THE UNIVERSITY
of EDINBURGH

Reconstructing the phylogenetic relationships of nematodes using draft genomes and transcriptomes

Georgios D. Koutsovoulos

Doctor of Philosophy

Institute of Evolutionary Biology

School of Biological Sciences

University of Edinburgh

2015

Declaration

I declare that this thesis is my own work, and that the work described here is my own except where explicitly stated. This work has not been submitted for any other degree or professional qualification.

Georgios D. Koutsovoulos, May 2015

Abstract

Nematoda is a very diverse animal phylum. Within Nematoda, species display a multitude of life styles, different reproductive strategies and parasitism has arisen independently several times. Furthermore, morphological conservation and a high rate of homoplasy have impeded the resolution of nematode systematics. To address these issues, single gene (usually the nuclear ribosomal small subunit gene) and mitochondrial gene phylogenies have been used, but the information contained within the sequence of these genes is not enough to resolve the topological relationships between clades that emerged during rapid cladogenesis.

Next generation sequencing data have been shown to produce high quality genomic and transcriptomic assemblies at low cost, as a result more and more nematode species are being sequenced. Sequences were gathered or generated for 53 nematode species from ESTs, gene predictions from full genome assemblies and transcripts from RNA-Seq experiments. These sequences were screened for orthologous gene clusters, which were concatenated into a supermatrix with thousands of aminoacid sites. The analysis of the supermatrix with maximum likelihood and Bayesian inference methods sheds light into the early splitting clades of the phylogenetic tree of nematodes and the derived clades III, IV and V. Furthermore, the phylogenetic relationships within the parastitic family Onchocercidae were resolved, unveiling the evolutionary history of these important taxa. Finally, data produced in this work will be useful for subsequent evolutionary studies of the phylum Nematoda.

Lay Summary

Nematodes are small worms that are present in almost all environments of our planet. Experts think that only a small fraction of nematode species has been identified and categorised, while most of the species remain to be discovered. Some of the species are free-living, while others are parasites of plants and animals. Human-parasitic nematodes infect more than one billion people worldwide, while cow-parasitic nematodes cause problems in rural areas, where these animals are crucial for day to day survival. This PhD aims to further our knowledge of the different types of nematodes, so that we can more effectively understand and control them. Therefore, a major part of the project was spent in generating the genomic and transcriptomic sequences of these species. With this information, I produced a more robust phylogenetic tree. This allows us to establish an evolutionary background for the different types of nematodes. Thus, the underlying processes of the evolution of parasitism and of speciation events can now be explored in a defined framework.

Acknowledgements

Firstly, I would like to thank my supervisor Mark Blaxter for the guidance and encouragement he has provided during my PhD programme. I was extremely lucky to have a supervisor who cared about my work providing advice and giving me the freedom to explore emerging hypotheses. His attitude about life has inspired me to be a better person and a better scientist.

Many thanks go to my second supervisor Andrew Rambaut and my postgraduate committee member Nick Colegrave for their advise and comments over the years. I would also like to thank the people of the Institute of Evolutionary Biology for the interesting scientific discussions over coffee or beer.

Thank you to past members of the Blaxter Lab, Ben Elsworth, Graham Thomas, Jack Hearn, John Davey, Michael Clarke and Pablo Fuentes, and I hope to work with you again in the future. Thank you to present members, Beate Nurnberger and Lewis Stevens, and especially Dominik Laetsch for keeping me sane during thesis writing and having a keen eye for grammatical errors. Special thanks to Sujai Kumar who belongs in both categories (your PERL one-liner expertise is greatly needed).

I thank the people from Edinburgh Genomics for providing the data needed for this study and especially the Bioinformatics team for the countless hours of discussion about bioinformatics algorithms and programming languages.

This study would not be possible without the external collaborators Byron Adams, Bernadette Connolly, Richard Davis, Benjamin Makepeace, Pascal Maser, Kenneth Pfarr, Einhard Shierenberg, Andre Pires da Silva, Vincent Tanya and Itai Yanai. Special thanks to Philipp Schiffer for allowing me to use data from multiple species for my phylogenetic analyses and endless discussions about phylogenetic networks (and sending chocolate to keep me awake).

I am very grateful to the BBSRC and the School of Biological Sciences for funding my PhD programme and providing funds for conference travel.

Completing this work would have been all the more difficult were it not for the support of my friends both in Greece and Scotland. I would also like to thank my flatmates Thanasis, Iraklis and Sakis for all the philosophical discussions about our place in the universe.

Thank you to my partner Rodanthi for the endless support and encouragement, especially during the last months. Your smile kept me going.

And finally, I would like to thank my parents, Dimitris and Evgenia, and my sisters, Vasiliki and Katerina, for all the support throughout my life. Without you this journey would not have been possible.

Contents

List of Figures

List of Tables

Acronyms

aa	amino acids
b	bases
bp	base pairs
BI	Bayesian Inference
CEGMA	Core Eukaryotic Genes Mapping Approach
Dayhoff	Dayhoff substitution matrix
ENA	European Nucleotide Archive
EST	Expressed Sequence Tag
GTR	General-Time Reversible substitution model
HMM	Hidden Markov Model
HSP	High-scoring Segment Pair
JTT	Jones, Taylor, and Thornton substitution model
kb	kilobases
kbp	kilobase pairs
LG	Le and Gascuel substitution matrix
LBA	Long Branch Attraction
M	million
Mb	megabases
Mbp	megabase pairs
MCMC	Markov chain Monte Carlo

ML	Maximum Likelihood
m	metres
mm	millimetres
μm	micrometres
MP	Mate Pair
MxP	Maximum Parsimony
MSA	Multiple sequence alignment
NCBI	National Centre for Biotechnology Information
NGS	Next-Generation Sequencing
NJ	Neighbour Joining
ORF	Open Reading Frame
PE	Paired End
Rfam	RNA families database
RNA-Seq	RNA-Sequencing
SE	Single End
SD	Standard Deviation
SH	Shimodaira-Hasegawa test
SRA	Short Read Archive
nSSU	nuclear ribosomal small subunit
WGA	Whole-Genome Amplification
WAG	Whelan and Goldman substitution matrix

Chapter 1

Introduction

1.1 Thesis structure

In this thesis I describe the generation and analysis of novel nematode genome and transcriptome data.

In this Chapter, I outline the biological and evolutionary significance of the phylum Nematoda. A historical review of the past phylogenies is described along with the outstanding questions still remaining which will be explored in this thesis.

Chapter 2 focuses on the bioinformatics aspect of the work. It details the problems associated with assembling a genome using short read technology, and the effects of choosing different parameters with the programs that were tested.

In Chapter 3, I introduce a new program that uses transcriptome evidence to scaffold regions of the genome. It also contains a section that assesses the different programs for annotating a genome and inferring the most reliable gene models.

In Chapter 4, the progress for each species in this work is detailed. Each species section contains additional information about its phylogenetic importance and a report about its status from the raw data to the final output files.

In Chapter 5, the phylogenetic analysis is described. The rationale for each step is explained and the results of the phylogenetic analysis are discussed. There are

two parts in the discussion. Firstly, I evaluate the methods used for their abilities to infer deep phylogenetic relationships. Secondly, I focus on a specific family of animal parasitic nematodes (Onchocercidae) and the implications of the analysis of this group.

The final chapter summarises the findings, and provides thoughts on possible future work based on the results presented in this thesis.

1.2 Nematodes are diverse

The phylum Nematoda is widely distributed in nearly every environment of the planet, and only a small proportion of the estimated 1 million species have been described (~26,000) [1]. Their reproductive modes vary from hermaphroditic, through multiple flavours of dioecy to asexual/parthenogenetic, and they can be categorised based on their lifestyles which range from free-living through phoretic to a fully parasitic lifestyle.

Almost every habitat (terrestrial, freshwater, and marine) is populated by free-living species, indicating a high adaptability to diverse environmental conditions. Based on the habitat and their taxonomic position, their size varies from 150 micrometres (μm) to 10 millimetres (mm) in soil and freshwater environments, and from 83 μm to 50 mm in marine environments [2]. Since different habitats contain different sources of nutrients, free-living nematodes have developed different feeding habits which can consist of algae, fungi, bacteria and even meiofaunal species (including other nematodes).

Equally, parasitic species exhibit a variety of life styles and target hosts. They are capable of infecting plants, insects, higher animals and humans, with varying degrees of pathogenicity. Furthermore, depending on the range of the possible hosts, some parasitic species are encountered in specific regions, while others are cosmopolitan. Based on the final host and their method of transmission, animal parasites range from 1 mm to 6 metres (m) in length, while plant parasites range from 0.25 mm to 12 mm in length [2, 3]. The animal parasites also differ in the complexity of their transmission, some are monoxenous, others heteroxenous, and others are secondarily monoxenous [4]. In addition, some animal parasitic species are host specific, while others can target a variety of related hosts. Displaying similar diversity, plant parasites can feed on different parts of the plant (including roots, stems, leaves, flowers and seeds), and can have various levels of host specificity.

1.3 Nematodes are important

Nematodes are important in various fields of Biology. They can be studied as an indicator group in biodiversity assessment and biomonitoring, since they are present in almost every ecosystem [5]. Long-term river pollution can be monitored using relations between nematode communities' structure and the level and source of contamination [6]. Similarly, soil quality can be assessed by counting the number of free-living nematode species in different families [7].

Human-parasitic nematodes are estimated to infect more than ~1 billion people worldwide accounting for the loss of at least 40 million disability-adjusted life years [8]. The symptoms of infection differ between different species, ranging from asymptomatic infections (pinworm, low-grade ascariasis) to river blindness and elephantiasis (filariasis). Currently, there are no effective vaccines for these parasitic diseases, and the treatments only deal with the disease while there is a constant uptake of the specified drugs. Closely related species that infect animal model organisms can be used to understand the specifics of parasite biology and help with the development of more effective vaccines and treatments [9].

Animal-parasitic nematodes inflict serious economic damage to livestock production. The strongylid parasite *Haemonchus contortus* infects flocks of sheep worldwide, and displays high pathogenicity. Resistance to antihelmintic drugs is compromising control of the nematode. A number of different species infect cattle and cause a severe damage in rural areas, where these animals are crucial for day to day survival. Almost every higher animal is susceptible to nematode infection.

A more indirect effect on human communities is the presence of plant parasitic nematodes. More than 4,000 plant-parasitic species have been described [10], and a few of them are responsible for substantially decreasing crop yields, and thus contribute to famine in developing countries. For instance, the genus *Globodera* is responsible for parasitising potatoes worldwide, and the estimated cost from yield losses in UK is £50 million per year, while the estimated cost of human agriculture

losses for all plant parasitic nematodes is £58 billion worldwide [11]. Furthermore, members of the genus *Bursaphelenchus* threaten forest ecosystems, since they are responsible for the death of pine trees.

Free-living nematodes are also important as biological models. The first metazoan with a complete sequenced genome was the bacterivorous nematode *Caenorhabditis elegans* [12]. In 2003, the complete genome of *Caenorhabditis briggsae* was determined allowing comparisons between the two genomes, advancing the understanding of the evolutionary forces that mould nematode genomes [13]. The following years, six additional *Caenorhabditis* genomes became available, *C. remanei* [14] in 2007, *C. angaria* [15], *C. brenneri* [14] and *C. japonica* [14] in 2010, *C. tropicalis* [14] in 2011, and lastly *C. sp5* [16] in 2013. Additionally, a collective effort has started to try and sequence all 38 described species of *Caenorhabditis* to further understand the diversity between members of this genus (Mark Blaxter, pers. comm.).

Knowledge of the evolutionary relationships between nematode species will permit the effective use and control of nematodes.

1.4 History of nematode phylogeny

Nematode systematics has historically been based on morphological traits. However, the description of many nematode species is subject to the expertise of the nematologist limited by the equipment quality. As a result, some genera contain many poorly identified species and a revision of whole genera would be necessary to allow correct species identification [17]. The diversity of morphological traits has further hindered the identification process and the physical characteristics of the species have not been for the most part connected to the underpinning biology. The limited representation and emphasis on a few morphological characters made the reconstruction of nematode phylogeny a difficult process. Finally, the limited number of nematode fossils that exist lack useful information about the origin of the phylum, and therefore are of limited use in resolving phylogenetic relationships within Nematoda [18]. Many

scientists have contributed to nematode systematics over the years, and their work helped to advance the field of nematology. The most influential classifications that shaped nematode taxonomy to its current state are described below.

The first classifications were based on morphological and ecological observations. In 1922, Micoletzky grouped nematode species based on stomatal characters, identifying five different families [19]. However, these groupings proved to be artificial and were quickly challenged. Chitwood BG and Chitwood MB [20] and Chitwood BG [21] observed that nematodes with phasmids share additional characteristics and thus must represent a monophyletic group. They classified two groups within the Nematoda, Adenophorea (gland bearers) and Secernentea (secretors) [22] (Fig. 1.1a). Later in 1963, Maggenti suggested the paraphyly of Adenophorea based on aspects of pharynx structure and excretory system. He suggested that these markers can identify evolutionary relationships [23]. In 1976, Andrássy used a number of morphological characters to split Adenophorea into Torquentia and Penetrantia giving them the same rank with Secernentea, proposing three monophyletic groups [24]. Three years later, Lorenzen created the first taxonomy using cladistic principles, also suggesting the paraphyly of Adenophorea [25].

In 1998, Blaxter et al. used nuclear ribosomal small subunit (nSSU) sequences to produce the first phylogenetic framework of the Nematoda [26] (Fig. 1.1b). The authors proposed 5 major clades, confirming the paraphyly of Adenophorea, additionally observing that parasitism has evolved multiple times independently. De Ley and Blaxter [27, 28] updated nematode systematics to accommodate these changes using morphological and molecular data. In 2006, a 12-clade division was proposed by Holterman et al. using phylogenetic methods for more than three hundred nearly full-length nSSU sequences from a wide range of nematode species [29] (Fig. 1.1c). The nSSU phylogenetic trees from [26] and [29] are congruent. The additional information from sampling more taxa was only able to provide a more clear subdivision of Adenophorea. Since 2006, a large number of molecular-based phylogenetic studies using nSSU sequences have been conducted (e.g. [5, 30, 31, 32, 33]). In 2009,

van Megen et al. analysed nSSU sequences from ~1,200 taxa producing the most species-complete phylogenetic tree [34]. Their findings were consistent with the 12-clade division, noting however that a multigene approach is needed for resolving deep phylogenetic relationships.

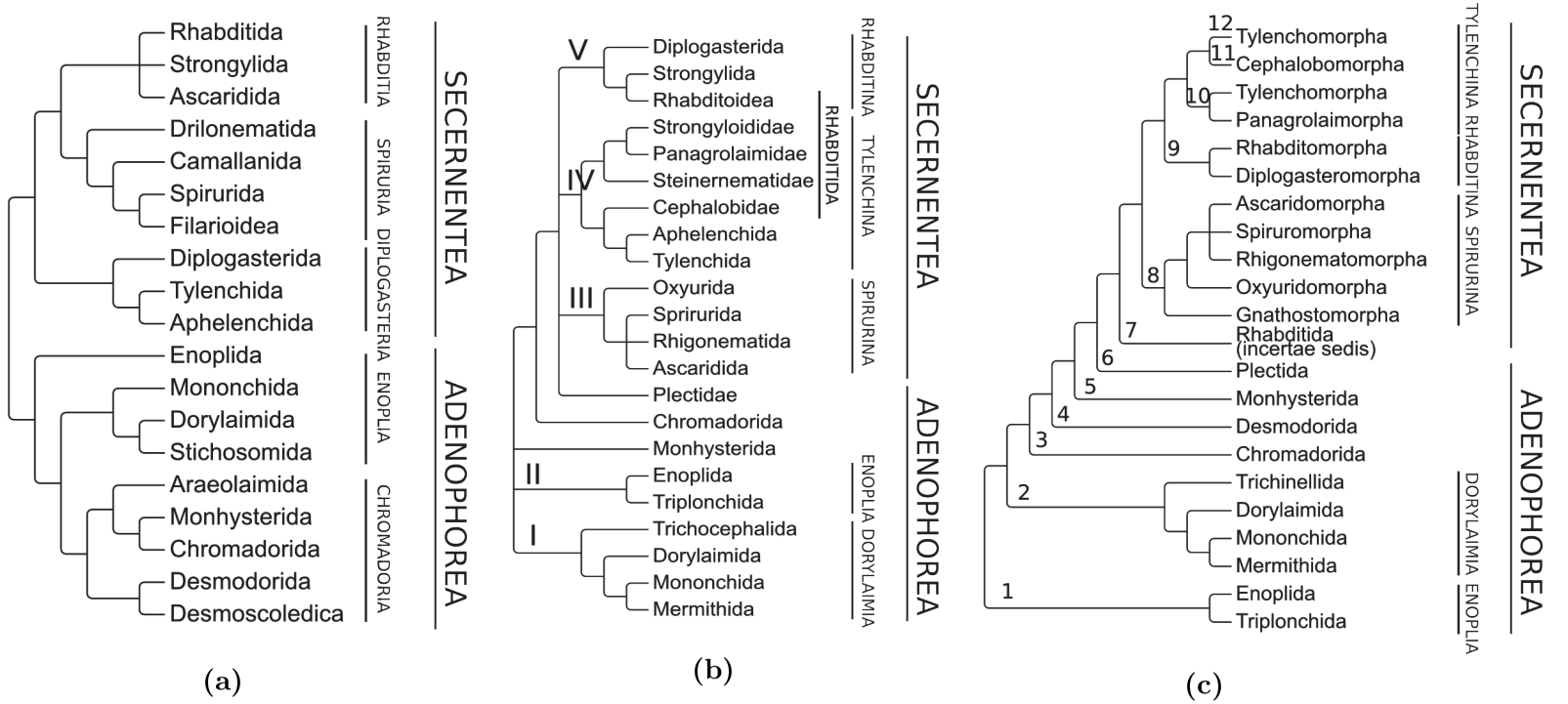


Figure 1.1 Phylogenetic relationships of the phylum Nematoda, (a) Chitwood BG and Chitwood MB (1937) divided the tree into two classes [21], (b) Blaxter et al. (1998) identified five major clades [26], and (c) Holterman et al. (2006) proposed a division into twelve clades [29].

1.5 Multigene approach

There are two approaches that may achieve a better resolution of the nematode tree. The first approach is to add more taxa in the phylogenetic analysis. The phylogenetic tree by van Megen et al. [34] contains data from ~1,200 species. An updated version of that tree containing more than 3,000 taxa and covering most of the described diversity of this phylum is in preparation (Johannes Helder, pers. comm.). These phylogenetic trees are inferred using a single gene, nSSU. Even though a large number of taxa are being used, deep phylogenetic issues are likely to remain, since the single gene approach based on nSSU sequences suffers from problems such as functionally constrained diversity and different substitution patterns [35]. Phylogenies based on a single, highly conserved gene can have additional problems, as the history of the specific gene may differ from the species phylogeny and informative characters may be hard to obtain. Since the number of nematode taxa will soon exceed the number of nucleotide sites in the nSSU alignment (i.e. in [34] 2,967 aligned positions including gaps for 1,225 taxa), the resolutive power of nSSU has been reached.

The second approach is to add more genes to the analysis, and sample multiple genes from phylogenetically informative taxa. Using multiple genes will eliminate random errors and reduce noise when constructing phylogenetic relationships, and multigene approaches have been able to resolve deep animal relationships (e.g. [36, 37]). However, multigene approaches that use concatenated datasets are prone to misleading node support and erroneous identified orthologous genes [38]. Until now, most multigene datasets were constructed using Expressed Sequence Tags (ESTs), which can be produced by sequencing one or both inserts of a transcribed cDNA sequence. Advances in genome and transcriptome sequencing (NGS) can provide nearly complete datasets containing the proteome of a species [39]. Thus, the extensive dataset that multigene approaches need can be derived from data generated by the new sequencing techniques. Compared to EST datasets where the problem was the amount of missing data [36], NGS approaches will yield more informative data.

Thus, the effect of missing data is smaller [40].

It is however hard to sample nematode species across the whole phylum which will aid the phylogenetic analysis due to their position. These phylogenetically informative species are under-represented since most nematodes being studied are biased towards animal and plant parasites due to their significance, and also towards members of the Clade V due to their proximity to the model organism *C. elegans*. Although the whole diversity of the Nematoda cannot be sampled, there is a good representation of taxa in the derived clades III, IV and V, and a basic representation of taxa in the basal clades I and II. For most of the species in this study, the genome was sequenced to have a more complete proteome from the species. In some cases, genomic sequences were not available and thus only the transcriptome was used.

Finally, outgroup choice can severely influence tree topology and can be a determining factor in resolving controversial phylogenies [41]. As outgroup for the phylogenetic analyses three evolutionary closest phyla were chosen (Nematomorpha, Tardigrada, and Arthropoda). Three arthropod species with available NGS data were chosen (*Drosophila melanogaster*, *Bombyx mori* and *Tetranychus urticae*). NGS data from the tardigrade *Hypsibius dujardini* and the nematomorph *Gordius* were generated for this study, since no genomic or transcriptomic datasets were available for this two phyla. Unfortunately, sequences from other close phyla (e.g Onychophora, Priapulida) were not available.

1.6 Phylogenetic and NGS status

Fig. 1.2 shows the current phylogenetic structure based on multiple nSSU phylogenies described previously. Phylogenetic analyses using the nSSU gene recovered many traditional monophyletic groups within Nematoda (e.g. Dorylaimida, Rhabditida) and identified new relationships between taxonomic groups (e.g. the Plectida and Rhabditida as sister clades). However, the limitations of the nSSU marker can be seen as the phylogenetic tree contains multiple polytomies. Nematode nSSU se-

quences do not contain enough phylogenetic signal to resolve deep nodes and groups with short branches [5].

Phylogenetically informative nematode taxa were sampled and added to the existing published datasets to answer these questions,

- ***Relationships between the three classes.***

  The division of the Nematoda in three monophyletic clades, Dorylaimia (Clade I or Clade 2), Enoplia (Clade II or Clade 1) and Chromadoria (Group C or Clades 3-7, and Clades III-V or Clades 8-12) has been supported by previous nSSU phylogenies. However, the relationships between the three clades are not clear. Both the Dorylaimia and the Enoplia have been proposed as the earliest splitting lineage. To resolve the topology at the base of the Nematoda, one taxa was sampled from the Enoplia and two taxa from the Dorylaimia.

  Embryological and morphological data suggest that Enoplia is the most basal clade of the phylum Nematoda [42]. However, single-gene phylogenies have failed to robustly support this topology. The resolution of the basal node may provide information for the habitat of the ancestral nematode, since the Enoplia and the Dorylaimia mostly occupy different ecological niches.

- ***Relationships between Rhabditida suborders.***

  The order Rhabditida is divided in three suborders, the Spirurina (Clade III or Clade 8), the Tylenchina (Clade IV or Clades 10-12), and the Rhabditina (Clade V or Clade 9). Although almost all sequencing projects are targeted towards species from this order, the relationships between the suborders remains unresolved. Previous nSSU phylogenies have hinted that the suborder Spirurina is the earliest splitting lineage, but a recent multigene study suggested that the Tylenchina diverged first [43]. To resolve the relationships within the order Rhabditida, two taxa were sampled from the sister order Plectida.

  More than 50% of the named nematode species belong to the order Rhabditida

[44], with most species being parasitic. The evolutionary relationships between the suborders will shed light in our knowledge about the evolution of parasitism.

- ***Relationships between Onchocercidae genera.***

  The family Onchocercidae contains most of the important parasites of humans and livestock. nSSU sequences do not have enough signal to resolve the phylogenetic positions of the genera within the family. Until recently, it was thought that the endosymbiont Alphaproteobacterium *Wolbachia pipientis* was present in nematodes only in species of the family Onchocercidae. The presence of *Wolbachia* is crucial for the survival of the species, since antibiotic treatments aimed at *Wolbachia* also kill the nematode host [45]. However, some species have since lost the *Wolbachia* symbiosis but traces can be found as *Wolbachia* nuclear insertions in the nematode genome.

  The evolutionary history of the symbiosis can be obtained by the evolutionary history of the nematode species. Effective control of this parasitic nematodes can be achieved by understanding the basis of the symbiosis. Furthermore, it will be feasible to detect if the symbiosis occurred as a single event and if loss of the symbiosis occurred multiple times independently.

- ***Relationships between Tylenchina infraorders.***

  Previous analyses have shown the paraphyly of Tylenchomorpha (i.e. superfamilies Aphelenchoidea and Tylenchoidea) and Panagrolaimorpha (i.e superfamilies Strongyloidoidea and Panagrolaimoidea). nSSU phylogenies recovered Strongyloidoidea as the earliest splitting lineage, and Aphelenchoididae as sister clade to Panagrolaimoidea. However, the latter topology was considered an artefact attributed to elevated AT-contents in both taxa [29, 27].

  To check the robustness of multigene protein datasets to underlying differences in GC within taxa, one additional species from Aphelenchoididae was sampled.

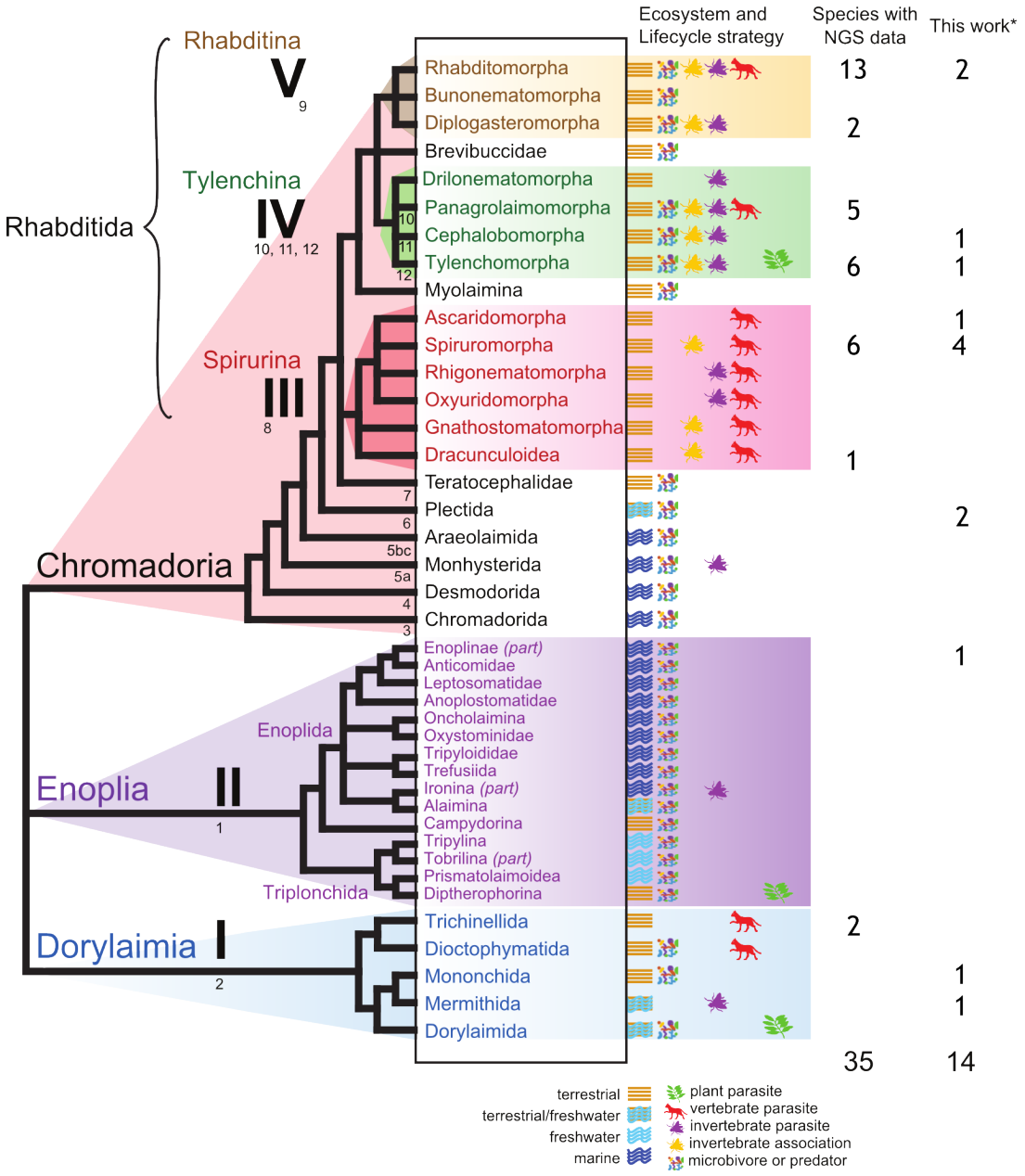


Figure 1.2 Current phylogenetic structure of the Nematoda, based on nSSU analyses. Clades I-IV were defined by Blaxter et al. [46], Clades 1-12 were defined by Holterman et al. [29]. For each group named, the ecosystem and trophic habits are indicated by small icons. For each clade the number of available NGS datasets and number of NGS datasets produced in this work are given. Figure modified from [46].
⋆or directly involved in assembly or annotation.

Chapter 2

An Introduction to Assembly

2.1 Introduction

In this chapter, the bioinformatics algorithms used to obtain the proteome of the species that had their genome or transcriptome sequenced are outlined. Raw genomic or transcriptomic reads were obtained from sequencing centres and analysed to produce the dataset used in subsequent phylogenetic analyses. Although transcriptomes are easier to assemble and use, there is the chance that parts of the proteome are missing. This is due to the fact that transcripts are specific to the particular time and locality that the sample was collected.

Most datasets were derived from genomic DNA. However, to obtain the protein sequences the whole genome has to be assembled and annotated. Difficulties during genome assembly include low quality DNA, sub optimal sequencing libraries, contamination and sample heterozygosity.

2.2 Genome Assembly

The first metazoan to have its genome assembled was the nematode *Caenorhabditis elegans* [12]. It took ten years and several labs working with Sanger sequencing [47] to complete the genome assembly into chromosome level contigs. Although

Sanger sequencing can produce high quality assemblies, the cost and resources needed prevent the extensive use of this technology for sequencing projects. Next generation sequencing (NGS) technologies offer a cheap alternative to whole genome sequencing projects.

2.2.1 Sequencing platforms

NGS technologies utilise a whole-genome shotgun technique, but without cloning the fragments into a vector [48]. The reads produced by NGS are shorter (i.e reads can be up to 300 bases (b) from Illumina machines) but the output contains many millions of reads. This results in a higher genome coverage, but the smaller read length and larger amount of reads requires computationally and memory expensive algorithms.

The reduced read length leads to an additional problem when assembling the reads into a genome. Repeated sequences can break the contiguity of the genome assembly if the length of the repeat is longer than the read length. To circumvent this problem, pairs of reads can be sequenced from opposite ends of a bigger fragment. For example, a fragment with 500 base pairs (bp) length, that has a read of 100 b from each side sequenced can resolve repeats of up to 400 bp in length.

NGS projects that use only Paired End (PE) libraries, are fragmented proportionally to the repeat content of the genome analysed. These projects result in draft assemblies with multiple genomic fragments (contigs) rather than complete chromosome assemblies. The number of contigs are typically in the range of tens of thousands, depending on the complexity of the genome sequenced. Although the genome is not finished in terms of contiguity, draft assemblies can be used in a variety of gene centric ways (phylogenetics, presence or absence of selected genes, novel gene identification etc.). Higher quality assemblies can be produced using various scaffolding methods, which are discussed in subsection 2.2.3.

Each NGS technology has a different set of advantages and disadvantages. The

most common differences are the read length, error profile and cost that makes each one more suitable for a different purpose. The Illumina platforms use the sequencing by synthesis technology. Read length can be up to 300 b and the errors are commonly miscalled bases. Roche 454 platforms use the pyrosequencing technology, read length can be up to 1,000 b and the most common errors are homopolymer length miscalls. Life Technologies' SOLiD platforms use sequencing by ligation technology, read length can be up to 60 b and the most common problem is the sequencing of palindromic regions (Table 2.1).

Table 2.1 NGS technologies

Platform	Maximum read length	Cost per million bases ($)	Error profile	Applications
Illumina Miseq	2 x 300 b	0.5	Miscalled bases (~0.5%)	Small genomes, RNA-Seq
Illumina Hiseq	2 x 150 b	0.04	Miscalled bases (~0.5%)	Big genomes
Roche GS FLX Titanium XL+	1,000 b	10	Homopolymer errors, miscalled bases (~ 0.01%)	RNA-Seq
Life Technologies SOLiD 5500xl	2 x 60 b	0.13	Sequencing of palindromic regions	SNP sequencing

2.2.2 Short Read Assembly

All the species assembled here were sequenced using Illumina platforms. Typically, each run yielded more than ~15 gigabases of raw sequencing data with an approximate 0.5% error rate. Short read assembly algorithms have to be able to work with a high load of sequencing data using minimal computer resources, deal with sequencing errors efficiently and resolve repeats smaller than the fragment length. Older algorithms used for Sanger reads were computationally not capable of using large amounts of short reads, which led to the introduction of more efficient methods. The most commonly used algorithm for assembling short-reads is the de Bruijn Graph approach [49]. Each read of length l is split into overlapping strings of a particular length k (k-mers) with $k < l$. For example, a read that is 50 b long can be split into 30 overlapping 21-mers. Each k-mer prefix and suffix is represented as a node in the graph structure, connected by an edge which represents the k-mer. Once the whole graph is constructed, the algorithm traverses each edge once, and outputs the genomic sequence. Widely used programs that use the de Bruijn Graph approach are ABySS [50], CLC-bio [51], SOAPdenovo [52], SPAdes [53], Velvet [54].

In practice, de Bruijn graph approaches are not straightforward. There are some assumptions which do not hold for next-generation sequencing, such as that the raw data contains all the k-mers present in the genome, that all k-mers contain no errors and that each k-mer appears at most once in the genome [55].

- ***Generating all k-mers present***. The reads generated by Illumina machines to a high coverage have only a fraction of the full length read-mers captured. However, by breaking the reads into overlapping shorter k-mers, nearly all k-mers will be represented. This ensures that all the k-mers of the genome can be detected and effectively used to reconstruct the genome.

- ***Handling errors in reads***. Errors in the sequenced read create bulges and tips in the de Bruijn graph. Error-corrected reads (discussed below) simplify

the de Bruijn graph construction before beginning the assembly. Bulges (alternative paths starting from the same node and ending in the same node) in the de Bruijn graph are removed after the full graph is constructed by the difference in coverage between the paths [56]. Tips (alternative paths with no end or start node) are usually clipped based on coverage and length.

- ***Handling DNA repeats***. In addition to finding all k-mers present in the genome, short-read assemblers also find the number of occurrences for each k-mer. If the true number of k-mer occurrences can be found, the assembler can create multiple directed paths and resolve the orientation of the regions that are interleaved by repeated regions in the genome. Thus the graph is resolved from the addition of multiple edges by balancing the number of indegree and outdegree of the nodes. Practically, obtaining the true count of k-mer occurrences is difficult with the current sequencing technologies. Usually, the insert size information presented in PE sequencing is used to connect sequences when the insert size is bigger than the span of the repeat. A correct traversal through the graph occurs when one read maps before the start of the repeat while its pair maps after the end of the repeat. Therefore, the contiguity of the assembly is improved and repeated regions get resolved.

- ***Handling unsequenced regions***. Unsequenced regions or regions with sequencing errors introduce breaks in the assembly. A high value of k will reduce the number of bulges in regions with high coverage and small number of errors. Low-coverage regions will still give contigs with gaps. The size of these can still be determined by the use of paired reads that can span the gap, resulting in finding the correct spacing, order and orientation of the contigs.

- ***Effect of k-mer size***. The most important parameter in a de Bruijn graph assembler is the k-mer size. Small k-mers allow more overlapping sequence while increasing the amount of ambiguity. A high k-mer has the opposite effect. Datasets with high coverage benefit from a high k-mer as this will

decrease ambiguous repeats and will result in a more contiguous assembly (Fig. 2.1)

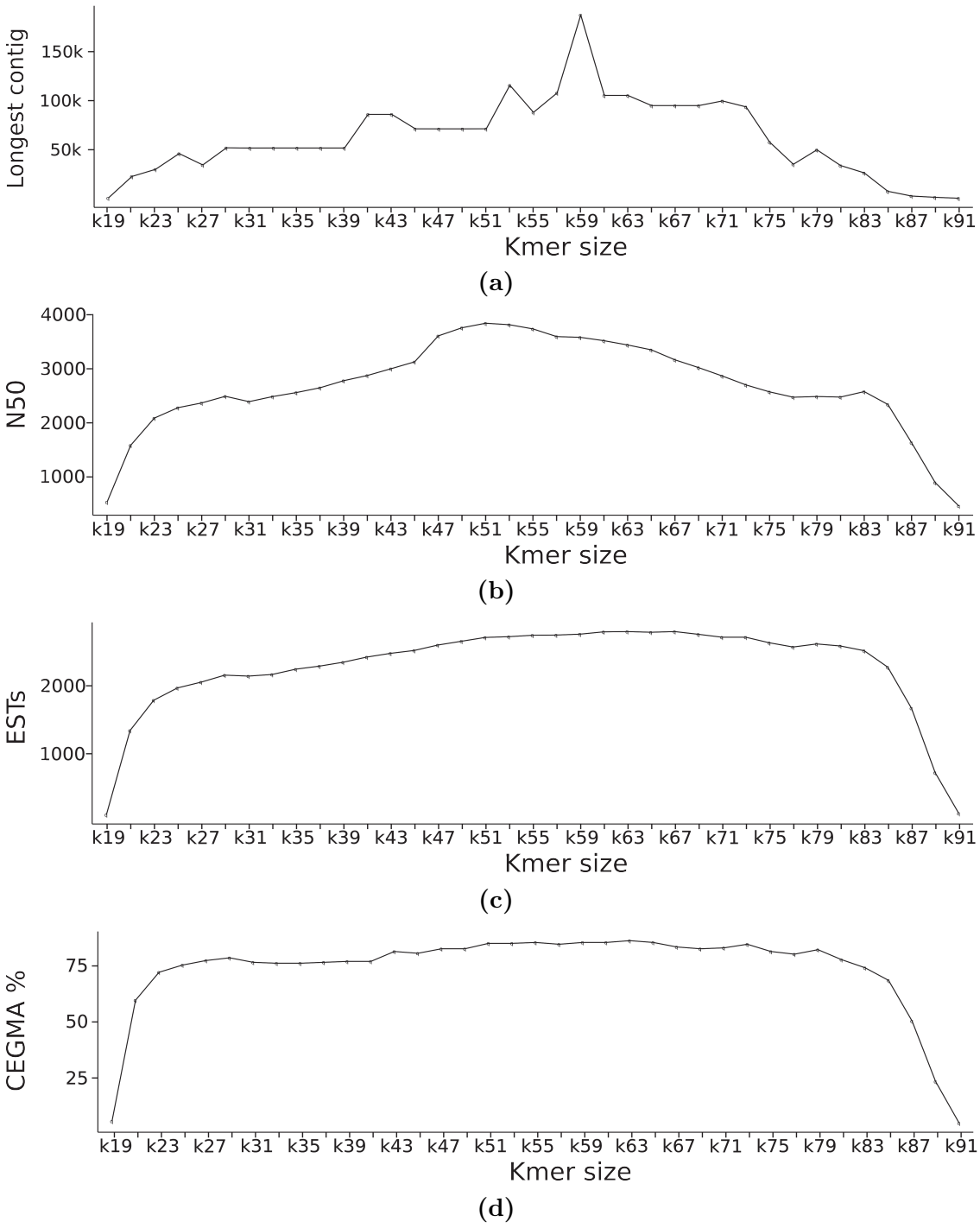


Figure 2.1 Effect of kmer size on four assembly metrics (discussed in 2.2.5.8), two numerical metrics (a) longest contig and (b) N50, and two biological metrics (c) number of ESTs present in full length and (b) CEGMA completeness. Metrics maximise at different kmer sizes, (a) k59 (b) k51 (c) k67 and (d) k63.

2.2.3 Scaffolding

Normally, PE libraries have low insert size (i.e 180 b to 600 b) and thus they can only bridge contigs when repeats are smaller than the insert size of the library. Resolving longer repeats requires the use of Mate Pair (MP) libraries with insert size from 2 kilobases (kb) to 20 kb. Although these libraries are harder to construct, they can be used to bridge regions that have repeats which are smaller than the MP insert size. MP library construction requires higher amounts of high molecular weight DNA than PE library construction, because the fragments needed for its construction have to be long (depending on the protocol, up to 20 kb). The long fragments are circularised and fragmented. Only the region around the connection of both ends is selected and sequenced. The length of the circle is the insert size of the library. Typically, an amount of PE reads are also sequenced with the MP reads which can generate misassemblies if not dealt properly.

Another approach is to use high-resolution restriction maps generated by optical mapping technologies. Single DNA molecules are immobilised and elongated onto a positively charged glass surface. Next, the DNA molecules are digested by a restriction enzyme and stained with a fluorescent dye. The cut sites are detected as gaps in DNA images fusing fluorescence microscopy, and the fragments are measured. Restriction maps for the assembly are obtained by converting the scaffolds by *in silico* restriction enzyme digestion. Matches in lengths of the sequence-derived fragments and optical fragments result into the linkage of scaffolds into super-scaffolds. Optical restriction maps vary on size depending on the quality of DNA (e.g average size of 360 kb [57]).

Two more approaches are discussed in later sections, scaffolding with long reads generated by new technologies (see subsection 2.2.4) and scaffolding with transcripts (see section 3.1).

2.2.4 Advances in NGS technology

Newer technologies have even longer reads and different error profiles, but couldn't be tested because they were at an early stage when sequencing was conducted. The most promising are PacBio (read length of up to 30 kb) and MinIon (in testing phase). The new single molecule sequencing technology PacBio can create long fragments (mean length ~8.5 kb) that can be used for scaffolding. These long reads are difficult to use for *de novo* assembly of large genomes because of high error rates (15% to 30%) and difficult error profiles (mismatches, insertions and deletions), but have been used efficiently for bridging adjacent contigs. If the error rates drop in the next years, these technologies will be the gold standard for assembling new genomes.

Whole-genome amplification (WGA) is a promising method to generate sufficient DNA for sequencing from small quantities of DNA, i.e from single nematode specimens. This will result in better assemblies from wild isolates, since pooling of multiple individuals with high heterozygosity can be avoided. Furthermore, it would eradicate the errors of misidentifying species from which the genomic DNA is pooled. WGA still needs to be tested for sequencing coverage bias and the proportion of chimeric contigs generated from the reads.

2.2.5 Typical assembly workflow

The typical workflow of a sequencing project from raw data to genome assembly consists of the following steps. Raw reads are checked for their quality. Low quality bases and adapters are trimmed. Bases are error corrected. The insert size of the library is calculated. Reads are checked for contamination and are removed. Coverage is normalised digitally if necessary, and the final assembly is produced. The final assembly is further refined by additional information (Fig. 2.2).

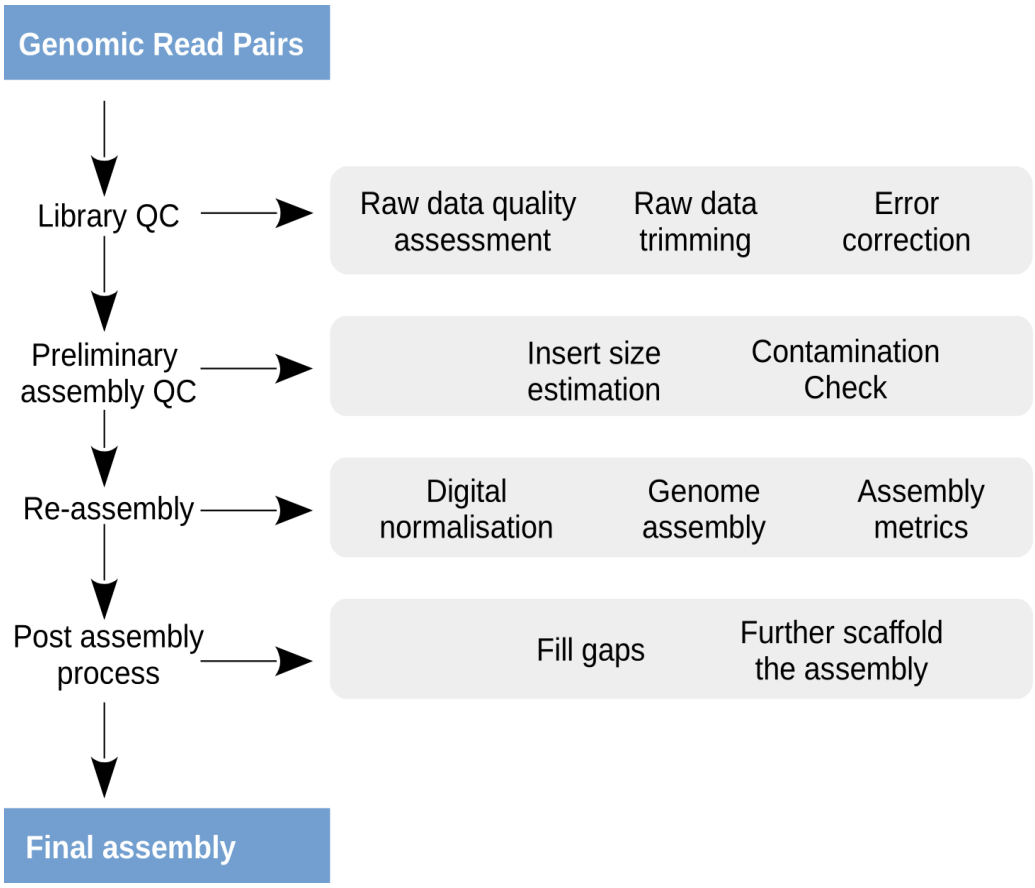


Figure 2.2 Standard workflow for assembling raw genomics reads

2.2.5.1 Raw data quality analysis

Illumina platforms generate reads in cycles. Each cycle corresponds to a single base position in the read. A quality score is assigned to each base which corresponds to a probability of error. Usually, the quality starts dropping in the last cycles. The raw reads are then delivered in fastq format [58], which has the read sequence and the measured quality for each position of the read.

An initial quality check is performed to the raw reads to find potential problems with the sequencing run. The overall quality of the run can be assessed by the distribution of qualities per position for all the reads. The GC distribution of the reads can show potential contamination, which can be taken into account in later steps of the pipeline. Overrepresented k-mers are useful to identify potential adapters still present in the raw sequences, and number of N's per position can show failed cycles.

2.2.5.2 Raw data trimming

The quality of a read usually drops in the last cycles [39], and adapters can still be present if the fragment that is being sequenced is shorter than the read length. Low quality bases need to be removed, because they cause unnecessary complexity in the construction of the de Bruijn graph and may contribute to the assembly of erroneous contigs. Adapter sequences, if not trimmed, will bridge different sequences together creating chimeric contigs.

Fastq files have the quality displayed for each base as an ASCII character, which encodes a Phred quality range from 0 to 40. Phred quality scores (Q) are logarithmically related to the base error probabilities (P), with $Q = -10 \log_{10} P$. For example, a Phred quality of 20 corresponds to an error probability of 0.01. Typically, bases are trimmed below a Phred quality of 20, and reads are discarded if they are below a certain length. Reads that contain N's even after trimming are usually discarded, because it is indicative of a potential problematic read.

2.2.5.3 Error correction

After trimming, miscalled bases can be present in the reads. Error correcting programs try to correct these bases based on a frequency table of k-mers assuming a uniform coverage. The reads are split into overlapping k-mers and a frequency table is created. Then, low frequency k-mers are corrected based on error profile calculated from the data and similar high frequency k-mers. Error-correcting programs use the quality values of each position to determine the error profile (e.g. Quake [59], Shrec [60]). An alternative approach that does not rely on uniform coverage was proposed in [61].

Error-free reads can improve the genome assembly and reduce compute resources required, because the assemblers produce simpler de Bruijn graphs. Furthermore, error correction is very useful when used on high coverage datasets. However, it is possible that a small number of real bases will be erroneously corrected.

2.2.5.4 Library insert size estimation

PE libraries are created by selecting a specific length of the fragments to be sequenced (180 b to 600 b). The insert size of the library is the distance between the 5' end of the paired reads. Although, a specific insert size library is requested by the sequencing centre, errors can still occur. Hence, it is recommended to calculate the insert size based on the observed distance between the reads in a Single End (SE) assembly (e.g. libraries for *Rhabditis* sp. SB347, see 4.3.9).

Reads are assembled as though they were single end (i.e paired reads are assumed to not be connected to the same fragment). Although this results in a more fragmented assembly, reads that are mapped to the same contig contain the distance information between them. This distance information is used across the whole dataset to calculate the insert size distribution of the library.

Most de Bruijn graph assemblers can calculate the insert size while assembling the genome (e.g. [50, 54]), therefore this step can be skipped. However, it is recom-

mended, because problems can be identified with the library and whether a bigger insert size library is required. PE read insert size distributions are assumed by the assemblers to follow a normal distribution, and by plotting the distance information additional errors of the library can surface (e.g. the 600 bp library for *Enoplus brevis*, see 4.3.1). Furthermore, in some cases the insert size is smaller than the read length of both reads, and it is better to merge the overlapping reads into longer reads prior to assembling the genome [62].

2.2.5.5 Contamination check

Using the SE assembly or the PE assembly with the recalculated insert size, we can screen the assembly for contaminants. This is a much faster approach than screening reads for contamination. Different organisms have a different distribution of k-mers and thus will assemble independently. All the contigs of the assembly that are above a certain length (usually 1000 bp) are compared to the NCBI nucleotide (nt) database using the BLAST+ suite [63]. Each contig is taxonomically annotated based on its best BLAST hit. The contigs are then plotted in a Taxon Annotated GC Coverage (TAGC) scatter plot, with contig GC in X-axis and log contig coverage in y-axis, and colour-coded based on their taxonomic affiliation [64].

This visualisation helps with identifying clusters of contigs belonging to different species than the target species. The GC axis helps distinguish different organisms due to the inherit organism GC bias. Coverage reveals differences on the effective stoichiometry of different genomes in the input DNA. Contigs belonging to different clusters of GC or coverage or both can be easily seen in a TAGC plot and flagged as contamination (various TAGC plots can be seen in chapter 3). All the reads that map to these contigs are removed to produce a cleaner raw dataset.

2.2.5.6 Digital normalisation

Extremely high coverage datasets create a novel problem in short-read assemblies. A lot of the k-mers will be new due to sequencing errors and due to the amount of erroneous k-mers, error correction programs cannot cope. Additionally, de Bruijn graph assemblers will require prohibitive amounts of machine memory. Digital normalisation was proposed as a solution to the problem [65].

Digital normalisation works on datasets that have been generated through sequencing removing high-coverage reads. The coverage is normalised uniformly across the genome to a specified value (i.e ~50X), reducing sampling variation by removing reads and the possible errors contained within them. The reduction of reads results in lower computational requirements for *de novo* assembly.

Digital normalisation is not a mandatory step, and should rather be used only in cases where de Bruijn graph assemblers are incapable of handling the amount of data generated. Furthermore, using digital normalisation on polymorphic and repetitive genomes can result in worse assemblies.

2.2.5.7 Genome Assembly with reduced dataset

Gathering all the information from the previous steps, and reducing the dataset either by removing contaminant reads or with digital normalisation or both, the final dataset is created to be used for assembling the genome. The choice of assembler is mostly left to personal preference, although studies have assessed the performance of different assemblers based on different datasets (discussed in subsection 2.2.6).

Different assemblers offer the choice of different parameters with the most important being k-mer size (discussed in 2.2.2). The choice of k-mer size and other parameters are dependant on the dataset and should be tailored to different organisms. Although it is impossible to test every parameter, the default values should be adequate for almost all datasets.

2.2.5.8 Genome Assembly metrics

There are two ways of assessing the quality of the assemblies produced by different programs or parameters. The first one is the contiguity of the assembly, and other assembly-derived metrics. The second one is assessing the biological accuracy of each assembly.

Contiguity of the assembly can be assessed with the N50 metric, which is the size of contig (or scaffold) in an assembly such that 50% of the assembly is in contigs of that size or larger. A higher N50 usually suggests a better assembly. However, nothing can be said about the true quality of the assembly, because N50 values can be inflated by chimeric contigs. A better use of the metric is to compare it to the expected gene size. For example, if the N50 of an assembly is 6,000 bp and the expected gene size is 800 bp, that would mean that most genes would be in a single contig and thus exon-complete when annotating the genome. Other metrics when comparing different assemblies are the number of N's present, the GC percentage of the assembly and the number of contigs.

A biological metric of genome assembly quality is the number of full length genes present in each assembly. The Core Eukaryotic Genes Mapping Approach (CEGMA) [66] has a set of 248 highly conserved genes from six evolutionary distant eukaryotic species (*Saccharomyces cereviseae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*), that can be searched in the assembly. CEGMA uses Hidden Markov Models (HMMs) created from the alignments of the proteins. A higher percentage of complete CEGMA genes found indicates a better assembly. Although in some cases a 100% completeness is not possible due to to the difference between the HMM profiles and the genes present or the evolutionary loss of the gene, it is useful for comparing different assemblies and also getting an initial idea about the completeness of each assembly.

A more informative approach is to search the assembly for completeness of Ex-

pressed Sequence Tags (ESTs) and transcripts from transcriptome assemblies from the same species, and proteins from closely related species. Using the BLAST+ suite the assembly can be searched for these sequences and be assessed. More contiguous assemblies will have more transcripts and ESTs present in full length.

2.2.5.9 Post-assembly processing

The best assembly, based on the metrics, can be further improved by filling gaps created during the PE stage and scaffold the contigs using additional evidence (e.g. MP libraries, transcripts). Although regions of the genome that have repeats can be resolved with PE information, the repeated regions are not identified and are displayed by a string of N's. These repeated regions can be resolved by mapping the reads back to the assembly. The string of N's will then be replaced by the actual repeat sequence (e.g program GapFiller [67]). Once most of the gaps are filled, the assembly can be used for scaffolding. Either by MP libraries or transcripts, contigs can be bridged improving the contiguity of the assembly.

2.2.6 Performance of assemblers

The accuracy of short read assemblers was evaluated in previous studies with simulated and real datasets. The first Assemblathon study [68] assessed different assemblers on a simulated genome derived from human chromosome 13. Next, PE and MP reads were generated for the simulated genome. Assemblies were submitted by different groups independently, and the Assemblathon team, which consisted from the groups that submitted an assembly, evaluated each assembly based on contiguity metrics. Different assemblers performed variably in different metrics, but the most consistent were ALLPATHS-LG [69], SOAPdenovo [52] and SGA [70]. Interestingly, there were extreme differences between the assemblers tested, indicating the importance of assembler choice.

The GAGE study [71] assessed eight highly used assemblers on four genomes

(*Staphylococcus aureus*, *Rhodobacter sphaeroides*, human chromosome 14, and *Bombus impatiens*). Only *B. impatiens* didn't have a reference genome. The authors showed the problem of only using the N50 metric to assess performance, because a large N50 may mean higher number of errors. ALLPATHS-LG was again the most consistent assembler. SOAPdenovo produced assemblies with similar N50, but they were with far more contiguity errors.

The second Assemblathon study [72] assessed different assemblers on three vertebrate species (*Melopsittacus undulatus*, *Maylandia zebra*, and *Boa constrictor constrictor*) without a reference genome. Fosmid sequences and optical maps were generated to evaluate the assembly contiguity. Assemblies were submitted by different groups independently. Reads were generated from Illumina, Roche 454 and PacBio technologies. As in the first Assemblathon study, different assemblers performed differently in different metrics. Furthermore, it appears that assemblers vary in performance for different species.

From these three studies, ALLPATHS-LG was the most consistent assembler. This algorithm requires an overlapping PE library of 180 bp insert size, and a high quality MP library. This combination of libraries was not available for any of the species in this study, and thus ALLPATHS-LG could not be tested. Assemblies from SGA and SOAPdenovo were generated for some of the species but they had worse assembly metrics compared to Velvet [54] and ABySS [50] assemblies. Overall Velvet and ABySS assemblies were similar in terms of contiguity and gene presence.

2.3 Transcriptome assembly

The *de novo* assembly of a transcriptome is a different process than the *de novo* assembly of a genome. Many assumptions that are used in short-read genome assemblers cannot be applied to transcriptome datasets. For example, genomes are expected to have a uniform coverage and one locus should result in one contig. The transcriptome assembler should be able to produce one contig per isoform rather

than per locus, and transcripts will have different coverage based on their expression levels. As a result, RNA-seq *de novo* assemblers are extensions of genome assemblers, to accommodate the different approach required. The transcriptomes of the species described in this study were assembled with the program Trinity [73] which was developed solely for transcriptome assembly from Illumina libraries. Trinity uses three underlying programs (Fig. 2.3).

- ***Inchworm*** extracts all overlapping 25-mers from the RNA-Seq reads and generates transcripts based on 24-mer extensions. The dominant isoform is usually recovered in full length at this stage, and the alternatively spliced isoforms are reported partially (only the unique regions).

- ***Chrysalis*** uses the raw read information to cluster Inchworm contigs into components. Contigs that share reads or paired read links are grouped together. These components contain contigs that originated from the same gene (alternative isoforms) or from closely related genes (paralogous transcripts). Each component is then treated separately as an assembly problem; a de Bruijn graph is constructed and the reads are partitioned among the components. Partitioning the reads into separate components allows for parallel processing of following computations.

- ***Butterfly*** operates on the individual graphs, identifying full length transcripts for all the isoforms and removing transcripts that belong to paralogous genes. Butterfly traverses the graphs tracing connectivity based on read sequencing and information from PE reads. If connectivity cannot be established, the graph is split into sub-graphs to be processed separately. In the end all the transcripts are reported along with the component information.

After the transcriptome assembly, abundance estimation of transcripts can be calculated by RSEM [74]. Abundance estimates are reported in fragments per kilobase of transcript per million fragments mapped (FPKM). A number of transcripts

that were assembled with Trinity may have low or zero expression values indicating questionable biological significance. In addition, some isoforms may be represented at a low percentage within a component. The transcripts predicted by Trinity can then be filtered based on specific constrains imposed.

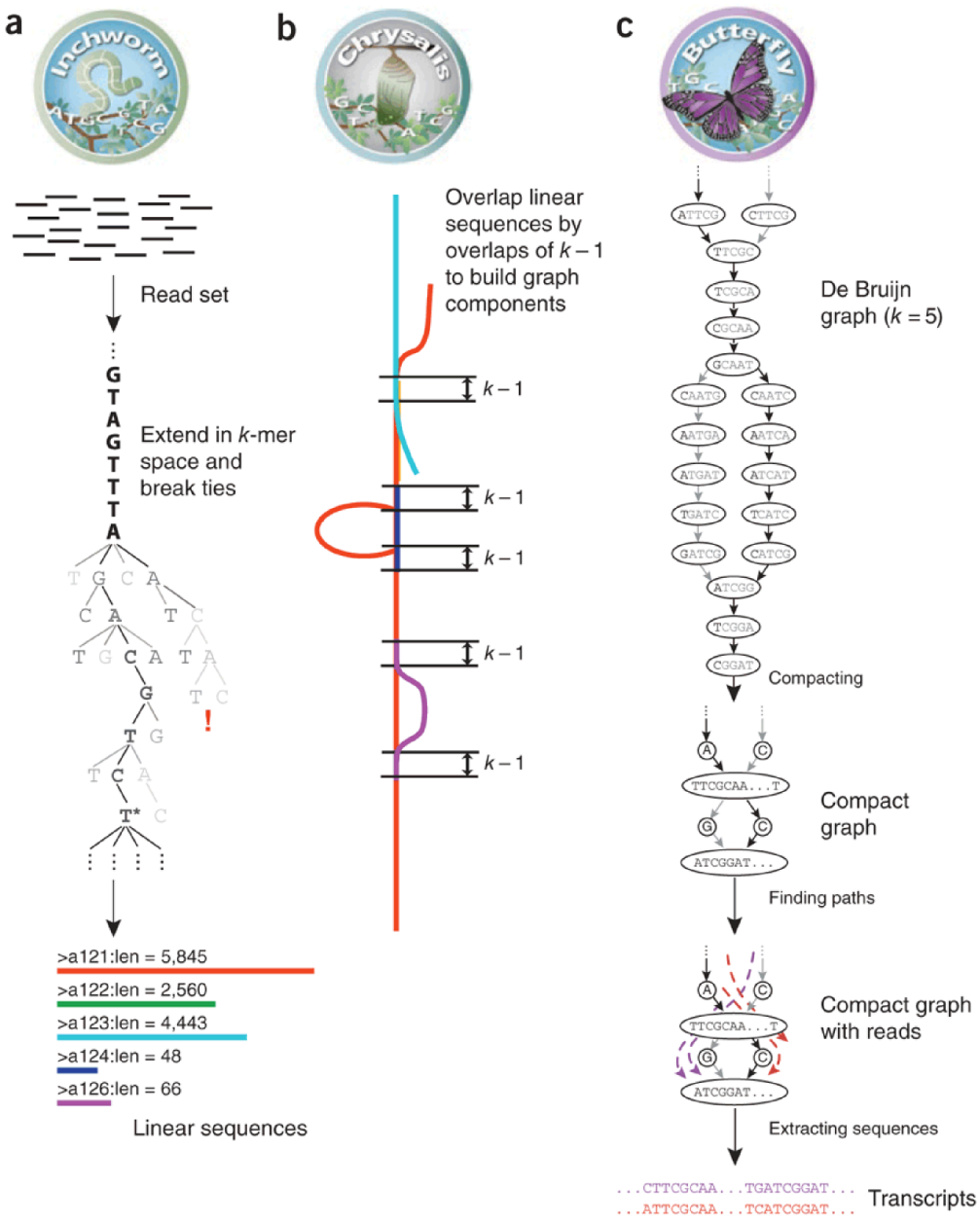


Figure 2.3 Overview of the three programs inside Trinity. (a) Inchworm performs the initial assembly, (b) Chrysalis creates each component as a de Bruijn graph, and (c) Butterfly traverses all component graphs. Figure from [73]

Chapter 3

Post-Assembly Scaffolding and Gene Finding

To produce a more contiguous assembly, I improved a previous algorithm for scaffolding contigs based on transcript alignments in the genome. The algorithm is outlined in the first section and although it is still under active development, the results are promising. In the second section I test the existing annotation pipelines and provide a combined pipeline approach that gives better results than using the programs separately.

3.1 SCUBAT v2 program

Genome sequencing projects often have the goal of identification of genes that will allow functional studies and evolutionary comparison between related species. Full length gene models can be more easily found in long genomic fragments. As discussed previously, short read genome assemblies produce contigs/scaffolds of varying length. A more complete genome can be achieved by using libraries with different insert sizes to connect contigs over repeat regions of the genome. Large-insert libraries can drastically improve the contiguity of the assembly, but making these libraries is still a complex, expensive and time-consuming procedure. In organisms of small size, the

amount of DNA required is not enough from one specimen and multiple specimens have to be pooled together which creates additional problems due to heterozygosity. As an alternative, transcripts spanning multiple contigs can be used as evidence to further improve assembly continuity. For example, in *C. elegans* intronic regions can be as long as 20 kilobase pairs (kbp) and thus transcripts can help bridge contigs based on the exon continuity.

Adjacent exons in two different contigs can be used to bridge these contigs together. The procedure can be compared to that of the long insert libraries which are used in scaffolding. The evidence from the transcripts mapped to exonic regions is used, instead of the MP reads mapped to regions randomly across the assembly. Although intron size distributions vary between organisms, it is possible to use the transcripts alone to scaffold NGS sequences.

The program SCUBAT (or SCUBAT v1) was created by Ben Elsworth [75] to use transcript information to bridge contigs. It uses the output of BLAT aligner [76] and merges the contigs with CAP3 assembler [77]. Similarly, the program L_RNA_scaffolder [78] also uses the output of BLAT but merges the contigs internally.

SCUBAT v2 aims to use transcripts to scaffold genomic contigs but uses the BLAST+ aligner instead which can also use proteins as queries. SCUBAT v2 algorithm was written from scratch in python to allow the use of BLAST XML output. In summary, SCUBAT v2 algorithm filters the BLAST output of aligning the transcripts to the genome assembly, and then creates potential connections between contigs based on the exonic evidence. Finally, scaffolding occurs between optimally selected contigs and a new genome assembly file is generated. The program was applied to a custom assembly of *C. elegans* genome using transcripts from a custom assembled *C. elegans* RNA-Seq dataset. The program is available on GitHub [79].

3.1.1 SCUBAT v2 algorithm

The program requires two files, the genome assembly file and a BLAST XML file of transcripts aligned to the genome assembly. The algorithm can be broken down in 3 steps, filtering the BLAST file, creating connections between contigs, and scaffolding the contigs creating a new assembly.

3.1.1.1 BLAST XML file

The user must provide a BLAST XML file of transcripts aligned to the genome assembly. The transcription-based linking information can be derived from either nucleic acids (i.e. transcripts) or proteins (i.e. proteome). The XML format was chosen because it contains all the information required in an easily parsable way. Furthermore, XML formats remain constant between different versions of the BLAST+ suite and can be scaled to include additional information without breaking the structure.

3.1.1.2 Filter BLAST results

BLAST results have to be filtered before proceeding because they contain contradictory information. Common problems are High-scoring Segment Pairs (HSPs) extending beyond exon boundaries, HSPs containing each other in the same contig region, HSPs being found with low similarity in other contigs and HSPs being found in two different regions of the same contig (Fig. 3.1). The filtered results will contain the most probable scenario of the alignment. While the BLAST results are being filtered, transcripts are divided into 4 categories.

- ***Unaligned transcripts***. These transcripts can either be misassemblies during transcriptome assembly or they align to an unsequenced part of the genome. Since they contain no information, they are discarded.

- ***Transcripts with one HSP***. Only the information about the exon size is retained.

- ***Transcripts with multiple HSPs in the same contig***. Information about exon and intron size can be extracted.

- ***Transcripts that have HSPs in different contigs***. The information from these transcripts will be used to scaffold contigs. Exon size information is retained, and intron size information is extracted if any HSPs are present in the same contig.

Filtering the results simplifies the problem, and connections can be more easily created. The exon and intron size information can be used later for assessing the likelihood of each connection proposed.

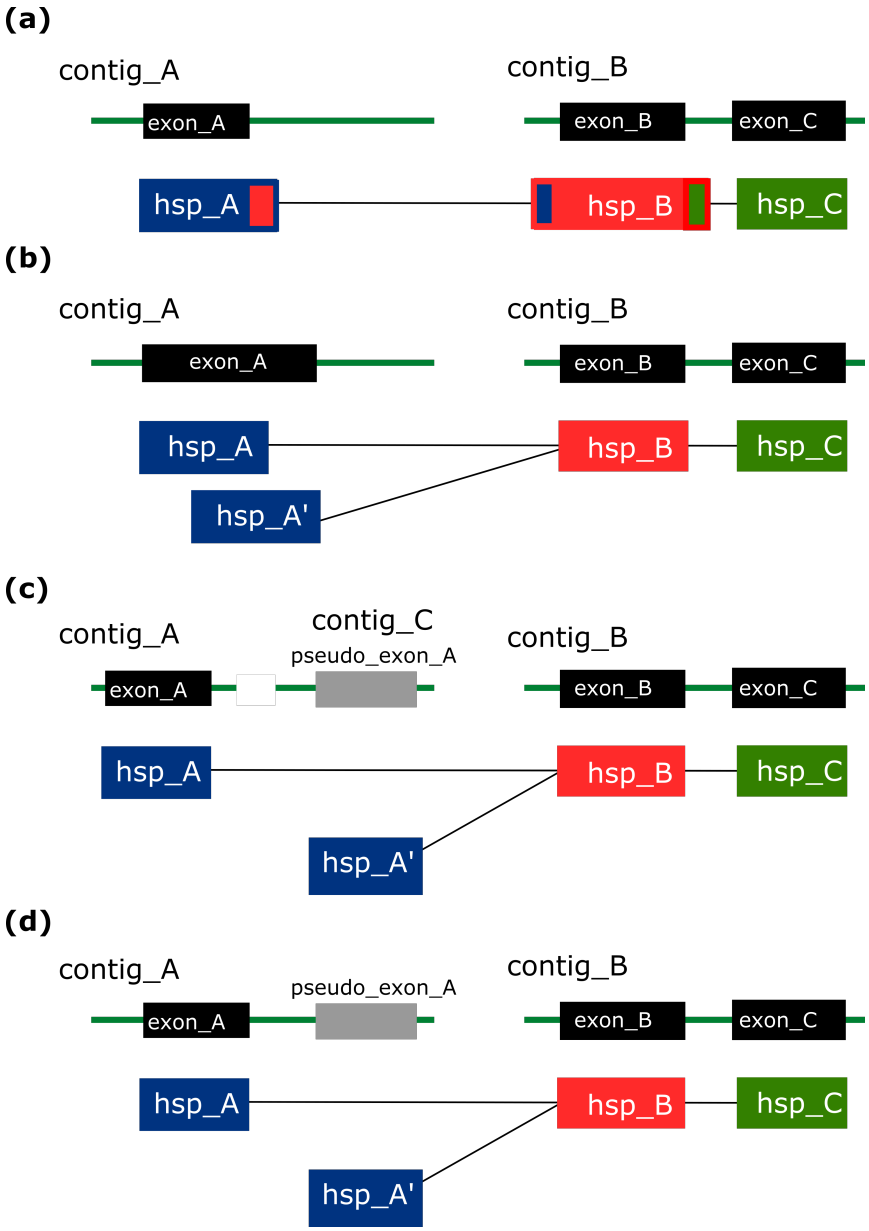


Figure 3.1 HSP filtering from BLAST output. Contigs are shown as green lines, exons are shown as black boxes, pseudo-exons are shown as grey boxes, HSP links are shown as black lines and HSPs are shown as blue, red and green boxes. (a) If HSPs overlap a few bases, both HSPs are kept, otherwise the one with the lowest identity is removed, (b) If HSPs overlap in the same region, they are merged, (c) and (d) the HSP which aligned better is retained

3.1.1.3 Create connections

A connection is created when different HSPs of a transcript (exons) align to multiple contigs and the proposed intron size is smaller than the maximum intron size allowed. The maximum intron size allowed is derived from the intron size distribution of aligned transcripts or defined by the user. Before a connection is created, the minimum intron size of that connection is calculated. The minimum intron size is the distance between the exons if the contigs were adjacent plus a number specified by the user (insert size of the library used for the genome assembly, if applicable). Next, the connections are filtered to only allow one incoming and one outgoing connection per contig.

3.1.1.4 Scaffold the contigs

The contigs of the assembly are then scaffolded based on the final proposed connections. This part of the algorithm is currently in active development and will be incorporated in the main program at a later date.

3.1.2 *C. elegans* test dataset

To test the effectiveness of the algorithm, real data from *C. elegans* experiments were used. Real data are better than simulated datasets, since a chromosome level assembly exists for *C. elegans* and raw read datasets can be downloaded from the public repositories. In that regard, we can assess how the algorithm works with both genomic and transcriptomic misassemblies.

Raw genomic reads with accession number ERR089813 were downloaded from European Nucleotide Archive (ENA). The reads were trimmed with fastq-mcf and assembled with Velvet (k = 51). The assembly spans 95,535,178 bp in 4,901 contigs above 500 bp with a N50 size of 49,990. The genome assembly was mapped to the *C. elegans* assembly from WormBase (WS236) [80] with MUMmer [81].

Raw RNA-Seq reads with accession number SRR504322 were downloaded from

ENA. The reads were trimmed with fastq-mcf and assembled with Trinity. Only the contigs that had a minimum of 300 bp Open Reading Frame (ORF) were considered in subsequent analyses. This resulted in 16,686 contigs with a median size of 483.

3.1.3 Results

The SCUBAT v2 algorithm was tested against SCUBAT v1 and L_RNA_scaffolder. The results of each program (contig connections) were verified by the order of contigs compared to the chromosome level *C. elegans* assembly. The programs were run with two different parameter settings: identity and transcript coverage cutoff of 90% and 95%. As accuracy measures, sensitivity (Sn) and specificity (Sp) were used. The sensitivity is defined as the number of correctly predicted connections divided by the number of correct connections that can be inferred given the dataset. Specificity is the number of correctly predicted connections divided by the number of predicted connections. A predicted connection is considered correct if the contigs are adjacent in the MUMmer output.

Table 3.1 shows the results of each program. In both cutoffs SCUBAT v1 predicted the most connections, although at a low specificity. Since SCUBAT v1 and L_RNA_scaffolder use the same aligner, the latter program appears to produce better results. SCUBAT v2 and L_RNA_scaffolder both had similar sensitivity and specificity, with SCUBAT v2 having better values. When the thresholds are raised, SCUBAT v2 predicts fewer true connections, while L_RNA_scaffolder predicts more correct connections. This difference can be attributed to the different aligner that the programs use. Figure 3.2 shows the number of shared connections that each program proposed for 90% cutoffs. It appears that a combination of SCUBAT v2 and L_RNA_scaffolder will produce the best results (combined: Sn=88.36%, Sp=99.49%). For that reason, a new option was added to SCUBAT v2 that can also output the proposed connections in a format that can be used with L_RNA_scaffolder.

Table 3.1 Comparison of the results from the three scaffolding programs on the *C. elegans* dataset

Identity & Transcript coverage cutoffs	Program	Connections	Correct	Sn	Sp
90%	SCUBAT v1	397	377	84.34%	95.21%
	SCUBAT v2	379	378	84.56%	99.73%
	L_RNA_scaffolder	341	339	75.83%	99.41%
95%	SCUBAT v1	394	376	82.85%	95.43%
	SCUBAT v2	373	373	81.97%	100%
	L_RNA_scaffolder	370	366	80.87%	98.91%

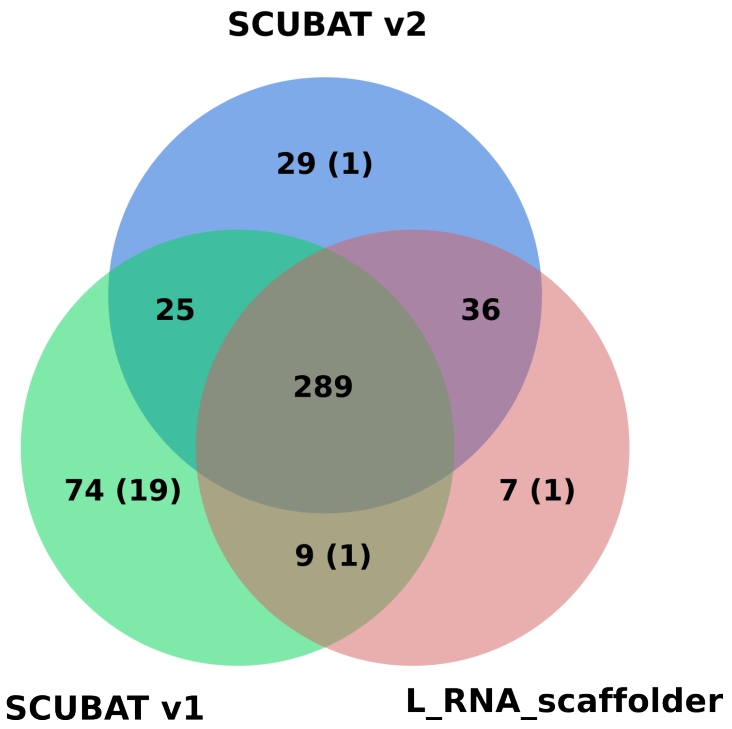


Figure 3.2 Venn diagram illustrating the number of shared proposed connections between the three programs (90% cutoffs). The numbers in brackets show the number of erroneous proposed connections.

3.2 Annotation

After the final draft assembly is generated, the process of annotating the genome can start. If the N50 is low, additional sequencing may be required, otherwise genes will not be predicted correctly. High-quality annotation requires human genome annotators who will review the genes predicted by various sources of evidence and fix intron-exon boundaries. Since time and resources were limited, an annotation pipeline created by Sujai Kumar and named two-pass MAKER2 [16], was used to automate the process for draft nematode assemblies. This section includes an overview of the annotation process, a summary of the pipeline and its efficiency when applied in a model organism.

3.2.1 Annotation process

Usually, the first step in annotation is to predict the gene structures in the genome. Automated annotation pipelines that deal with prokaryote genomes give good results since the gene structure in these organisms is less complex [82]. Eukaryote genomes have more complex gene structures: intron-exon boundaries are hard to predict and intron sizes vary from a few to thousand of bases [83]. Furthermore, untranslated regions (UTRs) in transcripts are hard to identify. Ultimately, the annotation pipeline should be able to predict correctly all the exons, introns and UTRs for each gene model. There are three ways to predict genes in the assembly with computational methods [84].

- ***Evidence alignment***. ESTs and RNA-seq data are aligned to the genome assembly. These sequences are from the same organism whose genome is being annotated. Proteins from other organisms are also included, since proteins are more conserved than nucleotide sequences in evolutionarily distant species. The SwissProt dataset includes well-curated proteins which can be incorporated in the annotation pipeline [85]. RNA-Seq datasets can either be used as

transcripts after *de novo* assembly or directly as reads.

- ***Ab initio***. *Ab initio* algorithms use mathematical models to predict gene models in genome assemblies. These tools usually need a set of high quality genes to tune their models for different species. Codon frequencies and intron and exon length distributions allow the algorithms to identify exonic regions and predict correct intron-exon boundaries. Although some of the tools come with predefined parameter files for a few species, it is better to retrain the program unless the species to be annotated is very closely related to the pre-calculated species files.

- ***Evidence driven***. These programs use external evidence (i.e. ESTs, proteins) to improve the *ab initio* predictions. Using multiple sources has the advantage of producing a more complete annotation. However, this approach requires the use of multiple programs, and the results are collated together to produce a final gene annotation. Usually, these programs are grouped together in annotation pipelines.

Next, the gene models are functionally annotated based on their similarity to known genes. The program annot8r [86] can assign GO terms, EC numbers and KEGG pathways to gene predictions. The program InterProScan [87] scans the genes for Pfam domains and other protein signature databases. The program BLAST2GO [88] also assigns GO terms, and can incorporate Pfam domains for better accuracy.

Finally, the RNAs present in the genome can also be annotated. The tRNAscan program [89] is used to identify all the tRNAs present. RNA genes and ncRNAs are identified by screening the genome against Rfam database using covariance models [90].

3.2.2 Performance of annotation programs

The accuracy of annotation programs was evaluated in three Genome Annotation Assessment Project (GASP) studies. The programs were assessed in ENCODE regions of the human genome (eGASP [91]), in the *C. elegans* genome (nGASP [92]), and in human, *C. elegans* and *D. melanogaster* genomes using RNA-Seq data (rGASP [93]). In all three studies Augustus [94], Fgenesh++ [95] and mGENE [96] had similar accuracy, with Augustus performing a bit better. Overall, it appears that annotation pipelines have trouble with genes with many exons, very short exons, very long introns, weak translation start signals and weak splice sites.

3.2.3 Two-pass MAKER2 pipeline

A more comprehensive review of the pipeline can be found in [16]. In summary, multiple evidence is used in the first pass of the MAKER2 pipeline [97], and *ab initio* programs are retrained using the output of the first pass. Then, the MAKER2 pipeline is rerun with the updated parameter files. The workflow has the following steps (Fig. 3.3).

- ***SNAP models***. CEGMA is run on the genome assembly, and the genes found are used to train SNAP [98].
- ***GeneMark models***. GeneMark [99] is run on the genome assembly.
- ***First-pass MAKER2***. The first pass of MAKER2 uses the HMM models from SNAP and GeneMark, ESTs and transcripts (if present), and proteins from SwissProt and closely related species (if available).
- ***SNAP retrained models***. The genes obtained from the previous step are used to retrain SNAP.
- ***Augustus models***. Augustus is trained using the gene models from first pass MAKER2.

- ***Second-pass MAKER2***. It uses the same evidence files as the first pass, the retrained SNAP models and the addition of the Augustus models.

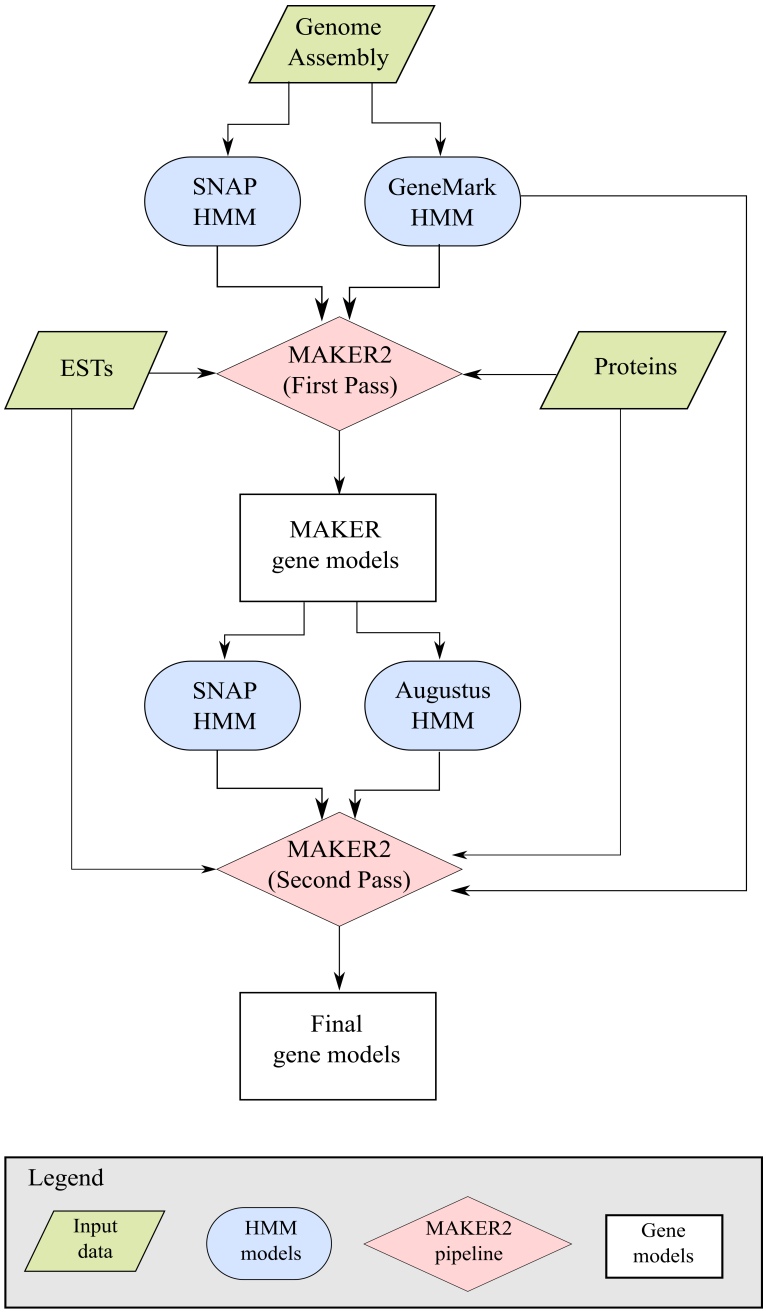


Figure 3.3 Two-pass MAKER2 pipeline.

3.2.4 Pipeline evaluation

Using the same dataset as in subsection 3.1.2, the pipeline was evaluated at 3 different stages,

- First-pass MAKER2 predictions
- Augustus predictions
- Two-pass MAKER2 predictions

The predictions for each stage were evaluated against the annotated features of the *C. elegans* chromosome-level assembly. Three features (gene, exon, nucleotide) were evaluated using the Eval program [100]. As accuracy measures, sensitivity (Sn) and specificity (Sp) were used. For each feature the sensitivity was calculated as the number of correctly predicted features divided by the number of annotated features, while the specificity was calculated as the number of correctly predicted features divided by the number of predicted features. A predicted exon is correct if the prediction of the splice sites matches exactly the annotated position of an exon. A predicted gene is correct if all exons are correctly predicted and the same number as the annotated gene.

Table 3.2 shows the accuracy values for three steps of the pipeline. As expected the two-pass MAKER2 produces better results than the one-pass MAKER2 since SNAP models are retrained and Augustus models are added. Interestingly, the Augustus predictions trained from first-pass MAKER2 models appear to be better in every metric. Figure 3.4 shows the number of shared exons between the predicted exons from two-pass MAKER2, Augustus and annotated *C. elegans* exons. Around 50,000 exons are predicted incorrectly by two-pass MAKER2, while ~20,000 exons are predicted incorrectly by Augustus. In both stages they failed to predict ~16,000 exons.

Table 3.2 Evaluation of three stages of the two-pass MAKER2 annotation pipeline.

	One-pass MAKER2	Augustus	Two-pass MAKER2
Gene Sn (%)	24.79	41.90	33.26
Gene Sp (%)	21.65	47.04	24.86
Exon Sn (%)	53.94	74.61	63.80
Exon Sp (%)	61.15	83.20	63.15
Nucleotide Sn (%)	81.79	92.60	88.77
Nucleotide Sp (%)	95.25	95.49	94.25

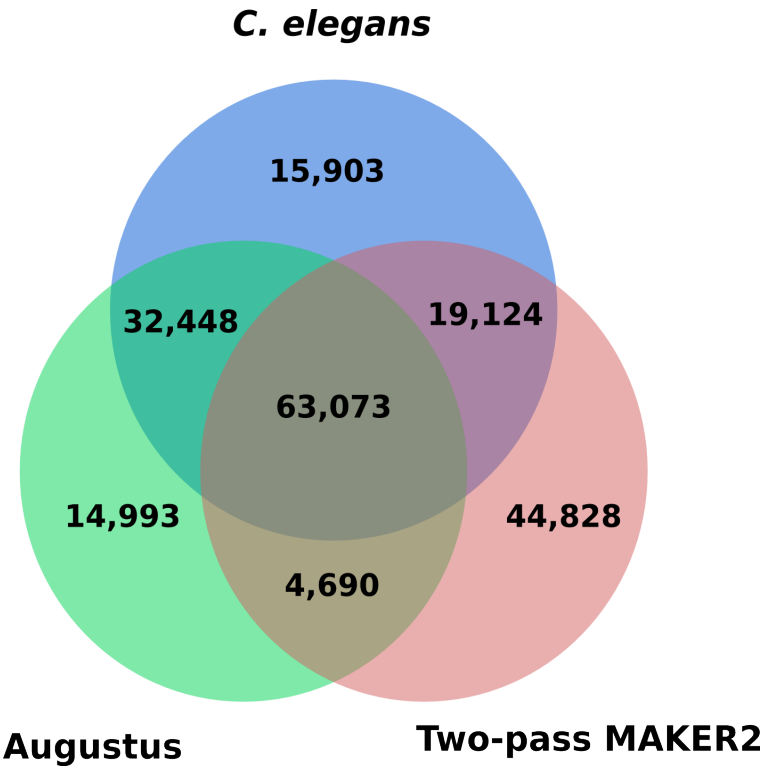


Figure 3.4 Venn diagram illustrating the number of shared exons between the predictions from Augustus, MAKER2 and annotated *C. elegans* exons.

Since MAKER2 incorporates multiple models from different programs as well as external transcript/protein evidence, one would expect to predict more correct features than individual programs. In the second pass, Augustus models are also included in the two-pass MAKER2 pipeline, but the accuracy drops quite substantially compared to Augustus predictions. This result indicates that other gene models of the pipeline are selected instead of the Augustus models. The predictions of SNAP, GeneMark and Augustus were tested against the annotated *C. elegans* exons. Exons were considered as equal between the datasets if they overlapped at least 90% of the exon length. This measure was helpful to identify differences between the programs. Figure 3.5 displays the amount of overlapping exons that were shared between the datasets. It appears that the three programs have differences in the way they call exons. It seems that there is a gradual elasticity from Augustus to GeneMark to SNAP, with Augustus being the most conserved and SNAP being the most relaxed. The difference in predictions may be the reason that the two-pass MAKER2 pipeline is less accurate than Augustus predictions. When a feature is being called within MAKER2, all the different evidence is being evaluated into one result. This means that it is possible that the SNAP models are swaying MAKER2 towards inaccurate predictions.

Both the differences of two-pass MAKER2 and Augustus predictions, and the differences between the individual programs prompted me to decide that better annotations could be obtained for the genomes of this study if the pipeline was stopped at the Augustus predictions. These predictions will then be used as the final predictions.

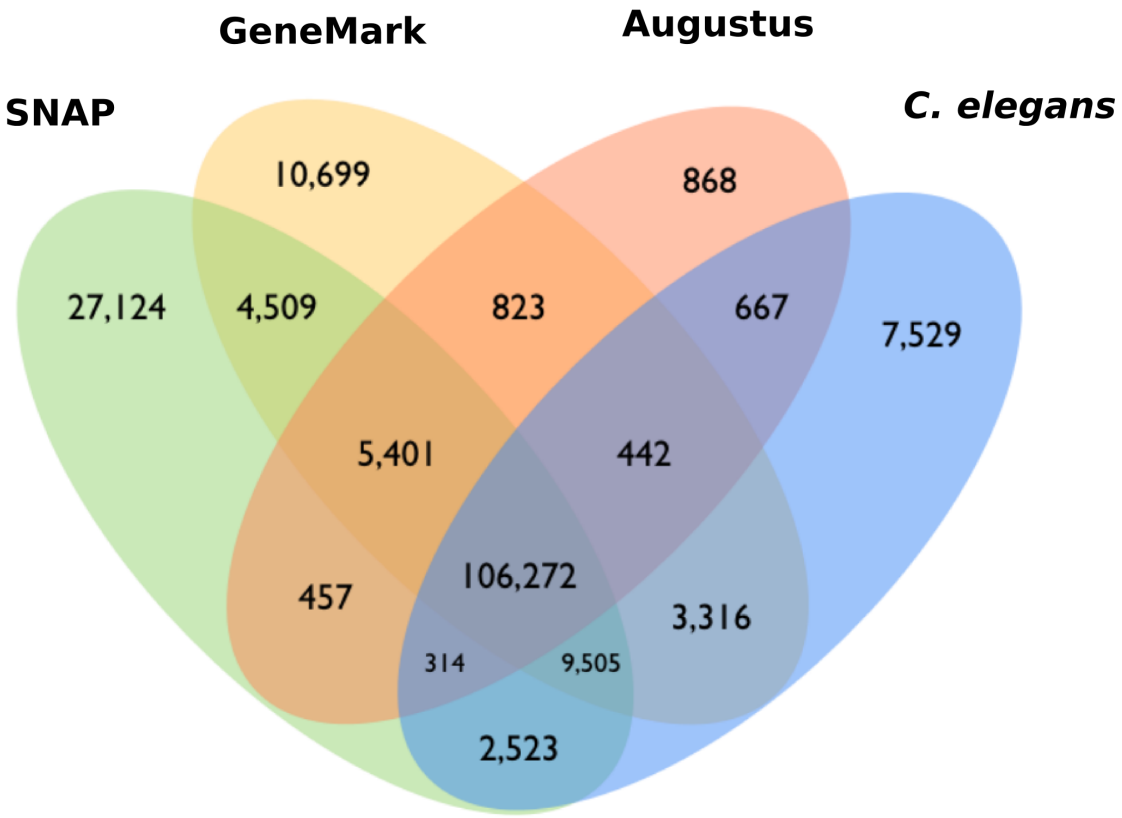


Figure 3.5 Venn diagram illustrating the number of shared overlapping exons between the predictions from SNAP, GeneMark, Augustus and annotated *C. elegans* exons.

Chapter 4

Species data analyses

In this chapter, the results for each of the species assembled in this study are presented.

4.1 Workflow details

Improvements in technology and bioinformatics algorithms occur at regular intervals. During the PhD programme, NGS technologies became more accurate at generating more and longer reads, and a large variety of software was introduced to tackle the new bioinformatics challenges. Different species were worked on at different periods and thus a variety of genomic libraries and software was used. Overall, I tried to have a flexible workflow, where each step serves a specific purpose but can be achieved using a variety of tools (details in 2.2.5).

FastQC [101] was used to assess the quality of the genomic and transcriptomic raw reads. For all datasets a phred quality of 20 was chosen and read pairs were discarded if one read was below 51 b. A number of different read filtering tools were tested, but in the end two programs were used for trimming, fastq-mcf [102] and Trimmomatic [103].

A preliminary SE assembly was generated for the genomic datasets using CLC-bio [51]. The SE assembly was used to estimate the insert size of the libraries, and

identify possible contaminants. The insert size of each library was estimated by mapping the reads back to the assembly using CLC-bio, and plotting the distance in a histogram using scripts from [104]. The correct orientation for PE libraries is Forward-Reverse (FR), while for MP libraries it is Reverse-Forward (RF). Contaminants were identified by generating a TAGC scatterplot using scripts from [105].

Although error-correction algorithms were not used in these analyses, they are highly recommended for future assemblies. Instead, digital normalisation was used in almost all datasets. This step generated more contiguous assemblies by removing low quality and high-error reads. The digitally normalised reads were assembled using Velvet [54] or ABySS [50]. Contig coverage was obtained by mapping all the reads using CLC-bio back to the assembly.

The terms contig and scaffold have multiple meanings when describing assemblies. For clarity, these two terms are defined below. The term contig can be used to describe any contiguous sequence that has been assembled by overlapping reads. In this chapter, it is used to define contiguous sequences from SE and PE assemblies. Thus, contigs may contain short gaps of known length based on the insert size of the library. The term scaffold is used to define contigs bridged together by MP libraries or other external evidence (i.e. optical mapping). Furthermore, in assembly comparison tables, scaffold includes both definitions. Finally, in cumulative length curves, the term scaffold includes both definitions, while contig is defined as scaffolds broken at sites with a gap of at least 10 b long (consecutive Ns). All the assembly statistics were calculated for contigs above 500 bp.

Trinity [73] generates thousands of probable transcripts from RNA-Seq experiments. An abundance estimation filter was applied in order to remove erroneous transcripts. The abundance of each transcript was calculated with RSEM, and transcripts were removed if they had extremely low expression levels (below 1 FPKM) or low isoform representation (below 1%). For each component the highest expressed transcript was selected for downstream analyses, with ORFs above 300 bp. Transcriptome assemblies were assessed by percentage of CEGMA [66] genes present. The

ORFs from raw Trinity output were searched against CEGMA hmm profiles using HMMER [106], with e-value threshold $1e^{-100}$.

Tables 4.1 and 4.2 show the genome and transcriptome assembly statistics for the species described in this chapter. These assemblies were generated with new data that have been obtained as part of this study. The difference in quality between assemblies for different species can be attributed to different quality of libraries used, different types of libraries and different levels of repeat content and heterozygosity of each species.

However, although the assemblies have different quality, most of the genes predicted were useful for phylogenetic analyses. Even though more data is needed to identify the complete proteome, draft assemblies and subsequent phylogenetic analyses provide a first glimpse of the evolutionary history of the phylum.

Table 4.1 Comparison of genome assemblies and annotations for the species described in this chapter

Species	Group	Genome size (Mb)	N50 (b)	Number of scaffolds	Number of genes predicted	Median transcript length (bp)
Hypsibius dujardini	Tardigrada	139.92	48,302	30,122	23,021	870
Romanomermis culicivorax	Clade I	322.77	17,632	62,537	10,206	387
Enoplus brevis	Clade II	126.60	753	165,525	–	–
Plectus murrayi	Group C	185.02	10,183	33,505	–	–
Plectus sambesii	Group C	239.12	9,794	73,716	–	–
Ascaris suum	Clade III	334.00	290,558	260	15,446	864
Dirofilaria immitis	Clade III	88.30	22,560	71,281	16,021	771
Onchocerca gutturosa	Clade III	109.56	10,567	25,605	19,916	456
Acanthocheilonema viteae	Clade III	77.35	25,808	6,796	10,397	975
Pseudaphelenchus vindai	Clade IV	53.83	5,784	28,977	6,073	858
Dictyocaulus viviparus	Clade V	169.39	22,560	17,715	14,306	834
Rhabditis SB347	Clade V	70.27	243,406	1,006	21,105	930

Table 4.2 Comparison of transcriptome assemblies for the species described in this chapter

Species	Group	Method	Number of transcripts	Median transcript length (bp)
Gordius sp.	Nematomorpha	Illumina	8,888	1,260
Prionchulus punctatus	Clade I	Illumina	21,290	662
Enoplus brevis	Clade II	Illumina	20,196	887
Plectus murrayi	Group C	Illumina	24,055	821
Setaria labiatopapillosa	Clade III	Illumina	14,283	1,049
Acrobeloides nanus	Clade IV	Illumina	9,549	495
Rhabditis SB347	Clade V	Illumina	13,225	988

Table 4.3 Location of the photos of the species described in this chapter

Hypsibius dujardini	http://tardigrades.bio.unc.edu/
Gordius sp.	http://www.naturefg.com/pages/an-annelida.htm
Enoplus brevis	http://www.boldsystems.org/index.php/Taxbrowser_Taxonpage?taxid=231885
Prionchulus punctatus	http://nematode.unl.edu/specay919267.htm
Plectus murrayi	https://bishwoadhikari.wordpress.com/research/
Setaria labiatopapillosa	http://parasitipedia.net
Onchocerca gutturosa	http://www.riverblindness.eu/onchocerciasis/filarial-species/
Dictyocaulus viviparus	http://www.uniprot.org/taxonomy/29172
Rhabditis sp. SB347	http://wweb.uta.edu/faculty/apires/research.html
Romanomermis culicivorax	http://i.somethingawful.com/news/2005/09/12-parasite1.jpg
Ascaris suum	http://en.wikipedia.org/wiki/Ascaris_suum
Dirofilaria immitis	http://pixshark.com/dirofilaria-immitis-microfilariae.htm
Acrobeloides nanus	http://yanailab.technion.ac.il/YanaiLab/

4.2 Outgroup species

Prior to this study, the most closely related phylum to the Nematoda with NGS datasets was the Arthropoda. Concerning evolutionary distance, the Arthropoda are very divergent from the Nematoda. As discussed earlier, outgroup species can influence the topology of the phylogeny. Therefore, two additional datasets from two species from two phyla were generated. *Hypsibius dujardini* from the phylum Tardigrada, and *Gordius* sp. from the phylum Nematomorpha. Nematomorpha are believed to be the closest phylum to the Nematoda, while the position of the Tardigrada is still under debate.

4.2.1 *Hypsibius dujardini*

Phylum . . . . . Tardigrada Brisson, 1762
Class . . . . . Eutardigrada Marcus, 1927
Order . . . . . Parachela Steindachner, 1881
Family . . . . . Hypsibiidae Pilato, 1969
Genus . . . . . Hypsibius Ehrenberg, 1840
Species . . . . . Hypsibius dujardini Doyere, 1840

H. dujardini is a member of the phylum Tardigrada which consists of microscopic animals, also known as water bears. Tardigrades are aquatic organisms, most commonly found in mosses, and are able to survive in extreme conditions of temperature, pressure, radiation and even outer space. They are capable of different types of cryptobiosis, surviving for years in hostile environments [107].

Raw data

The DNA sample was provided by Mark Blaxter. Two genomic libraries, a PE library of 300 bp (lib300) insert size and a MP library of 4000 bp (lib4000) insert size, were generated by Edinburgh Genomics, and sequenced on HiSeq2000 using 100 b PE

and MP sequencing. A RNA-Seq dataset was provided by Itai Yanai (Haifa, Israel). Raw sequence fastq files are not publicly available yet. The quality of Illumina reads was assessed with FASTQC, and over-representation of adapters was detected. Raw reads were filtered with fastq-mcf. The trimmed reads were then digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.4).

Genome Assembly and Annotation

The insert size distribution of lib300 on the preliminary SE assembly (assembly name: Hd_CLC_SE) has a median of 292 (96 Standard Deviation (SD)), and the insert size distribution of lib4000 has a median of 1133 (1460 SD) (Fig. 4.1). The TAGC plot shows contamination from various bacterial species (Fig. 4.3).

The digitally normalised reads were used in Velvet with a kmer size of 51. Three rounds of contamination removal were performed. Reads mapping to contaminant contigs were removed, and the remaining reads were assembled anew. Next, gaps within contigs were filled using GapFiller. Finally the MP library was used to scaffold the genome with SSPACE [108]. The final assembly (assembly name: Hd_Velvet_PE) spans 140 megabases (Mb) with median coverage of 86X. In terms of contiguity, the assembly is of moderate quality with an N50 of 48 kb (Fig. 4.2). There are 30,122 scaffolds above 500 bp. The assembly has a GC content of 45.1% and 92% CEGMA completeness (Table 4.5).

Transcriptome Assembly

A total of 94,006 contigs were generated by Trinity. Gene and isoform expression levels were calculated with RSEM and 17,199 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.7).

Synopsis

Hd_Velvet_PE's N50 was sufficient for annotating the genome, and the genes predicted were useful for phylogenetic analyses. Furthermore, the transcripts predicted from the RNA-Seq experiment show a high degree of CEGMA completeness (91.5%), and were useful for improving the genome annotation. The MP library was useful for scaffolding the genome. The genome assembly and annotations are available in a Badger instance [109] at http://badger.bio.ed.ac.uk/H_dujardini/. Further improvement to the assembly could be achieved by a cleaner sample and extra MP libraries.

Table 4.4 Read data for *H. dujardini*

Library	Raw		Post Trimming		Post Khmer	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib300	74.3	15.0	67.4	12.8	26.8	5.1
MP lib4000	58.8	11.9	44.5	4.9	—	—
RNA-Seq	175.6	35.5	144.5	28.1	—	—

Table 4.5 Comparison of assemblies for *H. dujardini*

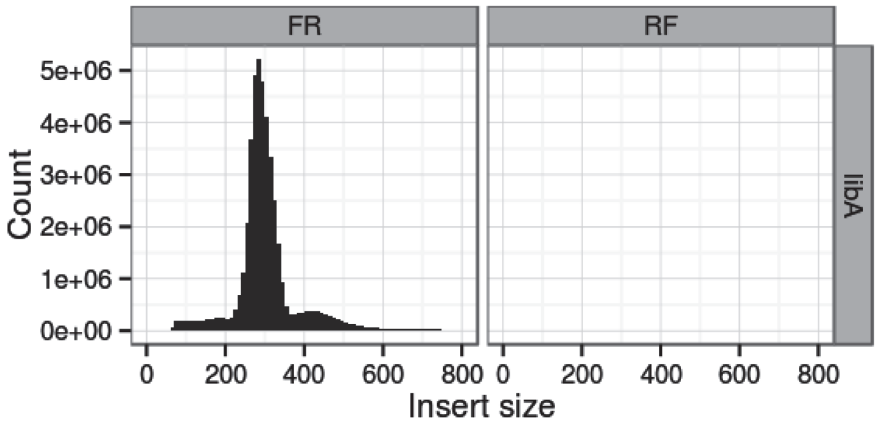
	Hd_CLC_SE	Hd_Velvet_PE
Number of scaffolds	264,712	30,122
Longest scaffold (bp)	277,783	594,143
Assembly span (bp)	184,593,598	139,919,165
Number of N's (bp)	33,616	3,548,300
Mean scaffold length (bp)	1,408	4,645
Scaffold N50 (bp)	1,725	48,302
GC content (%)	49.8	45.1
CEGMA completeness (%)	40.3	92.0
Transcriptome completeness (%)	7.5	88.0

Table 4.6 Genome annotation statistics for *H. dujardini*

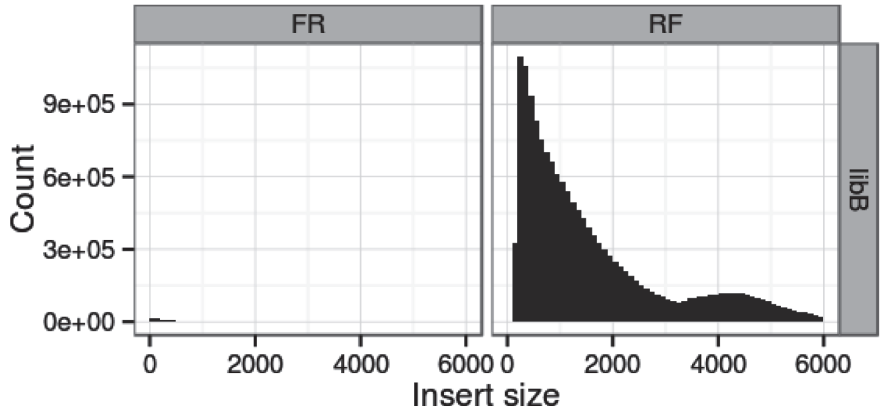
	Predicted
Number of genes	23,021
Longest transcript (bp)	30,894
Median transcript length (bp)	870
Median exon length (bp)	153
Median exons per gene	4
Median intron length (bp)	216

Table 4.7 RNA-Seq assembly statistics for *H. dujardini*

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	94,006	60,857	17,199	12,466
Longest transcript (bp)	28,057	19,134	25,732	15,696
Median transcript (bp)	690	960	1,042	1,053



(a) libA



(b) libB

Figure 4.1 Insert size estimations for the two libraries (a) lib300 and (b) lib4000 of *H. dujardini*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

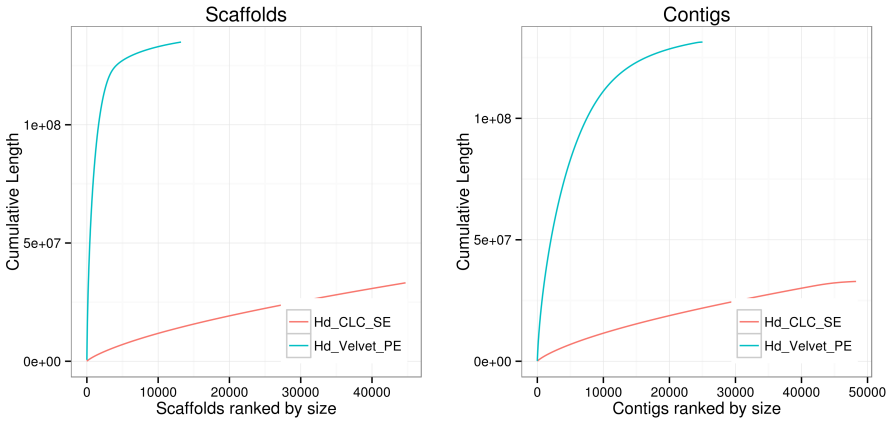


Figure 4.2 Scaffold and contig length cumulative curves for *H. dujardini* assemblies. Steeper curves indicate better assembly contiguity.

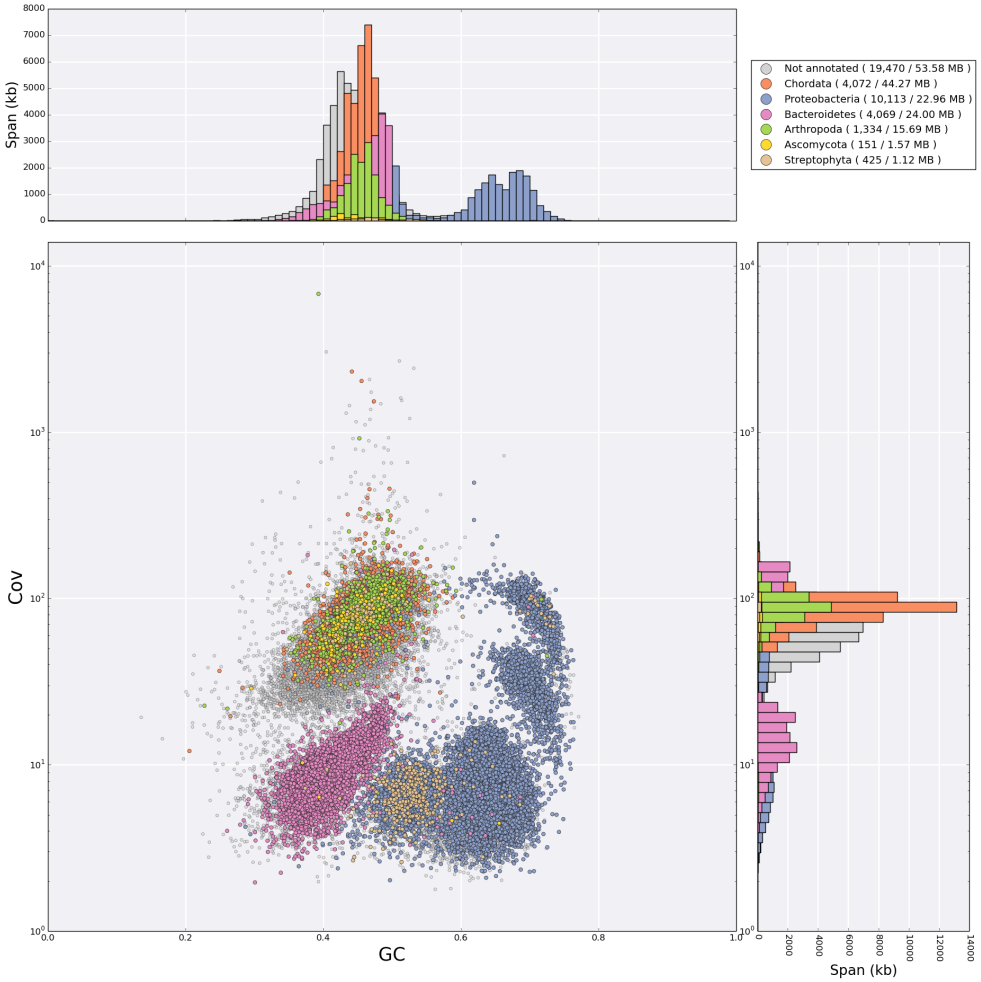


Figure 4.3 TAGC plot for Hd_CLC_SE prior to contamination removal. Multiple bacterial contaminants are visible in distinct groups.

4.2.2 *Gordius* sp.

Phylum ..... Nematomorpha Vejdovsky, 1886
Class ..... Gordioida Rauther, 1930
Superfamily ... Gordioidea Rauther, 1930
Family ..... Gordiidae May, 1919
Genus ..... Gordius Linnaeus, 1758
Species ..... Gordius sp.

Gordius sp. larvae are parasites of arthropods, while the adult stage is free-living. The life cycle of nematomorphs is similar to the order Mermithida (phylum Nematoda). Furthermore, they share several apomorphies with the nematodes, and these two phyla are widely accepted as a monophyletic unit [110].

Raw data

Two *Gordius* sp. RNA samples (one male and one female) were provided by Philipp Schiffer from the University of Cologne. Two short-insert PE RNA-Seq libraries were prepared by Edinburgh Genomics on an Illumina HiSeq2000 instrument using 100 b PE sequencing. Raw sequence fastq files are not publicly available yet. The quality of the reads was assessed with FASTQC, and no problems were detected. Raw reads were filtered using Trimmomatic (Table 4.8).

Transcriptome Assembly

A total of 69,831 contigs were generated by Trinity for the male sample. Gene and isoform expression levels were calculated with RSEM and 9,180 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.9).

A total of 29,726 contigs were generated by Trinity for the female sample. Gene and isoform expression levels were calculated with RSEM and 8,216 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.10).

A total of 75,839 contigs were generated by Trinity when the two samples were pooled together. Gene and isoform expression levels were calculated with RSEM and 8,888 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.11).

Synopsis

The transcripts predicted from the pooled dataset show a high degree of CEGMA completeness (90.3%) and were useful for phylogenetic analyses. The transcriptome assemblies are not publicly available yet.

Table 4.8 Read data for *Gordius* sp.

Library	Raw		Post Trimming	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
Male RNA-Seq	131.4	26.3	128.3	25.4
Female RNA-Seq	77.2	15.4	75.8	15.0

Table 4.9 RNA-Seq assembly statistics for male *Gordius* sp.

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	69,831	31,100	9,180	6,408
Longest transcript (bp)	14,420	11,403	14,420	11,403
Median transcript (bp)	526	909	1,156	1,206

Table 4.10 RNA-Seq assembly statistics for female *Gordius* sp.

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	29,726	16,504	8,216	6,575
Longest transcript (bp)	14,195	11,382	14,195	11,382
Median transcript (bp)	625	933	1,286	1,206

Table 4.11 RNA-Seq assembly statistics for pooled *Gordius* sp.

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	75,839	29,381	8,888	6,675
Longest transcript (bp)	14,420	11,364	14,420	11,364
Median transcript (bp)	448	894	1,260	1,218

4.3 Nematode species

Phylum . Nematoda Potts, 1932

4.3.1 *Enoplus brevis*

Class . . . . . Enoplia Pearse, 1942
Order . . . . . Enoplida Filipjev, 1929
Suborder . . . . . Enoplina Chitwood and Chitwood, 1937
Superfamily . . . . Enoploidea Dujardin, 1845
Family . . . . . Enoplidae Dujardin, 1845
Genus . . . . . Enoplus Dujardin, 1845
Species . . . . . Enoplus brevis Bastian, 1865

E. brevis is a member of the order Enoplida, an early splitting clade of the phylogenetic tree of the phylum Nematoda. It is the first species from Clade II with NGS data, and phylogenetically important to resolve the topology at the base of the tree. The earliest branching clade may give a hint of the ancestral habitat of the phylum.

E. brevis is a free-living marine nematode found in estuarine mud within the depth of oxygen penetration. Embryonic development of *E. brevis* varies drastically from *Caenorhabditis elegans*, with no visible cell lineage until the 8-cell stage [111]. Furthermore, embryogenesis advances very slowly with long reproductive cycles [112]. This development indicates an adaptive mechanism to the potential environmental changes in the living conditions of the species.

Raw data

E. brevis specimens were collected from the island Sylt located in northern Germany by Einhard Schierenberg. Two PE genomic libraries of 300 bp (lib300) and 600 bp (lib600) insert size and one short-insert PE RNA-Seq library were prepared by

Edinburgh Genomics, and sequenced on Illumina HiSeq2000 using 100 b PE sequencing. Raw sequence fastq files are not publicly available yet. The quality of Illumina reads was checked with FASTQC, genomic reads from the 600 bp library contained an excess amount of {CA}X dimers reducing the abundance of informative k-mers. Raw reads were filtered using Trimmomatic. The trimmed reads were then digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.12).

Genome Assembly

The insert size distribution of lib300 on the preliminary SE assembly (assembly name: Eb_CLC_SE) has a median of 271 bp (58 SD), and the insert size distribution of lib600 has a median of 216 bp (202 SD) (Fig. 4.4). Lib600 has a large SD because the fragments selected for the library seem to be in two sizes, at 200 bp and at 550 bp. In addition, fragment insert size estimations for lib600 are probably underrepresented because Eb_CLC_SE's N50 is also low (702 bp). The TAGC plot shows a very fragmented assembly with no observable contamination (Fig. 4.6).

The digitally normalised reads were used in Velvet with a kmer size of 51 (assembly name: Eb_Velvet_PE). Similar to Eb_CLC_SE, the N50 is also low (Fig. 4.5, Table 4.13). Eb_Velvet_PE has a median coverage of 7.

Transcriptome Assembly

A total of 257,040 contigs were generated by Trinity. Gene and isoform expression levels were calculated with RSEM and 20,196 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.14).

Synopsis

Both assemblies were highly fragmented. Therefore, it was decided to not advance into the annotation pipeline. However, the transcriptome assembly shows a high degree of CEGMA completeness (95.5%) and the transcripts were useful for phylo-

genetic analyses. The fragmentation of the assemblies can be attributed to high level of polymorphism which was identified by aligning the contigs against each other. One possible solution could be the use of Whole-Genome Amplification (WGA) to lower the level of polymorphism.

Table 4.12 Read data for *E. brevis*

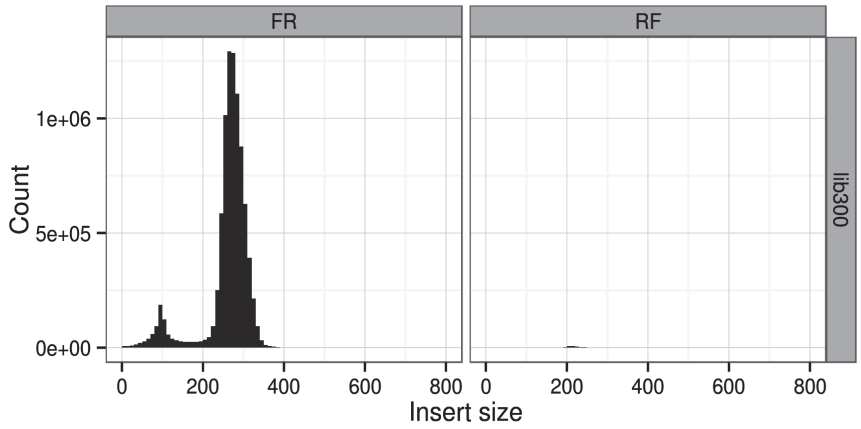
Library	Raw		Post Trimming		Post Khmer	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib300	89.0	17.8	79.4	15.3	39.4	7.6
PE lib600	71.9	14.4	56.2	10.5	26.1	4.9
RNA-Seq	189.4	37.9	180.3	35.5	—	—

Table 4.13 Comparison of assemblies for *E. brevis*

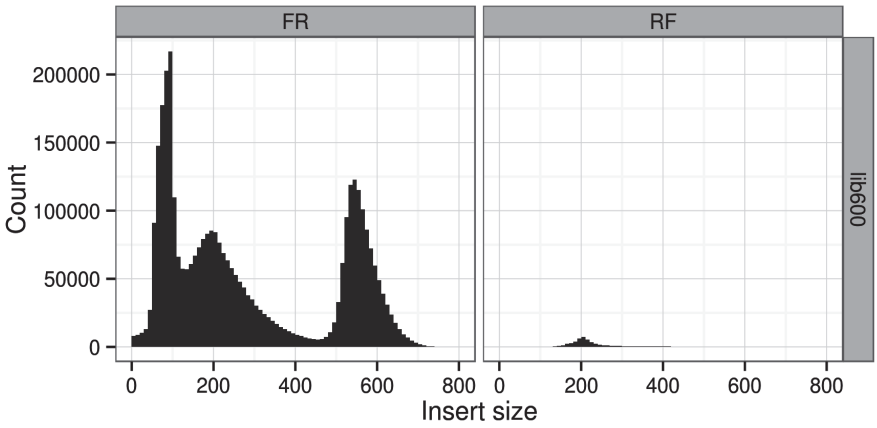
	Eb_CLC_SE	Eb_Velvet_PE
Number of scaffolds	233,733	165,525
Longest scaffold (bp)	16,008	14,390
Assembly span (bp)	172,568,232	126,600,175
Number of N's (bp)	0	8,748,362
Mean scaffold length (bp)	738	764
Scaffold N50 (bp)	702	753
GC content (%)	40.2	40.7

Table 4.14 RNA-Seq assembly statistics for *E. brevis*

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	257,040	89,204	20,196	13,683
Longest transcript (bp)	20,592	19,401	20,592	19,401
Median transcript (bp)	382	681	887	903



(a) lib300



(b) lib600

Figure 4.4 Insert size estimations for the two libraries (a) lib300 and (b) lib600 of *E. brevis*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

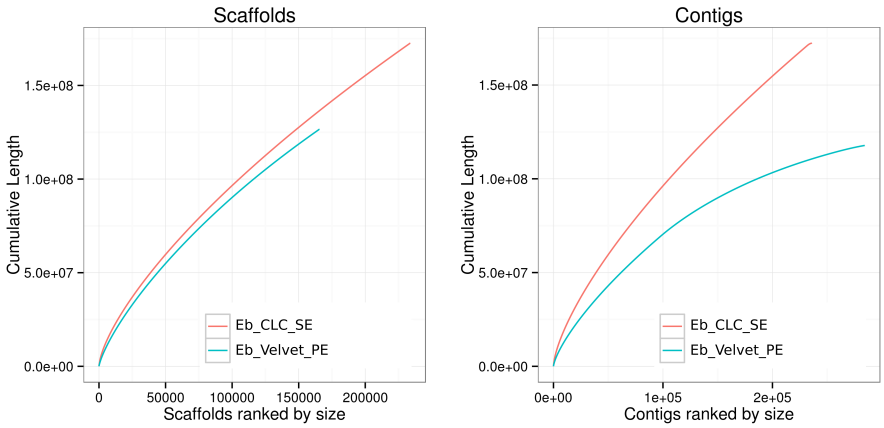


Figure 4.5 Scaffold and contig length cumulative curves for *E. brevis* assemblies. Steeper curves indicate better assembly contiguity.

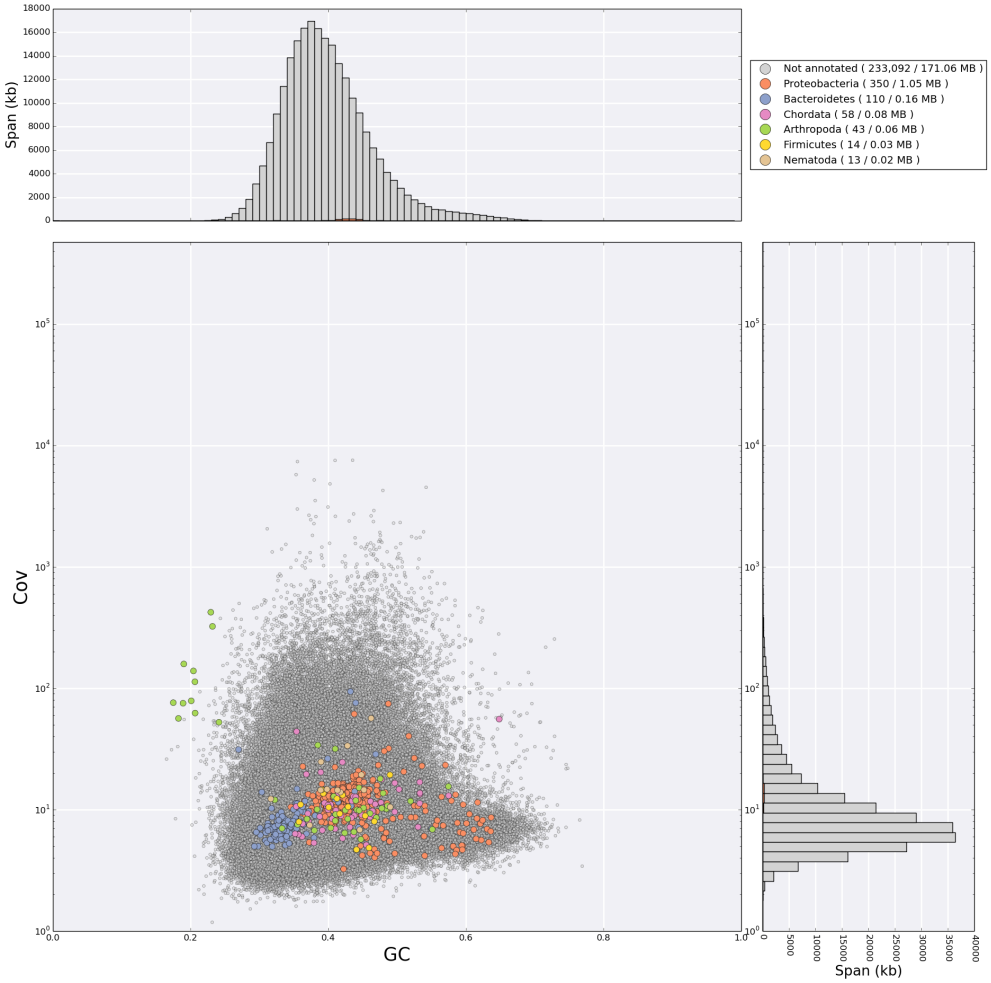


Figure 4.6 TAGC plot for Eb_CLC_SE

4.3.2 *Prionchulus punctatus*

Class . . . . . Dorylaimia Inglis, 1983
Order . . . . . Mononchida Jairajpuri, 1969
Suborder . . . . . Mononchina Kirjanova and Krall, 1969
Superfamily . . . . Mononchoidea Chitwood, 1937
Family . . . . . Mononchidae Chitwood, 1937
Genus . . . . . Prionchulus Cobb, 1916
Species . . . . . Prionchulus punctatus Cobb, 1917

P. punctatus belongs to the order Mononchida which prior to this study was not represented with any NGS data. *P. punctatus* is thus crucial to resolve the relationships between orders of Clade I, but also the relationships between the three classes.

P. punctatus is a free-living nematode found widely in European and North American soils. The members of the order Mononchida are predatory species that feed on nematodes, rotifers, protozoans, algae and fungal spores. These nematodes have a larger mouth cavity and *P. punctatus* has been shown to reduce the populations of plant-parasitic nematodes in pots [113]. Although there is the potential that mononchid nematodes can be used as biocontrol agents for plant-parasitic nematodes, their lack of prey specificity is inefficient for controlling specific pest nematodes [114].

Raw data

A *P. punctatus* RNA sample was provided by Bernadette Connolly from the University of Aberdeen. One short-insert PE RNA-Seq library was prepared by Edinburgh Genomics on an Illumina HiSeq2000 instrument using 101 b PE sequencing. Raw sequence fastq files are not publicly available yet. The quality of reads was assessed with FASTQC, and no problems were detected. Raw reads were filtered using Trimmomatic (Table 4.15).

Transcriptome assembly

A total of 109,445 contigs were generated by Trinity. Gene and isoform expression levels were calculated with RSEM and 21,290 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.16).

Synopsis

The transcripts predicted from the RNA-Seq experiment show a high degree of CEGMA completeness (94.7%) and were useful for phylogenetic analyses. The transcriptome assembly is not publicly available yet.

Table 4.15 Read data for *P. punctatus*

Library	Raw		Post Trimming	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
RNA-Seq	75.1	15.0	63.7	11.5

Table 4.16 RNA-Seq assembly statistics for *P. punctatus*

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	109,445	37,238	21,290	14,651
Longest transcript (bp)	16,595	9,588	16,595	9,588
Median transcript (bp)	296	588	662	741

4.3.3 *Plectus murrayi*

Class.....Chromadoria Pearse, 1942
Order.....Plectida Malakhov, 1982
Superfamily...Plectoidea Örley, 1880
Family.....Plectidae Örley, 1880
Genus.....Plectus Bastian, 1865
Species.....Plectus murrayi Yeates, 1970

P. murrayi belongs to the order Plectida, and is consider a close outgroup to the Rhabditida order (Clades III, IV, and V) [5]. Prior to this study, only the order Rhabditida from the class Chromadoria had species with NGS data. The relationships between the suborders of the Rhabditida have not been resolved, possibly because of the lack of intermediate taxa between the Rhabditida and Clades I and II. The phylogenetic information from plectid species may play a key role to achieve a better resolution.

P. murrayi is free-living bacterivorous nematode endemic to terrestrial Antarctica and occurs in the McMurdo Dry Valleys in places with high soil moisture. Due to the extreme weather conditions, species in Dry Valleys may require several years to complete their life cycles [115]. Genomic resources can help in studying extreme environmental survival, and the resistance to environmental stresses in adverse conditions.

Raw data

One PE genomic library of 500 bp insert size (lib500) and one short-insert PE RNA-Seq library were generated from *P. murrayi* culture maintained by Byron Adams in Brigham Young University. All reads are 100 b long. Raw sequence fastq files are not available yet. The quality of Illumina reads was assessed with FASTQC, genomic reads from the first pair had over-representation of adapter contamination.

Raw reads were filtered using fastq-mcf. The trimmed reads were then digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.17).

Genome Assembly and Annotation

The insert size distribution of the lib500 on the preliminary SE assembly (assembly name: Pm_CLC_SE) has a median of 450 bp (117 SD), with a significant number of paired reads below 400 bp insert size (Fig. 4.7). The TAGC plot shows a Proteobacteria cluster at around 20X coverage and 68% GC, while genomic sequences show a wide range of GC between 40% and 55% at ~80X coverage and a wide range of coverage at 40% GC (Fig. 4.9).

The digitally normalised reads were used in Velvet with a kmer size of 51 (assembly name: Pm_Velvet_PE). The Velvet assembly (without contamination removal) spans 167.4 Mb with median coverage of 75X. In terms of contiguity, the assembly is of moderate quality with an N50 of 10 kb (Fig. 4.8). There are 33,505 contigs above 500 bp. The assembly has a GC content of 44.1% and 85.9% CEGMA completeness (Table 4.18).

Transcriptome assembly

A total of 74,719 contigs were generated by Trinity. Gene and isoform expression levels were calculated with RSEM and 24,055 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.19).

Synopsis

High level of polymorphism was identified in *P. murrayi* contigs, by aligning the contigs against each other. Therefore, it was decided to assemble the genome using dipSPAdes [116], a polymorphic aware de Bruijn graph assembler. Unfortunately, the assembly has not finished in a manageable timetable to be included in this thesis. Instead the transcripts predicted from the RNA-Seq experiment were used in the

downstream phylogenetic analyses, since the transcriptome assembly showed a high degree of CEGMA completeness (95.1%).

Table 4.17 Read data for *P. murrayi*

Library	Raw		Post Trimming		Post Khmer	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib500	93.3	18.7	86.6	17	43.8	8.6
RNA-Seq	59.3	11.9	57.3	11.3	—	—

Table 4.18 Comparison of assemblies for *P. murrayi*

	Pm_CLC_SE	Pm_Velvet_PE
Number of scaffolds	77,366	33,505
Longest scaffold (bp)	251,162	167,436
Assembly span (bp)	199,970,644	185,018,654
Number of N's (bp)	0	8,907,930
Mean scaffold length (bp)	2,584	5,522
Scaffold N50 (bp)	4,342	10,183
GC content (%)	45.7	44.1
CEGMA completeness (%)	87.5	85.9
Transcriptome completeness (%)	73.1	82.7

Table 4.19 RNA-Seq assembly statistics for *P. murrayi*

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	74,719	48,576	24,055	17,691
Longest transcript (bp)	20,179	19,767	15,018	14,463
Median transcript (bp)	717	939	821	912

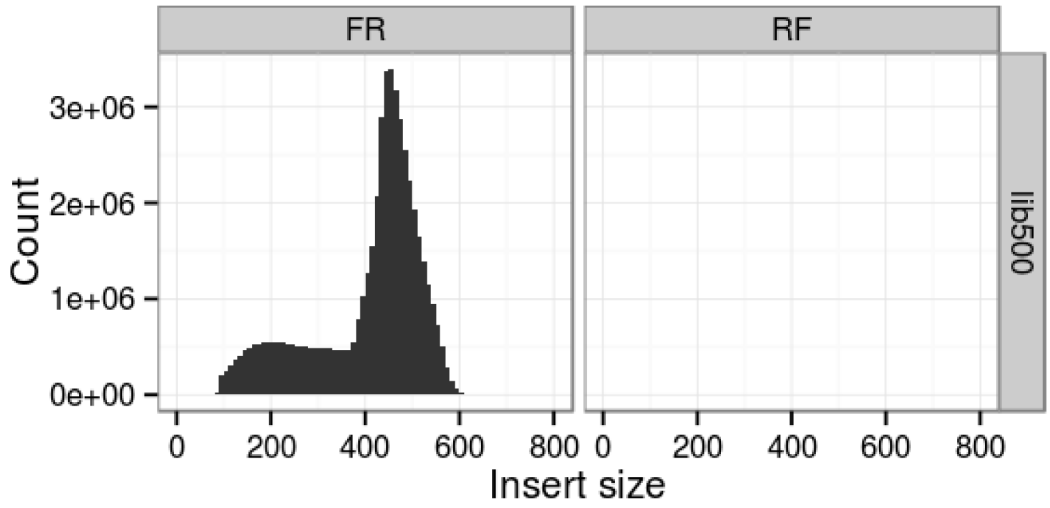


Figure 4.7 Insert size estimation for lib500 of *P. murrayi*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

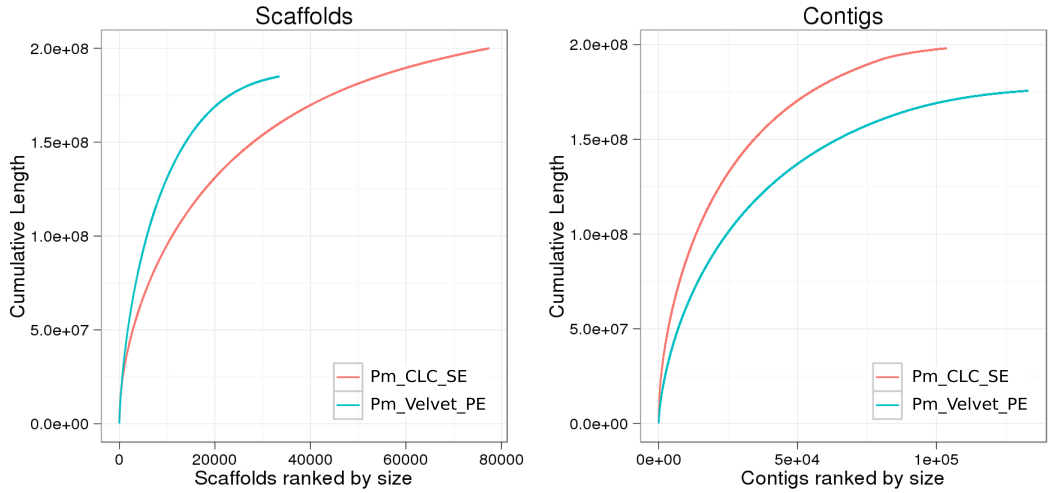


Figure 4.8 Scaffold and contig length cumulative curves for *P. murrayi* assemblies. Steeper curves indicate better assembly contiguity.

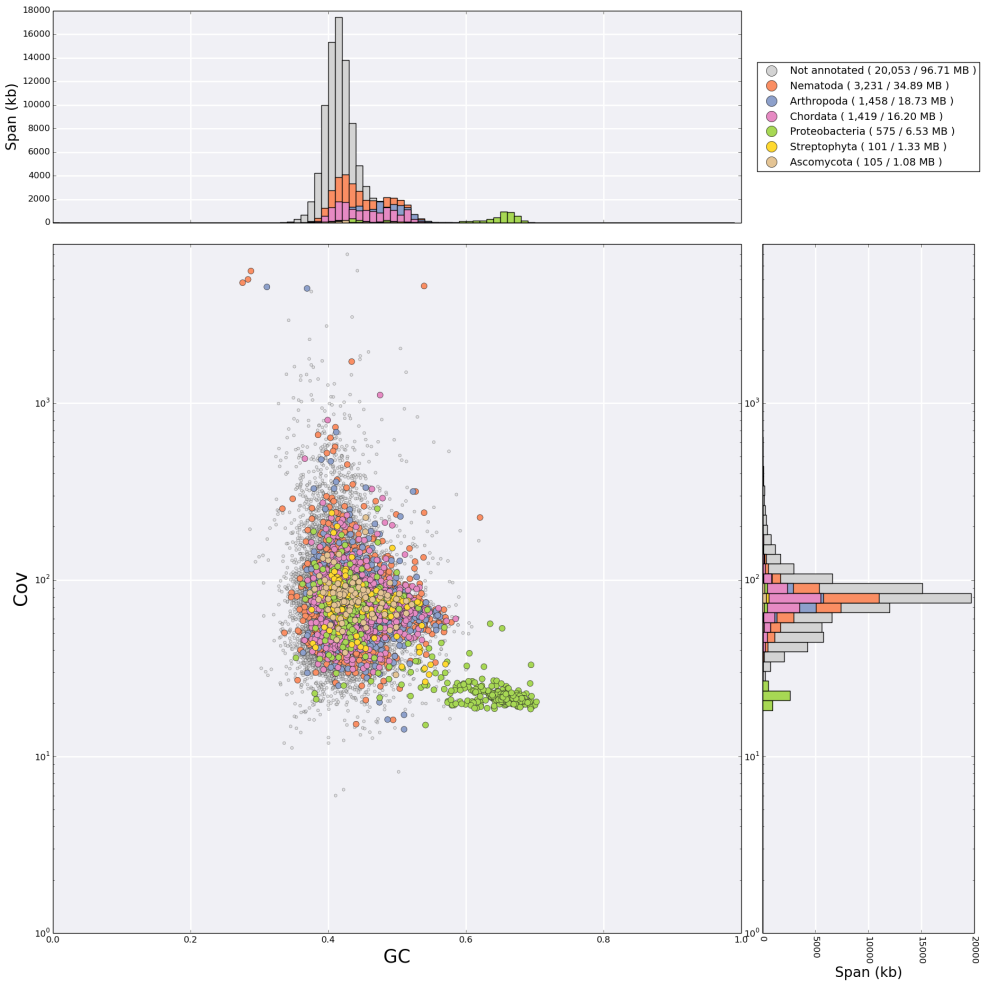


Figure 4.9 TAGC plot for Pm_Velvet_PE prior to contamination removal. Contigs with low coverage and above 60% GC belong to Proteobacteria. The GC percentage spread of the nematode contigs (80X coverage) was also present in *Plectus sambesii* (Fig. 4.12).

4.3.4 *Plectus sambesii*

Class.....Chromadoria Pearse, 1942
Order.....Plectida Malakhov, 1982
Superfamily...Plectoidea Örley, 1880
Family.....Plectidae Örley, 1880
Genus.....Plectus Bastian, 1865
Species.....Plectus sambesii Micoletzky, 1916

P. sambesii also belongs to the order Plectida and it is a free-living bacterial feeder with global distribution found predominantly in high moisture soil. The species can be cultured in lab conditions, and are being used for comparative embryonic studies due to their fast life cycle and high reproductive rate [117].

Raw data

Genomic DNA of *P. sambesii* was provided by Philipp Schiffer from the University of Cologne. Two short-insert PE genomic libraries of 400 bp (lib400) and 600 bp (lib600) insert size were prepared by Edinburgh Genomics on HiSeq2000, across two lanes per library, using 100 b PE sequencing. Raw sequence fastq files are not publicly available yet. The quality of Illumina reads was checked with FASTQC, showing a double peak at the GC% distribution of reads indicating high levels of contamination. Raw reads were filtered using Trimmomatic. Table 4.20 shows the reduction on raw data due to trimming.

Genome Assembly

The insert size distribution of lib400 on the preliminary SE assembly has a median of 341 bp (41 SD), and the insert size distribution of lib600 has a median of 586 bp (141 SD) (Fig. 4.10). The TAGC plot shows a heavily contaminated sample with lots of distinct clusters of bacteria at various coverage values and GC percentages (Fig. 4.12). Median coverage for the nematode sequences was measured at 8.77.

The reads were assembled with CLC-bio (assembly name: Ps_CLC_PE) using the insert size distribution parameters obtained in the previous step. Bacteria contaminants assembled better resulting in an inflated N50 of 9,794 (Fig. 4.11) and bigger than expected assembly span (239 Mb) (Table 4.21). The longest contig has 1,784,255 bp and belongs to a Bacteria species. The assembly has a GC content of 53.1 which, with reference to the TAGC plot, is in between the bacteria and the nematode contigs.

Synopsis

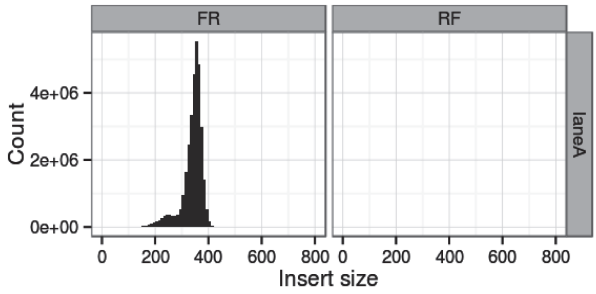
Due to high levels of contamination the nematode assembly has low coverage, resulting in high fragmentation. Therefore, it was deemed inadequate for advancing into the annotation pipeline. Initially, *P. murrayi* transcripts were going to be used to extract similar sequences from *P. sambesii*. However, an RNA-Seq dataset was generated by Philipp Schiffer, and the transcriptome assembly was used in the phylogenetic analyses. Unfortunately, it will be very difficult to generate an adequate genome assembly with the current libraries and thus were not used in the phylogenetic analyses. A cleaner DNA sample and new libraries are required.

Table 4.20 Read data for *P. sambesii*

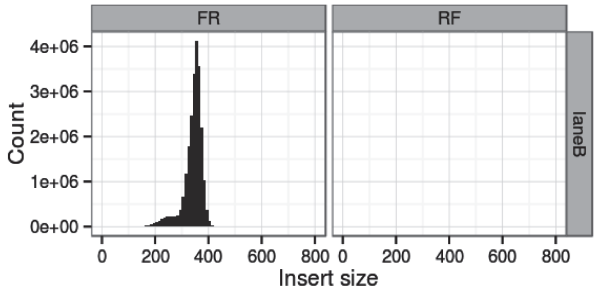
Library	Raw		Post Trimming	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib300 LaneA	34.5	6.9	34.0	6.7
PE lib300 LaneB	25.4	5.1	24.7	4.9
PE lib600 LaneC	24.5	4.9	23.7	4.6
PE lib600 LaneD	25.4	5.1	23.8	4.7

Table 4.21 Comparison of assemblies for *P. sambesii*

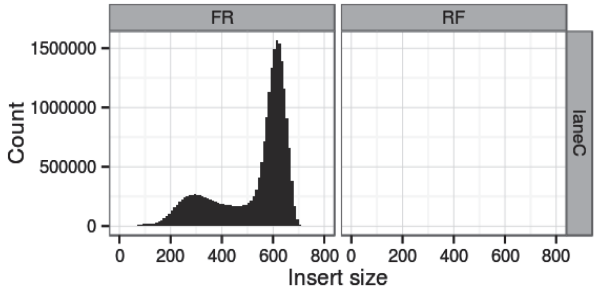
	Ps_CLC_PE
Number of scaffolds	73,716
Longest scaffold (bp)	1,784,255
Assembly span (bp)	239,122,110
Number of N's (bp)	1,027,808
Mean scaffold length (bp)	3,243
Scaffold N50 (bp)	9,794
GC content (%)	53.1
CEGMA completeness (%)	56.5



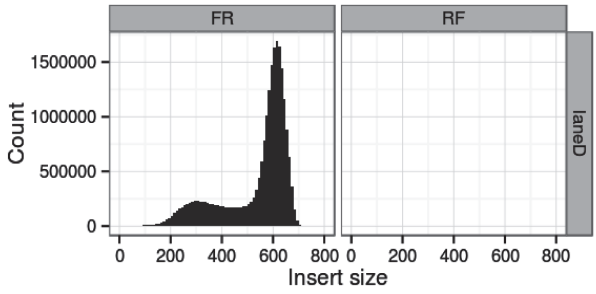
(a) lib400 laneA



(b) lib400 laneB



(c) lib600 laneC



(d) lib600 laneD

Figure 4.10 Insert size estimations for the two libraries lib400 ((a) laneA, (b) laneB) and lib600 ((c) laneC, (b) laneD) of *P. sambesii*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

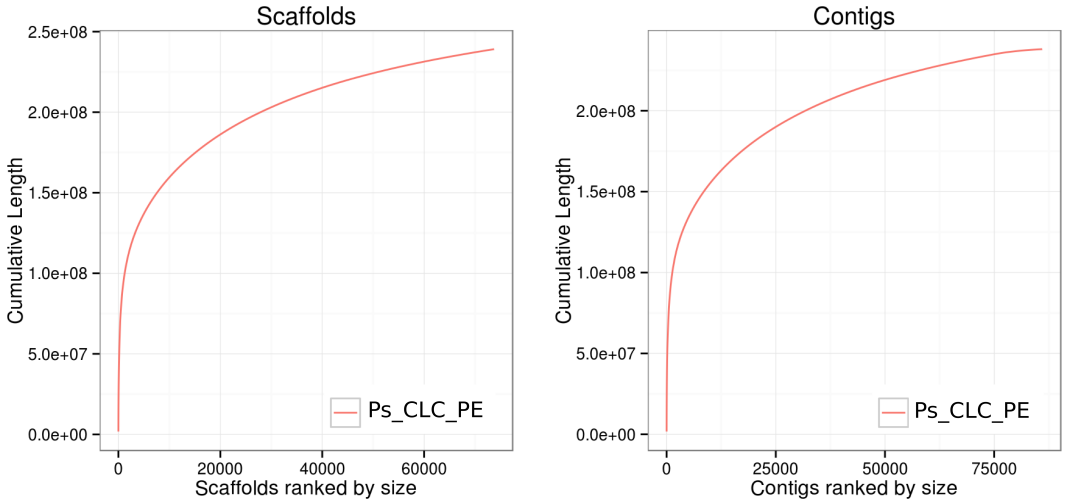


Figure 4.11 Scaffold and contig length cumulative curves for *P. sambesii* assembly. Steeper curves indicate better assembly contiguity.

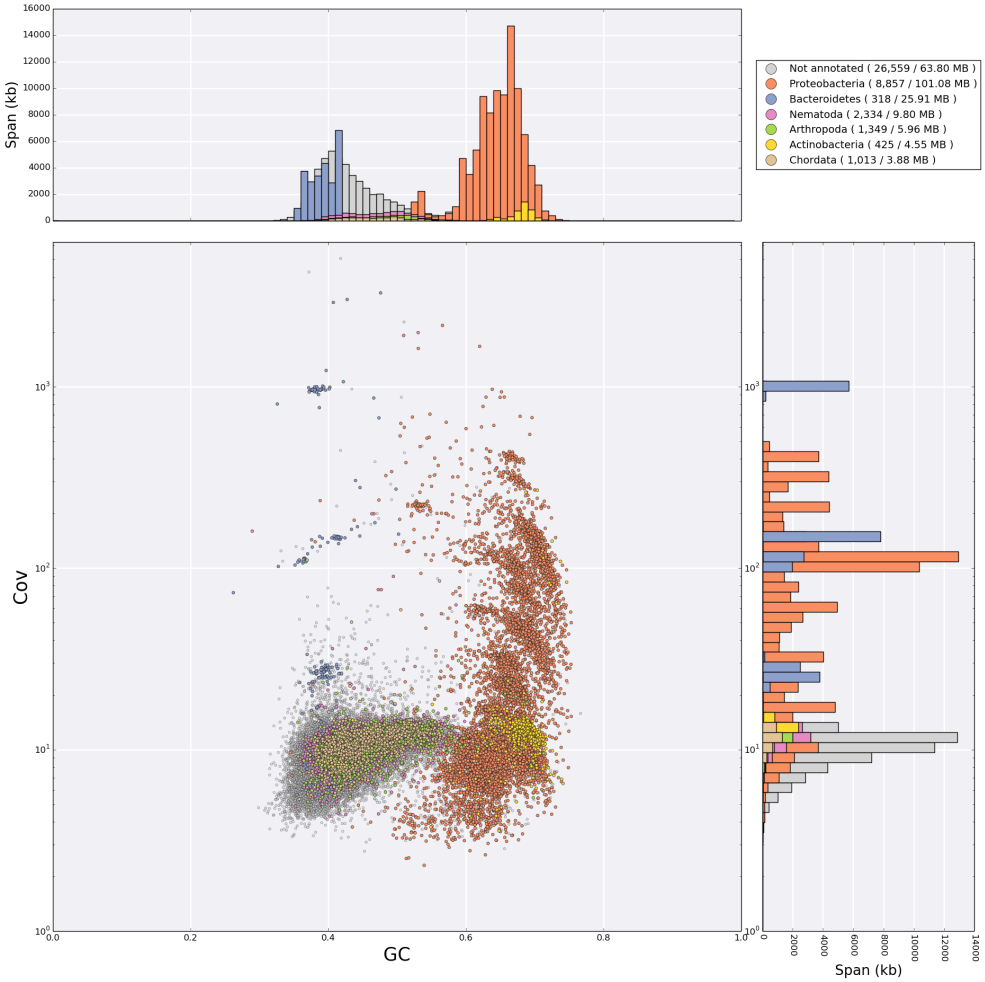


Figure 4.12 TAGC plot for Ps_CLC_PE. Multiple Bacteria contaminants are visible in distinct clusters.

4.3.5 *Setaria labiatopapillosa*

Class ..... Chromadoria Pearse, 1942
Order ..... Rhabditida Chitwood, 1933
Suborder ..... Spirurina Linnaeus, 1758
Infraorder ..... Spiruromorpha De Ley and Blaxter, 2002
Superfamily ..... Filarioidea Weinland, 1858
Family ..... Setariidae Skrjabin and Schikhobalova, 1945
Genus ..... Setaria Viborg, 1795
Species ..... Setaria labiatopapillosa Alissandrini, 1838

S. labiatopapillosa is a member of the superfamily Filarioidea which includes animal parasitic species. In contrast to other members of the superfamily, these nematodes probably lack *Wolbachia*, an endosymbiont bacterium. Phylogenetically, this species is interesting because it has been hypothesized that the acquisition of the *Wolbachia* species occurred after the split of the Setariidae [118]. Furthermore, it can be used as a close outgroup to the Onchocercidae.

This species infects cattle and the adults mature in the peritonal cavity of the host. It is not pathogenic [119]. Erratic migration of larvae may occur after transmission from the intermediate host (various mosquito species).

Raw data

S. labiatopapillosa RNA samples were provided by Benjamin Makepeace from the University of Liverpool from nematodes isolated in Cameroon. Three short-insert PE RNA-Seq libraries were prepared by Edinburgh Genomics on an Illumina HiSeq2000 instrument using 100 b PE sequencing. Raw sequence fastq files are not publicly available yet. The quality of the reads was assessed with FASTQC, and no problems were detected. Due to time constrains, only one sample (from female specimens) was considered for downstream analyses. Raw reads were filtered using Trimmomatic (Table 4.22).

Transcriptome assembly

A total of 107,776 contigs were generated by Trinity. Gene and isoform expression levels were calculated with RSEM and 14,283 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.23).

Synopsis

The transcripts predicted from the RNA-Seq experiment show a high degree of CEGMA completeness (92.3%), and thus were useful for phylogenetic analyses. Although only one sample was used, it also includes embryos and microfilariae. *Wolbachia* sequences were not present in the transcriptome assembly. However, genomic sequences are needed to verify the absence of the bacteria (present or past). The transcriptome assembly is not publicly available yet.

Table 4.22 Read data for *S. labiatopapillosa*

Library	Raw		Post Trimming	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
RNA-Seq	100.1	20	97.5	19.2

Table 4.23 RNA-Seq assembly statistics for *S. labiatopapillosa*

	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	107,776	90,893	14,283	8,642
Longest transcript (b)	15,824	9,171	13,177	9,171
Median transcript (b)	1,205	738	1,049	1,005

4.3.6 *Acanthocheilonema viteae*

Class ..... Chromadoria Pearse, 1942
Order ..... Rhabditida Chitwood, 1933
Suborder ..... Spirurina Linnaeus, 1758
Infraorder ..... Spiruromorpha De Ley and Blaxter, 2002
Superfamily ..... Filarioidea Weinland, 1858
Family ..... Onchocercidae Leiper, 1911
Genus ..... Acanthocheilonema Cobbold, 1870
Species ..... Acanthocheilonema viteae Krepkogorskaya, 1933

A. viteae, previously known as *Dipetalonema viteae*, is a parasite of rodents. Interestingly, *A. viteae* has probably lost the *Wolbachia* bacterial endosymbiont found in most members of the family Onchocercidae [120]. The phylogenetic position of *A. viteae* within the same family will provide insights into the mechanisms of the *Wolbachia* symbiosis.

The life cycle of the species can be completed under lab conditions, and it is a widely used model for human filariases. *A. viteae* is transmitted by soft ticks (*Ornithodoros moubata*) to the host where L3 stages settle under the skin and develop into adults.

Raw Data

The *A. viteae* DNA sample was provided by Kenneth Pfarr from the University of Bonn in Germany. One PE genomic library of 300 bp (lib300) insert size was prepared by Edinburgh Genomics, and sequenced on HiSeq2000 across three lanes, using 100 b PE sequencing. Sequence fastq files were submitted to the Short Read Archive (SRA) with accession number PRJEB1697. The quality of Illumina reads was checked with FASTQC, and no problems were detected. Raw reads were filtered using fastq-mcf. The trimmed reads were then concatenated and digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.24).

Genome Assembly and Annotation

The insert size distribution of lib300 on the preliminary SE assembly (assembly name: Av_CLC_SE) has a median of 275 bp (41 SD) (Fig. 4.13). The TAGC plot shows contamination with Chordata (*Macaca mulatta*) and Arthropoda (*Heliconius melpomene*) (Fig. 4.15). Furthermore, contigs with Proteobacteria sequences (genus *Wolbachia*) were present with similar GC and coverage as the nematode contigs.

The digitally normalised reads were used in ABySS with a kmer size of 39, and contigs with macaque or butterfly contamination were removed (assembly name: Av_ABySS_PE). Av_ABySS_PE spans 77.4 Mb with median coverage of 375X. In terms of contiguity, the assembly is of moderate quality with an N50 of 26 kb (Fig. 4.14). There are 6,796 contigs above 500 bp. The assembly has a GC content of 29.9% and 93.5% CEGMA completeness (Table 4.25).

Av_ABySS_PE was then annotated using the MAKER2-Augustus pipeline. The final step predicted 10,397 protein-coding genes, with a median length of 975 bp, median exon length of 136 bp and a median of 7 exons per gene (Table 4.26).

Synopsis

Macaque contamination was a consequence of *A. viteae* species feeding on macaque cells in tissue culture, while the butterfly contamination was probably introduced in the sequencing facility. Av_ABySS_PE's N50 was adequate for annotating the genome, and the genes predicted were useful for phylogenetic analyses. As expected, there was no *Wolbachia* genome present, but *Wolbachia* insertions could be identified in *A. viteae* genome indicating a former symbiosis. MP libraries and long reads from new sequencing platforms can improve the assembly further, and RNA-Seq experiments can improve the annotation. The genome assembly and annotations are available at http://nematod.es.

Table 4.24 Read data for *A. viteae*

Library	Raw		Post Trimming		Post Khmer[†]	
	Reads (M pairs)	Bases (Gbp)	Reads (M pairs)	Bases (Gbp)	Reads (M pairs)	Bases (Gbp)
PE lib300 laneA	39.2	7.8	38.7	7.6		
PE lib300 laneB	69.0	13.8	68.3	13.5	23.6	4.4
PE lib300 laneC	131.2	26.2	128.6	25.0		

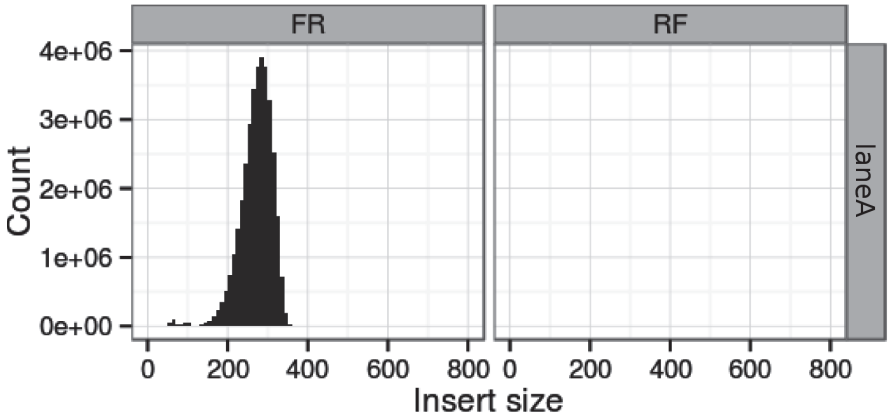
[†] All lanes concatenated together

Table 4.25 Comparison of genome assemblies for *A. viteae*

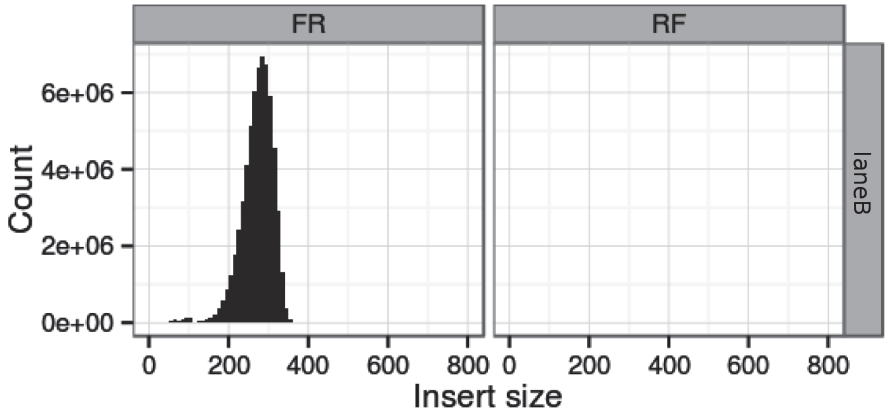
	Av_CLC_SE	Av_ABySS_PE
Number of scaffolds	14,310	6,796
Longest scaffold (bp)	146,491	172,453
Assembly span (bp)	75,109,148	77,350,906
Number of N's (bp)	0	173,632
Mean scaffold length (bp)	5,248	11,381
Scaffold N50 (bp)	12,142	25,808
GC content (%)	30 .0	29.9
CEGMA completeness (%)	88.7	93.5

Table 4.26 Genome annotation statistics for *A. viteae*

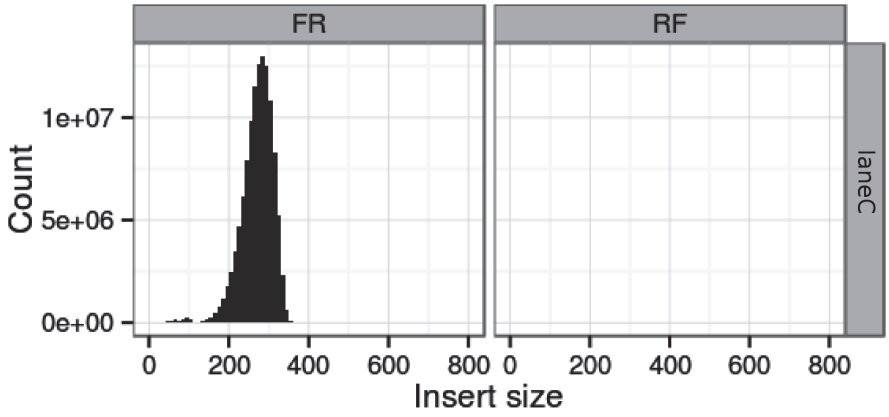
	Predicted
Number of genes	10,397
Longest transcript (bp)	21,495
Median transcript length (bp)	975
Median exon length (bp)	136
Median exons per gene	7
Median intron length (bp)	226



(a) laneA



(b) laneB



(c) laneC

Figure 4.13 Insert size estimations for the three lanes of lib300 (a) laneA, (b) laneB and (c) laneC of *A. viteae*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

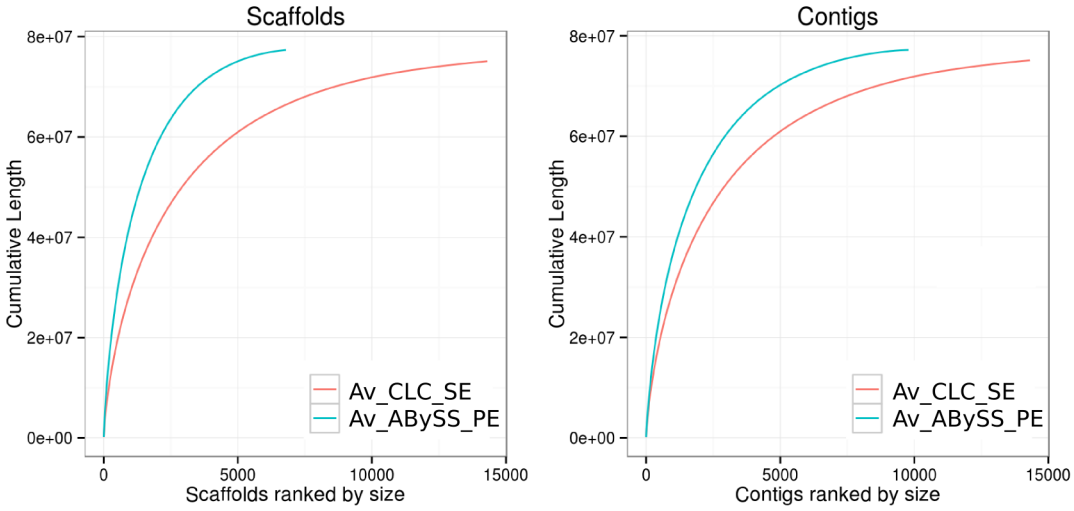


Figure 4.14 Scaffold and contig length cumulative curves for *A. viteae* assemblies. Steeper curves indicate better assembly contiguity.

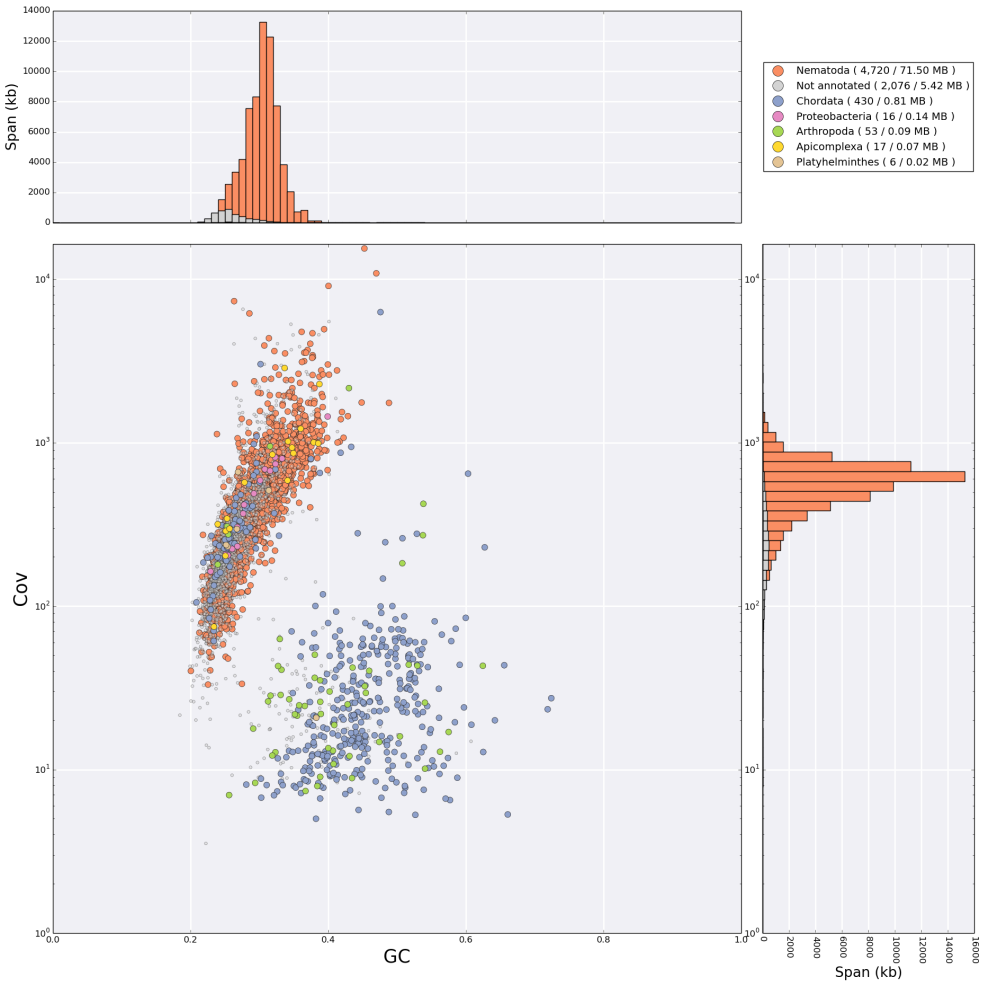


Figure 4.15 TAGC plot for Av_ABySS_PE prior to contaminant removal. Contigs with low coverage and in the 30%-60% GC range are the Chordata and Arthropoda contaminants.

4.3.7 *Onchocerca gutturosa*

Class ..... Chromadoria Pearse, 1942
Order ..... Rhabditida Chitwood, 1933
Suborder ..... Spirurina Linnaeus, 1758
Infraorder ..... Spiruromorpha De Ley and Blaxter, 2002
Superfamily ..... Filarioidea Weinland, 1858
Family ..... Onchocercidae Leiper, 1911
Genus ..... Onchocerca Diesing, 1841
Species ..... Onchocerca gutturosa Neumann, 1910

O. gutturosa is a member of the family Onchocercidae. The position of *O. gutturosa* within the genus *Onchocerca* (data from two additional species are used in the phylogenetic analyses, *O. ochengi* and *O. volvulus*) will provide valuable information about the relationships between the members of the genus. The position of *Onchocerca* species within the family is crucial for understanding the evolutionary forces that mould animal parasites.

The genus *Onchocerca* consists of a number of parasitic species, with various mammalian hosts (i.e. cattle, horses, humans). *O. gutturosa* nematodes are transmitted by black flies of species *Simulium ornatum* to the ligamentum nuchae and other parts of the body of the cattle. Bovine onchocerciasis is responsible for considerable economic damage in areas with high infection. In addition, *O. gutturosa* can be used as a potential model organism for the human parasite *O. volvulus*.

Raw data

The *O. gutturosa* DNA sample was provided by Benjamin Makepeace from the University of Liverpool. One PE genomic library of 400 bp (lib400) insert size was prepared by Edinburgh Genomics, and sequenced on HiSeq2000 using 100 b PE sequencing. Raw sequence fastq files are not publicly available yet. The quality of Illumina reads was assessed with FASTQC, and over-representation of adapter sequences was detected. Raw reads were filtered using fastq-mcf. The trimmed reads

were then digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.27).

Genome assembly and Annotation

The insert size distribution of lib400 on the preliminary SE assembly (assembly name: Og_CLC_SE) has a median of 377 bp (54 SD) (Fig. 4.16). The TAGC plot (Fig. 4.18) shows contamination with Chordata (*Bos taurus*), Proteobacteria (*Wolbachia*) and Bacteroidetes.

The digitally normalised reads were used in Velvet with a kmer size of 51, and contaminant contigs were removed (assembly name: Og_Velvet_PE). Og_Velvet_PE spans 109.6 Mb with median coverage of 121X. In terms of contiguity, the assembly is of moderate quality with an N50 of 10 kb (Fig. 4.17). There are 25,605 contigs above 500 bp. The assembly has a GC content of 28.9% and 83.5% CEGMA completeness (Table 4.28).

Og_Velvet_PE was then annotated using the one-pass MAKER2 pipeline. The final step predicted 19,916 protein-coding genes, with a median length of 456 bp, median exon length of 109 bp and a median of 7 exons per gene (Table 4.29).

Synopsis

Bos taurus contamination was a consequence of *O. gutturosa* species being extracted from cow tissues, while *Wolbachia* sequences belong to the endosymbiont bacteria. Bacteroidetes contamination was probably introduced afterwards in the preparation of the sample. The assembly has a comparable N50 to *O. ochengi* (12 kb), but significantly smaller than *O. volvulus* (25 megabase pairs (Mbp)). However, 84% of the core eukaryotic genes were predicted with CEGMA, indicating that the majority of genes are exon complete, and useful for phylogenetic analysis. A *Wolbachia* genome was present, indicating a live symbiosis (also present in *O. ochengi* and *O. volvulus*. MP libraries and long reads from new sequencing platforms can improve

the assembly further, and RNA-Seq experiments can improve the annotation. The genome assembly and annotations are not publicly available yet.

Table 4.27 Read data for *O. gutturosa*

Library	Raw		Post Trimming		Post Khmer	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib400	72.5	14.5	70.8	14.0	24.4	4.8

Table 4.28 Comparison of assemblies for *O. gutturosa*

	Og_CLC_SE	Og_Velvet_PE
Number of scaffolds	37,244	25,605
Longest scaffold (bp)	55,099	286,421
Assembly span (bp)	66,251,728	109,596,146
Number of N's (bp)	0	4,669,501
Mean scaffold length (bp)	1,778	4,280
Scaffold N50 (bp)	3,045	10,567
GC content (%)	30.4	28.9
CEGMA completeness (%)	54.4	83.5

Table 4.29 Genome annotation statistics for *O. gutturosa*

	Predicted
Number of genes	19,916
Longest transcript (bp)	16,761
Median transcript length (bp)	456
Median exon length (bp)	109
Median exons per gene	7
Median intron length (bp)	212

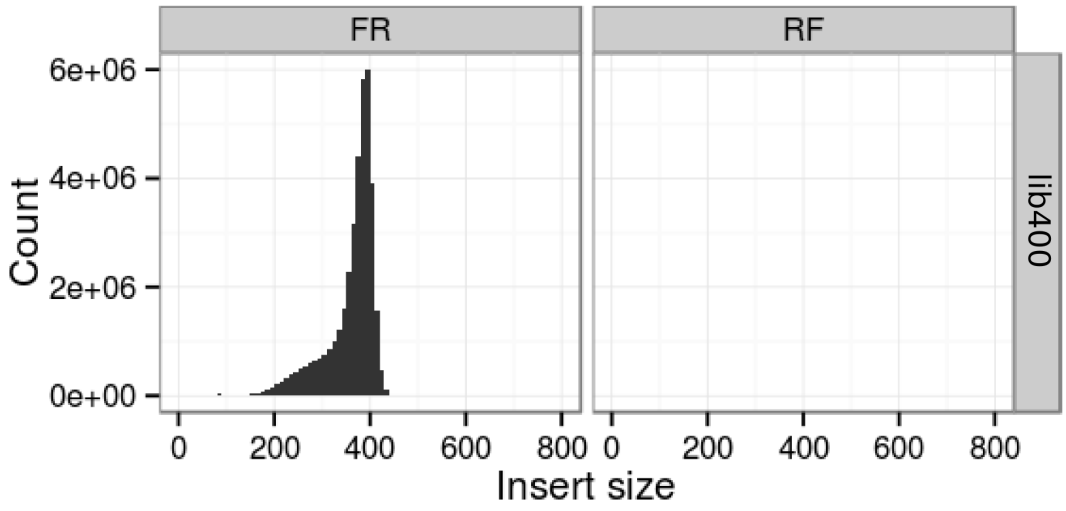


Figure 4.16 Insert size estimation for lib400 of *O. gutturosa*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

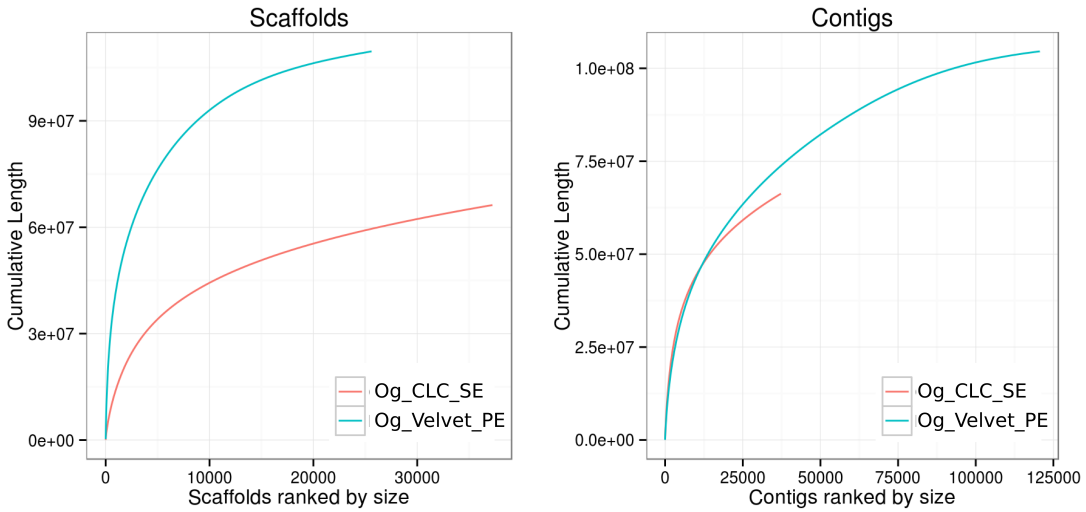


Figure 4.17 Scaffold and contig length cumulative curves for *O. gutturosa* assemblies. Steeper curves indicate better assembly contiguity.

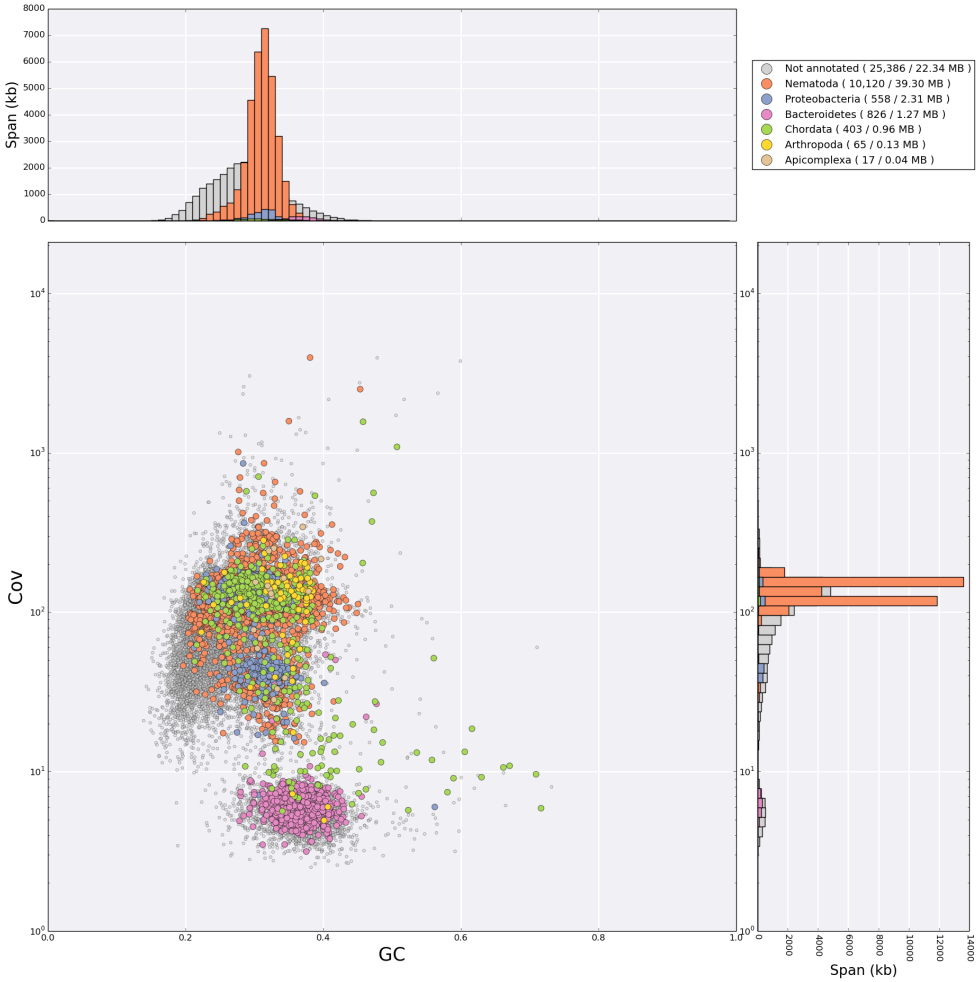


Figure 4.18 TAGC plot for Og_CLC_PE prior to contamination removal. Chordata contigs belong to host cow contamination. Proteobacteria belong to the genus *Wolbachia*, Bacteroidetes is likely a lab contamination.

4.3.8 *Dictyocaulus viviparus*

Class . . . . . Chromadoria Pearse, 1942
Order . . . . . Rhabditida Chitwood, 1933
Suborder . . . . . Rhabditina Chitwood, 1933
Infraorder . . . . . Rhabditomorpha De Ley and Blaxter, 2002
Superfamily . . . . . Strongyloidea Baird, 1853
Family . . . . . Trichostrongylidae Witenberg, 1925
Genus . . . . . Dictyocaulus Railliet and Henry, 1907
Species . . . . . Dictyocaulus viviparus Bloch, 1782

D. viviparus was chosen to broaden the phylogenetic sampling of the family Trichostrongylidea. Previous phylogenetic analyses have failed to resolve the topology of the three genera (*Dictyocaulus*, *Haemonchus*, *Nippostrongylus*) which are present in the phylogenomic dataset.

The strongylid *D. viviparus* causes parasitic bronchitis predominantly in cattle. It occurs worldwide in temperate areas and is responsible for significant economic losses, with symptoms in the affected animals varying from mild respiratory problems to more severe conditions. Adult worms occur in the bronchi of the host where they lay eggs in the lungs. The eggs may hatch in the lungs, but are usually coughed up and swallowed, and hatch as they pass through the alimentary tract of the host. Larvae mature in the faeces and invade in the sporangium of *Philobolus* sp., waiting to be dispersed during the sporulation [119].

Raw data

A male *D. viviparus* specimen was extracted from a cow being slaughtered at an abattoir in Ngaoundr, Cameroon, by Vincent N. Tanya (Institut de Recherche Agricole pour le Dveloppement, Cameroon). One short-insert PE genomic library of 400 bp (lib400) insert size was prepared by Edinburgh Genomics. It was sequenced on Illumina HiSeq2500 using 100 b PE sequencing. Raw sequence fastq files were submitted to the SRA with accession number PRJEB5116. The quality of Illumina reads was assessed with FASTQC, and no problems were detected. Raw reads were

filtered using fastq-mcf program. The trimmed reads were then digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.30).

Genome Assembly and Annotation

The insert size distribution of lib400 on the preliminary SE assembly (assembly name: Dv_CLC_SE) has a median of 383 bp (45 SD) (Fig. 4.19). The TAGC plot showed little to none bovine contamination with most of the contigs assigned to nematode classification. However a number of sequences are similar to *Wolbachia* sequences (Fig. 4.21). These sequences are not clustered together in a single region, but are scattered with different coverage values.

The digitally normalised reads were used in Velvet with a kmer size of 51, and gaps within contigs were filled using GapFiller (assembly name: Dv_Velvet_PE). Dv_Velvet_PE spans 169 Mb with median coverage of 80X. In terms of contiguity, the assembly is of moderate quality with an N50 of 22 kb (Fig. 4.20). There are 17,715 contigs above 500 bp. The assembly has a GC content of 34.5% and 90% CEGMA completeness. Biological accuracy was tested further by aligning a 454 transcriptome of *D. viviparus* [121] and 87% of the transcripts were mapped with 70% transcript coverage in one contig (Table 4.31).

The Dv_Velvet_PE assembly was then annotated using the MAKER2-Augustus pipeline. The final step predicted 14,306 protein-coding genes, with a median length of 834 bp, median exon length of 168 bp and a median of 7 exons per gene (Table 4.32).

Synopsis

Wolbachia sequences were identified within nematode contigs indicating a former symbiosis. Dv_Velvet_PE's N50 was adequate for annotating the genome, and the genes predicted were useful for phylogenetic analyses. MP libraries and long reads from new sequencing platforms can improve the assembly further. The genome

assembly and annotations are available at http://nematod.es. The genome of *D. viviparus* was published [122].

Table 4.30 Read data for *D. viviparus*

Library	Raw		Post Trimming		Post Khmer	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib400	84.3	16.9	82.4	16 .0	31.9	6.2

Table 4.31 Comparison of assemblies for *D. viviparus*

	Dv_CLC_SE	Dv_Velvet_PE
Number of scaffolds	53,968	17,715
Longest scaffold (bp)	132,281	196,229
Assembly span (bp)	158,778,867	169,388,535
Number of N's (bp)	0	448,921
Mean scaffold length (bp)	2,942	9,561
Scaffold N50 (bp)	7,069	22,560
GC content (%)	34.6	34.5
CEGMA completeness (%)	80.2	90
454 Transcriptome completeness (%)	57	87

Table 4.32 Genome annotation statistics for *D. viviparus*

	Predicted
Number of genes	14,306
Longest transcript (bp)	19,749
Median transcript length (bp)	834
Median exon length (bp)	168
Median exons per gene	7
Median intron length (bp)	172

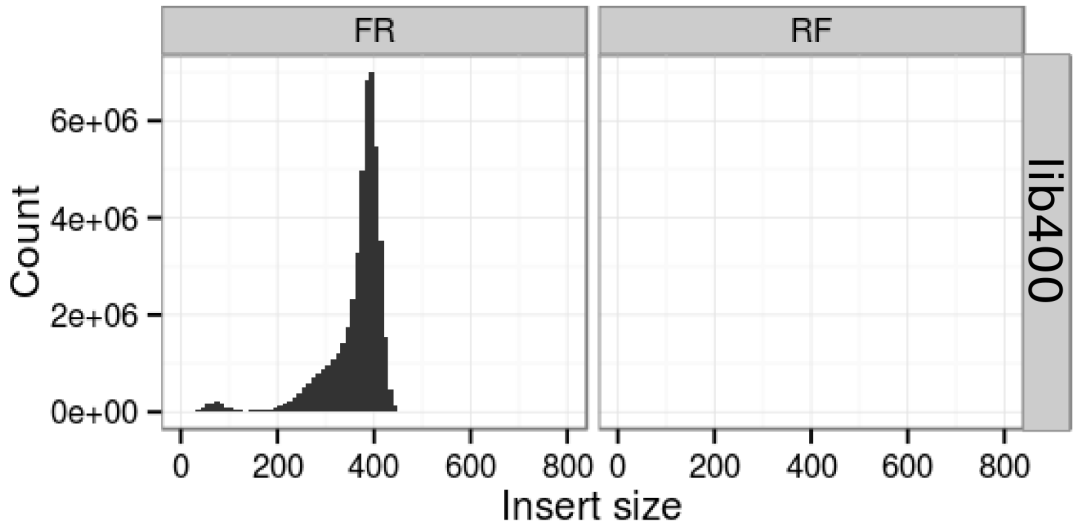


Figure 4.19 Insert size estimation for lib400 of *D. viviparus*. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

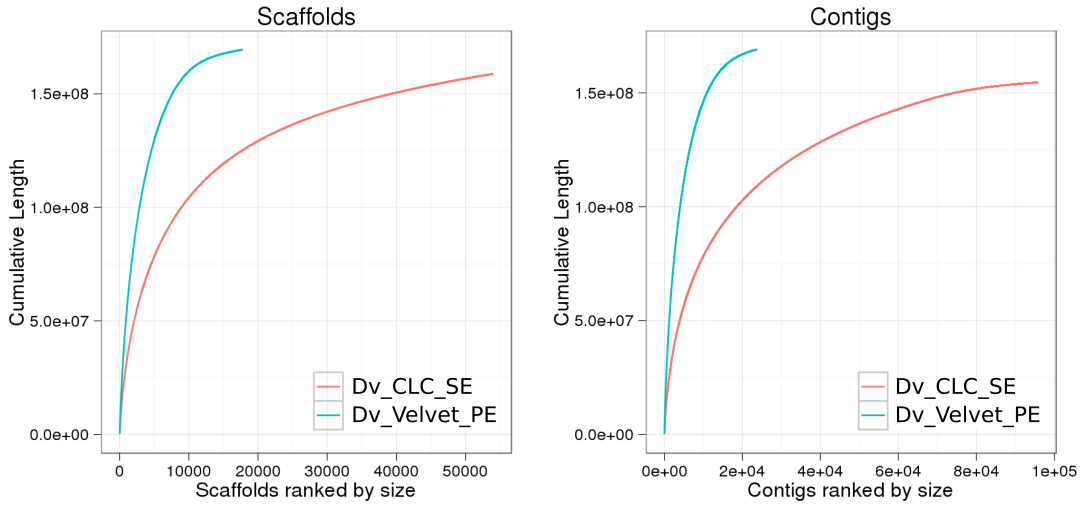


Figure 4.20 Scaffold and contig length cumulative curves for *D. viviparus* assemblies. Steeper curves indicate better assembly contiguity.

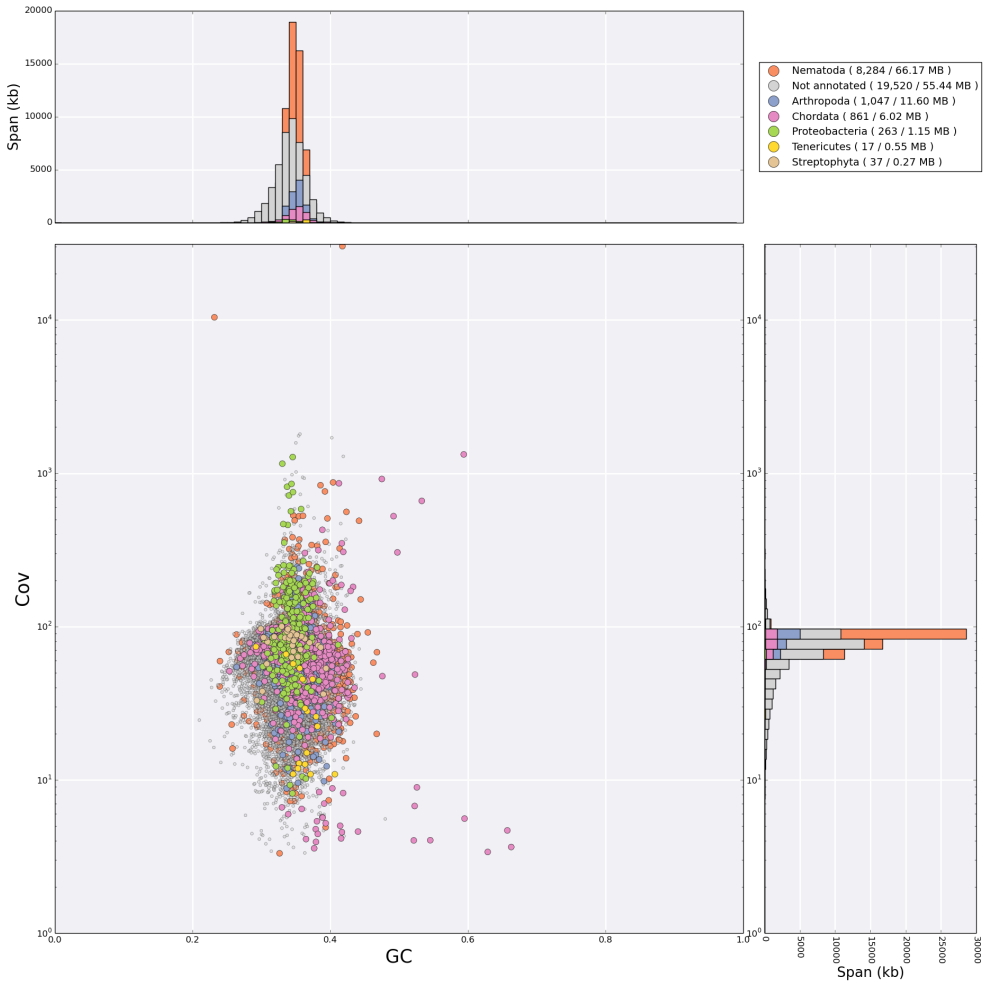
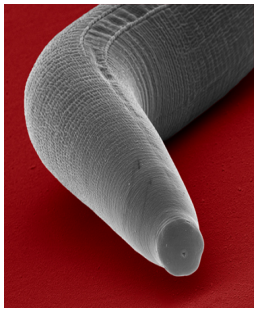


Figure 4.21 TAGC plot for Dv_Velvet_PE. *Wolbachia* insertions were identified in genomic contigs.

4.3.9 *Rhabditis* sp. SB347

Class . . . . . Chromadoria Pearse, 1942
Order . . . . . Rhabditida Chitwood, 1933
Suborder . . . . . Rhabditina Chitwood, 1933
Infraorder . . . . . Rhabditomorpha De Ley and Blaxter, 2002
Superfamily . . . . . Rhabditoidea Örley, 1880
Family . . . . . Rhabditidae Örley, 1880
Genus . . . . . Rhabditis Dujardin, 1845
Species . . . . . Rhabditis sp. SB347



Rhabditis sp. SB347 is a member of the family Rhabditidae which also includes the genus *Caenorhabditis*. Previous phylogenetic analyses have shown the family to be paraphyletic. The position of *Rhabditis* sp. SB347 will provide a better understanding of the relationships between the Rhabditomorpha in relation to *Caenorhabditis* genus.

Rhabditis sp. SB347 is a sexually polymorphic nematode that produces males, females and hermaphrodites under standard *C. elegans* culture conditions [123]. Trioecious mating systems are rare, and it has been suggested that this type of mating system acts as an intermediate step between hermaphroditism and gonochorism [124]. A similar mating system occurs in a number of parasitic species, with nematodes alternating between hermaphroditic parasites and gonochoristic free-living [125].

Raw data

Three short-insert PE genomic libraries of 250 bp (lib250), 450 bp (lib450), and 600 bp (lib600) insert size and one short-insert PE RNA-Seq library were generated, using DNA and RNA samples from an inbred line of *Rhabditis* sp. SB347 maintained by Andre Pires da Silva in University of Texas. All reads are 100 b long. Raw sequence fastq files are not publicly available yet. The quality of Illumina reads was assessed with FASTQC, and no problems were detected. Raw reads were filtered

using Trimmomatic. The trimmed reads were then digitally normalised to ~20X coverage using a kmer size of 20 with Khmer (Table 4.33).

Genome Assembly and Annotation

A preliminary SE assembly (assembly name: R_CLC_SE) using the trimmed reads was used to check the insert size distributions of the three libraries. Unexpectedly, all 3 libraries have a similar insert size of ~280 bp (lib250: 269 median and 77 SD, lib450: 295 median and 90 SD, lib600: 302 median and 93 SD) (Fig. 4.22). A PE using CLC-bio and Velvet was generated with above insert size estimations (assembly name: R_CLC_PE and R_Velvet_PE respectively). The TAGC plot shows contamination with two distinct species of Proteobacteria (*Pseudomonas fluorescens* and a *Ochrobactrum* species) (Fig. 4.24).

The digitally normalised reads were used in Velvet with a kmer size of 51 (assembly name: R_Velvet_PE_k), and contigs mapping to Proteobacteria were removed. R_Velvet_PE_k spans 70.3 Mb with median coverage of 235X. In terms of contiguity, the assembly is of good quality with an N50 of 243 kb (Fig. 4.23). There are 6,796 contigs above 500 bp. The assembly has a GC content of 29.9% and 93.5% CEGMA completeness (Table 4.34).

R_Velvet_PE_k was then annotated using the MAKER2-Augustus pipeline. The final step predicted 21,105 protein-coding genes, with a median length of 930 bp, median exon length of 116 bp and a median of 2 exons per gene (Table 4.26).

Transcriptome assembly

A total of 23,390 contigs were generated by Trinity. Gene and isoform expression levels were calculated with RSEM and 13,225 transcripts were retained. ORFs were identified with the in-built Trinity ORF finder (Table 4.36). The post RSEM transcripts were mapped to R_Velvet_PE_k to assess its contiguity, and 99.2% of the transcripts were present in one contig with 70% transcript coverage.

Synopsis

R_Velvet_PE_k's N50 was good for annotating the genome, and the genes predicted were useful for phylogenetic analyses. Furthermore, the transcripts predicted from the RNA-Seq experiment show a high degree of CEGMA completeness (94.3%), and were useful for improving the genome annotation. MP libraries and long reads from new sequencing platforms can improve the assembly further. The genome assembly, annotations and transcriptome assembly are not publicly available yet.

Table 4.33 Read data for *Rhabditis* sp. SB347

Library	Raw		Post Trimming		Post Khmer	
	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)	Reads (M pairs)	Bases (Gb)
PE lib250	35.0	7.0	30.1	5.8	6.1	1.2
PE lib450	44.8	9.0	37.8	7.3	6.2	1.2
PE lib600	48.1	9.6	40.1	7.7	6.2	1.2
RNA-Seq	24.2	4.8	20.5	4	—	—

Table 4.34 Comparison of assemblies for *Rhabditis* sp. SB347

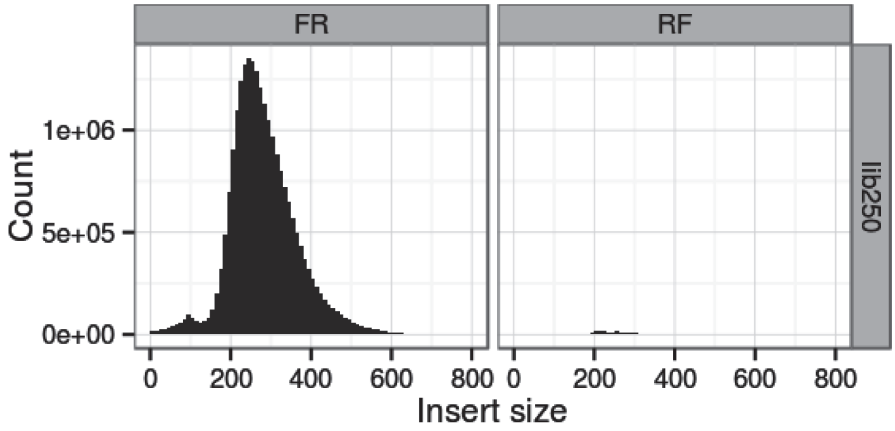
	R_CLC_SE	R_CLC_PE	R_Velvet_PE	R_Velvet_PE_k
Number of scaffolds	6,915	3,696	2,071	1,006
Longest scaffold (bp)	481,944	573,352	700,937	1,070,694
Assembly span (bp)	73,772,077	74,693,284	63,620,426	70,265,684
Number of N's (bp)	0	147,208	345,273	288,660
Mean scaffold length (bp)	10,668	20,209	30,719	69,846
Scaffold N50 (bp)	40,611	111,643	106,948	243,406
GC content (%)	38.2	38.1	34.1	36.3
CEGMA completeness (%)	99.2	99.2	98.8	99.2
Transcriptome completeness (%)	95.9	98.1	98.1	99.1

Table 4.35 Genome annotation statistics for *Rhabditis* sp. SB347

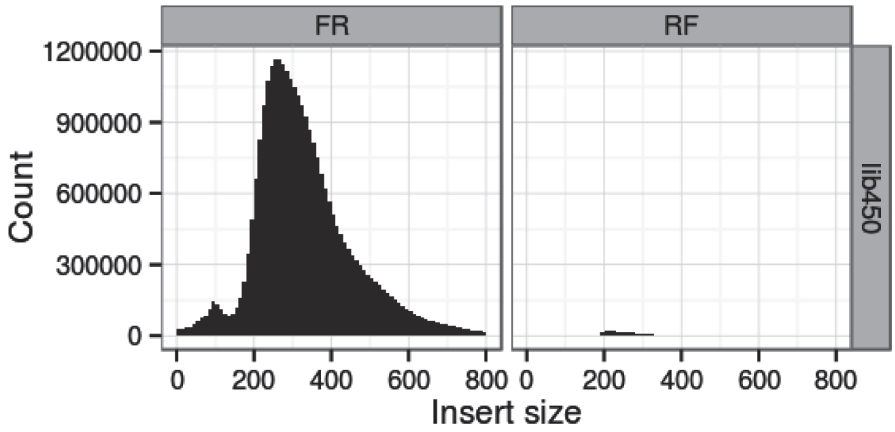
	Predicted
Number of genes	21,105
Longest transcript (bp)	20,163
Median transcript length (bp)	930
Median exon length (bp)	116
Median exons per gene	2
Median intron length (bp)	47

Table 4.36 RNA-Seq assembly statistics for *Rhabditis* sp. SB347

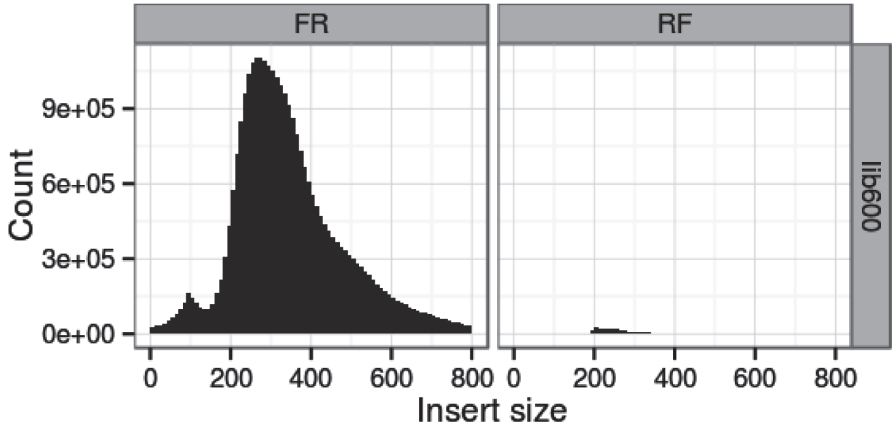
	Raw Trinity		Post RSEM	
	All	ORFs	All	ORFs
Number of transcripts	23,390	20,830	13,225	11,101
Longest transcript (bp)	17,026	15,258	17,026	15,258
Median transcript (bp)	1,113	939	988	984



(a) lib250



(b) lib450



(c) lib600

Figure 4.22 Insert size estimations for the three libraries (a) lib250, (b) lib450 and (c) lib600 of *Rhabditis* sp. SB347. The left histogram shows the distance between read pairs mapping in the FR orientation while the right histogram shows read pairs mapping in the RF orientation.

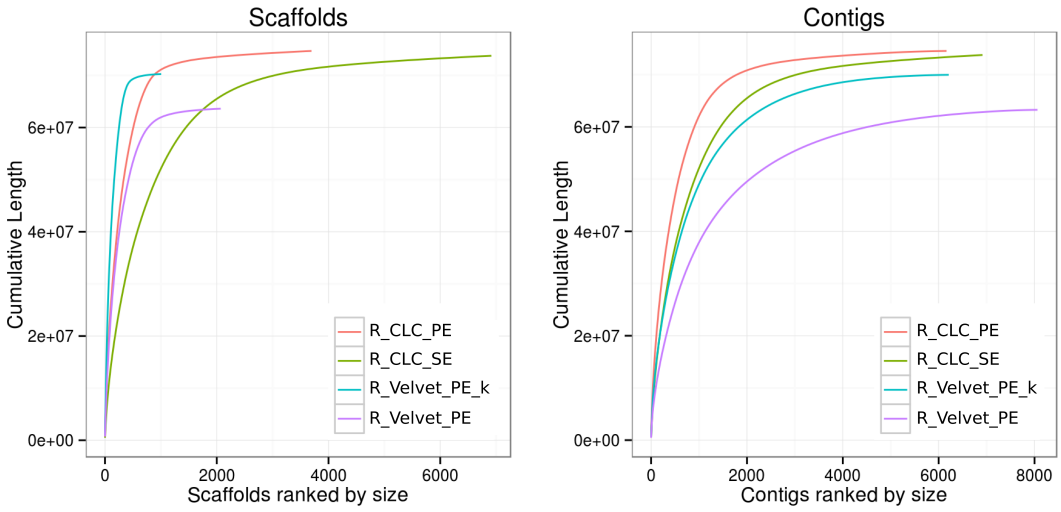


Figure 4.23 Scaffold and contig length cumulative curves for *Rhabditis* sp. SB347 assemblies. Steeper curves indicate better assembly contiguity.

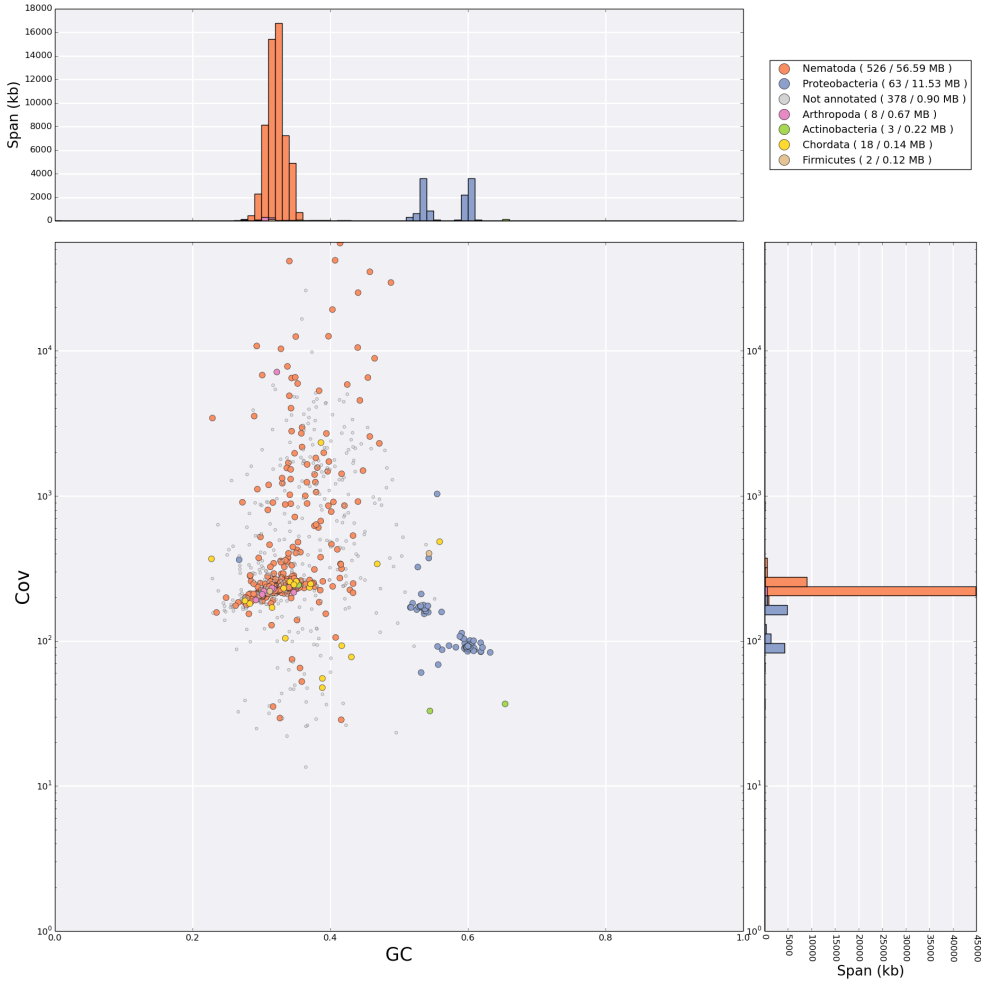
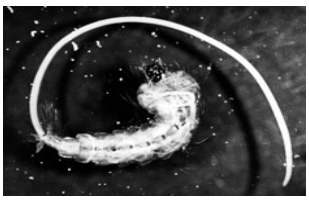


Figure 4.24 TAGC plot for R_Velvet_PE_k prior to contamination removal. Two different species of Proteobacteria can be seen at 50% (*Ochrobactrum* species) and 60% GC (*Pseudomonas fluorescens*).

4.3.10 Collaborative projects

During the course of the PhD programme, I was involved in genomic projects with different labs, in a joint effort to assemble and annotate nematode genomes. These projects acted as a springboard to discuss the bioinformatics methods and the difficulties of assembling and annotating a genome. It is a time consuming process from samples to a finished annotated genome. However, collaborations can speed up the process and provide a didactic work environment. The following five species were analysed together with other labs.

4.3.10.1 *Romanomermis culicivorax*

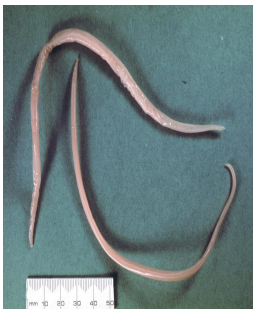


Class . . . . . Dorylaimia Inglis, 1983
Order . . . . . Mermithida Hyman, 1951
Suborder . . . . . Mermithina Andrássy, 1974
Superfamily . . . . Mermithoidea Braun, 1883
Family . . . . . Mermithidae Braun, 1883
Genus . . . . . Romanomermis Coman, 1961
Species . . . . . Romanomermis culicivorax Ross and Smith 1976

The assembly and annotation of *R. culicivorax* was generated in collaboration with Schierenberg's lab in Cologne, Germany. *R. culicivorax* belongs to order Mermithida and is the only mermithid with NGS data available. Phylogenetically, the species resides at the basal Clade I and at the time it was the second species from that Clade after *Trichinella spiralis* that was sequenced. Broader sampling from Clade I is useful for resolving the topology at the basal node. Assembly and annotation statistics are shown in Table 4.1. The genome of *R. culicivorax* was published [126] and is available at http://nematodes.org/genomes/.

4.3.10.2 *Ascaris suum*

Class . . . . . Chromadoria Pearse, 1942
Order . . . . . Rhabditida Chitwood, 1933
Suborder . . . . . Spirurina Linnaeus, 1758
Infraorder . . . . . Ascaridomorpha De Ley and Blaxter, 2002
Superfamily . . . . . Ascaridoidea Baird, 1853
Family . . . . . Ascarididae Baird, 1853
Genus . . . . . Ascaris Linnaeus 1758
Species . . . . . Ascaris suum Goeze, 1782



The annotation of *A. suum* was generated in collaboration with Davis' lab in Colorado, USA. Prior to the assembly of the *Setaria labiatopapillosa* transcriptome, *A. suum* was the closest sister species to the family Onchocercidae for which NGS data was available. Including *A. suum*, only two infraorders from the suborder Spirurina have been sampled. Assembly and annotation statistics are shown in Table 4.1. The genome of *A. suum* was published [127] and is available at http://nematodes.org/genomes/.

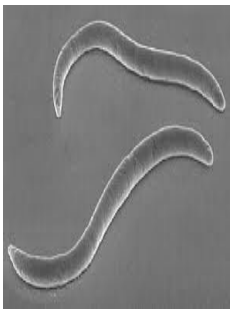
4.3.10.3 *Dirofilaria immitis*

Class . . . . . Chromadoria Pearse, 1942
Order . . . . . Rhabditida Chitwood, 1933
Suborder . . . . . Spirurina Linnaeus, 1758
Infraorder . . . . . Spiruromorpha De Ley and Blaxter, 2002
Superfamily . . . . . Filarioidea Weinland, 1858
Family . . . . . Onchocercidae Leiper, 1911
Genus . . . . . Dirofilaria Railliet and Henry, 1911
Species . . . . . Dirofilaria immitis Leidy, 1856

The assembly and annotation of *D. immitis* was generated in collaboration with Mäser's lab in Basel, Switzerland. *D. immitis* is a member of the family Onchocercidae and the relationships between the genera within the family will provide an evolutionary framework for the filarial nematodes. The assembly and annotation of *D. immitis* are discussed in detail in [16], and the results are shown in Table 4.1. The genome of *D. immitis* was published [128] and is available at http://nematodes.org/genomes/.

4.3.10.4 *Acrobeloides nanus*

Class ..... Chromadoria Pearse, 1942
Order ..... Rhabditida Chitwood, 1933
Suborder ..... Tylenchina Thorne, 1949
Infraorder ..... Cephalobomorpha De Ley and Blaxter, 2002
Superfamily ..... Cephaloboidea Filipjev, 1934
Family ..... Cephalobidae Filipjev, 1934
Genus ..... Acrobeloides Cobb, 1924
Species ..... Acrobeloides nanus De Man, 1880



The transcriptome assembly of *A. nanus* was generated in collaboration with Yanai's lab in Haifa, Israel. *A. nanus* is the only species from the infraorder Cephalobomorpha with NGS data. Therefore, it is phylogenetically important for the resolution of the topology of the infraorders from the suborder Tylenchina. Assembly statistics are shown in Table 4.2. The transcriptome assembly is not publicly available yet.

4.3.10.5 *Pseudaphelenchus vindai*

Class ..... Chromadoria Pearse, 1942
Order ..... Rhabditida Chitwood, 1933
Suborder ..... Tylenchina Thorne, 1949
Infraorder ..... Tylenchomorpha De Ley and Blaxter, 2002
Superfamily ..... Aphelenchoidea Fuchs, 1937
Family ..... Aphelenchoididae Skarbilovich, 1947
Genus ..... Pseudaphelenchus Kanzaki and Giblin-Davis, 2009
Species ..... Pseudaphelenchus vindai Kanzaki and Giblin-Davis, 2010

The annotation of *P. vindai* was generated in collaboration with Schierenberg's lab in Cologne, Germany. *P. vindai* belongs to the the family Aphelenchoididae together with *Bursaphelenchus xylophilus*. The NGS data from these two species are useful for resolving the relationships between members of the infraorder Tylenchomorpha. Assembly and annotation statistics are shown in Table 4.1. The genome assembly and annotations are not publicly available yet.

Chapter 5

Nematode phylogenomics

The term phylogenomics was originally coined by Jonathan Eisen, and it applied to annotation of gene functions using phylogenetic approaches [129]. Later, the term was also used for studies which use large portions of genomes to determine the phylogenetic relationships between species. In this chapter, the information gathered from assembling and annotating genomes and transcriptomes (chapter 4) with the addition of protein data for key additional taxa from public databases will be used in a phylogenetic pipeline to reconstruct nematode phylogeny (Tables 5.1 and 5.2).

The phylogenetic concepts are introduced, followed by the workflow and the phylogenomic results. Furthermore, the effect of intermediate taxa in a small scale phylogeny is tested. Finally, the use of different genes and data partitions is explored in a nematode family.

5.1 Introduction

The study of molecular phylogenetics is based on using molecular data (i.e DNA sequences, protein sequences or other molecular markers) to infer the evolutionary relationships between different organisms. The primary source of information to reconstruct the evolutionary history is the differential presence of random mutations in the gene sequences.

The most common mutations modify a single base in the DNA sequence and are called single nucleotide polymorphisms (SNPs). Amino acids are encoded by nucleotide triplets (codons), with most amino acids being encoded by multiple codons. As a result, single base changes can either lead to the same amino acid (synonymous mutations) or to a different amino acid (non-synonymous mutations). SNP changes in coding regions can impact the functions of the transcribed protein. These mutations can be beneficial, neutral or harmful for the organism. Beneficial mutations usually affect the structure or function of the protein and will be incorporated into the genome of future generations. Neutral mutations include all synonymous mutations, and the non-synonymous mutations that alter an amino acid that makes no significant difference to the function of the protein. Harmful mutations affect the structure or function of the protein, usually with deleterious effects in the organism.

Single nucleotide substitutions can either be transitions (A$\longleftrightarrow$G, T$\longleftrightarrow$C) or transversions ([A,G]$\longleftrightarrow$[T,C]). Usually, most mutations observed in coding regions are transitions, because transitions are more likely to be synonymous and therefore retain the structure and function of the protein [130].

5.1.1 Evolutionary history

The changes in molecular sequences can be used to infer the evolutionary history of populations and species. Usually if the organisms are closely related or the gene evolves at a slow rate, substitutions are assumed to be single events. If the taxa are evolutionarily distant or the gene evolves at a fast rate, observed substitutions can be the product of multiple events. However, the observed substitutions give no information about the number of evolutionary steps that occurred to result in the observed state [131].

In reality, the evolutionary processes underpinning change at homologous nucleotide sites are more complex. The observed states at homologous sites can be similar without having being inherited from the ancestral sequence (homoplasy).

Furthermore, multiple changes in distinct lineages can result randomly in the same base at a homologous site (convergent substitutions). Finally, after multiple substitutions a site can result in its ancestral state (back substitutions). Since phylogenetic relationships are inferred from homologous positions, the choice of molecular markers is important.

Additionally, homologous genes are assumed to have evolved independently from the last common ancestral gene before the branching of the lineages. Inferring the true orthology of gene sequences is a non trivial problem. Organisms undergoing speciation have been shown to exchange genomic material for some time after separation [132]. These processes can lead to different evolutionary histories at each locus, which will not necessarily reflect the evolutionary path of the species. Furthermore, alleles that show different survival rates can affect the historical trajectory of the species (lineage sorting). Horizontal gene transfer can further impede phylogenetic reconstruction, since these genes have been incorporated in the genome after transfer from a very different organism. Finally, paralogous genes that arise from gene duplications have different evolutionary pressures and subsequent loss of one copy may lead to an incongruence between the gene and species tree.

As a result, gene phylogenies may or may not reflect the actual species phylogenetic tree. Figure 5.1 shows two different evolutionary processes that can result in incongruence between the two trees. The topology of the gene tree (a) will result in a cluster between B and C because the gene diverged prior to the divergence of species (A,B) and C. Similarly, the topology of the gene tree (b) will result in the same incorrect cluster B and C. However, in this case it is a result of gene duplication at the base of the gene tree. The presence of a single copy at the present day requires at least three independent losses. The definition and choice of orthologous genes is thus crucial in the inference of phylogenetic relationships.

Depending on the evolutionary question, some genes can be more informative than others. Slowly-evolving genes are more helpful for resolving deep phylogenies with distantly related species, because positions can be more easily traced to the

ancestral sequence. On the other hand, rapidly-evolving genes are more helpful for closely related taxa since they are more likely to contain phylogenetic informative substitutions. Similarly, amino acid sequences are more informative for distantly related taxa while nucleotide sequences are more informative for closely related taxa. In this study, sequences were sampled across the phylum Nematoda and thus amino acid sequences were selected.

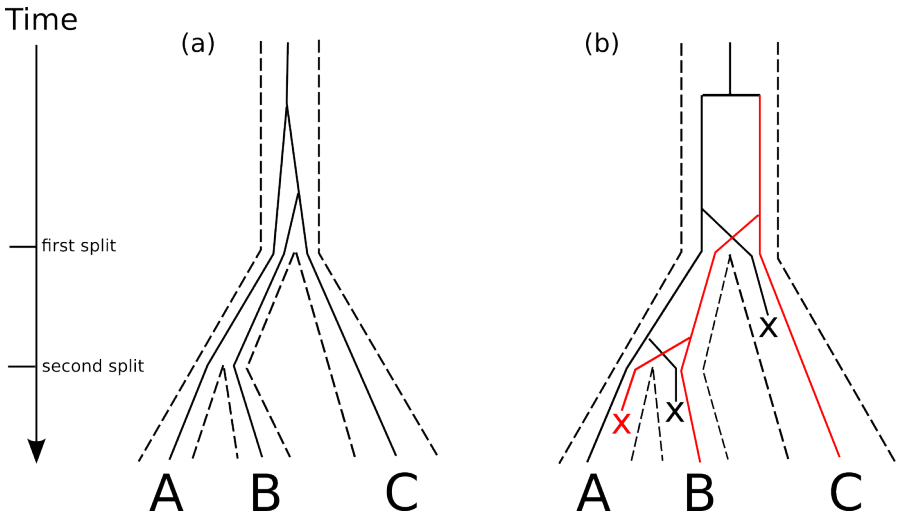


Figure 5.1 Two possible relationships between a species trees (dotted lines) and a gene tree (solid lines) that show incongruence. In (a), the divergence of the gene pre-dates the divergence of the species. In (b), a gene duplication at the base of the gene tree resulted in gene A being paralogous to genes B and C, after three gene losses. Modified from [131] and [133].

5.1.2 Orthology Assessment

In phylogenetic reconstruction the assumption is made that the sequences analysed are orthologous; they are directly descendant from a last common ancestral gene. If they are indirectly descendant because of gene duplication, then they are called paralogous. The term homologous encompasses both categories, generally meaning that these sequences are derived from a single gene in some ancestor. These terms are used in a boolean sense, for example two sequences can either be homologous or not.

In a phylogenetic framework, two sequences are homologous if they have *excess similarity*. Excess similarity is defined by the following statements:

- Unrelated sequences have similarity equal to random sequences.
- If similarity is not random then sequences must be not unrelated.
- Hence, not random similarity must reflect related sequences (i.e. homologues).

Paralogues that have arisen from ancient duplication are likely to have diverged functions, while orthologues are more likely to retain the same function. Thus, orthologous clustering is needed to infer accurate phylogenies. Furthermore, it can provide important information about the presence or absence of gene families and biological processes, and thus the likely biological differences between species.

Automatic clustering is the first step to identify potential orthologous clusters. Refinement of the clusters is needed to avoid the erroneous inclusions of paralogous sets in the phylogenetic analyses (see subsection 5.2.2).

5.1.3 Sequence Alignment

Sequence alignment is a comparison of the residues in two or more sequences, describing one-to-one correspondences between residues in different sequences in the order the residues occur in the sequences. Gap characters indicate no corresponding

residue. Homologous sequences can then be aligned to identify homologous residues within the sequences. The accuracy of the alignment has a great effect on phylogenetic inference.

Protein sequences align more correctly than nucleotide sequences, especially when the sequences are distantly related [134]. There are more characters (20 compared to 4) and the algorithm is less influenced by gap penalties. Furthermore, the use of protein sequences eliminates potential erroneous alignments within codons. Protein sequence alignment uses matrices derived from empirical datasets that define the cost for every possible amino acid replacement (e.g. BLOSUM 62 matrix [135]).

In most cases, more than two sequences need to be aligned. Multiple sequence alignments (MSAs) are gradually built by aligning the closest sequences first and successively adding more sequences. Once all sequences are aligned, most alignment algorithms will refine the alignment iteratively (e.g. MUSCLE [136]). Currently, most alignment algorithms use the phylogeny as a guide to order the alignment of pairs (e.g. MAFFT [137], Clustal Omega [138]).

The accuracy of the phylogenetic tree is heavily connected to the accuracy of the alignment. Although the accuracy of the alignment drops as the amino acid identity of the alignment decreases, >20% identity can result in an average of ~80% successfully aligned positions [139]. Orthologous sequences will usually have more than 20% identity and thus align accurately.

5.1.4 Alignment Trimming

Large scale phylogenomic datasets will likely contain a mix of slow and fast evolving genes (e.g [140]). Furthermore, regions of sequence with doubtful homology can be aligned, creating erroneous aligned sites. Trimming algorithms (e.g. Gblocks [141], trimAl [142]) are used to remove these positions from the alignment, retaining only conserved blocks which will more probably reflect the correct evolutionary history. In addition, misidentified homologous sequences can be removed by these algorithms

on the basis of low alignment quality. A more robust approach to remove erroneous sequences is to visually inspect the tree topologies of these alignments for long branches, as it likely that sequences with long branch (i.e. very highly diverged) have been misidentified as homologous. However, visual inspection and correction of every alignment in a large dataset is a time consuming process, and the use of these trimming algorithms is recommended instead.

5.1.5 Phylogenetic inference

The information of aligned sequences is then used to reconstruct the evolutionary history of the taxa present. The goal of molecular phylogenetics is to reconstruct the evolutionary tree given a set of sequences, and a phylogenetic method that relies on a defined substitution model for the aligned sites.

Phylogenetic relationships are visualised using trees as networks comprised of nodes and branches. Internal nodes (nodes) represent hypothetical ancestral taxa while terminal nodes (tips) are associated with an extant sequence or Operational Taxonomic Unit (OTU). If a root is specified (representing the most recent common ancestor of all the tips), then time flows from the base of the tree to the tips and branches now represent successive generations of taxa. Branch lengths are usually adjusted to the amount of change per aligned site, reflecting the evolutionary distance of the taxa. Internal nodes with two daughter branches are bifurcations, while a node is multifurcating (polytomy) when more than two daughter lineages have diverged simultaneously (or in small amount of time that is not resolvable).

In rooted trees, a clade is defined as a set of OTUs that includes all the descendants of a given internal node. In unrooted trees, clans are defined as groups of OTUs that form clades in at least one rooted phylogeny. A monophyletic character appears in all the members of a clade, while a paraphyletic or polyphyletic character appears in taxa that do not form a clade. As discussed in section 1.4, many of the earliest classifications of nematodes have been shown to be polyphyletic in comparison to

molecular phylogenetic trees of the phylum.

5.1.5.1 Phylogenetic methods

The most commonly used tree building methods are Neighbour Joining (NJ), Maximum Parsimony (MxP), Maximum Likelihood (ML) and Bayesian Inference (BI). Based on the objections to NJ and MxP methods (reviewed in [131, 143]), only ML and BI were considered in this study.

Maximum likelihood

In maximum likelihood we estimate the tree and the model parameters by maximizing the probability of observed data.

$$Pr(X|\theta)$$

- X, is the data (MSA).
- $\theta = (\tau, m, v)$, are the parameters; topology, substitution model and branch lengths.

The likelihood function was developed by R. A. Fisher in 1920 to estimate parameters in a model to maximise the likelihood of the selected model with the observed data. In 1981, Felsenstein applied the first maximum likelihood algorithm to DNA sequence data [144]. The model assumptions of ML are explicit and thus can be improved and evaluated [143]. Models can be estimated from the dataset, and thus be more effective than predefined models. However, the tree searches in ML are computationally expensive. Therefore, optimisation algorithms (e.g. hill climbing [145], genetic algorithms [146]) were introduced to lower the computational costs.

Non-parametric bootstrap methods are widely used in ML analyses to give an estimate of the support of each node [147]. The aligned sites are re-sampled with replacement usually 100 times (bootstrap replicates). Then, a tree is calculated for

every replicate, and a value is given to each node based on the number of times it was recovered in the bootstrap replicates. Generally, bootstrap values above 75% indicate well supported nodes [148].

Bayesian Inference

Before the analysis, prior distributions are assigned to the parameters and combined with the likelihood function in order to generate the posterior distribution.

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{\int f(X|\theta)f(\theta)d\theta} \propto f(X|\theta)f(\theta)$$

- X, is the data (MSA).
- $f(\theta|X)$, is the posterior probability.
- $f(X|\theta)$, is the likelihood.
- $f(\theta)$, is the prior distribution.
- $\theta = (\tau, m, v)$, are the parameters; topology, model and branch lengths.

The difficulty in Bayesian statistics is the estimation of the normalising constant which can not be calculated. Instead, a sample is generated from the posterior distribution using the Markov chain Monte Carlo (MCMC) algorithms [149] explained below.

To circumvent sampling on only local maxima, additional chains (called heated) are started with different acceptance ratios (i.e. accepting changes more often). The heated chains sample trees more efficiently, but only the trees sampled by the original chain (called cold) are used to calculate the posterior distribution. At regular intervals, two random chains are selected and an exchange in states is proposed. If the change is accepted, the two chains change states and continue sampling from the new positions.

Simplified MCMC algorithm

```
Set generation N = 1
Start at a random point
Set maximum number of generations MAX
for N ≤ MAX do
    Attempt a random move sampling from the parameter distribution
    Calculate the acceptance ratio (α = f′(X|θ)/f(X|θ),
    f′(X|θ) is the likelihood at the proposed position,
    f(X|θ) is the likelihood at the current position)
    if α ≥ 1 then
        Accept always
    else
        Accept with probability α
        If the move is not accepted, stay at the same position
    end if
    Increment N by 1
end for
```

Usually it is hard to estimate if the parameter space was sampled sufficiently to calculate the correct posterior probabilities. Instead, if the algorithm starts sampling only from a set of trees, then the algorithm can be stopped, assumed to have reached convergence. A more robust estimate of convergence can be achieved by starting multiple independent runs.

Bayesian statistics inherently calculate the support for each node. The posterior probability of a node being correct is the number of times the node is present in the posterior distribution of the trees sampled. However, studies have shown that posterior probabilities can be too high in real datasets [150, 151]. Generally, posterior probabilities above 0.95 indicate well supported nodes [152].

5.1.5.2 Evolutionary models

One of the first amino acid substitution matrix was generated by Dayhoff and Schwartz (Dayhoff matrix) [153]. It was calculated based on the alignments of closely related proteins by counting the changes in amino acids, normalising for divergence between sequence pairs. In 1992, Jones, Taylor, and Thornton applied

the same approach on a larger dataset (JTT matrix) [154]. Nine years later, Whelan and Goldman used a dataset consisting of 3,905 proteins split into 182 alignments, inferred phylogenies with NJ, optimised branch lengths by ML with JTT, and estimated the optimal substitution matrix (WAG matrix) [155]. Using an improved ML method and alignments from Pfam database [156], Le and Gascuel generated the LG replacement matrix [157]. Finally, similar to the general-time reversible model (GTR) for DNA sequences [158], recent phylogenetic programs can calculate the model for amino acids.

In addition to substitution matrices, the CAT model [159] was proposed to account for site-specific features based on the biochemical constrains in a protein. The accurate distribution of profiles requires large alignments, thus the model is only useful for large phylogenomic studies. Furthermore, the mixture of profiles is based on a Dirichlet process, which practically can only be implemented in BI algorithms [160].

The simplest model to account for site variation is the invariable-site model ($+I$), which splits the sites into two groups, one group has zero rate while the other group evolves at the same rate [161]. Yang [162] proposed a discrete gamma distribution ($+\Gamma$) with four categories to model rate heterogeneity, since a continuous gamma distribution is only computationally practicable in small datasets. To speed up calculations, Stamatakis [163] proposed an approximation of variable sites where each site is placed in rate categories based on the observed relative rate at each site ($+\mathbf{CAT}$, denoted with bold letters to distinguish from the CAT profile mixture model).

5.2 Workflow

I set out to produce a genome-scale phylogeny of the Nematoda. An analysis of this scale requires a robust workflow to maximise the correct aligned positions in the supermatrix. Issues can arise at all stages of the workflow. The genome and

transcriptome assemblies are not perfect, as they can contain chimeric contigs or very fragmented regions. The automated annotation pipeline will then collate chimeric protein predictions and fragmented genes, and even gene models that do not actually exist. Clustering the protein data from these sources can result in erroneous groups of genes which will be assumed to be homologous. The alignment of these clusters will have low alignment accuracy which will further impede correct phylogenetic reconstruction.

Graphical representation of the workflow developed is shown in figure 5.2. Protein datasets were collected from annotated genomes, transcriptome assemblies and ESTs (Tables 5.1 and 5.2). To remove recent paralogues, each protein file was clustered with cd-hit at 95% similarity threshold. After this redundancy removal there were 927,436 proteins from 58 species in the final file. OrthoMCL [164] was used to calculate the orthologous clusters.

An alternative workflow using CEGMA genes was also tested. Genes were extracted from genomic sequences using CEGMA, while genes from transcriptomes and ESTs were identified by BLAST+. The dataset consisted of 412 CEGMA gene sets, 72 of which were also present in the final dataset from the OrthoMCL workflow which contained 959 gene clusters (Table 5.4).

Table 5.1 Comparison of genome assemblies and annotations for the species used in the phylogenomic analyses

Species	Group	Genome size (Mb)	N50 (b)	Number of scaffolds	Number of genes predicted	Median transcript length (bp)	Source & publication (if available)	
Drosophila melanogaster	Arthropoda	159.44	24,543,557	8	26,950	1,404	FlyBase[a]	[165]
Bombyx mori	Arthropoda	479.42	3,998,728	40,280	14,623	867	SilkDB[b]	[166]
Tetranychus urticae	Arthropoda	90.82	2,993,488	640	18,393	1,151	Ensembl[c]	[167]
Hypsibius dujardini	Tardigrada	139.92	48,302	30,122	23,021	870	http://badger.bio.ed.ac.uk/H_dujardini	
Romanomermis culicivorax	Clade I	322.77	17,632	62,537	10,206	387	http://nematod.es	[126]
Trichinella spiralis	Clade I	63.51	6,373,445	6,819	16,380	577	WormBase[d]	[168]
Trichuris suis	Clade I	73.50	503,430	1,470	14,436	612	WormBase[d]	[169]
Enoplus brevis †	Clade II	126.60	753	165,525	–	–	This work[e]	
Plectus murrayi†	Group C	185.02	10,183	33,505	–	–	This work[e]	
Plectus sambesii†	Group C	239.12	9,794	73,716	–	–	This work[e]	
Ascaris suum	Clade III	334.00	290,558	260	15,446	864	http://nematod.es	[127]
Brugia malayi	Clade III	94.14	191,089	9,827	17,846	726	WormBase[d]	[170]
Dirofilaria immitis	Clade III	88.30	22,560	71,281	16,021	771	http://nematod.es	[128]
Loa loa	Clade III	91.37	174,388	5,770	15,444	642	WormBase[d]	[43]
Onchocerca volvulus	Clade III	96.43	25,485,961	708	12,995	825	WormBase[d]	
Onchocerca gutturosa	Clade III	109.56	10,567	25,605	19,916	456	This work[e]	
Onchocerca ochengi	Clade III	95.51	12,317	24,057	13,990	735	http://nematod.es	[171]
Litomosoides sigmodontis	Clade III	64.81	45,863	3,165	10,246	1,002	http://nematod.es	[172]
Acanthocheilonema viteae	Clade III	77.35	25,808	6,796	10,397	975	http://nematod.es	
Wuchereria bancrofti	Clade III	81.51	5,161	25,884	18,327	474	Broad Institute[f]	[43]

Comparison of genome assemblies and annotations for the species used in the phylogenomic analyses (continued)

Species	Group	Genome size (Mb)	N50 (b)	Number of scaffolds	Number of genes predicted	Median transcript length (bp)	Source & publication (if available)	
Bursaphelenchus xylophilus	Clade IV	74.56	949,830	5,526	18,074	789	WormBase[d]	[173]
Globodera pallida	Clade IV	124.67	121,687	6,872	16,417	753	Sanger Institute[g]	[174]
Globodera rostochiensis	Clade IV	93.09	86,303	4,398	13,650	966	BlaxterLab[h]	
Meloidogyne floridensis	Clade IV	96.67	3,698	58,696	11,975	342	http://nematod.es	[175]
Meloidogyne hapla	Clade IV	53.02	37,608	3,452	14,420	753	WormBase[d]	[176]
Meloidogyne incognita	Clade IV	82.09	12,786	9,533	19,212	774	WormBase[d]	[177]
Panagrellus redivivus	Clade IV	65.06	257,941	867	24,249	852	WormBase[d]	[178]
Pseudaphelenchus vindai	Clade IV	53.83	5,784	28,977	6,073	858	This work[e]	
Strongyloides ratti	Clade IV	52.64	359,029	2,184	8,188	1,089	WormBase[d]	
Caenorhabditis elegans	Clade V	100.29	17,493,829	7	20,520	1,020	WormBase[d]	[12]
Caenorhabditis angaria	Clade V	99.01	87,708	11,453	27,967	603	WormBase[d]	[15]
Caenorhabditis briggsae	Clade V	108.42	17,485,439	12	21,850	933	WormBase[d]	[13]
Caenorhabditis sp. 5	Clade V	131.80	25,228	15,261	46,280	930	http://nematod.es	[16]
Haemonchus contortus	Clade V	368.83	83,501	19,726	24,775	903	WormBase[d]	[179]
Heterorhabditis bacteriophora	Clade V	77.01	312,328	1,259	20,964	426	WormBase[d]	[180]
Necator americanus	Clade V	244.05	211,861	11,767	19,153	576	WormBase[d]	[181]
Pristionchus pacificus	Clade V	170.36	1,290,309	10,227	24,217	723	WormBase[d]	[182]
Pristionchus exspectatus	Clade V	177.64	142,227	4,412	24,642	789	WormBase[d]	[183]
Dictyocaulus viviparus	Clade V	169.39	22,560	17,715	14,306	834	http://nematod.es	[122]
Ancylostoma ceylanicum	Clade V	313.10	668,412	1,736	36,587	600	WormBase[d]	
Rhabditis SB347	Clade V	70.27	243,406	1,006	21,105	930	This work[e]	

† The genome assembly is preliminary

[a] http://flybase.org [184]

[b] http://silkworm.genomics.org.cn [185]

[c] http://www.ensembl.org [186]

[d] http://www.wormbase.org [80]

[e] Not publicly available

[f] http://www.broadinstitute.org/annotation/genome/filarial_worms

[g] http://www.sanger.ac.uk/resources/downloads/helminths/globodera-pallida.html

[h] Not publicly available

Table 5.2 Comparison of transcriptome assemblies and EST datasets for the species used in the phylogenomic analyses

Species	Group	Method	Number of transcripts	Median transcript length (bp)	Source & publication (if available)
Gordius sp.	Nematomorpha	Illumina	8,888	1,260	This work[a]
Prionchulus punctatus	Clade I	Illumina	21,290	662	This work[a]
Trichuris muris	Clade I	ESTs	3,067	531	NEMBASE4[b]
Xiphinema index	Clade I	ESTs	4,626	631	NEMBASE4[b]
Enoplus brevis	Clade II	Illumina	20,196	887	This work[a]
Plectus murrayi	Group C	Illumina	24,055	821	This work[a]
Plectus sambesii	Group C	Illumina	34,648	960	Philipp Schiffer[c]
Setaria labiatopapillosa	Clade III	Illumina	14,283	1,049	This work[a]
Anguillicola crassus	Clade III	454	6,990	471	AfterParty[d] [187]
Panagrolaimus sp. ES5	Clade IV	Illumina	30,865	822	Philipp Schiffer[c]
Panagrolaimus sp. PS1159	Clade IV	Illumina	39,721	837	Philipp Schiffer[c]
Propanagrolaimus sp. JUL765	Clade IV	Illumina	32,914	882	Philipp Schiffer[c]
Acrobeloides nanus	Clade IV	Illumina	9,549	495	This work[a]
Strongyloides stercocalis	Clade IV	ESTs	3,708	481	NEMBASE4[b]
Aphelenchus avenae[†]	Clade IV	ESTs	1,829	498	NCBI[e] [188]
Panagrolaimus superbus	Clade IV	ESTs	4,042	441	NEMBASE4[b]
Heterodera glycines	Clade IV	ESTs	9,270	584	NEMBASE4[b]
Rhabditis SB347	Clade V	Illumina	13,225	988	This work[a]
Cylicostephanus goldi	Clade V	454	10,193	438	AfterParty[d] [189]
Ancylostoma caninum	Clade V	ESTs	26,356	576	NEMBASE4[b]
Nippostrongylus brasiliensis	Clade V	ESTs	3,691	521	NEMBASE4[b] [190]

† I assembled the EST dataset using CLOBB2 [191] and CAP3 [77]

[a] Not publicly available

[b] http://www.nematodes.org/nembase4 [192]

[c] Not publicly available

[d] http://afterparty.bio.ed.ac.uk [193]

[e] http://www.ncbi.nlm.nih.gov/nucest/?term=txid70226

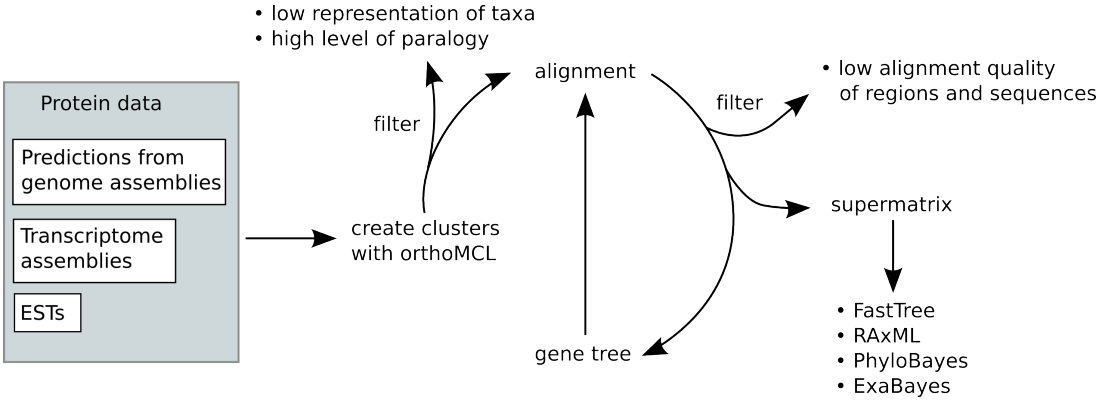


Figure 5.2 Phylogenomic workflow. Protein datasets are clustered with OrthoMCL and clusters with low number of taxa or high levels of paralogy are removed. Next, proteins in each cluster are aligned and the alignments are filtered to remove low quality aligned regions and low quality aligned sequences. Alignments with one sequence per species are added to the supermatrix. For the rest a gene tree is created, and clans with at least 30 species with one sequence per species were retained, the reduced sequence cluster is aligned and added to the supermatrix.

5.2.1 OrthoMCL

Of the initial 927,436 proteins, 4,051 poor-quality proteins were removed from the dataset using the default options of orthomclFilterFasta algorithm (proteins below 50 aa length or with 20% stop codons). BLAST+ was used within the remaining 923,385 proteins to identify putative orthologous relationships by reciprocal best similarity pairs. The BLAST+ results were loaded in a relational database, and the OrthoMCL algorithm converted the relationships into a graph. Each node is a protein sequence which connects to other nodes by a weighted edge representing their relationship. Edge weights are normalised BLAST scores for the connected nodes, initially calculated as $-\log_{10}$(e-value). Next, the graph is parsed by the MCL algorithm [194] separating diverged paralogs and distant orthologs. The inflation parameter of the MCL algorithm determines the granularity of the clusters. A high inflation value will increase the cluster tightness while a low value will result in more nodes being pulled together. A value of 3 was chosen to avoid chimeric clusters but also to allow for relatively distant sequences to cluster together. Except for the number of clusters, different inflation values do not show significant change in within cluster variation (Table 5.3). Using an inflation value of 3 resulted in 91,549 clusters containing at least 2 proteins.

Table 5.3 Effect of different inflation values on OrthoMCL clustering

Inflation value	No. of clusters	Mean proteins per species per cluster		Median proteins per species per cluster	
		Mean	Median	Mean	Median
1.5	73,706	8.3	3	1.08	1
2	82,556	7.4	3	1.06	1
3	91,549	6.7	3	1.04	1
4	97,705	6.2	3	1.03	1
5	102,179	5.9	3	1.02	1

5.2.2 Cluster filtering

Initially, clusters with fewer than 30 species were removed. Next, the number of proteins for each species within the cluster was calculated, and clusters with a mean greater than 2.5 and a median greater than 1 were removed to eliminate lineage-specific duplication. This constrain eliminates clusters that contain taxa with lineage specific duplications which will impede the identification of orthologous proteins. The most conservative filtering parameter was the minimum number of taxa required, which resulted in the exclusion of 86,681 clusters. After the filtering 4,737 clusters were retained. Then, the clusters were split into two groups. Clusters were added to the first group if each species within the cluster was represented by 1 protein sequence (922 clusters). The remaining 3,815 clusters comprised the second group.

The sequences for each cluster in the first group were aligned with Clustal Omega and trimmed with trimAl. The trimming algorithm removed columns that contained 20% gap positions or had a similarity score lower than 0.001, and sequences that had less than 20% aligned to 75% of the combined alignment of the other sequences. All the trimmed clusters were added in the final alignment.

Similarly, the sequences for each cluster of the second group were aligned with Clustal Omega and trimmed with trimAl (same parameters as above). Then, a phylogenetic gene tree was produced for each of the 2,075 clusters which had at most 2 sequences per species within them. The best model (from JTT, WAG, LG, and Dayhoff) was selected with prottest3 [195] and a phylogenetic tree was constructed for each cluster with RAxML [145]. RAxML calculated the best scoring ML tree once, and 100 bootstrap replicates were generated. Each tree was visually inspected with FigTree [196]. All nodes that had bootstrap support lower than 50 were collapsed to polytomies. Next, if sequences from the same species were monophyletic, the longer sequence was retained. If a clan with at least 30 species, all having at most one sequence, was identified, then the other clan was discarded. Only 37 clusters were retained with reduced sequence information and were added in the final alignment.

Table 5.4 shows changes in cluster sequence information after trimming. Although the parameters in trimAl may have been overly-conservative, a lot of likely uninformative sites and misclustered sequences were removed. Huge datasets can be filtered more conservatively, since a lot of information is still retained, and erroneous alignment sites can affect greatly the phylogenetic inference.

Changes in protein composition for each clustering for each species are listed in Appendix A.

Table 5.4 Number of proteins and aligned amino acid sites before and after filtering.

Groups	Clusters	Before filtering		After filtering	
		Proteins	Aligned sites	Proteins	Aligned sites
Group 1	922	39,645	960,126	29,996	160,619
Group 2	37	2,269	55,985	1,533	9,673
Total	959	41,914	1,016,111	31,529	170,292
CEGMA genes	412	20,143	633,895	17,762	114,805

5.3 Phylogenetic analyses

One ML program RAxML [145] and two BI programs (ExaBayes [197] and PhyloBayes [198]) were used to infer phylogenies. In addition, FastTree [199] with JTT+**CAT** model was used as an initial approximation of the phylogeny and the results are shown in Appendix C.

All the alignments from the 959 trimmed clusters were concatenated together in a supermatrix (SM1). The unpartitioned alignment was analysed using RAxML with GTR+Γ model and 100 bootstrap replicates (Fig. 5.3), and ExaBayes with LG+Γ for 112,300 generations (Fig. 5.4).

Species from EST datasets were removed from SM1 to form SM2, and the phylogenetic inference was performed with RAxML with GTR+Γ model and 100 bootstrap replicates, and PhyloBayes with GTR+CAT+Γ for 1,073 cycles (Fig. 5.5).

CEGMA clusters were concatenated together in a supermatrix (SM3). The unpartitioned alignment was analysed using RAxML with GTR+Γ model and 100 bootstrap replicates, ExaBayes with LG+Γ for 53,300 generations, and PhyloBayes with GTR+CAT+Γ for 263 cycles (Fig. 5.6).

In each case, the Bayesian inference algorithms (ExaBayes and PhyloBayes) were initiated with 2 independent runs with 4 MCMC chains each. The programs were stopped when the trace files (visualised with Tracer [200]) showed convergence for the 2 independent runs.

The likelihood score for each tree produced was calculated with GTR+Γ model (Table 5.5).

Table 5.5 Likelihood scores for the phylogenetic trees.

Dataset	Program	Tree	Likelihood score
SM1	RAxML	Fig. 5.3	**-3551917.339961**
	ExaBayes	Fig. 5.4	-3551940.109491
SM2	RAxML	Fig. 5.5	**-3495730.574051**
	PhyloBayes	Fig. 5.5	-3495964.919141
SM3	RAxML	Fig. 5.6	-2762365.406547
	ExaBayes	Fig. 5.6	**-2762365.406386**
	PhyloBayes	Fig. 5.6	-2762527.646351

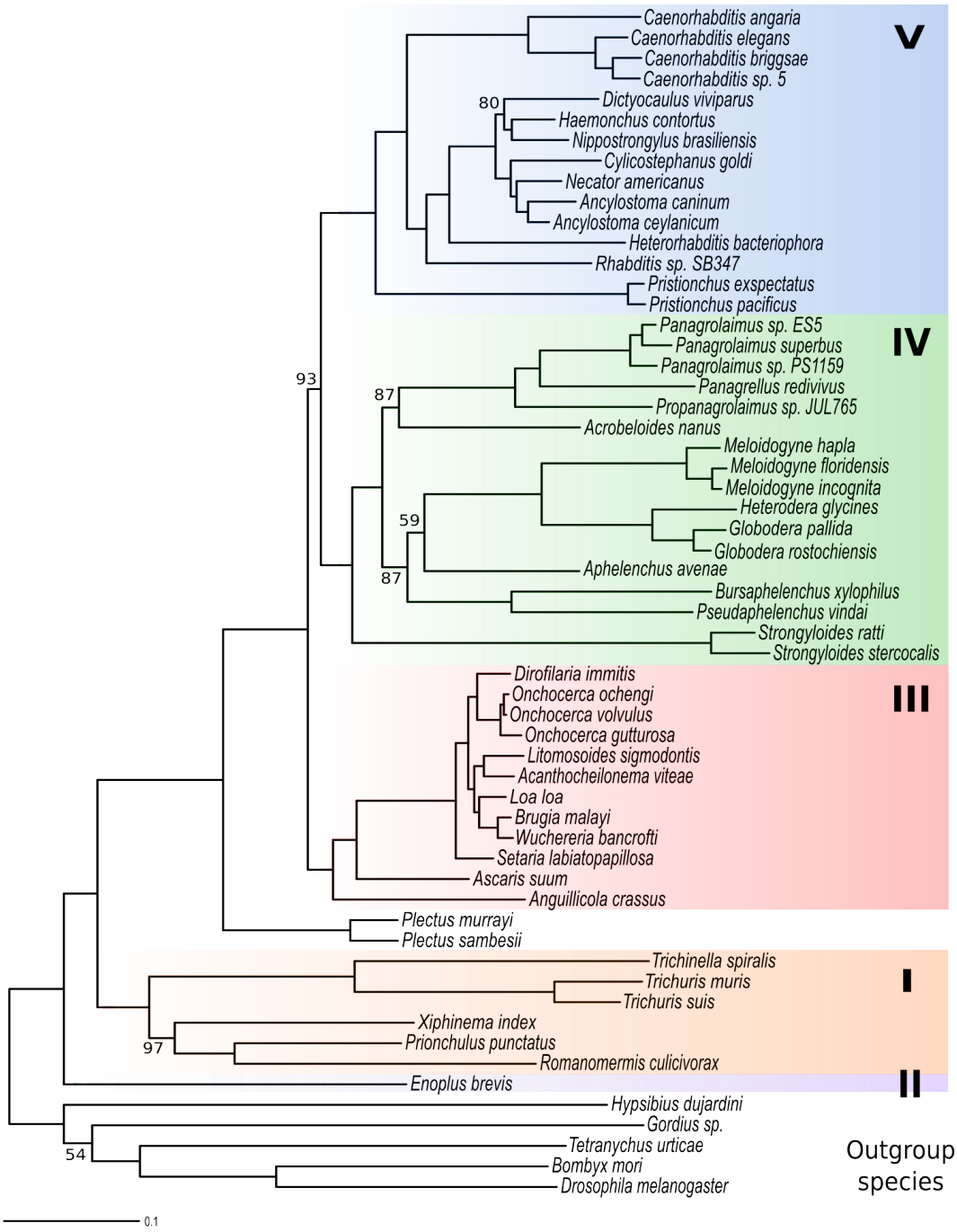


Figure 5.3 Phylogenetic tree obtained from RAxML using SM1. Bootstrap support values below 100 are shown. Coloured boxes represent the five clades identified by Blaxter et al. [26]

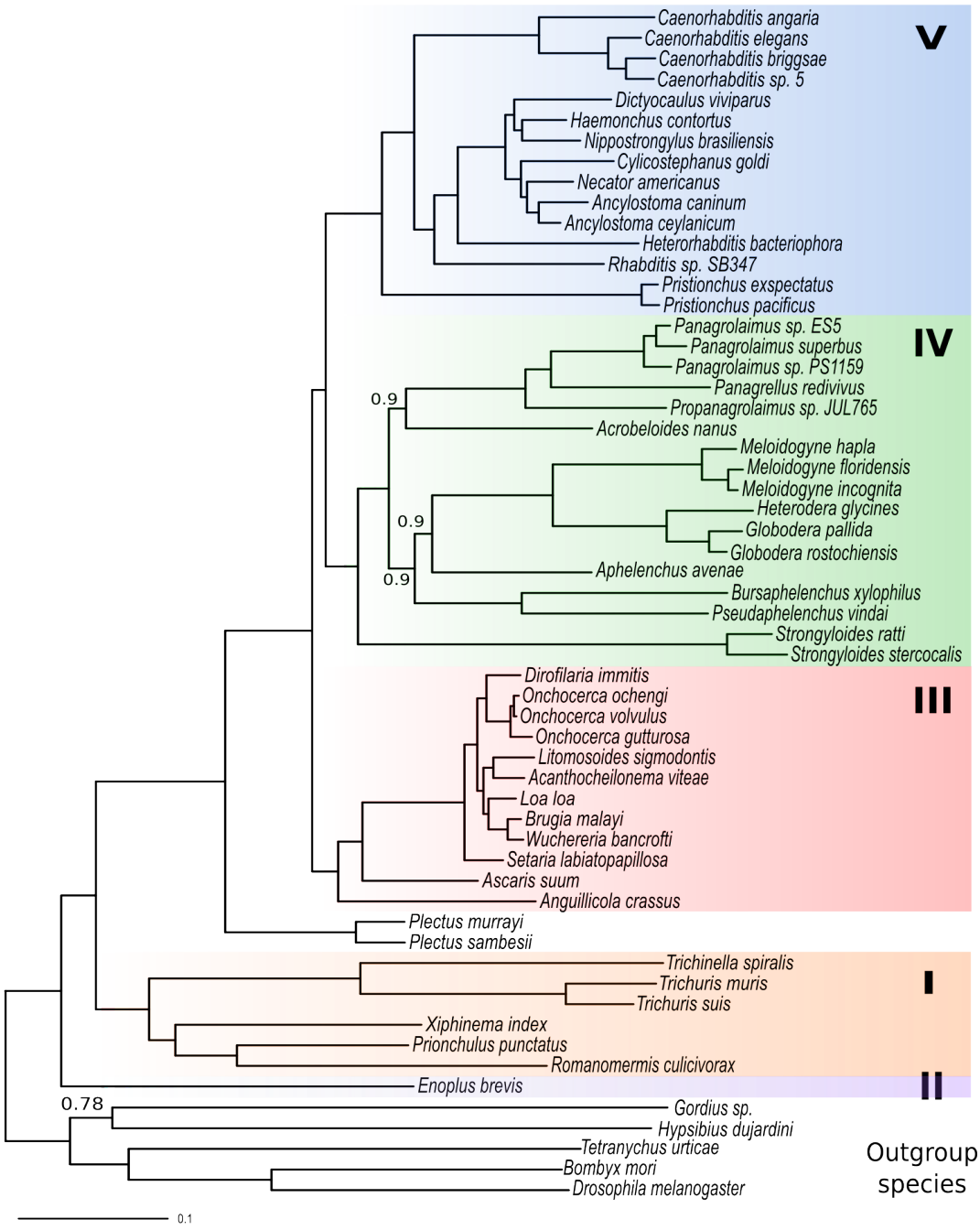


Figure 5.4 Phylogenetic tree obtained from ExaBayes using SM1. Posterior probabilities below 1 are shown. Coloured boxes represent the five clades identified by Blaxter et al. [26]

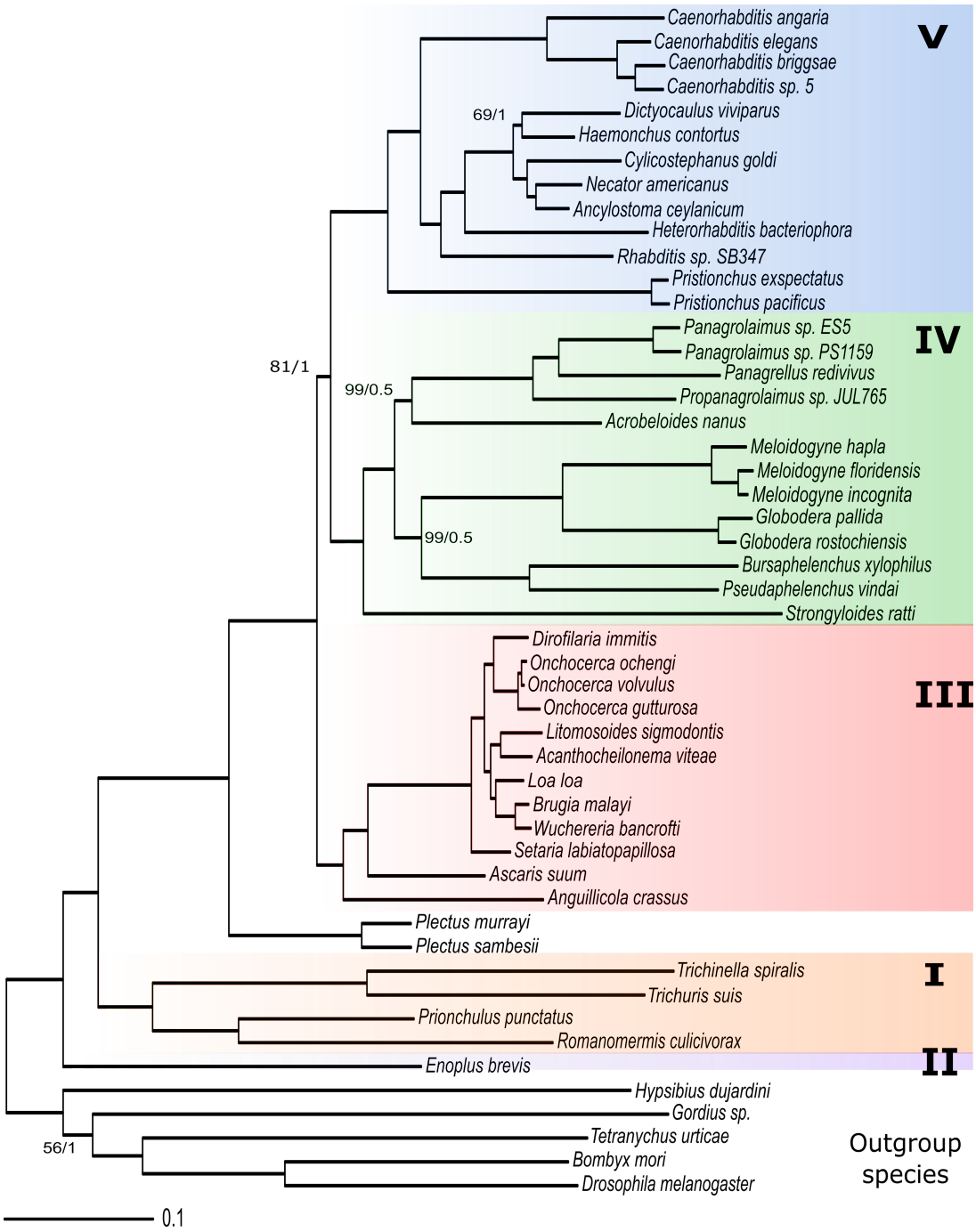


Figure 5.5 Phylogenetic tree using SM2. The tree shown was obtained with RAxML. Node support values below 100 (RAxML bootstraps), or below 1 (PhyloBayes posterior probabilities) are shown in the order RAxML/PhyloBayes. Coloured boxes represent the five clades identified by Blaxter et al. [26]

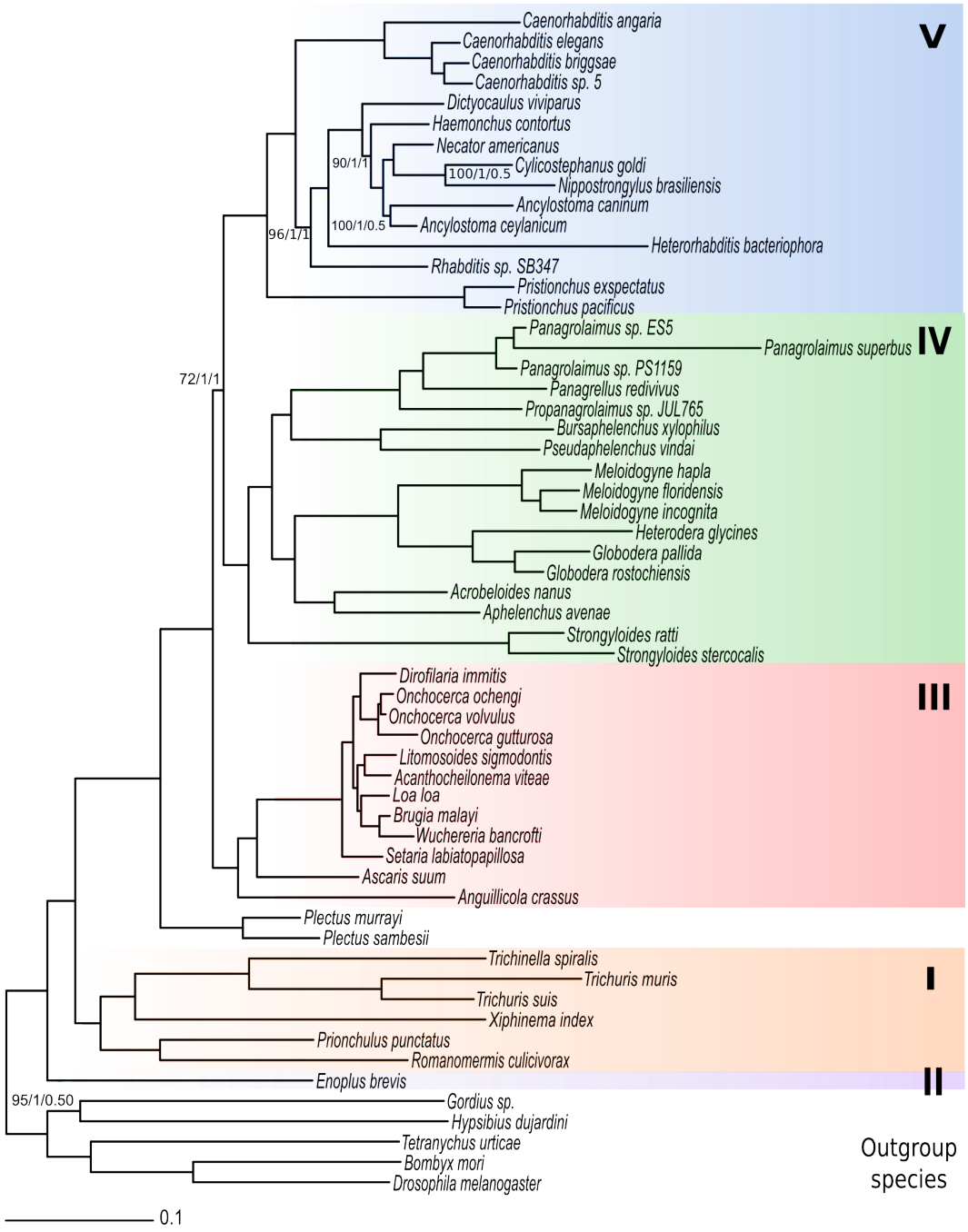


Figure 5.6 Phylogenetic tree using SM3. The tree shown was obtained with RAxML. Node support values below 100 (RAxML bootstraps), or below 1 (ExaBayes and PhyloBayes posterior probabilities) are shown in the order RAxML/ExaBayes/PhyloBayes. Coloured boxes represent the five clades identified by Blaxter et al. [26]

5.3.1 Phylogenetic Results

Nematode systematics (*sensu* De Ley and Blaxter [27, 28]) for the species used in the phylogenetic analyses are shown in Appendix B.

The relationship between the three classes of the phylum (Dorylaimia, Enoplia, and Chromadoria) was resolved in all analyses. The trees showed the Enoplia as the most basal clade, and the Dorylaimia and Chromadoria as sister clades. Enoplea (Dorylaimia and Enoplia) was always recovered as paraphyletic.

5.3.1.1 Outgroup species

Hypsibius dujardini was the only species present from Tardigrada. Although the Tardigrada historically are placed within the Panarthropoda with Onychophora and Arthropoda, previous analyses using multigene datasets were unable to consistently place Tardigrada within the Panarthropoda [36, 201].

Similarly, two alternative topologies were present in the phylogenetic trees from this work. In the first topology, Tardigrada and Nematomorpha form a monophyletic clade, while in the second topology Tardigrada and Nematoda are sister taxa. Correct placement of tardigrades will require sampling of additional taxa both within Tardigrada and from other phyla. Nematoda and Nematomorpha did not form a monophyletic clade in all analyses.

5.3.1.2 Class Enoplia (Clade II)

Previous phylogenetic analyses were not able to robustly resolve the basal position of Enoplia. Molecular phylogenies using nSSU supported Enoplia as the first splitting clade, but with low node support. The placement of Enoplia may be important to answer the habitat of the ancestral nematode. A marine ancestry was originally proposed by Filipjev [202] and is widely accepted by the scientific community. An alternative hypothesis erected for argument by De Ley and Blaxter proposed a terrestrial origin with migration to marine habitats [28].

Although the analyses here contain one species from class Enoplia, its placement was always at the basal node of the tree (Fig. 5.3-5.6). The position of *E. brevis* may be attributed to Long Branch Attraction (LBA) artefact due to under-sampling of the order Enoplida and close outgroup species. However, due to other evidence from embryonic development [42] and low-supported single-gene phylogenies [29], the possibility of LBA seems unlikely. In order to eliminate the likelihood of LBA placing Enoplia at the basal position, additional nematode species from this order and additional species from the outgroup phyla must be sampled and used in the phylogenetic analyses.

Ancestral state reconstruction of the habitat was calculated with BayesTraits v2 [203] using the MultiState ML analysis. Each species was assigned to the corresponding environment they inhabit and along with the phylogenetic tree of these analyses (Fig. 5.3-5.6), the probability of the ancestral state of the habitat being marine was calculated as one. This analysis supports a marine origin of the phylum with multiple migrations to the land. Assuming that nematodes diverged from the other Metazoa 700 to 1000 Million years ago (Mya), the split is likely to predate the colonisation of the land in the Silurian period (~430 Mya) [5].

5.3.1.3 Class Dorylaimia (Clade I)

Six Dorylaimia species were present in the analyses from 4 of 5 described orders (*Xiphinema index*; Dorylaimida, *Romanomermis culicivorax*; Mermithida, *Prionchulus punctatus*; Mononchida, *Trichinella spiralis*; Trichinellida, *Trichuris muris* and *Trichuris suis*; Trichinellida). The relationship between the orders has not been resolved robustly in molecular analyses previously, except for the sister relationship of Mermithida and Mononchida. Previous studies have also suggested Dorylaimida as sister order to orders Mononchida, Mermithida and Trichinellida [29].

Phylogenetic analyses using SM1 showed Mermithida as sister clade to Mononchida (Fig. 5.3-5.4), in congruence with singe-gene phylogenies. Dorylaimida was recovered as sister clade to Mononchida and Mermithida. The monophyletic order Trichinellida

was recovered as sister clade to Dorylaimida, Mononchida and Mermithida. Analysis using SM3 (Fig. 5.6) showed Dorylaimida as sister clade to Trichinellida instead. For the only member of Dorylaimia present (*X. index*) only EST sequences were available, which may reflect the incongruence between the two topologies.

5.3.1.4 Class Chromadoria

Chromadoria was represented with two orders in the analyses (Plectida and Rhabditida). The Plectida (*Plectus murrayi* and *Plectus sambesii*) were in all analyses monophyletic and placed as sister taxa to Rhabditida (Fig. 5.3-5.6).

All three suborders of the order Rhabditida (Spirurina, Tylenchina, and Rhabditina) were present in the analyses. So far, phylogenetic analyses have been unable to robustly resolve the relationships between these three suborders. Molecular phylogenetics using nSSU sequences had low node support for the topology of Spirurina as sister clade to Tylenchina and Rhabditina. The multigene analysis by Desjardins et al. [43] showed support for a different topology with Tylenchina as sister clade to Spirurina plus Rhabditina. As discussed later in section 5.4, this finding is an artefact due to the absence of informative intermediate taxa. All the phylogenetic trees in this study have recovered the original topology proposed by the single gene phylogenies (Fig. 5.3-5.6).

The vast majority of the nematode species included (44 out of 53) belong to the order Rhabditida. All three suborders of Rhabditida were monophyletic in all analyses (Fig. 5.3-5.6). The relationships between the species within each suborder of Rhabditida are discussed in detail below.

5.3.1.5 Suborder Spirurina (Clade III)

The parasitic suborder Spirurina was always recovered as a monophyletic with high node support in all phylogenetic trees (Fig. 5.3-5.6). The infraorders Ascaridomorpha (1 species; *Ascaris suum*) and Spiruromorpha (10 species; 2 families) were the

only ones present from the five described infraorders. In addition 1 species from the superfamily Dracunculoidea (*Anguillicola crassus*) was present. *A. crassus* was recovered at the basal position in Spirurina as previously shown in a three locus phylogeny [204].

All the species present from the infraorder Spiruromorpha were monophyletic, in congruence with previous phylogenetic studies. *Setaria labiatopapillosa* was recovered as sister taxon to family Onchocercidae, and is considered to be a close outgroup taxon to the family. Relationships within the family Onchocercidae (filarial nematodes) were resolved with the same topology in all analyses (Fig. 5.3-5.6).

5.3.1.6 Family Onchocercidae

Previous phylogenies based on a single gene locus were unable to resolve the relationships within the family due to little variation in the conserved regions of nSSU.

Filarial nematodes are responsible for a plethora of diseases in mammals including humans. Most filarial parasites have an endosymbiont bacteria of genus *Wolbachia*, which is important for the survival of infected nematodes. However, some filarial species have lost the *Wolbachia* endosymbiont, but have traces of *Wolbachia* nuclear insertions still present in their genomes. It is hypothesized that *Wolbachia* was acquired by the last common ancestor of the filarial nematodes. The absence of the symbiont in the closely related *Setaria labiatopapillosa* species will help the understanding of the biological basis of the endosymbiosis.

Furthermore the evolutionary relationships of the filarial species in the analysis show that the symbiosis with bacteria was lost multiple times independently (Fig. 5.7). The evolutionary history of the family is also important when devising vaccines for human parasites. Closely related taxa that parasitise other mammals can be screened for drug targets and the results can inform human drug discovery.

Two clades were identified within the family Onchocercidae. The first clade consists of the genera *Dirofilaria* and *Onchocerca*, while the second clade contains the genera *Litomosoides*, *Acanthocheilonema*, *Loa*, *Brugia* and *Wuchereria*.

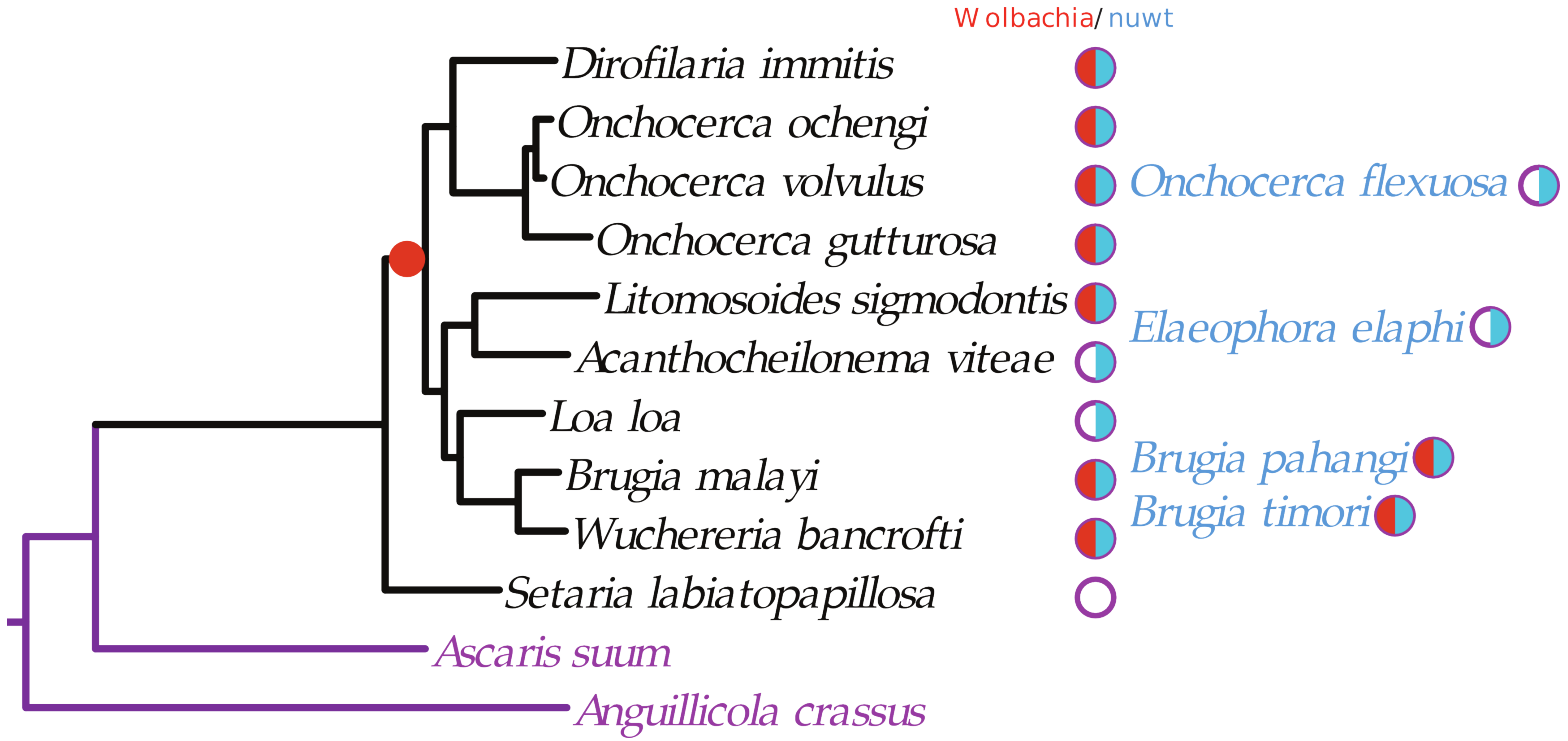


Figure 5.7 Phylogenetic tree of Clade III using SM1. Circles next to species names denote the presence or absence of *Wolbachia* endosymbiont and nuclear insertions of *Wolbachia* sequences (nuwts). Placement of newly sequenced taxa is indicated next to the phylogenetic tree. These results suggest that *Wolbachia* symbiosis was introduced in the last common ancestor of the family (red circle). Figure from Koutsovoulos and Blaxter (unpublished).

5.3.1.7 Suborder Tylenchina (Clade IV)

The suborder Tylenchina was always recovered as a monophyletic with high node support in all phylogenetic trees (Fig. 5.3-5.6).

The infraorder Panagrolaimorpha was represented by two superfamilies (Strongyloidoidea and Panagrolaimoidea) and was recovered as paraphyletic. Each superfamily was monophyletic with Strongyloidoidea (*Strongyloides* species) placed at the basal position of Tylenchina, and Panagrolaimoidea having conflicting topologies in the different analyses.

In some analyses Tylenchomorpha was monophyletic (Fig. 5.3-5.5) while in the others the relationships between the families Aphelenchidae, Aphelenchoididae and the superfamily Tylenchoidea were different (i.e Fig. 5.6). Overall it appears that the OrthoMCL dataset (SM1 and SM2) favours the monophyly of Tylenchomorpha while the CEGMA dataset (SM3) supports the paraphyly. OrthoMCL phylogenetic analyses show that the effects of taxon specific GC bias on reconstructing the evolutionary relationships of this group are minimal.

Similar patterns were observed for *Acrobeloides nanus*, the only species present from infraorder Cephalobomorpha. SM1 and SM2 analyses placed it as sister taxon to Panagrolaimoidea, while SM3 placed it as sister taxon to Tylenchoidea.

5.3.1.8 Suborder Rhabditina (Clade V)

The suborder Rhabditina was always recovered as a monophyletic with high node support in all phylogenetic trees (Fig. 5.3-5.6). The topology of Diplogasteromorpha and Rhabditomorpha found in SM1 analyses is in congruence with phylogenies using nSSU sequences [205], with the addition of Ancylostomatidae as sister clade to Strongylidae (*Cylicostephanus goldi*). The topology of Strongyloidea is different in SM3, which recovered Trichostrongylidae as paraphyletic.

5.3.2 Summary

Although the addition of more taxa may change the phylogenetic relationships, based on the observations from these analyses, previous studies and classifications of nematodes, the following statements can be made,

- ***Nematode classes.*** All three classes are monophyletic. Enoplia is the earliest splitting clade of nematodes. The Enoplia is as sister clade to the Dorylaimia and the Chromadoria.
- ***Rhabditida suborders.*** All three suborders are monophyletic. The Spirurina are sister taxa to the Tylenchina and Rhabditina.
- ***Suborder Spirurina.*** The Spiruromorpha and Ascaridomorpha are sister taxa.
- ***Family Onchocercidae.*** The relationships between the species present within the family are robustly resolved.
- ***Suborder Tylenchina.*** Panagrolaimorpha are paraphyletic with the Strongyloidea as sister taxa to the rest of Tylenchina. The Panagrolaimoidea are sister taxa to the Cephalobomorpha. Tylenchomorpha are monophyletic.
- ***Suborder Rhabditina.*** The Diplogasteromorpha and Rhabditomorpha are sister taxa. The Rhabditidae are paraphyletic.

5.4 Effects of Taxon sampling

As discussed in subsection 5.3.1.6, the phylogenetic analyses showed Clade III as a sister clade to Clades IV and V (Figure 5.8a). This finding is congruent with phylogenetic trees using nSSU. Although in these one-gene phylogenies the node support was not significant, the phylogenetic tress in this study using multiple loci have high node support. However, in [43] the authors showed high support for Clade

IV being a sister clade to Clades III and V (Figure 5.8b). In their analyses, no taxa was present from Group C while two taxa from order Plectida, which is a close outgroup to order Rhabditida, were present in the analyses described in this study.

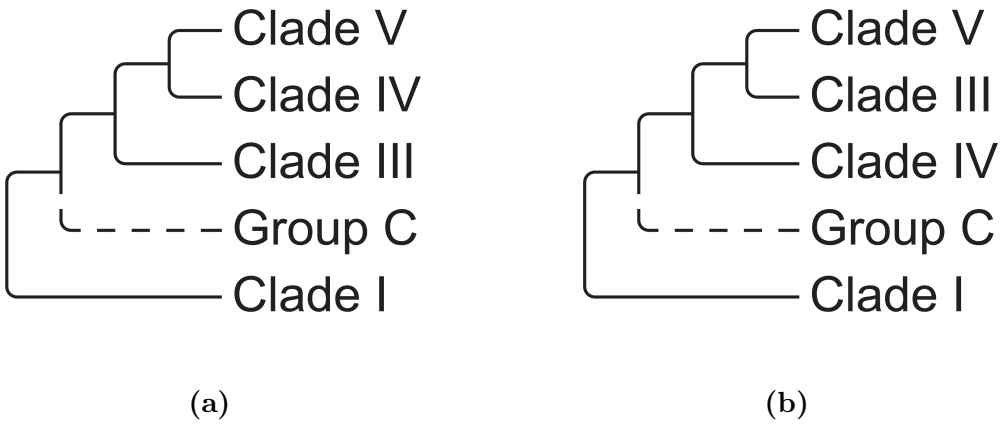
Hendy and Penny [206] have shown that the inclusion of intermediate taxa have a positive effect in parsimony algorithms converging in the correct tree. To test the effect of Plectids in the topology, the dataset (SM4) for the phylogenetic analysis from [43] was used with the addition of a plectid species. *Plectus murrayi* sequences were added into SM4 if the BLAST e-value was below $1e^{-50}$ (SM5). Next, SM4 and SM5 were used in RAxML to test the differences in topology and likelihood scores.

Table 5.6 shows the difference in likelihood scores for the two datasets tested. The inclusion of *P. murrayi* affects the topology of the order Rhabditida, and Fig. 5.8a has a better likelihood score than Fig. 5.8b. Even though only one taxon was added, the evolutionary history of the derived Clades III, IV and V is completely changed. These two topologies were also tested in SM1, and showed Fig. 5.8 having the better likelihood.

Previously, it was shown that datasets with few taxa and large number of aligned sites introduce systematic bias resulting in inaccurate phylogenies [207]. This is attributed to inaccurate estimations of the parameters due to low information in the dataset. However, the results shown here indicate that phylogenetically informative taxa have a bigger importance than the number of taxa present.

Table 5.6 RAxML likelihood scores for the different datasets (better scores in bold).

	Model	Likelihood score	
		Fig. 5.8a	Fig. 5.8b
SM4	LG+Γ	-1138695	**-1138444**
	LG+Γ+F	-1136915	**-1136666**
SM5	LG+Γ	**-1279525**	-1279532
	LG+Γ+F	**-1277661**	-1277668
SM1	LG+Γ	**-3572184**	-3572260
	LG+Γ+F	**-3566288**	-3566368



(a) **(b)**

Figure 5.8 Two alternative topologies for the order Rhabditida, (a) Clade III as the earliest splitting clade and (b) Clade IV as the earliest splitting clade. SM4 does not have any sequences from Group C

5.5 Effects of Gene sampling

The same workflow as in section 5.2 was used to infer orthologous clusters for the family Onchocercidae within the order Spirurina. The ascaridid *Ascaris suum* was used as outgroup and sequences from 7 filarial species were gathered. This analysis was conducted prior to generation of sequences from a closer outgroup species (*Setaria labiatopapillosa*) and two additional species in genus *Onchocerca*. However, the phylogenetic relationships do not change by the addition of sequences from these three species (described in subsection 5.3.1.6). OrthoMCL identified 1,809 single-copy orthologous clusters with sequences from every taxon present. The concatenation of the alignments after trimming resulted in a supermatrix with 971,074 amino acid sites (SM6). Although the number of clusters is ~2X more than in SM1, the number of aligned sites is ~6X more. This is mainly due to sequences aligning better, because of lower divergence between the species. The phylogenetic tree (Fig. 5.9) was reconstructed in RAxML (GTR+Γ model) and PhyloBayes (GTR+CAT+Γ model).

Most members of the family Onchocercidae carry the endosymbiont bacteria *Wolbachia pipientis*, and nuclear *Wolbachia* insertions can be identified in *Wolbachia*-free species of this family. The *Wolbachia* symbiosis in the last common ancestor of the family is further backed by co-evolution patterns between the nematode and the bacterium. This allows for an independent estimate of the phylogenetic relations between the members of the family and thus makes a good dataset to test the effect of gene sampling on phylogenetic reconstruction.

Gene clusters were picked randomly from the 1,809 identified orthologous clusters and concatenated together. The number of gene clusters picked was incremented by 5 in each iteration until maximum of 60 gene clusters was reached. For each iteration 10 different sets were picked to allow for sampling variation. The process is described in the pseudocode below.

Each gene set was analysed with three different partition patterns in RAxML.

Sampling gene clusters algorithm

```
Set N = 5
for N ≤ 60 do
    Create 10 sets of N random gene clusters
    Increment N by 5
end for
```

- ***No partitions.*** The concatenated matrix was used with no partitions between genes. The model chosen was GTR+Γ.

- ***Partitions with common GTR estimates.*** The concatenated matrix was partitioned per gene locus, common GTR estimates across all the partitions were used but with independent Γ distribution estimates for each partition.

- ***Partitions with different Models.*** The concatenated matrix was partitioned per gene locus. The model for each partition was chosen with prottest3.

Figure 5.10a shows the change in bootstrap values when more genes were included. All nodes except Node D converged to the same topology with node support greater than 90 when 20 genes are used, while Node D required at least 35 genes. Noticeably, nodes near the tips converged faster than the early-splitting nodes. In this dataset, 30 random orthologous clusters are enough to produce the correct phylogeny robustly. This finding can be extrapolated (with caution) to datasets with similar divergence to provide a minimum cutoff of genes need to resolve the phylogenetic relationships.

Figure 5.10b shows the difference of bootstrap values for Node D using different partition patterns. Although inferring the phylogeny without partitioning the dataset yields consistently higher bootstrap values, the effect seems to be minimal indicating that the underpinning data have the most effect in the phylogenetic reconstruction. Similar results for the other nodes were obtained. It seems that the choice of whether to partition the dataset or not has minimal effect on the outcome, with unpartitioned datasets performing slightly better.

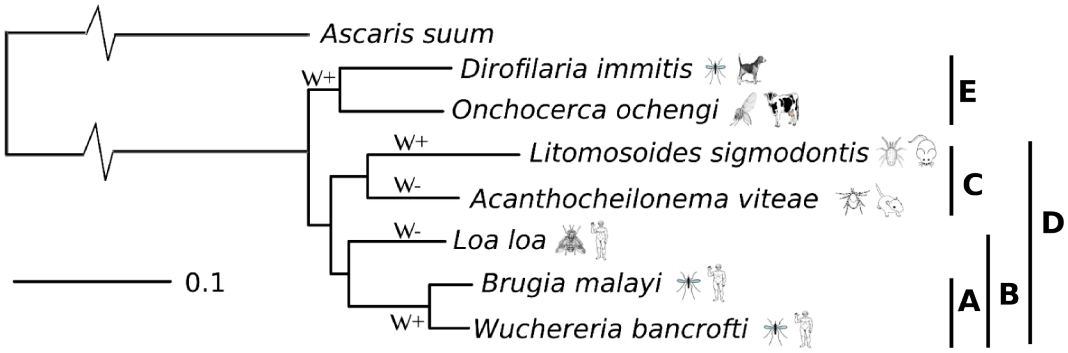
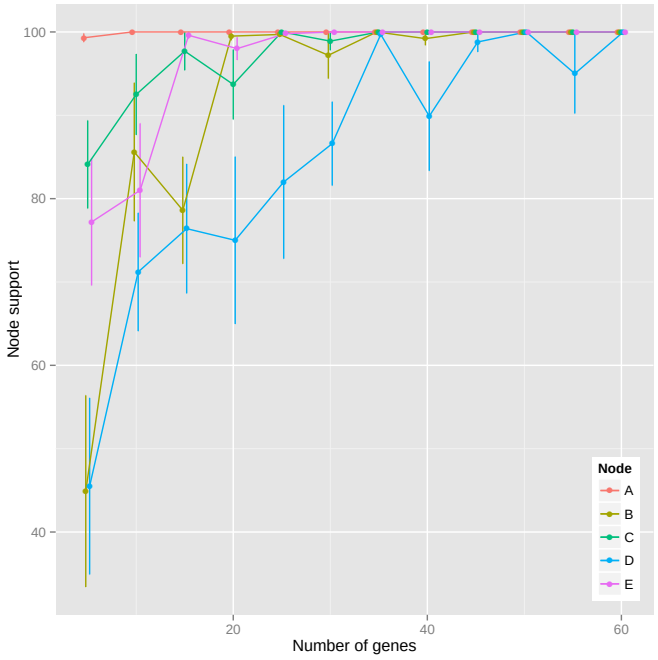
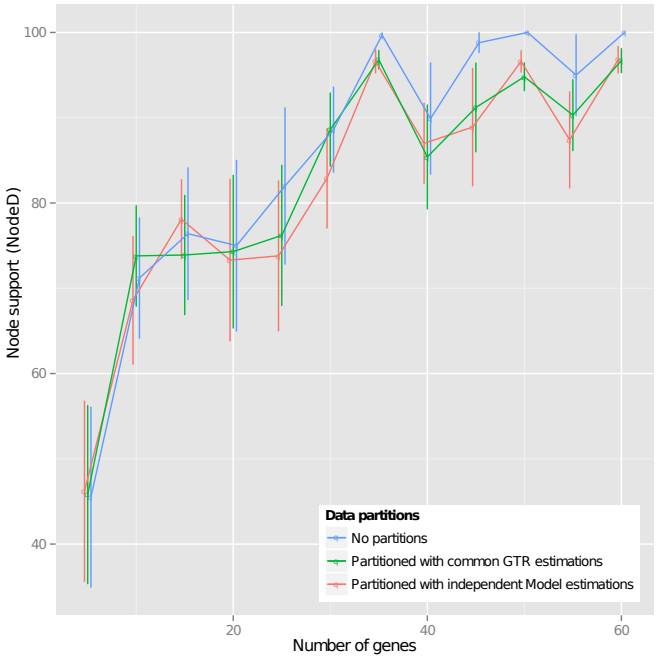


Figure 5.9 Phylogenetic tree of the family Onchocercidae from RAxML (all bootstrap values are 100) and PhyloBayes (all posterior probabilities are 1). Node names are shown on the side. Presence/absence of *Wolbachia* endosymbiont is denoted by W±. Vectors and hosts for each species are shown next to their name.



(a)



(b)

Figure 5.10 The number of genes affect the bootstrap support for (a) different nodes and (b) different data partitions.

Chapter 6

Discussion

6.1 Assembly and annotation

A more contiguous assembly could provide a more complete proteome, and thus more proteins can be clustered into orthologous groups, increasing the number of informative characters. In chapter 2, the workflow used for assembling and annotating genomes using short-read Illumina datasets is described. Only PE libraries were available for the species described in chapter 3 (except for *H. dujardini*), which resulted in fragmented assemblies. However, phylogenetically informative proteins could be identified within these assemblies, despite the draft status of the genome assemblies. In addition, assembly programs are being updated to use new technologies and longer reads. Furthermore, new assembly programs are being developed (e.g. SPAdes [53]) which use new algorithms and compute resources more efficiently. The combination of longer reads, new algorithms, and new techniques (e.g. WGA) will result in more complete genomes.

In section 3.1, I introduce a new algorithm (SCUBAT v2) which uses BLAST to scaffold contigs based on transcripts or proteins which are mapped to the assembly. The two pre-existing programs (SCUBAT v1 and L_RNA_scaffolder) use BLAT which can work only with nucleotide sequences. The three programs were tested on

a real dataset, using identical parameters. SCUBAT v2 yields significantly better results than SCUBAT v1, and similar results to L_RNA_scaffolder (albeit slightly better). Thus, SCUBAT v2 and L_RNA_scaffolder can complement each other producing a more contiguous assembly. This scaffolding approach is particularly useful in small specimens, since the amount of input DNA required for RNA-Seq libraries is lower than MP libraries. In addition, transcriptome data increase the accuracy of the genome annotations. Since the algorithm is still being actively developed, this approach has not yet been applied to the assemblies in this study. The genome assemblies outlined in chapter 3 will be scaffolded, once the program is complete and extensively tested.

The accuracy of the two-pass MAKER2 annotation pipeline was evaluated (subsection 3.2.4). The results show better accuracy at the Augustus step, indicating a conflict of annotation programs when used together to infer gene models. The performance of the pipeline can been used as an indication in terms of accuracy (on *C. elegans* and probably other nematodes). The expectations can be extrapolated to genomes with similar gene characteristics, but accuracy will most likely drop in more complex genomes and gene models. As our understanding of gene structure increases, annotation programs will be trained more efficiently. Full length transcripts in long reads could minimise the error of chimeric transcripts assembled and alternative spliced isoforms will be easier to predict.

6.2 Current status of nematode genomes

Currently, there are 37 nematode genomes available (including those in Table 5.1) and 5 presented here which will become available soon. An initiative to sequence 959 nematode genomes was started in 2012, and a wiki was created to coordinate collaborations and minimise duplication of efforts for the same species [208]. The 959 initiative notes that more than ~100 genomic projects are ongoing [46] and will be completed soon. Although the 959 genomes target has yet to be reached, as

sequencing technologies become more accessible this goal will be surpassed in a few years time. As seen in Figure 1.2, most sequencing projects have been targeted to the derived Clades III, IV and V, while only a few species of the basal Clades I and II have had their genome sequenced. An effort has to be made to sample species from under-represented portions of the phylogenetic tree in order to sample the full divergence of the phylum.

6.3 Current status of nematode phylogenetics

Prior attempts to resolve the deep phylogenetic relationships of Nematoda with multigene datasets have mostly supported the existing nSSU phylogeny [43, 128, 179]. However, the representation of taxa was limited. A previous phylogenetic analysis with 23 nematode species (from species described here and publicly available datasets) was published in 2014 [209]. The datasets analysed here were updated to include 53 nematode species.

Although the sample size can still be considered small compared to previous nSSU phylogenies, an effort was made to capture as much of the divergence of this large phylum as possible. During this study, NGS datasets were generated for the first time for the following taxonomic ranks

- ***Class (1):*** Enoplia.
- ***Order (4):*** Enoplida, Mermithida, Mononchida, and Plectida.
- ***Superfamily (6):*** Enoploidea, Mermithoidea, Mononchoidea, Plectoidea, Ascaridoidea, and Cephaloboidea.
- ***Genus (12):*** *Enoplus*, *Romanomermis*, *Prionchulus*, *Plectus*, *Ascaris*, *Dirofilaria*, *Acanthocheilonema*, *Setaria*, *Acrobeloides*, *Pseudaphelenchus*, *Rhabditis* and *Dictyocaulus*.

The phylogenetic analyses of the 53 nematode species has provided a better resolution both at deep nodes and between terminal nodes. Still, many groups of nematodes are missing representation while others are under-represented. For example, the resolution at the basal node was achieved with only one species from the class Enoplia (order Enoplida). A previous phylogenetic study focussing primarily on the Enoplida, showed a high level of divergence within the order [33]. In addition, most orders of Chromadoria do not have any NGS data yet. However, the phylogenetic relationships proposed here can act as a stepping stone towards a better understanding of the evolutionary forces that mould a phylum. The future for nematode systematics looks exciting.

6.3.1 Phylogenetic conclusions

The phylogenetic analyses described in this study advances the knowledge of the evolutionary history of the phylum. The questions that were outlined at the beginning of this project (section 1.6) were explored in accordance to the datasets sampled.

- The relationships of the three nematode classes (Enoplia, Dorylaimia and Chromadoria) have been resolved using a multigene dataset. Enoplia emerge as the earliest splitting clade, and subsequent analyses place the ancestral nematode in a marine environment. Thus, the approximation of the time of the colonisation of the land can be placed in the Silurian period.

- The Rhabditida suborders, which contain a plethora of important parasites, have been also robustly resolved. The knowledge of the phylogenetic relationships of these suborders allows the identification of common evolutionary histories between parasites. Plant parasitism is shown to have arisen at a later stage than animal parasitism.

- The phylogenetic relationships of the species of the family Onchocercidae have been resolved. Their symbiosis with the bacteria *Wolbachia* has been mapped

on the phylogenetic tree providing evidence for an ancestral acquisition of *Wolbachia* in the last common ancestor of all the species of the family.

- The relationships of the infraorders of the suborder Tylenchina have been resolved. Previous problems with difference in AT-content seem to not affect the amino acid dataset used in the phylogenetic analyses of this study.

6.4 Phylogenomics

Multigene datasets appear to be effective in resolving phylogenetic relationships. However, the methodologies have not been extensively tested. For example, Salichos and Rokas [38] identified problems with the resulting topologies. Nodes were highly supported when genes were concatenated together but some nodes were supported by very few individual gene trees. In the future, different phylogenetic algorithms may be proposed for multigene datasets, to account for the intricacies of the underlying data.

As shown in section 5.4, missing taxa between evolutionary distant species have an effect on the resulting topology. In that case, the different topology has severe implications on the evolutionary history of the order Rhabditida. In other cases, the effect may be minimal. However, the ramifications of missing taxa have not been properly tested for vast datasets. For example, a dataset may require more sampling to produce more robust phylogenies.

Furthermore, the view on the gene-tree/species-tree problem is already changing. Speciation is not a bifurcating process and genes between sister taxa may have a different history. Once chromosome complete genomes become more accessible, the genome architecture can provide a new prospective on species evolution.

6.5 Future work

This study will lead to two future projects. The first project is to finish the SCUBAT v2 algorithm. At the moment, the final part of scaffolding the contigs based on the transcriptome information is missing and is currently under development and testing. Furthermore, the algorithm will be looked further to improve the runtime and memory consumption. Finally, the code is being streamlined to provide more clarity and readability.

The second project is to use the information of the orthologous gene clusters produced in this study to identify differences in new sequence data from new nematode species. The information can be used for assessing completeness of genome assemblies and create gene models that can be used in annotation pipelines.

Appendices

Appendix A

Orthologous clustering details

Table A.1 Number of proteins in OrthoMCL and CEGMA clusters

Species	Group	Before OrthoMCL	4,737 clusters	959 clusters	CEGMA clusters
Drosophila melanogaster	Arthropoda	15,264	4,267	545	410
Bombyx mori	Arthropoda	14,384	3,471	490	356
Tetranychus urticae	Arthropoda	16,389	3,658	504	364
Hypsibius dujardini	Tardigrada	21,466	3,630	338	295
Gordius sp.	Nematomorpha	6,663	2,956	437	361
Romanomermis culicivorax	Clade I	47,001	4,453	465	292
Trichinella spiralis	Clade I	13,826	3,335	362	270
Trichuris suis	Clade I	13,287	3,128	390	269
Prionchulus punctatus	Clade I	14,584	4,353	545	365
Trichuris muris	Clade I	2,724	655	41	57
Xiphinema index	Clade I	4,069	1,210	100	128
Enoplus brevis	Clade II	13,244	3,430	505	370
Plectus murrayi	Group C	17,683	4,576	789	396
Plectus sambesii	Group C	19,833	5,410	822	397
Ascaris suum	Clade III	15,036	4,898	785	383
Brugia malayi	Clade III	13,754	4,916	858	391
Dirofilaria immitis	Clade III	11,977	4,557	794	361
Loa loa	Clade III	14,743	4,800	882	389
Onchocerca volvulus	Clade III	12,782	4,568	832	369
Onchocerca gutturosa	Clade III	19,630	4,980	593	304
Onchocerca ochengi	Clade III	12,935	4,901	723	338
Litomosoides sigmodontis	Clade III	10,039	4,580	779	365
Acanthocheilonema viteae	Clade III	10,118	4,524	707	319
Wuchereria bancrofti	Clade III	18,751	5,669	682	320
Setaria labiatopapillosa	Clade III	8,616	4,272	825	376
Anguillicola crassus	Clade III	6,801	2,340	183	204

Number of proteins in OrthoMCL and CEGMA clusters (continued)

Species	Group	Before OrthoMCL	4,737 clusters	959 clusters	CEGMA clusters
Bursaphelenchus xylophilus	Clade IV	17,693	4,879	795	377
Globodera pallida	Clade IV	14,798	3,791	401	281
Globodera rostochiensis	Clade IV	13,328	3,984	461	262
Meloidogyne floridensis	Clade IV	43,076	5,359	419	237
Meloidogyne hapla	Clade IV	13,230	3,808	549	318
Meloidogyne incognita	Clade IV	15,696	3,692	452	298
Panagrellus redivivus	Clade IV	24,271	4,999	846	384
Pseudaphelenchus vindai	Clade IV	5,973	3,455	413	354
Strongyloides ratti	Clade IV	7,627	2,650	301	362
Panagrolaimus sp. ES5	Clade IV	22,266	5,135	853	386
Panagrolaimus sp. PS1159	Clade IV	20,755	5,208	849	390
Propanagrolaimus sp. JUL765	Clade IV	17,882	5,052	857	389
Acrobeloides nanus	Clade IV	9,519	3,808	347	309
Strongyloides stercocalis	Clade IV	3,147	1,333	85	106
Aphelenchus avenae	Clade IV	1,793	666	57	64
Panagrolaimus superbus	Clade IV	3,589	776	68	62
Heterodera glycines	Clade IV	8,171	2,444	252	209
Caenorhabditis elegans	Clade V	20,937	5,275	905	412
Caenorhabditis angaria	Clade V	25,541	4,986	542	310
Caenorhabditis briggsae	Clade V	20,809	4,861	840	406
Caenorhabditis sp. 5	Clade V	35,096	5,274	817	384
Haemonchus contortus	Clade V	17,508	5,179	729	350
Heterorhabditis bacteriophora	Clade V	19,898	3,616	282	147
Necator americanus	Clade V	18,748	4,933	600	330
Pristionchus pacificus	Clade V	23,232	4,270	432	252
Pristionchus exspectatus	Clade V	23,283	4,057	417	230
Dictyocaulus viviparus	Clade V	14,124	4,845	613	316
Ancylostoma ceylanicum	Clade V	40,096	5,748	768	368
Rhabditis sp. SB347	Clade V	11,087	4,670	890	406
Cylicostephanus goldi	Clade V	9,521	3,350	273	230
Ancylostoma caninum	Clade V	22,169	3,830	339	266
Nippostrongylus brasiliensis	Clade V	2,923	683	101	118

Appendix B

Nematode systematics

Table B.1 Nematode species systematics

Class	Order	Suborder	Infraorder	Superfamily	Family	Genus	Species
Enoplia (Clade II)	Enoplida	Enoplina		Enoploidea	Enoplidae	Enoplus	*E. brevis*
Dorylaimia (Clade I)	Mermithida	Mermithina		Mermithoidea	Mermithidae	Romanomermis	*R. culicivorax*
	Dorylaimida	Dorulaimina		Dorylaimoidea	Longidoridae	Xiphinema	*X. index*
	Mononchida	Mononchina		Mononchoidea	Mononchidae	Prionchulus	*P. punctatus*
	Trichinellida			Trichinelloidea	Trichinellidae	Trichinella	*T. spiralis*
					Trichuridae	Trichuris	*T. muris*
							T. suis
Chromadoria	Plectida			Plectoidea	Plectidae	Plectus	*P. murrayi*
							P. sambesii
	Rhabditida	Spirurina (Clade III)		Dracunculoidea	Anguillicolidae	Anguillicola	*A. crassus*
			Ascaridomorpha	Ascaridoidea	Ascarididae	Ascaris	*A. suum*
			Spiruromorpha	Filarioidea	Onchocercidae	Brugia	*B. malayi*
						Dirofilaria	*D. immitis*
						Loa	*L. loa*
						Onchocerca	*O. gutturosa*
							O. ochengi
							O. volvulus
						Litomosoides	*L. sigmodontis*
						Acanthocheilonema	*A. viteae*
						Wuchereria	*W. bancrofti*
					Setariidae	Setaria	*S. labiatopapillosa*

Nematode species systematics (continued)

Class	Order	Suborder	Infraorder	Superfamily	Family	Genus	Species
Chromadoria	Rhabditida	Tylenchina (Clade IV)	Cephalobomorpha	Cephaloboidea	Cephalobidae	Acrobeloides	*A. nanus*
			Panagrolaimorpha	Panagrolaimoidea	Panagrolaimidae	Panagrellus	*P. redivivus*
						Panagrolaimus	*P.* sp. ES5
							P. sp. PS1159
							P. superbus
						Propanagrolaimus	*P.* sp. JUL765
				Strongyloidoidea	Strongyloididae	Strongyloides	*S. ratti*
							S. stercocalis
			Tylenchomorpha	Aphelenchoidea	Aphelenchidae	Aphelenchus	*A. avenae*
					Aphelenchoididae	Bursaphelenchus	*B. xylophilus*
						Pseudaphelenchus	*P. vindai*
				Tylenchoidea	Heteroderidae	Globodera	*G. pallida*
							G. rostochiensis
						Heterodera	*H. glycines*
					Meloidogynidae	Meloidogyne	*M. floridensis*
							M. hapla
							M. incognita
		Rhabditina (Clade V)	Diplogasteromorpha		Neodiplogasteridae	Pristionchus	*P. exspectatus*
							P. pacificus
			Rhabditomorpha	Rhabditoidea	Rhabditidae	Caenorhabditis	*C. angaria*
							C. briggsae
							C. elegans
							C. sp. 5
						Rhabditis	*R.* sp. SB347
				Strongyloidea	Ancylostomatidae	Ancylostoma	*A. caninum*
							A. ceylanicum
						Necator	*N. americanus*
					Heterorhabditidae	Heterorhabditis	*H. bacteriophora*
					Strongylidae	Cylicostephanus	*C. goldi*
					Trichostrongylidae	Dictyocaulus	*D. viviparus*
						Haemonchus	*H. contortus*
						Nippostrongylus	*N. brasiliensis*

Appendix C

FastTree

FastTree was used to calculate phylogenetic trees using the data from SM1 and SM3. The use of FastTree was meant to produce an initial fast approximation of the phylogenies, and the results were not considered for the phylogenetic conclusions. The alignments were analysed with JTT+**CAT** model, and the support for the nodes was calculated with the Shimodaira-Hasegawa (SH) test [210]. Usually, 1,000 alignments are re-sampled on the other three possible topologies around the split and the SH test is used to compare the topologies.

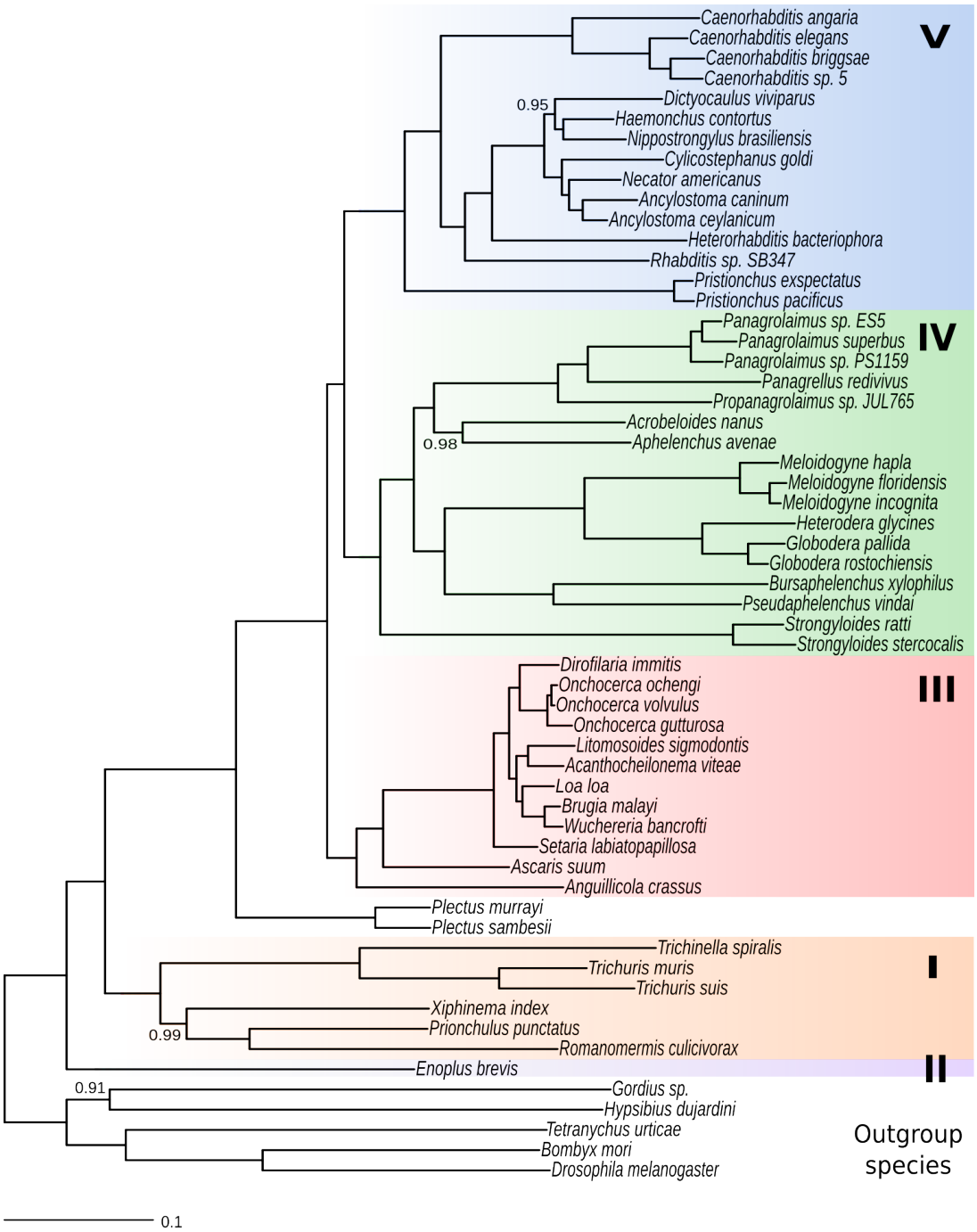


Figure C.1 Phylogenetic tree obtained from FastTree using SM1. SH support values below 1 are shown. Coloured boxes represent the five clades identified by Blaxter et al. [26]

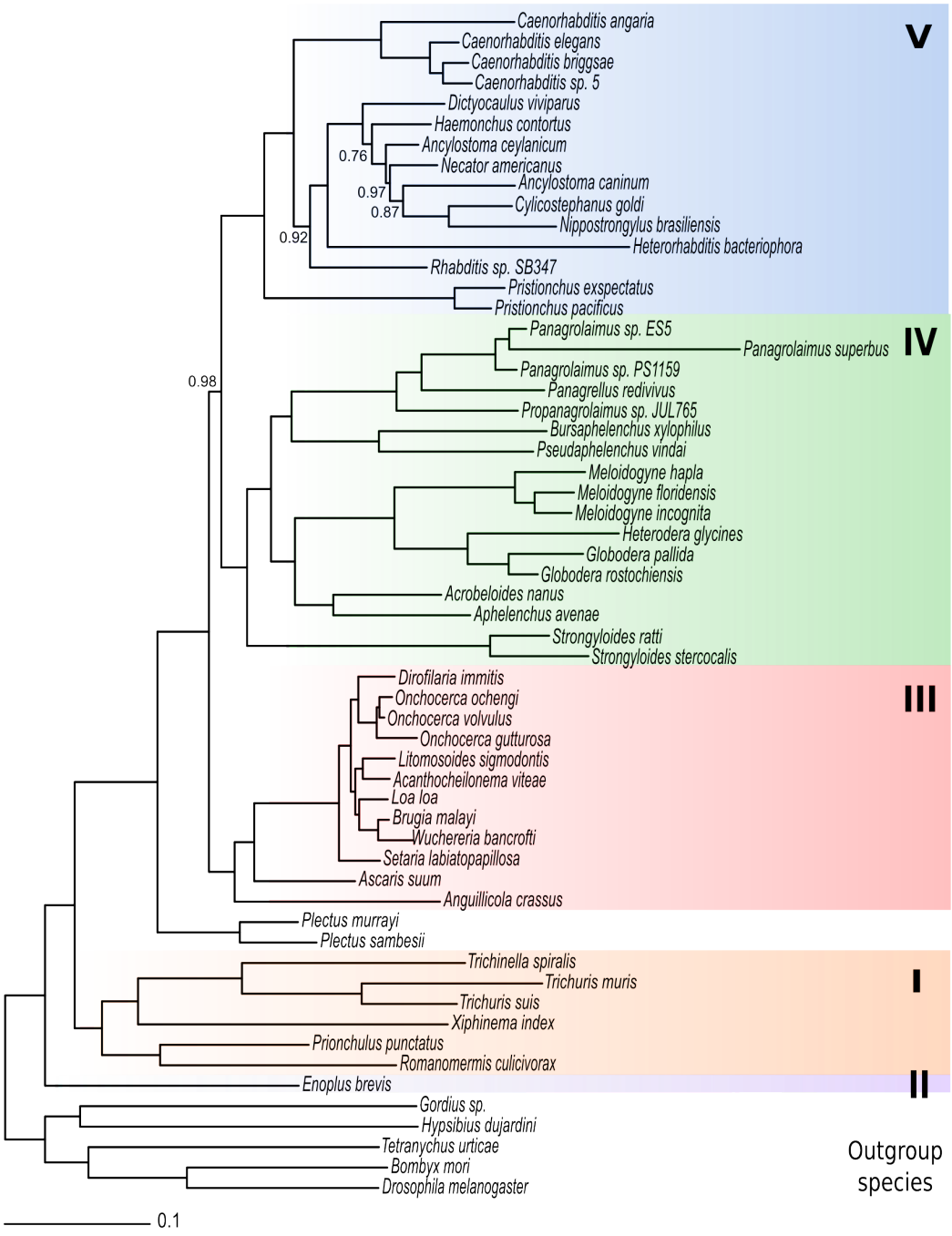


Figure C.2 Phylogenetic tree obtained from FastTree using SM3. SH support values below 1 are shown. Coloured boxes represent the five clades identified by Blaxter et al. [26]

Bibliography

[1] J.-P. Hugot, P. Baujard, and S. Morand, "Biodiversity in helminths and nematodes as a field of study: an overview," *Nematology*, vol. 3, pp. 199–208, Jul 2001.

[2] A. Maggenti, *General Nematology*. Springer-Verlag New York, Inc., 1981.

[3] G. N. Agrios, *Plant Pathology*. Elsevier Academic Press, 5th ed., 2005.

[4] D. L. Lee, *The Biology of Nematodes*, ch. 3. Taylor and Francis, 2002.

[5] B. H. Meldal, N. J. Debenham, P. De Ley, I. T. De Ley, J. R. Vanfleteren, A. R. Vierstraete, W. Bert, G. Borgonie, T. Moens, P. A. Tyler, and et al., "An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa," *Molecular Phylogenetics and Evolution*, vol. 42, pp. 622–636, Mar 2007.

[6] H. Wu, P. Chen, and T. Tsay, "Assessment of nematode community structure as a bioindicator in river monitoring," *Environmental Pollution*, vol. 158, pp. 1741–1747, May 2010.

[7] J. W. Doran, A. J. Jones, J. M. Blair, P. J. Bohlen, and D. W. Freckman, "Soil invertebrates as indicators of soil quality," *SSSA Special Publication*, 1996.

[8] P. J. Hotez, D. H. Molyneux, A. Fenwick, E. Ottesen, S. Ehrlich Sachs, and J. D. Sachs, "Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria," *PLoS Med*, vol. 3, p. e102, 01 2006.

[9] S. Lustigman, D. Abraham, and T. R. Klei, *Parasitic Helminths: Targets, Screens, Drugs and Vaccines*, ch. 23. Wiley-Blackwell, 2012.

[10] W. Decraemer and D. J. Hunt, "Structure and classification.," *Plant nematology*, pp. 3–32, 2006.

[11] J. M. Nicol, S. J. Turner, D. L. Coyne, L. den Nijs, S. Hockland, and Z. T. Maafi, *Genomics and Molecular Genetics of Plant-Nematode Interactions*, ch. 2. Springer, 2011.

[12] *Caenorhabditis elegans* Sequencing Consortium, "Genome sequence of the Nematode *C. elegans*: A platform for investigating biology," *Science*, vol. 282, pp. 2012–2018, Dec 1998.

[13] L. D. Stein, Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, and et al., "The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics," *PLoS Biology*, vol. 1, no. 2, p. e5, 2003.

[14] Washington University Genome Sequencing Center. http://genome.wustl.edu/genomes/category/invertebrates/.

[15] A. Mortazavi, E. M. Schwarz, B. Williams, L. Schaeffer, I. Antoshechkin, B. J. Wold, and P. W. Sternberg, "Scaffolding a *Caenorhabditis* nematode genome with RNA-seq," *Genome Research*, 2010.

[16] S. Kumar, *Next-generation nematode genomes*. PhD thesis, The University of Edinburgh, June 2013.

[17] A. Coomans, "Present status and future of nematode systematics.," *Nematology*, no. 573-582 Part 5, 2002.

[18] G. Poinar, "Trends in the evolution of insect parasitism by nematodes as inferred from fossil evidence," *J. Nematol.*, vol. 35, pp. 129–132, Jun 2003.

[19] H. Micoletzky, "Die freilebenden Erd-Nematoden, mit besonderer Berücksichtigung der Steiermark und der Bukowina, zugleich mit einer Revision sämtlicher nicht mariner, freilebender Nematoden in Form von Genus-Beschreibungen und Bestimmungsschlüsseln," *Archiv fr Naturgeschichte, Abteilung A*, vol. 87, pp. 1–650, 1922.

[20] B. G. Chitwood and M. B. Chitwood, "The characters of protonematodes," *Journal of parasitology*, vol. 20, p. 130, 1933.

[21] B. G. Chitwood, "A revised classification of the Nematoda," *Papers in Helminthology published in commemoration of the 30 year jubileum of K. J. Skrjabin AND of 15th anniversary of the All-Union Institute of Helminthology*, pp. 67–79, 1937.

[22] B. G. Chitwood, "The designation of official names for higher taxa of invertebrates," *Bulletin of Zoological Nomenclature*, vol. 15, no. 25/28, pp. 860–895, 1958.

[23] A. R. Maggenti, "Comparative morphology in nemic phylogeny," *In the lower Metazoa, comparative biology and phylogeny*, pp. 273–282, 1963.

[24] I. Andrássy, *Evolution As a Basis for the Systematization of Nematodes*. Akadémiai Kiadó, 1976.

[25] S. Lorenzen, *Entwurf eines phylogenetischen Systems der freilebenden Nematoden*. 1979.

[26] M. L. Blaxter, P. De Ley, J. R. Garey, L. X. Liu, P. Scheldeman, A. Vierstraete, J. R. Vanfleteren, L. Y. Mackey, M. Dorris, L. M. Frisse, *et al.*, "A molecular evolutionary framework for the phylum Nematoda," *Nature*, vol. 392, no. 6671, pp. 71–75, 1998.

[27] P. De Ley and M. Blaxter, "Systematic position and phylogeny," *The biology of nematodes*, pp. 1–30, 2002.

[28] P. De Ley and M. Blaxter, *A new system for Nematoda: combining morphological characters with molecular trees, and translating clades into ranks and taxa*, vol. 2, pp. 633–653. E.J. Brill, 2004.

[29] M. Holterman, A. van der Wurff, S. van den Elsen, H. van Megen, T. Bongers, O. Holovachov, J. Bakker, and J. Helder, "Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1792–1800, 2006.

[30] S. A. Nadler, R. Carreno, H. Mejía-Madrid, J. Ullberg, C. Pagan, R. Houston, and J.-P. Hugot, "Molecular phylogeny of clade III nematodes reveals multiple origins of tissue parasitism," *Parasitology*, vol. 134, no. 10, pp. 1421–1442, 2007.

[31] W. Bert, F. Leliaert, A. R. Vierstraete, J. R. Vanfleteren, and G. Borgonie, "Molecular phylogeny of the Tylenchina and evolution of the female gonoduct (Nematoda: Rhabditida)," *Molecular phylogenetics and evolution*, vol. 48, no. 2, pp. 728–744, 2008.

[32] M. Holterman, G. Karssen, S. Van Den Elsen, H. Van Megen, J. Bakker, and J. Helder, "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*, vol. 99, no. 3, pp. 227–235, 2009.

[33] H. Bik, P. J. Lambshead, W. K. Thomas, and D. Lunt, "Moving towards a complete molecular framework of the Nematoda: a focus on the Enoplida and early-branching clades," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 353, 2010.

[34] H. van Megen, S. van den Elsen, M. Holterman, G. Karssen, P. Mooyman, T. Bongers, O. Holovachov, J. Bakker, and J. Helder, "A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences," *Nematology*, vol. 11, no. 6, pp. 927–950, 2009.

[35] S. K. Stenroos and P. T. DePriest, "SSU rDNA phylogeny of cladoniiform lichens," *American Journal of Botany*, vol. 85, no. 11, pp. 1548–1559, 1998.

[36] C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, *et al.*, "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature*, vol. 452, no. 7188, pp. 745–749, 2008.

[37] J. C. Regier, J. W. Shultz, A. Zwick, A. Hussey, B. Ball, R. Wetzer, J. W. Martin, and C. W. Cunningham, "Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences," *Nature*, vol. 463, no. 7284, pp. 1079–1083, 2010.

[38] L. Salichos and A. Rokas, "Inferring ancient divergences requires genes with strong phylogenetic signals," *Nature*, vol. 497, pp. 327–331, May 2013.

[39] M. L. Metzker, "Sequencing technologies-the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2009.

[40] J. J. Wiens, "Missing data, incomplete taxa, and phylogenetic accuracy," *Systematic Biology*, vol. 52, no. 4, pp. 528–538, 2003.

[41] J. E. De La Torre-Bárcena, S.-O. Kolokotronis, E. K. Lee, D. W. Stevenson, E. D. Brenner, M. S. Katari, G. M. Coruzzi, and R. DeSalle, "The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data," *PLoS One*, vol. 4, no. 6, p. e5764, 2009.

[42] E. Schierenberg, "Unusual cleavage and gastrulation in a freshwater nematode: developmental and phylogenetic implications," *Development genes and evolution*, vol. 215, no. 2, pp. 103–108, 2005.

[43] C. A. Desjardins, G. C. Cerqueira, J. M. Goldberg, J. C. D. Hotopp, B. J. Haas, J. Zucker, J. M. Ribeiro, S. Saif, J. Z. Levin, L. Fan, *et al.*, "Genomics of *Loa loa*, a *Wolbachia*-free filarial parasite of humans," *Nature genetics*, vol. 45, no. 5, pp. 495–500, 2013.

[44] J. Hallan, "Synopsis of the described nematoda of the world," 2007. https://insects.tamu.edu/research/collection/hallan/.

[45] A. Hoerauf, K. Nissen-Phle, C. Schmetz, K. Henkle-Dhrsen, M. L. Blaxter, D. W. Bttner, M. Y. Gallin, K. M. Al-Qaoud, R. Lucius, and B. Fleischer, "Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility.," *J Clin Invest*, vol. 103, pp. 11–18, Jan 1999.

[46] M. Blaxter and G. Koutsovoulos, "The evolution of parasitism in Nematoda," *Parasitology*, pp. 1–14, 2014.

[47] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[48] S. Anderson, "Shotgun DNA sequencing using cloned DNase I-generated fragments," *Nucleic Acids Research*, vol. 9, no. 13, pp. 3015–3027, 1981.

[49] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proceedings of the National Academy of Sciences*, vol. 98, no. 17, pp. 9748–9753, 2001.

[50] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, "ABySS: a parallel assembler for short read sequence data," *Genome research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[51] CLC-bio website. http://www.clcbio.com.

[52] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, *et al.*, "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler," *Gigascience*, vol. 1, no. 1, p. 18, 2012.

[53] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, and et al., "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *Journal of Computational Biology*, vol. 19, pp. 455–477, May 2012.

[54] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–829, 2008.

[55] P. E. Compeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nature biotechnology*, vol. 29, no. 11, pp. 987–991, 2011.

[56] P. A. Pevzner, H. Tang, and G. Tesler, "De novo repeat classification and fragment assembly," *Genome research*, vol. 14, no. 9, pp. 1786–1796, 2004.

[57] Y. Dong, M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp, J. Wang, S. Yang, J. Liang, and et al., "Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*)," *Nature Biotechnology*, vol. 31, pp. 135–141, Dec 2012.

[58] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic acids research*, vol. 38, no. 6, pp. 1767–1771, 2010.

[59] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, "Quake: quality-aware detection and correction of sequencing errors," *Genome Biology*, vol. 11, no. 11, p. R116, 2010.

[60] J. Schröder, H. Schröder, S. J. Puglisi, R. Sinha, and B. Schmidt, "SHREC: a short-read error correction method," *Bioinformatics*, vol. 25, no. 17, pp. 2157–2163, 2009.

[61] S. I. Nikolenko, A. I. Korobeynikov, and M. A. Alekseyev, "BayesHammer: Bayesian clustering for error correction in single-cell sequencing," *BMC Genomics*, vol. 14, no. Suppl 1, p. S7, 2013.

[62] T. Magoc and S. L. Salzberg, "FLASH: fast length adjustment of short reads to improve genome assemblies," *Bioinformatics*, vol. 27, pp. 2957–2963, Sep 2011.

[63] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC Bioinformatics*, vol. 10, no. 1, p. 421, 2009.

[64] S. Kumar, M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots," *Frontiers in genetics*, vol. 4, 2013.

[65] M. Crusoe, G. Edvenson, J. Fish, A. Howe, E. McDonald, J. Nahum, K. Nanlohy, H. Ortiz-Zuazaga, J. Pell, J. Simpson, C. Scott, R. R. Srinivasan, Q. Zhang, and C. T. Brown, "The khmer software package: enabling efficient sequence analysis," 04 2014.

[66] G. Parra, K. Bradnam, and I. Korf, "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.," *Bioinformatics*, vol. 23, no. 9, pp. 1061–1067, 2007.

[67] M. Boetzer and W. Pirovano, "Toward almost closed genomes with GapFiller," *Genome Biology*, vol. 13, no. 6, p. R56, 2012.

[68] D. Earl, K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, and et al., "Assemblathon 1: A competitive assessment of de novo short read assembly methods," *Genome Research*, vol. 21, pp. 2224–2241, Sep 2011.

[69] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, and et al., "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 1513–1518, Dec 2010.

[70] J. T. Simpson and R. Durbin, "Efficient de novo assembly of large genomes using compressed data structures," *Genome Research*, vol. 22, pp. 549–556, Dec 2011.

[71] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, and et al., "GAGE: A critical evaluation of genome assemblies and assembly algorithms," *Genome Research*, vol. 22, pp. 557–567, Jan 2012.

[72] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, and et al., "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species," *Giga Sci*, vol. 2, no. 1, p. 10, 2013.

[73] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, and et al., "Full-length transcriptome assembly from RNA-seq data without a reference genome," *Nature Biotechnology*, vol. 29, pp. 644–652, May 2011.

[74] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[75] SCUBAT. https://github.com/elswob/SCUBAT.

[76] W. J. Kent, "BLAT—the BLAST-Like Alignment Tool," *Genome Research*, vol. 12, p. 656664, Mar 2002.

[77] X. Huang, "CAP3: A dna sequence assembly program," *Genome Research*, vol. 9, p. 868877, Sep 1999.

[78] W. Xue, J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong, Y.-Y. Kuang, and X.-W. Sun, "L_RNA_scaffolder: scaffolding genomes with transcripts," *BMC Genomics*, vol. 14, no. 1, p. 604, 2013.

[79] SCUBAT v2. https://github.com/GDKO/SCUBAT2.

[80] K. Yook, T. W. Harris, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, N. De La Cruz, A. Duong, R. Fang, *et al.*, "WormBase 2012: more genomes, more data, new website," *Nucleic acids research*, vol. 40, no. D1, pp. D735–D741, 2012.

[81] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," *Genome Biology*, vol. 5, no. 2, p. R12, 2004.

[82] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, and et al., "The RAST server: Rapid Annotations using Subsystems Technology," *BMC Genomics*, vol. 9, no. 1, p. 75, 2008.

[83] M. R. Brent and R. Guigo, "Recent advances in gene structure prediction," *Current Opinion in Structural Biology*, vol. 14, pp. 264–272, Jun 2004.

[84] M. Yandell and D. Ence, "A beginner's guide to eukaryotic genome annotation," *Nature Reviews Genetics*, vol. 13, pp. 329–342, Apr 2012.

[85] UniProt Consortium, "Activities at the Universal Protein Resource (UniProt)," *Nucleic acids research*, vol. 42, no. D1, pp. D191–D198, 2014.

[86] R. Schmid and M. L. Blaxter, "annot8r: GO, EC and KEGG annotation of EST datasets," *BMC Bioinformatics*, vol. 9, no. 1, p. 180, 2008.

[87] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, and et al., "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, pp. 1236–1240, Jan 2014.

[88] A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, pp. 3674–3676, Aug 2005.

[89] T. M. Lowe and S. R. Eddy, "tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence," *Nucleic Acids Research*, vol. 25, pp. 955–964, Mar 1997.

[90] E. P. Nawrocki and S. R. Eddy, "Infernal 1.1: 100-fold faster RNA homology searches," *Bioinformatics*, vol. 29, pp. 2933–2935, Sep 2013.

[91] R. Guigó, P. Flicek, J. F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V. B. Bajic, E. Birney, and et al., "EGASP: the human ENCODE genome annotation assessment project," *Genome Biology*, vol. 7, no. Suppl 1, p. S2, 2006.

[92] A. Coghlan, T. J. Fiedler, S. J. McKay, P. Flicek, T. W. Harris, D. Blasiar, The nGASP Consortium, and L. D. Stein, "nGASP - the nematode genome annotation assessment project," *BMC Bioinformatics*, vol. 9, no. 1, p. 549, 2008.

[93] RNA-seq Genome Annotation Assessment Project. http://www.gencodegenes.org/rgasp/.

[94] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "AUGUSTUS: ab initio prediction of alternative transcripts," *Nucleic Acids Research*, vol. 34, pp. W435–W439, Jul 2006.

[95] V. Solovyev, P. Kosarev, I. Seledsov, and D. Vorobyev, "Automatic annotation of eukaryotic genes, pseudogenes and promoters," *Genome Biology*, vol. 7, no. Suppl 1, p. S10, 2006.

[96] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, and et al., "*mGene*: Accurate SVM-based gene finding with an application to nematode genomes," *Genome Research*, vol. 19, pp. 2133–2143, Nov 2009.

[97] C. Holt and M. Yandell, "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects," *BMC Bioinformatics*, vol. 12, no. 1, p. 491, 2011.

[98] I. Korf, "Gene finding in novel genomes," *BMC Bioinformatics*, vol. 5, no. 1, p. 59, 2004.

[99] V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, and M. Borodovsky, "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training," *Genome Research*, vol. 18, pp. 1979–1990, Oct 2008.

[100] E. Keibler and M. R. Brent, "Eval: A software package for analysis of genome annotations," *BMC Bioinformatics*, vol. 4, no. 1, p. 50, 2003.

[101] Babraham Bioinformatics, "FastQC: A quality control tool for high throughput sequence data." http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[102] E. Aronesty, "ea-utils: Command-line tools for processing biological sequencing data," 2011.

[103] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, p. btu170, 2014.

[104] Assemblage scripts. https://github.com/sujaikumar/assemblage.

[105] TAGC scripts. https://github.com/DRL/tagc_plots.

[106] S. R. Eddy, "Accelerated profile HMM searches," *PLoS Comput Biol*, vol. 7, p. e1002195, Oct 2011.

[107] J. C. Wright, P. Westh, and H. Ramløv, "Cryptobiosis in Tardigrada," *Biological reviews*, vol. 67, no. 1, pp. 1–29, 1992.

[108] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, "Scaffolding pre-assembled contigs using SSPACE," *Bioinformatics*, vol. 27, no. 4, pp. 578–579, 2011.

[109] B. Elsworth, M. Jones, and M. Blaxter, "Badger-an accessible genome exploration environment," *Bioinformatics*, p. btt466, 2013.

[110] C. Nielsen, *Animal evolution: interrelationships of the living phyla*. Oxford University Press, 2012.

[111] D. Voronov and Y. Panchin, "Cell lineage in marine nematode *Enoplus brevis*," *Development*, vol. 125, no. 1, pp. 143–150, 1998.

[112] E. Schierenberg, "Three sons of fortune: early embryogenesis, evolution and ecology of nematodes," *BioEssays*, vol. 23, no. 9, pp. 841–847, 2001.

[113] R. Small, "The Effects of Predatory Nematodes On Populations of Plant Parasitic Nematodes in Pots," *Nematologica*, vol. 25, no. 1, pp. 94–103, 1979.

[114] Z. Khan and Y. H. Kim, "A review on the role of predatory soil nematodes in the biological control of plant parasitic nematodes," *Applied Soil Ecology*, vol. 35, no. 2, pp. 370–379, 2007.

[115] C. M. de Tomasel, B. J. Adams, F. G. Tomasel, and D. H. Wall, "The Life Cycle of the Antarctic Nematode Plectus murrayi Under Laboratory Conditions," *J. Nematol.*, vol. 45, pp. 39–42, Mar 2013.

[116] Y. Safonova, A. Bankevich, and P. Pevzner, "dipSPAdes: Assembler for highly polymorphic diploid genomes," in *Research in Computational Molecular Biology* (R. Sharan, ed.), vol. 8394 of *Lecture Notes in Computer Science*, pp. 265–279, Springer International Publishing, 2014.

[117] J. Schulze, W. Houthoofd, J. Uenk, S. Vangestel, and E. Schierenberg, "*Plectus* - a stepping stone in embryonic cell lineage evolution of nematodes," *EvoDevo*, vol. 3, no. 1, pp. 1–15, 2012.

[118] M. Casiraghi, O. Bain, R. Guerrero, C. Martin, V. Pocacqua, S. L. Gardner, A. Franceschi, and C. Bandi, "Mapping the presence of *Wolbachia pipientis* on the phylogeny of filarial nematodes: evidence for symbiont loss during evolution.," *Int J Parasitol*, vol. 34, pp. 191–203, Feb 2004.

[119] E. J. L. Soulsby and H. O. Monnig, *Helminths, arthropods, & protozoa of domesticated animals [by] E. J. L. Soulsby.* Williams and Wilkins Co Baltimore, 6th ed., 1968.

[120] S. N. McNulty, J. M. Foster, M. Mitreva, J. C. Dunning Hotopp, J. Martin, K. Fischer, B. Wu, P. J. Davis, S. Kumar, N. W. Brattig, B. E. Slatko, G. J. Weil, and P. U. Fischer, "Endosymbiont DNA in Endobacteria-Free Filarial Nematodes Indicates Ancient Horizontal Genetic Transfer," *PLoS ONE*, vol. 5, p. e11029, 06 2010.

[121] C. Cantacessi, R. B. Gasser, C. Strube, T. Schnieder, A. R. Jex, R. S. Hall, B. E. Campbell, N. D. Young, S. Ranganathan, P. W. Sternberg, and M. Mitreva, "Deep insights into *Dictyocaulus viviparus* transcriptomes provides unique prospects for new drug targets and disease intervention," *Biotechnology Advances*, vol. 29, no. 3, pp. 261–271, 2011.

[122] G. Koutsovoulos, B. Makepeace, V. N. Tanya, and M. Blaxter, "Palaeosymbiosis revealed by genomic fossils of *Wolbachia* in a strongyloidean nematode," *PLoS genetics*, vol. 10, no. 6, p. e1004397, 2014.

[123] M.-A. Flix, "Alternative morphs and plasticity of vulval development in a rhabditid nematode species.," *Dev Genes Evol*, vol. 214, pp. 55–63, Feb 2004.

[124] S. C. Weeks, C. Benvenuto, and S. K. Reed, "When males and hermaphrodites coexist: a review of androdioecy in animals.," *Integr Comp Biol*, vol. 46, pp. 449–464, Aug 2006.

[125] J. Chaudhuri, V. Kache, and A. Pires-daSilva, "Regulation of sexual plasticity in a nematode that produces males, females, and hermaphrodites," *Current Biology*, vol. 21, no. 18, pp. 1548–1551, 2011.

[126] P. H. Schiffer, M. Kroiher, C. Kraus, G. D. Koutsovoulos, S. Kumar, J. I. Camps, N. A. Nsah, D. Stappert, K. Morris, P. Heger, *et al.*, "The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in nematoda," *BMC genomics*, vol. 14, no. 1, p. 923, 2013.

[127] J. Wang, M. Mitreva, M. Berriman, A. Thorne, V. Magrini, G. Koutsovoulos, S. Kumar, M. L. Blaxter, and R. E. Davis, "Silencing of germline-expressed genes by DNA elimination in somatic cells," *Developmental cell*, vol. 23, no. 5, pp. 1072–1080, 2012.

[128] C. Godel, S. Kumar, G. Koutsovoulos, P. Ludin, D. Nilsson, F. Comandatore, N. Wrobel, M. Thompson, C. D. Schmid, S. Goto, *et al.*, "The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets," *The FASEB Journal*, vol. 26, no. 11, pp. 4650–4661, 2012.

[129] J. A. Eisen, "Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionaryanalysis," *Genome Research*, vol. 8, pp. 163–167, Mar 1998.

[130] P. Turner, A. McLennan, A. Bates, M. White, *et al.*, *BIOS Instant Notes in Molecular Biology*. Taylor & Francis, 2007.

[131] R. D. Page and E. C. Holmes, *Molecular evolution: a phylogenetic approach*. John Wiley & Sons, 2009.

[132] R. Nichols, "Gene trees and species trees are not the same," *Trends in Ecology & Evolution*, vol. 16, no. 7, pp. 358–364, 2001.

[133] W.-H. Li and D. Graur, *Fundamentals of molecular evolution*, vol. 48. Sinauer Associates Sunderland, MA, 1991.

[134] B. G. Hall, "Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 792–802, 2005.

[135] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992.

[136] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[137] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[138] F. Sievers and D. G. Higgins, "Clustal Omega, accurate alignment of very large numbers of sequences," in *Multiple Sequence Alignment Methods*, pp. 105–116, Springer, 2014.

[139] J. D. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs," *Nucleic acids research*, vol. 27, no. 13, pp. 2682–2690, 1999.

[140] N. Lartillot and H. Philippe, "Improvement of molecular phylogenetic inference and the phylogeny of Bilateria," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1496, pp. 1463–1472, 2008.

[141] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular biology and evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[142] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, 2009.

[143] Z. Yang and B. Rannala, "Molecular phylogenetics: principles and practice," *Nature Reviews Genetics*, vol. 13, no. 5, pp. 303–314, 2012.

[144] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *Journal of molecular evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[145] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014.

[146] D. Zwickl, *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, The University of Texas, 2006.

[147] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, pp. 783–791, 1985.

[148] D. M. Hillis and J. J. Bull, "An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis," *Systematic biology*, vol. 42, no. 2, pp. 182–192, 1993.

[149] B. Larget and D. L. Simon, "Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees," *Molecular Biology and Evolution*, vol. 16, pp. 750–759, 1999.

[150] Y. Suzuki, G. V. Glazko, and M. Nei, "Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics," *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 16138–16143, 2002.

[151] Z. Yang and B. Rannala, "Branch-length prior influences Bayesian posterior probability of phylogeny," *Systematic Biology*, vol. 54, no. 3, pp. 455–470, 2005.

[152] J. P. Huelsenbeck and B. Rannala, "Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models," *Systematic Biology*, vol. 53, no. 6, pp. 904–913, 2004.

[153] M. O. Dayhoff and R. M. Schwartz, "A model of evolutionary change in proteins," in *In Atlas of protein sequence and structure*, Citeseer, 1978.

[154] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Computer applications in the biosciences: CABIOS*, vol. 8, no. 3, pp. 275–282, 1992.

[155] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Molecular biology and evolution*, vol. 18, no. 5, pp. 691–699, 2001.

[156] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, *et al.*, "Pfam: the protein families database," *Nucleic acids research*, p. gkt1223, 2013.

[157] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix," *Molecular biology and evolution*, vol. 25, no. 7, pp. 1307–1320, 2008.

[158] S. Tavaré, "Some probabilistic and statistical problems in the analysis of dna sequences," *Lectures on mathematics in the life sciences*, vol. 17, pp. 57–86, 1986.

[159] N. Lartillot and H. Philippe, "A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Molecular biology and evolution*, vol. 21, no. 6, pp. 1095–1109, 2004.

[160] L. S. Quang, O. Gascuel, and N. Lartillot, "Empirical profile mixture models for phylogenetic reconstruction," *Bioinformatics*, vol. 24, no. 20, pp. 2317–2323, 2008.

[161] M. Hasegawa, H. Kishino, and T.-a. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial dna," *Journal of molecular evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[162] Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods," *Journal of Molecular evolution*, vol. 39, no. 3, pp. 306–314, 1994.

[163] A. Stamatakis, "Phylogenetic models of rate heterogeneity: a high performance computing perspective," in *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*, pp. 8–pp, IEEE, 2006.

[164] L. Li, C. J. Stoeckert, and D. S. Roos, "OrthoMCL: identification of ortholog groups for eukaryotic genomes," *Genome research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[165] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, *et al.*, "The genome sequence of *Drosophila melanogaster*," *Science*, vol. 287, no. 5461, pp. 2185–2195, 2000.

[166] Q. Xia, Z. Zhou, C. Lu, D. Cheng, F. Dai, B. Li, P. Zhao, X. Zha, T. Cheng, C. Chai, *et al.*, "A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*)," *Science*, vol. 306, no. 5703, pp. 1937–1940, 2004.

[167] M. Grbić, T. Van Leeuwen, R. M. Clark, S. Rombauts, P. Rouzé, V. Grbić, E. J. Osborne, W. Dermauw, P. C. T. Ngoc, F. Ortego, *et al.*, "The genome of *Tetranychus urticae* reveals herbivorous pest adaptations," *Nature*, vol. 479, no. 7374, pp. 487–492, 2011.

[168] M. Mitreva, D. P. Jasmer, D. S. Zarlenga, Z. Wang, S. Abubucker, J. Martin, C. M. Taylor, Y. Yin, L. Fulton, P. Minx, *et al.*, "The draft genome of the parasitic nematode *Trichinella spiralis*," *Nature genetics*, vol. 43, no. 3, pp. 228–235, 2011.

[169] A. R. Jex, P. Nejsum, E. M. Schwarz, L. Hu, N. D. Young, R. S. Hall, P. K. Korhonen, S. Liao, S. Thamsborg, J. Xia, *et al.*, "Genome and transcriptome of the porcine whipworm *Trichuris suis*," *Nature genetics*, vol. 46, no. 7, pp. 701–706, 2014.

[170] E. Ghedin, S. Wang, D. Spiro, E. Caler, Q. Zhao, J. Crabtree, J. E. Allen, A. L. Delcher, D. B. Guiliano, D. Miranda-Saavedra, *et al.*, "Draft genome of the filarial nematode parasite *Brugia malayi*," *Science*, vol. 317, no. 5845, pp. 1756–1760, 2007.

[171] A. C. Darby, S. D. Armstrong, G. S. Bah, G. Kaur, M. A. Hughes, S. M. Kay, P. Koldkjær, L. Rainbow, A. D. Radford, M. L. Blaxter, *et al.*, "Analysis of gene expression from the *Wolbachia* genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis," *Genome research*, vol. 22, no. 12, pp. 2467–2477, 2012.

[172] F. Comandatore, D. Sassera, M. Montagna, S. Kumar, G. Koutsovoulos, G. Thomas, C. Repton, S. A. Babayan, N. Gray, R. Cordaux, *et al.*, "Phylogenomics and analysis of shared genes suggest a single transition to mutualism in *Wolbachia* of nematodes," *Genome biology and evolution*, vol. 5, no. 9, pp. 1668–1674, 2013.

[173] T. Kikuchi, J. A. Cotton, J. J. Dalzell, K. Hasegawa, N. Kanzaki, P. McVeigh, T. Takanashi, I. J. Tsai, S. A. Assefa, P. J. Cock, *et al.*, "Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*," *PLoS pathogens*, vol. 7, no. 9, p. e1002219, 2011.

[174] J. A. Cotton, C. J. Lilley, L. M. Jones, T. Kikuchi, A. J. Reid, P. Thorpe, I. J. Tsai, H. Beasley, V. Blok, P. J. Cock, *et al.*, "The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode," *Genome biology*, vol. 15, no. 3, p. R43, 2014.

[175] D. H. Lunt, S. Kumar, G. Koutsovoulos, and M. L. Blaxter, "The complex hybrid origins of the root knot nematodes revealed through comparative genomics," *PeerJ*, vol. 2, p. e356, 2014.

[176] C. H. Opperman, D. M. Bird, V. M. Williamson, D. S. Rokhsar, M. Burke, J. Cohn, J. Cromer, S. Diener, J. Gajan, S. Graham, *et al.*, "Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism," *Proceedings of the National Academy of Sciences*, vol. 105, no. 39, pp. 14802–14807, 2008.

[177] P. Abad, J. Gouzy, J.-M. Aury, P. Castagnone-Sereno, E. G. Danchin, E. Deleury, L. Perfus-Barbeoch, V. Anthouard, F. Artiguenave, V. C. Blok, *et al.*, "Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*," *Nature biotechnology*, vol. 26, no. 8, pp. 909–915, 2008.

[178] J. Srinivasan, A. R. Dillman, M. G. Macchietto, L. Heikkinen, M. Lakso, K. M. Fracchia, I. Antoshechkin, A. Mortazavi, G. Wong, and P. W. Sternberg, "The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle," *Genetics*, vol. 193, no. 4, pp. 1279–1295, 2013.

[179] R. Laing, T. Kikuchi, A. Martinelli, I. J. Tsai, R. N. Beech, E. Redman, N. Holroyd, D. J. Bartley, H. Beasley, C. Britton, *et al.*, "The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery," *Genome Biol*, vol. 14, p. R88, 2013.

[180] X. Bai, B. J. Adams, T. A. Ciche, S. Clifton, R. Gaugler, K.-s. Kim, J. Spieth, P. W. Sternberg, R. K. Wilson, and P. S. Grewal, "A lover and a fighter: the genome sequence of an entomopathogenic nematode *Heterorhabditis bacteriophora*," *PLoS ONE*, vol. 8, no. 7, p. e69618, 2013.

[181] Y. T. Tang, X. Gao, B. A. Rosa, S. Abubucker, K. Hallsworth-Pepin, J. Martin, R. Tyagi, E. Heizer, X. Zhang, V. Bhonagiri-Palsikar, *et al.*, "Genome of the human hookworm *Necator americanus*," *Nature genetics*, vol. 46, no. 3, pp. 261–269, 2014.

[182] C. Dieterich, S. W. Clifton, L. N. Schuster, A. Chinwalla, K. Delehaunty, I. Dinkelacker, L. Fulton, R. Fulton, J. Godfrey, P. Minx, *et al.*, "The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism," *Nature genetics*, vol. 40, no. 10, pp. 1193–1198, 2008.

[183] C. Rödelsperger, R. A. Neher, A. M. Weller, G. Eberhardt, H. Witte, W. E. Mayer, C. Dieterich, and R. J. Sommer, "Characterization of genetic diversity in the nematode *Pristionchus pacificus* from population-scale resequencing data," *Genetics*, vol. 196, no. 4, pp. 1153–1165, 2014.

[184] S. E. S. Pierre, L. Ponting, R. Stefancsik, P. McQuilton, *et al.*, "Flybase 102advanced approaches to interrogating FlyBase," *Nucleic acids research*, vol. 42, no. D1, pp. D780–D788, 2014.

[185] J. Wang, Q. Xia, X. He, M. Dai, J. Ruan, J. Chen, G. Yu, H. Yuan, Y. Hu, R. Li, *et al.*, "SilkDB: a knowledgebase for silkworm biology and genomics," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D399–D402, 2005.

[186] P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, *et al.*, "Ensembl 2014," *Nucleic acids research*, p. gkt1196, 2013.

[187] E. Heitlinger, S. Bridgett, A. Montazam, H. Taraschewski, and M. Blaxter, "The transcriptome of the invasive eel swimbladder nematode parasite *Anguillicola crassus*," *BMC genomics*, vol. 14, no. 1, p. 87, 2013.

[188] N. Karim, J. T. Jones, H. Okada, and T. Kikuchi, "Analysis of expressed sequence tags and identification of genes encoding cell-wall-degrading enzymes from the fungivorous nematode *Aphelenchus avenae*," *BMC genomics*, vol. 10, no. 1, p. 525, 2009.

[189] K. Cwiklinski, J. Merga, S. L. Lake, C. Hartley, J. B. Matthews, S. Paterson, and J. E. Hodgkinson, "Transcriptome analysis of a parasitic Clade V nematode: Comparative analysis of potential molecular anthelmintic targets in *Cylicostephanus goldi*," *International journal for parasitology*, vol. 43, no. 11, pp. 917–927, 2013.

[190] Y. M. Harcus, J. Parkinson, C. Fernández, J. Daub, M. E. Selkirk, M. L. Blaxter, and R. M. Maizels, "Signal sequence analysis of expressed sequence tags from the nematode *Nippostrongylus brasiliensis* and the evolution of secreted proteins in parasites," *Genome Biology*, vol. 5, no. 6, p. R39, 2004.

[191] J. Parkinson, D. B. Guiliano, and M. Blaxter, "Making sense of EST sequences by CLOBBing them," *Bmc Bioinformatics*, vol. 3, no. 1, p. 31, 2002.

[192] B. Elsworth, J. Wasmuth, and M. Blaxter, "NEMBASE4: the nematode transcriptome resource," *International journal for parasitology*, vol. 41, no. 8, pp. 881–894, 2011.

[193] M. Jones and M. Blaxter, "afterParty: turning raw transcriptomes into permanent resources," *BMC bioinformatics*, vol. 14, no. 1, p. 301, 2013.

[194] S. M. Van Dongen, *Graph clustering by flow simulation.* PhD thesis, Utrecht University, 2000.

[195] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada, "Prottest 3: fast selection of best-fit models of protein evolution," *Bioinformatics*, vol. 27, no. 8, pp. 1164–1165, 2011.

[196] FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

[197] A. J. Aberer, K. Kobert, and A. Stamatakis, "Exabayes: Massively parallel Bayesian tree inference for the whole-genome era," *Molecular biology and evolution*, p. msu236, 2014.

[198] N. Lartillot, T. Lepage, and S. Blanquart, "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating," *Bioinformatics*, vol. 25, no. 17, pp. 2286–2288, 2009.

[199] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2–approximately maximum-likelihood trees for large alignments," *PloS one*, vol. 5, no. 3, p. e9490, 2010.

[200] A. Rambaut and A. J. Drummond, "Tracer v1.4." http://beast.bio.ed.ac.uk/Tracer, 2007.

[201] L. I. Campbell, O. Rota-Stabelli, G. D. Edgecombe, T. Marchioro, S. J. Longhorn, M. J. Telford, H. Philippe, L. Rebecchi, K. J. Peterson, and D. Pisani, "MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda," *Proceedings of the National Academy of Sciences*, vol. 108, no. 38, pp. 15920–15924, 2011.

[202] I. N. Filipjev *et al.*, "The classification of the free-living nematodes and their relation to parasitic nematodes.," *Smithsonian Miscellaneous Collection.*, vol. 89, pp. 1–63, 1934.

[203] M. Pagel, A. Meade, and D. Barker, "Bayesian estimation of ancestral character states on phylogenies," *Systematic biology*, vol. 53, no. 5, pp. 673–684, 2004.

[204] D. R. Laetsch, E. G. Heitlinger, H. Taraschewski, S. A. Nadler, and M. L. Blaxter, "The phylogenetics of Anguillicolidae (Nematoda: Anguillicoloidea), swimbladder parasites of eels," *BMC evolutionary biology*, vol. 12, no. 1, p. 60, 2012.

[205] A. J. Shannon, T. Tyson, I. Dix, J. Boyd, and A. M. Burnell, "Systemic RNAi mediated gene silencing in the anhydrobiotic nematode *Panagrolaimus superbus*," *BMC molecular biology*, vol. 9, no. 1, p. 58, 2008.

[206] M. D. Hendy and D. Penny, "A framework for the quantitative study of evolutionary trees," *Systematic Biology*, vol. 38, no. 4, pp. 297–309, 1989.

[207] T. A. Heath, S. M. Hedtke, and D. M. Hillis, "Taxon sampling and the accuracy of phylogenetic analyses," *J Syst Evol*, vol. 46, no. 3, pp. 239–257, 2008.

[208] S. Kumar, G. Koutsovoulos, G. Kaur, and M. Blaxter, "Toward 959 nematode genomes," in *Worm*, vol. 1, pp. 42–50, Landes Bioscience, 2012.

[209] M. Blaxter, G. Koutsovoulos, M. Jones, S. Kumar, and B. Elsworth, *Next-Generation Systematics*, ch. Phylogenomics of Nematoda. Cambridge University Press, Cambridge, UK, 2014.

[210] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular biology and evolution*, vol. 16, pp. 1114–1116, 1999.