



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Next-generation Nematode Genomes

Sujai Kumar

Doctor of Philosophy

Institute of Evolutionary Biology

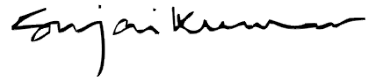
School of Biological Sciences

University of Edinburgh

2012

Declaration

I declare that this thesis is my own work, and that the work described here is my own except where explicitly stated. Some of the work was done in collaboration with others, and my contributions are described in the introductions to those chapters. This work has not been submitted for any other degree or professional qualification.

A handwritten signature in black ink that reads "Sujai Kumar". The signature is written in a cursive style with a long horizontal stroke at the end.

Sujai Kumar, 2012

Acknowledgements

This thesis is about next-generation sequence data, and those data would not have existed without the GenePool Genomics Facility of the University of Edinburgh. My thanks for the long hours put in by everyone, especially the scientific manager (Karim Gharbi).

It's been a great experience sharing office space, lab meetings, and countless lunches with both the GenePool bioinformatics team and the Blaxter Lab (Alex Marshall, Ben Elsworth, Gaganjot Kaur, Georgios Koutsovoulos, Graham Thomas, Jack Hearn, John Davey, Martin Jones, Pablo Fuentes, Stephen Bridgett, Timothée Cezard, and Urmi Trivedi). I have learnt so much from all of these people about bioinformatics and genomics. Many of the ideas in this thesis emerged out of hours of discussions with everyone on how to deal with the problems of next-gen data. Within the Institute of Evolutionary Biology, I'd also like to thank Peter Keightley and Graham Stone for their help and advice over the years.

I have enjoyed working with our external collaborators very much and am grateful to them for their insights and expertise: Asher Cutter, Erich Schwarz, and Marie-Anne Felix for *Caenorhabditis sp. 5*; David Lunt for *M. floridensis*; Adrian Wolstenholme, Christelle Godel, Christian Epe, Christoph Schmid, Claudio Bandi, Daniel Nilsson, Francesco Comandatore, Frédéric Bringaud, Pascal Mäser, Philipp Ludin, Ronald Kaminsky, and Susumu Goto for the *D. immitis* (and its *Wolbachia*) genome project; David Taylor, Judith Allen, and Simon Babayan for *L. sigmodontis*; Kevin Howe and Michael Paulini for help with WormBase; Philipp Schiffer for ideas on the 959 Nematode Genomes wiki and for testing many of the scripts in this thesis; and Matt Berriman and the Pathogen Genomics group at the Wellcome Trust Sanger Institute for hosting me for a week and helping me appreciate the challenges in managing large data sets and software systems.

For constantly keeping me up to date with the challenges in the field and for providing a place to swap bioinformatics-in-the-trenches stories, I'd like to thank the Scotland-wide Next Gen Bioinformatics User Group (NextGenBUG) and the bioinformatics journal club at the IEB. A general thank-you goes out to the open source community in general and the people who have written all the programs used in this thesis. I hope that some day I can contribute something more substantial than a bunch of scripts to this community.

The work in this thesis made extensive use of the resources provided by the Edinburgh Compute and Data Facility (ECDF - www.ecdf.ed.ac.uk). The ECDF is partially supported by the eDIKT initiative (www.edikt.org.uk). Many thanks to the ECDF team (especially Orlando Richards, Kenton D'Mellow, Nick Moir, and Brian Fletcher) for setting up and maintaining amazing levels of performance on the Eddie compute cluster.

The School of Biological Sciences of the University of Edinburgh funded my PhD studentship, and I am very grateful to them for giving me the opportunity to work in this

beautiful city with happy and brilliant people. I would also like to thank the Overseas Research Studentship scheme for funding the fee difference for non-EU students. Over the last three years, the following organisations and schemes have also provided funds for conference travel and research activities: NemaSym, a Research Coordination Network for the study of Nematode-Bacterium Symbioses, for travel to two research meetings in 2011 (Corvallis, OR, USA) and 2012 (Cold Spring Harbor, NY, USA); James Rennie Bequest (University of Edinburgh) for travel to the 'Evolution of Caenorhabditis and other nematodes' meeting in 2012 (Cold Spring Harbor, NY, USA); and the AWS in Education Research Grant for the use of Amazon Web Services.

I cannot imagine how I would have completed this thesis without the help of Linda Sims, my closest friend. She was the perfect thesis coach: keeping me on task, listening to outlines, and checking first drafts for grammatical mistakes. Many thanks to her for keeping me sane.

My immediate and extended family have been wonderfully supportive throughout my life, but I am especially thankful for the last few months where Rohit kept me laughing and Amma was always there whenever I needed to hear a reassuring voice. I'd like to dedicate this thesis to the memory of my father, Kuldip Kumar, who taught me to wander.

And finally, I'd like to thank Mark Blaxter for wearing many hats: as a PhD supervisor, he provided freedom and amazing intellectual scaffolding; as a feature creeper, he motivated all of us to think of ways to make everything work better; as an admirer of all things wiggly, he encouraged an appreciation for the smaller things in life; and as a wonderful human being, he has inspired me to be kinder and more generous.

Abstract

Introduction:

The first metazoan to be sequenced was a nematode (*Caenorhabditis elegans*), and understanding the genome of this model organism has led to many insights about all animals. Although eleven nematode genomes have been published so far and approximately twenty more are under way, the vast majority of the genomes of this incredibly diverse phylum remain unexplored. Next-generation sequencing has made it possible to generate large amounts of genome sequence data in a few days at a fraction of the cost of traditional Sanger-sequencing. However, assembling and annotating these data into genomic resources remains a challenge because of the short reads, the quality issues in these kinds of data, and the presence of contaminants and co-bionts in uncultured samples. In this thesis, I describe the process of creating high quality draft genomes and annotation resources for four nematode species representing three of the five major nematode clades: *Caenorhabditis sp. 5*, *Meloidogyne floridensis*, *Dirofilaria immitis*, and *Litomosoides sigmodontis*. I describe the new approaches I developed for visualising contamination and co-bionts, and I present the details of the robust workflow I devised to deal with the problems of generating low-cost genomic resources from Illumina short-read sequencing.

Results:

The draft genome assemblies created using the workflow described in this thesis are comparable to the draft nematode genomes created using Sanger sequencing. Armed with these genomes, I was able to answer two evolutionary genomics questions at very different scales. The first question was whether any non-coding elements were deeply conserved at the level of the whole phylum. Such elements had previously been hypothesised to be responsible for the phylum body plan in vertebrates, insects, and nematodes. I used twenty nematode genomes in several whole-genome alignments and concluded that no such elements were conserved across the whole phylum. The second question addressed the origins of the highly destructive plant-parasitic root-knot nematode *Meloidogyne incognita*. Comparisons with the newly sequenced *Meloidogyne floridensis* genome revealed the complex hybrid origins of both species, undermining previous assumptions about the rarity of hybrid speciation in animals.

Conclusions:

This thesis demonstrates the role of next-generation sequencing in democratising genome sequencing projects. Using the sequencing strategies, workflows, and tools described here, one can rapidly create genomic resources at a very low cost, even for unculturable metazoans. These genomes can be used to understand the evolutionary history of a genus or a phylum, as shown.

Table of Contents

Declaration.....	i
Acknowledgements	iii
Abstract.....	v
Table of Contents	vii
List of Figures.....	xi
List of Tables.....	xiii
Acronyms and definitions	xv
1 Introduction.....	1
1.1 Why sequence nematode genomes.....	1
1.2 Current status and the 959 Nematode Genomes initiative	5
1.3 Thesis structure.....	6
2 Assembling four nematode genomes	9
2.1 Introduction.....	9
2.1.1 Sequencing technologies for <i>de novo</i> genomes	9
2.1.2 Sequencing strategies	12
2.1.3 Short-read assembly concepts and algorithms.....	14
2.1.4 Assembly optimality criteria.....	21
2.2 Methods	25
2.2.1 Read quality control	25
2.2.2 Preliminary assembly.....	28
2.2.3 Stringent re-assembly.....	30
2.2.4 Post-assembly	34
2.3 Results	36
2.3.1 <i>Caenorhabditis sp. 5</i>	36
2.3.2 <i>Meloidogyne floridensis</i>	46
2.3.3 <i>Dirofilaria immitis</i>	52
2.3.4 <i>Litomosoides sigmodontis</i>	61
2.4 Discussion.....	70
2.4.1 NGS genomes are comparable to Sanger-sequenced genomes	70
2.4.2 Lessons learnt	72
2.4.3 Upcoming technologies	78
2.4.4 Summary	79

3	Annotating nematode genomes	83
3.1	Introduction	83
3.1.1	Annotation concepts	84
3.1.2	Review of existing tools	85
3.2	Methods	89
3.2.1	Computational gene prediction using MAKER2	89
3.2.2	Calculating gene prediction metrics	94
3.2.3	Functional annotation of protein-coding genes	94
3.2.4	Repeat-masking	95
3.3	Results	95
3.3.1	Gene models with the MAKER2 workflow	95
3.3.2	InterProScan annotations across 20 nematode genomes	99
3.3.3	tRNA predictions across 20 nematode genomes	102
3.4	Discussion	104
3.4.1	Automated gene prediction and functional annotation have limitations	104
3.4.2	<i>C. elegans</i> is not "the" nematode	105
3.4.3	Future improvements	106
4	Lack of deeply conserved non-coding elements in nematodes	109
4.1	Introduction	109
4.1.1	Previous CNE research on vertebrates, insects, and nematodes	109
4.1.2	CNEs might define the phylum body plan	114
4.1.3	Aims for this study	114
4.2	Methods	115
4.2.1	Genome and coding-region data	115
4.2.2	Identifying CNEs using whole-genome alignments	117
4.2.3	Identifying CNEs using MegaBLAST and clustering	121
4.3	Results	121
4.3.1	No CNEs were shared across clades	121
4.4	Discussion	125
4.4.1	An important negative result	125
4.4.2	Next steps	126
5	The <i>Meloidogyne floridensis</i> genome reveals complex hybrid origins of the root-knot nematodes	129
5.1	Introduction	129
5.1.1	Hypotheses for origins of <i>M. incognita</i> genomic duplicates	132
5.2	Methods	135
5.2.1	Nematode materials	135
5.2.2	Protein predictions and comparisons	135
5.2.3	Clustering	136

5.2.4	Phylogenetic analyses	136
5.3	Results	137
5.3.1	The genome of <i>M. floridensis</i>	137
5.3.2	Intra-genomic comparisons reveal high numbers of duplicate genes in <i>M. incognita</i> and <i>M. floridensis</i>	139
5.3.3	Distinguishing sibling from parent-child species relationships	139
5.3.4	Phylogenetic analysis of homologue relationships	143
5.4	Discussion.....	146
5.4.1	The <i>M. floridensis</i> genome reveals hybrid origins	146
5.4.2	Molecular genetic approaches to <i>Meloidogyne</i> diversity	148
6	What next for next-generation nematode genomes	151
6.1	What can we do with 959 Nematode Genomes	151
6.2	Will new technologies make assembly redundant.....	152
6.3	Can we generate more accurate annotations.....	156
6.4	Do CNEs really define genetic regulatory networks	158
6.5	What could fifteen <i>Meloidogyne</i> genomes tell us about hybrid speciation.....	159
6.6	Summary.....	160
	Bibliography.....	161
	Appendix A: Workflows and scripts	173
	Appendix B: Data files	177
	Appendix C: Publications.....	179

List of Figures

Figure 1.1	Systematic tree of Nematoda indicating current sequenced, in progress or proposed genome sequencing projects	4
Figure 2.1	Cartoon example of contig (or scaffold) cumulative length curve	23
Figure 2.2	Workflow for assembling nematode genomes from short-read Illumina sequencing	26
Figure 2.3	Biological accuracy measured using alignments with ESTs or Proteins	33
Figure 2.4	Removing short-fragment read-pairs before scaffolding contigs with mate-pairs	35
Figure 2.5	Taxon-annotated GC-coverage plot for <i>Caenorhabditis sp. 5</i>	40
Figure 2.6	Comparison of cumulative scaffold and contig length curves for <i>Caenorhabditis sp. 5</i> stringent re-assemblies.....	45
Figure 2.7	Taxon-annotated GC-coverage plot for <i>M. floridensis</i>	47
Figure 2.8	Comparison of cumulative scaffold and contig length curves for <i>M. floridensis</i> stringent re-assemblies.....	51
Figure 2.9	Taxon-annotated GC-coverage plot for <i>D. immitis</i>	55
Figure 2.10	Comparison of cumulative lengths of scaffolds, contigs, and blocks of Ns for different assemblies of <i>D. immitis</i>	58
Figure 2.11	Taxon-annotated GC-coverage plot for <i>L. sigmodontis</i>	64
Figure 2.12	Comparison of cumulative scaffold and contig lengths for <i>L. sigmodontis</i>	67
Figure 2.13	Assembly comparisons of 20 nematode genomes	71
Figure 3.1	Two-pass iterative MAKER2 pipeline	92
Figure 3.2	Comparison of gene model statistics across 20 nematode genomes	98
Figure 3.3	InterProScan annotations for 20 nematode proteomes	101
Figure 3.4	tRNA predictions for 20 nematode genomes.....	103
Figure 4.1	A model for the evolution of cis-regulatory elements involved in animal development.....	113
Figure 4.2	Workflow for finding CNEs using whole-genome multiple alignments	116
Figure 4.3	Cladogram depicting guide trees used in TBA+Multiz	119
Figure 4.4	Cartoon example of how coding regions were removed from multiple alignments.....	120
Figure 4.5	Length and identity of CNEs found at different nodes in nematode phylogeny	123
Figure 5.1	<i>Meloidogyne</i> phylogeny indicating positions of <i>M. floridensis</i> and tropical apomicts	131
Figure 5.2	Hypotheses for relationships between <i>M. floridensis</i> , <i>M. incognita</i> , and <i>M. hapla</i> , and the origins of duplicated gene copies.....	133
Figure 5.3	Intra-genomic duplication of protein-coding sequences	140
Figure 5.4	Venn diagram of clustering of proteins from three <i>Meloidogyne</i> species.....	141
Figure 5.5	Phylogenetic analysis of clustered CDS sets	145

List of Tables

Table 2.1	Sequencing costs, throughputs, read lengths and error profiles	11
Table 2.2	Classification of NGS assembly algorithms	17
Table 2.3	Assemblers tested	20
Table 2.4	Read data for <i>Caenorhabditis sp. 5</i>	38
Table 2.5	Read separation for <i>Caenorhabditis sp. 5</i> preliminary assembly	41
Table 2.6	Comparison of stringent re-assemblies for <i>Caenorhabditis sp. 5</i>	44
Table 2.7	Comparison of stringent re-assemblies for <i>M. floridensis</i>	50
Table 2.8	Read data for <i>D. immitis</i>	53
Table 2.9	Comparison of assemblies for <i>D. immitis</i>	57
Table 2.10	Comparison of assemblies for <i>Wolbachia</i> of <i>D. immitis</i>	60
Table 2.11	Read data for <i>L. sigmodontis</i>	62
Table 2.12	Comparison of stringent re-assemblies for <i>L. sigmodontis</i>	66
Table 2.13	Comparison of Velvet and ABySS stringent re-assemblies for <i>Wolbachia</i> of <i>L. sigmodontis</i>	69
Table 2.14	Upcoming sequencing technologies	80
Table 3.1	Differences in gene predictions using varying HMMs with SNAP	90
Table 3.2	Data used for annotation	93
Table 4.1	CNEs found in different groups of species.....	111
Table 4.2	TBA+Multiz guide trees	119
Table 4.3	Comparing CNEs found using different methods	124
Table 5.1	Summary statistics describing assemblies and protein predictions in <i>Meloidogyne</i> genomes	138
Table 5.2	Numbers of <i>M. floridensis</i> and <i>M. incognita</i> members in homeologue sets with one <i>M. hapla</i> member	142

Acronyms and definitions

3p/3'	three-prime
5p/5'	five-prime
959NG	959 Nematode Genomes
aa	amino acids
AFLP	amplified fragment length polymorphisms
b	bases
BAM	Binary (Sequence) Alignment Map - a binary compressed version of SAM
BED	a file format for storing a set of genome coordinates and intervals
bp	base pairs
CC50	the distance between correctly contiguous pairs in an assembly, such that at least 50% of all correctly contiguous pairs are that distance apart.
CDS	coding sequence
CEGMA	Core Eukaryotic Genes Mapping Approach - a software program that searches for a set of 248 core eukaryotic gene models.
CPNG50	NG50 (see below) for all contig-paths obtained when a test assembly is aligned to a reference sequence.
DBG	de Bruijn Graph
DDBJ	DNA Databank of Japan
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
EST	Expressed Sequence Tag
Fastq	format for short-read sequence data along with quality
GAGE	Genome Assembly Gold-standard Evaluations
GB	giga bytes
Gbp	giga base pairs
GC	Guanine-Cytosine, used as shorthand for GC-content
GC-cov	GC and coverage
GFF	Genome Feature Format, a file format specification for genome features
GFF3	Genome Feature Format, version 3
GO	gene-ontology, a set of terms used to describe the cellular component location, biological process, and molecular function of a gene
GOslim	a subset of the gene ontology (GO) that can be useful in assigning high-level annotations to gene sets
HSP	High-Scoring Pair (in a sequence alignment)
ITS	internal transcribed spacer
kbp	kilo base pairs
M	Million
Mbp	mega base pairs
miRNA	micro RNA
MP	Mate-Pairs
mtDNA	mitochondrial DNA
MYA	million years ago
N50	Size of contig (or scaffold) in an assembly such that 50% of the assembly is in contigs of that size or larger
NCBI	National Centre for Biotechnology Information
ncRNA	non-coding RNA
nDi	nuclear genome of <i>Dirofilaria immitis</i>
NG50	Size of contig (or scaffold) in an assembly such that 50% of the actual genome size (if known) is in contigs of that size or larger
NGS	Next-Generation Sequencing
nLs	nuclear genome of <i>Litomosoides sigmodontis</i>
nmwDi	nuclear, mitochondrial, and <i>Wolbachia</i> assemblies of <i>Dirofilaria immitis</i>
nmwLs	nuclear, mitochondrial, and <i>Wolbachia</i> assemblies of <i>Litomosoides sigmodontis</i>
PE	Paired-End
Pfam	Protein families database
QC	Quality Control

rDNA	ribosomal DNA
RAPD	random amplified polymorphic DNA marker
Rfam	RNA families database
RNA-Seq	RNA-Sequencing; typically carried out by selecting RNA and sequenced using NGS
rRNA	ribosomal RNA
RT qPCR	Real-Time quantitative Polymerase Chain Reaction
SAM	Sequence Alignment Map - a format for storing how reads map to a reference
SE	Single-End (i.e., unpaired reads)
SG	String-Graph
SNP	Single Nucleotide Polymorphism
SPNG50	NG50 (see above) for all scaffold-paths obtained when a test assembly is aligned to a reference sequence.
SRA	Short Read Archive
TAGC plot	Taxon-annotated GC-coverage plot
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
tLCA	time to Last Common Ancestor
tRNA	transfer RNA
UTR	Untranslated region
wBm	<i>Wolbachia</i> endosymbiont of <i>Brugia malayi</i>
wDi	<i>Wolbachia</i> endosymbiont of <i>Dirofilaria immitis</i>
WGA	Whole-Genome Amplification
WGS	Whole-Genome Shotgun
wLs	<i>Wolbachia</i> endosymbiont of <i>Litomosoides sigmodontis</i>
wMel	<i>Wolbachia</i> endosymbiont of <i>Drosophila melanogaster</i>

I Introduction

The phylum Nematoda, present in nearly every ecological niche on our planet, is the most numerous and diverse of all animal phyla [1, 2]. With the dramatic evolution of DNA sequencing technology in the last two decades, genome resources have been created for many nematodes. As a result, nematode research in the areas of developmental biology [3], genome biology [4, 5], evolutionary genomics [6], neurobiology [7], aging [8], health [9], and parasitology [10] have transformed our understanding of not just this phylum, but of all organisms. The goal of this thesis is to present the process of creating nematode genome resources using next-generation sequencing and to describe the testing of evolutionary hypotheses using these resources.

This introductory chapter begins by arguing why nematodes and their genomes are important and why we need more complete nematode genomes. A brief introduction to nematode phylogeny follows, along with a list of the genomes currently available or in progress. As part of the effort to sequence more genomes, the 959 Nematode Genomes (959NG) wiki—set up to coordinate sequencing efforts—is described next. The last part of the chapter outlines the structure of the rest of this thesis and its contributions to nematode genomics research. Sections of this chapter have been previously published in two review articles [11, 12] and in a paper describing the 959NG wiki [13].

I.1 Why sequence nematode genomes

Nematodes are ubiquitous and diverse

Nematodes are incredibly diverse in form and life habit despite their simple fundamental body plan. They are also commonly known as roundworms because of their tube-within-a-tube structure. The outer tube is made up of a cuticle, hypodermis, excretory system, neurons, and muscles, whereas the inner tube consists of the pharynx, intestine and gonads (in adults). Adult body sizes range from 200 μm (e.g. plant-parasitic *Sphaeronema* species [14]) to 7 m (*Placentanema gigantisma*, a spirurid parasite of sperm whales [15]). Lifespans range from a few days (e.g. *Caenorhabditis elegans* [16]) to several years (human filarial nematodes can live for decades [17]). Having evolved to thrive in diverse niches from interstitial microbivory to tissue-dwelling parasitism, the ubiquitous nature of nematodes is impressive: according to some estimates, there are one to two million nematode species (although only ~23,000 have been described) [2], and nematodes make up 80% of all individual animals [1].

... if all the matter in the universe except the nematodes were swept away, our world would still be dimly recognizable [...] we should find its mountains, hills, vales, rivers, lakes, and oceans represented by a film of nematodes [...] The location of the various plants and animals would still be decipherable, and, had we sufficient knowledge, in many cases even their species could be determined by an examination of their erstwhile nematode parasites

Nathan Cobb [18]

Nematodes affect health and economies

According to Chan [19], the three most damaging parasitic nematodes for humans are *Ascaris lumbricoides* (prevalent in 24% of the global population), *Necator americanus* (24%), and *Trichuris trichiura* (17%). Together, these helminths cause a loss in disability-adjusted life-years (DALYs) that is comparable to the worst diseases of our times, such as malaria, tuberculosis, and measles. In addition, human and veterinary filariasis and ascariasis are caused by approximately a dozen different nematodes, infecting approximately 120 M people worldwide [17]. Sequencing the genomes of these nematodes (or those of closely related nematodes that infect animal models) will speed the process of developing effective vaccines and treatments against them. Similarly, plant-parasitic root-knot nematodes affect ~5 % of global crops, translating to an annual loss of over 100 billion dollars [20].

On the positive side, some nematodes act as biocontrol agents for pests (e.g., *Deladenus siricidicola* for wood wasps and *Phasmarhabditis hermaphrodita* for slugs) by infecting these unwanted organisms. Overall, understanding the genes involved in nematode parasitism can help us design better vaccines and control techniques [21].

***C. elegans* is a model nematode**

When Sydney Brenner selected the free-living *C. elegans* as his preferred organism for genetic experiments in 1974 [16], few might have anticipated the effect this choice would have on modern biology. *C. elegans* was the first metazoan to have its complete developmental cell lineage mapped [3], its complete nervous system defined [7], and its

complete genome sequenced [4]. Several general biological principles and features were first discovered in *C. elegans*, such as RNA interference [22], microRNAs [23], and the use of green fluorescent protein as a marker for gene expression [24]. Similarly, signalling pathways in *C. elegans* for organ development turned out to have analogous pathways and genes in other animal groups [25], and the mechanisms of programmed cell death [26] were found to be applicable to human disease as well.

The *C. elegans* genome project [4] acted as a test bed for the human genome project [27]. To this day, it remains the only absolutely complete animal genome with no missing or ambiguous bases. Even though it is a compact genome (100 Mbp) compared to the 3 Gbp human genome, *C. elegans* is a complex genome with roughly the same number of protein-coding genes (19,735 [28]) as the latter (20,687 [29]), and exhibits many of the same genomic features. Combined with the long-standing genetic tools and genomic resources [30] available for *C. elegans*, it continues to be a model genome. Subsequent nematode genome sequencing projects were built on the success of the *C. elegans* genome, and as a result were able to exploit its extensive annotations to study the newer genomes.

One genome does not a phylum make

While *C. elegans* is an excellent model nematode and its genome—with its wealth of annotation—is an excellent model genome [31], *C. elegans* cannot be taken to represent all nematode genomes (Figure 1.1). We know that this species is quite derived within Nematoda [32] and that it lacks many genes shared between other nematodes and other Metazoa [33]. For example, nematode genomes have been sized from 20 Mbp to 500 Mbp (i.e., one fifth to five times that of *C. elegans*) [34], and sequenced nematode genomes range from *Meloidogyne hapla* [35] at 54 Mbp to *Ascaris suum* [36] at 273 Mbp. In addition, interesting genomic features have been found in other species, including chromatin diminution in *Ascaris suum* and other ascaridids (i.e., the germline has a larger genome than the soma [37]), aneuploid triploidy in the *Meloidogyne incognita* genome [38], and the presence of obligate, vertically-transmitted symbiont alphaproteobacterial *Wolbachia* and their genomes inside the cells of many filarial nematodes [39].

Apart from expanding our understanding of genome organisation and origins, working with a richer sampling of sequenced genomes opens the door to a better understanding of the phylogeny of Nematoda and the evolutionary dynamics of important traits—such as parasitism of plants and animals—and developmental modes. To date, the most comprehensive molecular phylogenies of Nematoda have been based on a single gene, the ~1600 bp nuclear small subunit rRNA locus [32, 40, 41], but this single locus is insufficient for robust resolution of the deep divergences in the phylum. Methods for generating large-scale multi-gene phylogenies now exist and can be applied even to draft genomes.

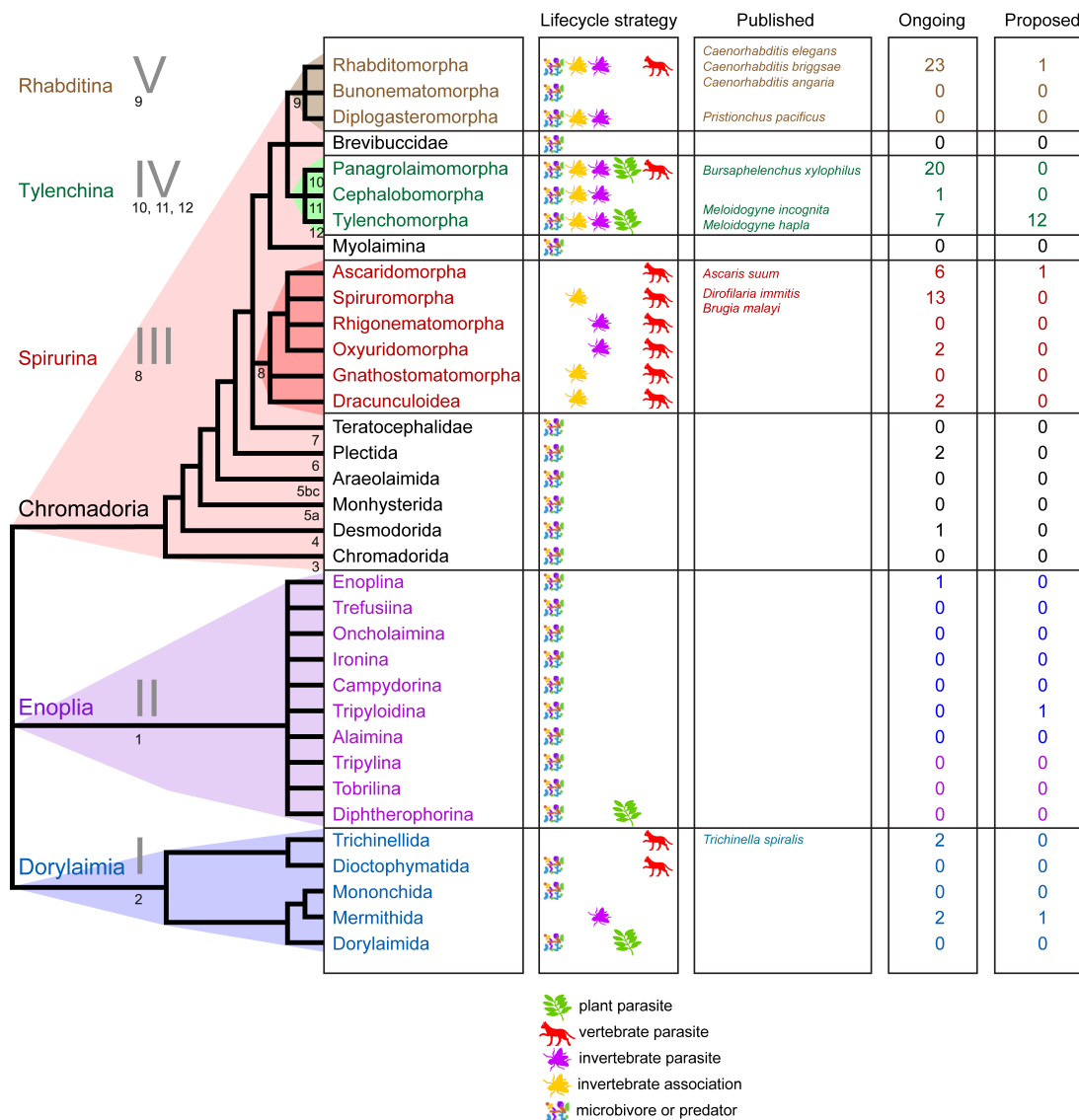


Figure 1.1 Systematic tree of Nematoda indicating current sequenced, in progress or proposed genome sequencing projects
 The systematic arrangement of Nematoda is based on De Ley and Blaxter [42]; the clades defined by Blaxter *et al.* [32] and van Megen *et al.* [41] are indicated using roman and arabic numerals respectively. For each major group we summarise the trophic ecology (microbivore, predator, fungivore, plant parasite, non-vertebrate parasite or associate, vertebrate parasite) and the number of species for which genome projects are reported in the 959 Nematode Genomes wiki as on September 1, 2012. The "Ongoing" column includes completed genome projects that have not yet been published.
 Figure updated from Kumar *et al.* [11]

Genomes instead of transcriptomes

In the past, an expressed sequence tag (EST) survey was the most cost-effective way to understand the protein-coding capabilities of a nematode [33, 43, 44]. However, next-generation sequencing (NGS) technologies have revolutionised and democratised the field of genome sequencing. Even small labs can now sequence their favourite nematodes in a few weeks for a few thousand dollars.

It is now a viable undertaking to produce whole-genome sequences for a nematode species of interest without the support of an industrial-scale genomics institution. Once the end point of many years of deliberation, selection, funding applications, and sequencing centre coordination efforts, genome sequencing can now be a starting point when one does not know much about an organism's biochemistry. For those interested in a near-complete gene catalogue for their chosen nematode, whole-genome sequencing is a more robust approach than transcriptome sequencing. While EST or transcriptome sequencing can only sample genes expressed at the time of harvest, and may poorly sample genes expressed at low levels, whole-genome sequencing yields access to all the genes. Even though the gene catalogue from a whole-genome sequencing project may still be incomplete (as some regions of the genome are difficult to sequence by any method), and some transcripts will be very hard to predict on the basis of genomic sequence alone, the results of whole-genome sequencing will be more complete than even the deepest transcriptome programme can deliver. More importantly, many transcriptome and EST sequencing projects use poly(A) enrichment which only provides an estimate of the protein-coding regions. As the modENCODE [45] project has shown, a large part of genome functionality lies in non-protein-coding regions, which are accessible in genomic resources.

1.2 Current status and the 959 Nematode Genomes initiative

Since the publication of the *C. elegans* genome in 1998 [4], ten other nematode genomes have been published. In the order of publication, they are: *Caenorhabditis briggsae* [5], *Brugia malayi* [46], *Meloidogyne incognita* [38], *Meloidogyne hapla* [35], *Pristionchus pacificus* [47], *Caenorhabditis angaria* [48], *Trichinella spiralis* [49], *Bursaphelenchus xylophilus* [50], *Ascaris suum* [36], and *Dirofilaria immitis* [51]. Of these, only *C. angaria*, *B. xylophilus*, *A. suum*, and *D. immitis* were sequenced using NGS, while the rest were sequenced using traditional capillary (Sanger) sequencing. Many more NGS genome projects are ongoing (Figure 1.1), but are at various stages of completion and have not been published yet.

Figure 1.1 also shows the phylogenetic context of the genomes that have been sequenced and are under way. Very few nematode genomes are ongoing or proposed in Clades I and II (as defined by Blaxter *et al.* [32]), with only one published Clade I genome. We anticipate that

this situation will change shortly as sequencing costs drop and individual labs generate genome sequences for the less represented branches of the nematode phylogenetic tree.

While biologists everywhere applaud the arrival of rapid and inexpensive sequencing, the explosion of data comes with a new problem: keeping track of which genomes are being sequenced, who is sequencing them, what stage the genome projects are at, and where one can get early access to the data. The nucleotide sequence archives (DDBJ [52], ENA [53], and GenBank [54]) are the de facto storehouses for complete and published genomes. However, as the bottleneck of a genome project has shifted from sequencing to analysis, which can take months to years, it is now imperative to have a place to share information about projects before they are published. Inspired by ArthropodBase (www.arthropodgenomes.org), we created the 959 Nematode Genomes (959NG) wiki in early 2010 to meet this need [13], which can be accessed at <http://959.nematodegenomes.org> [55].

The 959NG wiki announced a push to sequence, in the first instance, 959 nematode genomes [11]. Why (only) 959 genomes? The adult hermaphrodite *C. elegans* has 959 somatic cells, and one of the first major projects that turned *C. elegans* from a local curiosity into a key global research organism was the deciphering of the near-invariant developmental cell lineage that gives rise to these adult cells, starting from the fertilised zygote [3]. In an analogous way, we hope that a nematode phylogeny (the evolutionary lineage of the extant species) with 959 or more species will be similarly catalytic in driving nematode research programs across the spectrum of basic and applied science. Obviously, as sequencing technologies improve and become more accessible, we will move beyond this initial goal of 959, especially with over 23,000 described species and an estimated one to two million undescribed species in the phylum [1].

The wiki was developed on the Semantic MediaWiki (SMW) platform [56] because it allows pages to store properties and relationships to other pages. These properties and relationships can be queried by anyone. SMW is an extension to the popular MediaWiki platform that powers Wikipedia. We chose it for the 959NG web site because (i) users are familiar with wikis and comfortable with creating and editing pages and (ii) we were not sure at the outset about the information we wanted to capture for each species and its genome sequencing status. The SMW platform allows for a flexible data architecture and for new kinds of information to be stored and queried as sequencing technologies and the scope of genome projects change. For more details on the kinds of complex relationships and queries possible, see Kumar *et al.* [13].

1.3 Thesis structure

The rest of this thesis is organised as follows. Chapters 2 and 3 describe the workflows I developed to assemble and annotate four nematode genomes: *Caenorhabditis sp. 5*,

Meloidogyne floridensis, *Dirofilaria immitis*, and *Litomosoides sigmodontis*. In Chapter 2, I briefly introduce the problem of assembling short reads into genomes. Based on inputs from other members of the Blaxter Lab, I evaluated tools and wrote scripts that addressed every stage of the genome assembly process, and successfully generated high-quality draft genome assemblies. One of the main innovations presented in this thesis is a new way of visualising the contents of a genomic sequencing run based on a preliminary assembly and using that information to extract and re-assemble the genome(s) of interest. I also developed optimality criteria (and tools for testing these) for evaluating non-model organism genome assemblies for which no genomic resources exist. The results of applying this workflow are presented for each of the four genomes. The discussion section of this chapter lists some of the assembly insights gleaned from these projects. The chapter and its accompanying scripts and workflows (listed in Appendix A and available at <http://github.com/sujaikumar/assembly> [57]) should be a useful set of best-practice resources for anyone embarking on a nematode or other metazoan assembly project.

In Chapter 3, I describe how I used an existing annotation pipeline in a novel way to predict genes for the four newly assembled genomes. I put the annotation results in context by comparing these gene prediction sets with 16 other nematode gene prediction sets, the first such large-scale comparison. I also functionally annotated and compared all 20 genomes afresh because functional annotations were not available in a standard format for most of the existing nematode genome resources. These large-scale comparisons revealed interesting features that could point to biological or methodological differences among the genomes.

In Chapter 4, I describe how the 20 genomes collated in the previous chapter were used to test a recently proposed hypothesis about how conserved non-coding elements (CNEs) define a phylum body plan. Previous research on nematode CNEs had been based on just three genomes from one genus, so the set of 20 genomes provided a robust data set to see if the previous observations remained valid on a phylum-wide scale. I developed fast comparative genomics pipelines and tools to identify non-coding DNA across more than 20 genomes. My findings conclusively demonstrate that no CNEs were shared across the whole phylum, and therefore these elements are unlikely to be connected with the phylum body plan.

A second aspect of genome evolution is discussed in Chapter 5. In collaboration with David Lunt (University of Hull), Mark Blaxter, and Georgios Koutsovoulos (both University of Edinburgh), I assembled the genome of *M. floridensis* (as described in Chapter 2), and compared its coding sequences (CDSs) with those of two other *Meloidogyne* species to understand the possible hybrid origins of *Meloidogyne* species. In addition to the genome assembly, my contribution was the creation, analysis, and visualisation of thousands of phylogenies for CDS clusters made up of the three species *M. hapla*, *M. incognita*, and the newly sequenced *M. floridensis*. We concluded that hybrid speciation was more common and complex than previously thought.

Both Chapters 4 and 5 are examples of the kinds of studies that would not have been possible without complete genomic resources. The genomic resources that I created will be of use to the wider nematode community. The genome of the gonochoristic (equally divided male and female population) *Caenorhabditis sp. 5* can be used to understand the evolution of the androdioecious (hermaphrodite and rare male population) *C. briggsae* and *C. elegans* genomes. *M. floridensis* is a destructive polyphagous plant-parasitic root-knot nematode, and its genome will help further elucidate how it evolved and how we can control its effects. Both *D. immitis* and *L. sigmodontis* are filarial parasitic nematodes that contain endosymbiotic *Wolbachia* genomes. Understanding the genes, pathways, and regulatory sequences in these species will help us combat human and animal filariases in the future. The final section (Chapter 6) summarises the findings of each chapter and suggests future directions for the methods and results presented in this thesis.

2 Assembling four nematode genomes

2.1 Introduction

This chapter describes the creation of high-quality draft genome assemblies from next-generation sequence data for four nematode species—*Caenorhabditis sp. 5*, *Meloidogyne floridensis*, *Dirofilaria immitis*, and *Litomosoides sigmodontis*—using low-cost Illumina sequencing. The first part is an overview of sequencing technologies, sequencing strategies, and assembly software. This is followed by a detailed description of the workflow that I created and refined for assembling these four genomes, as well as the assembly metrics and software tools and visualisations that I developed as part of the workflow. The results of applying this workflow to the four genome projects are described next. These results demonstrate that it is possible, even from short-read data, to obtain whole-genome assemblies of "Improved High-Quality Draft" standard as described in Chain *et al.* [58].

The workflow in this chapter addresses many of the pitfalls of sequencing metazoan genomes, from sequence quality to contaminant issues, and can be a useful checklist of best practices for anyone sequencing a comparably sized genome (100-200 Mbp) using Illumina short reads. The sequencing technology and strategy sections of this chapter have been adapted from a previously published review article (Kumar *et al.* [11]). The idea of using a preliminary diagnostic assembly to visualise and then separate contaminants or co-bionts has been published in Kumar and Blaxter [59].

2.1.1 Sequencing technologies for *de novo* genomes

The sequencing of *C. elegans*, the first nematode and first animal genome to be sequenced [4], was completed over a decade ago using Sanger dideoxy technology [60]. At that time, sequencing the 100 Mbp genome to ten-fold depth required a decade of work and cost approximately \$10M. Once the sequencing was completed, similar resources were required to finish the genome. Sanger dideoxy sequencing (henceforth referred to as Sanger sequencing) is still considered the gold standard in terms of quality, but because of the high cost and time investment, it is unlikely that Sanger sequencing of nematode genomes will continue in the future.

Sequencing the *C. elegans* genome was based on an array of mapped and ordered large-insert genomic clones, which greatly facilitated assembly [61]. Most genome sequencing today avoids this time-consuming step by using only whole-genome shotgun (WGS) sequencing [62]. As a result, current genome projects typically result in draft genomes with multi-gene sized contigs rather than chromosome-sized sequences. If one's goal is to study chromosome organisation or long-range regulation, then the substantial additional effort required to place

scaffolds on a chromosomal map is necessary and unavoidable. However, many questions about phylogenetics, gene evolution, and shared or novel gene functions can be approached using high-quality draft genomes generated at a tiny fraction of the time and cost of a finished genome.

NGS technologies have dramatically reduced costs and increased throughput, with the trade-off of reduced read length compared with Sanger-sequenced reads (Table 2.1). If a genomic repeat is longer than a read, then the only way to attempt to resolve its position in a genome assembly is to use pairs of reads sequenced from opposite ends of fragments that are longer than the repeats. This strategy was used for Sanger-sequenced reads as well, but is even more important for NGS reads because the latter are shorter than dideoxy reads. Sophisticated assembly programs that use high sequencing depth and multiple libraries with different insert sizes to get around the problems of sequencing errors and repeats have been developed specifically for NGS data.

Each technology has a different range of read lengths and error profiles that affect its suitability for *de novo* genome sequencing projects. The sequencing-by-synthesis technology as used by the Illumina HiSeq2000 platform generates reads up to 100 bases (b) and is currently the workhorse of sequencing projects. The most common Illumina sequencing errors are miscalled bases, and, as a result, higher read-depths are recommended to consensus-correct such errors. Pyrosequencing reads (e.g., from the Roche 454 platform) can extend to 750 b but are more expensive than shorter-read technologies. Roche 454 data are also prone to homopolymer errors that can accumulate and confound assembly algorithms. Life Tech's SOLiD platform uses sequencing by ligation technology to generate short (~75 b) reads in colour-space in which each colour represents four possible di-nucleotide combinations. Although this 2-base encoding makes single nucleotide polymorphism (SNP) calls more accurate when these reads are mapped to a reference sequence, it is not recommended for *de novo* genome sequencing because a single colour-space error can cause the remaining sequence to be incorrectly represented in base-space. Newer technologies (the so-called generations 2.5 and 3) have even longer reads and different error characteristics, but they were not available when the work in this thesis was carried out. They are briefly described in the discussion section of this chapter.

Table 2.1 Sequencing costs, throughputs, read lengths and error profiles

Technology	Approximate read length (bases)	Error model	Recommended sequencing depth	Cost per base (£/€/ \$)	Cost per 100 Mbp genome (£/€/ \$)	Throughput (bases/ day/ instrument)	Time per 100 Mbp genome per instrument (days)
Sanger dideoxy [60] (e.g., ABI 3730)	1000–1500	Gold standard, accurate base quality, typical error probability 0.0001	8X	10^{-3}	10^6	10^6	10^3
Pyrosequencing [63] (e.g., Roche 454 FLX/FLX+)	400–1000	Homopolymer errors, typical error probability 0.001	20–30X	10^{-5}	2×10^4	5×10^8	5
Sequencing by synthesis (SBS) [64] (e.g., Illumina HiSeq2000)	100–150	Typical error probability 0.01, lower quality towards end of read	50–100X	10^{-7}	10^3	10^{10}	1
Sequencing by ligation [65] (e.g., ABI/Life Technologies SOLiD 5500xl)	50–75	Reads are output as colours with 2-base encoding; sequencing errors propagate without a reference sequence	Not recommended for <i>de novo</i> assembly	10^{-7}	10^3	10^{10}	1

Note: Costs and throughputs are order of magnitude estimates as of December, 2011.

2.1.2 Sequencing strategies

Depending on the resources available, previous genome projects of nematodes and other metazoans have used different combinations of sequencing technologies, insert lengths, and depths of coverage to exploit the best characteristics of each and minimise known classes of errors. In the early years of NGS, pyrosequencing (e.g., Roche 454) was more popular than sequencing-by-synthesis (e.g., Illumina GA, GAI, and GAIIx) for *de novo* sequencing of metazoan genomes because the former's read lengths (250–400 b) were 5–8 times longer than the read lengths of the latter (36–50 b), although the longer read lengths came at a higher cost per base and had more homopolymer errors. Today, Illumina sequencing can provide 100 b read lengths at one-hundredth the cost per base of 454 sequencing (Table 2.1) without the homopolymer errors of Roche 454 sequencing that can cause frame shifts. Therefore, Illumina sequencing is particularly useful for *de novo* genome sequencing of non-model organisms where we are interested in obtaining complete gene catalogues at a low cost. This section describes some of the combinations of sequencing strategies used previously and describes the resource constraints that led us to use Illumina paired-end sequencing for the four genome projects described in this thesis.

Both Illumina and 454 sequencing can be used with different library preparation protocols. For *de novo* assembly, short-insert (200–700 b) paired-end (PE) libraries are often complemented by long-insert (1–20 kilobase) mate-pair (MP) libraries. While PE data derive from directly captured genome fragments and are thus largely free of chimaeras, construction of MP libraries involves additional manipulations, including circularisation of long DNA fragments that can result in high proportions of chimeric or aberrantly short virtual inserts. MP data are typically used for scaffolding contigs generated from PE data, which are generated in higher coverage. Deep sequencing of the transcriptome can also yield scaffolding information, linking genome sequence contigs that contain exons for a gene that cannot be joined by genome sequence data because of repeats [48].

In 2011 and 2012, genome sequences were published for five nematode species. Each project used different sequencing strategies. The genome of *Trichinella spiralis* was determined using traditional Sanger sequencing [49] with a 33-fold base coverage in the final assembly. Bacterial artificial chromosome clones and multiple-size insert clone libraries were used to scaffold the 64 Mbp genome. The *Bursaphelenchus xylophilus* [50] genome was sequenced using Illumina PE and Roche 454 single-end (SE) reads for basic contig generation and Roche 454 MP reads for scaffolding contigs. For *Caenorhabditis angaria* [48], Illumina PE (from libraries with multiple insert sizes from 200–450 bp) totalling 170-fold coverage were used, and then deep transcriptome data (Illumina RNA-Seq) were used to improve this assembly. This was the first genome project to use RNA-Seq reads to scaffold genomic contigs. Two versions of the *Ascaris suum* genome have been released. As part of an extensive transcriptome sequencing project, Wang *et al.* [66] generated an assembly using Roche 454

and Illumina data from short insert libraries and mate-pair data from 5.5 kilobase libraries sequenced using Sanger sequencing. Jex *et al.* [36] used a mix of Illumina PE 170 bp and 500 bp PE reads, scaffolded with Illumina MP data from 800 bp, 2 kbp, 5 kbp and 10 kbp libraries. Interestingly, these long-insert MP libraries were generated from DNA that was whole-genome amplified using strand-displacing isothermal amplification, a technology that holds great promise for future nematode genome projects where starting materials may be limiting. The *D. immitis* genome [51] was assembled using two libraries of PE reads and one library of MP reads, but the MP library was excessively contaminated with short PE fragments and had to be used as an SE library to prevent mis-assemblies. Although already published, this genome was re-assembled (as described in this thesis) using a new process to remove the contaminating short fragments, improving the genome considerably.

Which sequencing strategy is best for a draft nematode genome between 50 and 300 Mbp in size? If starting DNA material is not limiting and the sequencing budget is about £10,000, then the best possible assembly would probably be obtained using the strategy described by Gnerre *et al.* [67]. The ALLPATHS-LG assembler described in that paper uses a combination of Illumina overlapping PE libraries (e.g., a pair of 100 bp reads from a 180 bp fragment so that the ends of the reads overlap) to build high quality contigs, Illumina short-insert PE libraries (200-700 bp) to resolve small repeats, and one or more long-insert MP libraries (Illumina or Roche 454) to build scaffolds across longer repeats. Using this strategy (Illumina PE, Illumina MP 3 kbp, and Illumina MP 6–14 kbp) and the ALLPATHS-LG assembler, the Broad Institute recently completed a 2.4 Gbp assembly of the *Chinchilla lanigera* mammalian genome (GenBank accession AGCD00000000) with a scaffold N50 of 21.9 Mbp (i.e., more than half the genome was in scaffolds of this size or longer). Given that the longest *C. elegans* chromosome is only 20.9 Mbp, MP sequencing and the ALLPATHS-LG assembler appears to be the best way to proceed. This assembly may have a few scaffolding errors, but, overall, ALLPATHS-LG has been shown to perform well on large mammalian genomes (see next section on comparing assemblers).

MP libraries are essential for establishing long-range contiguity and resolving repeats longer than a few hundred bp. Although MP sequencing is more expensive and the libraries take much longer to prepare, MP sequencing is highly recommended because it provides critical information for spanning repeats and providing more contiguous assemblies. MP libraries do have two major problems—the library protocols do not always work, and they require large amounts of starting material. Both Illumina and Roche 454 have been used for MP libraries, and although the 454 MP protocol has thus far proved more robust, both protocols can erroneously generate chimeric constructs and an abundance of short-insert fragment pairs that can confound assembly algorithms that expect clean and accurate data. In fact, the authors of the leafcutter ant genome [68], which has a scaffold N50 of 5.2 Mbp, acknowledge that their scaffold sizes are likely to be inflated due to chimeric joins in their 454 MP libraries. Additionally, because most nematodes are less than 1 mm long and yield only

picograms of DNA per individual, MP libraries are difficult to construct, as they need many micrograms of DNA as starting material.

Aside from not being able to make MP libraries, having only tiny amounts of DNA from individual nematodes is also a problem because many individuals have to be pooled for some projects, leading to high levels of heterozygosity in a sample (approximately 1 in 200 bp is expected to differ between two individuals of the same species in a population). A smaller worm will require more individuals leading to a more heterozygous sample. Nematode sequencing projects in the past have inbred worms for many generations to get clonal populations. However, that may not be an option for many current projects because of a lack of resources and time, and because some nematodes are hard to culture. Of the four species in this thesis, only *Caenorhabditis sp. 5* was self-crossed, but only for seven generations.

Whole-genome amplification (WGA) has previously been used to generate sufficient quantities of DNA from tissues of single *A. suum* for MP libraries [36]. This opens the prospect of using WGA on single nematode specimens, although the mass of DNA input from *A. suum* used by the BGI team (200 ng) is much more than is present in most individual nematodes (one *C. elegans* adult contains ~200 pg). Proof that amplification does not overly bias sequencing coverage or generate chimaeras that mislead assembly algorithms would be a major advance. Sequencing from single nematodes will reduce the assembly issues arising from extremes of heterozygosity observed in wild populations and will allow researchers to select specimens directly from environmental samples.

The data for the genomes assembled in this thesis were generated before these general rules for data generation were developed. MP data were generally not available (except for *D. immitis*) and only a very limited budget was available for each genome. Therefore, Illumina short-insert PE sequencing was chosen as the most inexpensive and reliable sequencing strategy for our *de novo* nematode genome sequencing projects.

2.1.3 Short-read assembly concepts and algorithms

One lane of Illumina HiSeq2000 sequencing can currently generate 40 gigabases of 100 b PE sequence data. Assembling these data into a nematode-sized genome is not trivial, especially when one considers that even high quality raw data have a 1–2% error rate. Additionally, metazoan genomes have repetitive regions that range from a few bp in size (e.g., microsatellites) to ~10 kbp in size (e.g., retrotransposons). It is impossible to place short reads correctly if they belong to a repeated region longer than the sequenced DNA fragments. Thus, the main challenges for any short-read assembly algorithm are how well it deals with sequencing errors, and how it resolves repeats using pairs from fragments that are longer than the repeated elements. In this section, I provide a brief overview of the

different types of *de novo* assembly algorithms and describe the algorithms used in assembling our four nematode genome projects.

Some terms have multiple specific meanings in the genome assembly literature and so I define them explicitly here. The term *contig* is generally used to indicate any contiguous sequence that has been assembled by overlapping individual reads. However, a more precise terminology was used by Myers *et al.* [69]: a *unitig* describes a sequence assembled unambiguously by overlapping individual reads, with no gaps or overlaps from any other reads; a *contig* is the consensus sequence made up of many reads or unitigs, and may have small gaps of known length indicating unknown bases; and a *scaffold* indicates the correct placement of unitigs and contigs based on longer fragment read pairs, but also includes sequence gaps ("N"s) that represent unknown amounts of missing sequence.

There are four main classes of assembly algorithms for NGS as summarised in Table 2.2. A brief description of each type is presented here:

1. Greedy Extension (GE). Each read or contig is extended by adding the best overlapping read or contig on each step. Memory requirements are linear because only the best overlap is stored. Unfortunately, because the algorithm only considers the best overlaps, local maxima might be found in preference to globally optimal solutions. These were among the first NGS-specific algorithms for *de novo* assembly but are no longer used because the time taken to find the best overlap increases as $O(N^2)$, where N is the number of reads.
2. Overlap Layout Consensus (OLC). The overlap between every pair of reads is calculated and represented on a graph with the reads represented as nodes and the overlaps as edges. Traversal algorithms compute the best paths through this graph and output the consensus as contigs. OLC algorithms performed well for Sanger-sequenced genomes with a few million 800–1000 b reads at 8X-10X coverage. However, both the time and space requirements of these algorithms scale quadratically with respect to the number of reads, making it impractical to process NGS data sets with hundreds of millions of short-reads despite vastly improved computational resources.
3. de Bruijn Graph (DBG). Pevzner *et al.* [70] proposed a data representation for assembly where each read is split into overlapping k -mers. For example, if k is 31, then bases 1–31 form the first k -mer, bases 2–32 form the second k -mer, and so on. A de Bruijn graph is created where each k -mer is a node, and two nodes are connected by an edge if there is a $k - 1$ b perfect overlap between them. Bubbles and hanging tips are formed in the DBG when there are sequencing errors. These bubbles and tips are "popped" or "pruned" if they have low coverage k -mers. DBG assembly algorithms do not require exponentially more time as the number of reads increases because the graph-building step uses a constant-time lookup for placing each k -mer. DBG assemblers also use different

strategies to reduce the memory requirements of tracking and representing millions of k-mers. As a result of these optimisations, almost all DBG assemblers today can finish a typical *de novo* assembly of a 100 Mbp nematode from 200 M pairs of 100 bp PE sequencing in a few hours on a machine in less than 100 GB RAM.

4. String Graph (SG). Myers [71] proposed an SG structure for storing read overlaps and using flow analysis to find the paths through repeat edges representing unique sequence. Conceptually, this is similar to a k-mer based DBG structure except that the reads are not split. To get around the time-consuming problem of finding all possible overlaps as in OLC assemblers, a compressed read data structure known as a Ferragina-Manzini index (FM-Index) was used by both the SG assemblers that have been published thus far [72, 73].

Table 2.2 Classification of NGS assembly algorithms

Type	Characteristics	Useful for	Examples
1. Greedy Extension (GE)	<ul style="list-style-type: none"> • Reads or contigs added one by one till no more can be added • Can get stuck at local maxima 	<ul style="list-style-type: none"> • Bacterial genomes with few reads and lower complexity 	SSAKE [74], SHARCGS [75], VCAKE [76]
2. Overlap Layout Consensus (OLC)	<ul style="list-style-type: none"> • Reads as nodes and overlaps as edges • Initial pairwise comparisons are computationally expensive 	<ul style="list-style-type: none"> • Longer reads such as Roche 454 • Small read sets up to a few M reads 	Celera [69], Newbler [63], CAP3 [77], Phrap [78], MIRA [79], CABOG [80]
3. de Bruijn Graph (DBG)	<ul style="list-style-type: none"> • Reads split into overlapping k-mers • k-mers as nodes and k - l perfect overlaps as edges 	<ul style="list-style-type: none"> • Short-reads (36–150 b) in large quantities up to hundreds of millions of reads 	Velvet [81], ABySS [82], SOAPdenovo [83], CLC [84], Cortex [85]
4. String Graph (SG)	<ul style="list-style-type: none"> • Generalisation of DBG, using whole reads instead of read k-mers • Uses less memory than DBG, but more time to build indexes 	<ul style="list-style-type: none"> • Longer NGS reads, and combining reads from multiple platforms • Keeping polymorphisms intact 	SGA [72], Fermi [73]

For a comprehensive review of GE, OLC, and DBG assembler algorithms till 2010, see Miller *et al.* [86]. Currently, almost all Illumina sequencing projects have used DBG assemblers, as DBGs can handle millions of short reads and perform well even on highly repetitive plant and animal genomes, provided MP libraries are available. In theory, SG assemblers could be even more accurate and memory efficient, but DBG assemblers have been around longer and therefore are more widely used.

Systematic comparisons of the performance and accuracy of all of these assemblers were rare till 2010. However, three studies in 2011 [87-89] and one in 2012 [90] evaluated several assemblers. Of these, the most comprehensive, systematic, and unbiased assessment was performed by the Assemblathon study [89]. Unlike the other studies where each assembler was tested by the same set of authors, Assemblathon instead released a synthetic but realistic data set (consisting of both PE and MP short-reads) and invited submissions of assembled sequences from groups around the world. The Assemblathon team then evaluated each assembly based on existing and new metrics to assess contiguity and errors. All Assemblathon metrics were designed for evaluating a test assembly against a known reference sequence, such as the scaffold NG50 (SNG50), contig-path NG50 (CPNG50), scaffold-path NG50 (SPNG50), correct-contiguity 50 (CC50), intrachromosomal joins, and interchromosomal joins. Although no single assembler topped every metric, ALLPATHS-LG, SGA, and SOAPdenovo consistently performed well on most metrics. The study also found significant differences between the best and worst assemblies, and therefore the choice of assembly algorithm matters. The best assemblers were capable of generating ~100 kbp regions without any errors or gap (contigs), and roughly ~1 Mbp regions without errors, but with gaps (scaffolds).

In another recent assembly assessment study, Genome Assembly Gold-standard Evaluations (GAGE), Salzberg *et al.* [27] tested assemblers on read sets from four organisms, only one of which, *Bombus impatiens* (bumblebee), had no known reference sequence. As in the Assemblathon, the other three read sets all had known reference genomes: *Staphylococcus aureus* (2.9 Mbp), *Rhodobacter sphaeroides* (4.6 Mbp), and human chromosome 14 (88.3 Mbp). On *S. aureus*, ALLPATHS-LG performed best with a scaffold N50 of 1.09 Mbp without any errors, although Bambus2 [91] followed closely with 1.08 Mbp and no errors. MSR-CA [92] had a larger scaffold N50 of 2.4 Mbp, but had 3 errors, therefore the corrected N50 (calculated after splitting scaffolds at errors) was only 1.02 Mbp. For the eukaryotic read set (human chromosome 14), ALLPATHS-LG had an almost chromosome-sized scaffold N50 of 81.6 Mbp. Because this assembly had 45 errors, the error-corrected scaffold N50 was 4.7 Mbp, which was still much better than SOAPdenovo, the next best assembler, with an error-corrected scaffold N50 of 0.214 Mbp.

Based on these two comparisons, ALLPATHS-LG consistently emerged as the best assembly software. Unfortunately, ALLPATHS-LG could not be used on our nematode read sets as it required a short-insert library with overlapping pairs and a high-quality MP library, neither

of which were available for our samples. Preliminary attempts to run SGA and SOAPdenovo on our data sets resulted in considerably worse assemblies than with other DBG assemblers, so they are not included in our comparisons. Table 2.3 shows the assemblers tested for the genome assemblies in this thesis. These attempts also show that different data sets do better with different assemblers, and that more research is needed to define the characteristics of data that make one assembler more suitable than another. Only DBG assemblers were used on our genomes. At least three different assemblers were tested on each data set with a small range of best-guess parameters, as it would be prohibitively expensive and time-consuming to test every available assembler with an exhaustive parameter set. Even so, the assembly optimality criteria and the workflow described in this chapter include many general principles that are not specific to any one assembly algorithm, and should improve any assembly.

Table 2.3 Assemblers tested

Assembler	Used for	Reason
CLC [84]	Preliminary assembly, Stringent reassembly	<ul style="list-style-type: none">• Extremely fast• Low memory usage• Familiarity with behaviour and parameters
Velvet [81]	Stringent reassembly	<ul style="list-style-type: none">• Used in many genome projects• Familiarity with behaviour and parameters
ABYSS [82]	Stringent reassembly	<ul style="list-style-type: none">• Useful intermediate files and visualisation tools• Low memory usage per node on a compute cluster• Used in many genome projects• Familiarity with behaviour and parameters
ALLPATHS-LG [67]	No	<ul style="list-style-type: none">• Needs overlapping read-pairs from small-insert fragments and high-quality mate-pair libraries
SGA [72]	No	<ul style="list-style-type: none">• Preliminary tests resulted in highly fragmented assemblies on our data• Preliminary tests showed that the index-building stage took several days to run on our computing resources
Ray		<ul style="list-style-type: none">• Preliminary tests showed that the assembler took several days to run without completing on our computing resources
SOAPdenovo [83]	No	<ul style="list-style-type: none">• Preliminary tests resulted in erroneous scaffolds

2.1.4 Assembly optimality criteria

De novo assembly is not a one-button solution, where one data set gives one assembly. For any input read set, many assemblies can be generated using different assemblers, different parameters for each assembler, different pre- and post-processing steps, and even different combinations of assemblers. Sequencing projects and publications typically report just one assembly even though they may have internally performed many assemblies.

Although the Assemblathon study [89] defined several new and very useful metrics for evaluating assemblies, none of the metrics were applicable to the task of evaluating assemblies where the true sequence is not known, and thus could not be used for picking an optimal assembly for our genome sequencing projects. They did find that the scaffold N50 was highly correlated with the scaffold NG50, SPNG50, and CPNG50, so the scaffold N50 was used as one measure of contiguity. The GAGE study [27] also did not define any new way of evaluating *Bombus impatiens* assemblies for which they had no reference genome, and the only advice was to "interpret scaffold and contig N50s with caution".

As part of my efforts to assemble four nematode genomes, I defined assembly optimality criteria that let us objectively choose assemblies for further analysis. Apart from the scaffold N50, most of our criteria are novel: I have not seen them defined elsewhere in other genome sequencing projects. However, they are fairly obvious and we would be very surprised if other researchers had not already used them (and perhaps not reported for reasons of brevity).

Given a choice of assemblies from a read set, the optimal assembly will have the highest sequence contiguity. Typically, sequence contiguity is measured using the scaffold N50. A larger N50 is normally assumed to be better, but one can get higher N50 values by discarding short sequences from the set of assembled sequences, or by incorrectly concatenating contigs and scaffolds. We found that visualising the cumulative scaffold and contig lengths across the entire assembly sorted by sequence size provided a richer view of assembly contiguity [93]. Other researchers have used length histograms in the past, but our visualisation of cumulative lengths not only showed which assembly had the longest scaffolds, but also indicated which had an abundance of short scaffolds, and which had the greatest span (i.e., the sum of the lengths of all assembled sequence). The perl script for plotting cumulative length curves (see cartoon example in Figure 2.1) for scaffolds, contigs, and runs of Ns in an assembly is listed in Appendix A (scaffold_stats.pl). The results for each genome later in this chapter demonstrate the value of visualising these metrics rather than simply reporting them as a single number.

The most optimal assembly should also have a cumulative span and longest sequence that is in line with expectations. An independent estimate of genome size should be used where possible (e.g., from flow cytometry), but if no such data are available, then the genome size

of a closely related genome can be used to estimate genome size. For example, we would expect a filarial onchocercid nematode genome assembly to span approximately the same size (~95 Mbp) as that of the previously sequenced *Brugia malayi*, which is closely related. A considerably smaller or larger assembly is not impossible, but should be checked. While checking assembly span, it is also very useful to check the span of uncalled bases (Ns) as they might indicate scaffolding errors. Such checks are described in more detail in the results section for *D. immitis* later in this chapter, where such a check was necessary for spotting gross mis-assemblies when using mate-paired data. Carrying out similar checks on other public nematode genome assemblies by other research groups also enabled us to identify a sub-optimal assembly of *Caenorhabditis* sp. 11 (unpublished, available at ftp://ftp.wormbase.org/pub/wormbase/releases/WS230/species/c_sp11), where two large scaffolds measuring 33 Mbp and 21 Mbp were found in a genome assembly with a total span of only 79 Mbp. This finding was highly unlikely given that most *Caenorhabditis* species have genomes around 100 Mbp with a maximum chromosome size of ~20 Mbp.

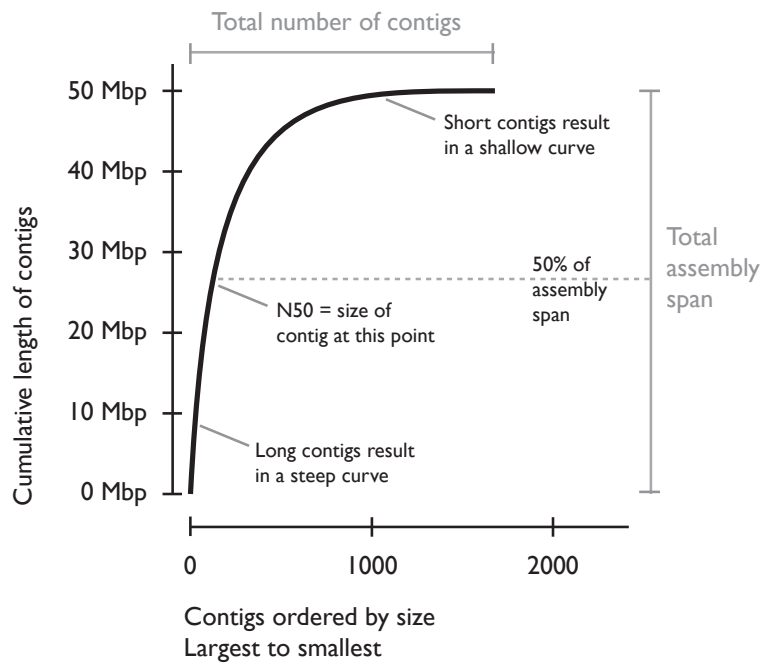


Figure 2.1 Cartoon example of contig (or scaffold) cumulative length curve

To ensure optimal biological accuracy, it is necessary to determine which of the competing assemblies captures the most biologically meaningful sequence. One way of doing this is to use the Core Eukaryotic Genes Mapping Approach [94] or CEGMA. CEGMA uses a set of 248 eukaryotic genes that are found across 6 completely sequenced eukaryotic genomes comprising yeasts (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*), a plant (*Arabidopsis thaliana*), invertebrates (*C. elegans*, *Drosophila melanogaster*), and a vertebrate (*Homo sapiens*). The CEGMA software tool uses sensitive Hidden Markov Models (HMMs) to identify which of these genes are present in a given assembly. A value of above 90% would be expected for a good assembly, and a higher CEGMA completeness percentage indicates a more complete and accurate assembly. The CEGMA protocol is not perfect; even *C. elegans* does not achieve 100% completeness using CEGMA version 2.1 (Figure 2.13) because the HMM profiles for each gene use a six-species alignment, and therefore the HMM profile for a particular gene may be just different enough that the original *C. elegans* gene is not found. The primary purpose of CEGMA, however, is to allow full-length gene models of highly conserved genes to be identified in a new draft genome. These models can then be used to train gene finders for annotation as shown in Chapter 3.

A second approach to assessing biological accuracy is to align the competing draft assemblies to Expressed Sequence Tags (ESTs) for that species, if available, or align to ESTs and protein sequences from closely related species. ESTs are a good representation of the transcriptome of a species and are available for many species, especially nematodes [33, 95, 96]. Although genomes diverge, many proteins (and their coding sequences) are expected to be largely conserved, even for species that diverged as long ago as 30–50 million years ago (MYA). Even if the absolute numbers of ESTs or protein sequences aligned are low, the number of matches found can be used as a relative score because more ESTs or protein sequences will align to the more optimal assembly. For ESTs from the same species, at least 90% of the total span of the EST sequences should be aligned to a genome assembly. Alignments that cover 100% of the EST set are unlikely even if the genome assembly is complete because many ESTs are single-coverage Sanger sequences with sequencing errors that may prevent alignments, and because the library preparation of ESTs may result in chimeric joins between different genes [97]. The methods section describes specific details of how these alignments were selected and filtered for testing contiguity and completeness, and the results section shows the use of these criteria to pick an optimal assembly from several choices.

2.2 Methods

Figure 2.2 shows the workflow I developed and refined to create draft genomes for four nematode species. I first describe the rationale behind each step of the workflow as well as the specific tools and settings used for each step. The latter part of this chapter describes the results of applying this workflow to the four genomes I worked on. Because the workflow evolved over time, not all steps were performed for all four projects. The *Caenorhabditis sp. 5* and *M. floridensis* assemblies were frozen in 2011 but new tools (better scaffolding algorithms in ABySS) and data-reduction approaches (digital normalisation) have been developed since then, which were applied to *D. immitis* and *L. sigmodontis*.

2.2.1 Read quality control

Quality assessment

Sequencing providers typically provide raw reads from the Illumina platform as fastq files [98]. FastQC version 0.10.0 [99] was used to display quality per base, number of Ns per base, GC content across the read, and over-represented k-mers to visualise the error profile of each run before proceeding with other quality control steps. Illumina sequencing proceeds in cycles, with a cycle corresponding to a position on a read. Therefore, the terms cycles and positions can be used interchangeably. By default, FastQC shows every position for the first 10 cycles, and then groups every 5 cycles. This is acceptable for 454 sequencing or longer read sequencing (as it is difficult to visualise hundreds of read positions on a single chart). However, for Illumina sequencing which is typically 100–150 b, every cycle should be visualised, as it is possible for a single cycle to be adversely affected and have very low quality or a high proportion of Ns (as we found in the case of *L. sigmodontis*, see below). By averaging that information over a window of 5 cycles, some information about a particular cycle is lost. A detailed checklist of what to look for in FastQC reports for genome assembly is provided in Appendix A (Use FastQC to check raw read qualities before assembly).

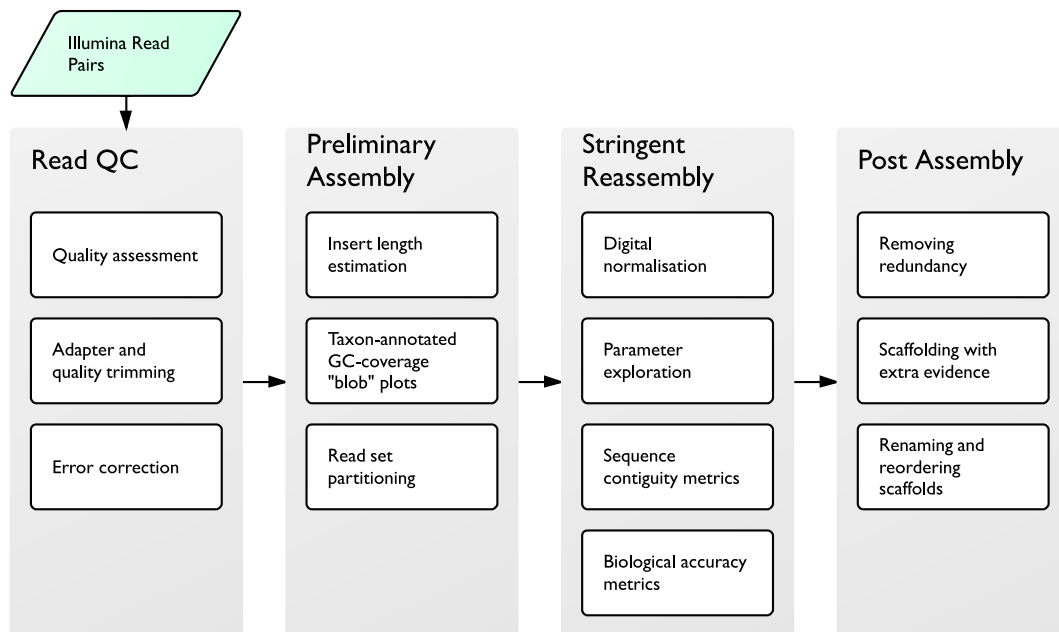


Figure 2.2 Workflow for assembling nematode genomes from short-read Illumina sequencing

Adapter and quality trimming

Illumina reads typically have lower quality bases towards the 3' end of each read [100], and may also have adapter sequences at the 3' end if the fragment being sequenced is shorter than the read length. Adapters and low quality bases should therefore be trimmed from the 3' end of each read or from the 5' end as well if the FastQC plots above indicate any aberrations. Adapter trimming (>10 b match with >90% identity) was done before quality trimming where bases with a Phred quality below 20 were trimmed from the 3' end. Phred quality scores (Q) are a negative log transform of the error probability (P) given by

$$Q = -10 \log_{10} P$$

Therefore, a Phred quality of 20 corresponds to an error probability of 0.01. After trimming, a read pair was discarded if either of the reads in the pair was shorter than a certain length threshold or if it had an N, as the presence of an N typically indicated that the whole read might have problems. However, if FastQC showed that a whole cycle had failed, then reads with Ns were not discarded and were instead corrected as described below.

Initially, I wrote custom shell scripts to perform the trimming and read filtering described above, because existing tools were too slow or too restrictive in that they only allowed trimming one adapter at a time. Currently, I prefer the tools Scythe (<https://github.com/ucdavis-bioinformatics/scythe>) [101] for adapter trimming and Sickle (<https://github.com/ucdavis-bioinformatics/sickle>) [102] for quality trimming and filtering because they are fast (written in C), open-source, regularly maintained, and used in production environments as parts of pipelines (Appendix A: Adapter- and quality-trim Illumina fastq reads using sickle and scythe in one command with no intermediate files).

Error correction

Illumina sequencing faltered on a particular cycle in two cases, and resulted in very low quality bases and many Ns at that position in the read (positions 16 and 65 in the forward reads of two separate runs of the *L. sigmodontis* 600 b libraries). If the affected reads had been discarded because of the presence of an N, 25% of the data would have been lost. To avoid losing useful sequence data, an error correction tool was used.

Reference-free error correcting programs split reads into overlapping k-mers (or seed-spaced k-mers), build a k-mer frequency table, and flag any k-mers with very low frequency as having errors. Low frequency k-mers are then converted to the nearest high frequency k-mer to correct the error. The following tools use this approach to correct errors: Quake [103], DecGPU [104], Shrec [105] and SOAPec (Ruibang Luo, BGI, pers. comm.). SOAPec with default settings was used, as it was the fastest and most memory efficient tool (~16 GB memory and ~6 hours on a 3.0 GHz single-processor machine for ~130 M read pairs).

On the two projects where error correction was tested (*L. sigmodontis* and *D. immitis*), preliminary tests resulted in more contiguous and complete assemblies in both cases. Error correction is therefore recommended even in cases where Illumina sequencing did not fail.

2.2.2 Preliminary assembly

Most *de novo* genome assembly projects move straight from the read cleaning step to the assembly step, trying a few different assemblers and assembler parameters to get the longest and most accurate contigs and scaffolds. We developed the idea of first generating a preliminary assembly from the cleaned reads and then using that to address two issues before proceeding to a final assembly: the presence of contaminants (or co-bionts) in the sample; and inaccurate read-pair insert-size assumptions.

Briefly, the preliminary assembly was used to generate contigs, which were screened after visualising their GC content, read coverage, and best taxon matches in public databases. By separating these contigs (and the read sets that mapped to these contigs) that showed evidence of likely contaminant or co-biont origin, the assembly problem was simplified and more accurate re-assemblies of the partitioned read sets were generated. This approach was useful for not only removing bacterial contamination but also for assembling endosymbiont *Wolbachia* genomes, as described later in this chapter.

To generate the preliminary assembly, the `clc_novo_assemble` tool from CLCBio's Assembly Cell suite (version 4.06) was used. Contigs were generated without considering read pairing information, as the goal was not to get the longest contigs or scaffolds, but to visualise sequence composition by reducing hundreds of millions of reads to a manageable number of contigs. Although other assemblers were tested (Table 2.3), only this assembler was used for the preliminary assembly step as it was the fastest and most memory efficient programme.

Insert length estimation

Pairs of reads from each library were interleaved and mapped back to the preliminary assembly as a reference sequence to estimate library insert sizes accurately. `Clc_ref_assemble` (from CLCBio's Assembly Cell suite version 4.06 [84]) was used in single-end mapping mode, although any other mapping tool could also have been used. Reads were not mapped as pairs because, in that mode, mapping tools typically either require that the insert size range be specified beforehand, or they make the assumption that the insert size is normally distributed.

After the reads were mapped, a custom script (Appendix A: `clc_len_cov_gc_insert.pl` or `sam_len_cov_gc_insert.pl`) was used to extract pair distances and orientation for each library, and plot the information to identify whether any libraries had evident issues such as bimodal distributions or insert sizes that were shorter than expected.

Taxon-annotated GC-coverage plots

Once all the reads had been mapped back to the preliminary assembly, the GC content and read coverage of each contig in the assembly was calculated using the same script as above. A simple scatterplot of the GC versus coverage for each contig showed each species in a mixed sample as its own separate cluster if the two species were present in different molar quantities. To visualise the identity of these clusters, a random sample of contigs was selected and a MegaBLAST search was performed against the NCBI nt database (NCBI's BLAST+ tool suite, version 2.2.25, using default settings [106]). Random sampling was used to reduce the computation time. A custom script was used to parse these hits and assign a taxon name to each contig (Appendix A: `blast_taxonomy_report.pl`). Each contig on the GC-coverage plot was colour-coded by its best matching taxon and visualised using the `ggplot2` tool [107] in R [108] (see workflows in Appendix A). Colour-coding at the level of taxonomic order was found to be most useful for visualising the species present, although a more specific taxonomic level (e.g., genus or species) or a broader level (e.g., phylum) might be optimal for other data.

Read-set partitioning

After the taxon-annotated GC-coverage (TAGC) plot had been used to visualise possible contaminants or co-bionts, a series of negative and positive filters were applied to the preliminary contigs to choose the set of contigs that represented the species of interest. In all four cases, the nematode being sequenced was the primary genome of interest. However, in both the filarial nematodes *L. sigmodontis* and *D. immitis*, endosymbiont *Wolbachia* bacteria were the secondary genomes of interest. Using this method of read-set partitioning, it was possible to separate and assemble the *Wolbachia* genomes successfully. For more details on the motivation behind and advantages of this method, see Kumar and Blaxter [59].

Negative filters identify contigs that definitely do not belong to the genome of interest. An example of a negative filter was the removal of contigs with very high GC content or very low coverage on the basis of the TAGC plot (see *Caenorhabditis sp. 5* and *M. floridensis* results for examples). If the sample had bacterial contaminants that needed to be removed, the negative filter could be a more sensitive search against a specific bacterial contaminant database. Many metazoan sequencing projects automatically screen all raw reads against large bacterial databases. However, screening preliminary contigs rather than individual reads against specific databases is both quicker and more sensitive in identifying data from contaminant species that have not been sequenced previously. Positive filters were also applied in some cases. For example, to assemble the genome of the *Wolbachia* endosymbiont of *L. sigmodontis*, a data set of all *Wolbachia* sequences available in NCBI was created and preliminary contigs that matched this database were selected.

Once the contigs of interest had been identified, the reads mapping to these contigs were extracted along with their read pairs and put into separate files (Appendix A: Extract reads

and their pairs that map to a set of desired contigs in the preliminary assembly). This new read set was re-assembled stringently.

2.2.3 Stringent re-assembly

Cleaned and partitioned read data were re-assembled stringently using different parameters. Unlike the preliminary assembly stage where the goal was to visualise and estimate the characteristics of the raw data (coverage, insert lengths, and species composition), the goal now was to obtain contigs and scaffolds that were as long and accurate as possible. Accurate insert length and coverage estimates obtained from the preliminary assembly were provided to the assembly software.

Digital normalisation

The idea of normalising raw sequence data to improve genome assemblies is very new [109]. High coverage read sets (e.g., >500X read coverage in *L. sigmodontis*) need to be normalised because sequencing errors cause novel k-mers to be generated from the raw reads. These erroneous k-mers require additional memory and can confound DBG assemblers because alternate paths that visit these k-mers in the DBG have to be taken into account. Erroneous k-mers are typically present at very low frequencies relative to the real k-mers generated from error-free reads and a normalisation approach removes the low frequency reads and k-mers prior to the assembly step.

Khmer [109] was used to normalise data on the two most recently assembled species: *D. immitis* and *L. sigmodontis*. Digital normalisation using khmer solved the problem of excessive and uneven genome coverage in whole-genome shotgun NGS. A "high-pass" filter was performed first, breaking up reads into k-mers and removing all reads that did not contribute new k-mers beyond a given coverage (50X in *L. sigmodontis*, 20X in *D. immitis*, corresponding to one-tenth of their read coverages respectively). In *L. sigmodontis*, where the nematode was sequenced at very high coverage, a second, "low-pass" filter was also run, removing reads that had an average k-mer coverage below 5X.

The khmer scripts currently treat each read in a pair as a separate entity and might discard only one of the reads in a pair, causing the assembly software to lose pairing information which is valuable for scaffolding across repeats. Therefore, a custom script was written that pulls in the read pair for every unpaired khmer filtered read (Appendix A: khmer_re_pair.pl).

Parameter exploration

For each final read set, a range of assemblers and assembly parameters was tried. Table 2.3 lists all the assemblers tested and the three de Bruijn graph assemblers that were chosen for stringent reassemblies: CLC, Velvet, and ABySS. The most important parameter in a de

Brujin graph assembler is the k-mer. As a general rule, a longer k-mer will give more accurate assemblies. However, shorter k-mers will give more contiguous assemblies if read coverage is low or if there are many sequencing errors in the read data.

Along with the k-mer, some of the other parameters tested were coverage cutoffs (in ABySS and Velvet), expected coverage (Velvet), and the minimum number of read pairs required to link contigs into scaffolds (ABySS and Velvet). CLC does not allow any coverage or linkage parameters to be changed. The Velvet manual [81] suggests doing a first-pass assembly without setting any coverage parameters and using the resulting coverage histogram to set sensible expected coverage and coverage cutoff values. If allowed to set these values automatically, Velvet uses the median length-weighted k-mer coverage as the expected coverage, and sets the coverage cutoff to half the expected coverage. ABySS does not allow an expected coverage parameter to be set, but it calculates the median coverage and uses the square root of that value to set the coverage cutoff if no coverage cutoff has been specified.

Sequence contiguity metrics

A genome assembly was created for each assembler and parameter set used in the previous step. For each assembly, all scaffolds smaller than 200 bp were removed in keeping with GenBank requirements for WGS projects [110]. Contigs and scaffolds in these sequences were assessed on sequence contiguity metrics using a custom script, as no other tool provided the same metrics at different length cutoffs (Appendix A: scaffold_stats.pl). The script also provided a visualisation of the cumulative length of all scaffolds sorted in descending order of length (as in the cartoon example in Figure 2.1). Along with scaffold lengths, this script also provided metrics on the number of Ns inserted by scaffolding algorithms such as the number of blocks of Ns, the total number of Ns, and the "N" N50. These metrics were very useful because some assemblers incorrectly inserted large blocks of Ns that resulted in very long but likely biologically incorrect scaffolds.

Biological accuracy metrics

The biological accuracy of each assembly was assessed using the following three types of evidence:

1. Core Eukaryotic Genes Mapping Approach (CEGMA), version 2.1, [94]. The CEGMA software searches for 248 single-copy orthologous eukaryotic genes that have been found in all or almost all of the six species used for building the core eukaryotic gene set. The expectation is that any complete genome assembly should have >90% of these genes, and that the better assembly will have better CEGMA scores. CEGMA also reports the average number of copies found for its single-copy genes. The better assembly should have an average copy number closer to 1.0.

2. Alignments to ESTs from the same species. An EST file was created by combining Sanger ESTs, 454 transcriptome assemblies, and RNA-Seq transcriptome assemblies wherever available. For *L. sigmodontis*, a high quality 454 transcriptome assembly was available [93]. For *D. immitis* and *Caenorhabditis sp. 5*, Illumina PE RNA-Seq data were assembled using SOAPdenovo-trans version 1.1.05 [111] with default settings and accurate insert sizes (estimated using the SE assembly plus mapping approach described previously). Further optimisation of the RNA-Seq transcriptome assemblies using other assemblers and parameters would have been necessary if the goal had been to elucidate full length cDNA sequences and their corresponding protein structures. However, such optimisations were beyond the scope of this thesis which concentrates on genome assembly. For the simpler goal of comparing one genome assembly to another, this first-pass RNA-Seq assembly was considered sufficient as the number of contigs and total span of each RNA-seq assembly (Table 3.2) was in line with expectations. BLAT [112] was used with default settings to align the EST fasta file as the query set to each of the genome assemblies being compared. Custom scripts were used (Appendix A: Align ESTs and proteins to assemblies to measure completeness) to calculate the total span of transcript sequences aligned to each assembly. To assess if assembly A was more contiguous than assembly B, only the best alignment for each query sequence was considered. Alignments where >70% of the query sequence was covered with hits were counted as contiguous alignments. Percentage contiguity was calculated as the total number of aligned bases in the query sequences from contiguous alignments, divided by the total number of bases in the query set. To assess completeness, all alignments for each query sequence were considered. Alignments where >50% of the query sequence was covered with hits were counted as complete alignments. As with percentage contiguity above, percentage completeness was calculated as the total number of aligned bases in query sequences from complete alignments, divided by the total number of bases in the query set. The better assembly would be expected to have a higher EST contiguity and EST completeness percentage (Figure 2.3).
3. Alignments to protein sequences from a closely related species. Protein fasta files of closely related species were obtained for each species: *C. briggsae* for *Caenorhabditis sp. 5*, *M. incognita* for *M. floridensis*, and *B. malayi* for *L. sigmodontis* and *D. immitis*. TBLASTN [113] was used to query the protein fasta file against the genome assemblies being compared. As with the EST alignments above, custom scripts were used to combine all TBLASTN HSPs that hit the same query sequence, Both protein contiguity and protein completeness percentages were calculated (Appendix A: Align ESTs and proteins to assemblies to measure completeness). A higher protein contiguity and protein completeness percentage indicated a better assembly (Figure 2.3).

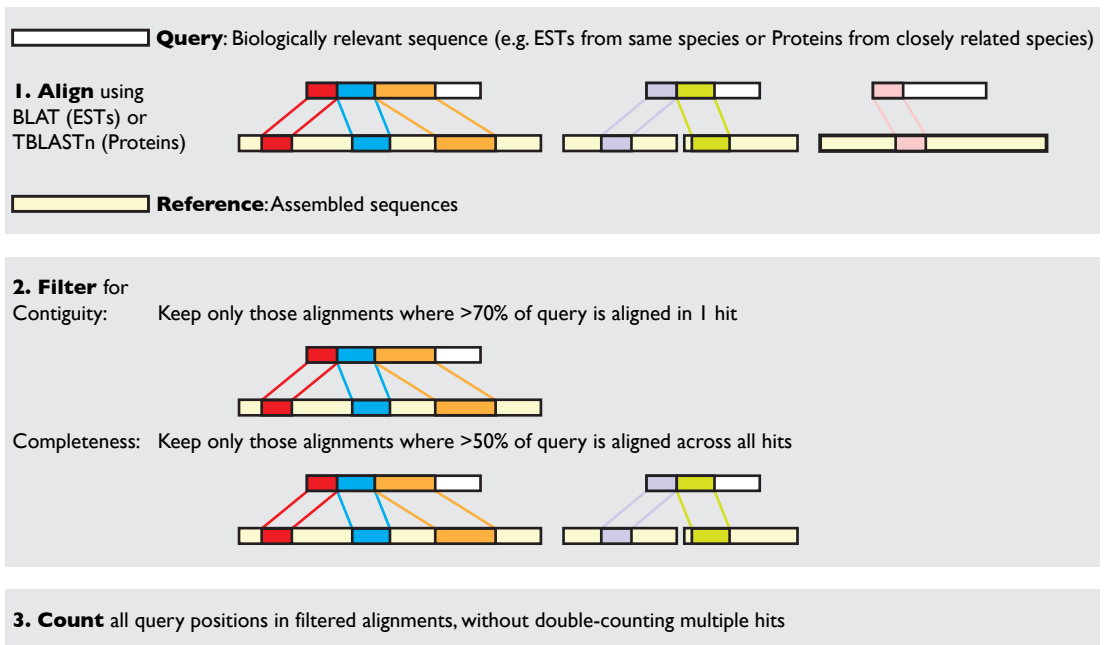


Figure 2.3 Biological accuracy measured using alignments with ESTs or Proteins

2.2.4 Post-assembly

Using a combination of the sequence contiguity and biological accuracy metrics described above, a final assembly was selected and additional steps performed as needed.

Removing redundancy

Assembly algorithms may create nearly identical contigs and scaffolds if a series of closely spaced SNPs or higher frequency sequencing errors are present. CD-HIT-EST (version 4.5.5, [114]) was used with the `-c 99` parameter to remove sequences that were completely contained inside other sequences and were more than 99% identical to the longer sequence over the full length of the shorter sequence (i.e., a global alignment). In the case of *M. floridensis*, the amount of sequence remaining for different values of `-c` from 95 to 99 was tested and `-c 97` was used instead of `-c 99`, as there was a clear inflection in the curve at that value.

Scaffolding with mate-pairs

Longer scaffolds for *D. immitis* were built using a sequencing library that was prepared as a long-insert MP library but turned out to have >50% short-insert pairs. Assembly programs normally incorporate MP data while building scaffolds by mapping the reads back to the assembled sequences and using the insert lengths supplied by the user to estimate how sequences should be arranged in a scaffold and how many Ns should be inserted while bridging two contigs. In our test assemblies, it was found that all the assemblers (SOAP, ABySS, Velvet, and CLC) and stand-alone scaffolders (such as SSPACE [115]) erroneously inserted millions of Ns because they could not distinguish contaminating short-read pairs from true long-insert read pairs.

To minimise the possibility of using short-insert reads incorrectly to bridge contigs, all read-pairs mapped to the same contig were removed. Then, any reads (and their pairs) that mapped to within 600 bp of the end of a contig were removed, because contaminating short-insert read pairs ranged in size from 200–500 b. The remaining read pair mappings were provided to SSPACE (version 1.0, [115]) using strict settings: at least 50 links between two contigs (`-k 50`) were required, and the maximum link ratio between two conflicting contig pairs had to be low (10%), so that only the contig pair with an overwhelming abundance of links was chosen for bridging (Appendix A: Find mate-paired reads that do not map to the ends of contigs and use those to conservatively scaffold the contigs).

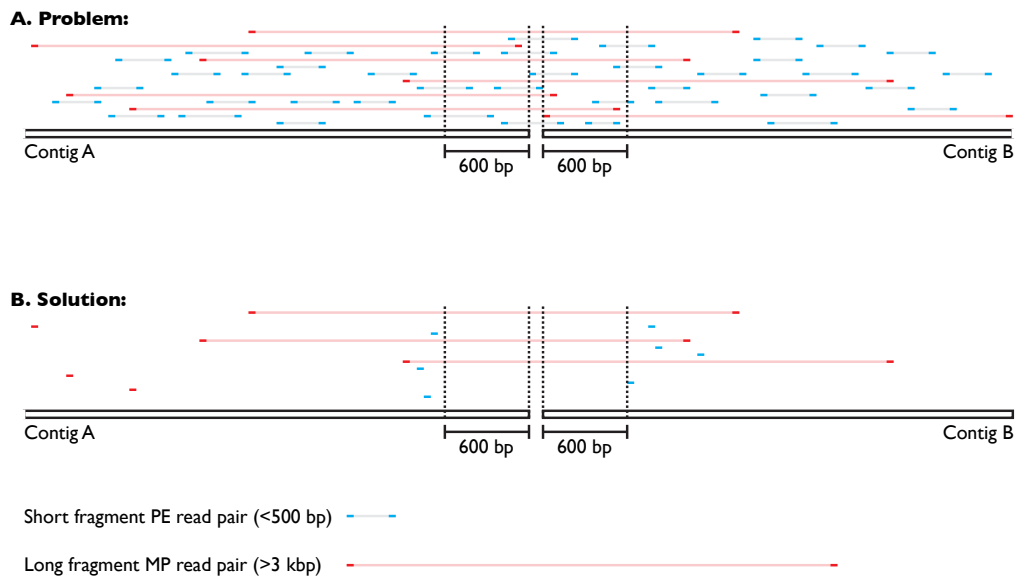


Figure 2.4 Removing short-fragment read-pairs before scaffolding contigs with mate-pairs

A. Problem:

1. Contigs A and B are only 100 bp apart in reality
2. Read pairs map to ends of both contigs
3. Scaffolding algorithm cannot distinguish short (<500 bp) fragment read pairs (in blue) from long fragment (>3 kbp) read pairs (in red)
4. If short fragments outnumber long fragments, scaffolding algorithm assumes contigs are 3 kbp apart

B. Solution:

1. Remove all read pairs that map to same contig
2. Remove all reads that map within 600 bp of contig ends
3. Some reads from long fragments may be removed, but all short fragment reads will be removed
4. Scaffolding algorithm will now be biased by read pairs that come from long fragments

Standardising file formats and scaffold names

Completed genome assembly files were reordered with the longest sequences first, and all scaffolds were renamed with a standard prefix to make downstream analyses easier (e.g., nLs.2.1.scaf00001; "nLs" refers to the nuclear *L. sigmodontis* genome; "2" refers to the read set used, and "1" is the assembly iteration for that read set). The sequence file is named using the prefix (e.g., "nLs.2.1.fna"; ".fna" is a conventional file extension for nucleotide fasta files). In Chapter 3 (Annotating nematode genomes), the prefixes will be extended to include gene and transcript names as well. ABySS, in particular, generates scaffold names that begin with a number, which is allowed in the fasta format standard, but some tools such as CEGMA version 2.4 require an alphabetical starting character. Assembly programs (such as ABySS) can also generate the full range of IUPAC nucleotide bases (such as using R to represent A or G), and these were also changed to Ns as some downstream analysis programs only work with DNA sequences if they have A, C, G, T, or N bases. One might argue that there is no need to version each sequence header if the sequence filename is correctly versioned, but this protocol proved very useful when exploring many assemblies with many read sets. Because several assemblies for each read set were generated, it was critical to generate an assembly freeze at each stage before moving on to the next stage, and this versioning system helped track how the data were generated.

2.3 Results

2.3.1 *Caenorhabditis sp. 5*

Caenorhabditis sp. 5 is usually found in eastern Asia, and is a small (~1 mm long), bacterivorous, and transparent nematode [116]. Out of all lab-culturable *Caenorhabditis* species, *Caenorhabditis sp. 5* is the most closely related to *C. briggsae* [117] and is of particular interest to developmental and evolutionary biologists because it follows the ancestral gonochoristic breeding system (i.e., dioecious, with male and female individuals), unlike *C. elegans* and *C. briggsae*, which are both androdioecious (i.e., the population has male and hermaphrodite individuals). The emergence of androdioecy is not yet understood, but the *Caenorhabditis* genus is an excellent test bed for studying this rare breeding system. *Caenorhabditis sp. 5* also exhibits an exceptionally high level of polymorphism [116], and this hyperdiversity may be a feature of gonochoristic species. Along with the genomes of other *Caenorhabditis* species currently being sequenced (see <http://nematodegenomes.org/Caenorhabditis>), this genome is expected to contribute to the understanding of how gene and genome features relate to breeding strategy.

Sample, Sequencing, and Read QC

Caenorhabditis sp. 5 (strain JU800) was grown up by Asher Cutter (University of Toronto) and DNA was extracted from a sucrose- and detergent-cleaned plate culture of nematodes using proteinase K and phenol-chloroform. The standard Illumina protocol was used for generating two PE libraries with insert sizes 300 bp and 600 bp and sequenced on an Illumina HiSeq2000 instrument using 101 base PE sequencing.

Raw reads were adapter- and quality-trimmed, and paired reads were discarded if either read in a pair was shorter than 35 bases, leaving 26.2 gigabases of sequence in 134.3 M read pairs (Table 2.4). The raw sequence data are available at the Short Read Archive with accession number ERP001495.

Table 2.4 Read data for *Caenorhabditis sp. 5*

Strain, Library	Num of reads and bases	Type of seq	Trimming steps	Post trimming Reads and bp
JU800, PE 300 bp	88.6 M pairs, 17.9 gigabases	HiSeq 101 b PE	Adapter removal from 3' end; Quality < 20 b from 3' end	85.2 M pairs, 16.8 gigabases
JU800, PE 600 bp	52.4 M pairs, 10.6 gigabases	HiSeq 101 b PE	Adapter removal from 3' end; Quality < 20 b from 3' end	49.1 M pairs, 9.4 gigabases

Preliminary assembly, taxon-annotated GC-coverage plot and read separation

A preliminary assembly of the 300 bp and 600 bp libraries taken together gave 86,272 contigs totalling 183.9 Mbp. MegaBLAST with default parameters was used to query 10,000 randomly selected contigs against the NCBI nt database, and the results were used to assign taxon order names to the contigs. The TAGC plots in Figure 2.5 show extensive bacterial contamination. Along with the large blue cluster on the left (order Rhabditida) representing *Caenorhabditis sp. 5*, at least 7 bacterial orders were present in this sample. Two of the orders (Actinomycetales and Pseudomonadales) had more than one cluster each, implying multiple species from that order. The 300 bp library also had higher read coverage for each contig compared to the 600 bp library, as expected from the raw data read counts, although the log scale on the y-axis makes it difficult to discern this.

Some low-level bacterial contamination was expected because the nematodes had been fed on *E. coli* (order Enterobacteriales, median coverage ~5X, GC content ~0.5). However, bacterial clusters present in even higher molar concentrations than the nematode of interest were unexpected. Further analysis using various binning techniques [118-121] would be necessary for a full-scale metagenomic analysis. For the goal of obtaining a draft nematode genome, the taxon composition and coverage information gleaned from this diagnostic plot were enough to proceed with data separation and stringent re-assembly.

Table 2.5 shows the contig removal steps performed on the preliminary assembly and the number of contigs and reads remaining as the result of each step. Based on the TAGC plot (Figure 2.5 - All Libraries Combined), contigs below 10X coverage were first removed because they were identified as bacterial contaminants. Bacterial clusters to the top right of the plot could also have been removed using GC-coverage cutoffs, but a more conservative approach was taken and only contigs that hit bacterial databases were removed. Bacterial order identifications from the TAGC plots were used to create a Bacteria-specific subset of NCBI's nt database. A Nematoda-subset of the nt database was also created. Contigs were queried against both databases and only contigs with high-confidence hits to the Bacteria database were removed. If a contig hit both databases, it was only labelled bacterial if the hit score to the Bacteria database was at least 50 more than the hit to the Nematoda database (Appendix A: Separate contigs based on which taxon-specific blast database they hit better). Reads (and their pairs) that mapped to the final list of putative nematode contigs were re-assembled as described below.

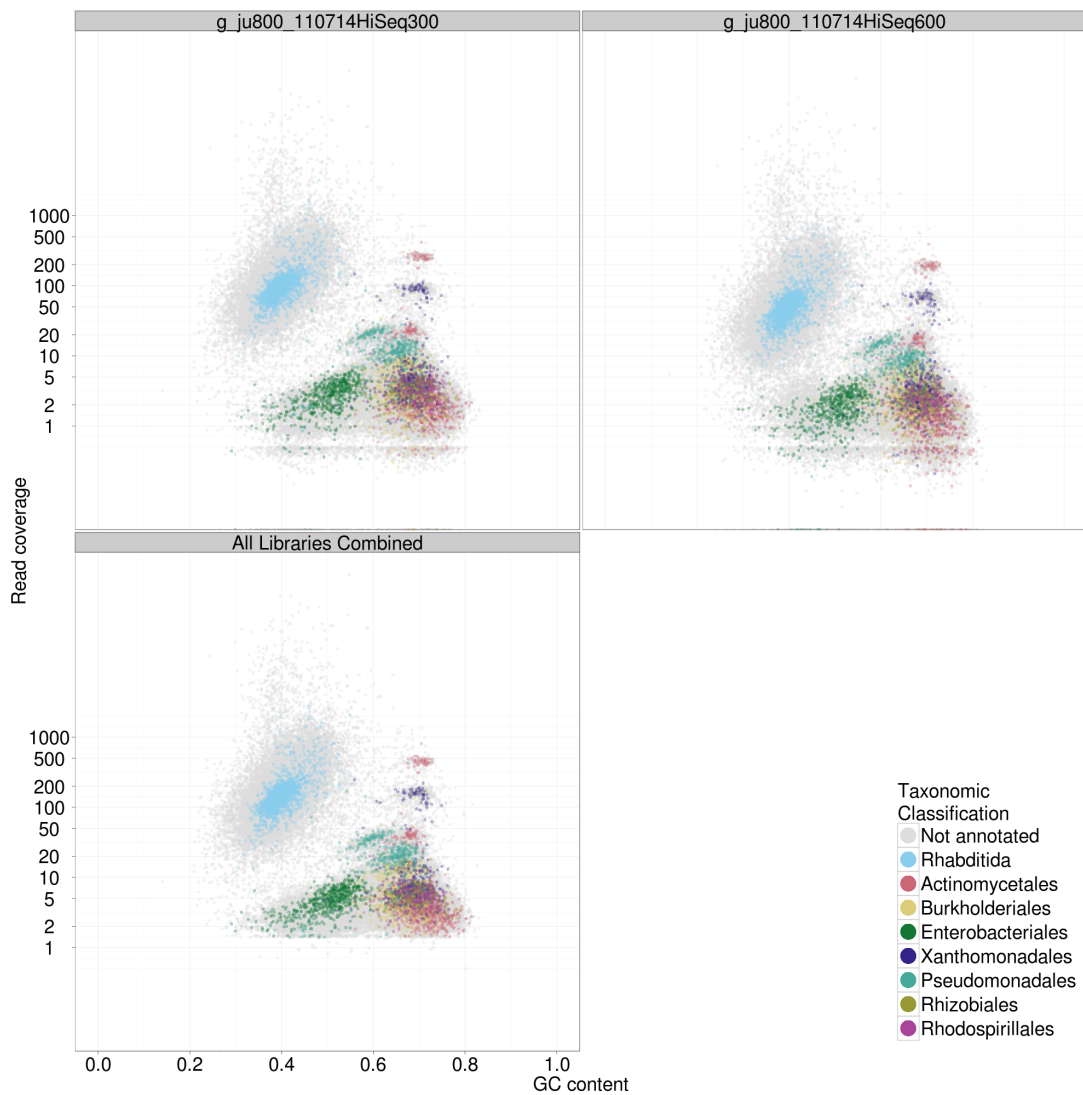


Figure 2.5 Taxon-annotated GC-coverage plot for *Caenorhabditis sp. 5*
 Each dot in these plots represents a contig with a particular GC-content (x-axis) and read coverage (y-axis). The preliminary assembly of the read set yielded 86,272 contigs. Out of 10,000 randomly selected contigs, only 6,440 were assigned a taxonomic order based on a Blastn search against the NCBI nt database. Only taxa representing at least 1% of all annotated contigs are shown.

All three plots in this figure use the same contig set with the same annotations. The only difference is that each plot shows the read coverage of the specified library. The 300 bp library (Top Left) has the same relative composition as the 600 bp library (Top Right) but with higher coverage.

The large Rhabditida cluster on the left is the genome of interest (*Caenorhabditis sp. 5*) whereas the remaining clusters are contaminants that were removed before stringent re-assembly.

Table 2.5 Read separation for *Caenorhabditis sp. 5* preliminary assembly

Process	Number of contigs remaining	Span of contigs remaining	Reads mapped to contigs (both libraries)
Preliminary assembly	86,272	183,862,303	268,537,136
Remove low coverage (cov <10)	44,279	161,073,470	264,718,396
Remove hits to bacterial databases	35,882	129,310,004	234,031,150 (117,015,575 pairs)

Note: The reads mapped in the last row were re-paired, i.e., if only one read of a PE fragment mapped to a contig, it's pair was also pulled in. This set was used for stringent re-assembly.

Stringent re-assembly and post-assembly

To assist with picking the most optimal stringent re-assembly, a set of transcriptome contigs was generated from one lane of Illumina RNA-Seq data, as no Sanger ESTs were available for this species. Asher Cutter (University of Toronto) kindly provided 20.5 M pairs of 38 b Illumina GAIIx PE reads of a 100 bp insert RNA-Seq library from *Caenorhabditis. sp. 5*. These reads were assembled using SOAPdenovo-Trans [111] into 30,756 contigs spanning 12.7 Mbp with an N50 of 792 bp. This assembly was considered a substitute for an EST set and was used both for testing the contiguity of different assemblies and also for annotating the final draft genome as shown in Chapter 3. A protein set of 21,961 sequences from the closely related *C. briggsae* genome (WormBase release WS230) was used to test protein sequence completeness.

Three different DBG assemblers were used to stringently re-assemble the nematode-only reads obtained after the read-separation steps described in the previous section. Although many parameter sets were tried for Velvet, ABySS, and CLC, only the best assembly for each program is reported in Table 2.6 and Figure 2.6. This table shows why biological accuracy metrics are important. The scaffold N50 metric used most in the literature does not tell the full story about sequence contiguity. VelvetK51 is clearly the assembly with the longest scaffolds (steepest cumulative scaffold length curve), but it also has the shortest contigs of the three (shallowest cumulative contig length curve) as seen in Figure 2.6. Coupled with the lower mapping of protein and EST sequences, VelvetK51 is the least optimal assembly.

Velvet's better scaffold metrics but poorer gene-centric metrics indicate that it might be putting together contigs into scaffolds incorrectly and therefore it cannot recover genes that span across multiple contigs as well as CLC can. One possible reason for poorer scaffolding by Velvet could be the default setting for the minimum number of read pairs needed to join contigs into scaffolds (default value 5). This default parameter value has been unchanged since the earliest versions of the software, when less sequence data were generated for each project, and may be too aggressive for current NGS projects that generate much more data. CLC does not allow users to set this parameter but it possibly uses a higher default value, although this is not documented. As with most assembly projects, we only changed a few parameters that we thought were relevant at the time. Unfortunately, we arrived at this explanation after we had chosen the CLC assembly and proceeded with further analysis, otherwise we would have rerun Velvet with different settings.

Overall, CLC performed best on all biological accuracy metrics even though it did not have the best scaffold or contig metrics. This assembly was post-processed to rename the contigs and made available on WormBase as *c_sp5.WS230* without any redundancy filtering.

Summary

Although highly polymorphic [116] and surrounded by many other bacterial species in the sample, we were able to successfully assemble the *Caenorhabditis sp. 5* genome from two Illumina PE libraries. Compared to the other *Caenorhabditis* species sequenced using Sanger sequencing (Figure 2.13), this assembly had a lower scaffold N50 than all the other species, but a higher contig N50 than *C. japonica*, and a higher CEGMA completeness than *C. remanei* and *C. japonica*.

Table 2.6 Comparison of stringent re-assemblies for *Caenorhabditis sp. 5*

	VelvetK51	ABySSK51	CLC
Longest Scaffold	375,732	276,948	383,975
Number of scaffolds	16,384	15,086	15,261
Assembly span	130,934,954	136,855,721	131,797,386
Mean scaffold length	7,991	9,071	8,636
Scaffold N50	31,176	22,090	25,228
GC content	39.40%	39.40%	39.50%
CEGMA completeness	91.94%	95.16%	96.37%
<i>C. briggsae</i> protein contiguity	63.93%	66.35%	67.14%
<i>C. briggsae</i> protein completeness	73.81%	75.78%	76.08%
<i>Caenorhabditis sp. 5</i> EST contiguity	86.95%	91.24%	91.82%
<i>Caenorhabditis sp. 5</i> EST completeness	91.07%	94.11%	94.46%

Note: Figures in bold indicate the best metric in that row.

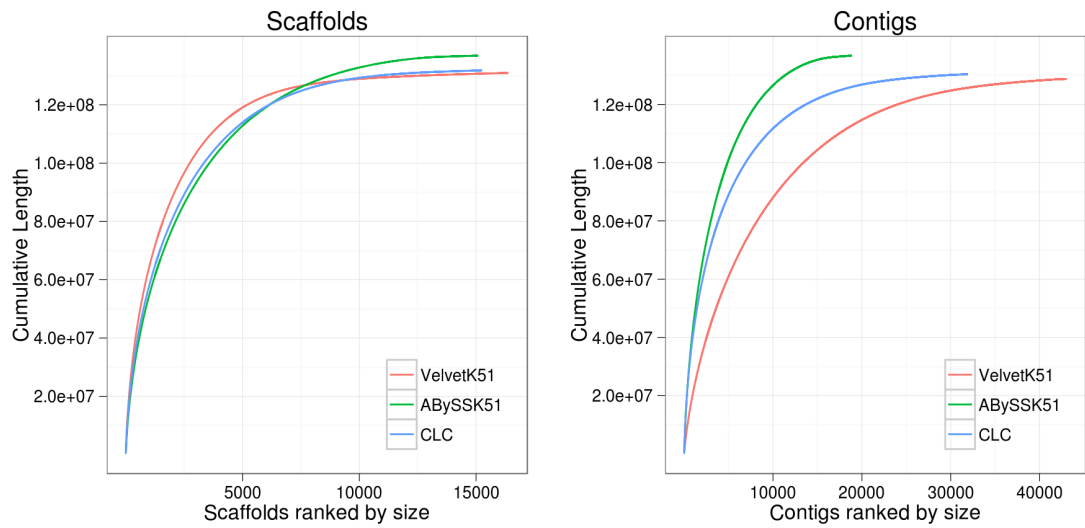


Figure 2.6 Comparison of cumulative scaffold and contig length curves for *Caenorhabditis sp. 5* stringent re-assemblies
 The curve with the steepest starting curves has the longest sequences.
 ABySSK51 had longer contigs and a larger assembly span, but shorter scaffolds.
 VelvetK51 had the longest scaffolds.

2.3.2 *Meloidogyne floridensis*

The genus *Meloidogyne* consists of many dangerous, multi-host, plant-parasitic root-knot nematodes, including *M. floridensis*. *M. floridensis* is a tropical apomict (i.e., it reproduces asexually) like *M. incognita*, and was hypothesised to be involved in the hybrid origin of *M. incognita* because several individual genes sequenced for these two species were identical [122]. Chapter 5 has more details on the background of this sequencing project.

Sample, Sequencing, and Read QC

DNA was extracted from a sample isolated by Thomas Powers (University of Nebraska) and Janete Brito (Florida Department of Agriculture and Consumer Services) and sequenced by the GenePool Genomics Facility of the University of Edinburgh. Because only a few adult female worms were available, the amount of starting material was very limited and only one library could be prepared. A 260 bp insert library was prepared using standard Illumina protocols and sequenced using one lane of an Illumina HiSeq2000 with 101 base paired-end sequencing. This library was submitted to the Short Read Archive (SRA) with accession ERP001338 and contained 71.9 M pairs totalling 14.5 gigabases of raw sequence data. The sequences were adapter trimmed and quality filtered to yield 70.2 M pairs totalling 13.2 gigabases.

Preliminary assembly, taxon-annotated GC-coverage plot and read separation

A preliminary assembly of the cleaned read library resulted in 204,416 contigs spanning 142.8 Mbp. Figure 2.7 shows a TAGC plot of this assembly. The large "blob" or cluster on the left has hits to three nematode orders: Tylenchida, Spirurida, and Rhabditida. The remaining clusters are bacterial, all belonging to the class Alphaproteobacteria. The GC-coverage cluster around 0.65 GC and 200X coverage had very few hits to known bacterial sequences, implying that it might be a co-biont of *M. floridensis* that has not previously been sequenced.

Based on this plot, *M. floridensis* was expected to be present at a median read coverage around 100X, and any nematode contigs with read coverage lower than 5X were presumed to represent low-coverage assemblies of reads with sequencing errors. Contigs with GC and coverage matching Alphaproteobacterial clusters were removed conservatively: GC>.2 cov <5, GC>.3 cov <10, GC>.4 cov <25, GC>.5 cov <50, and GC>.6 cov <200. To identify bacterial contaminants among the remaining contigs, an Alphaproteobacteria subset of NCBI's nt database was created and contigs were removed that aligned to this database (using MegaBLAST) with high stringency settings (1e-20). The final set of preliminary contigs spanned 86.3 Mbp in 94,252 contigs. Only reads mapping to these contigs and their pairs were selected and used for re-assembly.

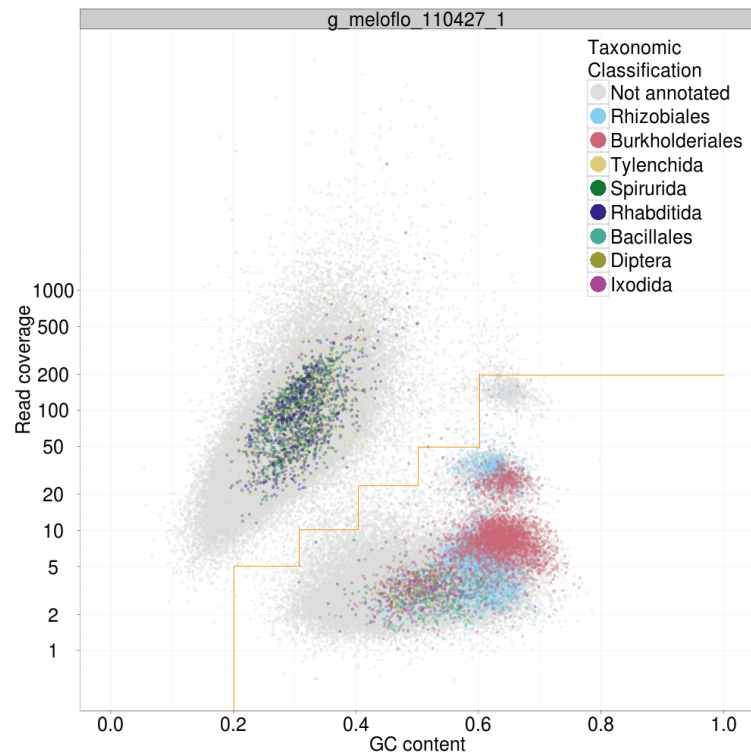


Figure 2.7 Taxon-annotated GC-coverage plot for *M. floridensis*
 The large cluster on the left has contigs annotated as Tylenchida, Spirurida, and Rhabditida (all are nematode orders). The remaining clusters indicate bacterial and other contaminants (Ixodida). The "Not annotated" cluster towards the top right of the plot most likely represents a bacterial order that has not been previously sequenced.

Orange lines indicate the GC-coverage cutoffs used to remove contaminant contigs conservatively as a first step. Matches to Bacteria-specific databases were used to remove the remaining contaminant contigs.

Stringent re-assembly and post-assembly

A stringent re-assembly of the cleaned read set was performed using coverage information estimated from the preliminary assembly above. Several assemblies were tried, three of which are reported in Table 2.7: clc-novo-assemble Version 3.22 with insert size between 200 and 360 (labelled CLC); Velvet (version 1.1.04 [81]) with k-mer 51 with parameters `-exp_cov 50`, `-cov_cutoff 5`, and `-ins_length 260` (labelled VelvetK51); and Velvet k-mer 55 with parameters `-exp_cov 45`, `-cov_cutoff 4.5`, and `-ins_length 260` (labelled VelvetK55).

Although CLC had the longest scaffold, it performed poorly on all other sequence length and biological accuracy metrics. VelvetK51 performed best on the scaffold N50 metric. However, Figure 2.8 shows how the scaffold N50 can be misleading, even for assessing sequence length, as the curve of VelvetK55 is steeper than VelvetK51, implying longer sequences overall. We chose VelvetK55 because it performed comparably on scaffold N50, and performed best on two biological accuracy metrics: CEGMA completeness and *M. hapla* completeness.

Compared to *Caenorhabditis sp. 5* stringent assemblies where Velvet had longer scaffolds but CLC had better gene-centric metrics, CLC had longer scaffolds and Velvet had better gene-centric metrics in the case of *M. floridensis*. The most likely reason for this reversal is the lower sequencing depth for *M. floridensis* which would prevent contigs from being joined spuriously even with a lower default value for the Velvet parameter controlling the minimum number of read pairs needed to scaffold contigs.

Scaffolds shorter than 100 b were removed and CD-HIT-EST was used to remove redundant sequences that were more than 97% identical across their full length to another sequence. The redundancy-filtered assembly is available as nMf.1.0 from <http://meloidogyne.nematod.es>. It spans a total of 99,886,934 bp in 81,111 scaffolds, with a scaffold N50 size of 3,516, and a mean scaffold size of 1,231 bp. These numbers are not as high as those of the *Caenorhabditis sp. 5* assembly because only one short-insert library was used (260 bp). To get a more contiguous assembly, longer insert PE sequencing (500–800 bp) and long-insert MP sequencing (2–8 kbp) would have to be used. Version nMf.1.1 of the assembly (used in Chapter 3) was created by removing all scaffolds smaller than 200 bp from version nMf.1.0, reducing the number of contigs by 22,415, and reducing the assembly span by only 3.2 Mbp. This was done so that the minimum scaffold size would be consistent with the other three species in this thesis.

Summary

The *M. floridensis* genome assembly has a low scaffold N50 of only 3516 bp (version nMf.1.0) and does not seem to be complete, as it has only 60.08% CEGMA completeness. However, as seen in Chapter 5, this assembly was adequate for extracting coding sequences that matched

two other species of the same genus, *M. hapla* and *M. incognita*. The amount of starting material was very limited and the sample was contaminated by Bacteria (at least one of which had not been sequenced previously). Despite these constraints, a usable genome assembly was obtained following the workflow described earlier in this chapter.

Table 2.7 Comparison of stringent re-assemblies for *M. floridensis*

	CLC	VelvetK51	VelvetK55
Longest Scaffold (bp)	83,915	38,991	40,762
Number of scaffolds	92,202	100,286	96,751
Assembly span (bp)	95,003,082	100,996,550	102,424,455
Mean scaffold length (bp)	1,030	1,007	1,058
Scaffold N50 (bp)	1,616	3,454	3,385
GC content (%)	29.30	29.70	29.80
CEGMA completeness (%)	42.74	57.66	60.08
<i>M. hapla</i> protein contiguity (%)	27.66	35.68	35.36
<i>M. hapla</i> protein completeness (%)	44.70	54.98	55.04

Note: Figures in bold indicate the best metric in that row.

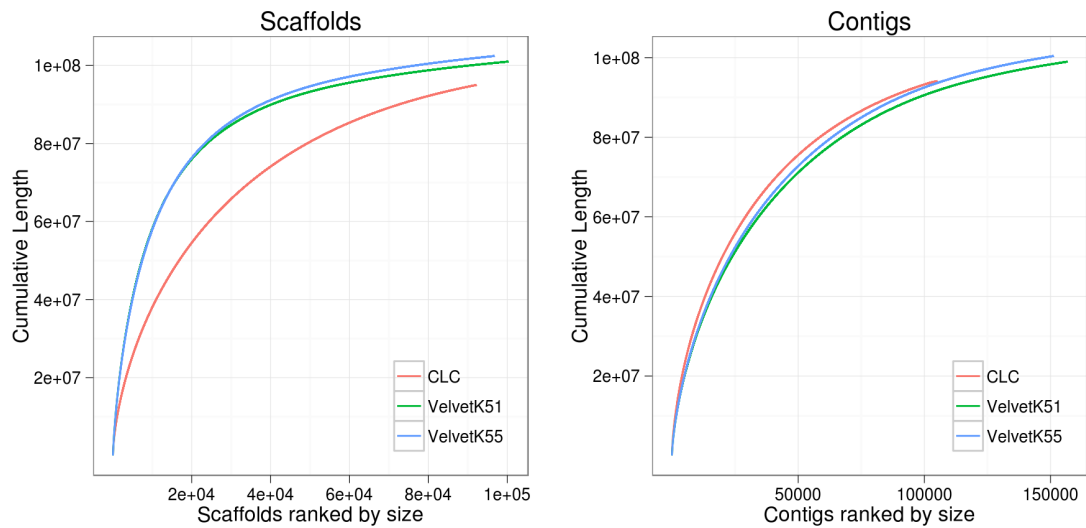


Figure 2.8 Comparison of cumulative scaffold and contig length curves for *M. floridensis* stringent re-assemblies

For each assembly, all scaffolds were ordered by length, and the cumulative length of all scaffolds shorter than or equal to a given scaffold was plotted. The end point of each curve represents the total assembly span and the number of scaffolds in that assembly, whereas the initial slope of each curve reflects the proportion of longer scaffolds. For the plot on the right, each scaffold was split into contigs at every run of 10 or more "N"s. The cumulative contig length curves on the right show that the CLC assembly had longer contigs but shorter scaffolds than the VelvetK51 and VelvetK55 assemblies.

2.3.3 *Dirofilaria immitis*

D. immitis is a filarial nematode parasite of dogs. Also known as the heartworm, it is transmitted by mosquitoes in warmer climatic zones, and is spreading rapidly across Southern Europe and the Americas. Because there is no vaccine, and chemotherapy is prone to complications, the genome and proteome of *D. immitis* are important for identifying potential drug targets, immune modulators, and vaccine candidates. This project was one of the first instances of a collaboration initiated by the 959NG wiki. A preliminary assembly of the strain being sequenced at the Blaxter Lab was put online from where it was downloaded by Pascal Maeser's team at the Swiss Tropical and Public Health Institute. Maeser's team found that their strain was almost identical to the preliminary assembly, and contacted the Blaxter Lab to pool our sequencing resources to get a better genome assembly. The *D. immitis* genome was published in 2012 [51] using a previous version of the genome (version 1.3). The assembly process described in this section used digital normalisation and MP data to generate a more contiguous version of the assembly (version nDi.2.2).

Samples, Sequencing, and Read QC

The Athens (Georgia, USA) strain of the dog heartworm *D. immitis* was sequenced at the GenePool Genomics Facility, Edinburgh. The Pavia (Italy) strain was sequenced at Fasteris, Switzerland. One 4 kilobase long-insert MP library (Pavia) and two short-insert PE libraries (Pavia and Athens) were used. The raw data from these three libraries totalled 31.5 gigabases in 164.4 M read-pairs (Table 2.8) and were submitted to the Short Read Archive (accession ERP000699). All libraries were adapter- and quality-trimmed (Q20), reads shorter than 35 b and reads with Ns were removed, and reads from the 4 kilobase mate-pair (MP) library was trimmed to 50 b to reduce the possibility of chimeric reads. SOAPec with default settings was used to error-correct all reads using k-mer frequencies as described previously. The final trimmed data set consisted of 18.7 gigabases in 145.6 M pairs.

Table 2.8 Read data for *D. immitis*

Strain, Library	Num of reads and bases	Type of seq	Trimming and error correction steps	Post trimming Reads and b	After khmer digital normalisation
Athens, PE 108 bp	40.4 M pairs 8.2 gigabases	GAllx 101 b PE	Adapter and quality trimming and ec	38.4 M pairs 7.3 gigabases	6.4 M pairs 1.2 gigabases
Pavia, PE 340 bp	29.8 M pairs 4.5 gigabases	GAllx 76 b PE	Adapter and quality trimming and ec	19.2 M pairs 2.6 gigabases	12.8 M pairs 1.7 gigabases
Pavia, MP 3.6 kbp	94.2 M pairs 18.8 gigabases	HiSeq2000 100 b PE	Adapter and quality trimming; trim to 50b and ec. Used as single-end.	176.1 M reads 8.8 gigabases	14.7 M reads 0.7 gigabases

Preliminary assembly and taxon-annotated GC-coverage plots

A preliminary SE assembly followed by read mapping and insert-size estimation revealed that the Athens PE library had an actual median fragment length of only 108 bp (with a std. dev. of 15 bp) instead of 250 bp as expected. In the Pavia MP lib, 70% of read pairs mapped to the same contig. Of these, only 45% mapped as out-facing read pairs (RF) orientation with 2–5 kbp insert lengths, and 55% mapped as in-facing read pairs (FR) with 200–500 bp insert lengths. The MP library protocol was the cause of this short-fragment FR contamination. When the MP library was initially sheared and size-selected (4 kbp in this case), the long, linear fragments were circularised with a biotin molecule marking the point where the ends of the fragment meet in a circle. The circularised library was sheared to small (200–500 bp) fragments and the biotin-labelled fragments were selected. All biotin-labelled fragments represented the ends of the 4 kbp MP fragments, whereas the remaining fragments were short PE fragments. The biotin-selection step was an enrichment, and seems to have been inefficient in this case, leading to an excess of short fragments. The MP library was initially used as an SE library because the insert lengths could not be determined reliably. Subsequently, a novel filtering step was used that removed all short-fragment pairs to generate a better scaffolded assembly.

Figure 2.9 shows the taxon-annotated GC-coverage (TAGC) plots for a preliminary assembly of *D. immitis*. All three libraries—Pavia PE, Pavia MP, and Athens PE—showed the presence of contigs from the *Wolbachia* endosymbiont of the nematode (contigs labelled Rickettsiales). The *Wolbachia* genome was present at approximately 10 times the copy number of the nuclear genome in the Athens PE library because the Athens sample was from a female worm [123]. The Athens strain also showed the presence of low-coverage lab contamination, which were expected to be removed during stringent re-assembly with coverage cutoffs. Therefore, no attempt was made to separate the reads belonging to those contigs.

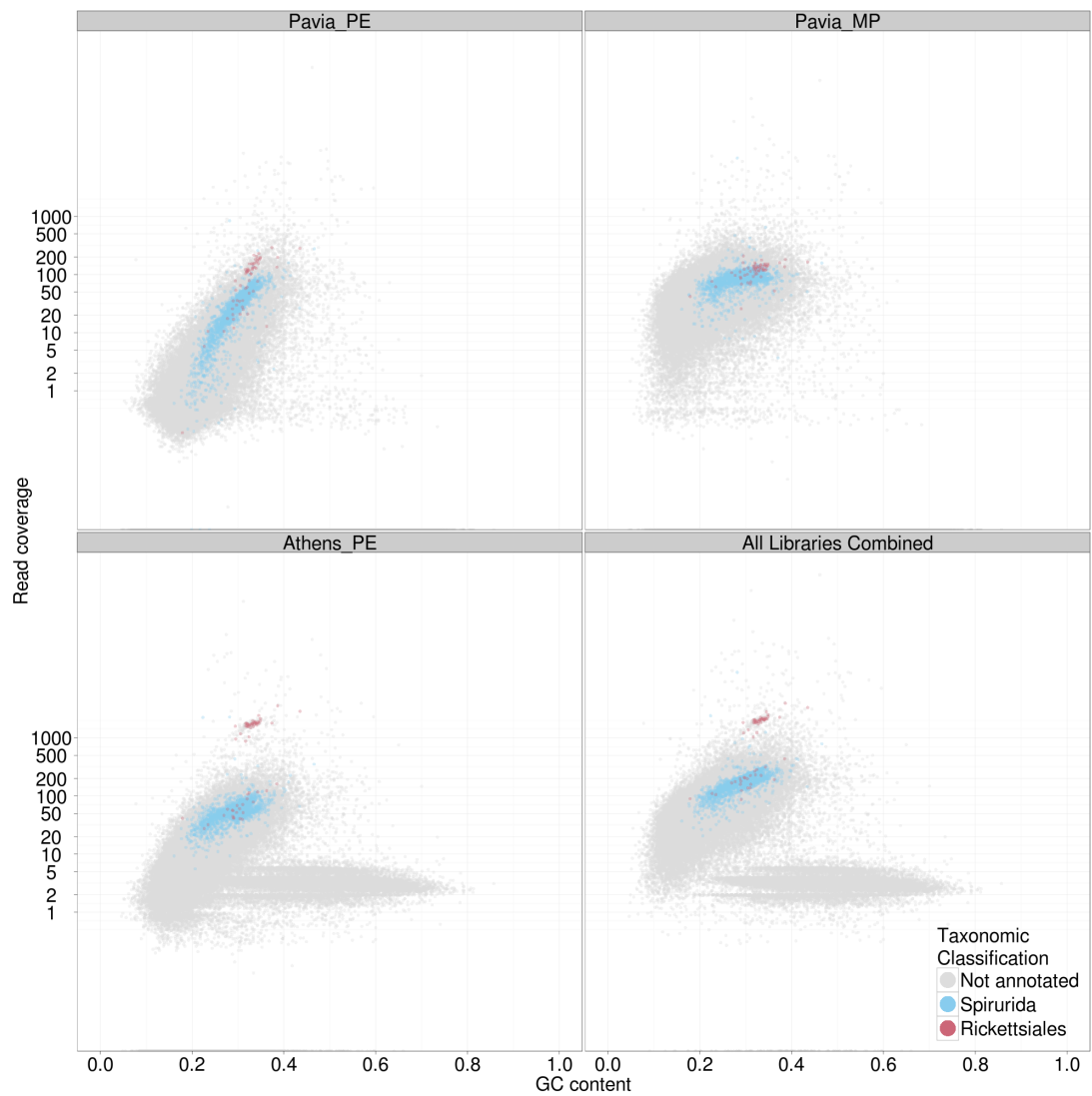


Figure 2.9 Taxon-annotated GC-coverage plot for *D. immitis*
 91,982 preliminary contigs are shown in each plot. Only 2,354 were assigned taxa based on a Blastn search against the NCBI nt database, and only taxa representing >1% of all annotated contigs are shown. The Athens PE library displays a low-coverage contaminant cluster. The Rickettsiales cluster represents the *Wolbachia* endosymbiont and is present at a much higher coverage in the Athens PE library compared to the Pavia PE and MP libraries.

Digital normalisation, stringent re-assembly, and scaffolding with mate-pairs

A first attempt at assembling the *D. immitis* genome was made in 2010, before a formal workflow existed for checking contaminants, coverage, and insert lengths. Table 2.9 shows the difference in assemblies as we moved from a naive approach to the workflow steps described in the methods section of this chapter.

The first two columns of Table 2.9 show two assemblies without digital normalisation. Velvet1.0.13MP shows a naive assembly done using Velvet (version 1.0.13) with the MP library provided as a long-insert library without any filtering. Although this assembly had the best sequence contiguity metrics with a very high scaffold N50 and longest scaffold size, it had the lowest biological accuracy metrics. This assembly also had 18 Mbp of Ns compared to only 0.1 Mbp in the ABySS1.2.3 assembly. Short-insert fragments dominated the MP library and caused mis-assemblies because Velvet expected that two contigs linked by read pairs from the MP library were approximately 4 kbp apart, whereas, in reality, the contigs were only 200–500 bp apart and oriented in reverse. These glaring mis-assemblies were avoided in the second assembly (Column 2 – ABySS1.2.3) by treating the MP library treated as an SE library (i.e., no pairing information was used from this library). The ABySS1.2.3 assembly was released as a data freeze version 1.3 and used in the *D. immitis* genome publication [51].

Column 3 (khmerABySS1.3.3) describes an assembly performed with a newer version of ABySS (version 1.3.3) using digitally normalised data. A single-pass khmer filter was used to remove all reads with a coverage greater than 20X. A low-pass filter for removing reads below a certain coverage was not used because tests showed that it removed too much data, resulting in more fragmented assemblies in this read set. This assembly had better sequence contiguity than ABySS1.2.3, and better biological accuracy metrics than both the previous assemblies. It was selected as an assembly freeze with the label nmwDi.2.0 ("nmw" indicated that this assembly consisted of the nuclear, mitochondrial, and *Wolbachia* genomes).

The last column (khmerABySS1.3.3MP) was the result of aggressively filtering the MP library for short-insert pairs, and using the rest to scaffold the nmwDi.2.0 assembly with SSPACE [115], as described in the Methods section. By removing all reads from the MP library that mapped to the same contig, or reads that mapped within 600 bp of the ends of the contigs in nmwDi.2.0, only long-fragment true MP reads were likely to remain. Using SSPACE, both the sequence contiguity and the biological accuracy of this scaffolded assembly were improved compared to khmerABySS1.3.3. Although SSPACE inserted approximately 3 Mbp of Ns into this assembly, these Ns were unlikely to be erroneously inserted (as in Velvet1.0.13MP) because all short-insert read pairs had already been removed. This assembly set was labelled nmwDi.2.1 and was used for identifying and re-assembling the *Wolbachia* and mitochondrial genomes. All assembly and annotation data for *D. immitis* are available at the genome data page <http://dirofilaria.nematod.es>.

Table 2.9 Comparison of assemblies for *D. immitis*

	Velvet1.0.13MP	ABYSSI.2.3	khmerABYSSI.3.3	khmerABYSSI.3.3MP
Assembly freeze label	–	Version 1.3	nmwDi.2.0	nmwDi.2.1
Longest Scaffold (bp)	1,196,450	167,771	254,425	1,085,577
Number of scaffolds	3,542	31,291	21,153	16,081
Assembly span (bp)	93,686,223	84,240,606	86,183,400	89,251,192
Mean scaffold length (bp)	26,450	2,692	4,074	5,550
Scaffold N50 (bp)	133,465	10,584	17,684	71,590
Span of Ns (bp)	18,596,734	147,335	321,550	3,428,067
GC content (%)	28.90	28.30	28.10	28.10
CEGMA completeness (%)	89.11	94.76	95.97	96.77
<i>B. malayi</i> protein contiguity (%)	60.45	64.81	68.98	72.82
<i>B. malayi</i> protein completeness (%)	72.83	74.65	77.80	80.09
<i>D. immitis</i> EST contiguity (%)	80.12	83.00	87.83	92.79
<i>D. immitis</i> EST completeness (%)	87.77	89.28	92.43	95.20

Note: Figures in bold indicate the best metric in that row.

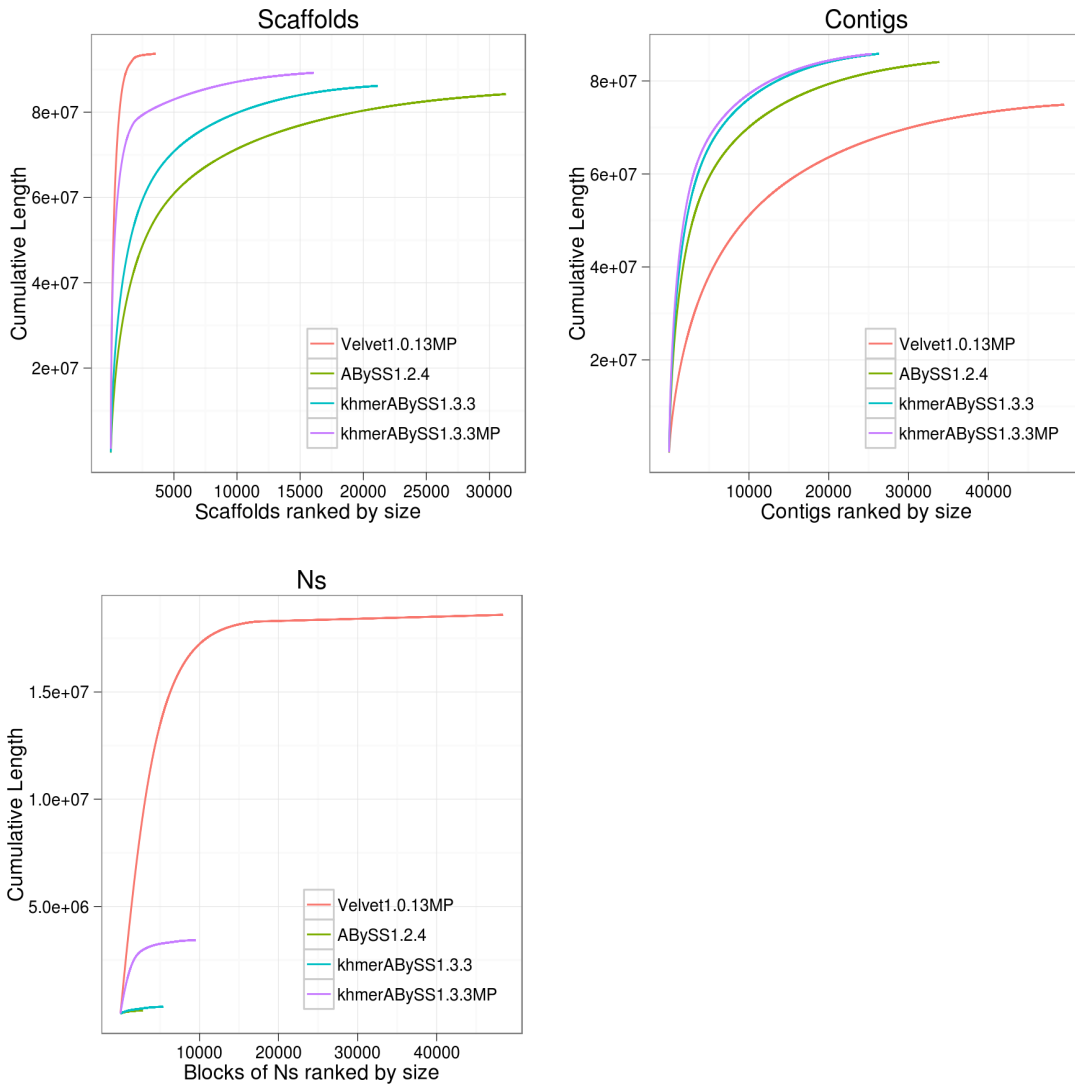


Figure 2.10 Comparison of cumulative lengths of scaffolds, contigs, and blocks of Ns for different assemblies of *D. immitis*

Assembling the *Wolbachia* genome of *D. immitis*

The genome of the *Wolbachia* endosymbiont of *D. immitis* (*wDi*) was present at approximately 10X the coverage of the *D. immitis* nuclear genome in the Athens PE library (Figure 2.9 c). MegaBLAST matches to known *Wolbachia* sequences from other species and a sequence coverage cutoff of 500X were used to identify 16 putative *wDi* scaffolds spanning 1,195,342 bp. All Athens PE reads mapping to these contigs were extracted along with their pairs. High coverage data sets such as this one (~1500X–2000X, estimated from Figure 2.9) can give suboptimal assemblies because excessive sequence errors can complicate the DBG [81, 124]. The read set was reduced to ~300X coverage by randomly sampling 20% of the read pairs, and reads were re-assembled stringently using Velvet and ABySS to test many parameter sets; a subset of assemblies is presented in Table 2.10. Khmer was also run with a high-pass filter of 20X coverage, but both Velvet and ABySS generated less contiguous assemblies with khmer-ed reads compared to the random subset. The Velvet assembly with a k-mer of 39 and a coverage cutoff of 38 was found to have the best sequence contiguity and biological accuracy when compared to the protein set of *Wolbachia* of *B. malayi* (*wBm*). The re-assembly was further scaffolded using aggressively filtered MP read pairs using the same protocol followed in the previous section, resulting in a new assembly with just two scaffolds, 919,954 bp and 1,058 bp in length, respectively. This *wDi* assembly was labelled *wDi.2.2* and is available on the genome data page at <http://dirofilaria.nematod.es>.

Post-assembly removal of mitochondrial and *Wolbachia* genomes

The final *wDi* genome assembly (*wDi.2.2*) was easy to identify for removal from the *nmwDi.2.1* combined assembly using MegaBLAST. However, identifying and removing the mitochondrial genome was not as straightforward because it did not assemble into a single contig as expected due to variations between the Athens and Pavia strains. On separately assembling both strains, the mitochondrial genomes were identified as single contigs in each assembly, and these contigs were used to identify and remove the mitochondrial genome from the *nmwDi.2.1* combined assembly. The final nuclear-genome assembly set for *D. immitis* was labelled *nDi.2.2* and is available online at <http://dirofilaria.nematod.es>.

Summary

The genome of *D. immitis* was published [51] using the same read sets as the ones described in this section. However, that assembly (version 1.3) was generated using a simpler process because the workflow described in this chapter had not yet been developed. The new workflow, using digital normalisation [109] and aggressive MP filtering, considerably improved the assembly contiguity, CEGMA completeness, and protein and EST alignment metrics (Table 2.9). The resulting assembly version (*nDi.2.2*) was used for subsequent analyses in this thesis in Chapters 3 and 4.

Table 2.10 Comparison of assemblies for *Wolbachia* of *D. immitis*

	ABySS k=43 n=3	ABySS k=49 n=3	Velvet k=31 cc=33	Velvet k=39 cc=38	Velvet k=39 cc=47	Velvet k=45 cc=33
Longest Scaffold (bp)	226900	174580	270292	225898	225977	174654
Number of scaffolds	73	74	20	10	15	10
Assembly span (bp)	970058	1036659	919558	918411	918473	918486
Mean scaffold length (bp)	13288	14008	45977	91841	61231	91848
Scaffold N50 (bp)	127800	126139	130452	174763	82955	126400
Span of Ns (bp)	1514	1537	450	817	1028	970
GC content (%)	0.328	0.328	0.327	0.327	0.327	0.327

Note: Figures in bold indicate the best metric in that row. k is the k-mer size; n specifies minimum number of pairs needed to join contigs; cc indicates coverage cutoff

2.3.4 *Litomosoides sigmodontis*

L. sigmodontis is a filarial nematode that parasitises rodents. Although it grows in cotton rats in the wild, it has been successfully grown in laboratory mice and now serves as a model for filarial infections. Research on *L. sigmodontis* is key to developing vaccines against filariasis, for testing drugs before clinical trials, and for understanding the basic biology of host-parasite interactions between nematodes and their filarial hosts. The genome sequence of *L. sigmodontis* is needed to provide a complete catalogue of the genes (and thus proteins) of the parasite, which can be screened using computational techniques to direct drug and vaccine research. The genome of the *Wolbachia* of *L. sigmodontis* (*wLs*) is also of interest as these alphaproteobacteria are essential endosymbionts of the nematode.

Sample, Sequencing, and Read QC

L. sigmodontis DNA was extracted from nematodes grown in gerbils by Simon Babayan (University of Edinburgh). Two short-insert paired-end libraries of 300 and 600 bp insert sizes were prepared by the GenePool Genomics Facility and sequenced on an Illumina HiSeq2000 instrument. The 600 bp library was run twice because the number of reads generated the first time was below expectation for a HiSeq2000 run. Sequence fastq files were submitted to the Short Read Archive with accession number ERP001496.

Table 2.11 shows the number of reads before and after cleaning. Additionally, the FastQC checklist (Appendix A) revealed a quality failure on cycle 66 in lib600-1 and on cycle 16 in lib600-2 on the forward reads. Approximately 35% of reads had Ns at these positions, and the standard quality and adapter trimming workflow would have thrown away all reads with an N in them, resulting in a massive loss of information. To avoid losing this information, we did not remove reads with Ns, but instead used the error-correcting tool, SOAPec, which corrected basecalls using high frequency k-mers.

Reads were digitally normalised using khmer [109] because of high sequence coverage, and a preliminary test showed that the normalised data had better sequence contiguity and biological accuracy metrics than un-normalised data. In Table 2.11, the last column demonstrates the dramatic reduction in sequencing depth after using khmer with a k-mer size of 25, a 20X high-pass k-mer coverage filter, and a 5X low-pass k-mer coverage filter.

Table 2.11 Read data for *L. sigmodontis*

Library	Num of reads and bases	Type of seq	Trimming and error correction steps	Post trimming Reads and bp	After khmer digital normalisation
Edinburgh, PE 300 bp	144.3 M pairs 29.2 gigabases	Illumina HiSeq 101 b PE	Adapter removal from 3' end; Quality < 20 b from 3' end; Error correction	132.4 M pairs, 26.2 gigabases	28.4 M pairs, 5.6 gigabases
Edinburgh, PE 600 bp	102.3 M pairs 21.3 gigabases	Illumina HiSeq 101 b PE	Adapter removal from 3' end; Quality < 20 b from 3' end; Error correction	73.6 M pairs, 14.3 gigabases	26.3 M pairs, 5.2 gigabases

Preliminary assembly and taxon-annotated GC-coverage plots

The TAGC plots in Figure 2.11 show low-coverage lab contamination, which was expected to be removed during stringent re-assembly with appropriate coverage cutoffs. The main nematode cluster (order Spirurida, as expected) showed the presence of the *Wolbachia* genome (order Rickettsiales). The *Wolbachia* genome is not present in multiple copies per nuclear genome and has approximately the same coverage as the nematode genome (500X – 1000X in Figure 2.11 C). Therefore, read coverage differences could not be used to separate the two organisms. First, the whole read-set was assembled. Next, *Wolbachia* sequences were extracted from this assembly using matches to known *Wolbachia* sequences, and then were re-assembled stringently.

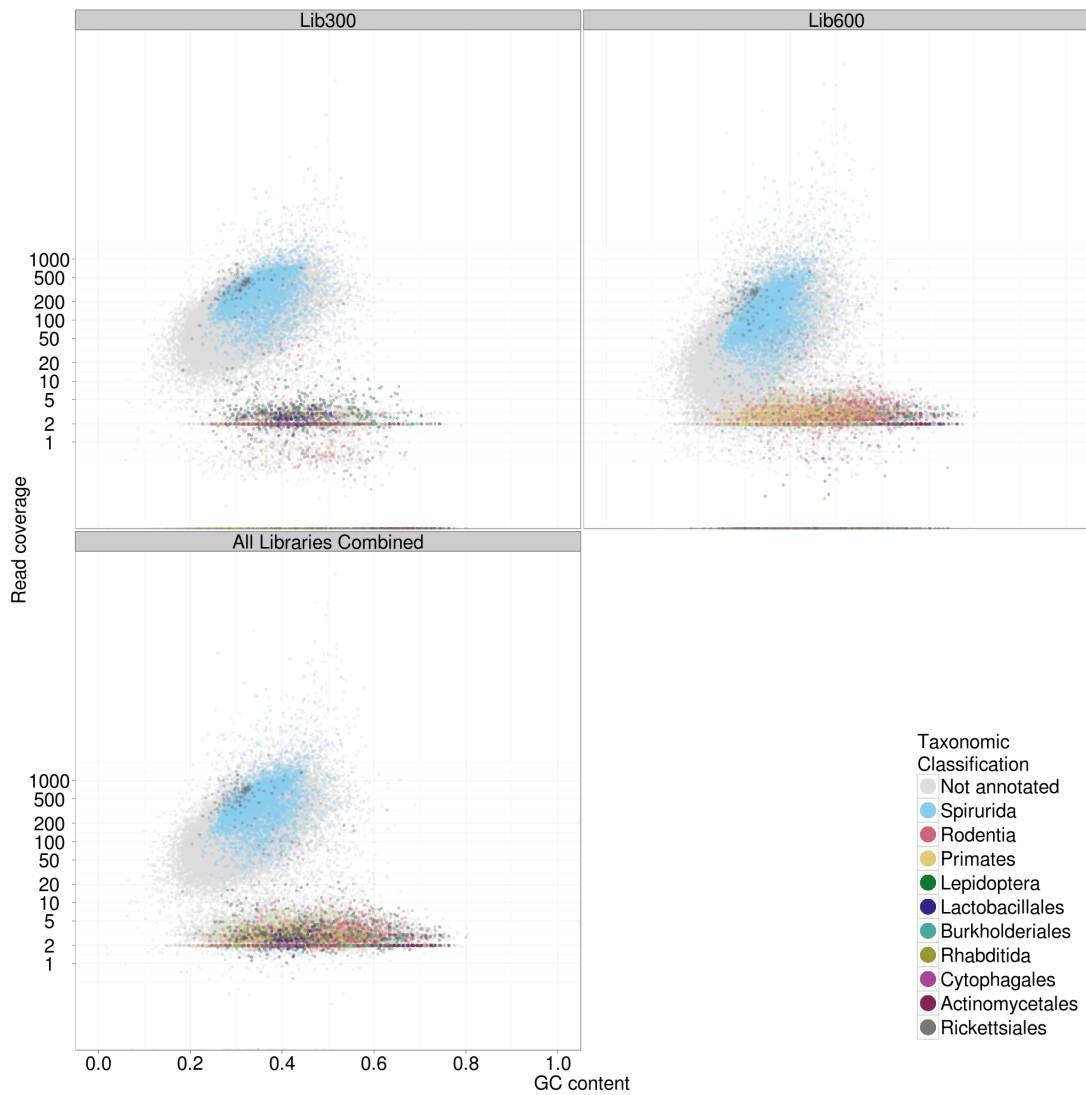


Figure 2.11 Taxon-annotated GC-coverage plot for *L. sigmodontis*

A preliminary assembly of 71,958 contigs is shown. A Blastn search of all contigs against the NCBI nt database resulted in 17,016 annotated contigs. The low-coverage contaminant cluster consists mostly of Rodentia (from the *L. sigmodontis* gerbil host), Primates and Lepidoptera (contaminants, possibly from other sequencing projects), and some bacterial sequences. The Rickettsiales cluster around ~500X coverage represents the Wolbachia endosymbiont of *L. sigmodontis*.

Stringent Re-assembly

Several re-assemblies were generated using both un-normalised and normalised data sets, with different assemblers and varying parameters. To test the biological accuracy of these assemblies, protein alignments to a related proteome were performed using the NCBI RefSeq set of 11,472 *B. malayi* proteins downloaded from GenBank, and EST alignments were performed using an *L. sigmodontis* transcriptome assembly generated previously using 454 sequencing [93].

The assemblies in Table 2.12 and Figure 2.12 represent a subset of all generated assembly sets. CLC chose its own k-mer value and did not have any coverage cutoff. As a result, CLC did not discard any of the short, low-coverage contaminants. These short scaffolds lowered the mean scaffold size and can be seen at the right end of the CLC cumulative scaffold length curve, where over 5000 scaffolds contributed less than 2 Mbp to the total assembly span. The next two columns, ABySSn10 and ABySSn5, show the effect of varying the "n" parameter in ABySS, which controls the minimum number of pairs needed to connect unitigs into contigs and scaffolds. Lowering this number from 10 to 5 increases the contiguity of the assembly set with a ~5 kbp increase in the scaffold N50, and a small increase in biological accuracy metrics. The final two columns were generated using khmer-normalised read sets assembled with ABySS (khmerABySS). The normal khmer protocol can discard a single read from a pair if it does not contribute new k-mers, thus losing read pairing information. The khmer_re_pair.pl script (Appendix A) restored the missing read for such pairs and the pairs were re-assembled using ABySS (khmerABySSrp). This last assembly had the best contiguity and biological accuracy metrics overall and was chosen as an assembly freeze. The assembly was labelled nmwLs.2.0: "nmw" was used to indicate that it consisted of nuclear, mitochondrial, and *Wolbachia* contigs; "2" referred to the fact that this was the second *L. sigmodontis* read-set; and "0" referred to the assembly iteration using that read-set. All genome and annotation data sets are available on the genome data page at <http://litomosoides.nematod.es>.

One of the advantages of khmer digital normalisation was that it reduced the initial data set substantially and therefore allowed many more assembly parameters to be explored in a short period of time. Many k-mer values from 25 to 70 were tested, and the final values chosen were 51, 47, 41 and 41 for the four assemblies ABySSn10, ABySSn5, khmerABySS, and khmerABySSrp, respectively. Khmer-ed read sets tended to have lower optimal k-mers because they had lower coverage overall.

Table 2.12 Comparison of stringent re-assemblies for *L. sigmodontis*

	CLC	ABySSn10	ABySSn5	khmerABySS	khmerABySSrp
Longest Scaffold (bp)	289,472	320,777	408,395	408,309	402,953
Number of scaffolds	9,083	4,075	3,497	3,345	3,178
Assembly span (bp)	65,885,602	65,799,421	65,505,800	66,503,273	65,887,609
Mean scaffold length (bp)	7,254	16,147	18,732	19,881	20,732
Scaffold N50 (bp)	36,556	37,699	42,366	43,269	47,550
GC content (%)	33.9	34.0	34.0	34.0	34.0
CEGMA completeness (%)	94.35	94.35	94.76	94.76	94.35
<i>B. malayi</i> protein contiguity (%)	67.43	67.67	68.51	68.42	68.82
<i>B. malayi</i> protein completeness (%)	76.21	76.44	76.99	77.10	77.32
<i>L. sigmodontis</i> EST contiguity (%)	90.67	90.86	91.39	91.91	92.21
<i>L. sigmodontis</i> EST completeness (%)	94.38	94.15	94.45	95.08	95.18

Note: Figures in bold indicate the best metric in that row.

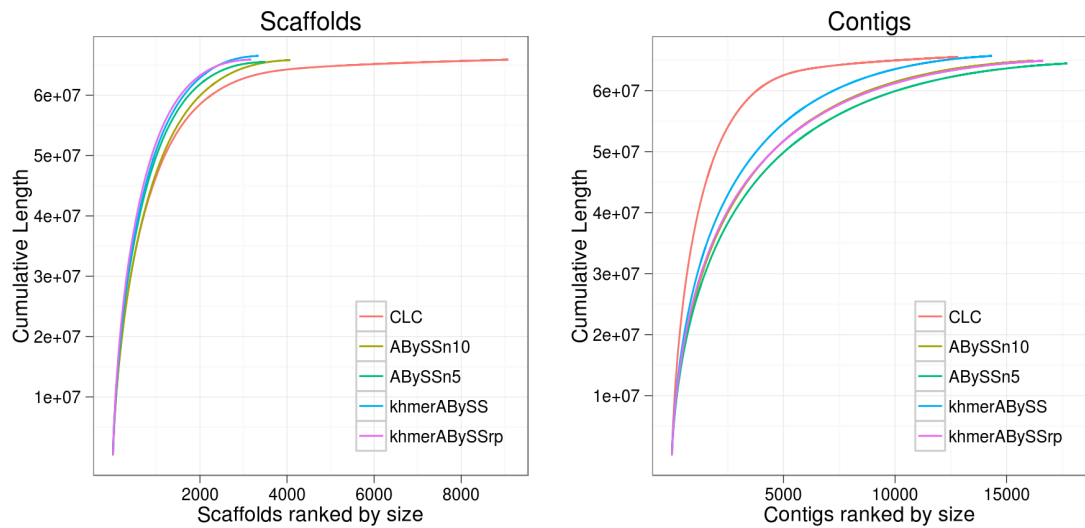


Figure 2.12 Comparison of cumulative scaffold and contig lengths for *L. sigmodontis*
 Almost all the assemblies had comparable scaffold length, although CLC had slightly shorter scaffolds, and slightly longer contigs than the rest. CLC did not have a coverage cutoff option. The shallow part of the CLC cumulative curve in the first plot represents short low-coverage scaffolds (from the contaminant cluster in Figure 2.11) that ABySS removed automatically.

Extracting the *Wolbachia* genome from the *L. sigmodontis* assembly

A set of all known *Wolbachia* genomic sequences from GenBank was selected as a query set, spanning 13.3 Mbp, including 4 finished genomes. Using this query set, 18 putative *Wolbachia* contigs spanning 1,077,004 bp were identified from the stringent assembly in the previous section (labelled nmwLs.2.0) using MegaBLAST [106] with default parameters. All reads that mapped to these 18 contigs were extracted along with their pairs and re-assembled stringently. Unlike the *Wolbachia* in *D. immitis*, neither khmer nor random sampling were used to reduce the number of reads because the expected coverage was not as hi

Both Velvet and ABySS were used to test many different k-mer values, coverage cutoffs, and minimum number of pairs needed to bridge contigs. The best assemblies from each program are shown in Table 2.13. Both are nearly identical in their span and comparison to the *Wolbachia* of *B. malayi* (*wBm*), but the ABySS assembly with a k-mer of 83, default coverage cutoff, and a minimum 3 read pairs joining contigs (WolABySSk83n3) was finally chosen because it assembled in fewer scaffolds than the Velvet assembly. This assembly freeze was labelled wLs.2.0 and is available on the genome data site and at the server for Rapid Annotation using Subsystems Technology (RAST) [125] as Job ID 54213.

Post-assembly

Assembly nmwLs.2.0 consisted of the nuclear, mitochondrial, and *Wolbachia* genomes of *L. sigmodontis* combined. The final steps in obtaining a nuclear genome assembly of *L. sigmodontis* were to separate the *Wolbachia* and mitochondrial contigs from this combined assembly. *Wolbachia* contigs in the nmwLs.2.0 assembly were identified using MegaBLAST against the *Wolbachia* genome created above (wLs.2.0). It is quite common for mitochondrial genomes to assemble into single contigs even with short-read NGS because of their low repeat content and higher coverage, and the mitochondrion of *L. sigmodontis* (mLs) was no exception. In this case, mLs was present as two consecutive copies in a single contig from the nmwLs.2.0 assembly, an assembly artefact due to the circular nature of mLs. One of the mitochondrial copies was labelled mLs.2.0. The final nuclear genome assembly with the wLs.2.0 and mLs.2.0 genomes removed was labelled nLs.2.1 and is available at <http://litomosoides.nematod.es> along with the mitochondrial and *Wolbachia* genomes.

Summary

The *L. sigmodontis* assembly benefitted from almost every stage of the workflow described in this chapter. An initial FastQC assessment of the raw reads revealed a sequencing failure on some cycles, which were corrected using k-mer-frequency-based error-correction tools. The TAGC plot revealed that the genome of the *Wolbachia* endosymbiont of *L. sigmodontis* was present at approximately the same number of copies as the nuclear genome, and therefore coverage differences could not be used to separate the two genomes. The final nuclear

genome assembly had high CEGMA completeness and had a high scaffold N50 (47.5 kbp) despite using only PE reads and no long-insert MP reads.

Table 2.13 Comparison of Velvet and ABySS stringent re-assemblies for *Wolbachia* of *L. sigmodontis*

	WolVelvetK85	WolABySSk83n3
Longest Scaffold (bp)	619,435	605,213
Number of scaffolds	17	10
Assembly span (bp)	1,040,727	1,048,936
Mean scaffold length (bp)	61,219	104,893
Scaffold N50 (bp)	619,435	605,213
GC content (%)	32.1	32.1
wBm protein contiguity (%)	90.96	90.98
wBm protein completeness (%)	91.78	91.76

Note: Figures in bold indicate the best metric in that row.

2.4 Discussion

2.4.1 NGS genomes are comparable to Sanger-sequenced genomes

The results of this chapter demonstrate that, with the right data cleaning and assembly strategy, it is possible to obtain high quality draft nematode genomes that are as good as some of the nematode genomes sequenced in the last five years using Sanger sequencing, for less than a thousandth of the cost (Figure 2.13). Spending more money and time on long-insert libraries will obviously result in even better assemblies, but inexpensive short-read sequencing can provide valuable and complete genomic resources for a large number of species in a short period of time, shifting our gene-centric perspective to a genome-centric one.

Figure 2.13 shows an overview of genome assembly statistics for all published nematode genomes, several complete nematode genomes (from WormBase), and the four genomes assembled reported in this chapter (shaded grey). With the exception of *M. floridensis*, the genomes reported in this chapter have higher CEGMA completeness than many Sanger-sequenced genomes. Among clade III Onchocercidae, the scaffold N50s of the newly sequenced *D. immitis* and *L. sigmodontis* genomes are higher than that of the Sanger-sequenced *B. malayi* genome, although the contig N50s are lower. *Caenorhabditis sp. 5* has a higher contig N50 than the *C. japonica* genome.

High scaffold N50 values in the other previously sequenced species are due to MP and fosmid libraries [126]. With additional high-quality MP sequencing, all the genomes in this chapter can become more contiguous. However, for most projects, the goal is to obtain an assembly that captures multi-gene sized segments, and this table shows that that goal can be met even with low-cost Illumina PE sequencing. The scaffold N90 column shows that 90% of the assembly for most NGS genomes is in scaffolds of size 4 kbp or longer, except for *D. immitis*, *C. angaria*, and *M. floridensis*. Using the 20 proteomes collected in Chapter 3, the median gene size was calculated to range between 1.03 kbp (*Caenorhabditis sp. 11*) and 3.85 kbp (*A. suum*). Therefore almost all genes should be present at full length in these assemblies.

One anomaly in Figure 2.13 is that *Caenorhabditis sp. 11* has a scaffold N50 (21.9 Mbp) that is longer than *C. elegans* (17.5 Mbp). Considering that the total *Caenorhabditis sp. 11* assembly span is only 79.3 Mbp, and the assembly is made up of 665 scaffolds, it seemed very unlikely that this 20.9 Mbp scaffold N50 size was accurate. On checking the assembly file using the scaffold_stats.pl script, it was found to contain large blocks of Ns (6 kbp at a time) inserted by the Newbler assembly algorithm [127], and two long scaffolds of length 33.3 Mbp and 20.9 Mbp. The mis-assembly was reported and the *Caenorhabditis sp. 11* sequencing consortium will redo the assembly using better tools.

Species	Technology	Status	Version	GC	Num scaffolds	Assembly span (bp)	Min scaffold (bp)	Scaffold N50 (bp)	Scaffold N90 (bp)	Contig ^A N50 (bp)	Span of Ns (bp)	CEGMA Comp. / Partial (%)
<i>T. spiralis</i>	Sanger	Published [46]	WS230	0.339	6,863	63,525,422	139	6,373,445	2,047	76,808	4,986,938	93.15 / 93.15
<i>A. suum</i>	Illumina PE/MIP	Published [34]	WS230	0.38	29,831	272,782,664	100	407,899	80,017	24,530	7,446,858	93.15 / 95.97
<i>D. immitis</i>	Illumina PE/MIP	Published [48]	nDI.2.2	0.28	16,061	88,309,529	200	71,281	1,636	15,147	3,430,609	96.77 / 97.18
<i>B. malayi</i>	Sanger	Published [43]	WS230	0.306	27,210	95,814,443	200	37,841	931	18,799	6,592,564	92.74 / 93.95
<i>L. sigmodontis</i>	Illumina PE	Ongoing	nLs.2.1	0.341	3,165	64,813,410	200	45,863	10,481	9,771	999,760	94.35 / 95.56
<i>A. viteae</i>	Illumina PE	Ongoing	nAv.1.0	0.299	6,796	77,350,906	500	25,808	5,125	18,572	175,459	93.55 / 95.16
<i>S. ratii</i>	Sanger; Roche 454; Illumina PE	Complete	WS230	0.249	2,184	52,638,471	1,001	359,029	13,671	359,029	11,106	93.55 / 93.55
<i>B. xylophilus</i>	Illumina PE; Roche 454 PE/MIP	Published [47]	WS230	0.404	5,527	74,561,461	444	949,830	4,294	18,150	1,475,733	94.76 / 97.18
<i>M. hapla</i>	Sanger	Published [33]	WS230	0.274	3,452	53,017,507	689	37,608	6,538	37,608	0	92.74 / 94.35
<i>M. incognita</i>	Sanger	Published [36]	INRA	0.314	2,995	86,061,872	63	62,516	11,923	12,900	3,966,853	72.58 / 75.81
<i>M. floridensis</i>	Illumina PE	Ongoing	nMf.1.1	0.296	58,696	96,673,063	200	3,698	622	1,234	1,778,970	60.08 / 72.18
<i>P. pacificus</i>	Sanger	Published [44]	WS230	0.428	18,083	172,494,865	47	1,244,534	85,679	18,213	19,302,620	90.32 / 93.15
<i>C. angaria</i>	Illumina PE	Published [45]	WS230	0.363	33,559	79,761,545	100	9,453	1,194	1,372	4,502,764	93.15 / 97.18
<i>C. japonica</i>	Sanger	Complete	WS230	0.392	18,817	166,256,191	186	94,149	1,850	10,166	12,198,257	85.48 / 93.55
<i>C. elegans</i>	Sanger	Published [4]	WS230	0.354	7	100,286,070	13,794 ^B	17,493,793	13,783,700	17,493,793	0	97.98 / 98.39
<i>C. brenneri</i>	Sanger	Complete	WS230	0.386	3,305	190,369,721	1,224	381,961	26,605	31,887	20,276,083	97.18 / 98.79
<i>C. sp. I</i>	Roche 454 PE/MIP	Ongoing	WS230	0.377	665	79,321,433	301	20,921,866 ^C	81,937	22,519	2,824,241	97.18 / 98.39
<i>C. remanei</i>	Sanger	Complete	WS230	0.38	3,670	145,442,736	2,000	43,512	11,656	30,878	7,036,533	95.56 / 97.58
<i>C. briggsae</i>	Sanger	Published [5]	WS230	0.374	12	108,419,665	25,117	17,485,439	14,578,851	45,098	3,003,126	96.37 / 97.18
<i>C. sp. 5</i>	Illumina PE	Ongoing	WS230	0.395	15,261	131,797,386	200	25,228	4,876	12,809	1,406,227	96.37 / 98.39

Figure 2.13 Assembly comparisons of 20 nematode genomes
 Roman numerals indicate Blaxter clades [32]. **A.** Contig N50 was calculated by splitting scaffolds at runs of 10 or more Ns. **B.** The *C. elegans* minimum size is smaller than *C. briggsae* because the former includes the mitochondrial genome. **C.** Large value indicative of gross mis-assembly.

2.4.2 Lessons learnt

The workflow presented in this chapter for assembling draft genomes from short-reads has evolved in the time that it took to assemble these four genomes, and will presumably continue to do so as sequencing and assembly technologies change. Some of the basic principles are likely to remain, however, and are summarised in this section. The quality checks, visualisations, and recommendations presented here are a summary of the best practices for creating low-cost genome assemblies, even for unculturable samples collected from the environment. Perhaps the most useful innovation is the idea of a preliminary assembly to assess sequence composition and coverage parameters. Using this chapter as a guide, it should also be fairly straightforward to separately assemble genomes of endosymbionts and other co-bionts from the same sample without requiring time-consuming and expensive laboratory cleaning and separation procedures. As the *D. immitis* re-assembly showed, it should also be possible to improve existing assemblies using the MP filtering protocol described here. Although all the recommendations in this discussion assume an "average" 100 Mbp sized nematode genome, they should be valid for any similarly sized metazoan genome.

Sequence longer-insert pairs where possible

Illumina PE sequencing is the most cost-effective and reliable technology currently available. Roche 454 sequencing can provide longer reads and therefore better assemblies, but the reads are more expensive and prone to indel errors. As a result of the indel errors, Roche 454 genome sequences are also more prone to frame-shift errors when gene-prediction algorithms are run on these sequences. Illumina-only assemblies for eukaryotic genomes have been shown to work well [89], provided many high-quality genomic libraries at a variety of insert sizes are used. The Illumina Genomic DNA Sample Prep Guide [128] suggests that the Illumina technology is robust for fragments up to 800 bp using short-insert PE sequencing. However, the *L. sigmodontis* read set indicated that the 600 bp library had some smaller inserts and was lower quality than the 300 bp library. Therefore, even though longer PE libraries can, in theory, bridge more repetitive sequences, a mix of shorter (300 bp) and longer (600–800 bp) libraries is advised.

Long-insert mate-pair (MP) libraries of insert sizes 2–8 kbp improve scaffolding dramatically at extra cost for library preparation, but without any extra sequencing cost, as the MP library can be multiplexed on the same sequencing lane. However, obtaining high-quality MP libraries continues to be a problem for nematode genomics, as they require large amounts of starting material, are hard to prepare, and can be heavily contaminated with short fragments. This chapter described a method for aggressively removing short-insert read pairs and using only reliable long-insert pairs to successfully scaffold the *D. immitis* genome. Without this filtering step, scaffolding algorithms were erroneously treating the short-insert

pairs as long-insert pairs, and mistakenly inserting large blocks of Ns to give much larger, but incorrect assemblies. If accurate estimates of the insert size (following our workflow) show that there is substantial short-insert contamination in a mate-pair library, then our method is recommended rather than the built-in scaffolding algorithms in programs like ALLPATHS-LG, ABySS, Velvet, or CLC.

Finally, although this chapter is about generating a genome sequence, we also recommend high coverage (e.g., at least one lane of Illumina HiSeq2000) sequencing of an RNA-Seq library made with as many tissues or stages as possible. This library can not only act as a valuable resource for genome annotation (as discussed in the next chapter), but can also be used by tools such as ERANGE [48] and SCUBAT (<https://github.com/e1swob/SCUBAT>) [129] to scaffold genomic contigs, although that was not attempted for the four projects in this chapter. Transcriptome assemblies from RNA-Seq data also acted as a biological accuracy metric and helped us evaluate the completeness and contiguity of assemblies in the case of the *Caenorhabditis sp. 5*, *D. immitis*, and *L. sigmodontis* genomes.

Discard or correct low-quality reads

For three of the four genome projects in this chapter—*C. sp. 5*, *M. floridensis*, and *L. sigmodontis*—older Illumina reads sequenced three years ago using the Illumina GA and GAIIx platforms were available. In all three cases, the old data consisted of shorter reads (maximum 50 b in length), lower quality, and shorter insert sizes. Assemblies with these data were highly fragmented, and therefore these data were discarded when we sequenced these genomes afresh. Sequencing costs have dropped to the point that it makes more sense to discard low quality data than to try and incorporate them into the assembly. Lower quality reads with sequencing errors not only introduce artefacts into the genome assembly, but also are impractical because assemblers require exponentially more time and memory to deal with sequence variations that are not real.

Error correction of raw reads is also becoming more commonplace. We used the SOAPec tool for error-correction on *D. immitis* and *L. sigmodontis*, the two most recent genomes in our set, and found that our assemblies improved. Without error-correction, a large proportion of otherwise high-quality *L. sigmodontis* 600 bp library data would have been discarded because of a sequencing error on one cycle. Shortly after these genomes were assembled, a study comparing several error correction programs was published [130] and its recommendations (Reptile [131], HiTEC [132], and ECHO [133]) will be used for future projects.

Digital normalisation is a new technique that has emerged in the last few months and has also improved assemblies by producing read sets with more uniform coverage. Because coverage information proved very useful for identifying contaminants and co-bionts in our preliminary assemblies, normalisation was used only for stringent re-assemblies. Although both *D. immitis* and *L. sigmodontis* improved with normalised reads, it is possible that

assembly algorithms that use long-range MP sequences and coverage information to resolve repeats would be misled by the normalised reads. More research is needed into the possible adverse effects of normalisation in such cases.

*90% of the drama in most bioinformatics
is error correction*

Ewan Birney (European Bioinformatics Institute,
<https://twitter.com/ewanbirney/status/202489451283349504>)

Overall, we see a trend where massive amounts of NGS data are being pared down using error correction and digital normalisation. These steps are computationally expensive and may take several days to run, but the end result is a smaller, higher quality data set that allows more assembly parameters to be explored.

Mate-pair data require extra care

The results section for *D. immitis* shows that when the PE-contaminated 4 kbp MP data provided by the sequencing centre was used as-is, without filtering, a genome assembly with very large, but very incorrect, scaffolds was produced. As a result of this experience, all libraries are now mapped back to a preliminary assembly to check the actual insert sizes and read orientations before proceeding.

We know of only one other effort to remove short-insert fragments from MP libraries [134], which has reportedly been incorporated as the De Novo Classifier (DNC) pipeline in the Celera assembler [69]. The DNC method could not be used because it was released after we had finished assembling the *D. immitis* genome, but it will be tested for future projects. One advantage of our method over DNC is that our filtering can be used as a standalone process as part of any assembly workflow or pipeline, whereas the DNC method can only be run as part of the Celera assembler.

Preliminary diagnostic assemblies are essential

Perhaps the most important workflow step in this chapter, and our strongest recommendation for any genome project, would be to perform a preliminary diagnostic assembly. The TAGC plot generated from a preliminary assembly not only helps identify contaminants or co-bionts in the sample, but also provides extra information on coverage that makes a better re-assembly possible. There have been rare instances in the literature of using GC-coverage scatterplots to visualise metagenomic read sets [135], but no one had previously used these plots as a diagnostic step to identify and separate organisms, followed by a more stringent re-assembly to get a higher quality genome. The preliminary assembly

also helped determine accurate insert-size distributions for our libraries, which are especially important in the case of MP libraries, as seen above.

In the past, genome sequencing was reserved for model organisms where clean samples from cultured organisms were available. As sequencing has become faster and cheaper, it is now easier to sequence an organism in its context (either inside a host, or along with its co-bionts) than to use extra resources to separate the organism in a sample. Some bacterial contamination was expected in the cases of *Caenorhabditis sp. 5* and *M. floridensis*, but we were surprised to see the diversity and abundance of bacterial species (Figure 2.7 and Figure 2.5). Many of the contaminants in these samples were present at high coverage and would be easy to re-assemble on their own, leading to new types of studies on the microbial environments of the nematodes, as we pointed out in [59].

Other genome sequencing centres routinely screen their eukaryotic genome projects for bacterial reads by mapping all sequenced output to a set of bacterial databases (Sahar Abubucker, The Genome Institute at St Louis, pers. comm.). Although the mapping approach will get rid of most of the bacterial contaminants, visualising the characteristics of a preliminary assembly is more likely to pick up novel and unexpected contaminants. Additionally, performing the contamination identification and filtering at the level of longer contigs rather than short 100 b reads, more sensitive searches against specific databases (as described in the Methods section) are possible. One example of the value of this approach was seen recently in the Blaxter Lab research group when a TAGC plot for the filarial nematode *Acanthocheilonema viteae* showed a large low-coverage cluster labelled "Primate", causing a colleague to exclaim "There's a monkey in my worm!" On checking with the sample provider, it was discovered that the worms had indeed been maintained on a culture of *Macaca mulatta* (Rhesus macaque) cells. If only a bacterial screen had been done, this contamination would have gone unnoticed until the genome annotation stage. Because of the preliminary diagnostic assembly, it was possible to do a sensitive search for all primate (macaque) contigs and remove the reads mapping to these contigs before proceeding with the rest of the project.

There's a monkey in my worm!

Georgios Koutsovoulos (University of Edinburgh, pers. comm.)

For the filarial nematodes *D. immitis* and *L. sigmodontis*, the goal was to separately assemble the genomes of the *Wolbachia* bacterial endosymbiont along with the nuclear genome of the nematodes. Preliminary assemblies and TAGC plots made it easy to use coverage as a filter for extracting *wDi* contigs which were present at more than ten times the concentration of

the nematode genome. In *L. sigmodontis*, although the *w*Ls genome was not present at a very different concentration from the nematode genome, all the *w*Ls contigs were above a certain coverage in the preliminary assembly, and that information was used to select putative *Wolbachia* contigs from the combined assembly. So far, these preliminary assemblies have only been used as diagnostic aids to help decide what contaminant databases to check and what coverage parameters to use. In the future, more sensitive metagenomic binning techniques [136] will be used to make it easier to computationally separate contigs belonging to different organisms.

Read separation improves re-assembly

In all four genome projects, partitioning the reads of the organisms of interest and re-assembling them separately improved the assemblies. For example, in the stringent re-assembly of the combined nuclear (*n*Ls) and *Wolbachia* (*w*Ls) genomes of *L. sigmodontis*, 13 putative *w*Ls contigs had an N50 of 169 kbp. When the reads mapping to these contigs were extracted and re-assembled, the resulting *w*Ls re-assembly improved, spanning only 10 contigs with an N50 and longest contig of 605 kbp. To date, only one other published assembly strategy describes a process where a first-pass or preliminary assembly was used to extract reads that were re-assembled a second time to get a more accurate and contiguous assembly [136]. However, the code described in that study has not yet been released, so it was not possible to test it.

Digital normalisation can improve assemblies

Digital normalisation using the khmer tool helped reduce the read-sets for the three projects where it was tried: *D. immitis*, *Wolbachia* of *D. immitis* (*w*Di), and *L. sigmodontis*. Normalisation also tends to remove reads with errors that contribute low-frequency unique k-mers, thus improving assembly quality. Smaller data sets enabled us to try multiple assemblers and parameters, and also improved the assembly quality in two of the three cases.

In the case of *w*Di, which had a sequencing depth of between 1500X–2000X, random sampling of 20% of the read set seemed to work better than using khmer to reduce the sequencing depth to 20X. For such extremely high sequencing depths it is possible that using khmer biases the selected read set and gets rid of useful read pairs that would have helped bridge small repeat regions. Khmer is order-dependent in that read pairs encountered earlier in the normalisation process are preferentially retained. If the original sequence file had poorer quality reads towards the start of the file as a result of a sequencing artefact, then khmer would discard later high-quality read pairs as they would not contribute enough new k-mers to the filtered read set. In this particular case, it is also harder to compare the two subsets because the randomly sampled set had ~300X coverage whereas the khmer set had ~20X coverage. A more systematic exploration of parameters is needed to quantify the optimal way to choose subsets of data for improving assemblies.

Different assemblers and parameters must be tested for each genome

Given the variety of genome characteristics, sequencing strategies, and sequencing error rates, no single assembly program or set of assembly parameters will work best across all cases. As the results in this chapter show, CLC and Velvet had opposing metrics on *Caenorhabditis sp. 5* and *M. floridensis* respectively, most likely because of the higher sequencing depth in the case of the former. Unfortunately, this means that any *de novo* genome assembly project would require testing several assemblers and parameter sets. Even for a single assembly program, different versions perform differently. For example, the latest versions of CLC (4.06beta.67189) and ABySS (1.3.3) used in the *L. sigmodontis* project have much better scaffolding algorithms than the previous versions that were tested on the same data (CLC 3.22.55705 and ABySS 1.2.7). The choice of k-mer and coverage cutoff in DBG assemblers also makes a big difference to the assembly quality. High-quality data with high coverage benefit from larger k-mers and coverage cutoffs. To add to the range of variables, a subset of reads can give better assemblies, as discovered during the assembly of the *Wolbachia* of *D. immitis* (*wDi*).

Thus, the number of different assemblies possible increases exponentially. Most genome publications do not report multiple assemblies even if they have internally tried different assemblers and parameters. Future projects on similar organisms or using similar inputs might benefit from the reporting of alternative assemblies, although it would be impossible to be exhaustive. Perhaps projects like the Assemblathon could include a repository for submitting some basic metrics for each assembly. This information could then be mined in the future to understand the process better.

Longer contigs and scaffolds are not enough

An objective set of metrics is needed to evaluate the different assemblies generated for each genome project. Typically, most genome publications for non-model organisms with no other genomic resources use the contig N50 (the contig size N at which 50% of the genome assembly is in contigs longer than N) or scaffold N50 as a measure of the contiguity of the assembly. Although sequence contiguity is important, it should not be the only metric used to compare assemblies. Along with contig and scaffold N50s, the size distribution and span of runs of "Ns" in the scaffolded genome were also measured. In *D. immitis*, these measures of Ns helped identify the problem of large-scale mis-assemblies caused by incorrect mate-pair library insert length assumptions. Cumulative sequence length plots for scaffold, contigs, and Ns also helped clarify the characteristics of the different assemblies, allowing for more informed judgements when picking the best assembly.

Assessing the biological accuracy of genome assemblies is more important than sequence contiguity, but is difficult to estimate when there is no reference sequence available. CEGMA completeness scores were used to see which assembly recaptured the most full-length core eukaryotic genes. This score can also be used as a crude absolute metric to assess overall

genome completeness as it is expected that >90% of the 248 core eukaryotic genes will be present at full length in any eukaryotic genome. Although three of the four projects described in this chapter were >94% CEGMA complete, *M. floridensis* performed poorly on this measure with only 60.08%. This issue needs further investigation because the *M. incognita* genome sequenced previously [38] used longer Sanger dideoxy reads, but also had a low CEGMA completeness of only 73.39%, whereas the Sanger-sequenced *M. hapla* genome [35] from the same genus had a more expected CEGMA score of >90%. As more nematode genomes are sequenced, a set of core nematode genes [49] would be very useful as a more sensitive metric than the core eukaryotic genes used in CEGMA.

Each assembly was also aligned to EST/cDNA sequences from the same genome and to protein sequences from closely related genomes, where available, to assess the biological accuracy of the assemblies. Protein and EST alignments are not absolute metrics, but both can be used to assess which assembly is relatively better than the others, using the steps described in this chapter and in [93]. Neither of the two major NGS assembly comparison studies [27, 89] addressed the issue of genome quality assessment in the absence of a known reference sequence, and this continues to remain an area of active research.

2.4.3 Upcoming technologies

The first generation of high-throughput sequencing was represented by capillary-based Sanger sequencing. Illumina's sequencing-by-synthesis, Roche 454's pyrosequencing, and ABI SOLiD's sequencing-by-ligation are commonly known as second-generation technologies. In the last year, three new "desktop" sequencing instruments have been released: Illumina MiSeq, Ion Torrent, and 454 GS Junior [137]. These are sometimes referred to as generation 2.5 because they use the same chemistries as second-generation sequencing, but improve on them by providing faster turnaround and/or longer reads. The trade off is lower throughput and higher cost-per-base, but these machines are easier to install and run in small labs, as they do not require many technicians to operate them (Table 2.14).

The term third-generation sequencing has been used to describe technologies where very long reads (>1 kbp) can be obtained from single DNA fragments at near real-time DNA synthesis speeds. PacBioRS [138] was the first of these third-generation sequencers to be released publicly. Longer reads from single-molecule technologies like PacBioRS have higher error rates (~15%) for the raw reads, but these can be corrected using high-fidelity Illumina short-reads to give >99.9% base-call accuracy. These high-quality corrected long-reads more than quintupled the median contig lengths obtained from second-generation assemblies [139]. The higher cost-per-base and lower throughput also mean that they are useful only in cases where a highly contiguous assembly is needed for studying chromosome rearrangements or long-scale structural variants.

The most promising single-molecule technology uses protein or solid-state nanopores that allow individual molecules of DNA to pass through them, and reads off single bases in real-time. The Oxford Nanopore system is an example of this technology, and the manufacturers claim that there are no hard limits to the read lengths [140]. If this technology works, most of the issues regarding sequencing and assembly strategies discussed in this chapter will become obsolete. Because no one outside of the company has seen real data from these machines yet, it is impossible to say if we can look forward to the day when obtaining the genome of a non-model nematode will be as simple as pressing a button.

2.4.4 Summary

The workflow described in this chapter was used to assemble four nematode genomes from Illumina short-read sequencing. This is currently the most cost- and time-effective way of generating "Improved High-Quality Draft" standard genomic resources for non-model organisms.

I relied heavily on the genome assembly community and used their software extensively. On my own, I would not have known how to approach the problem of assembling millions of short-reads if it had not been for the brilliant programmers who wrote fast and memory-efficient assemblers shortly after NGS technologies emerged.

Table 2.14 Upcoming sequencing technologies

Name	Read length	Error model	Cost per Megabase	Expected use for genome sequencing	Release
454 GS Junior	~500–600 b	Homopolymer errors; Lower quality towards end of read	\$31	Scaffolding and finishing <i>de novo</i> genomes	Shipped 2011
PacBio RS	Some reads as long as 15 kilobases. Most reads are ~1 kilobase	15% indels, errors distributed along read	\$2	Scaffolding and finishing <i>de novo</i> genomes	Shipped 2011
Ion Torrent Personal Genomics Machine (PGM)	100–120 b	Homopolymer errors; Lower quality towards end of read	\$0.63–\$23 (depending on chip)	Small scale sequencing. Testing mate-pair libraries.	Shipped 2011
MiSeq	150 b	Same as Illumina HiSeq (lower quality towards end of read). Typically 1% error per read	\$0.50	Small scale sequencing. Testing mate-pair libraries.	Shipped January 2012
Ion Proton	100–120 b, high throughput	Homopolymer errors; Lower quality towards end of read (assuming same as PGM)	–	Large genome projects	Early access as of April 2012
Oxford Nanopore	"No limit" ¹	4% (unverified ²)	–	Instant <i>de novo</i> genome sequencing; Large genome projects	Announced, expected end 2013

Note: Cost data taken from Loman [137] and Quail et al. [141]

¹<http://www.nanoporetech.com/news/press-releases/view/39>

²<http://www.rsc.org/chemistryworld/News/2012/February/oxford-nanopore-genome-sequencing.asp>

My contributions address the issues of data quality, visualising and removing contamination, and assessing assembly quality. The steps presented here were the result of repeated attempts to solve problems in genome assembly, and hopefully the tools I developed (Appendix A) will help others avoid some of the pitfalls I encountered. I strongly recommend the creation of preliminary diagnostic assemblies for all genome projects. To the best of my knowledge, no one else has suggested visualising genome data this way and using the information from taxon-annotated GC-coverage plots to improve subsequent re-assemblies.

Previous genome sequencing projects relied on large amounts of starting DNA material from cloned, cultured organisms. The data separation technique described here enables the sequencing of uncloned, unculturable samples taken directly from the environment. This is a significant advance, as it allows individual researchers to generate genomic resources quickly from field samples without months of lab work. It also allows the simultaneous sequencing of multiple organisms (parasites, endosymbionts, other co-bionts, etc.) along with the main nematode of interest, and this might lead to a better understanding of how nematodes interact with their environments.

More research is needed in the field of assembly quality assessment, and I hope to contribute to that in the future by developing better tools for quantifying genome accuracy and completeness. Assessing biological accuracy of genomic resources is very important, as all subsequent annotations and analyses rely on the genome sequence. After seeing how frighteningly easy it is to obtain large-scale misassemblies using an MP library contaminated with short-fragment read pairs (as in the case of *D. immitis*), I am somewhat wary of NGS genome projects that used MP libraries without checking the validity of the assemblies (e.g., *Caenorhabditis sp. 11* and perhaps others).

Sequencing technologies and assembly tools change rapidly, but many of the recommendations will remain valid even when low-cost and high-throughput third-generation sequencing becomes the norm. Combined with the annotation steps in the next chapter, this chapter demonstrates that it is possible for even individual graduate students to create highly useful genomic resources for metazoan genomes that can be used for broad-ranging evolutionary studies.

3 Annotating nematode genomes

3.1 Introduction

Assembling a draft genome from NGS is only the first step towards creating a usable genome resource. The next step is to identify regions of interest on the genome such as protein-coding genes and RNA genes. This chapter describes how a combination of bioinformatics pipelines was used to annotate the draft genomes generated in the previous chapter.

In the past, accurate gene-finding for a model organism was typically a massive undertaking involving dozens of people who searched for open reading frames (ORFs), aligned closely related proteins and ESTs, searched for valid intron-exon junctions, and then combined this information to create high-quality gene models. Gene models for model organisms such as humans and *C. elegans* are typically the result of many person-decades of effort that started even before the genomes for these organisms had been sequenced [142]. Our goal was much more modest as we were trying to create low-cost genomic resources for non-model organisms. We wanted to develop a best-practice workflow for automatically annotating our newly sequenced draft nematode genomes, which would incorporate as much relevant biological information as possible such as protein sequences from related species as well as ESTs and transcriptome assemblies from the same species.

My contribution was to establish a workflow that achieved a basic level of automated annotation by adapting the MAKER2 pipeline [142] as described in the Methods. The four genomes assembled in the previous chapter were annotated using this workflow and the results were utilised in the study on conserved non-coding elements described in Chapter 4.

Additionally, this chapter includes an overview of 20 sets of nematode gene predictions in one place. We generated annotations for our 4 genomes and compared them to 16 publicly available nematode genome annotations. Functional annotations were generated afresh for all 20 genomes, and this data set should prove to be a valuable resource for anyone interested in pan-Nematoda comparisons.

The first part of this chapter provides a brief overview of annotation concepts and terminologies, and some of the methods in the literature previously used for nematode gene annotation. The methods section describes the specific workflow used for gene prediction and gene annotation for the four nematode genomes in this thesis. The results section presents general statistics for each of the predicted gene sets and comparisons with other existing nematode gene prediction sets. Finally, this chapter finishes with a discussion of

some insights that large-scale comparative annotations can provide, as well as some of the limitations of automated annotation and the effect they can have on subsequent analysis.

3.1.1 Annotation concepts

Genome annotation encompasses several different processes, each of which can be achieved in different ways. The first step is usually protein-coding *gene prediction*, where coding sequences are identified on the genome. This process, also known as *structural annotation*, is better understood for prokaryotes, and most automated prokaryote annotation pipelines perform well [125, 143], although completely automated annotation may propagate some errors from previous reference genomes [144]. Nematodes are eukaryotes with introns and exons, and these features continue to be difficult to identify accurately, despite decades of research [145, 146]. The final product of the gene prediction step is a gene model that ideally identifies the locations of the five-prime untranslated region (UTR), each exon and intron, and the three-prime UTR, for each protein-coding gene.

Once the gene loci have been identified, the second step is *functional annotation* to identify what the genes do. In an automated pipeline, this step can be achieved in two ways: by sequence similarity to genes with known function in other species [147], and by identifying protein domains and signatures with known functions inside each gene (such as Zinc fingers, F-box domains, etc.). Common genes that take part in known biochemical pathways can also be identified at this stage and Enzyme Classification (EC) numbers assigned if the genes encode a known enzyme [148]. The typical end product of this step is a set of descriptors for each gene, including a canonical gene name (where applicable), a list of protein domains, and a list of Gene Ontology terms [149] or EC numbers associated with each gene.

The remaining steps of most genome annotation projects are *repeat-masking* and *RNA annotation*. Although listed last here because most researchers are interested in protein-coding genes and their functions, repeat-masking is usually performed first on the draft genome. The advantage of finding repeats first is that repeats can constitute as much as 80% of eukaryotic genomes [150], making it simpler to find the protein-coding genes once the repeat regions have been masked. Repeats can be simple repeats (such as homopolymers or microsatellites) or longer repetitive genome elements such as transposons and retrotransposons. Annotating a genome also involves identifying the RNA genes, which can be of many types: transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and other non-coding RNAs (ncRNAs) such as micro-RNAs (miRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs).

There are two main ways of performing structural annotation. Eukaryotic genes can be predicted using *ab initio* methods or using *evidence-based* methods. *Ab initio* methods predict genes using the underlying genome sequence and a set of Hidden Markov Models (HMMs)

that define the sequence characteristics of intron-exon boundaries, coding exon composition, splice site composition, and start and stop codons. The HMMs record a set of states and the transition probabilities between the states. These HMMs differ for each species because each species can have different state probabilities (e.g., a codon bias) and different transition probabilities (e.g., the probability of a given codon being followed by another specified codon). Therefore, the HMMs need to be tuned for each species using a high-quality set of a few hundred gene models. Unfortunately, non-model organisms have very few gene-level resources unlike model organisms where many genes may have already been elucidated. A better way is needed to automatically obtain a set of training genes for *ab initio* predictors, and the workflow described in this chapter addresses this issue.

Evidence-based methods use alignments to the genome from known coding sequences from the same species (in the form of EST or cDNA sequences) or known protein sequences (e.g., from closely related species). The locations where the EST or protein-coding sequences align are assumed to represent exons. As more genome sequences become available and more RNA-Seq evidence is generated, the quality of evidence-based gene predictions will improve.

Finally, the importance of assessing how well a gene model matches the evidence points to the need for a reliable metric. Currently, the "gold-standard" of genome annotation is human annotation, where an experienced annotator examines all the evidence for a particular model and determines the best gene structure. However, non-model organisms typically have very limited resources and cannot afford teams of experienced annotators. To be able to use and compare automated annotation tools, we have to know how well they perform. One of the metrics used for assessing the performance of an automated (or manual) annotation is the *Annotation Edit Distance* (AED), which assigns a number from 0 to 1 to measure how distant a gene model annotation is from the aligned evidence (an AED of 0 implies a perfect alignment, and an AED of 1 implies that none of the evidence sequences overlapped with that model). The AED [151] is the only metric for assessing the quality of an annotation when the true gene models are not known.

3.1.2 Review of existing tools

In this section, I briefly review the software typically used for genome annotation (especially gene prediction) and describe the reasons for choosing the tools used in the workflow described in the methods section. For a more comprehensive review of gene prediction concepts and programs, see [152].

Gene prediction

Because *C. elegans* was the first metazoan to be sequenced, many eukaryotic gene prediction algorithms have been tested on this species. The nGASP study [153] compared 20 gene-

prediction programs on a 10 Mbp test set of *C. elegans* and found that combiners—programs that incorporated results from multiple *ab initio* and evidence-based gene-prediction programs—performed best.

Nematode genome annotation projects in the last 2 years have used multiple *ab initio* gene predictors for each species:

1. *T. spiralis* [49]: SNAP [154] (trained with *B. malayi*), SNAP (trained with *C. elegans*), EAnnot [155], and FgenesH [156] (trained with *C. elegans*)
2. *B. xylophilus* [50]: Augustus [157], SNAP, and GeneMark.HMM [158], combined using EVidence Modeler [159]
3. *A. suum* [36]: Augustus, GlimmerHMM [160], SNAP, and TopHat+Cufflinks [161], combined using Glean [162] and custom scripts.
4. *D. immitis* [51]: Augustus, SNAP, combined using MAKER [142].

The general trend seems to be that no single software is accepted as the best protocol, and that all research groups merge the results from multiple tools. Both *ab initio* and evidence-based predictions are usually combined, although there does not seem to be any consensus on how best to combine predictions. Additionally, it is now possible to use Illumina sequencing to rapidly generate large amounts of RNA-Seq data, but most tools do not have an easy way to incorporate these data. TopHat and Cufflinks are two recent tools that can utilise RNA-Seq reads by first mapping the reads to the genome and then using coverage information to link exons together into gene models.

The other way to use high-throughput RNA-Seq short-reads is to first assemble them into transcripts and then use these sequences as evidence to find gene models using EST alignment programs such as PASA [163] and Exonerate [164]). Several transcriptome assembly programs have been developed over the past 3 years that work on RNA-Seq short-reads: Oases [165], Trans-ABYSS [166], SOAPdenovo-Trans [111], and Trinity [167]. We used SOAPdenovo-Trans for assembling the small RNA-Seq datasets generated for two of our projects because the program was fast, easy to use, and provided sensible results. Better assemblies could have been generated if we had systematically used other assemblers and assembly parameters, and developed new transcriptome assembly assessment metrics to test them, but these tasks were beyond the scope of our immediate goal of obtaining a first-pass automated annotation.

A complete eukaryotic annotation pipeline requires many pieces of software that can work together to carry out gene-prediction, repeat-masking, and evidence alignment. Two such pipelines are provided by the National Centre for Biotechnology Information (NCBI) and European Bioinformatics Institute (EBI) respectively. NCBI's Gnomon pipeline [168] is only used internally for their own genome annotation projects and is not available for external users. EBI's Ensembl Automatic Gene Annotation System [169] is used for generating and

upgrading annotations for the approximately 100 eukaryotic species currently present (as of February 2013) in the Ensembl genome browser. EBI generously makes its Ensembl gene annotation pipeline available for other groups to install and run on external genome projects as well. Another advantage of the Ensembl pipeline is that all the genome and annotation data would have been available in a standardised database format along with the application programming interfaces (APIs) for manipulating this information. However, installing and configuring the pipeline and its components is a time-consuming process that requires experienced bioinformaticians and system administrators, and so we were unable to use it on our genomes.

One possibility was to develop our own annotation pipeline that called existing programs in the right order using a workflow system. However, we did not have the resources to build such a system or to conduct a systematic evaluation of all available annotation tools, parameter combinations, and tool combinations (e.g., to determine if Augustus and Cufflinks together are better than Augustus alone, etc.). Therefore, we decided to proceed with a pipeline that, on paper, seemed like it should be the most effective. We chose the MAKER2 annotation pipeline [151], which was designed as a comprehensive Perl script that not only utilises multiple *ab initio* and evidence-based gene predictors inside it, but also evaluates the quality of each prediction and reports its AED. MAKER2 is also the only eukaryotic annotation tool supported by the Generic Model Organism Database (GMOD) consortium [170]. The GMOD project ensures that open-source software tools for creating and managing genome databases are inter-operable. Despite the "model organism" in its name, the project is very useful for small lab groups such as ours that are trying to develop genomic resources for non-model organisms. Other advantages of MAKER2 are that it is well-documented, well-supported (with a helpful online community), installs easily, and is easily parallelisable to take advantage of cluster computers. In addition, it allows existing annotations to be reused, offers flexible, well-formatted output in the form of GFF3 files [171], and comes with scripts and utilities to convert the output into files usable by many other programs.

Functional annotation

The goal of functional annotation is to assign a canonical gene name to each gene prediction where possible, or at least ascertain some idea of the gene's functionality by looking at the individual domains. Unlike gene prediction, there are not as many functional annotation tools that are widely used. The most commonly used tool for assigning GO terms and EC numbers to gene predictions is the Java-based Blast2GO program [172], also used in this thesis. To identify protein domains, most genome projects perform Pfam searches [173]. We chose InterProScan [174], which searches Pfam domains along with several other protein signature searches. InterProScan output is provided as is for our four genomes, but is also utilised as an input to Blast2GO above, as Blast2GO can use these results to assign GO terms more accurately.

Repeat-finding

Unlike functional annotation tools, repeat-finding tools continue to be under active development. Currently, the most popular repeat finding tool seems to be RepeatMasker [175], which is also built into the MAKER2 pipeline. RepeatMasker finds simple (low complexity) repeats and interspersed repeats by comparing the input genomic sequence against a regularly updated library of repeat elements. To identify new repeat elements in a newly assembled draft genome, other tools need to be used [150]. We used RepeatModeler [176] (developed by the RepeatMasker team) to identify new repeat elements.

RNA annotation

To find tRNAs, we used the industry-standard tRNAscan program [177]. For finding RNA genes and other non-coding RNA (ncRNA), we used the Rfam database [178] and searched it using rfam_scan.pl, the Rfam-supplied perl script. This script uses sequence similarity and covariance models to identify some RNA families, but it is not comprehensive. Therefore, anyone interested in a particular kind of RNA should use more specialised state of the art RNA finding algorithms for that kind of RNA, along with additional sequence evidence (e.g., small RNA-Sequence data if one is interested in miRNAs).

Genome delivery

Once a genome is annotated, its gene models and functional annotation can be delivered as GFF3 files, which store information about each feature location, and additional information such as gene names, gene ontology terms, and alternate transcripts, in a standard format. GFF3 files are flexible, easily parsed, and can be converted into EMBL, GenBank and other annotation file types. However, such files may be many millions of lines long and can only be explored and mined using other programs or scripts. To allow biologists to directly view their genes of interest and to search by gene name or gene function, the genomes need to be delivered in a more user-friendly format. Genome browsers enable the genome sequence and its annotations to be displayed graphically. With more genomes being sequenced and the advent of web 2.0 technologies that provide richer user experiences, genome browsers have seen more active development than any other type of tool in the genome annotation process. Although several different standalone and web-based genome browser tools are available [179], we chose GBrowse [180], which is installed on a server and accessed through a web browser.

3.2 Methods

3.2.1 Computational gene prediction using MAKER2

Figure 3.1 provides an overview of how we used MAKER2 (version 2.25) iteratively. Briefly, the first pass uses all the evidence available to obtain a set of gene models. These gene models are used to retrain the HMMs for the *ab initio* predictors (Augustus version 2.5.5 and SNAP version 2010-07-28). The second pass uses the newly trained *ab initio* gene finders and the existing evidence to obtain a final set of predictions. As the figure shows, the first pass uses EST sequences and Protein sequences as evidence and uses two *ab initio* gene finders—SNAP and GeneMark (GeneMark-ES version 2.3e). The initial HMM for SNAP (henceforth referred to as SNAP1) was trained using complete core eukaryotic gene models found by the CEGMA (version 2.1) [94] program, whereas GeneMark uses a self-training algorithm that does not need a separate training set. Because CEGMA only looks for 248 genes, this preliminary SNAP1 HMM was not very accurate. After MAKER2 was run with these inputs (commands and specific options in Appendix A: Predict genes using a two-pass (iterative) MAKER2 workflow), the resulting first-pass gene models were used to retrain SNAP and to train Augustus.

Table 3.1 is an example of the effect of different SNAP HMMs on one 384 kbp test scaffold from *Caenorhabditis sp. 5*. The transcript sequences predicted using different HMMs were aligned to the closely related *C. briggsae* transcript set using MegaBLAST. The number of full length alignments ($\geq 95\%$ of the *C. briggsae* transcript length) are shown. Rows 2 and 3 in Table 3.1 show the difference between a SNAP HMM trained on CEGMA genes alone (CEGMA SNAP) and one trained using all genes from a first-pass MAKER2 run (SNAP1). Although SNAP1 predicted only one extra gene, it aligned to 16 full length *C. briggsae* CDSs compared to 5 for CEGMA SNAP. The default *C. elegans* model shipped with SNAP (ELEGANS SNAP) aligned to more *C. briggsae* transcripts than CEGMA SNAP, but also over-predicted genes. The last row shows the result of running the full MAKER2 two-step workflow as described next. The full workflow recovers more *C. briggsae* alignments and has fewer over-predictions compared to the other methods.

Table 3.1 Differences in gene predictions using varying HMMs with SNAP

HMM used for SNAP	Number of predicted genes	Number of exons	Number of full length (>=95%) <i>C. briggsae</i> CDSs aligned
HMM for <i>C. elegans</i> shipped with SNAP (ELEGANS SNAP)	116	544	11
HMM trained from CEGMA run on <i>Caenorhabditis sp. 5</i> genome (CEGMA SNAP)	90	403	5
HMM trained from full length gene models of first-pass MAKER2 run (SNAP1)	91	533	16
Two-step MAKER2 workflow using SNAP HMM (SNAP2) trained from full length gene models of first-pass MAKER2 run, along with Augustus and GeneMark HMMs, and EST and protein evidence sets.	81	518	24

Note: All SNAP runs were performed on the longest *Caenorhabditis sp. 5* scaffold as a test set (384 kbp). The last row shows the effect of running the full two-step pipeline on the same scaffold.

The second round of MAKER2 was run with the same evidence sets as the first round and with the same GeneMark model, but with the newly trained Augustus model and the retrained SNAP model. Additional constraints on this final output included: minimum protein length of 30 amino acids (aa), extra steps taken to determine alternate splicing, a maximum intron size of 2000, and single exon gene models reported only if they were longer than 200 nucleotides. These cutoffs were based on *C. elegans* (WS230) where 0.2% of proteins are below 30 aa, 2.4% of introns are longer than 2000 bp, and 0.7% of all genes are single exon genes <200 bp. During the second round of MAKER2 predictions, *ab initio* predictions that did not have any protein or EST alignment evidence (i.e., gene models with an AED score of 1.00) were kept in the GFF files for future reference, although they were not reported in the official releases.

For each of the four species that we annotated, we used the best evidence sets available. Protein sequences from closely related organisms were available for all four species. EST sets in the form of Sanger-sequenced ESTs, Roche-454 transcriptome assemblies, or RNA-Seq assemblies were available for three species. Table 3.2 shows the EST and protein evidence used in the annotation process for each of the four species. No EST or cDNA data were available for *M. floridensis*, therefore we used CDS sequences from *M. incognita*, another species of the same genus, as MAKER2 allows the use of an alternate EST set where coding sequences from closely related species are aligned using the protein-space nucleotide alignment tool tblastx [113]. *L. sigmodontis* cDNA Roche 454 reads were assembled into transcripts as described in [93]. *D. immitis* and *Caenorhabditis sp. 5* Illumina RNA-Seq data were assembled using SOAPdenovo-Trans [111]. The config files are shown in Appendix A (Assemble RNA-Seq reads using SOAPdenovo-Trans).

MAKER2 produced gene models in GFF3 format [171] and provided CDS transcript and protein fasta files with one sequence each for each of the mRNAs predicted in the GFF3 file. For example, if there were 18,000 genes with 20,000 mRNA transcripts (i.e. 2,000 of the mRNAs predicted were alternate transcripts), then the protein and nucleotide fasta files would also have 20,000 sequences each. The GFF3 and transcript files were renamed using the `maker_map_ids`, `map_gff3_ids` and `map_fasta_ids` scripts provided with the program.

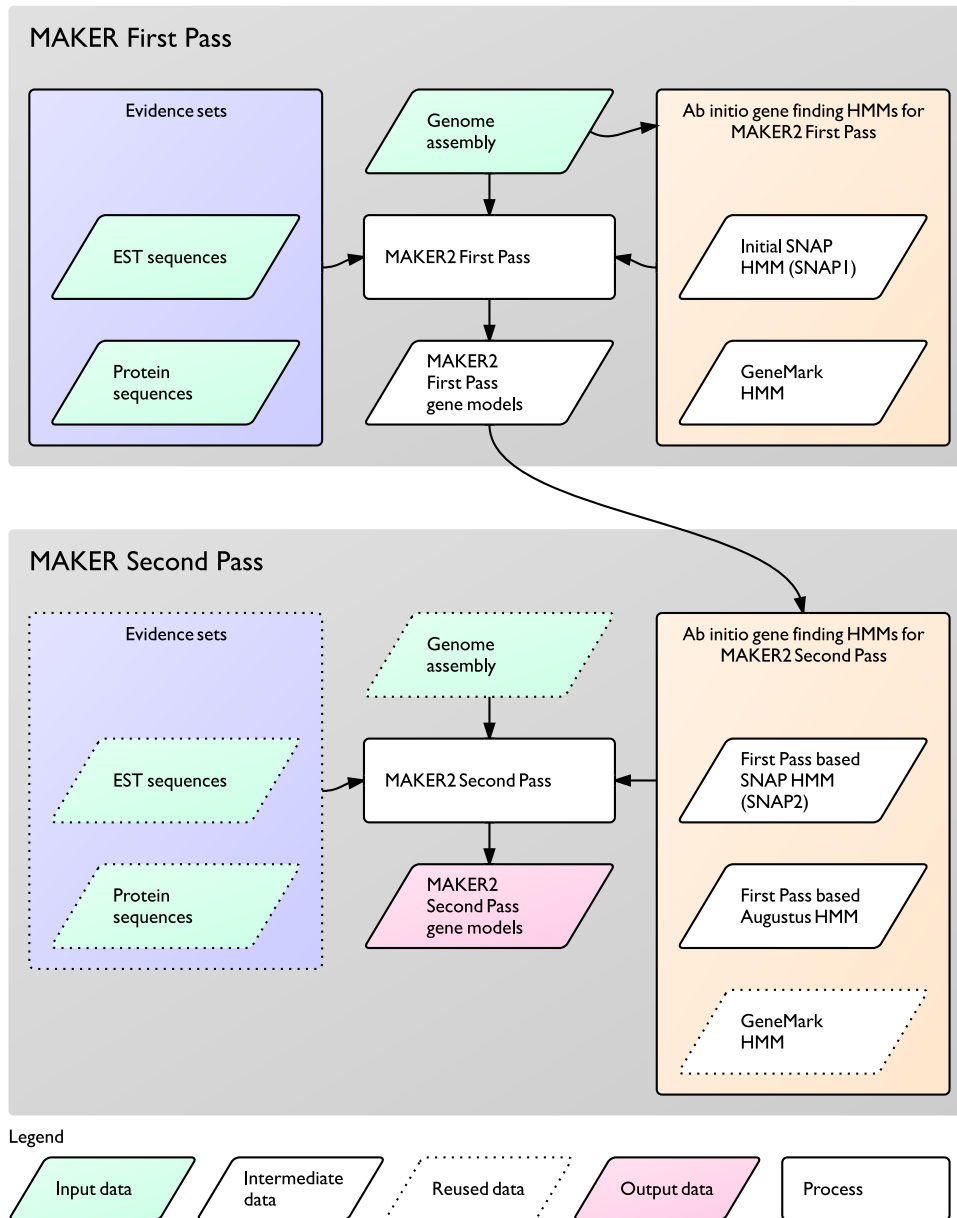


Figure 3.1 Two-pass iterative MAKER2 pipeline
 MAKER2 uses evidence (ESTs and proteins) along with *ab initio* gene finding HMMs to predict gene models. MAKER2 first pass uses SNAP1 HMMs obtained from running CEGMA on the genome assembly, along with EST and protein sequences to predict gene models. First-pass gene models are used to train new SNAP (SNAP2) and Augustus HMMs. MAKER2 second pass uses these new HMMs and the same evidence as the first pass to predict better gene models.

Table 3.2 Data used for annotation

	Genome version	EST evidence used	Protein set used
<i>Caenorhabditis sp. 5</i>	c_sp5.WS230	Mixed stage and mixed-sex library (41.1 M Illumina RNA-Seq reads, 1.6 gigabases) assembled into 30,756 transcripts spanning 12.7 Mbp	21,961 <i>C. briggsae</i> proteins (WS230); Swissprot
<i>M. floridensis</i>	nMf.1.1	21,232 <i>M. incognita</i> CDS transcripts from WormBase WS230	20,359 <i>M. incognita</i> and 13,072 <i>M. hapla</i> proteins (WS230); Swissprot
<i>D. immitis</i>	nDi.2.2	Female + male libraries (61.9 M Illumina RNA-Seq reads, 3.3 gigabases) assembled into 35,544 transcripts spanning 25.1 Mbp; 4005 GenBank ESTs	11,472 <i>B. malayi</i> Refseq proteins; Swissprot
<i>L. sigmodontis</i>	nLs.2.1	Female + male + microfilaria libraries (764 k 454 cDNA reads, 310.8 Megabases) assembled into 15,832 transcripts spanning 15.9 Mbp [93]; 2695 GenBank ESTs	11,472 <i>B. malayi</i> Refseq proteins; Swissprot

3.2.2 Calculating gene prediction metrics

Gene prediction metrics such as exon lengths, intron lengths, and exons per gene were easy to obtain for our four genomes because MAKER2 outputs well-formed and consistent GFF3 files. For the remaining 16 nematode genomes, we downloaded protein and annotation data from the WormBase ftp site (release WS230) except for *M. incognita* where gene models were obtained from Etienne Danchin of INRA, France; and *A. viteae* for which all data were obtained from <http://acanthocheilonema.nematod.es>. Most of the WormBase annotation files were hard to parse because the file formats were not consistent (some files were in GFF2 format instead of GFF3) and because multiple GFF source-feature entries matched the protein names. In several cases, the number of entries in the protein fasta file did not match the number of mRNA transcripts in the GFF files. As a result, different scripts had to be written to extract gene-model information in each file (Appendix A: Standardise nematode genomes and annotations).

To calculate CEGMA % completeness and average copy number, CEGMA [94] version 2.1 was used with default settings on each genome assembly. For gene numbers, GFF features labelled "gene" were counted, and for protein numbers the number of sequences in the protein fasta file for each species were counted. Rather than use the more common exons-per-gene metric, we were forced to use exons-per-transcript because the gene features in the GFF file often did not match the number of proteins. Means and medians for exon lengths were calculated by treating each exon as a separate entity, even if exons overlapped. However, the total exon span was calculated by merging overlapping exons so that no genome position was double counted. Exons were only used for these metrics if they belonged to protein-coding genes and not if they derived from RNA genes. Intron lengths were obtained from GFF files from entries with the feature "intron" if available, else intron lengths were calculated by subtracting all "exon" feature intervals from "mRNA" or "Coding_transcript" feature intervals for each mRNA.

3.2.3 Functional annotation of protein-coding genes

The command line version of Blast2GO (b2g4pipe version 2.5 [172]) was used to assign gene names and gene ontology (GO) terms to the genes predicted using MAKER2. First, MAKER2-formatted protein sequences were renamed using utility scripts provided with the software distribution. These sequences were then queried against the NCBI nr database using blastp (version 2.2.25 from the NCBI Blast+ suite) with expect value $1e-5$ and a maximum of 50 target sequences, and the results were stored in xml format. By default blastp uses an expect value of 100 and returns up to 500 target sequence matches, so these parameters were chosen to ensure that only relevant high-identity protein matches were found. The same protein sequences were also searched for protein domains using

InterProScan (version 4.8 [174]), which internally used the programs blastprodom, hmmpfam, hmmpanther, hmmtigr, hmmsmart, gene3d, seg, and coils. InterProScan output was also saved in xml format, and the two xml result sets were combined using Blast2GO. Blast2GO provided functional annotations in the form of a canonical gene name and GO terms for each sequence (where available) based on the best blastp hits to nr and the InterPro domains found. GO terms were also converted to a smaller generic GOslim set using the map2slim script from the GO-Perl library (version 0.13) [181]. These names and GO terms were inserted into the Notes and Alias field of each mRNA entry in the MAKER2-generated GFF3 files using custom scripts (Appendix A: Add functional annotations to a genome).

3.2.4 Repeat-masking

Draft genome assemblies were repeat masked from within the MAKER2 pipeline using RepeatMasker (v 3.0 [175]), searching all known repeat sequences (RepBase Library dated 2012-04-18 [182]). These repeats were reported in the GFF3 files output by MAKER2. Additionally, RepeatModeler [176] version 1.0.5 was run separately on each draft genome to obtain a file of species specific repeats.

3.3 Results

3.3.1 Gene models with the MAKER2 workflow

Using MAKER2, we identified gene models in each of our four draft genomes with evidence from EST alignments or protein sequence alignments from closely related species. The settings chosen were quite strict in that each prediction required at least some evidence in the form of alignments to EST or protein sequences. Figure 3.2 shows some of the characteristics of the gene predictions for these 4 species (shaded grey) compared to gene predictions for 16 other publicly available nematode genomes. The broad phylogeny and clade numbers in Figure 3.2 and other figures in this chapter are taken from [32], the detailed tree of the genus *Caenorhabditis* is from [117], and the previously unresolved topology of the Onchocercidae (in Clade III) is by Georgios Koutsovoulos (University of Edinburgh, pers. comm.).

Genome assembly sizes demonstrate a general trend of larger genomes in Clade V, with *P. pacificus* and six of the eight *Caenorhabditis* species being larger than 100 Mbp, whereas most of the remaining nematode species are smaller than 100 Mbp. The two exceptions among the *Caenorhabditis* species—*C. angaria* [48] and *Caenorhabditis* sp. 11—are both species that were sequenced using short-read NGS, which might account for the smaller assembly sizes as a result of repetitive regions being collapsed during assembly. Outside Clade V, the only

genome in this list that is greater than 100 Mbp is that of *A. suum*, which is known for its chromatin diminution [37].

One way of checking the completeness of a newly sequenced and assembled genome is to use the core eukaryotic genes mapping approach (CEGMA) as described in Chapter 2. CEGMA percentage completeness values for three of the four genomes sequenced for this thesis (*D. immitis*, *L. sigmodontis*, and *Caenorhabditis sp. 5*) are in the high 90s, which is higher than some of the genomes obtained using Sanger-capillary sequencing (*T. spiralis*, *B. malayi*, *M. hapla*, *M. incognita*, *P. pacificus*, and *C. japonica*). As reported in the previous chapter, the newly sequenced *M. floridensis* and previously sequenced *M. incognita* genomes have much lower CEGMA values (72.18% and 75.81% completeness of partial models) indicating incomplete assemblies or unusual core gene structures. To check whether the genes are missing or fragmented due to poor assemblies, a further study will be needed where the extra *M. hapla* CEGMA gene set (which is 92.74% complete, Figure 3.2) should be used as a query to search for matches in the *M. incognita* and *M. floridensis* genomes. If all the fragments are found across different contigs, then the low CEGMA numbers are likely to be due to fragmented assemblies. If no matches are found, then the CEGMA genes are most probably missing in these two *Meloidogyne* draft genomes.

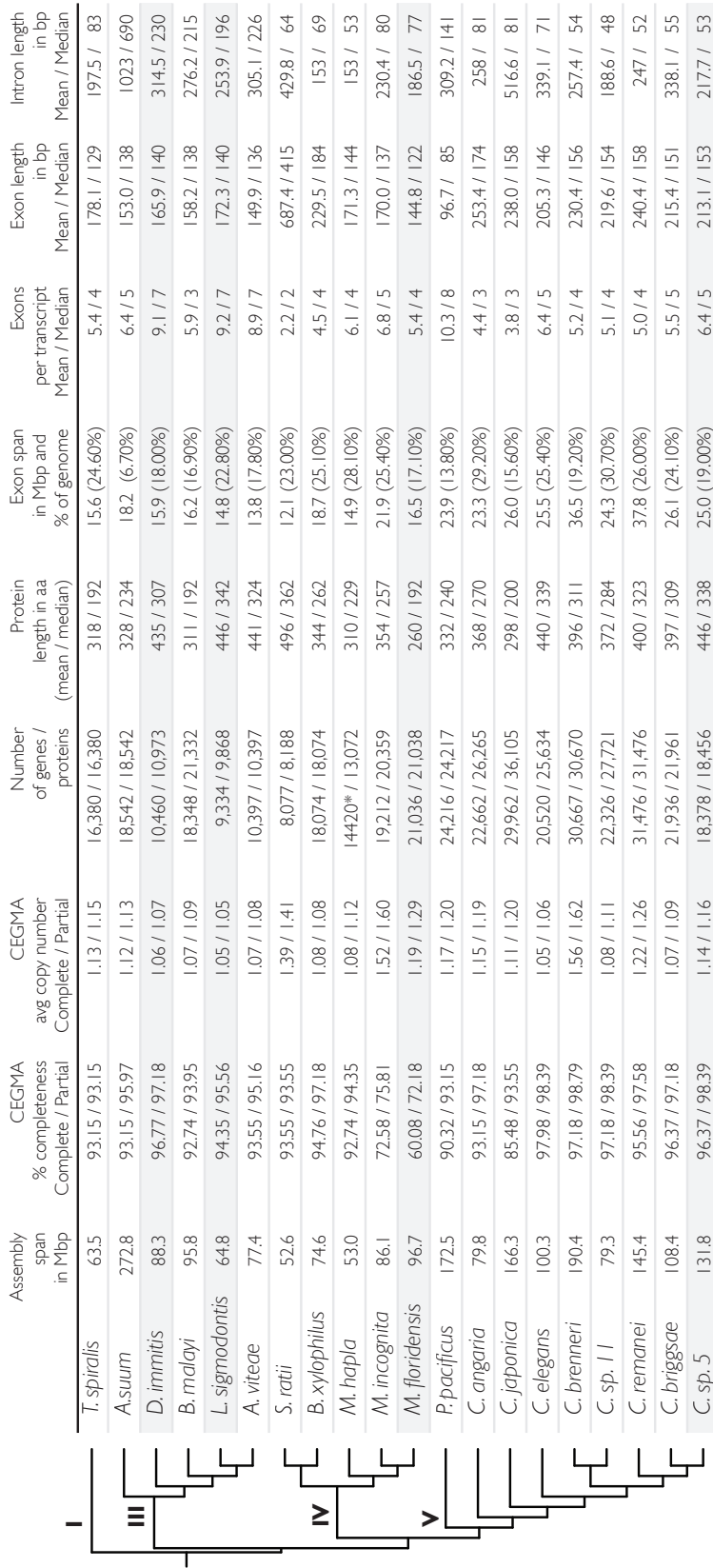
CEGMA also provides an average copy number of highly conserved, putatively single-copy core eukaryotic genes. This number is very close to 1.0 in two of our four genomes (1.06 and 1.05 in *D. immitis* and *L. sigmodontis*). In *Caenorhabditis sp. 5* and *M. floridensis*, the average copy numbers are 1.14 and 1.19, respectively, indicating that some of these core genes are present in more than one copy. A higher CEGMA average copy number could indicate a biological variation, but is more often the result of independent assemblies of haploid components of a diploid genome as might be encountered in a highly heterozygous DNA sample (if obtained from a non-clonal wild population of nematodes). Three of the genomes available on WormBase have relatively high CEGMA average copy numbers: *S. ratti* (1.39), *M. incognita* (1.52), *C. brenneri* (1.56), and *C. remanei* (1.22), which may also indicate that these assemblies do not accurately represent haploid genomes.

The number of genes predicted in each of our four species is comparable to the numbers predicted in genomes from the same clade. Almost all the Clade V species have a minimum of 21,000 protein-coding genes. The newly sequenced and annotated *Caenorhabditis sp. 5* genome is the only exception with 18,456 predictions. The lower number could be due to the strict requirement that all predictions have at least some evidence in the form of an EST or protein alignment. If purely *ab initio* gene models had been allowed, 40,813 *Caenorhabditis sp. 5* proteins would have been predicted, most of which would have been spurious. In Clade IV, 21,038 proteins were predicted for *M. floridensis*, which is comparable to the 20,359 proteins predicted in *M. incognita*. The two new clade III species *D. immitis* and *L. sigmodontis* have relatively fewer predicted proteins (10,973 and 9,868 respectively) compared to the 21,332 predictions in the closely related onchocercid filarial nematode *B. malayi*. However, it

is important to note that the *B. malayi* genome was published in 2007 [46] with only 11,460 nuclear proteins. The approximately 10,000 extra recent gene predictions are all based on a computational pipeline without additional evidence and may therefore be an over-prediction. Similarly, the *A. suum* annotation with 18,449 genes might be an over-prediction as only 14,783 were supported by transcriptome data [36].

Some other gene prediction metrics also seem to be clade specific. Three of the four Clade III Onchocercidae have a median of seven exons per transcript, although *B. malayi* only has a median of three exons. The smaller number of exons for *B. malayi* is an artefact of the excess *ab initio* predictions that have fewer exons per transcript. The original 11,460 *B. malayi* nuclear CDSs published in 2007 [46] had a median of five exons per transcript, and the mean and median protein lengths were also higher for the original set of *B. malayi* predictions at 371 aa and 272 aa respectively, indicating that the new predictions were both shorter and had fewer exons per transcript. A subset of the genus *Caenorhabditis* (*elegans*, *brenneri*, *sp. 11*, *remanei*, *briggsae*, and *sp. 5*) also has a median of four or five exons per transcript, and this number drops to three for the two *Caenorhabditis* species outside this group.

Median exon-lengths are also in a narrow band from 122 bp to 184 bp with the exception of *P. pacificus* at 85 bp and *S. ratti* at 415 bp. The extremes in numbers are most likely a result of methodological biases during the gene-finding process. Future comparative studies on nematode exon-lengths could confirm this by using the same gene-finding algorithm on all the species being analysed. With the exception of these two species, the mean exon-lengths for the *Caenorhabditis* species are all considerably higher than the other ten species, indicating a more positively skewed exon-length distribution in the genus *Caenorhabditis*. The median intron-lengths are equally clade specific, with the *Caenorhabditids* and Clade IV nematodes ranging from 48 bp to 79 bp. However, Clade III seems to have very different median intron lengths, ranging from 195 to 226 for the four Onchocercidae, rising to 689 bp for *A. suum*. Lastly, although absolute exon spans are higher overall for Clade V nematodes (>23 Mbp), the percentage span relative to the size of the genome shows no such trend. The only outlier is *A. suum* with just 6.7% of the genome spanned by exons.



Note: * indicates missing entries in the protein fasta file. The gene number is correct.

Figure 3.2 Comparison of gene model statistics across 20 nematode genomes

3.3.2 InterProScan annotations across 20 nematode genomes

Figure 3.3 shows the top 100 InterPro signatures identified by the InterProScan program across 20 nematode proteomes. Each row represents an InterPro signature and the shades in the heatmap indicate the number of proteins in each species with that InterPro signature. The most frequent InterPro signatures (by total count across all 20 species) are shown at the bottom of the heatmap and, as expected, include common domains such as protein kinase, NAD(P) binding, and Zinc finger domains.

Caenorhabditis species have many more protein domain annotations than the other nematodes, which is not surprising given that Figure 3.2 showed that these species also had more protein predictions than the rest. One way of looking for interesting differences in protein domain membership is to scan the heatmap for domains that are over-represented in a species that generally has an under-representation of domains. For example, the column for *S. ratti* is generally lighter in colour than all the other columns because it has fewer proteins than the others (8,188, compared to the 20-species average of 20,604). However, some domains are over-represented in *S. ratti* compared to other nematodes such as IPR016040 NAD(P)-binding domain, IPR011991 Winged helix–turn–helix transcription repressor DNA–binding, IPR003593 ATPase AAA+ type core, IPR001128 Cytochrome P450, IPR002213 UDP–glucuronosyl/ UDP–glucosyltransferase, and IPR012336 Thioredoxin–like fold. Visualisations like Figure 3.3 are a first step towards identifying species or clade differences that can be elaborated by closely examining the over- or under-represented protein domains and seeing what biological functions they typically correlate with.

A striking feature of this heatmap is the very high F-box domain count for *C. remanei* proteins (1157 and 1110 for the last two rows: IPR001810 and IPR012885). F-box domains are associated with the ubiquitin protein degradation pathway (found in most eukaryotic tissues) and are also present in higher numbers in *C. elegans*, *C. brenneri*, and *Caenorhabditis sp. 11*. Although the shades are quite light, all 20 species have at least 3 proteins annotated with this domain (median: 33), but it is unusual to see >1000 proteins in *C. remanei* annotated with the F-box domain. As with the gene models in the previous section, extreme counts could point to interesting biological differences or methodological biases and need to be investigated further. Redoing the genome annotations for all *Caenorhabditis* species using the same pipeline would eliminate methodological reasons for the difference.

The column for *C. japonica* indicates a possible failure of repeat-masking for transposable elements as the following domains are present in excess compared to other species and are related to retroviral and retrotransposon activity: IPR000477 Reverse transcriptase; IPR001878 Zinc finger, CCHC–type; and IPR001584 Integrase, catalytic core. *T. spiralis* is similarly enriched for IPR001878 and IPR001584, and possibly also indicates unmasked retrotransposons.

C. briggsae and *Caenorhabditis sp. 5* are closely related and have similar numbers of proteins, but the former is androdioecious (self-fertilizing hermaphrodites with rare males), whereas the latter is gonochoristic (equal ratio of males and females). Differences in protein domain counts between the two species therefore might hint at the mechanism underlying reproduction differences. Some of the InterPro domains that differ between these species are: IPR002290 Serine/threonine dual-specificity protein kinase; IPR020635 Tyrosine-protein kinase; IPR013069 BTB/POZ; IPR000210 BTB/POZ-like; and IPR011333 BTB/POZ fold. BTB/POZ is a protein-protein interaction motif that is associated with Zinc finger proteins, including those involved in repressing transcription by modifying chromatin, and this difference could indicate a change in chromatin regulation.

All Clade V nematodes and the *elegans* group in particular (*C. elegans*, *C. brenneri*, *Caenorhabditis sp. 11*, *C. remanei*, *C. briggsae*, and *Caenorhabditis sp. 5*) are abundant in proteins containing 7TM GPCR serpentine receptor domains, which is expected as these are chemosensory proteins essential for these free-living species. *B. xylophilus*, a Clade IV plant-parasitic nematode, is also enriched in these domains as noted in Kikuchi *et al.* [50], but other nematodes appear to have a less rich sensory repertoire.

C-type lectins, associated with anti-infection mechanisms in many cases, are diminished in number (median: 12) in the onchocercid filarial nematodes (*D. immitis*, *B. malayi*, *L. sigmodontis*, and *A. viteae* in Clade III) compared to the free living Clade V species (median: 185). One possible explanation is that filarial nematodes do not need as elaborate an immune system because they reside in the protected environment of a host.

All of the above observations are crude hypotheses based on a visual examination of Figure 3.3. We did not test any of these hypotheses or explore them further as that was not the primary motivation for annotating these genomes. These annotations and all other intermediate files are available at the data source listings in Appendix B and should be of use to anyone interested in genome-level comparisons of nematodes.

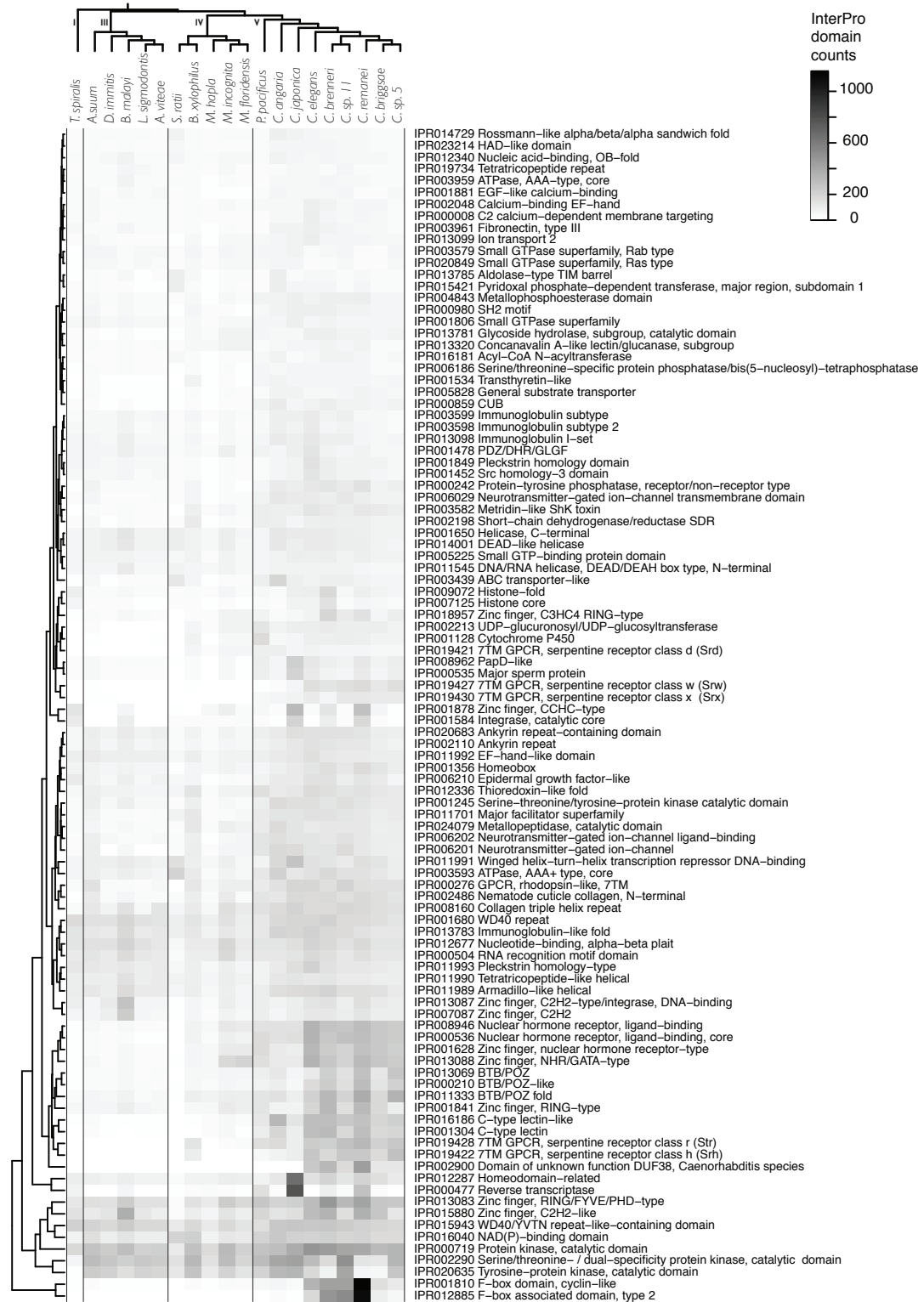


Figure 3.3 InterProScan annotations for 20 nematode proteomes

3.3.3 tRNA predictions across 20 nematode genomes

The results of running tRNAscan (version 1.3 [177]) across all 20 genomes are shown in Figure 3.4. The GC content of synonymous tRNAs was calculated by dividing the count of tRNAs with G or C in the third-base position by the total count of synonymous tRNAs for each amino acid in each species (only amino acids with 2 or 4 alternate codons were considered). The GC content of the genome and the mean GC content of synonymous tRNAs is shown below the heatmap in Figure 3.4. The two are correlated (Pearson's $\rho = 0.49$, $p < 0.05$), which agrees with Cutter et al.'s [183] finding that the major codons in nematode ESTs are correlated with overall GC composition.

All 9 species in Clade V have more copies of almost every type of tRNA compared to the 11 species in Clades I, III, and IV. The greater number of tRNA predictions is not solely due to the genome size; compared to most Clade III species, *C. angaria* and *Caenorhabditis sp. 11* have smaller genomes but more tRNAs. *A. suum* in Clade III is the largest genome (272.8 Mbp) and yet it only has 255 tRNA predictions in total compared to the much smaller *Caenorhabditis sp. 11* (79.3 Mbp), which has 538 tRNA predictions. The two branches in Clade IV are notable because the *Meloidogyne* branch has very few predictions (with the exception of threonine ACC), whereas the other branch (*S. ratti* and *B. xylophilus*) has predictions comparable to Clade V. Therefore, the difference in number of predictions is unlikely to be an artefact of being closely related to the model *C. elegans* genome as the two branches are equally distant from *C. elegans*. *M. incognita* and *M. floridensis* have more threonine ACC predictions (186 and 95 respectively) than any other codon in any other species (except arginine_ACG in *P. pacificus*). This abundance could have a biological explanation or it could be an artefact of a transposable element that contains a threonine tRNA or threonine-tRNA-like structure. We checked if the number of threonines in the protein predictions for these species was similarly in excess, but both *M. incognita* and *M. floridensis* had a threonine amino acid count that was close to the average for all amino acids in these two proteomes. Like other genome features in this chapter, tRNA predictions also seem to be clade-specific. As all the tRNA predictions were done using the same method, these differences are unlikely to be methodological and are likely correlated with other phenomenological differences. These results might aid future investigations into the differences in protein-producing capabilities of nematodes.

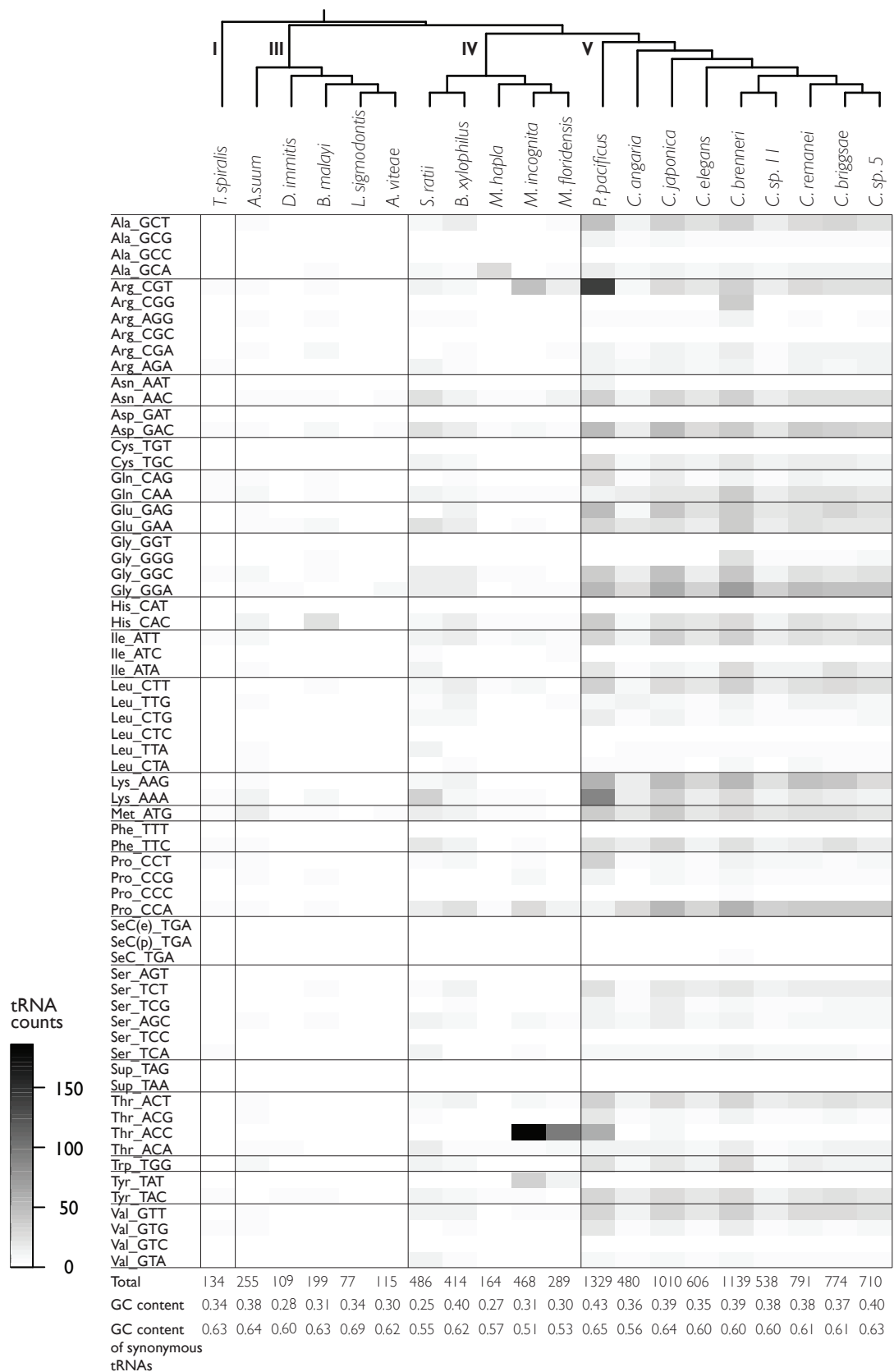


Figure 3.4 tRNA predictions for 20 nematode genomes

3.4 Discussion

3.4.1 Automated gene prediction and functional annotation have limitations

Typically, genome annotation projects will use the gold standard of manual annotation for a few hundred genes and use these well-annotated genes to train gene finders for the rest of the genome. However, for our goal of rapidly creating low-cost genomic resources for many species, we needed reliable automated procedures that would give a usable first-pass annotation. Our two-pass MAKER2 workflow achieves this goal, although end-users of these protein predictions should be aware that these predictions may not be complete and that a missing gene could be the result of poor assembly or poor annotation. Thus, only aggregate patterns should be trusted. If a specific gene is of interest, a more sensitive search against the genome assembly should be carried out using the BLAST server at <http://nematodegenomes.org>.

Each gene-finding tool also has many parameters to explore, which exponentially increases the number of options available. This problem of exploring parameter space is a recurring theme in all bioinformatics analyses, but is especially problematic in the case of annotating non-model genomes because the truth is not known. The nGASP project [153] attempted to address this problem by evaluating alternative methods against a well-documented gene-set from the model nematode *C. elegans*. However, that study is already 4 years old and RNA-Seq data from NGS have the potential to vastly improve annotation efforts. Efforts are under way in our lab to determine the most useful automatic gene-finding workflow using *ab initio* predictors and RNA-Seq data.

A major limitation of all automated gene prediction tools is that it is very hard to determine alternatively spliced transcripts. However, as NGS becomes more accessible and more RNA-Seq data is commissioned for differential expression studies, it should become easier to identify alternate transcripts using tools like TopHat and Cufflinks [161, 184].

The MAKER2 workflow described here may turn out to be sub-optimal in a more systematic evaluation, but it currently has the advantage of being the only eukaryotic annotation tool supported by the GMOD consortium, which ensures that its output is standardised and can be easily used by other software tools in GMOD. The two-pass use described here was first suggested in an online discussion group [185] and seems to be the best way to incorporate all known information, such as multiple *ab initio* predictions, as well as evidence in the form of EST alignments from the same species and protein alignments from closely related species. The workflow also addresses the problem of obtaining a good training set for the *ab initio* predictors, Augustus and SNAP by using the gene models from the first-pass as training data for the second pass.

The main problem associated with using this method is to decide if the second-pass MAKER2 run should keep *ab initio* predictions without any evidence (i.e. with an AED of 1.00). If we use MAKER2 with evidence (low AED values), then novel proteins will be left out. If we use MAKER2 and allow predictions without evidence (high AED values), we get approximately double the number of expected proteins in some cases, although some of them may be true. Additionally, the appropriate AED cutoff for each species might be different and would need to be determined empirically using the suggestions in [142].

One of my main goals of annotating these four genomes was to identify coding regions that I could reliably remove in order to analyse conserved non-coding elements (Chapter 4). For my purposes, I conservatively only kept gene predictions that had some evidence (AED <1). However, a secondary goal was also to generate a genomic resource that could be used by other people interested in the phenotypes of these nematodes, for which they might want as many gene predictions as possible. My solution was to generate the more sensitive output with an excess of predictions and to write scripts that filtered the protein and GFF files based on an AED cutoff. The GFF and protein files are available via <http://nematod.es> (Appendix B) and the scripts and commands for extracting subsets of predictions are also available (Appendix A).

3.4.2 *C. elegans* is not "the" nematode

Although *C. elegans* has been called "the" worm [31], the advent of next-gen sequencing and comparative genomics is rapidly making it obvious that the genomic characteristics of *C. elegans* are possibly limited to only the genus *Caenorhabditis* and do not extend to the rest of Clade V, let alone the phylum.

This chapter presents an analysis of gene model features and functional annotations of 20 species of nematodes. In our simple tabulation of gene model characteristics, many features showed clade-specific trends: Clade V species had more protein predictions than the other clades; intron lengths had a very large median within Clade III compared to the rest of the phylum; and most Clade III species also had fewer protein predictions and a higher number of exons per gene with the exception of *B. malayi*. We currently offer these observations without any explanation, but we believe that the clade-specific nature of these features should correspond to phenotypic correlations. Such pan-nematode gene and genome comparisons will provide rich opportunities for future research.

In addition, comparisons across genomes can identify exceptions to a trend. Such exceptions could be due to methodological differences or due to a genuine biological phenomenon. For instance, we know that the excessive *B. malayi* protein predictions compared to other Clade III nematodes are most likely due to an over-prediction of gene products. On the other hand, it is also possible that the exceptions are real and are biologically significant. Either way, a comparative approach can be very useful.

3.4.3 Future improvements

Automated genome annotation can be improved by using more complete evidence sets. Previously, generating cDNA evidence in the form of Sanger-sequenced ESTs was a time consuming project that could take several months. With NGS, a single lane of Illumina sequencing can provide as much as 40 gigabases of RNA-Seq data in a week, which can be assembled into transcripts and used to improve gene models on a draft genome.

In our genomes, *L. sigmodontis*, *D. immitis*, and *Caenorhabditis sp. 5* all had limited RNA-Seq or EST evidence data, and *M. floridensis* had no such data (Table 3.2). *L. sigmodontis* was perhaps the most complete with data from 3 life stages: male, female, and microfilaria. It was also the best assembled transcript set because it used 454 sequencing which generates longer 200-400 b reads. If more RNA-Seq data is generated for these species in the future (e.g., for differential expression studies), then these data can be incorporated to give better gene models. Short-read RNA-Seq data also need to be re-assembled using the best available transcriptome assembly programs. Unlike the Assemblathon effort, no systematic independent reviews of transcriptome assemblers have been conducted so far. The current best program seems to be Trinity [167] according to a crowd-sourced competition site [186]. Better draft genome assemblies also yield better gene predictions because they recover gene models that might be missed due to the gene regions being split into separate contigs or scaffolds. For example, the previously released *D. immitis* genome assembly [51] had 31,291 scaffolds and 10,179 protein predictions, whereas the improved assembly described in this thesis (obtained by using filtered MP reads) was less fragmented with only 16,061 scaffolds and ~800 more protein predictions.

The drawback of using NGS to rapidly improve genome assemblies and annotations is that it can be hard to keep track of the assemblies and the annotations associated with them. In the past, nematode genome projects, such as *C. japonica*, typically had one genome assembly release and one annotation release. With sequencing, assembly and annotation technologies changing rapidly, even a small research group might generate many assembly versions and many annotations of each assembly in an attempt to generate better results. Although end-users will only see occasional releases, we soon discovered the need for systematically versioning all our data files. We used a three-part naming convention for all four of our genomes that simplified keeping track of dataset, assembly, and annotation versions as we worked on improving each stage of the process. For example, the *D. immitis* nuclear genome release nDi.2.2 uses two digits: the first "2" describes the dataset iteration (the raw data was error-corrected and digitally normalised in this iteration), and the second "2" describes the assembly iteration. A third digit is used to define the annotation iteration. Thus, the first run of the two-pass MAKER2 workflow on this assembly was labelled nDi.2.2.1. We hope this system will make it easier for other groups to manage their genomes as well.

One last issue that needs improvement is consistency in public data sets. The methods section in this chapter briefly described inconsistencies in the data files available at WormBase. Although the WormBase team was very prompt and helpful in addressing these problems, simply identifying the problems took a considerable amount of time. Programs like MAKER2 explicitly have data management as one of their goals and are therefore very useful in generating consistent and valid outputs that others can use easily. As described in Chapter 1, the nematode genome community is a collaborative federation rather than a single project, and thus it is hard to get everyone to agree to a standard workflow for generating genome assemblies and annotations. In fact, it would probably not be desirable to have just one way of creating these genomic resources because that would slow the adoption and innovation of new methods. However, we hope that the community will agree on a consistent and valid genome and annotation format that will make future comparisons across all nematode genomes, such as the ones in this thesis, easier.

4 Lack of deeply conserved non-coding elements in nematodes

4.1 Introduction

Many complete metazoan genome sequences have become available in the last decade permitting comparative analyses on the non-protein-coding and non-RNA-coding parts of their genomes. These studies found several non-coding sequences that were far more conserved than expected which implied that such sequences were being maintained by selection and therefore were functional [187, 188]. Previous studies on conserved non-coding elements (CNEs) identified elements that shared high levels of identity even between species that diverged hundreds of millions of years ago [189]. CNEs were found close to genes that were related to development, and some CNEs also acted as tissue-specific enhancers during development [190]. Although CNEs had high levels of identity within a phylum, they were not shared across phyla (e.g., vertebrates did not share any CNEs with nematodes), leading to the speculation that perhaps CNEs were highly specific to a phylum and defined the phylum in some way [191, 192].

In this chapter, I explore the hypothesis put forth previously that CNEs are associated with animal phylum body plans and that they may have arisen in a "big bang" at the time of the Cambrian "explosion" when many phyla with diverse body plans first emerged [191]. The chapter begins by providing an overview of the research on CNEs and the different groups of species in which CNEs have been studied. Previous analyses of CNEs in nematodes used three species from the genus *Caenorhabditis*: *C. elegans*, *C. briggsae* and *C. remanei*, as these were the only complete genomes available at the time [191]. Thanks to Sanger and NGS, twenty complete nematode genomes are now available (including the four genomes described in this thesis) and I used all twenty to search for Nematoda-wide CNEs. A new, more sensitive method for identifying CNEs is described in the Methods section. No CNEs were found that spanned the whole phylum, although clade-specific CNEs do exist. The chapter finishes with a discussion of the absence of phylum-wide CNEs, and how CNEs are not likely to be related to the phylum body plan. This negative result demonstrates the value of sequencing non-model organisms using NGS.

4.1.1 Previous CNE research on vertebrates, insects, and nematodes

Different research groups have used different terms for CNEs, such as Ultra-Conserved non-coding Region (UCR) [193], Conserved Non-coding Sequences (CNS) [194], and Conserved Non-Genic sequence (CNG) [195]. A more accurate term might be "non-exonic" as used by

Lowe *et al.* [196] to imply that the functionality of these elements is not due to their exons being transcribed into an RNA product. Drake *et al.* [194] also showed that these CNEs are selectively constrained and therefore not simply mutation cold spots. The term CNE was first used by Woolfe *et al.* [188] and is used in this chapter because the analysis presented here builds on their work as well as the work by Vavouri *et al.* [191] and Vavouri and Lehner [192] who also use the same term. For this study, the term "non-coding" includes non-protein-coding and non-RNA-coding regions.

In the studies mentioned here, the process of identifying CNEs required filtering out regions that were known to code for protein or RNA coding exons. To ensure that a putative CNE was not simply an unannotated exon, it was checked against the exons of all the species being analysed. Therefore, an exon would have to be unannotated in all the species being considered (including well-annotated genomes such as *C. elegans* in the case of nematodes) for it to slip through the filtering process. Additional checks were also carried out. For example, in Vavouri *et al.* [191], all initial candidate CNEs were queried against the European Molecular Biology Laboratory (EMBL) EST database to ensure that no sequences matching any known coding sequences were kept. Despite these filters and checks, a few unannotated exons might have been incorrectly identified as CNEs but they are likely to be the rare exception and unlikely to account for the larger signal. Some CNEs could have been present in the 5' and 3' UTRs of genes. If such CNEs are highly conserved and under selection, then it is possible that they have a regulatory function as well.

CNEs have been found in different groups of species using various methods (Table 4.1). Despite diverging over 450 million years ago (MYA), humans and pufferfish have nearly 1,400 elements with a mean length of ~200 bp and a mean identity of 84% [188]. Some CNEs in this set are over 500 bp long with greater than 90% identity and are more conserved than coding sequences between these two species. Woolfe *et al.* also discovered that most of these elements are near genes that act as developmental regulators and that 23 of the 25 elements picked for testing showed significant enhancer activity. In a study on ultra-conserved elements, 256 non-exonic elements longer than 200 bp with 100% identity were shared between humans, mice, and rats [187]. These ultra-conserved, non-exonic elements flanked a set of genes that was significantly enriched for genes with early developmental roles. Glazov *et al.* [197] found 20,301 intronic and intergenic ultra-conserved elements (≥ 50 bp, 100% identity) shared between two *Drosophila* species, with the genes closest to these elements enriched for GO terms related to transcription factors.

Table 4.1 CNEs found in different groups of species

Species compared	Number of CNEs and level of identity	Approximate last common ancestor
Vertebrates [188]: <i>Homo sapiens</i> (Human), <i>Takifugu rubripes</i> (Pufferfish)	1,373 elements; 84% identity; Average length ~200 bp	450 MYA
Mammals [187]: <i>Homo sapiens</i> (Human), <i>Mus musculus</i> (Mouse), <i>Rattus norvegicus</i> (Rat)	256 elements; 100% identity; Length > 200 bp	55 MYA
Fruit Flies [197]: <i>Drosophila melanogaster</i> , <i>Drosophila pseudoobscura</i>	20,301 elements; 100% identity; Length > 50 bp	25-55 MYA
Nematodes [191]: <i>Caenorhabditis elegans</i> , <i>Caenorhabditis briggsae</i> , <i>Caenorhabditis remanei</i>	2,084 elements; MegaBLAST word seed size 30bp (W30) with e-value threshold 0.001; Average length 69 bp	30 MYA

Note: Adapted from Kumar (2007) [198]

Vavouri *et al.* [191] performed the first systematic search for CNEs in nematodes. Using the three complete nematode genomes available at the time, they identified 2,084 CNEs shared between *C. elegans*, *C. briggsae*, and *C. remanei* with a mean length of 69 bp and a mean identity of 96% (between *C. elegans* and *C. briggsae*). 990 elements were 100% identical between all three and 93% of the total sequence in these elements was found to be under purifying selection. They found that nematode, insect, and human CNEs are associated with genes involved in development and transcription regulation and, to some extent, with genes related to cell-signalling, although the latter association was weaker in vertebrates. Confirming previous findings in humans, they found further evidence that CNEs act as *cis*-regulatory enhancers that encode transcription factor binding sites (TFBSs). More importantly, they discovered that 40 of the 156 CNE-associated genes in humans had direct orthologs in *C. elegans* and *D. melanogaster*, and all these orthologous genes were associated with CNEs in both species. Thus, not only were CNEs closer to genes related to development, they were often near developmental genes with the same function in evolutionarily diverse species.

If CNEs are indeed the genomic substrate on which proteins and other molecules bind to regulate the expression of developmental or other genes, then it is important to remember that there may be other such elements that do not retain genomic conservation across species. The functional event is the binding of a molecule to the regulatory region on the genome. Therefore, if the molecule and the binding region both co-evolve, it is possible for the same event to be conserved across species at the functional level without genome level conservation of sequence for that event. Thus, although the presence of CNEs in a set of species might indicate shared GRNs among those species, the absence of highly conserved sequences does not necessarily indicate that they do not share similar GRNs.

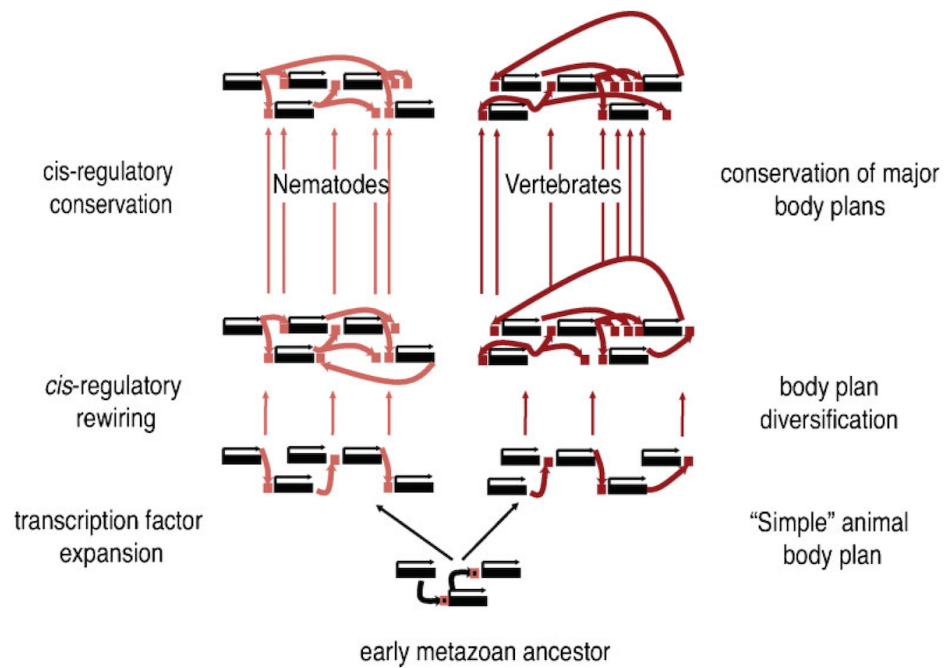


Figure 4.1 A model for the evolution of cis-regulatory elements involved in animal development

According to this model, duplication and rewiring of the regulatory toolkit of the common animal ancestor gave rise to a diverse set of complex regulatory elements that formed the core developmental programs of the major animal phyla. Since then, animal body plans have been largely conserved. This conservation may be reflected in a set of highly conserved cis-regulatory elements controlling the expression of developmental genes. Figure and caption taken from Vavouri and Lehner [192]

4.1.2 CNEs might define the phylum body plan

Based on their observations of CNEs in nematodes, insects, and vertebrates, Vavouri *et al.* [191] proposed that CNEs might be responsible for phylum body plans. This hypothesis was elaborated in their opinion piece [192] and is summarised as follows:

1. Even though the core developmental genes are often the same across phyla, the interactions between these genes are different for different phyla. These interactions make up the Genetic Regulatory Networks (GRNs) that correspond to phylum body plans.
2. A GRN is defined by transcription factors, regulatory elements, and enhancers that allow the output of one gene in the GRN to control one or more other genes in the GRN.
3. CNEs act as regulatory elements and enhancers by providing TFBSs for core developmental genes.
4. CNEs are highly conserved within a phylum, but are not shared across phyla.
5. Therefore, CNEs are the substrate that allows different body plans to be defined (i.e., when CNEs change, the phylum body plan changes).

This is a very elegant hypothesis that could, according to the authors, also explain the Cambrian "explosion" of metazoan diversity (Figure 4.1). Once an early metazoan ancestor developed a toolkit of genes that worked well together, an expansion in transcription factors and *cis*-regulatory regions (CNEs) could allow for these genes to interact in a vast number of combinations and allow many viable topologies to emerge.

4.1.3 Aims for this study

The hypothesis that CNEs are components of the GRN which defines a phylum body plan was developed using only a few genomes from each phylum. Only three complete nematode genomes were available for Vavouri *et al.*'s [191] analysis. The subsequent availability of 17 more complete genomes from the phylum Nematoda (four of which are described in this thesis) provided the perfect data set for studying CNEs in more detail.

We used 20 nematode genomes (as in Chapter 3) to answer the following questions:

1. Do nematodes have CNEs that span the whole phylum?
2. If there are phylum-wide CNEs, are they near the same kinds of genes as the CNEs discovered previously, or does the pattern of enrichment in CNE-associated genes change?
3. If there are phylum-wide CNEs, do they show any patterns of identity, gain, or loss that could explain when they arose? For example, can we tell if all phylum-

wide CNEs arose just once at the base of the phylum (~550 MYA) or whether new ones are constantly emerging.

The primary goal of this study was to look for CNEs across the phylum Nematoda using the hard definition of a CNE: sequences that are highly conserved across species and do not result in a protein or RNA product. As noted previously, the absence of CNEs does not necessarily indicate that GRNs are not shared, only that CNEs are not shared. The implications of the absence of CNEs are further elaborated in the Discussion section.

4.2 Methods

To find CNEs across a given group of species, we developed a new workflow that used whole-genome multiple alignments and was more sensitive than the MegaBLAST-based method used by Vavouri *et al.* [191]. An overview of the workflow and the details of each step are described below. Very briefly, we first found conserved elements using whole-genome alignments and then removed the coding regions from the alignments to obtain conserved non-coding elements.

4.2.1 Genome and coding-region data

To find coding regions for each species (Figure 4.2-A), genome, protein, and annotation files for fifteen species were downloaded from the WormBase FTP site (release WS230, [30]) as in Chapter 3. All scripts and workflows for processing the raw files are available in Appendix A. Data for the four species assembled and annotated in Chapters 2 and 3—*Caenorhabditis sp. 5*, *M. floridensis*, *D. immitis*, and *L. sigmodontis*—were added to this collection from their respective genome pages at <http://nematod.es>. An additional Clade III onchocercid nematode species *Acanthocheilonema viteae* was also assembled and annotated by colleagues at the Blaxter Lab with the help of the tools and pipelines described in Chapters 2 and 3. Data for this species were downloaded from <http://acanthocheilonema.nematod.es>. All files were renamed using short species names: *C. elegans* (ce), *C. briggsae* (cbg), *Caenorhabditis sp. 5* (csp5), *C. remanei* (cr), *C. brenneri* (cbn), *Caenorhabditis sp. 11* (csp11), *C. japonica* (cj), *C. angaria* (ca), *P. pacificus* (pp), *S. ratti* (sr), *B. xylophilus* (bx), *M. hapla* (mh), *M. incognita* (mi), *M. floridensis* (mf), *A. suum* (as), *B. malayi* (bm), *L. sigmodontis* (ls), *A. viteae* (av), *D. immitis* (di), and *T. spiralis* (ts). These short names were prefixed to all chromosome, contig, scaffold, and protein names, to make data management and tracking easier.

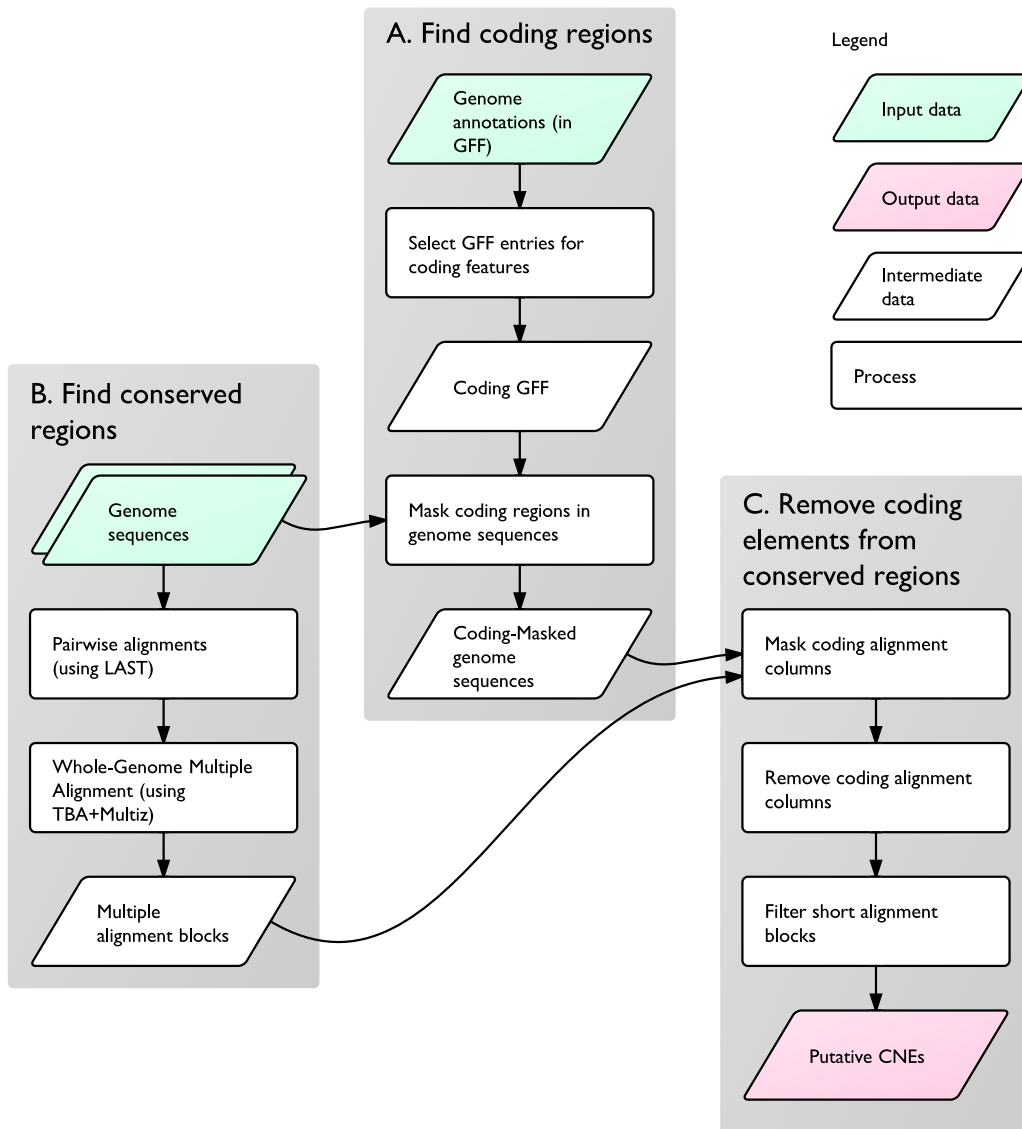


Figure 4.2 Workflow for finding CNEs using whole-genome multiple alignments

Coding intervals for each species were obtained from annotation GFF files (see Chapter 3 for annotation sources) and by running Rfamscan [199] and tRNAscan [177] separately on each genome (not all genome GFF files included information about RNA-Sequences). To get protein-coding GFF entries, GFF annotation files for each genome were examined to identify source-feature entries that represented coding regions because each genome used different source-feature combinations such as "WormBase, exon" or "curated, CDS" or "Augustus, CDS". The final list of GFF source-feature combinations used for identifying coding features is given in Appendix A (Identify CNEs using whole-genome-alignments). Protein- and RNA-coding entries were extracted for each species and placed in a new coding GFF file. Because the original annotation GFFs for each species were created by different groups using different programs, the transcript ID for each coding GFF feature was standardised by adding an extra attribute "Coding_name" (e.g., each exon feature from the *C. elegans* transcript "4R79.1a" would have "Coding_name=4R79.1a" appended). We used these standardised coding GFF files for each species to soft-mask all the genome fasta files with lowercase letters for coding regions (Appendix A: interval_mask.pl).

4.2.2 Identifying CNEs using whole-genome alignments

Pairs of genomes were aligned to each other using the LAST [200] alignment software, and the alignments were stored as Multiple Alignment Format (MAF) files (Figure 4.2-B). Pairwise MAF files were filtered to remove regions that aligned more than once so that only single-coverage regions remained. The TBA+Multiz [201] pipeline (TBA version 12 and Multiz version 11.2) was used to generate several whole-genome multiple alignments using phylogenetic guide trees. The guide trees corresponding to each analysed node in Figure 4.5 were rearranged so that the better assembled and annotated species occurred earlier in the list where possible. This was done because the TBA+Multiz pipeline uses the first species in each pair as the reference sequence (Table 4.2).

Each guide tree was used to generate a whole-genome multiple alignment of all the species in that branch of the nematode phylogeny (Figure 4.3). Alignments were stored as MAF files containing many alignment blocks each, where each alignment block corresponded to a conserved element. Figure 4.4-A shows a cartoon example of a short alignment block for the Onchocercidae alignment. Each block begins with an "a" and a score, followed by lines prefixed with an "s" that store the sequences in the alignment from each species. The scoring depends on the program used for alignment. The steps outlined below for processing these alignment MAF files to get CNEs correspond to the processing steps in Figure 4.2-C. Each step was implemented as a separate script rather than as a monolithic program because these scripts manipulate MAF files in a generic way and can be used for purposes other than finding CNEs.

Lower-case coding masked genome files were used to replace coding regions in each alignment block with lowercase bases (Figure 4.4-B). Alignment columns were marked as coding (i.e., all nucleotides in that column were converted to lowercase) even if only one sequence in that column had a coding (lowercase) base (Appendix A: maf_insert_lc.pl). Figure 4.4-C shows the result of this step.

Lower-case columns were removed from the alignment (Appendix A: maf_remove_lc.pl), resulting in split alignment blocks if the coding regions were in the middle of an alignment (Figure 4.4-D). The final split alignment blocks were output with new genome coordinates, each with new block scores that were scaled by the proportion of the length of the split block relative to the length of the original alignment block. The split blocks represent an alignment of conserved non-coding elements.

Some of the resulting CNE alignment blocks were very short or mostly made up of padding characters "-" as a result of the splitting step. CNE blocks that were shorter than 30 columns or had less than 50% relative identity were discarded (Appendix A: maf_select.pl). The relative identity of an alignment block was defined as the number of columns where more than half the sequences matched the consensus (except columns where the consensus was a padding base) divided by the total number of columns in that alignment. The consensus of a column was defined as the character (base or padding character) that occurs in at least half the rows of a column. Therefore it was possible for a column to have no consensus. This script printed the length, absolute identity, and relative identity of each CNE alignment block. The script also filtered out CNE blocks that did not contain every species that was used to create the initial alignment. The final CNE file in MAF format was converted to a set of GFF files for each species storing the genomic coordinates of each CNE (Appendix A: maf_to_gff.pl).

Table 4.2 TBA+Multiz guide trees

Branch	Guide tree (modified Newick format used by TBA+Multiz)
Elegans group	(ce (((cbg csp5) cr) (cbn csp11)))
Caenorhabditis	((((ce (((cbg csp5) cr) (cbn csp11))) cj) ca)
Clade V	(((((ce (((cbg csp5) cr) (cbn csp11))) cj) ca) pp)
Meloidogyne	(mh (mi mf))
Clade IV	((sr bx) (mh (mi mf)))
Onchocercidae	((bm (ls av) di)
Clade III	(as ((bm (ls av) di))
Nematoda (all 20 species)	((((((((ce (((cbg csp5) cr) (cbn csp11))) cj) ca) pp) ((sr bx) (mh (mi mf)))) (as ((bm (ls av) di))) ts)

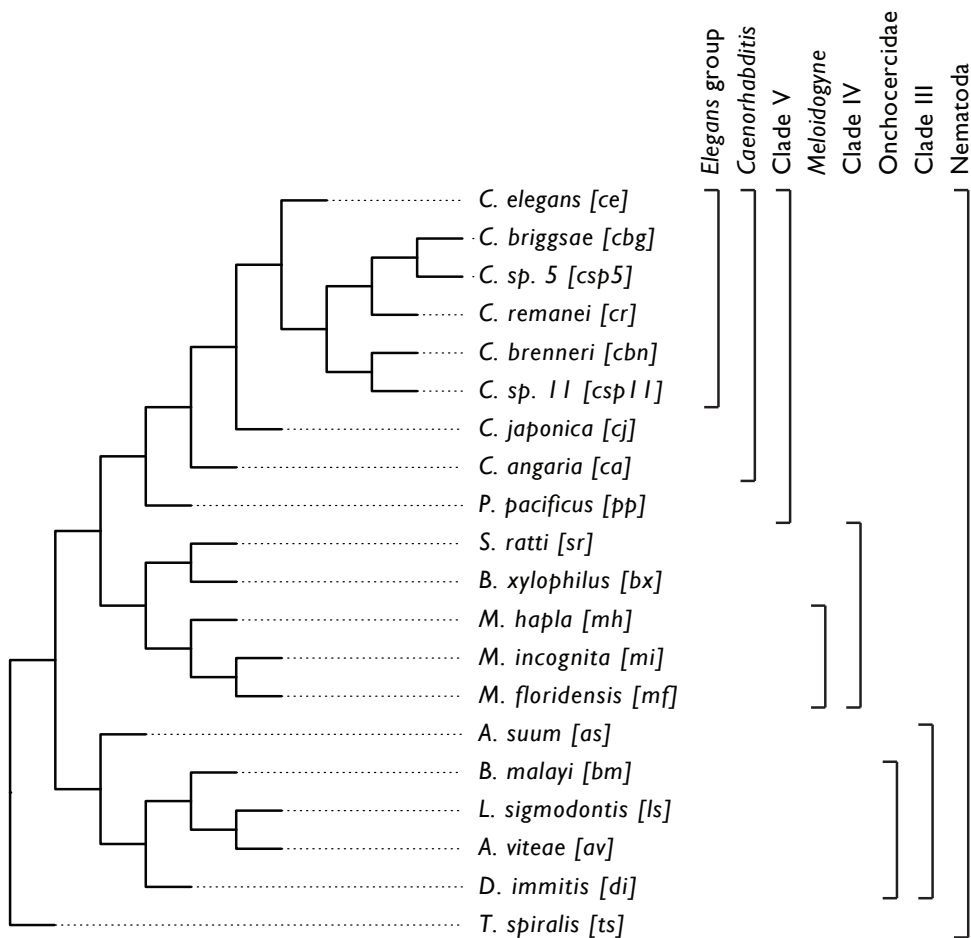


Figure 4.3 Cladogram depicting guide trees used in TBA+Multiz

A: TBA+Multiz output:

```
a score=100
s ls.scaf01598      990 39 + 8922 AAAGGCTT---GGAGATGATAACAACGGGCATAAACATCGAT
s bm.contig14747   1148 40 + 28583 aaaaGCTT--AAGAGATGATAACAACGGGCATAAACATTGAT
s av.scaf05297     1143 39 + 1825 AAAGGCTT---GGAGATGATAAAAACGGGCATAAACATCGAT
s di.scaf00076    183539 39 + 193403 ---AACTTAAAGATGATGATAACAACGGGCATAAACATCAAT
```

B: Lowercase masking of coding sequences in each species (**bold** letters indicate changes from previous step):

```
a score=100
s ls.scaf01598      990 39 + 8922 AAAGGCTT---GGAGATGATAACAACGGGCATAAACATCGAT
s bm.contig14747   1148 40 + 28583 AAAAGCTT--AAGAGatgataacaacgggcataaACATTGAT
s av.scaf05297     1143 39 + 1825 AAAGGCTT---GGAGATGATAAAAACGGGCATAAACATCGAT
s di.scaf00076    183539 39 + 193403 ---AACTTAAAGATGATGATAACAACGGGCATAAACATCAAT
```

C: Lower case masking of all alignment columns with coding (lowercase) bases:

```
a score=100
s ls.scaf01598      990 39 + 8922 AAAGGCTT---GGAGatgataacaacgggcataaACATCGAT
s bm.contig14747   1148 40 + 28583 AAAAGCTT--AAGAGatgataacaacgggcataaACATTGAT
s av.scaf05297     1143 39 + 1825 AAAGGCTT---GGAGatgataaacgggcataaaACATCGAT
s di.scaf00076    183539 39 + 193403 ---AACTTAAAGATatgataacaacgggcataaACATCAAT
```

D: Removing coding (lowercase) alignment columns to create split CNE blocks:

```
a score=35.71
s av.scaf05297     1143 12 + 1825 AAAGGCTT---GGAG
s bm.contig14747   1148 13 + 28583 AAAAGCTT--AAGAG
s di.scaf00076    183539 12 + 193403 ---AACTTAAAGATG
s ls.scaf01598      990 12 + 8922 AAAGGCTT---GGAG
```

```
a score=19.05
s av.scaf05297     1174 8 + 1825 ACATCGAT
s bm.contig14747   1180 8 + 28583 ACATTGAT
s di.scaf00076    18 3570 8 + 193403 ACATCAAT
s ls.scaf01598      1021 8 + 8922 ACATCGAT
```

Figure 4.4 Cartoon example of how coding regions were removed from multiple alignments

4.2.3 Identifying CNEs using MegaBLAST and clustering

We used a second way of identifying CNEs that corresponded closely to the method used by Vavouri *et al.* [191]. As in that study, MegaBLAST [106] was used to align genome sequences using a word seed size of 30 and an e-value cutoff of 0.001. Alignments that overlapped coding regions were removed. Unlike the previous study, we did not perform any subsequent filtering to remove tRNAs, repeats, or blast matches to the Rfam and miRNA databases, as our Coding GFFs (described above) already included tRNA, Rfam, and miRNA annotations.

To identify CNEs shared by all 3 species in their study, Vavouri *et al.* [191] aligned *C. briggsae* and *C. remanei* separately to the reference *C. elegans* genome and only selected those *C. elegans* regions that hit both the other genomes. We needed a more flexible way to identify shared CNEs because we wanted to analyse different branches of the nematode phylogeny and because we were also interested in CNEs that were possibly absent in the reference genome for that branch. Therefore, we clustered the blast results using simple single-linkage clustering (Appendix A: link_blast.pl). Our clustering script duplicated the functionality of BLASTCLUST [202], which also uses MegaBLAST and single-linkage clustering, but was more useful because BLASTCLUST does not run with more than ~10,000 sequences.

Clusters were filtered out if they did not include every species present in the phylogenetic branch being analysed. An additional "strict" non-coding filter was also tried where a cluster was removed from the analysis if any one of its members overlapped a coding region. This step exploits the fact that some species are better annotated than others and assumes that if one of those sequences is a coding sequence, then it is highly likely that the remaining aligned sequences are also coding sequences and should be removed from the set of CNEs. The final clusters file was converted into BED files with CNE coordinates for each genome using a command line perl script (Appendix A: Identify CNEs using MegaBLAST and clustering).

4.3 Results

4.3.1 No CNEs were shared across clades

We identified many thousands of CNEs within closely related groups of species by first creating multiple whole-genome alignments and then masking known coding and RNA-Sequences. 10,516 CNEs were found shared by all members of the *elegans* group—*C. elegans*, *C. brenneri*, *Caenorhabditis sp. 11*, *C. remanei*, *C. briggsae*, and *Caenorhabditis sp. 5*—with an alignment length cutoff of 30 and with 50% as the minimum relative identity (Figure 4.5). However, adding just two more species from the same genus (*C. japonica* and *C. angaria*)

reduced the number of CNEs to only 166. When *P. pacificus* was included, the total number of CNEs shared by all members of Clade V dropped to just 6, one of which matched a known snoRNA sequence. Taking Clades IV and V together, no CNEs were found shared by all members of both clades.

Similarly, in Clade IV, the three *Meloidogyne* species shared almost 60,000 CNEs, but all five species in Clade IV shared only 123 CNEs. The pattern was repeated in Clade III with the Onchocercidae sharing 28,923 CNEs, but with all five species in Clade III sharing only 249 CNEs. One possible explanation for the hyper-abundance of CNEs in the three *Meloidogyne* species is that the coding regions in these species have not been as well annotated, and therefore more regions are marked as non-coding. Although *M. floridensis* was not as rigorously annotated as *M. incognita* and *M. hapla*, our method of removing all alignment columns (Figure 4.4-C) identifies likely coding regions even in genomes that are not as comprehensively annotated. Therefore the lack of annotation is unlikely to be the sole reason for finding many *Meloidogyne* CNEs. A more likely explanation for the large number of *Meloidogyne* CNEs is that they all belong to the same genus and diverged more recently than the other groups. Alternatively, *M. incognita* may be an interspecific hybrid with *M. floridensis* as a possible parent (this hypothesis on *M. incognita* origins is explored in more detail in Chapter 5).

Although there are no accurate estimates of the time to last common ancestor (tLCA) for each group of species, the branch depths in Figure 4.5 can be treated as crude proxies for the tLCA. The number of CNEs at a node is negatively correlated with the mean branch depth for species at that node. The branches at the Clade IV-Clade V node are deep enough that no CNEs were found shared across clades, even though our method for finding CNEs was much more sensitive than the method used previously [191].

As expected, using the less sensitive MegaBLAST method [191] found even fewer CNEs in each category. Table 4.3 shows the number of CNEs found using MegaBLAST and clustering for the same nodes as in Figure 4.5. The "strict" workflow removed CNEs even if one sequence in the cluster overlapped a coding sequence. Removing putative coding sequences in this manner is analogous to removing coding columns in the whole-genome alignment method. However even in the non-strict case, far fewer CNEs were found compared to the more sensitive whole-genome alignment search. The first row in Table 4.3 includes an additional group *C. elegans*-*C. briggsae*-*C. remanei* that is not a branch in Figure 4.5, but has been listed so that this method can be compared with Vavouri *et al.*'s [191] analysis using these three species with the same MegaBLAST cutoffs. We found 2,972 CNEs shared between these species compared to the 2,084 found in the previous study. The ~900 extra CNEs identified using these settings were the result of less aggressive filtering of putative coding sequences. However, despite the presence of extra CNEs in our method, no CNEs were found shared across clades, confirming the absence of phylum-wide CNEs.

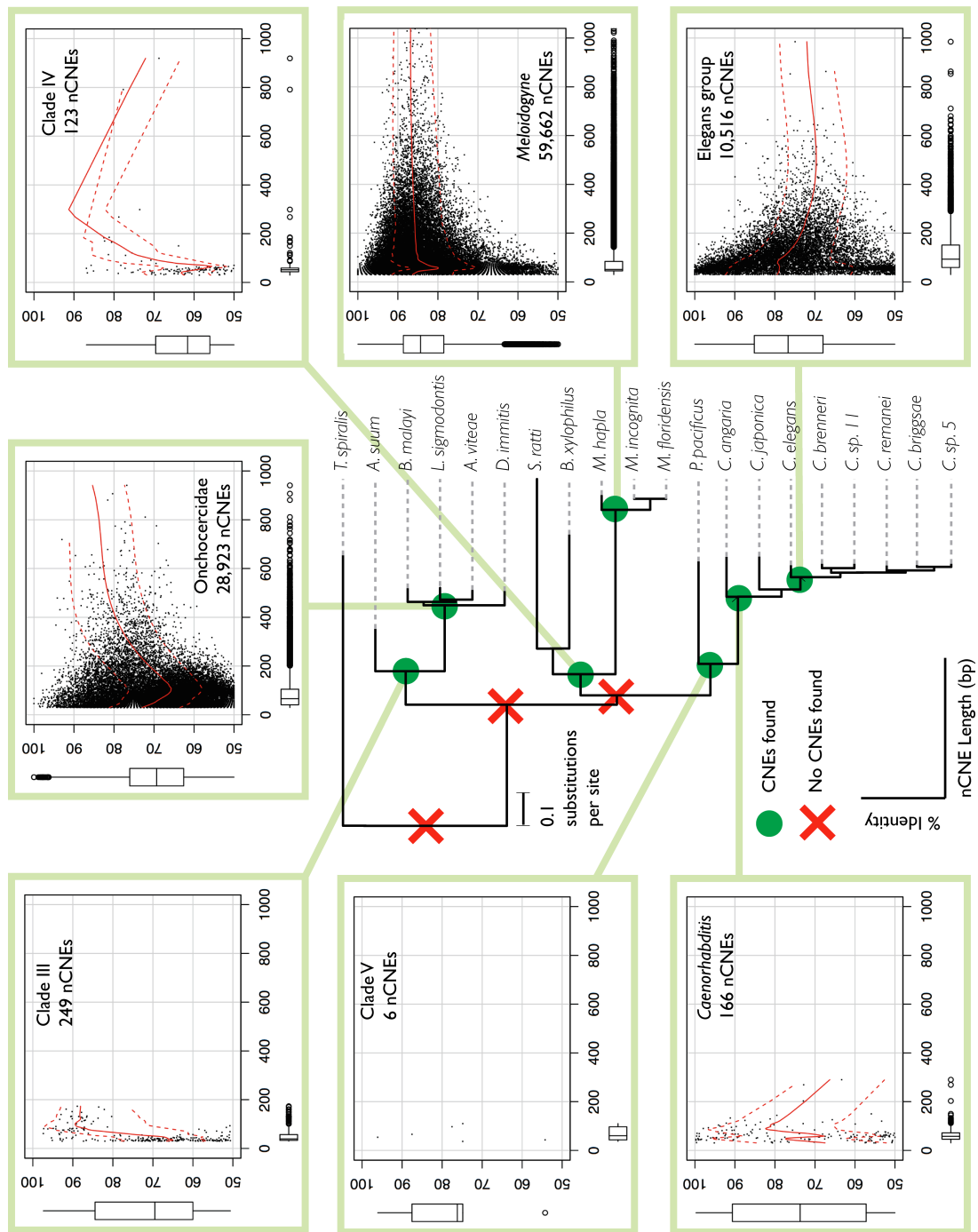


Figure 4.5 Length and identity of CNEs found at different nodes in nematode phylogeny

CNEs shared across all species at a given node in the phylogeny above were identified using whole-genome alignments of all species at that node. The length of each CNE is plotted against its relative identity across all species in the alignment (see Methods for how relative identity is defined). Green circles indicate nodes where CNEs were found and red crosses indicate that no CNEs were shared by all species at that node. Branch lengths were calculated by aligning 181 shared CDSs in all species and using RAXML [203] with the PROTGAMMAGTR model (courtesy G. Koutsovoulos, University of Edinburgh, pers. comm.)

Table 4.3 Comparing CNEs found using different methods

Nematode phylogeny branch	CNEs found using whole-genome multiple alignments	Length of CNEs bp Mean (std.dev)	Relative identity of CNEs % Mean (std.dev)	CNEs found using MegaBLAST + clustering	CNEs found using MegaBLAST + clustering (strict)
<i>C. elegans</i> - <i>C. briggsae</i> - <i>C. remanei</i>	27,174	87 (73)	70 (13)	2,993	2,972
<i>Elegans</i> group	10,516	118 (84)	76 (12)	1,477	1,459
<i>Caenorhabditis</i>	166	67 (36)	74 (17)	3	1
Clade V	6	68 (30)	78 (14)	2	0
<i>Meloidogyne</i>	59,662	86 (91)	83 (8)	2,572	2,261
Clade IV	123	74 (109)	64 (9)	0	0
Onchocercidae	28,923	87 (71)	70 (9)	243	235
Clade III	249	52 (31)	71 (13)	3	1

4.4 Discussion

4.4.1 An important negative result

Despite using a more sensitive method for finding CNEs than previous studies [187, 191, 197], we found no CNEs that were shared across all 20 nematode genomes. Additionally, we found no CNEs that were shared outside clade boundaries. Even within a clade, the number of CNEs dropped rapidly as older branch points in the nematode phylogenetic tree were analysed. When we used a MegaBLAST-based method that was more similar to previous studies and demanded higher sequence conservation as a pre-requisite for finding matches, we found even fewer CNEs within a clade: no CNEs were found shared by members of Clade III or Clade V, and only one CNE was shared by all species in Clade IV.

This finding reduces support for Vavouri and Lehner's hypothesis [192] that CNEs are the leftover traces of the process of GRN-rewiring that led to diverse animal body plans (Figure 4.1). However, the lack of CNEs shared across the nematode phylum or even across clades does not falsify the GRN-rewiring hypothesis. It is conceivable that regulatory regions such as promoters, enhancers, and TFBSs (i.e., the CNEs) have co-evolved along with the molecules that interact with these sites in such a way that the functionality of the interaction remains intact even though the underlying genome sequences (and interacting molecules) have changed. Rewired GRNs might still be responsible for the diversity in animal body plans but their traces in the form of CNEs are not present across the phylum, and are thus harder to detect.

Although no CNEs were found shared across clades, clade-specific CNEs were found for Clades III, IV, and V, which were shared by all species within a clade. Only a few hundred such elements existed at the level of the clade compared to tens of thousands of CNEs for more recently diverged species. The number of CNEs in each group seemed to be a function of tLCA. For the nodes in the phylogenetic tree where the tLCA for all species was low (e.g., family Onchocercidae and the *Elegans* group within genus *Caenorhabditis*), a large number of CNEs were found. For groups of species with a longer tLCA (e.g., Clade III, Clade IV, and Clade V), the number of CNEs dropped as the tLCA increased. *Meloidogyne* CNEs are interesting because they are present in large numbers compared to the other groups with a similar tLCA, and one possible reason could be that *M. incognita* is a hybrid species with *M. floridensis* as one parent, as discussed in the next chapter.

The absence of phylum-wide CNEs in nematodes was disappointing because we could not extend our analysis to understanding exactly how CNEs defined a phylum. However, this negative result is an important reminder that what is true for a couple of genomes in one genus is not necessarily true once more genomes are added to the analysis. This finding

underlines the need for more sequencing of non-model organisms to discover general patterns, rather than genus-specific ones.

The previous evidence for CNEs being shared and playing a role in the development of vertebrates is well documented and compelling. However, invertebrates, such as nematodes, do not have the same level of CNE conservation. On revisiting the insect-CNE study [197], we realised that the dramatic drop in CNE counts for more diverged species was not limited to nematodes. Glazov *et al.* had reported 21,301 ultra-conserved non-coding elements shared between two *Drosophila* species, but when *Anopheles gambiae* (from the same order, Diptera) was added to the analysis, only 2 such elements remained [197]. Therefore, it is possible that phylum-wide high-identity CNEs are limited to the vertebrates, and that they are the exception rather than the rule.

4.4.2 Next steps

Although no CNEs are shared across the phylum Nematoda, CNEs do exist at lower taxonomic levels and are evolutionarily constrained [191]. Vertebrate CNEs regulate development by acting as binding sites for transcription factors [204], although this mechanism has not yet been experimentally verified in nematodes or insects. Using comprehensive data from the modENCODE studies on *C. elegans* [45] and *D. melanogaster* [205], it should be possible to characterise the function of many of these "local" CNEs. For example, Cheng *et al* [206] were able to recreate gene regulatory networks in *C. elegans* using modENCODE chromatin immunoprecipitation sequencing (ChIP-Seq) and RNA-Seq data, and it would be informative to see which noncoding sequences participating in the regulatory network were highly conserved across all *Caenorhabditis* genomes.

The absence of phylum-wide CNEs in nematodes could indicate that regulatory sequences are less constrained in nematodes than in vertebrates. A more sensitive search for such elements might reveal more about how the different branches of the nematode phylum evolved. One of the limitations of the whole-genome alignment method for finding CNEs in this chapter is that only single-copy elements can be identified (a constraint of the whole-genome alignment protocol). The MegaBLAST-clustering method allows multiple-copy elements to be identified and clustered, but requires higher identity matches. Future research should focus on a combination of the two methods that can find more divergent multi-copy regulatory elements.

Future studies on the GRN-rewiring hypothesis could first identify GRNs in the species being compared and then look for similarities or differences at the level of GRNs, rather than at the level of sequence conservation alone. A pre-requisite for such a study would be high-quality gene predictions for each species, as well as extensive gene expression data and other genomic data such as ChIP-Seq, to aid the reconstruction of GRNs. Armed with these tools,

we can better approach the question of whether the incredible diversity of animal body plans is a result of rewiring GRNs.

5 The *Meloidogyne floridensis* genome reveals complex hybrid origins of the root-knot nematodes

Meloidogyne root-knot nematodes (RKNs) can infect most agricultural plant species and are among the most important of all plant pathogens. This chapter presents a study testing the role of *M. floridensis* in the hybrid origins of the highly destructive *M. incognita*. By sequencing the genome of *M. floridensis* and comparing the coding sequences of these two species, we elucidated the hybrid origins of not one, but both species.

The manuscript form of this chapter is being prepared for submission. I assembled the *M. floridensis* genome (as described in Chapter 2), performed all the data analysis, and wrote and edited several parts of the manuscript. Apart from the workflows developed by Georgios Koutsovoulos (University of Edinburgh) to generate phylogenetic trees, all the programs and scripts used in this study were written by me and are listed in Appendix A. The study was conceived by David Lunt (University of Hull) and Mark Blaxter (University of Edinburgh) and they were the lead authors of this manuscript.

5.1 Introduction

The tropical RKNs of the genus *Meloidogyne* are globally important crop pathogens. The most damaging species within this group are the apomicts *M. incognita*, *M. arenaria* and *M. javanica*. They are highly polyphagous, with the ability to infect most crop species, including all those producing the majority of the world's food supply. The damage attributable to RKNs is ~5% of world agriculture [20, 207, 208].

The tropical apomict RKNs possess aneuploid diploid or hypotriploid genomes and reproduce by obligatory mitotic parthenogenesis (Figure 5.1). They have previously been suggested to be hybrid taxa, and phylogenetic analysis of nuclear loci supports this conclusion [122, 209-212]. Hybrid speciation is thought to be relatively uncommon in animals compared to plants [213, 214], and often cannot be well investigated by standard molecular approaches.

Meloidogyne floridensis is a plant pathogenic root-knot nematode that was originally characterised as *M. incognita*, but has since been re-investigated and described as a separate species on the basis of morphology and a unique *esterase* isozyme pattern [215, 216]. Despite the fact that both nuclear ribosomal DNA (rDNA) and mitochondrial DNA (mtDNA) sequences place it within the phylogenetic diversity of the tropical apomict species [217, 218], *M. floridensis* is a diploid that reproduces through meiotic parthenogenesis (automixis). With the exception of *M. floridensis*, all of the "Group 1" RKNs [42, 218] are apomicts, unable

to reproduce by meiosis, lacking bivalent chromosomes, and exhibiting extensive aneuploidy. This phylogenetic distribution of reproductive modes (*M. floridensis* phylogenetically nested within the diversity of the apomict RKNs) is unanticipated as it implies the physiologically unlikely route of re-emergence of meiosis from within the obligate mitotic parthenogens. An alternative explanation for these observations is that the observed phylogenetic relationships have not arisen from a typical ancestor-descendent bifurcating process, but instead have been shaped by reticulate evolution and transfer of genes by interspecific hybridisation.

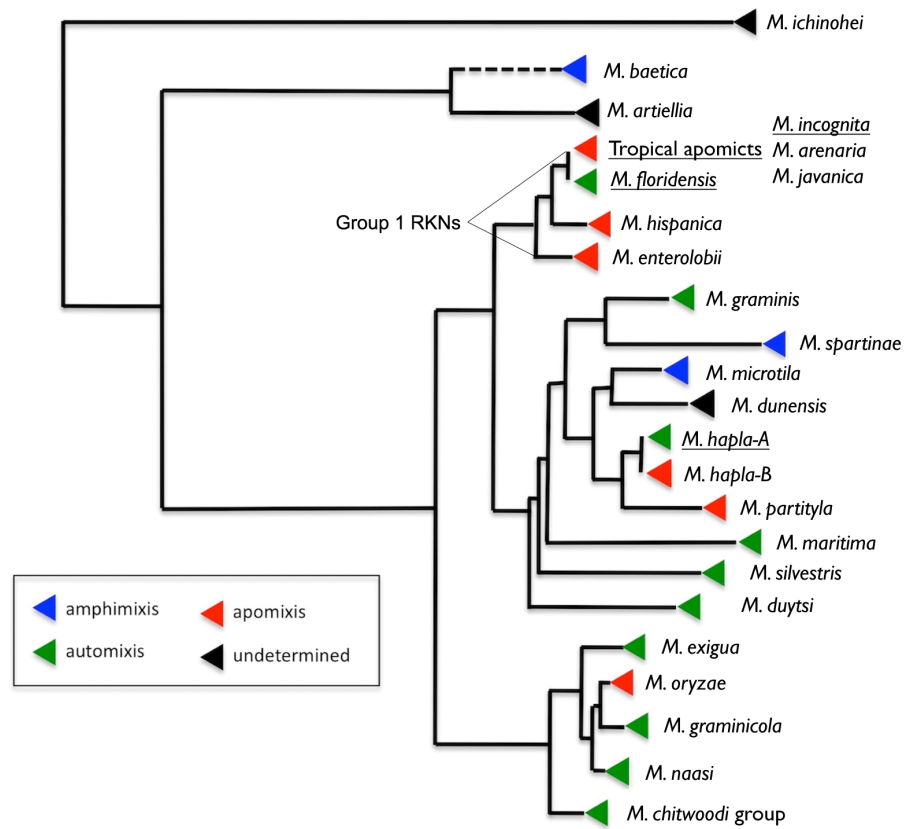


Figure 5.1 *Meloidogyne* phylogeny indicating positions of *M. floridensis* and tropical apomicts

Amphimixis: sexual reproduction between males and females involving gametes produced by meiosis

Apomixis: mitotic parthenogenesis, no meiosis or outbreeding can occur

Automixis: meiotic parthenogenesis predominates although females also occasionally outbreed with males.

Underlines indicate genomes that have been sequenced.

Figure and text adapted from Lunt and Blaxter (pers. comm.)

5.1.1 Hypotheses for origins of *M. incognita* genomic duplicates

The *M. incognita* genome [38] revealed that many of the genes of this species are present as two (or more) divergent copies. The origin of these divergent copies is controversial. The nuclear gene phylogenies of Lunt [122] indicate that the parental taxa of the apomict RKNs were closely related and derived from within the cluster of Group 1 *Meloidogyne* species after the divergence of *M. enterolobii* (= *M. mayaguensis*). Since this matches the phylogenetic position of *M. floridensis*, and this species is known to reproduce via sexual recombination, as the parental species also must have done, we set out to test by comparative genome sequencing and analysis if *M. floridensis* was one of the progenitors of the tropical apomicts. The genomes of *M. incognita* [38] and the outgroup *M. hapla* [35] have already been sequenced, and *M. floridensis* was sequenced during the course of this thesis (underlined species in Figure 5.1 indicate sequenced genomes). Using these three species, four hypotheses were generated to account for the origins of genomic duplicates in *M. incognita*; each makes specific predictions regarding the number of copies of genes that we expect to find while comparing these species.

Hypothesis A: No hybridisation

Genomic duplicates in *M. incognita* may have originated by a process of 'endoduplication' (Figure 5.2-A). Endoduplication can refer to two distinct processes, although their genomic outcomes are similar. In a first process, the entire *M. incognita* genome might have doubled to become tetraploid. The homeologous copies (paralogous copies that arise from polyploidy rather than individual gene duplication) could then diverge, and the extant pattern of partial retention of duplicated loci could be the result of gene loss as diploidy was regained. This process would leave many areas of the newly diploidised genome possessing divergent copies. An alternative process could be that in apomictic species such as *M. incognita*, former alleles, now independent from the homogenising effects of recombination, can independently accumulate mutations over long periods of time [219] resulting in highly divergent homologous loci (alleles) within a diploid genome [38].

Under this hypothesis, *M. floridensis* is a sister to *M. incognita*, and only one copy of the "X" genome in Figure 5.2-A would be expected in the *M. floridensis* assembly. *M. incognita* has undergone whole-genome endoduplication and the duplicated genes ("Z+Z") in *M. incognita* are diverging under Muller's ratchet.

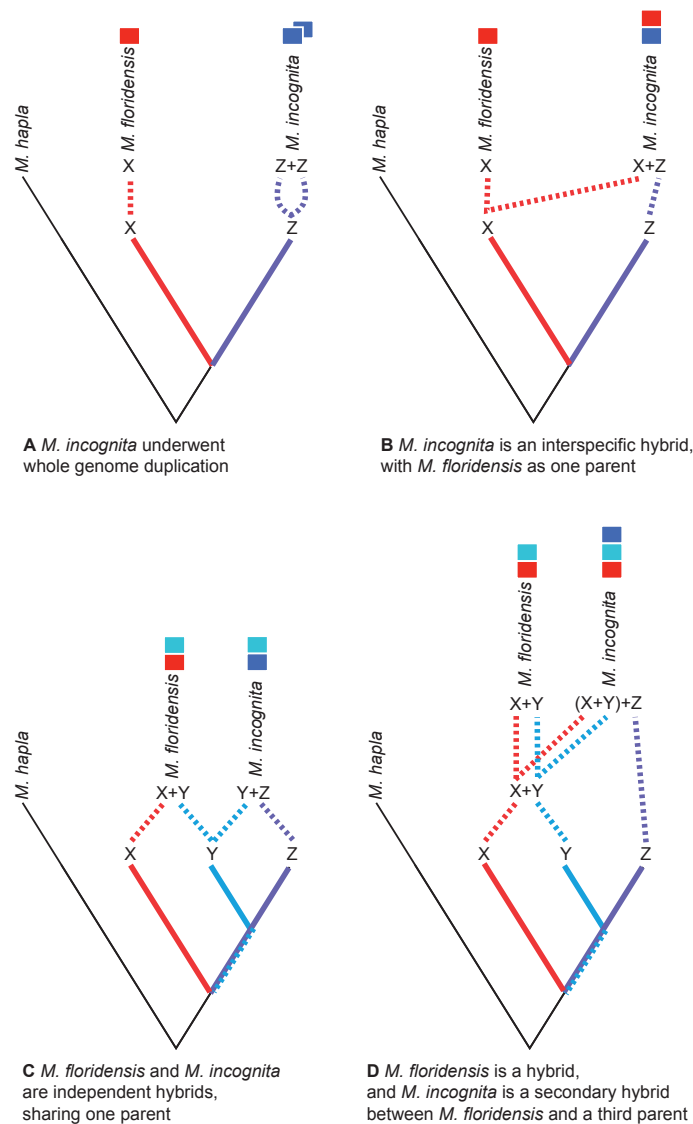


Figure 5.2 Hypotheses for relationships between *M. floridensis*, *M. incognita*, and *M. hapla*, and the origins of duplicated gene copies

M. hapla is a diploid species distantly related to *M. incognita* and *M. floridensis*. Species "X", "Y" and "Z" are postulated ancestral parents that could have given rise to *M. incognita* and *M. floridensis*.

A. No hybridisation: *M. floridensis* is a sister to *M. incognita*, and has one copy of the "X" genome. *M. incognita* has undergone whole-genome endoduplication. The duplicated genes ("Z+Z") in *M. incognita* are diverging under Muller's ratchet.

B. *M. floridensis* is a diploid parent of the hybrid *M. incognita*: Ancestor "X" gave rise to the diploid species *M. floridensis*, and also interbred with "Z" to yield *M. incognita*, which thus carries two copies of each gene ("X+Z"). Only *M. incognita* is predicted to carry two homeologues of many genes.

C. *M. incognita* and *M. floridensis* are hybrid siblings sharing a common parent: Both *M. incognita* ("Y+Z") and *M. floridensis* ("X+Y") are hybrid species, and share one parent ("Y"). Both *M. incognita* and *M. floridensis* are predicted to carry two homeologues of many genes.

D. *M. floridensis* is a hybrid parent of the hybrid *M. incognita*: Both *M. floridensis* ("Y+Z") and *M. incognita* ("X+Y+Z") are hybrid species, but *M. incognita* is a triploid hybrid between "X+Y" (the hybrid *M. floridensis* ancestor) and "Z". *M. incognita* is predicted to carry three, and *M. floridensis* is predicted to carry two, homeologues of many genes.

Hypothesis B: *M. floridensis* is a diploid parent of the hybrid *M. incognita*

Hypothesis B restricts the hybrid taxa to the apomict Group 1 species, and places *M. floridensis* as one of the hybridising parental species (Figure 5.2-B). This model predicts that, where divergent homeologous sequences are detected in the *M. incognita* genome, *M. floridensis* would possess one homeologue. The *M. floridensis* genome itself would be substantially different from that of *M. incognita*, not possessing divergent homeologous blocks but rather displaying normal allelic variation, perhaps more similar to that of *M. hapla*.

In Figure 5.2-B, if ancestor "X" gave rise to the diploid species *M. floridensis*, and also interbred with "Z" to yield *M. incognita*, we would expect to find two copies of each gene ("X+Z") in *M. incognita*, but only one copy of each gene in *M. floridensis*.

Hypothesis C: *M. incognita* and *M. floridensis* are hybrids sharing a common parent

Alternatively *M. floridensis* might be an independent hybrid that shares one parental taxon with *M. incognita*, and thus represents a 'sibling' taxon (Figure 5.2-C). Both *M. incognita* ("Y+Z") and *M. floridensis* ("X+Y") are hybrid species and share one parent ("Y"). Under this hypothesis both *M. incognita* and *M. floridensis* are predicted to carry two homeologues of many genes.

Hypothesis D: *M. incognita* and *M. floridensis* are hybrids. *M. floridensis* is a parent

Finally, *M. floridensis* may itself be a hybrid, but still have played a role as a parent of *M. incognita* by a subsequent hybridisation event. In Figure 5.2-D, both *M. floridensis* ("Y+Z") and *M. incognita* ("X+Y+Z") are hybrid species, but *M. incognita* is a triploid hybrid between "X+Y" (the hybrid *M. floridensis* ancestor) and "Z". Therefore, *M. incognita* is predicted to carry three, and *M. floridensis* is predicted to carry two, homeologues of many genes.

Summary

The different possible histories of hybridisation, as well as endoduplication, may be distinguished by the collection of sufficient homologous loci across the putatively hybrid and hybridising species and a robustly diploid outgroup. Each hypothesis predicts a different evolutionary relationship between gene copies, and thus gene-by-gene phylogenetic analyses should discriminate between the models. The *de novo* assembled genome of *M. floridensis* was used to identify and analyse a large number of sets of homologous sequences in *M. floridensis*, *M. incognita*, and the more distantly related automict *M. hapla*. We use both gene copy number distributions and gene phylogenies to test these different scenarios.

In both hypotheses A and B, *M. floridensis* should exhibit only one copy of each gene. Divergent copies of *M. incognita* genes should be more similar to each other than to *M.*

floridensis genes under hypothesis A, whereas under hypothesis B we expect *M. floridensis* genes to cluster with one copy of *M. incognita* genes more closely than the other, indicating that *M. incognita* is the product of an interspecific hybridisation.

In hypotheses C and D, *M. floridensis* would also be a product of an interspecific hybridisation, just as were the apomicts, but that this hybrid species remained diploid. In these scenarios the *M. floridensis* genome will, like *M. incognita*, show substantial sequence divergence between homeologues. It may also possess regions where one parental copy has been eliminated, and the remaining diversity is simple allelism. In hypothesis C, the parents of *M. floridensis* need not be the same as those of the apomicts, although the phylogenetic position of *M. floridensis* implies that at least one of them may have been identical or very closely related. The different putative hybrid origins of *M. incognita* predict two (hypothesis C, Figure 5.2-C) or three (hypothesis D, Figure 5.2-D) homeologous copies in *M. incognita*, modified by loss events.

5.2 Methods

5.2.1 Nematode materials

DNA from female egg masses of *M. floridensis* isolate 5 was generously sourced and provided from culture by Dr Tom Powers (University of Lincoln, Nebraska, USA) and Dr Janete Brito (Florida Department of Agriculture and Consumer Services, Gainesville, USA).

5.2.2 Protein predictions and comparisons

M. floridensis was sequenced and the draft genome assembled as described in Chapter 2. The analyses described in this chapter were carried out before our annotation pipeline had been fully developed. However, because we were interested in comparing only coding sequences conserved with *M. hapla* and *M. incognita*, protein alignment methods were used to extract sequences of interest, rather than carry out a full protein prediction and annotation effort. The protein2genome model in Exonerate version 2.2.0 [164] was used to align all *M. hapla* and *M. incognita* proteins to the *M. floridensis* draft genome (Appendix A: Extract protein coding CDSs from *M. floridensis* using exonerate). Coding sequences (CDSs) were extracted from the *M. floridensis* genome that aligned to at least 50% of the length of the query protein sequences. If multiple *M. hapla* or *M. incognita* query protein sequences aligned to overlapping loci on the *M. floridensis* genome, only the longest locus was chosen as a putative *M. floridensis* CDS. The CDSs for all three species were trimmed after the first stop codon, and only sequences with a minimum of 50 amino acids were retained for further analysis.

To assess the level of self-identity among CDSs in each species, Blastn version 2.2.25+ [220] was used and the top scoring hit (e-value $1e-5$) for each sequence to a CDS other than itself was selected if the length of the alignment was longer than 70% of the query sequence.

5.2.3 Clustering

Inparanoid version 4.1 [221] and QuickParanoid <http://pl.postech.ac.kr/QuickParanoid> [222] were used with default settings to assign proteins from the three *Meloidogyne* species to orthology groups. While assessing the level of duplication within the CDS sets (Figure 5.3), several *M. incognita* CDS sequences were observed to be identical or nearly identical (>98% identity). These are most likely derived from allelic variants rather than gene duplications (which show a separate peak between 95 and 97% identity). To simplify the construction of orthologous gene clusters, these near identical sequences in each species were reduced using CD-HIT-EST [114], removing any CDSs that were at least 98% identical across their whole length to another CDS.

5.2.4 Phylogenetic analyses

For each InParanoid cluster, Clustal Omega version 1.0.3 [223] was first used to align the protein sequences. Tralign (from the Emboss suite version 6.2.0 [224]) was then used along with the protein alignment as a guide to align the nucleotide CDS sequences. Finally, RAxML version 7.2.8 [203] was used to create maximum likelihood trees for each set of aligned CDS sequences in three steps: (i) finding the best ML tree by running the GTRGAMMA model for 10 runs; (ii) getting the bootstrap support values for this tree by running the same model until the autoMRE convergence criterion was satisfied; (iii) using the bootstrap trees to draw bipartitions on the best ML tree. The resulting trees were imported into the R Ape package version 2.8 [225] to count the number of trees with the same topology (Appendix A: Use RAxML and Ape to create and analyse multiple phylogenies).

5.3 Results

5.3.1 The genome of *M. floridensis*

M. floridensis is known to be diploid [215], and we assumed that the isolate sequenced here was diploid. The *M. floridensis* genome was assembled using 11.1 gigabases of cleaned data from 116 M reads (an estimated ~100X coverage), using Illumina HiSeq2000 100 b paired-end sequencing of 250 bp fragments. The genome version used here is an earlier version (nMf.1.0 with 100 bp as the minimum contig size) than the one reported in Chapters 2 and 3 (nMf.1.1 with 200 bp minimum contig size).

This assembly is ~100 Mbp (Table 5.1), larger than either of the other two *Meloidogyne* species published thus far. However the 86 Mbp *M. incognita* assembly [38] may be incomplete, and *M. incognita* may have a significantly larger genome (~140 Mbp) than currently published (Etienne Danchin, Institut national de la recherche agronomique, pers. comm.). Genome sizes derived from whole-genome shotgun assemblies should be interpreted cautiously because recent segmental duplications and repeat families with high identity are largely unresolvable using short reads and small insert sizes, and would be collapsed by assembly algorithms. The *M. floridensis* genome assembly is less contiguous than those of *M. hapla* and *M. incognita* (reflected in the lower N50 values). Such fragmentation is a known limitation of using a single small-insert paired-end library, but despite this lack of contiguity, the assembly yielded over 15,000 protein sequences that were more than adequate for the purpose of this study. We note that both the *M. incognita* and the *M. floridensis* genomes have low scores (60-75%) when assessed using CEGMA [94], compared to the 94% scored by the *M. hapla* assembly (and assemblies of other nematode genomes). It is not clear whether this is an artefact of assembly incompleteness and/or a biological feature of these genomes.

Table 5.1 Summary statistics describing assemblies and protein predictions in *Meloidogyne* genomes

Species	<i>M. hapla</i>	<i>M. incognita</i>	<i>M. floridensis</i>
Genome version	WormBase WS227 [35]	INRA scaffolds [38]	nMf.1.0
Maximum scaffold length	360,446	447,151	40,762
Number of scaffolds	3,452	2,995	81,111
Assembled size (bp)	53,017,507	86,061,872	99,886,934
Scaffold N50 (bp)	37,608	62,516	3,516
GC%	27.4	31.4	29.7
CEGMA completeness Full / Partial	92.74 / 94.35	75.00 / 77.82	60.08 / 72.18
Predicted proteins	13,072	20,359	15,327
Predicted proteins used for clustering and inferring phylogenies (after filtering for length >50 aa, see Methods)	12,229	17,999	15,121

5.3.2 Intra-genomic comparisons reveal high numbers of duplicate genes in *M. incognita* and *M. floridensis*

Analysis of the distribution of within-genome coding sequence (CDS) matches (Figure 5.3) identified an unexpected excess of apparent duplication in *M. floridensis*. While the CDS set of *M. hapla* had a relatively low rate of duplication and no excess of duplicates of any particular divergence, both *M. incognita* and *M. floridensis* had many more duplicates and a peak of divergence between duplicates at 95 to 97% identity. *M. incognita* showed an additional peak at ~100% identity most likely due to a failure to collapse allelic copies of some genes. Because of the way we constructed our draft genome assembly, collapsing high-identity assembly fragments before analysis, *M. floridensis* lacked a similar near-complete identity peak. These data strongly suggest that *M. floridensis*, like *M. incognita*, may be a hybrid species, with contributions from two distinct parental genomes.

5.3.3 Distinguishing sibling from parent-child species relationships

Several hypotheses that might explain the observed levels of within-genome divergent duplicates in *M. incognita* and *M. floridensis* were identified (Figure 5.2). Expectations of relative numbers of (homeologous) gene copies per species, and the phylogenetic relationships of these homeologue sets differ between the hypotheses. The CDSs of the three species were clustered using InParanoid, after removing all CDS encoding peptides less than 50 amino acids in length. We defined 11,587 clusters that contained CDSs from more than one species, and 4018 with representatives from all three species (Figure 5.4), a number and proportion congruent with other comparisons between nematode species with complete genomes (e.g., Mitreva *et al.* identified 2501 clusters containing representatives from four complete nematode genomes [49]). Clusters that had a single *M. hapla* member were identified and classified by the numbers of *M. incognita* and *M. floridensis* genes they contained (Table 5.2). This subset of clusters should contain a significant proportion of the homologue sets where the ancestral gene was single-copy.

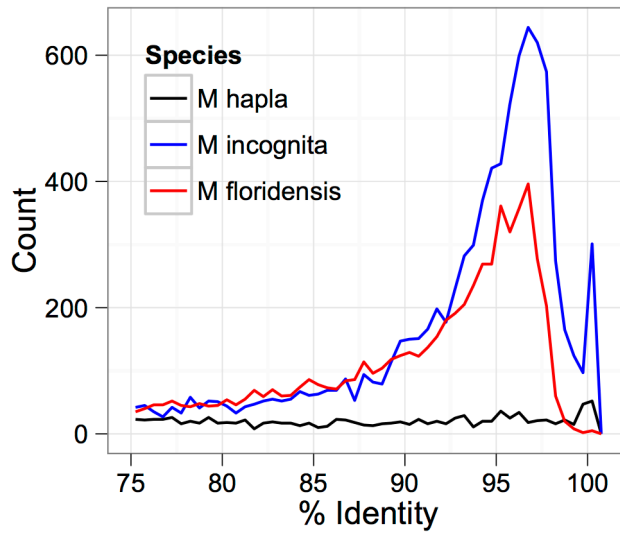


Figure 5.3 Intra-genomic duplication of protein-coding sequences
 Each coding sequence from each of the three target genomes (*M. hapla*, *M. incognita* and *M. floridensis*) was compared to the set of genes from the same species. The percentage identity of the best matching (non-self) coding sequence was calculated, and is plotted as a frequency histogram. Both *M. incognita* and *M. floridensis* show evidence of presence of many duplicates, while *M. hapla* does not.

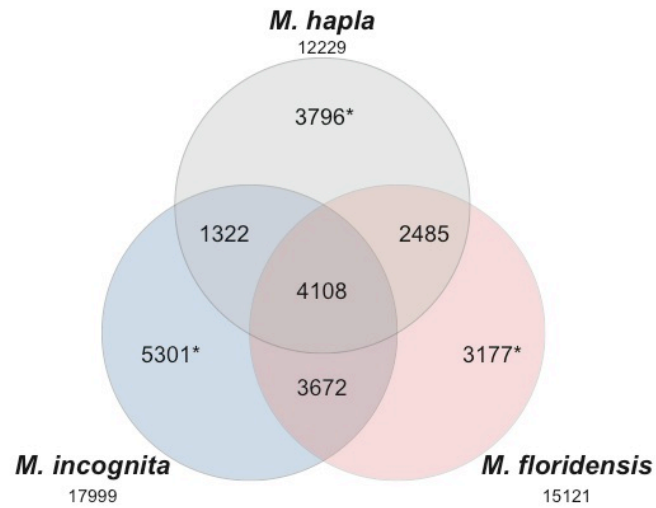


Figure 5.4 Venn diagram of clustering of proteins from three *Meloidogyne* species
 The complete proteomes of the three *Meloidogyne* species were clustered using InParanoid+QuickParanoid. This Venn diagram shows the numbers of clusters that had multiple species membership, and the numbers of proteins that were unique to each species (numbers marked with *). The total number of proteins input for each species are given under the species' name.

Table 5.2 Numbers of *M. floridensis* and *M. incognita* members in homeologue sets with one *M. hapla* member

	0 <i>M. incognita</i> members	1 <i>M. incognita</i> member	2 <i>M. incognita</i> members	3 <i>M. incognita</i> members	>3 <i>M. incognita</i> members
0 <i>M. floridensis</i> members	0	907	327	44	17
1 <i>M. floridensis</i> member	2196	2189	920	102	40
2 <i>M. floridensis</i> members	226	257	156	36	21
3 <i>M. floridensis</i> members	17	17	20	7	14
>3 <i>M. floridensis</i> members	8	11	6	4	21

The process of idiosyncratic gene loss (or failure to capture a gene in the draft sequencing and assembly) is evident in the numbers of genes that have one *M. hapla* representative and no members from either *M. incognita* (column 1 of Table 5.2) or *M. floridensis* (row 1 of Table 5.2). Here it is striking that the clusters that contain only one *M. hapla* and one *M. floridensis* member (Mh1:Mf1:Mi0) outnumber clusters that have one *M. hapla* and one *M. incognita* member (Mh1:Mf0:Mi1) by approximately two to one. This suggests that the *M. floridensis* genome draft is a good substrate for these analyses (it contains homologues of many conserved genes apparently lost from, or missing in the draft assembly of, the *M. incognita* genome), and that the *M. incognita* draft is either incomplete or has experienced greater rates of gene loss. The numbers of genes present in clusters that have more than two members, but lack one of *M. floridensis* or *M. incognita* (for example the 226 Mh1:Mf2:Mi0 clusters) reveal the likely extent of within-lineage duplication and divergence (and a component of stochastic loss of several homeologues in the missing species). There is no particular excess of these classes of cluster in *M. incognita*, arguing against a within-lineage, whole-genome duplication (i.e. against hypothesis A, Figure 5.2-A).

Despite prediction failures and idiosyncratic loss in the previously sequenced *M. incognita* genome, the striking feature of the enumeration of membership of clusters (Table 5.2) is the excess of clusters where *M. incognita* contributed more members than did *M. floridensis*. Thus there are 920 clusters in the class Mh1:Mf1:Mi2, but only 257 in the class Mh1:Mf2:Mi1, and 102 clusters in the class Mh1:Mf1:Mi3 compared to 17 in the class Mh1:Mf3:Mi1. This finding argues for the ancestral presence in *M. incognita* of at least one more genome copy than in *M. floridensis*, i.e. that *M. incognita* is likely to be a degenerate triploid hybrid (hypothesis D, Figure 5.2-D). It is possible that some of the clusters in the Mh1:Mf1:Mi0 and Mh1:Mf0:Mi1 sets arose from *M. floridensis* and *M. incognita* being derived from different, divergent parents.

5.3.4 Phylogenetic analysis of homologue relationships

A second set of predictions from the hypotheses in Figure 5.2 concerns the phylogenetic relationships of the resulting sets of homologous clusters. Each hypothesis predicts a particular set of relationships between gene copies in each species. It is possible, by adding postulated duplications and losses, to develop complex narratives that make all phylogenetic topologies congruent with every hypotheses. However, we used the frequencies of observation of each topology to identify which hypothesis was best supported. For example, the topology (Mh,((Mi,Mf),(Mi,Mf))) is congruent with hypothesis B (diploid hybrid), but requires the assumption of a duplication event, whereas the same topology is congruent with hypothesis D without any additional assumptions. Following Occam's razor, we are looking for the hypothesis that requires the least complex set of duplications and losses to be congruent with all gene sets.

Figure 5.5 shows clusters from Table 5.2 that provide informative topologies. Clusters with only one copy from each species were ignored, as were clusters with more than five total members because the number of possible topologies exponentially expanded and topology frequencies became negligible. For each informative set of clusters, the majority topology supported hypothesis D (hybrid parent, Figure 5.2-D), i.e., that *M. floridensis* is a hybrid, and was one parent species of a hybridisation event that gave rise to the triploid *M. incognita*. Thus for the 920 Mh1:Mf1:Mi2 clusters, the alternate topology in which one *M. incognita* member is closer to the *M. floridensis* member than it was to the other *M. incognita* member was favoured in 78% of the clusters supporting a hybrid rather than a duplicate scenario, while in 201 clusters (22%), it instead appeared to have arisen by duplication within *M. incognita*. In the Mh1:Mf2:Mi2 cluster set, one third of the clusters supported the topology where there were two independent sister relationships between *M. incognita* and *M. floridensis* genes supporting hybridisation. A further 48% (48 plus 29 out of 156) of the trees were congruent with a triploid status for *M. incognita* where gene loss (or lack of prediction) had removed one *M. incognita* representative. Similar detailed examinations of other gene sets gave similar outcomes in the hybrid status support (Figure 5.2-D).

5.4 Discussion

The genome structure and content of tropical *Meloidogyne* is revealed by our analyses to have had complex origins. It is likely that hybridisation, ploidy change, and partial returns to diploidy have all played a role in the evolution of diversity in this genus. The molecular evolutionary patterns revealed by comparative genomics however give us tools to conduct detailed analysis of these histories. This approach will allow us to understand the evolution of these polyphagous pathogens.

5.4.1 The *M. floridensis* genome reveals hybrid origins

Our first draft assembly of the genome of *M. floridensis* revealed a relatively typical nematode genome. While an independent estimate of genome size for this species is not available, it is likely, given the discussion of hybrid origins below, that it will be between one and two times that of *M. hapla* (i.e., between 53 Mbp and 106 Mbp). Our assembly, at 100 Mbp, was thus towards the higher end of the estimated range. The assembly was fragmented (in 81,111 scaffolds with an N50 of 3.5 kbp) but was sufficient for the experimental goals of this study. Refinement of the assembly, using larger-insert mate pair, or long single molecule reads, would undoubtedly improve the biological completeness of the product. Our assembly of *M. floridensis* scores poorly in terms of content of core, conserved eukaryotic genes. However, the published *M. incognita* genome, while having much better assembly statistics (only 2,995 scaffolds, and an N50 that is ~15 times the length achieved for *M. floridensis*), has similar poor scores in CEGMA analysis. Whether this is a reflection of shared, divergent biology, or, as we suspect, poor, fragmented assembly, will require additional sequencing data, re-assembly and re-assessment. A detailed biological analysis of the *M. floridensis* genome will be published in a separate report.

The phylogenetic position of *M. floridensis* and its automictic reproductive mode suggested that it was possible that this species, or an immediate ancestor, was parental to the tropical apomicts, i.e., it was one partner in the hybrid origins of the group (hypothesis B and D described previously). The other hypothesis relevant here is that *M. floridensis* is not directly parental to the apomicts, but rather a hybrid sibling, also created by hybridisation (hypothesis C above, with the additional restriction that one parent is very likely to be shared between the putative hybrid species, as some loci were found to be nearly identical between *M. incognita* and *M. floridensis* [122]). In order to distinguish between hypothesis B (diploid parent), hypothesis C (hybrid sibling) and hypothesis D (hybrid parent), we examined the sequence diversity within each species' genome and the phylogenetic trees for each cluster of homologues between the three species.

Intra-genomic divergence of coding loci

Information concerning the hybrid status of *M. floridensis* can be gained from comparing the pattern of gene duplication within its genome to that of other RKN species, since *M. incognita* has been suggested previously to have hybrid origins whereas *M. hapla* never has [122, 209-212]. A hybrid would be expected to have an excess of divergent duplicates compared to a non-hybrid. The genome of *M. hapla* allowed us to examine the intra-genomic duplication pattern of a closely related taxon not suggested to be of hybrid origin. In this case there was a relatively low number of divergent duplicates, and these had a wide range of divergences. While there was a slight excess of duplicates with high identity, the distribution overall is consistent with an ongoing, rare process of stochastic duplication followed by gradual divergence (Figure 5.3).

In contrast, the intra-genomic sequence comparisons of both *M. incognita* and *M. floridensis* displayed many more divergent duplicated CDS than found in *M. hapla* (Figure 5.3). While there was a peak of high-identity duplicates in *M. incognita*, this was absent in *M. floridensis*, likely because we collapsed high identity segments (as putative allelic copies) during assembly. Most striking was the presence in both species of a peak frequency of diverged duplicates at ~96% identity. Diverged duplicates were described previously [38, 122]. Ongoing individual gene duplication events—which we propose has generated the *M. hapla* distribution—could not have produced these patterns. Instead, the distributions are congruent with a single, major historical event of gene duplication followed by divergence, where variation in the rates of evolution of individual loci has resulted in variation in observed identity in the extant genomes. The mass duplication event could be a whole-genome endoduplication event or a hybridisation event that brought together homeologous loci that had been evolving independently since their last common ancestor. While these two alternative scenarios cannot be distinguished on the basis of duplicate divergence data alone, this analysis does suggest that the genomes of both *M. floridensis* and *M. incognita* have been shaped by major duplication events.

Integrating phylogenomic analyses

To distinguish between endoduplication and hybridisation origins of these CDS divergence patterns, the phylogenetic histories of sets of homologues from the three *Meloidogyne* genomes were examined. Choosing clusters that had only one *M. hapla* member, and were thus more likely to have been single copy in the last common ancestor of the three species, we compared support on a gene-by-gene basis for tree topologies that would support or refute the hybrid versus endoduplication scenarios (Figure 5.2, Table 5.2, and Figure 5.5).

Hypothesis A (no hybridisation; endoduplication of the *M. incognita* genome) could be robustly excluded as a source of duplicate CDSs as we observed that *M. floridensis* CDSs are frequently more closely related to one of the *M. incognita* CDSs than they are to each other. If *M. incognita* had duplicated its own genome, these duplicate CDSs would be expected to

have a monophyletic relationship with each other. Hypothesis B was similarly excluded because intra-genomic comparisons of CDSs in the *M. floridensis* genome revealed that it also possesses divergent duplicates, and phylogenetic analyses indicated that these, just like the *M. incognita* sequences, are not monophyletic by species.

The most parsimonious explanation of the duplicate divergence and phylogenetic data is that both *M. floridensis* and *M. incognita* are hybrid species, and the duplicate CDSs are homeologues rather than within-species paralogues. We could distinguish between hypotheses C (hybrid siblings: the two species are step-sisters) and D (*M. floridensis* represents one of the parents of a triploid hybrid *M. incognita*) using phylogenetic analyses of clustered CDSs. For clusters containing two *M. floridensis* homologues and two *M. incognita* homologues, the topology supporting shared hybrid ancestry was more frequently recovered than topologies supporting independent hybridisation events. In addition, and critically, there was an excess of clusters where there were more *M. incognita* members than there were *M. floridensis* members, as would be expected from a species originally triploid but now losing duplicated genes stochastically. In these clusters, the extra *M. incognita* CDS was less likely to be sister to one of the other *M. incognita* CDSs than it was to be a sister to an *M. incognita*-*M. floridensis* pair. These data suggest that the triplicate loci in *M. incognita* are the three homeologues that have resulted from a second hybridisation event involving the hybrid *M. floridensis* and an unidentified second, likely non-hybrid parent (hypothesis D, Figure 5.2-D).

5.4.2 Molecular genetic approaches to *Meloidogyne* diversity

Molecular approaches to understanding the diversity of apomictic RKN have a long history and include studies of isozymes, mtDNA, ribosomal internal transcribed spacer (ITS), rDNA genes, random amplified polymorphic DNA markers (RAPDs), amplified fragment length polymorphisms (AFLPs), and other marker systems (see Blok and Powers 2009 [226] for a review). However, if some *Meloidogyne* species are in fact hybrids, this presents particular problems for the standard molecular approaches used to characterise diversity. These typically assume that species or isolates have diverged following a bifurcating, tree-like, evolutionary pathway. Hybridisation violates this assumption and produces more complex evolutionary histories that can either be misrepresented by single locus markers, or else produce intermediate or equivocal signal from multi-locus approaches. For example, a major reason that mtDNA and rDNA sequencing have been useful in evolutionary ecology is that they are effectively haploid. Hybrid taxa, which often retain just one of their parental species' genotypes at these loci [227], present particular problems for these approaches. While carefully benchmarked marker approaches may have utility in diagnostics, they will not be able to accurately reflect the complex evolutionary pathway of hybrid *Meloidogyne* species where different loci are likely to have experienced very different histories. Incongruence between markers is therefore to be expected as a true reflection of history,

rather than due to a lack of analytical power, and current estimates of phylogenetic relationships between hybrid taxa will need to be re-evaluated.

Genomic approaches to the RKN system hold many advantages, including documenting the genomic changes associated with host-specialisation, extreme polyphagy, and interaction with plant defence systems. An interesting and important question now is whether all apomictic RKN species have monophyletic origins, with species divergence perhaps related to aneuploidy, or are instead the result of repeated hybridisations of the same or similar parental lineages. Different patterns of origin may determine the extent to which control strategies may be broadly or only locally applicable. We are now perhaps close to the time where RKN isolates can be characterised not only with a trivial name (e.g., *M. incognita* race X) but rather a detailed list of genome-wide variants and their known association with the environment, response to nematicides, and virulence against a range of plant host species and genotypes—an approach that will surely be extremely valuable in optimising agricultural success.

Understanding the evolutionary history of *Meloidogyne* species is a priority since only by this route can the evolution of pathogenicity, resistance, emergence of new pathogens, horizontal transfer of genes, and geographic spread of one of the world's most important crop pathogens be properly understood. We caution therefore that although classical loci may be valuable for rapid diagnostics, population genomics must be embraced in order to really advance our understanding and our ability to intervene robustly.

6 What next for next-generation nematode genomes

Thanks to modern sequencing technologies and the methods described in this thesis, genomic and transcriptomic resources for nematodes can be rapidly obtained, enabling exciting research. The two results described here on the lack of conservation of non-coding elements across the phylum Nematoda and the unusual evolutionary history of a newly sequenced *Meloidogyne* species are just two examples of the kinds of evolutionary genomics studies possible. This concluding chapter summarises the previous chapters and offers possible future directions for each topic, along with a description of the specific research that I would like to pursue in the future.

6.1 What can we do with 959 Nematode Genomes

Chapter 1 began by describing the nematodes that have been sequenced so far and why more such genomes are needed. Although ten nematode genomes have already been published, genomic data are publicly available for about 20 additional genomes (including the 4 described in this thesis). Only 20 genomes were selected for the analyses in chapters 3 and 4 of this thesis because they were the only ones with complete genome, proteome, and gene-model data available on WormBase at the time.

The 959NG wiki [13] was developed to keep track of completed genomes as well as genomes that have been proposed or are underway. The wiki has helped foster at least two collaborations, including the genome of *Dirofilaria immitis* described in this thesis, where the combined data sets from two research groups not only led to a better genome assembly, but also to the finding that the two geographically distant strains had almost no genomic variability [51].

Where does the 959NG wiki go from here? We think it has the potential to be more than just a list of genome projects. Because the species are organised phylogenetically and the tree is stored in a user-editable format, 959NG can be an always up-to-date, definitive reference for nematode phylogenies. For example, sequencing *D. immitis* allowed us to resolve the phylogeny of the Onchocercidae using multi-gene methods [51]. Also, unlike traditional database driven sites where the database design has to be fixed, 959NG uses a Semantic MediaWiki architecture underneath and any amount of structured and unstructured data can be added as pages and properties by any user.

Although not implemented yet, I would like to add information such as tRNA counts (as described in Chapter 3), which would enable users to ask queries such as "which species have tRNA counts that are more than 2 std-devs above or below the Clade mean". Information on the draft genomes could also be included, such as scaffold N50 or the

CEGMA completeness values of each assembly, enabling researchers interested in cross-species analyses to pick genomes that meet certain completeness criteria. Thus, 959NG could become a resource for storing and querying all kinds of genome properties.

Currently, only one nematode genome (*T. spiralis*) from clades I and II has been published. As sequencing and analyses become easier, hundreds of nematode genomes are likely to become available in the next few years, increasing the diversity of the nematodes being sequenced, providing a more complete overview of this ancient phylum.

Assuming we reach our first goal of 959 nematode genomes, I would personally like to carry out two types of studies:

1. Identify more functional elements using comparative genomics and selection signatures as done in recent studies on mammals [228] and vertebrates [196]. Additional annotations using other technologies such as ChIP-Seq would also help to determine functional elements. It would be especially interesting to identify functional elements that are restricted to specific branches of the phylogeny. By pinpointing when certain functional elements arose in the phylogeny, we could develop deeper insights into the process of genome evolution.
2. Correlate gene sets with life-histories and trophic mechanisms. Whole-genome sequences will allow us to obtain nearly complete gene catalogues for each species, enabling us to identify sets of genes that are correlated with particular characteristics such as parasitism [21]. These results might aid in the development of vaccines against parasitic nematode pathogens.

Both types of studies are currently possible with a few genomes. However, with hundreds of genomes, the results will be highly robust and not easily affected by individual events.

6.2 Will new technologies make assembly redundant

Chapter 2 described the workflow for assembling nematode genomes using low cost short-read sequencing and the results of applying that workflow to convert raw sequence data to complete contaminant-free assemblies for four species. With the exception of *D. immitis*, only paired-end (PE) libraries were available for these nematodes. Despite these limited resources, essentially complete (according to CEGMA) high-quality multi-gene sized contigs and scaffolds for three of the four genomes were obtained. The exception was the *M. floridensis* assembly, which had a low CEGMA completeness with only 72.18% of core genes partially found. Surprisingly, even the previously published Sanger-sequenced *M. incognita* was only 77.82% CEGMA complete (using partial genes). Given that the two species are closely related (Chapter 5), it is possible that these low numbers indicate a biological reason or methodological bias that would be fascinating to explore further.

The assembly workflow described in Chapter 2 addresses several NGS *de novo* assembly issues and how to deal with them: quality and adapter trimming, read error-correction, contaminant and co-biont visualisation, read separation, digital normalisation, parameter exploration during re-assembly, and avoiding mis-assemblies using PE-contaminated MP data. Some steps of this workflow are specific to the Illumina short-read sequencing technology used in this thesis. However, one of the main innovations proposed in this thesis is the taxon-annotated GC-coverage plot, which should be applicable to any NGS project. GC-coverage plots have been used in the metagenomics literature before [229], but using them along with a preliminary assembly to visualise contaminants and using the visualisations to guide read separation and re-assembly are new approaches. Making TAGC plots, also known informally as "blob" plots, for every genome sequencing project is highly recommended, as they provide a quick way of visualising what is in the raw data (and at what coverage) before spending a lot of time optimising assemblies using different software or parameters. At the Blaxter Lab, we now create preliminary assemblies and do a "blobology" run on every genome sequencing project, and our colleagues in several labs around the world are using this approach regularly as well.

*... blobology's the best hope I've had of
getting out the crud!*

Erich Schwarz (Cornell University, pers. comm.)

"Blobology" is also important because it provides a way to systematically process non-clonal samples from the wild. Future small-scale genome sequencing projects will most likely resemble constrained metagenomics projects, providing exciting opportunities to use and adapt metagenomics tools. In two of the four species in Chapter 2 (*Caenorhabditis sp. 5* and *M. floridensis*), bacteria were clearly present in the samples at very high coverage. Further investigation is needed to determine if these bacteria are more than just food for the nematodes. In the other two species described in this thesis (*D. immitis* and *L. sigmodontis*), the endosymbiotic *Wolbachia* genomes were clearly visible in the TAGC plots. In both cases, an estimate of the *Wolbachia* genome coverage was critical in helping us reassemble their genomes in very few pieces. I would like to develop these tools further by automating the creation of TAGC plots and implementing state-of-the-art metagenomic clustering and binning algorithms so that individual genomes can be extracted rapidly from a mixed sample.

In such a rapidly changing field, any workflow steps or protocols are likely to have short shelf-lives. Improvements in sequencing technologies and informatics tools will certainly yield better and better assemblies. Yet, any advancement in technology should be subject to scrutiny regarding its best fit for a particular project or budget. For example, the single-molecule PacBioRS sequencing platform was recently used to generate long reads that

significantly improved genome assembly and scaffolding when used in combination with second-generation technologies like Illumina [139, 230, 231]. However, even 10X coverage of a 100 Mbp genome currently costs ~£2000, so Illumina short reads (both PE and MP) continue to be the most cost effective technology for labs on a budget.

In light of the rapid pace of technological advances, it is worthwhile to consider how genome sequencing would change if we assumed that single-molecule platforms such as PacBioRS and Oxford Nanopore would soon start producing high-throughput long-reads (>10 kilobases) at a reasonable cost. Perhaps the most important change would be that researchers would no longer have to spend weeks running different assembly programs with different parameters to figure out which assembly is best, based on a variety of metrics. The assembly process would be much faster, emphasising data management, annotation, and analysis as the new bottlenecks.

With longer reads and quicker assemblies, the future would most likely see population genome sequencing rather than individual genome sequencing. Different data structures would be needed from those in use today to keep track of the variations in the genome, such as the FastG assembly graph structure [232] or the coloured graph in Cortex [85]. Keeping track of variations during the assembly and analysis process would be easier than trying to tease apart that information from pre-assembled sequences. Standardised ways to store genomes and assemblies would be needed to enable easier comparisons. The Ensembl system [233] is one example of how a well-designed data storage system can simplify analysis by providing robust APIs to query genomic data.

Longer reads would also make it easier to separate reads from different organisms in environmental samples. I would like to look at different nematode-bacterial symbioses as small-scale metagenomic projects to understand the metabolic capabilities and dependencies of each specific association. As more such symbioses are studied, it should become possible to study the evolutionary history of these associations and it would be fascinating to understand if the associations are ancient or recent adaptations in the face of environmental changes. In the case of filarial parasitic nematodes such as *D. immitis*, *L. sigmodontis*, and *B. malayi*, understanding the bacterial *Wolbachia* associations with these nematodes may also provide insights into controlling nematode infections in humans and animals.

Although future improvements in sequencing and informatics tools would be very welcome, just managing the current number of choices can be overwhelming at times. When NGS short-reads for *de novo* genomes first became popular in 2008, only a handful of assemblers [81, 83] could deal with the millions of reads generated. Last year, over 20 prominent assembly algorithms were used by teams participating in the Assemblathon [89] (although many more algorithms exist), and this year, the tool to watch seems to be SPAdes, which is yielding >100 kbp N50 *E. coli* assemblies with PE reads alone. Tools for error correction [83, 103-105], digital normalisation [109], and metagenome assembly [234, 235] have also

improved. However, one of the biggest challenges of bioinformatics is deciding what tools work best. Each published tool typically shows a few cases where it is as good as or better than other competing tools, but there is still a critical need for independent and unbiased studies that compare *de novo* genome assembly [27, 89], *de novo* transcriptome assembly [93], and gene prediction [153], for every stage of the process of creating genomic and transcriptomic resources. Because new tools are constantly emerging, these comparisons will also have to be updated. Assemblathon2 is already underway, but following an "annual update" approach for each category of tools would be very resource-intensive. One solution to this problem could be to make tool comparisons a part of bioinformatics training programmes; students could learn about tools by running them and recording their results on diverse data sets. The results could be recorded in a standard format that makes tool and algorithm comparisons easier. I would like to establish a bioinformatics protocol-evaluation platform (as mentioned in the Discussion section in Chapter 2) where users can share their attempts with different tools for assembly, read pre-processing, error-correction, and scaffolding.

The workflow in this thesis is one attempt to record and share the best practices for a small lab assembling and annotating metazoan genomes between 50–100 Mbp. However, metrics are still needed to determine how well the tools perform when the true sequence is not known. Assessing genome assemblies using sequence length metrics like the scaffold N50 can be very misleading, as shown in Chapter 2. Therefore, a combination of metrics (N50, CEGMA, and alignments to known EST and protein sequences) were used to evaluate assemblies. While the use of EST and protein sequence alignments seemed fairly obvious, it is interesting to note that the approach has not been documented in any recent NGS genome papers to date as a way of comparing alternative assemblies. Since the work in this thesis was conducted, newer methods for assessing genome assemblies without reference sequences have also been developed, which should make the process of choosing assemblies even more objective and unbiased [236, 237].

All of the tools, scripts, and commands that were run for this thesis have been made available at <http://github.com/sujaikumar> and are listed in Appendix A. In keeping with the Unix philosophy [238], almost all of the scripts do just one thing, and do it well. In addition, all scripts are designed to work well as part of a toolchain, accepting input on the standard input stream, and providing output on the standard output stream or in well-defined user-specifiable files. Given their modular nature, they can be easily ported to workflow systems like Galaxy [239], Taverna [240], and GeneProf [241]. As it becomes easier to deploy some of these workflow systems to an on-demand computing platform like Amazon's EC2 [242], it will likely be possible in the very near future for small groups with no bioinformatics expertise and no capital investment in high-end computational infrastructure to carry out all of these tasks at a low, per-use cost.

6.3 Can we generate more accurate annotations

Once the four genomes were assembled, they were annotated to create useful genomic resources as described in **Chapter 3**. Protein-coding genes and RNA-Sequences were predicted (structural annotation), and putative functions were assigned where possible (functional annotation). The MAKER2 pipeline [142] was used because it combined *de novo* prediction tools and evidence from sequence alignments to ESTs and protein sequences. RNA annotation was carried out using Rfamscan [199] and tRNAscan [177]. Functional annotation was carried out using Blast2GO and InterProScan [174]. Some of the shortcomings of automated annotations were discussed. Although not perfect, such efforts provide a good starting point for analysing these genomes and are useful as long as the results are used with the understanding that they may not be complete.

To put these annotations into context, the functional annotation was redone using the same methods for 16 additional publicly available genomes, and the results were compared. This is the first such phylogenetically-organised comparison for nematode genomes. Protein-coding gene metrics, InterPro annotations, and tRNA counts all seem to have clade-specific biases, but several exceptions were also found. These exceptions could point to interesting biological differences or methodological biases and errors (such as gene over-prediction). Either way, further investigations would likely prove interesting. Apart from the 4 genomes discussed in this thesis, each of the other 16 nematode genomes was annotated separately, using separate programs, different levels of human annotation, and different contexts (more evidence is available for recent genomes compared to previously sequenced genomes). Therefore it is very possible that some of the differences are artefacts of the specific annotation process used. Moving forward, it would be worthwhile to run all the genomes again through the same automated annotation pipeline, using the best evidence available, and note whether these exceptions remain. This is similar to the solution used by Ensembl, which uses a consistent and standardised annotation pipeline across its 70 metazoan genomes (Release 68, [233]).

As with genome assembly protocols, we need a set of best practices for genome annotation. With more time and computational resources available, I would like to systematically develop and evaluate other annotation pipelines. There are two main ways to generate competing annotations that can then be assessed using a set of standard metrics such as the AED [243] or matches to known gene models. The first is to run existing pipelines oneself and the second is to invite annotation submissions by developers and end-users in a competition format like the Assemblathon [89]. I believe an ongoing crowd-sourced competition (an "Annotathon") will be very useful to everyone interested in annotating eukaryotic genomes once we have established a set of shared metrics and shared annotation evaluation tools. By documenting the results of multiple annotation pipelines and

parameters on many types of organisms with different levels of input resources, we should be able to establish and popularise best practices in the field of genome annotation.

I would also like to improve annotations for nematode genomes using additional data such as RNA-Seq and ChIP-Seq. The modEncode project [45] aims to identify all functional genome elements for the model organisms *C. elegans* and *D. melanogaster*. As technologies improve and costs drop, such efforts will become more accessible even for non-model organisms. RNA-Seq data is also useful because it can be used to elucidate alternate transcripts.

Apart from the process of gene-prediction and functional element classification, the databases and standards for storing these annotations need some improvements and standardisation. One important operational problem is the storage of and access to multiple annotations for each species. Annotation systems must also be flexible enough to incorporate data from such studies without changing the previous gene nomenclatures except when there is a conflict. Similarly, it must be able to easily transfer annotations for highly conserved regions from closely related species. In addition, the levels of annotation are also important. Some gene models might be highly reliable because they are hand-curated, have transcript evidence, and are predicted by gene-finding algorithms. Other gene models might be less reliable as they only have computational evidence. A good system must be able to store these annotations in a way that allows choosing gene-sets of different qualities for different purposes. In this thesis, I made an attempt to store two levels of annotation: genes with sequence alignments from ESTs and proteins, and genes with only computational predictions. In the future, I would like to help specify revisions to the standards for genome annotation so that additional information can be stored and utilised for making novel inferences.

For storing annotation data, I used GFF3 plain-text files, which, although convenient to read and write, were not as powerful as relational databases for storing and querying relationships between annotation features. For future projects, I would like to use databases for storing annotations. Fortunately, the existing GFF3 annotation files can be migrated to databases easily. Unfortunately, there seem to be many competing choices for genome annotation databases such as Chado [244], BioSQL [245], and Ensembl [233]. More work will be needed to determine which of these will suit our needs best. Currently, it seems as if Chado would be the easiest, Ensembl would be the most powerful and come with a set of mature tools and APIs, and BioSQL would be the most flexible (allowing any sequence data and annotations to be stored, not just genomes). Some effort will also be needed to convince the International Nucleotide Sequence Database Collaboration (INSDC) to allow multiple annotations to be stored for each genome sequence in the public databases.

6.4 Do CNEs really define genetic regulatory networks

Chapters 4 and 5 build on the resources created in Chapters 2 and 3. The 20 genomes analysed and compared for Chapter 3 became the dataset for **Chapter 4**. Armed with these genomic resources, an elegant hypothesis by Vavouri and Lehner [192] on the connection between deeply conserved non-coding elements (CNEs) and the phylum body plan was addressed.

Previous research had found that CNEs are deeply conserved within a phylum, are not shared across phyla, and often regulate the same core genes across different phyla. These observations led Vavouri and Lehner to propose that CNEs possibly define the genetic regulatory network that represents each phylum's body plan. CNEs in nematodes had been studied previously using only three species from the same genus. However, when the search for CNEs was expanded to 20 nematode species spread across four clades in this thesis, the results showed no CNEs shared across clade boundaries, even though a more sensitive method (than the one previously used) was tried.

Some CNEs were seen at the clade, family, and genus levels, and they were located near developmental genes. However the number of CNEs dropped rapidly as older nodes were analysed, leading to the conclusion that no non-coding elements were deeply conserved, and were therefore highly unlikely to be linked to the phylum body plan. A re-examination of Glazov's findings in insect CNEs [197] indicated that although thousands of CNEs were found between two *Drosophila* species, only two CNEs were shared across the order Diptera. This finding showed that although CNEs might play a role in defining the vertebrate body plan [187, 188], they are unlikely to be associated with phylum body plans in both the invertebrate phyla studied so far. Although this study had a negative result, it demonstrates the power of sequencing multiple genomes in a phylum in overturning a hypothesis formed using only three species from one genus. Future studies in finding CNEs could explore more sensitive tools for comparing multiple genomes and should examine nematode clade-specific CNEs further to see if they follow the same trends as vertebrate CNEs.

The lack of non-coding elements conserved across the phylum does not disprove the underlying hypothesis that perhaps diverse animal body plans are a consequence of gene regulatory networks (GRNs) being rewired in different ways. A GRN is defined by the interactions between proteins (and other molecules) and gene regulatory regions that lead to specific patterns of gene expression needed for the development of any organism. It is therefore possible that two GRNs have identical structures and functions whereas the regulatory regions have no sequence conservation. For example, a transcription factor (TF) and a transcription factor binding site (TFBS) might co-evolve together in a way that the interaction does not change even though the TFBS sequence changes. Thus, even though no CNEs were found shared across the phylum, it is possible that GRNs are shared across the phylum and are distinct from the GRNs for a different phylum.

I would like to test this theory in the future by first documenting key GRNs for several nematode and non-nematode species and then testing the GRN structures to see if they are shared across all nematodes but not shared outside the phylum. The first step will be to improve genome annotation data (using tools such as ChIP-Seq) to identify the regulatory regions and the molecules that bind to them. Once the GRNs are clearly defined, shared GRN structures could be further analysed to determine their evolutionary dynamics. Do the TFBSs or TFs change such that one evolves faster or do they coevolve? Are regulatory regions likely to be duplicated so that the same TF can bind to multiple locations and serve as a modular cassette of gene expression? Answers to questions such as these will help us understand how the great diversity of metazoan body plans came to be.

6.5 What could fifteen *Meloidogyne* genomes tell us about hybrid speciation

Where Chapter 4 took a wide-range view of the whole phylum, **Chapter 5** restricted itself to three species to understand the origin of gene duplicates in a single genus. Lunt [122] had observed that the most damaging plant-parasitic root-knot nematodes (RKNs) were all aneuploid diploids that reproduced by obligatory mitotic parthenogenesis (apomictically) and that they likely arose as a result of recent interspecific hybridisations. Abad *et al.* [38] had previously sequenced the *M. incognita* genome and observed that many genes were present as duplicates. Lunt sequenced several genes in many *Meloidogyne* species and discovered that some *M. floridensis* sequences were identical to some of the *M. incognita* duplicates, and proposed that *M. incognita* was the result of a recent hybridisation.

To test the possible hybrid origins of *M. incognita*, the complete genome sequences of *M. floridensis* (generated as described in Chapter 2), *M. incognita*, and the outgroup diploid *M. hapla* were used. The results showed that the CDSs in *M. floridensis* had the same pattern of divergent self-identity as the *M. incognita* sequences, hinting that *M. floridensis* might also be a recent interspecific hybrid. Using thousands of homologous clusters of CDSs from all three species, a clear excess of *M. incognita* CDSs in clusters where only one *M. hapla* CDS was present (assuming these clusters represent the single-copy genes in the ancestral lineage) were identified. On analysing the phylogenies of these clusters, most trees put one *M. incognita* CDS as a sister to an *M. floridensis* CDS, and many clusters had an excess of *M. incognita* branches, providing support to the model that *M. floridensis* is a hybrid and was one parent species of a hybridisation event that gave rise to the triploid *M. incognita*. This result would not have been possible without the sequence of the *M. floridensis* genome. Even though the *M. floridensis* genome assembly was not as contiguous as the other two assemblies, full-length or nearly full-length alignments to CDSs from both *M. hapla* and *M. incognita* were recovered and used to study the hybrid origins within this genus.

The future of *Meloidogyne* genome research is exciting, with a project just starting at the Blaxter Lab that will sequence 15 species in the genus and create genomic and transcriptomic resources to investigate the effects of organismal reproductive mode on genome content and diversity. *Meloidogyne* species reproduce by amphimixis (meiotic sexual reproduction), apomixis (mitotic parthenogenesis), and automixis (meiotic parthenogenesis). These modes are somewhat specific to branches in the *Meloidogyne* phylogenetic tree (Figure 5.1) and a comparative genomics approach should be able to unearth the genic and genomic correlates of reproduction mode.

Additional *Meloidogyne* genomes will also help us understand the possible reasons for the unexpectedly low recovery of core eukaryotic genes in *M. incognita* and *M. floridensis*. Low CEGMA scores could be due to excessive repetition in the core eukaryotic genes leading to poor assemblies in those regions. Alternately, it is also possible that the core eukaryotic genes are not as essential as previously assumed. If any of the upcoming *Meloidogyne* genomes also demonstrate low CEGMA scores, then it will be fascinating to test a phylogenetically diverse set of genomes to test different hypotheses for the cause of these low scores.

With additional genomes, it will become increasingly necessary to automate how we tally the number of gene trees that support a particular hypothesis. In Figure 5.5, it was possible for each tree to support multiple hypotheses on the hybrid origins of the species involved, depending on the number of gene losses or gains that would be needed to arrive at each CDS cluster. Manually computing the possible combinations of gene losses or gains was cumbersome and I would like to work on an algorithmic approach to this problem.

6.6 Summary

In summary, this thesis demonstrates the creation and use of next-generation nematode genome resources. Half a decade ago, generating a genome assembly for a nematode of interest and annotating it would have been prohibitively expensive and time-consuming, and only large labs or consortiums could have afforded to do so. Today, with NGS, the sequencing costs for a 100-300 Mbp nematode can be minimal (on the order of a thousand £/\$/€), and the technologies are only getting better and less expensive. Hopefully, the tools and processes described here will encourage even small labs without genomics expertise to consider sequencing the genomes of their nematodes of interest to answer their specific questions. With more nematode (and indeed, all other) genomes, comparative genomics studies like the ones described in this thesis will become easier and more robust.

Bibliography

1. Lamshead PJD: **Recent Developments in Marine Benthic Biodiversity Research.** *Oceanis* 1993, **19**:5-24.
2. Platt H: **Foreword.** In: *The Phylogenetic Systematics of Free-Living Nematodes.* Edited by Lorenzen S. London: The Ray Society; 1994: 5-6.
3. Sulston J, Schierenberg E, White J, Thomson J: **The embryonic cell lineage of the nematode *Caenorhabditis elegans.*** *Developmental Biology* 1983, **100**:64-119.
4. C. elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282**:2012-2018.
5. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A *et al*: **The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics.** *PLoS Biology* 2003, **1**:E45.
6. Cutter AD, Dey A, Murray RL: **Evolution of the *Caenorhabditis elegans* Genome.** *Molecular Biology and Evolution* 2009, **26**:1199-1234.
7. White JG, Southgate E, Thomson JN, Brenner S: **The Structure of the Nervous System of the Nematode *Caenorhabditis elegans.*** *Philosophical Transactions of the Royal Society of London B, Biological Sciences* 1986, **314**:1-340.
8. Crittenden SL, Eckmann CR, Wang L, Bernstein DS, Wickens M, Kimble J: **Regulation of the mitosis/meiosis decision in the *Caenorhabditis elegans* germline.** *Philosophical Transactions of the Royal Society of London B, Biological Sciences* 2003, **358**:1359-1362.
9. Wang MC, O'Rourke EJ, Ruvkun G: **Fat Metabolism Links Germline Stem Cells and Longevity in *C. elegans.*** *Science* 2008, **322**:957-960.
10. Brooker S: **Estimating the global distribution and disease burden of intestinal nematode infections: adding up the numbers--a review.** *International Journal for Parasitology* 2010, **40**:1137-1144.
11. Kumar S, Koutsovoulos G, Kaur G, Blaxter M: **Toward 959 nematode genomes.** *Worm* 2012, **1**:42-50.
12. Blaxter M, Kumar S, Kaur G, Koutsovoulos G, Elsworth B: **Genomics and transcriptomics across the diversity of the Nematoda.** *Parasite Immunology* 2012, **34**(2-3):108-120.
13. Kumar S, Schiffer PH, Blaxter M: **959 Nematode Genomes: a semantic wiki for coordinating sequencing projects.** *Nucleic Acids Research* 2012, **40**:D1295-D1300.
14. Siddiqi MR: **Tylenchida: parasites of plants and insects,** 2nd edn. New York: CABI; 2000.
15. Anderson RC: **Nematode parasites of vertebrates: their development and transmission,** 2nd edn. New York: CABI; 2000.
16. Brenner S: **The genetics of *Caenorhabditis elegans.*** *Genetics* 1974, **77**:71-94.
17. Nutman TB: **Lymphatic filariasis.** London: Imperial College Press; 2000.
18. Cobb NA: **Nematodes and their relationships.** In: *Yearbook of the United States Department of Agriculture* 1914. 1915: 457-490.
19. Chan MS: **The global burden of intestinal nematode infections--fifty years on.** *Parasitology Today* 1997, **13**(11):438-443.
20. Trudgill DL, Blok VC: **Apomictic, polyphagous root-knot nematodes: exceptionally successful and damaging biotrophic root pathogens.** *Annual Review of Phytopathology* 2001, **39**:53-77.
21. Blaxter M: **Nematoda: genes, genomes and the evolution of parasitism.** *Adv Parasitol* 2003, **54**:101-195.
22. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans.*** *Nature* 1998, **391**:806-811.
23. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans.*** *Nature* 2000, **403**(6772):901-906.
24. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC: **Green fluorescent protein as a marker for gene expression.** *Science* 1994, **263**(5148):802-805.
25. Horvitz HR, Sternberg PW: **Multiple intercellular signalling systems control the development of the *Caenorhabditis elegans* vulva.** *Nature* 1991, **351**:535-541.

26. Hengartner MO, Ellis R, Horvitz R: **Caenorhabditis elegans gene ced-9 protects cells from programmed cell death.** *Nature* 1992, **356**:494-499.
27. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
28. Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH: **Genomics in C. elegans: so many genes, such a little worm.** *Genome Research* 2005, **15**(12):1651-1660.
29. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Fretze S, Harrow J, Kaul R *et al*: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57-74.
30. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Cruz Ndl, Duong A, Fang R *et al*: **WormBase 2012: more genomes, more data, new website.** *Nucleic Acids Research* 2012, **40**:D735-D741.
31. Blaxter M: **Nematodes: The Worm and Its Relatives.** *PLoS Biology* 2011, **9**(4):e1001050.
32. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM *et al*: **A molecular evolutionary framework for the phylum Nematoda.** *Nature* 1998, **392**:71-75.
33. Wasmuth J, Schmid R, Hedley A, Blaxter M: **On the extent and origins of genic novelty in the phylum Nematoda.** *PLoS Neglected Tropical Diseases* 2008, **2**(7):e258.
34. Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD: **Eukaryotic genome size databases.** *Nucleic Acids Research* 2007, **35**:D332-D338.
35. Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S *et al*: **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* 2008, **105**:14802-14807.
36. Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z *et al*: **Ascaris suum draft genome.** *Nature* 2011, **479**:529-533.
37. Müller F, Bernard V, Tobler H: **Chromatin diminution in nematodes.** *Bioessays* 1996, **18**:133-138.
38. Abad P, Gouzy J, Aury J-M, Castagnone-Sereno P, Danchin EGJ, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC *et al*: **Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita.** *Nature Biotechnology* 2008, **26**:909-915.
39. Fenn K, Conlon C, Jones M, Quail MA, Holroyd NE, Parkhill J, Blaxter M: **Phylogenetic Relationships of the Wolbachia of Nematodes and Arthropods.** *PLoS Pathogens* 2006, **2**(10):e94.
40. Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J: **Phylum-Wide Analysis of SSU rDNA Reveals Deep Phylogenetic Relationships among Nematodes and Accelerated Evolution toward Crown Clades.** *Molecular Biology and Evolution* 2006, **23**:1792-1800.
41. van Megen H, van den Elsen S, Holterman M, Karssen G, Mooyman P, Bongers T, Holovachov O, Bakker J, Helder J: **A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences.** *Nematology* 2009, **11**:927-950.
42. De Ley P, Blaxter M: **Systematic position and phylogeny.** In: *The Biology of Nematodes.* Edited by Lee DL. London: Taylor and Francis; 2002: 1-30.
43. Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH *et al*: **A transcriptomic analysis of the phylum Nematoda.** *Nature Genetics* 2004, **36**:1259-1267.
44. Elsworth B, Wasmuth J, Blaxter M: **NEMBASE4: The nematode transcriptome resource.** *International Journal for Parasitology* 2011, **41**:881-894.
45. Gerstein MB, Lu ZJ, Nostrand ELV, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K *et al*: **Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project.** *Science* 2010, **330**:1775-1787.
46. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D *et al*: **Draft genome of the filarial nematode parasite Brugia malayi.** *Science* 2007, **317**:1756-1760.
47. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P *et al*: **The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism.** *Nature Genetics* 2008, **40**:1193-1198.

48. Mortazavi A, Schwarzh EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW: **Scaffolding a *Caenorhabditis nematode* genome with RNA-seq.** *Genome Research* 2010, **20**:1740-1747.
49. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton L, Minx P *et al*: **The draft genome of the parasitic nematode *Trichinella spiralis*.** *Nature Genetics* 2011, **43**:228-235.
50. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, Takanashi T, Tsai IJ, Assefa SA, Cock PJA *et al*: **Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen *Bursaphelenchus xylophilus*.** *PLoS Pathogens* 2011, **7**(9):e1002219.
51. Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S *et al*: **The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets.** *The FASEB Journal* 2012:fj.12-205096.
52. Miyazaki S, Sugawara H, Ikeo K, Gojobori T, Tateno Y: **DDBJ in the stream of various biological data.** *Nucleic Acids Research* 2004, **32**(suppl 1):D31-D34.
53. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tv^orraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R *et al*: **The European Nucleotide Archive.** *Nucleic Acids Research* 2011, **39**(suppl 1):D28-D31.
54. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Research* 2011, **39**:D32-D37.
55. Kumar S, Blaxter M: **959 Nematode Genomes** [<http://959.nematodegenomes.org>] Accessed April 2 2010
56. **Semantic MediaWiki** [<http://semantic-mediawiki.org>] Accessed April 2 2010
57. Kumar S: **Assemblage** [<http://github.com/sujaikumar/assemblage>] Accessed May 11 2012
58. Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C *et al*: **Genome project standards in a new era of sequencing.** *Science* 2009, **326**:236-237.
59. Kumar S, Blaxter ML: **Simultaneous genome sequencing of symbionts and their hosts.** *Symbiosis* 2011, **55**(3):119-126.
60. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.
61. Coulson A, Sulston J, Brenner S, Karn J: **Toward a physical map of the genome of the nematode *Caenorhabditis elegans*.** *Proceedings of the National Academy of Sciences of the United States of America* 1986, **83**(20):7821-7825.
62. Anderson S: **Shotgun DNA sequencing using cloned DNase I-generated fragments.** *Nucleic Acids Research* 1981, **9**(13):3015-3027.
63. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376-380.
64. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**(7218):53-59.
65. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K *et al*: **A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning.** *Genome Research* 2008, **18**(7):1051-1063.
66. Wang J, Czech B, Crunk A, Wallace A, Mitreva M, Hannon GJ, Davis RE: **Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles.** *Genome Research* 2011, **21**:1462-1477.
67. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S *et al*: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **108**:1513-1518.
68. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A *et al*: **The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle.** *PLoS Genetics* 2011, **7**(2):e1002007.

69. Myers E, Sutton G, Delcher A, Dew I, Fasulo D, Flanigan M, Kravitz S, Mobarry C, Reinert K, Remington K *et al*: **A whole-genome assembly of *Drosophila***. *Science* 2000, **287**:2196-2204.
70. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:9748-9753.
71. Myers EW: **The fragment assembly string graph**. *Bioinformatics* 2005, **21 Suppl 2**:ii79-85.
72. Simpson JT, Durbin R: **Efficient de novo assembly of large genomes using compressed data structures**. *Genome Research* 2012, **22**:549-556.
73. Li H: **Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly**. *Bioinformatics* 2012, **28**:1838 - 1844.
74. Warren RL, Sutton GG, Jones SJ, Holt RA: **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics* 2007, **23**:500-501.
75. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing**. *Genome Research* 2007, **17**:1697-1706.
76. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD: **Extending assembly of short DNA sequences to handle error**. *Bioinformatics* 2007, **23**:2942-2944.
77. Huang X, Madan A: **CAP3: A DNA Sequence Assembly Program**. *Genome Research* 1999, **9**:868-877.
78. Green P: **Phrap** [<http://www.phrap.org/>] Accessed March 23 2010
79. Chevreux B, Wetter T, Suhai S: **Genome Sequence Assembly Using Trace Signals and Additional Sequence Information**. In: *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99*. 1999: 45-56.
80. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates**. *Bioinformatics* 2008, **24**:2818-2824.
81. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Research* 2008, **18**:821-829.
82. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABYSS: a parallel assembler for short read sequence data**. *Genome Research* 2009, **19**:1117-1123.
83. **SOAPdenovo** [<http://soap.genomics.org.cn/soapdenovo.html>] Accessed February 15, 2012 2012
84. **CLC bio: CLC Assembly Cell User Manual** [<http://www.clcbio.com/index.php?id=1331>] Accessed January 20 2010
85. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs**. *Nature Genetics* 2012, **44**(2):226-232.
86. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data**. *Genomics* 2010, **95**:315-327.
87. Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B: **A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies**. *PLoS ONE* 2011, **6**(3):e17915.
88. Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng H-WW: **Comparative studies of de novo assembly tools for next-generation sequencing technologies**. *Bioinformatics* 2011, **27**:2031-2037.
89. Earl DA, Bradnam K, John JS, Darling A, Lin D, Faas J, Yu HOK, Vince B, Zerbino DR, Diekhans M *et al*: **Assemblathon 1: A competitive assessment of de novo short read assembly methods**. *Genome Research* 2011, **21**:2224-2241.
90. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M *et al*: **GAGE: A critical evaluation of genome assemblies and assembly algorithms**. *Genome Research* 2012, **22**:557-567.
91. Koren S, Treangen TJ, Pop M: **Bambus 2: Scaffolding Metagenomes**. *Bioinformatics* 2011, **27**:2964-2971.
92. **MSR-CA Genome Assembler, draft manual 1.0** [http://www.genome.umd.edu/SR_CA_MANUAL.htm] Accessed September 3 2012
93. Kumar S, Blaxter M: **Comparing de novo assemblers for 454 transcriptome data**. *BMC Genomics* 2010, **11**:571.

94. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**:1061-1067.
95. Parkinson J, Whitton C, Schmid R, Thomson M, Blaxter M: **NEMBASE: a resource for parasitic nematode ESTs.** *Nucleic Acids Research* 2004, **32**:D427-D430.
96. Martin J, Abubucker S, Heizer E, Taylor CM, Mitreva M: **Nematode.net update 2011: addition of data sets and tools featuring next-generation sequencing data.** *Nucleic Acids Research* 2012, **40**(D1):D720-D728.
97. Parkinson J, Blaxter M: **Expressed sequence tags: analysis and annotation.** *Methods in Molecular Biology* 2004, **270**:93-126.
98. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Research* 2010, **38**:1767-1771.
99. Andrews S: **Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data** [<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>] Accessed January 15, 2012 2012
100. Metzker ML: **Sequencing technologies - the next generation.** *Nature Reviews Genetics* 2010, **11**(1):31-46.
101. **Scythe** [<https://github.com/ucdavis-bioinformatics/scythe>] Accessed February 2 2011
102. **Sickle** [<https://github.com/ucdavis-bioinformatics/sickle>] Accessed February 2 2011
103. Kelley D, Schatz M, Salzberg S: **Quake: quality-aware detection and correction of sequencing errors.** *Genome Biology* 2010, **11**:R116.
104. Liu Y, Schmidt B, Maskell D: **DecGPU: distributed error correction on massively parallel graphics processing units using CUDA and MPI.** *BMC Bioinformatics* 2011, **12**:85.
105. Schröder J, Schröder H, Puglisi SJ, Sinha R, Schmidt B: **SHREC: a short-read error correction method.** *Bioinformatics* 2009, **25**:2157-2163.
106. Zhang Z, Schwartz S, Wagner L, Miller W: **A Greedy Algorithm for Aligning DNA Sequences.** *Journal of Computational Biology* 2000, **7**:203-214.
107. Wickham H: **ggplot2: elegant graphics for data analysis:** Springer New York; 2009.
108. R. Core Team: **R: A Language and Environment for Statistical Computing.** Vienna, Austria; 2012.
109. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH: **A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data.** *arXiv* 2012:1203.4802 [q-bio.GN].
110. National Centre for Biotechnology Information: **How to Submit WGS Genomes** [<http://www.ncbi.nlm.nih.gov/genbank/wgs.submit>] Accessed February 10, 2013 2013
111. **SOAPdenovo-Trans** [<http://soap.genomics.org.cn/SOAPdenovo-Trans.html>] Accessed February 15, 2012 2012
112. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Research* 2002, **12**:656-664.
113. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25**:3389-3402.
114. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**:1658-1659.
115. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2010, **27**:578-579.
116. WANG G-X, REN S, REN Y, AI H, CUTTER AD: **Extremely high molecular diversity within the East Asian nematode *Caenorhabditis* sp. 5.** *Molecular Ecology* 2010, **19**:5022-5029.
117. Kiontke K, Felix MA, Ailion M, Rockman M, Braendle C, Penigault JB, Fitch D: **A phylogeny and molecular barcodes for *Caenorhabditis*, with numerous new species from rotting fruits.** *BMC Evolutionary Biology* 2011, **11**:339.
118. Kelley D, Salzberg S: **Clustering metagenomic sequences with interpolated Markov models.** *BMC Bioinformatics* 2010, **11**:544.
119. Parks DH, MacDonald NJ, Beiko RG: **Classifying short genomic fragments from novel lineages using composition and homology.** *BMC Bioinformatics* 2011, **12**:328.

120. Saeed I, Tang S-LL, Halgamuge SK: **Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition.** *Nucleic Acids Research* 2012, **40**(5):e34.
121. Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO: **Application of tetranucleotide frequencies for the assignment of genomic fragments.** *Environmental Microbiology* 2004, **6**:938-947.
122. Lunt DH: **Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins.** *BMC Evolutionary Biology* 2008, **8**:194.
123. Kozek WJ: **What is new in the Wolbachia/Dirofilaria interaction?** *Veterinary Parasitology* 2005, **133**(2-3):127-132.
124. Loman NJ: **Tips for de novo bacterial genome assembly** [<http://pathogenomics.bham.ac.uk/blog/2009/09/tips-for-de-novo-bacterial-genome-assembly/>] Accessed September 1 2012
125. Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M *et al*: **The RAST Server: Rapid Annotations using Subsystems Technology.** *BMC Genomics* 2008, **9**:75.
126. Quail MA, Matthews L, Sims S, Lloyd C, Beasley H, Baxter SW: **Genomic libraries: II. Subcloning, sequencing, and assembling large-insert genomic DNA clones.** *Methods in Molecular Biology* 2011, **772**:59-81.
127. The Genome Center at Washington University: **Caenorhabditis sp11 JU1373-3.0.1 ASSEMBLY** [http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_sp11_JU1373/assembly/Caenorhabditis_sp11_JU1373-3.0.1/ASSEMBLY] Accessed June 22 2012
128. Illumina: **Preparing samples for sequencing genomic DNA 1003806 Rev. B.** San Diego: Illumina; 2008.
129. Elsworth B: **SCUBAT (Scaffolding Contigs Using BLAT And Transcripts)** [<https://github.com/elswob/SCUBAT>] Accessed March 12 2012
130. Yang X, Chockalingam SP, Aluru S: **A survey of error-correction methods for next-generation sequencing.** *Briefings in Bioinformatics* 2012:bbs015.
131. Yang X, Dorman KS, Aluru S: **Reptile: representative tiling for short read error correction.** *Bioinformatics* 2010, **26**(20):2526-2533.
132. Ilie L, Fazayeli F, Ilie S: **HiTEC: accurate error correction in high-throughput sequencing data.** *Bioinformatics* 2011, **27**:295-302.
133. Kao W-CC, Chan AH, Song YS: **ECHO: a reference-free short-read error correction algorithm.** *Genome Research* 2011, **21**(7):1181-1192.
134. Walenz B, Sutton G, Miller J: **SourceForge.net: Pair classification within Illumina mate pair data - wgs-assembler** [http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Pair_classification_within_Illumina_mate_pair_data] Accessed March 12 2012
135. Wooley JC, Godzik A, Friedberg I: **A Primer on Metagenomics.** *PLoS Computational Biology* 2010, **6**(2):e1000667.
136. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV: **Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota.** *Science* 2012, **335**:587-590.
137. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms.** *Nature Biotechnology* 2012, **30**(5):434-439.
138. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: **Real-time DNA sequencing from single polymerase molecules.** *Methods in Enzymology* 2010, **472**:431-455.
139. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED *et al*: **Hybrid error correction and de novo assembly of single-molecule sequencing reads.** *Nature Biotechnology* 2012, **30**(7):693-700.
140. Oxford Nanopore Technologies: **Oxford Nanopore Technologies - About Us - For Customers** [<http://www.nanoporetech.com/about-us/for-customers>] Accessed January 20 2013
141. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC Genomics* 2012, **13**(1).

142. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12**:491.
143. Markowitz VM, Chen IMM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P *et al*: **IMG: the Integrated Microbial Genomes database and comparative analysis system.** *Nucleic Acids Research* 2012, **40**:D115-D122.
144. Richardson EJ, Watson M: **The automatic annotation of bacterial genomes.** *Briefings in Bioinformatics* 2012:bbs007.
145. Brent MR, Guigó R: **Recent advances in gene structure prediction.** *Current Opinion in Structural Biology* 2004, **14**(3):264-272.
146. Zhang MQ: **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 2002, **3**(9):698-709.
147. Valencia A: **Automatic annotation of protein function.** *Current Opinion in Structural Biology* 2005, **15**(3):267-274.
148. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Research* 2000, **28**(1):304-305.
149. The Gene Ontology Consortium: **Creating the Gene Ontology Resource: Design and Implementation.** *Genome Research* 2001, **11**(8):1425-1433.
150. Lerat E: **Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs.** *Heredity* 2010, **104**:520-533.
151. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M: **MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Research* 2008, **18**:188-196.
152. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nature Reviews Genetics* 2012, **13**:329-342.
153. Coghlan A, Fiedler T, McKay S, Flicek P, Harris T, Blasiar D, The nGASP Consortium, Stein L: **nGASP - the nematode genome annotation assessment project.** *BMC Bioinformatics* 2008, **9**:549.
154. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**(1):59.
155. Ding L, Sabo A, Berkowicz N, Meyer RR, Shotland Y, Johnson MR, Pepin KH, Wilson RK, Spieth J: **EAnnot: a genome annotation tool using experimental evidence.** *Genome Research* 2004, **14**(12):2503-2509.
156. Salamov AA, Solovyev VV: **Ab initio Gene Finding in Drosophila Genomic DNA.** *Genome Research* 2000, **10**(4):516-522.
157. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7**:62.
158. Borodovsky M, Lomsadze A: **Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES.** *Current Protocols in Bioinformatics* 2011, **Chapter 4**:Unit 4.6.1-10.
159. Haas B, Salzberg S, Zhu W, Pertea M, Allen J, Orvis J, White O, Buell CR, Wortman J: **Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biology* 2008, **9**:R7.
160. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20**(16):2878-2879.
161. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nature Protocols* 2012, **7**(3):562-578.
162. Elsik C, Mackey A, Reese J, Milshina N, Roos D, Weinstock G: **Creating a honey bee consensus gene set.** *Genome Biology* 2007, **8**(1):R13.
163. Haas BJ, Delcher AL, Mount SM, Wortman JR, Jr RKS, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Research* 2003, **31**(19):5654-5666.
164. Slater GSC, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6**:31.
165. Zerbino DR, Schulz M: **Oases: De novo transcriptome assembler for very short reads.** 2010.
166. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman S, Mungall K, Lee S, Okada HM, Qian J *et al*: **De novo assembly and analysis of RNA-seq data.** *Nature Methods* 2010, **7**(11):909-912.

167. Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nature Biotechnology* 2011, **29**(7):644-652.
168. National Centre for Biotechnology Information: **Gnomon - The NCBI eukaryotic gene prediction tool** [<http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml>] Accessed February 12 2013
169. Curwen V, Eyraas E, Andrews D, Clarke L, Mongin E, Searle S, Clamp M: **The Ensembl Automatic Gene Annotation System**. *Genome Research* 2004, **14**(5):942-950.
170. GMOD Consortium: **Generic Model Organism Database** [<http://gmod.org/>] Accessed 2012
171. Stein L: **The Sequence Ontology - Resources - GFF3** [<http://www.sequenceontology.org/gff3.shtml>] Accessed 2012
172. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite**. *Nucleic Acids Research* 2008, **36**:3420-3435.
173. Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al*: **The Pfam protein families database**. *Nucleic Acids Research* 2012, **40**:D290-D301.
174. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S *et al*: **InterPro in 2011: new developments in the family and domain prediction database**. *Nucleic Acids Research* 2012, **40**:D306-D312.
175. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0** [<http://www.repeatmasker.org>] Accessed February 2 2012
176. Smit AFA, Hubley R: **RepeatModeler Open-1.0** [<http://www.repeatmasker.org>] Accessed February 2 2012
177. Lowe TM, Eddy SR: **tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence**. *Nucleic Acids Research* 1997, **25**:0955-0964.
178. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR *et al*: **Rfam: updates to the RNA families database**. *Nucleic Acids Research* 2009, **37**:D136-D140.
179. **Genome Browser** [http://en.wikipedia.org/wiki/Genome_browser] Accessed August 2 2012
180. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The Generic Genome Browser: A Building Block for a Model Organism System Database**. *Genome Research* 2002, **12**(10):1599-1610.
181. Mungall C: **go-perl-0.13** [<http://search.cpan.org/~cmungall/go-perl/>] Accessed July 12 2012
182. Marín RM, Vaníček J: **Efficient use of accessibility in microRNA target prediction**. *Nucleic Acids Research* 2011, **39**:19-29.
183. Cutter AD, Wasmuth JD, Blaxter ML: **The Evolution of Biased Codon and Amino Acid Usage in Nematode Genomes**. *Molecular Biology and Evolution* 2006, **23**(12):2303-2315.
184. Roberts A, Pimentel H, Trapnell C, Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq**. *Bioinformatics* 2011, **27**:2325-2329.
185. Holt C: **[maker-devel] training SNAP with ests and cegma proteins - Google Groups** [<https://groups.google.com/d/msg/maker-devel/wbILWRVQ7r0/K9jngyzMLuEJ>] Accessed February 4 2012
186. **The DREAM Project - Alternative Splicing Challenge** [<http://www.the-dream-project.org/result/alternative-splicing>] Accessed December 10 2012
187. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent JJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**:1321-1325.
188. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K *et al*: **Highly conserved non-coding sequences are associated with vertebrate development**. *PLoS Biology* 2005, **3**:e7.
189. McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G: **Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis**. *Genome Research* 2006, **16**:451-465.
190. Ritter DL, Li Q, Kostka D, Pollard KS, Guo S, Chuang JH: **The Importance of Being Cis: Evolution of Orthologous Fish and Mammalian Enhancer Activity**. *Molecular Biology and Evolution* 2010, **27**:2322-2332.

191. Vavouri T, Walter K, Gilks W, Lehner B, Elgar G: **Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans.** *Genome Biology* 2007, **8**:R15.
192. Vavouri T, Lehner B: **Conserved noncoding elements and the evolution of animal body plans.** *BioEssays* 2009, **31**:727-735.
193. Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes.** *BMC Genomics* 2004, **5**:99.
194. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET *et al*: **Conserved noncoding sequences are selectively constrained and not mutation cold spots.** *Nature Genetics* 2005, **38**:223-227.
195. Dermitzakis ET, Reymond A, Antonarakis SE: **Conserved non-genic sequences - an unexpected feature of mammalian genomes.** *Nature Reviews Genetics* 2005, **6**:151-157.
196. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D: **Three Periods of Regulatory Innovation During Vertebrate Evolution.** *Science* 2011, **333**:1019-1024.
197. Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS: **Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing.** *Genome Research* 2005, **15**:800-808.
198. Kumar S: **Evolutionary dynamics of conserved non-coding DNA elements: Big bang or gradual accretion?** *MSc Informatics Thesis.* Edinburgh: University of Edinburgh; 2007. [https://publish.inf.ed.ac.uk/publications/thesis/online/IM070463.pdf]
199. Griffiths-Jones S: **Annotating Non-Coding RNAs with Rfam.** *Current Protocols in Bioinformatics* 2005, **9**:12.15.11-12.15.12.
200. Swindell SR, Plasterer TN: **SEQMAN.** In: *Sequence Data Analysis Guidebook.* vol. 70: Humana Press; 1997: 75-89.
201. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFAA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED *et al*: **Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner.** *Genome Research* 2004, **14**:708-715.
202. NCBI: **BLASTCLUST - BLAST score-based single-linkage clustering** [<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>] Accessed June 11 2007
203. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**:2688-2690.
204. Woolfe A, Elgar G: **Organization of Conserved Elements Near Key Developmental Regulators in Vertebrate Genomes.** In: *Advances in Genetics.* vol. 61; 2008: 307-338.
205. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negro N, Eaton ML, Landolin JM, Bristow CA, Ma L *et al*: **Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE.** *Science* 2010, **330**(6012):1787-1797.
206. Cheng C, Yan K-K, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M *et al*: **Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data.** *PLoS Computational Biology* 2011, **7**(11):e1002190.
207. Taylor AL, Sasser JN: **Biology, identification and control of root-knot nematodes (Meloidogyne species).** Department of Plant Pathology, North Carolina State University, United States Agency for International Development.; 1978.
208. Sasser JN, Carter CC: **Overview of the International Meloidogyne Project 1975-1984.** In: *An Advanced treatise on Meloidogyne.* Edited by Barker KR, Carter CC, Sasser JJ. Raleigh, NC: Dept. of Plant Pathology. North Carolina State Univ; 1985: 19-24.
209. Dalmaso A, Bergé JB: **Enzyme polymorphism and the concept of parthenogenetic species, exemplified by Meloidogyne.** In: *Concepts in Nematode Systematics.* Edited by Stone AR, Platt HM, Khalil LF. London: Academic Press; 1983: 187-196.
210. Triantaphyllou AC: **Cytogenetics, cytotaxonomy and phylogeny of root-knot nematodes.** In: *An Advanced treatise on Meloidogyne.* Edited by Barker KR, Carter CC, Sasser JJ: Dept. of Plant Pathology. North Carolina State Univ; 1985: 113-126.
211. Hugall A, Stanton J, Moritz C: **Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic Meloidogyne.** *Molecular Biology and Evolution* 1999, **16**:157-164.
212. Castagnone-Sereno P: **Genetic variability and adaptive evolution in parthenogenetic root-knot nematodes.** *Heredity* 2006, **96**:282-289.
213. Mallet J: **Hybrid speciation.** *Nature* 2007, **446**:279-283.

214. Arnold ML: **Natural Hybridization and Evolution**. Oxford: Oxford University Press; 1997.
215. Handoo ZA, Nyczepir AP, Esmenjaud D, van der Beek JG, Castagnone-Sereno P, Carta LK, Skantar AM, Higgins JA: **Morphological, Molecular, and Differential-Host Characterization of *Meloidogyne floridensis* n. sp. (Nematoda: Meloidogynidae), a Root-Knot Nematode Parasitizing Peach in Florida**. *Journal of Nematology* 2004, **36**:20-35.
216. Jeyaprakash A, Tigano MS, Brito J, Carneiro RMDG, Dickson DW: **Differentiation of *Meloidogyne floridensis* from *M. arenaria* using high-fidelity PCR amplified mitochondrial AT-rich sequences**. *Nematopica* 2006, **36**:1-12.
217. Tigano M, Carneiro R, Jeyaprakash A, Dickson D, Adams B: **Phylogeny of *Meloidogyne* spp. based on 18S rDNA and the intergenic region of mitochondrial DNA sequences**. *Nematology* 2005, **7**(6):851-862.
218. Holterman M, Karssen G, van den Elsen S, van Megen H, Bakker J, Helder J: **Small subunit rDNA-based phylogeny of the Tylenchida sheds light on relationships among some high-impact plant-parasitic nematodes and the evolution of plant feeding**. *Phytopathology* 2009, **99**:227-235.
219. Judson OP, Normark BB: **Ancient asexual scandals**. *Trends in Ecology & Evolution* 1996, **11**:41-46.
220. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**:403-410.
221. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis**. *Nucleic Acids Research* 2010, **38**:D196-D203.
222. Kim T: **QuickParanoid - A Tool for Ortholog Clustering** [<http://pl.postech.ac.kr/QuickParanoid>] Accessed October 2 2011
223. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J *et al*: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Molecular Systems Biology* 2011, **7**:539.
224. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends in Genetics* 2000, **16**:276-277.
225. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language**. *Bioinformatics* 2004, **20**:289-290.
226. Blok VC, Powers TO: **Biochemical and molecular identification**. In: *Root-knot nematodes*. Edited by Perry RN, Moens M, Starr JL. Wallingford: CABI; 2009: 98-118.
227. Seehausen O: **Hybridization and adaptive radiation**. *Trends in Ecology & Evolution* 2004, **19**:198-207.
228. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E *et al*: **A high-resolution map of human evolutionary constraint using 29 mammals**. *Nature* 2011, **478**:476-482.
229. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ *et al*: **Symbiosis insights through metagenomic analysis of a microbial consortium**. *Nature* 2006, **443**:950-955.
230. Zhang X, Davenport KW, Gu W, Daligault HE, Munk CC, Tashima H, Reitenga K, Green LD, Han CS: **Improving genome assemblies by sequencing PCR products with PacBio**. *BioTechniques* 2012, **53**(1):61-62.
231. Bashir A, Bansal V, Bafna V: **Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance**. *BMC Genomics* 2010, **11**:385.
232. The FASTG Format Specification Working Group: **FASTG** [<http://fastg.sourceforge.net/>] Accessed February 20 2013
233. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2012**. *Nucleic Acids Research* 2012, **40**:D84-D90.
234. Peng Y, Leung HCM, Yiu SM, Chin FYL: **Meta-IDBA: a de Novo assembler for metagenomic data**. *Bioinformatics* 2011, **27**:i94-i101.
235. Namiki T, Hachiya T, Tanaka H, Sakakibara Y: **MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads**. *Nucleic Acids Research* 2012:gks678.
236. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: Quality Assessment Tool for Genome Assemblies**. *Bioinformatics* 2013:btt086.

237. Vezzi F, Narzisi G, Mishra B: **Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathon**. *arXiv* 2012, **1210.1095v1** [q-bio.GN].
238. Raymond ES: **The Art of UNIX Programming**: Addison-Wesley Professional; 2003.
239. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biology* 2010, **11**(8):R86-R86.
240. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services**. *Nucleic Acids Research* 2006, **34**(Web Server issue):W729-732.
241. Halbritter F, Vaidya HJ, Tomlinson SR: **GeneProf: analysis of high-throughput sequencing experiments**. *Nature Methods* 2011, **9**(1):7-8.
242. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J: **Galaxy CloudMan: delivering cloud compute clusters**. *BMC Bioinformatics* 2010, **11**(Suppl 12):S4.
243. Eilbeck K, Moore B, Holt C, Yandell M: **Quantitative measures for the management and comparison of annotated genomes**. *BMC Bioinformatics* 2009, **10**(1):67.
244. Mungall C, Emmert D, The FlyBase Consortium: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information**. *Bioinformatics* 2007, **23**(13):i337-i346.
245. **BioSQL** [<http://www.biosql.org/>] Accessed February 20 2013

Appendix A: Workflows and scripts

Chapter 2: Assembling four nematode genomes

All the workflows and scripts are available at <http://github.com/sujaikumar/assemblage> and are listed here:

Workflows:

- Make a taxon-annotated GC cov "blob" plot
- Sub-sample sequences at random for test read sets
- Adapter- and quality-trim Illumina fastq reads using sickle and scythe in one command with no intermediate files
- Create a preliminary assembly using ABySS
- Create a preliminary assembly using CLC
- Map reads to an assembly to get insert-size and coverage information using Bowtie 2
- Map reads to an assembly to get insert-size and coverage information using CLC
- Assign high-level taxon ids to contigs
- Make blobology plots with R
- Make taxon-specific blast databases
- Separate contigs based on which taxon-specific blast database they hit better
- Extract reads (and their pairs) that map to a set of desired contigs in the preliminary assembly
- Find mate-paired reads that do not map to the ends of contigs and use those to conservatively scaffold the contigs
- Run khmer to digitally normalise reads
- Assemble RNA-Seq reads using SOAPdenovo-Trans
- Align ESTs and protein sequences to genome assemblies to assess assembly contiguity and completeness

Scripts:

- `blast_separate_taxa.pl` - Takes two tabular blast files as input (same query sequences, different databases), and assigns query sequences to one or the other or both based on parameters.
- `blast_taxonomy_report.pl` - Takes tabular blast results to any of the NCBI preformatted databases, and returns a report specifying which taxa had hits for each query.

- `blastm8_filter.pl` - Selects only those tabular blast hits that meet specified criteria of length, percentage identity, query or database sequence coverage. Can select topmost hits, and can combine HSPs in a hit while calculating coverage.
- `blobology.R` - R script for making GC-coverage blob plots from tabular data
- `bowtie2_extract_reads_mapped_to_specific_contigs.pl`
- `clc_cas_to_sspace_tab.bash` - Converts CLC's CAS mapping format to the TAB format used by SSPACE.
- `clc_len_cov_gc_insert.pl` - Generates preliminary assembly info based on mapping reads back to assembly using CLC
- `fastaqual_multiline_to_singleline.pl` - Converts multiline fasta files to single line fasta files (i.e. two lines for each entry - one line for the fasta header, and one for the sequence).
- `fastaqual_select.pl` - Selects entries from a fasta file based on length, order, include lists, exclude lists, specified intervals, regular expressions, etc.
- `fastq2fasta.pl`
- `khmer_re_pair.pl` - Uses original interleaved fasta file to pull out the pairs for each read in a khmer output file.
- `multi2single` - converts blocks of lines (-l number of lines) into tab separated columns
- `plot_insert_freq_txt_binned.R` - R script for plotting insert-length data files.
- `sam_len_cov_gc_insert.pl` - Generates preliminary assembly info based on mapping reads back to assembly in SAM/BAM format.
- `scaffold_stats.pl` - Calculates metrics for scaffolds, contigs, and runs-of-Ns in a genome assembly fasta file.
- `seq_st_en_merge_overlapping.pl` - Takes sequence names and start-stop intervals, merges overlapping intervals, and returns only non-overlapping intervals.
- `shuffleSequences_fastx.pl` - Interleaves forward and reverse reads into one file
- `unshuffleSequences_fastx.pl` - Separates forward and reverse reads from an interleaved file.

Chapter 3: Annotating nematode genomes

All the workflows and scripts are available at <http://github.com/sujaikumar/assemblage> and are listed here:

Workflows:

- Predict genes using a two-pass (iterative) MAKER2 workflow
- Run CEGMA to generate a SNAP HMM
- Run GeneMark

- Run MAKER2 first pass and second pass
- Extract AED < 1 genes only from MAKER2 output
- Standardise nematode genomes and annotations
- Calculate gene prediction metrics
- Add functional annotations to a genome
- Convert InterProScan annotations to a heatmap
- Convert tRNA annotations to a heatmap

Chapter 4: Lack of deeply conserved non-coding elements in nematodes

Workflows:

- Identify CNEs using whole-genome-alignments
- Identify CNEs using MegaBLAST and clustering

Scripts:

- interval_mask.pl - Masks (to uppercase or lowercase) an input fasta file based on a file of intervals (specified as "seqid \tstart \tend" or "seqid_start_end")
- maf_insert_lc.pl - Takes a MAF alignment file and a fasta input file. Replaces sequence in MAF file with the corresponding substring extracted from the fasta file.
- maf_remove_lc.pl - Takes a MAF alignment file as input. Converts all rows in an alignment column to lowercase even if only one row has a lowercase base in that column. Removes lowercase columns to split alignment blocks. New alignment scores for each split block are scaled by the length of the split block relative to the original alignment block.
- maf_select.pl - Takes a MAF alignment file as input. Selects only those alignment columns that meet the length, absolute identity, and relative identity cutoffs.
- maf_to_gff.pl - Takes a MAF alignment file as input. Outputs a GFF file with coordinates of each block for each species in the alignment (assumes species id is a prefix to each sequence id)
- link_blast.pl - Alternative to BLASTCLUST. Takes tabular blast output and performs single linkage clustering.
- fgrep.pl - Alternative to fgrep (grep with patterns specified in a file). Allows -v option to remove entries from original file.

Chapter 5: The *Meloidogyne floridensis* genome reveals complex hybrid origins of the root-knot nematodes

Workflows:

- Extract protein-coding CDSs from *M. floridensis* using exonerate
- Calculate and plot CDS self-identity
- Cluster proteins using InParanoid and QuickParanoid
- Use RAxML and APE to create and analyse multiple phylogenies

Appendix B: Data files

Chapter 2: Assembling four nematode genomes

Caenorhabditis sp. 5

Raw reads: <http://www.ebi.ac.uk/ena/data/view/ERP001495>

Assemblies: <http://csp5.nematod.es>

M. floridensis

Raw reads: <http://www.ebi.ac.uk/ena/data/view/ERP001338>

Assemblies: <http://meloidogyne.nematod.es>

D. immitis

Raw reads: <http://www.ebi.ac.uk/ena/data/view/ERP000699>

Assemblies: <http://dirofilaria.nematod.es>

L. sigmodontis

Raw reads: <http://www.ebi.ac.uk/ena/data/view/ERP001496>

Assemblies: <http://litomosoides.nematod.es>

Chapter 3: Annotating nematode genomes

All data files are available at

<http://dx.doi.org/10.6084/m9.figshare.96089>

- 20 nematode genomes (nucleotide fasta files)
- 20 nematode proteomes (protein fasta files)
- 20 nematode coding.gff files with transcript names
- 20 Blast2GO annotation files for each nematode proteome
- 20 proteomes with InterProScan annotations
- tRNA counts for 20 nematode genomes
- tRNA locations for 20 nematode genomes (GFF format)
- Rfamscan output for 20 nematode genomes (GFF format)

Chapter 4: Lack of deeply conserved non-coding elements in nematodes

- Whole-genome multiple alignment files for specific nodes in the nematode phylogeny: Clade III, Onchocercidae, Clade IV, *Meloidogyne*, Clade V, *Caenorhabditis*, *Elegans* group
- CNE multiple alignment files for specific nodes in the nematode phylogeny (whole-genome multiple alignments with coding regions removed)

- Tab delimited files with length and relative identity for each CNE
- Pairwise MegaBLAST alignments for all 20 genomes
- MegaBLAST based clusters of CNEs

Chapter 5: The *M. floridensis* genome reveals complex hybrid origins of the root-knot nematodes

- Protein sets used for *M. hapla*, *M. incognita*, and *M. floridensis* after truncating at stop codons and filtering short proteins (protein fasta files)
- CDS transcript files corresponding to proteins in *M. hapla*, *M. incognita*, and *M. floridensis* (nucleotide fasta files)
- Tab-delimited file with self-identity scores for each CDS in each species
- InParanoid results (pair-wise clustering)
- QuickParanoid results (orthologous clusters across three species)
- Phylogenetic trees for each QuickParanoid cluster

Appendix C: Publications

The following six papers were published as part of my PhD research. Excerpts or ideas from these papers were used in this thesis as indicated:

1. Kumar S, Blaxter M: **Comparing de novo assemblers for 454 transcriptome data.** *BMC Genomics* 2010, **11**:571.
This paper first demonstrated the utility of cumulative contig length curves (Chapter 2). The paper also described the use of EST and protein sequence alignments to compare competing assemblies. The transcriptome assembly of *L. sigmodontis* is described here; this assembly was used in both Chapters 3 and 4 of the thesis.
2. Kumar S, Blaxter ML: **Simultaneous genome sequencing of symbionts and their hosts.** *Symbiosis* 2011, **55**(3):119-126.
This paper first described the use of TAGC plots ("blob" plots) for visualising DNA sample composition, read separation, and genome re-assembly (Chapter 2).
3. Kumar S, Schiffer PH, Blaxter M: **959 Nematode Genomes: a semantic wiki for coordinating sequencing projects.** *Nucleic Acids Research* 2012, **40**:D1295-D1300.
The 959 Nematode Genomes wiki was announced in this paper. The paper also described the suitability of the Semantic MediaWiki platform for the 959NG community collaboration effort. Chapter 1 used excerpts from this paper.
4. Kumar S, Koutsovoulos G, Kaur G, Blaxter M: **Toward 959 nematode genomes.** *Worm* 2012, **1**:42-50.
This paper reviewed nematode genome projects and excerpts were used in Chapter 1.
5. Blaxter M, Kumar S, Kaur G, Koutsovoulos G, Elsworth B: **Genomics and transcriptomics across the diversity of the Nematoda.** *Parasite Immunology* 2012, **34**(2-3):108-120.
This paper reviewed nematode genome and transcriptome projects. Excerpts from this paper were used in Chapter 1 to describe the diversity of the Nematoda.
6. Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S *et al*: **The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets.** *The FASEB Journal* 2012:fj.12-205096.
Chapter 2 describes improvements over the version of the *D. immitis* genome that was published in this paper.

I also helped analyse data for four other papers that were not related to this thesis:

7. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML: **Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines.** *Genome Research* 2009, **19**(7):1195-1201.
8. Ferguson L, Lee S, Chamberlain N, Nadeau N, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, Kee T-J *et al*: **Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus.** *Molecular Ecology* 2010, **19**:240-254.
9. Hunt P, Martinelli A, Modrzynska K, Borges S, Creasey A, Rodrigues L, Beraldi D, Loewe L, Fawcett R, Kumar S *et al*: **Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites.** *BMC Genomics* 2010, **11**:499.
10. Wang J, Mitreva M, Berriman M, Thorne A, Magrini V, Koutsovoulos G, Kumar S, Blaxter M, Davis R: **Silencing of Germline-Expressed Genes by DNA Elimination in Somatic Cells.** *Developmental Cell* 2012, **23**(5): 1072-1080.