Journal* **29**, 727–739.

Knippschild, U., Gocht, A., Wolff, S., Huber, N., Lohler, J. & Stoter, M. (2005), 'The casein kinase I family: participation in multiple cellular processes in eukaryotes.', *Cellular Signalling* **17**, 675–689.

Knudsen, B., Forsberg, R. & Miyamoto, M. M. (2010), 'A Computer Simulator for Assessing Different Challenges and Strategies of de Novo Sequence Assembly', *Genes* **1**, 263–282.

Kovacs, A. & Zhang, H. (2010), 'Role of autophagy in Caenorhabditis elegans.', *FEBS letters* **584**, 1335–1341.

Kozak, M. (1983), 'Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles.', *Microbiological reviews* **47**, 1–45.

Kranner, I. & Birtic, S. (2005), 'A modulating role for antioxidants in desiccation tolerance.', *Integrative and Comparative Biology* **45**, 734–740.

Krogh, A. (1929), 'The Progress of Physiology.', *Science* **70**, 200–204.

Kubota, H. (2009), 'Quality control against misfolded proteins in the cytosol: a network for cell survival.', *Journal of Biochemisty* **146**, 609–616.

Kumar, S. & Blaxter, M. L. (2010), 'Comparing de novo assemblers for 454 transcriptome data', *BMC Genomics* **11**, 571.

Kyte, J. & Doolittle, R. (1982), 'A simple method for displaying the hydropathic character of a protein.', *Journal of Molecular Biology* **21**, 3433–3434.

Lamark, T. & Johansen, T. (2010), 'Autophagy: links with the proteasome.', *Current Opinion in Cell Biology* **22**, 192–198.

Lander, E. & International (2001), 'Initial sequencing and analysis of the human genome', *Nature* **409**(6822), 860–921.

Lee, S., Koh, H., Park, D., Song, B., Huh, T. & Park, J. (2002), 'Cytosolic NADP(+)-dependent isocitrate dehydrogenase status modulates oxidative damage to cells.', *Free Radical Biology and Medicine* **32**, 1185–1196.

Lemire, B. (2005), 'Mitochondrial genetics.', *WormBook : the Online Review of C. elegans Biology* pp. 1–10.

Leterrier, M., Del Rio, L. & Corpas, F. (2007), 'Cytosolic NADP-isocitrate dehydrogenase of pea plants: genomic clone characterization and functional analysis under abiotic stress conditions.', *Free Radical Research* **41**, 191–199.

Lewis, S. C., Dyal, L. A., Hilburn, C. F., Weitz, S., Liau, W. S., LaMunyon, C. W. & Denver, D. R. (2009), 'Molecular evolution in Panagrolaimus nematodes: origins of parthenogenesis, hermaphroditism and the Antarctic species P. davidi', *BMC Evolutionary Biology* **9**, 15.

Li, D., Li, J., S, O., Wu, S., Wang, J., Xu, X., Zhu, Y. & He, F. (2005), 'An integrated strategy for functional analysis in large-scale proteomic research by gene ontology', *Progress in Biochemistry and BioPhysics* **32**, 1026 – 1029.

Liang, P., Amons, R., Macrae, T. H. & Clegg, J. S. (1997), 'Molecular Characterization of a Small Heat Shock/a- Crystallin Protein in Encysted Artemia Embroyos', *Exposure* **232**, 225–232.

Lildballe, D., Pedersen, D., Kalamajka, R., Emmerson, J., Houben, A. & Grasser, K. (2008), 'The expression level of the chromatin-associated HMGB1 protein in-

fluences growth, stress tolerance, and transcriptome in Arabidopsis.', *Molecular Biology* **384**, 9–21.

Lindquist, S. (1988), 'The Heat-Shock Proteins', *Annual Review of Genetics* **22**, 631 – 677.

Liu, J., Sturrock, R. & Ekramoddoullah, A. (2010), 'The superfamily of thaumatin-like proteins: its origin. evolution. and expression towards biological function.', *Plant Cell Reports* **29**, 419–436.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012), 'Comparison of Next-Generation Sequencing Systems', *Journal of Biomedicine and Biotechnology* **2012**, 1–11.

Lottaz, C., Iseli, C., Jongeneel, C. V. & Bucher, P. (2003), 'Modeling sequencing errors by combining Hidden Markov models', *Bioinformatics* **19**, ii103–ii112.

Lü, B., Gong, Z., Wang, J., Zhang, J. & Liang, J. (2007), 'Microtubule dynamics in relation to osmotic stress-induced ABA accumulation in Zea mays roots.', *Journal of Experimental Botany* **58**, 2565–72.

Lynch, M. & Blanchard, J. L. (1998), 'Deleterious mutation accumulation in organelle genomes.', *Genetica* **102-103**, 29–39.

Ma, X., Zhu, Y., Li, C., Shang, Y., Meng, F., Chen, S. & Miao, L. (2011), 'Comparative transcriptome sequencing of germline and somatic tissues of the Ascaris suum gonad.', *BMC genomics* **12**, 481.

Mader, S. (2001), *Biology*, 7th editio edn, McGraw Hill, London.

Majoros, W. H., Pertea, M. & Salzberg, S. L. (2005), 'Efficient implementation of a generalized pair hidden Markov model for comparative gene finding.', *Bioinformatics (Oxford, England)* **21**, 1782–1788.

Malik, A. & Storey, K. (2009), 'Activation of antioxidant defense during dehydration stress in the African clawed frog.', *Gene* **422**, 99–107.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. (2005), 'Genome sequencing in microfabricated high-density picolitre reactors.', *Nature* **437**, 376–380.

Margulis, L. (1981), *Symbiosis in Cell Evolution*, San Francisco.

Martin, J. a. & Wang, Z. (2011), 'Next-generation transcriptome assembly.', *Nature Reviews. Genetics* **12**, 671–682.

Martínez, A., Portero-Otin, M., Pamplona, R. & Ferrer, I. (2010), 'Protein targets of oxidative damage in human neurodegenerative diseases with abnormal protein aggregates.', *Brain Pathology (Zurich, Switzerland)* **20**, 281–97.

Mason, C. E. & Elemento, O. (2012), 'Faster sequencers, larger datasets, new challenges', *Genome Biology* **13**, 314.

Mattimore, V. & Battista, J. R. (1996), 'Radioresistance of Deinococcus radiodurans: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation.', *Journal of Bacteriology* **178**, 633–637.

McNulty, S., Mullin, A., Vaughan, J., Tkach, V., Weil, G. & Fischer, P. (2012), 'Comparing the mitochondrial genomes of Wolbachia-dependent and independent filarial nematode species', *BMC Genomics* **13**, 145.

Meldal, B. H. M., Debenham, N. J., De Ley, P., De Ley, I. T., Vanfleteren, J. R., Vierstraete, A. R., Bert, W., Borgonie, G., Moens, T., Tyler, P. A., Austen, M. C., Blaxter, M. L., Rogers, A. D. & Lambshead, P. J. D. (2007), 'An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa.', *Molecular Phylogenetics and Evolution* **42**, 622–636.

Mereschkowski, K. (1903), *Das irdische Paradies oder ein Winternachtstraum.*, Jahrhundert, Jahrhundert.

Merida, I., Avila-Flores, A. & Merino, E. (2008), 'Diacylglycerol kinases: at the hub of cell signalling.', *Biochemistry* **409**, 1–18.

Meyer, Y., Buchanan, B., Vignols, F. & Reichheld, J. (2009), 'Thioredoxins and gutaredoxins: unifying elements in redox biology.', *Annual Review of Genetics* **43**, 335–376.

Montiel, R., Lucena, M. A., Medeiros, J. & Simo, N. (2006), 'The Complete Mitochondrial Genome of the Entomopathogenic Nematode Steinernema carpocapsae : Insights into Nematode Mitochondrial DNA Evolution and Phylogeny', *Galleria Rassegna Bimestrale Di Cultura* pp. 211–225.

Morimoto, R. I. (2008), 'Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging.', *Genes & Development* **22**, 1427–1438.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. (2007), 'KAAS: an automatic genome annotation and pathway reconstruction server', *Nucleic Acids Research* **35**, W182–W185.

Mounolou, J.-C. & Lacroute, F. (2005), 'Mitochondrial DNA: an advance in eukaryotic cell biology in the 1960s.', *Biology of the Cell / under the auspices of the European Cell Biology Organization* **97**, 743–748.

Mowla, S. B., Thomson, J. a., Farrant, J. M. & Mundree, S. G. (2002), 'A novel stress-inducible antioxidant enzyme identified from the resurrection plant Xerophyta viscosa Baker.', *Planta* **215**, 716–726.

Myers, E. W. (2000), 'A Whole-Genome Assembly of Drosophila', *Science* **287**, 2196–2204.

Nagayasu, E., Ogura, Y., Itoh, T., Yoshida, A., Chakraborty, G., Hayashi, T. & Maruyama, H. (2013), 'Transcriptomic analysis of four developmental stages of Strongyloides venezuelensis.', *Parasitology international* **62**, 57–65.

Navarre, C. & Goffeau, A. (2000), 'Membrane hyperpolarization and salt sensitivity induced by deletion of PMP3, a highly conserved small protein of yeast plasma membrane.', *The EMBO Journal* **19**, 2515–2524.

Neumann, S., Reuner, A., Brummer, F. & Schill, R. (2009), 'DNA damage in storage cells of anhydrobiotic tardigrades.', *Comparative Biochemistry and Physiology - Part A: Molecular and Integrative Physiology* **153**, 425–429.

Ngo, J. & Davies, K. (2009), 'Mitochondrial Lon protease is a human stress protein.', *Free Radical Biology and Medicine* **46**, 1042–1048.

Nicholas, W. L. & Stewart, A. C. (1985), 'Cryptobiotic nematodes in dry soil from Kinchega National Park, NSW.', *Australian Nematologists' Newsletter* **35**, 489–491.

Nicol, P., Gill, R., Fosu-Nyarko, J. & Jones, M. G. K. (2012), 'de novo analysis and functional classification of the transcriptome of the root lesion nematode,

Pratylenchus thornei, after 454 GS FLX sequencing.', *International Journal for Parasitology* **42**, 225–237.

Oda, Y., Okada, T., Yoshida, H., Kaufman, R., Nagata, K. & Mori, K. (2006), 'Derlin-2 and Derlin-3 are regulated by the mammalian unfolded protein response and are required for ER-associated degradation.', *Journal of Cell Biology* **172**, 383–393.

Okimoto, R., Macfarlane, J. L., Clary, D. & Wolstenholme, D. R. (1992), 'The mitochondrial genomes of two nematodes, Caenorhabditis elegans and Ascaris suum', *Genetics* **130**, 471–498.

Oliver, M. (1996), 'Desiccation tolerance in vegetative plant cells.', *Physiol Plant* **97**, 779–787.

Opperman, C. H., Bird, D. M., Williamson, V. M., Rokhsar, D. S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T. D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B. R., Thomas, V. P. & Windham, E. (2008), 'Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism.', *Proceedings of the National Academy of Sciences of the United States of America* **105**, 14802–14807.

Padmanabhan, S., Mukhopadhyay, A., Narasimhan, S. D., Tesz, G., Czech, M. P. & Tissenbaum, H. a. (2009), 'A PP2A regulatory subunit regulates C. elegans insulin/IGF-1 signaling by modulating AKT-1 phosphorylation.', *Cell* **136**, 939–951.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. a., Smirnova, T., Nosrat, B., Markowitz, V. M. & Kyrpides, N. C. (2012), 'The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.', *Nucleic acids research* **40**(Database issue), D571–9.

Pan, Q., Shai, O., Lee, L. J., B., F. J. & Blencowe, B. J. (2008), 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nature Genetics* **40**, 1413–1415.

Paradis, S. & Ruvkun, G. (1998), 'Caenorhabditis elegans Akt/PKB transduces insulin receptor-like signals from AGE-1 PI3 kinase to the DAF-16 transcription factor.', *Genes and Development* **12**, 2488–2498.

Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A. & Blaxter, M. (2004), 'PartiGene–constructing partial genomes.', *Bioinformatics (Oxford, England)* **20**, 1398–1404.

Parkinson, J., Guiliano, D. B. & Blaxter, M. (2002), 'Making sense of EST sequences by CLOBBing them', *BMC Bioinformatics* **8**, 1–8.

Parra, G., Bradnam, K. & Korf, I. (2007), 'CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.', *Bioinformatics (Oxford, England)* **23**, 1061–1067.

Parra, G., Bradnam, K., Ning, Z. M., Keane, T. & Korf, I. (2009), 'Assessing the gene space in draft genomes', *Nucleic Acids Research* **37**, 289–297.

Pellicer, J., Fay, M. F. & Leitch, I. J. (2010), 'The largest eukaryotic genome of them all?', *Botanical Journal of the Linnean Society* **164**, 10–15.

Perry, R. N. (1999), 'Desiccation survival of parasitic nematodes.', *Parasitology* **119 Suppl**, S19–30.

Pickup, J. (1990*a*), 'Seasonal variation in the cold-hardiness of a free-living predatory antarctic nematode.', *Polar Biology* **10**, 307–315.

Pickup, J. (1990*b*), 'Seasonal variation in the cold hardiness of three species of free-living antarctic nematodes.', *Functional Ecology* **4**, 257–264.

Poinar, G. O. & Sarbu, S. M. (1994), 'Chronogaster troglodytes sp. n. (Nemata : Chronogasteridae) from Movile Cave, with a review of cavernicolous nematodes', *Fundamental and Applied Nematology* **17**, 231–237.

Pujol, N., Zugasti, O., Wong, D., Couillault, C., Kurz, C. L., Schulenburg, H. & Ewbank, J. J. (2008), 'Anti-fungal innate immunity in C. elegans is enhanced by evolutionary diversification of antimicrobial peptides.', *PLoS Pathogens* **4**, e1000105.

Race, H. L., Herrmann, R. G. & Martin, W. (1999), 'Why have organelles retained genomes ?', *Trends in Genetics* **15**, 364 – 370.

Rao, K. V., Eswaran, M., Ravi, V., Gnanasekhar, B., Narayanan, R. B., Kaliraj, P., Jayaraman, K., Marson, A., Raghavan, N. & Scott, a. L. (2000), 'The Wuchereria bancrofti orthologue of Brugia malayi SXP1 and the diagnosis of bancroftian filariasis.', *Molecular and Biochemical Parasitology* **107**, 71–80.

Rasmussen, M., Alexander, R. T., Darborg, B. V., Mø bjerg, N., Hoffmann, E. K., Kapus, A., Pedersen, S. F., Kapus, A. & Sf, P. (2008), 'Osmotic cell shrinkage activates ezrin / radixin / moesin (ERM) proteins : activation mechanisms and physiological implications', *American Journal of Physiology - Cell Physiology* pp. 197–212.

Reardon, W., Chakrabortee, S., Pereira, T. C., Tyson, T., Banton, M. C., Dolan, K. M., Culleton, B. a., Wise, M. J., Burnell, A. M. & Tunnacliffe, A. (2010), 'Expression profiling and cross-species RNA interference (RNAi) of desiccation-induced transcripts in the anhydrobiotic nematode Aphelenchus avenae.', *BMC Molecular Biology* **11**, 6.

Rebecchi, L., Cesari, M., Altiero, T., Frigieri, A. & Guidetti, R. (2009), 'Survival

and DNA degradation in anhydrobiotic tardigrades.', *The Journal of Experimental Biology* **212**, 4033–4039.

Rensburg, C. J. V., Bert, C.-p. W. & Swart, A. (2010), 'Nematodes from the Bakwena Cave in Irene, Souh Africa. Promoter: Prof. Wilfrida Decraemer', *Africa* pp. 2009–2010.

Richter, K., Haslbeck, M. & Buchner, J. (2010), 'The heat shock response: life on the verge of death.', *Molecular Cell* **40**, 253–266.

Roach, J. C., Boysen, C., Wang, K. & Hood, L. (1995), 'Pairwise end sequencing: a unified approach to genomic mapping and sequencing', *Genomics* **26**, 345 – 353.

Rocha, E. (2008), 'The Organization of the Bacterial Genome', *Annual Review of Genetics* **42**, 211–233.

Roche (2012), '454 Sequencing'.

Rosario, K., Duffy, S. & Breitbart, M. (2009), 'Diverse circovirus-like genome architectures revealed by environmental metagenomics', *Journal of General Virology* **90**, 2418–2424.

Roy, S. W. & Gilbert, W. (2006), 'The evolution of spliceosomal introns: patterns, puzzles and progress', *Nature Review Genetics* **7**, 211–221.

Rudolphi, C. (1810), *Entozoorum sive Vermium Intesstinalium Historia Naturalis*, volumn 1 edn, Apud Treuttel et Wurtz Bibliopolas, Amsterlaedami.

Sagan, L. (1967), 'On the origin of mitosing cells', *Journal of Theoretical Biology* **14**, 255–274.

Sales, K., Brandt, W., Rumbak, E. & Lindsey, G. (2000), 'The LEA-like protein HSP 12 in Saccharomyces cerevisiae has a plasma membrane location and protects membranes against desiccation and ethanol-induced stress.', *Biochimica et Biophysica Acta - Biomembranes* **1463**, 267–278.

Salomons, F. F. a., Menendez-Benito, V., Bottcher, C., McCray, B. B. a., Taylor, J. P., Dantuma, N. P. N., Menéndez-Benito, V. & Böttcher, C. (2009), 'Selective accumulation of aggregation-prone proteasome substrates in response to proteotoxic stress.', *Molecular and cellular biology* **29**(7), 1774–1785.

Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998), 'Microbial gene identification using interpolated Markov models', *Nucleic Acids Research* **26**, 544–548.

Sambrook, J. & Russell, D. (2011), *Molecular Cloning: A Laboratory Manual*, 3rd edn, Cold Spring Harbor Laboratory, NewYork.

Sanger, F. & Nicklen, S. (1977), 'DNA sequencing with chain-terminating', *Biochemistry* **12**, 5463–5467.

Sato, N., Funayama, N., Nagafuchi, A., Yonemura, S. & Tsukita, S. (1992), 'A gene family consisting of ezrin, radixin and moesin. Its specific localization at actin filament/plasma membrane association sites.', *Journal of Cell Science* **103 (Pt 1**, 131–43.

Satou, Y., Hamaguchi, M., Takeuchi, K., Hastings, K. E. M. & Satoh, N. (2006), 'Genomic overview of mRNA 5-leader trans-splicing in the ascidian Ciona intestinalis', *Nucleic Acids Research* **34**, 3378–3388.

Schatz, M., Witkowski, J. & McCombie, W. (2012), 'Current challenges in de novo plant genome sequencing and assembly', *Genome Biology* **13**, 243.

Scheffler, I. E. (2000), 'A century of mitochondrial research : achievements and perspectives', *Science*.

Schmid, R. & Blaxter, M. L. (2008), 'annot8r: GO, EC and KEGG annotation of EST datasets.', *BMC Bioinformatics* **9**, 180.

Schroder, M. (2008), 'Endoplasmic reticulum stress responses.', *Cellular and Molecular Life Sciences* **65**, 862–894.

Sedding, D. (2008), 'FoxO transcription factors in oxidative stress response and ageing–a new fork on the way to longevity?', *Biological Chemistry* **389**, 279–283.

Shadel, G. S. & Clayton, D. a. (1997), 'Mitochondrial DNA maintenance in vertebrates.', *Annual Review of Biochemistry* **66**, 409–435.

Shannon, A. J., Browne, J. A., Boyd, J., Fitzpatrick, D. A. & Burnell, A. M. (2005), 'The anhydrobiotic potential and molecular phylogenetics of species and strains of Panagrolaimus (Nematoda, Panagrolaimidae).', *The Journal of Experimental Biology* **208**, 2433–2445.

Sharon, M., Kozarova, A., Clegg, J., Vacratsis, P. & Warner, A. (2009), 'Characterization of a group 1 late embryogenesis abundant protein in encysted embryos of the brine shrimp Artemia franciscana.', *Biochemistry and Cell Biology* **87**, 415–430.

Shen, Y. F., Sarin, S., Liu, Y., Hobert, O. & Pe'er, I. (2008), 'Comparing Platforms for C. elegans Mutant Identification Using High-Throughput Whole-Genome Sequencing', *PLoS ONE* **3**, e4012.

Shendelman, S., Jonason, A., Martinat, C., Leete, T. & Abeliovich, A. (2004), 'DJ-1 is a redox-dependent molecular chaperone that inhibits alpha-synuclein aggregate formation.', *PLoS biology* **2**, e362.

Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. (2004), 'Advanced sequencing technologies: Methods and Goals', *Nature Review: Genetics* **5**, 335–349.

Slack, G. & Gascoyne, R. (2013), 'Next-generation Sequencing Discoveries in Lymphoma.', *Advances in Anatomic Pathology* **20**, 110 – 116.

Smith, G. R. & Shanley, D. P. (2010), 'Modelling the response of FOXO transcription factors to multiple post-translational modifications made by ageing-related signalling pathways.', *PloS ONE* **5**, e11092.

Sohlenius, B. (1972), 'Nematodes from Surtsey I', *Surtsey Research Progress Report* **6**, 1973–1973.

Sohlenius, B. (1988), 'Interactions between two species of Panagrolaimus in agar curtures', *Nematologica* **34**, 208–217.

Spieth, J. & Blumenthal, T. (1985), 'Caenorhabditis elegans Vitellogenin', *Microbiology* **5**, 2495–2501.

Srinivasan, J., Dillman, A., Macchietto, M., Heikkinen, L., Lakso, M., Fracchia, K., Antoshechkin, I., Mortazavi, A., Wong, G. & Sternberg, P. W. (2013), 'The Draft Genome and Transcriptome of Panagrellus redivivus are Shaped by the Harsh Demands of a Free-Living Lifestyle.', *Genetics* pp. 0–0.

Stengel, F., Baldwin, A. J., Painter, A. J., Jaya, N., Basha, E., Kay, L. E., Vierling, E., Robinson, C. V. & Benesch, J. L. P. (2010), 'Quaternary dynamics and plasticity underlie small heat shock protein chaperone function.', *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2007–2012.

Stokoe, D., Engel, K., Campbell, D., Cohen, P. & Gaestel, M. (1992), 'Identifica-

tion of MAPKAP kinase-2 as a major enzyme responsible for the phosphorylation of the small mammalian heat-shock proteins.', *FEBS letters* **313**, 307–313.

Storch, J. & Xu, Z. (2009), 'Niemann-Pick C2 (NPC2) and intracellular cholesterol trafficking.', *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biololgy of Lipids* **1791**, 671–678.

Strange, K., Denton, J. & Nehrke, K. (2006), 'Ste20-Type kinases: evolutionarily conserved regulators of ion transport and cell volume.', *Physiology* **21**, 61–68.

Sultana, T., Kim, J., Lee, S.-H., Han, H., Kim, S., Min, G.-S., Nadler, S. a. & Park, J.-K. (2013), 'Comparative analysis of complete mitochondrial genome sequences confirms independent origins of plant-parasitic nematodes.', *BMC Evolutionary Biology* **13**, 12.

Sun, Y., Bojikova-Fournier, S. & MacRae, T. (2006), 'Structural and functional roles for beta-strand 7 in the alpha-crystallin domain of p26. a polydisperse small heat shock protein from Artemia franciscana.', *FEBS Journal* **273**, 1020–1034.

Tarr, D. E. K. & Scott, A. L. (2005), 'MSP domain proteins.', *Trends in Parasitology* **21**, 224–31.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. & Natale, D. A. (2003), 'The COG database: an updated version includes eukaryotes', *BMC Bioinformatics*.

Tavaria, M., Gabriele, T., Kola, I. & Anderson, R. L. (1996), 'A hitchhiker's guide to the human Hsp70 family', *Cell Stress and Chaperones* **1**, 23 – 28.

The Blaxter Lab (2013), 'Nematodes.org'.

The C. elegans Sequencing Consortium (1998), 'Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology', *Science* **282**, 2012–2018.

Tompa, P. (2005), 'The interplay between structure and function in intrinsically unstructured proteins.', *FEBS letters* **579**, 3346–54.

Tsuji, N., Suzuki, K., Kasuga-aoki, H., Isobe, T., Arakawa, T. & Matsumoto, Y. (2003), 'Mice Intranasally Immunized with a Recombinant 16-Kilodalton Antigen from Roundworm Ascaris Parasites Are Protected against Larval Migration of Ascaris suum', *Infection and immunity* **71**, 5314–5323.

Tunnacliffe, A. & Wise, M. J. (2007), 'The continuing conundrum of the LEA proteins.', *Die Naturwissenschaften* **94**, 791–812.

Tuskan, G. A. (2006), 'The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)', *Science* **313**, 1596–1604.

Tyson, T., O'Mahony Zamora, G., Wong, S., Skelton, M., Daly, B., Jones, J. T., Mulvihill, E. D., Elsworth, B., Phillips, M., Blaxter, M. & Burnell, A. M. (2012), 'A molecular analysis of desiccation tolerance mechanisms in the anhydrobiotic nematode Panagrolaimus superbus using expressed sequenced tags.', *BMC Research Notes* **5**, 68.

Tytgat, T., Vercauteren, I., Vanholme, B., De Meutter, J., Vanhoutte, I., Gheysen, G., Borgonie, G., Coomans, A. & Gheysen, G. (2005), 'An SXP/RAL-2 protein produced by the subventral pharyngeal glands in the plant parasitic root-knot nematode Meloidogyne incognita.', *Parasitology Research* **95**, 50–4.

Unseld, M., Marienfeld, J., Brandt, P. & Brennicke, A. (1997), 'The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides.', *Nature* **15**, 57–61.

van West, P., Kamoun, S., van't Klooster, J. & Govers, F. (1999), 'Ric1. a Phytophthora infestans gene with homology to stress-induced genes.', *Current Genetics* **36**, 310–315.

Vanier, M. & Millat, G. (2004), 'Structure and function of the NPC2 protein.', *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biololgy of Lipids* **1685**, 14–21.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. a., Holt, R. a., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, a. G., Nadeau, J., McKusick, V. a., Zinder, N., Levine, a. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, a., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, a., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, a. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. a., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, a. K., Narayan, V. a., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, a., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, a., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, a., Woodage, T., Ali, F., An, H., Awe, a., Baldwin, D., Baden, H., Barnstead, M.,

Barrow, I., Beeson, K., Busam, D., Carver, a., Center, a., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, a., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, a., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, a., Mi, H., Lazareva, B., Hatton, T., Narechania, a., Diemer, K., Muruganujan, a., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, a., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, a., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, a., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, a., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, a., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, a., Zandieh, a. & Zhu, X. (2001), 'The sequence of the human genome', *Science* **291**(5507), 1304.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002), 'Comparative assessment of large-scale data sets of protein-protein

interactions.', *Nature* **417**, 399–403.

Vos, M., Hageman, J., Carra, S. & Kampinga, H. (2008), 'Structural and functional diversities between members of the human HSPB. HSPH. HSPA. and DNAJ chaperone families.', *Biochemistry* **47**, 7001–7011.

Wallin, I. (1927), *Symbionticism and the Origin of Species.*, Williams & Wilkens, Baltimore, MD.

Wang, E. T. (2008), 'Alternative isoform regulation in human tissue transcriptomes', *Nature* **456**, 470–476.

Wang, J., Czech, B., Crunk, A., Wallace, A., Mitreva, M., Hannon, G. J. & Davis, R. E. (2011), 'Deep small RNA sequencing from the nematode Ascaris reveals conservation, functional diversification, and novel developmental profiles.', *Genome Research* **21**, 1462–1477.

Wang, S., Zheng, H., Dissanayake, S., Cheng, W., Tao, Z., Lin, S. & Piessens, W. (1997), 'Evaluation of recombinant chitinase and SXP1 antigens as antimicrofilarial vaccines.', *American Journal for Tropical Medicine and Hygenie* **56**, 474–481.

Wang, X., Li, W., Zhao, D.F. Liu, B., Shi, Y., Chen, B., Yang, H., Guo, P., Geng, X. Shang, Z. & Al., E. (2010*a*), 'Caenorhabditis elegans transthyretin-like protein TTR-52 mediates recognition of apoptotic cells by the CED-1 phagocyte receptor.', *Nature Cell Biology* **12**, 655–669.

Wang, Z., Abubucker, S., Martin, J., Wilson, R. K., Hawdon, J. & Mitreva, M. (2010*b*), 'Characterizing Ancylostoma caninum transcriptome and exploring nematode parasitic adaptation', *BMC Genomics* p. 11.

Ward, S., Burke, D., Sulstor, J., Coulson, A., Albertson, D., Ammons, D., Klass, M. & Hogan, E. (1988), 'Genomic organization of major sperm protein genes and pseudogenes in the nematode Caenorhabditis elegans.', *Journal of Molecular Biology* **199**, 1–13.

Wasmuth, J. D. & Blaxter, M. L. (2004), 'prot4EST : Translating Expressed Sequence Tags from neglected genomes', *BMC Bioinformatics* **14**, 1–14.

Wasmuth, J., Schmid, R., Hedley, A. & Blaxter, M. (2008), 'On the Extent and Origins of Genic Novelty in the Phylum Nematoda', *PLoS Neglected Tropical Diseases* **2**, 14.

Waterston, R. H. (2002), 'Initial sequencing and comparative analysis of the mouse genome', *Nature* **420**, 520–562.

Watson, J. (1953), 'Molecular structure of nucleic acids', *Nature* **171**, 737–738.

Weber, J. L. & Myers, E. W. (1997), 'Human Whole-Genome Shotgun Sequencing', *Genome Research* **7**, 401–409.

Wharton, D. A. (1996), 'Water loss and morphological changes during desiccation of the anhydrobiotic nematode Ditylenchus dipsaci', *The Journal of experimental biology* **1093**(Pt 5), 1085–1093.

Wharton, D. A. & Marshall, C. J. (2009), 'How do terrestrial Antarctic organisms survive in their harsh environment?', *Journal of Biology* **8**, 39.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pontius, J. U., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A.,

Wagner, L. & Yaschenko, E. (2005), 'Database resources of the National Center for Biotechnology Information.', *Nucleic Acids Research* **33**, D39–45.

Whitmarsh, A. J. (2010), 'A central role for p38 MAPK in the early transcriptional response to stress Commentary', *BMC Biology* pp. 8–10.

Williams, M. S. R. (1986), 'The use of scanning electron microscopy in the taxonomy of Panagrolaimus (Nematoda: Panagrolaimidae)', *Nematologica* **32**, 89–97.

Winston, P. & Bates, D. (1960), 'Saturarted solutions for the control of humidity in biological research', *Ecology* **41**, 232–237.

Winter, A., McCormack, G. & Page, A. (2007), 'Protein disulfide isomerase activity is essential for viability and extracellular matrix formation in the nematode Caenorhabditis elegans.', *Developmental Biology* **308**, 449–461.

Womersley, C. (1987), *A reevaluation of strategies employed by nematode anhydrobiotes in relation to their natural environment*, Hyattsville: Society of Nematologists.

Wood, Z., Schroder, E., Harris, J. & Poole, L. (2003), 'Structure. mechanism and regulation of peroxiredoxins.', *Trends in Biochemical Sciences* **28**, 32–40.

Wright, A. F., Murphy, M. P. & Turnbull, D. M. (2009), 'Do organellar genomes function as long-term redox damage sensors?', *Trends in Genetics* **25**, 253–261.

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. (1985), 'Mitochondrial origins.', *Proceedings of the National Academy of Sciences of the United States of America* **82**, 4443–4447.

Zaslaver, A., Baugh, L. R., Sternberg, P. W. & W, P. (2011), 'Metazoan Operons Accelerate Recovery from Growth-Arrested States', *Cell* **146**, 981–992.

Zerbino, D. R. & Birney, E. (2008), 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs.', *Genome Research* **18**, 821–829.

Zhang, J., Li, X., Mueller, M., Wang, Y., Zong, C., Deng, N., Vondriska, T. M., Liem, D. a., Yang, J.-I., Korge, P., Honda, H., Weiss, J. N., Apweiler, R. & Ping, P. (2008), 'Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria.', *Proteomics* **8**, 1564–1575.

Zheng, Y., Zhao, L., Gao, J. & Fei, Z. (2011), 'iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences.', *BMC Bioinformatics* **12**, 453.

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., Matz, M. V., Meleshkevitch, E., Moroz, L. L., Lukyanov, S. A. & Shagin, D. A. (2004), 'Simple cDNA normalization using kamchatka crab duplex-specific nuclease.', *Nucleic Acids Research* **32**, e37.

Zollner, N. (1982), 'Purine and pyrimidine metabolism', *Proceedings of the Nutrition Society* **41**, 329 – 342.