THE UNIVERSITY
of EDINBURGH

On the evolution of effector gene families in potato cyst nematodes

Dominik R. Laetsch

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy

Institute of Evolutionary Biology

School of Biological Sciences

University of Edinburgh

2017

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.

Dominik R. Laetsch

August 2017

Abstract

Potato cyst nematodes (PCN) are economically relevant plant parasites that infect potato crops. The genomes of three PCN species are available and genome data have been generated for several populations of PCN, to address questions related to the molecular basis of plant parasitism.

In this thesis, I employ approaches of comparative genomics to highlight differences and similarities between PCNs and other nematode species. I present two new software solutions to address challenges associated with the field of comparative genomics: BlobTools, a taxonomic interrogation toolkit for quality control of genome assemblies, and KinFin, a solution for the analysis of protein orthology data. I apply both software solutions to genomic datasets of nematodes, platyhelminths, and tardigrades. Based on KinFin analysis of plant parasitic nematodes, I identify protein families in PCNs likely to be involved in host-parasitic interaction, termed effectors, and discuss their functions. I highlight examples of horizontal gene transfer from bacteria to plant parasitic nematodes. Through genomic data of European and South American populations of PCNs, I address variation in populations, infer phylogenetic relationships, and try to estimate the effect of selection on effector genes identified through KinFin. Furthermore, I estimate the rate of variation across the reference genomes of two PCNs.

Lay summary

Potato cyst nematodes (PCNs) are small worms that infect potato crops and cause reduction in crop yield in the UK and across the world. As cysts they can persist in the soil for many years, even in the absence of a suitable host, which limits the success of control and eradication measures. They use small molecules, proteins which are encoded by genes in their genomes, to establish infection and to prevent the immune system of the host from detecting them. Understanding how these proteins differ from the proteins of other worms, for example those that do not infect potatoes, could inform detection and controls measures. This PhD thesis aims to further our understanding of the proteins, and the underlying genes, of PCNs by comparing them between different populations of PCNs and to those of other worms, both free-living and parasitic. During this thesis I developed two software solutions. One is aimed at identifying contamination in genomic data, which is a common phenomenon when working with organisms that live in the soil. The other allows comparison of sets of proteins between organisms, which I use to identify both evolutionary conserved and species-specific protein sets in PCNs, focussed on those involved in the parasitic interaction with the host. Finally, I investigate variation in populations of PCNs based on genomic data, identify patterns in the variation, and formulate hypotheses regarding the population structures of PCNs. These results can now be explored further by other researchers using the methodological innovations I developed.

Acknowledgements

I would like to thank my supervisors Mark Blaxter, Vivian Blok, Peter Cock, and Graham Stone. Mark Blaxter has been an constant source of inspiration, knowledge, and support. Working with him has made me both a better scientist and, more importantly, a better person. He made me aware of the beauty of nature, language, and art and I will be forever grateful for the opportunity to work with him. I thank Vivian Blok and Peter Cock for their support and for giving me the freedom to develop my own scientific interests. Vivian Blok has taught me the importance of being realistic in my expectations and that biology is often more complicated than one imagines. Peter Cock has inspired me to become a better programmer and has shown me the importance of form.

I thank my current employer, Laura Ross, for her endless patience and trust, especially during the last months of my thesis. Thanks also go to my postgraduate committee member, Ally Phillimore, for his advise and comments over the years.

Past and present members of the Blaxter Lab have turned the past years into an unique and joyful experience. By working and living with Georgios Koutsovoulos, my believes, opinions, and thoughts have been challenged on a daily basis. I will be forever grateful for being his friend. Carlos Caurcel has kept me sane and I feel honoured to count him among my friends. I thank Richard Challis, Sujai Kumar,

Reuben Nowell, Laura Salazar, Elisabet Sjokvist and Lewis Stevens for creating the best office environment I could possibly imagine. I would also like to thank Martin Jones for teaching me how to program and for showing me the importance of passing on knowledge in a clear and concise way.

I would like to thank the people at the Institute of Evolutionary Biology and the James Hutton Institute for passing on their knowledge over coffee and tea, especially Jack Hearn, Daniel Barker, Amy Buck, Alex Twyford, Darren Obbard, Leighton Pritchard, and John Jones.

I am grateful to the University of Edinburgh and the James Hutton Institute for funding my PhD programme and providing funds for travel and research.

I would also like to thank my friends Christian Reyes-Knoche and Daniel Weickgenannt who have not heard from me in months and probably wonder whether I am still alive. I will work hard for making up for the past years.

Finally, I would like to thank my parents, Mira and Wolfgang, and my grandmother, Anna, for all the love and support they have given me. Not one day goes by where I do not miss being close to them. My sanity would have been lost years ago were it not for them.

Contents

List of Tables

List of Figures

List of Abbreviations

b	base(s)
CEGMA	Core Eukaryotic Genes Mapping Approach
ENA	European Nucleotide Archive
GFF3	General Feature Format
HSP	high-scoring segment pair
kb	kilobase(s)
Mb	megabase(s)
MCL	Markov Clustering
MSA	multiple sequence alignment
NCBI	National Centre for Biotechnology Information
BED	Browser Extensible Data
PCN	potato cyst nematode
PE	paired end
MP	mate pair
RBBH	reciprocal-best-BLAST-hit
RFA	representative functional annotation
RKN	root-knot nematode
SD	standard deviation
SE	single end
SNP	single nucleotide polymorphism
MNP	multiple nucleotide polymorphism
SRA	Short Read Archive
VCF	Variant Call Format
WGA	whole genome amplification
WGS	whole genome sequencing
WTSI	Wellcome Trust Sanger Institute

Chapter 1

Introduction

"Io ritornai da la santissima onda
rifatto sì come piante novelle
rinovellate di novella fronda,
puro e disposto a salire a le stelle."

- Dante Aligheri, *La Comedia, Purgatorio Canto XXXIII*

1.1 Thesis structure

In this thesis, I present analyses of genomic data of economically important parasites of potato crops, potato cyst nematodes (PCNs) of the genus *Globodera*.

I developed two software solutions in order to overcome common challenges associated with genome sequencing data of non-model organisms. I illustrate their functionality through use cases and apply them to genomic data of PCNs in order

to study evolutionary patterns of gene/protein families involved in host-parasite interactions. These proteins are termed effectors.

In this Chapter, I outline the challenges associated with genome sequencing data of non-model organisms and describe the biology of the organisms on which this thesis is focussed.

In Chapter 2, I present BlobTools, a modular toolkit for the taxonomic interrogation and visualisation of genome assemblies for the purpose of quality control. The software was developed to address the issue of contamination in genomic datasets and its functionality is illustrated based on three use cases.

In Chapter 3, I discuss available methods for the analysis of gene/protein families and present KinFin, a software solution for taxon aware analysis of clustered protein data. I explain how I formalised the problem of analysis of protein clustering data and illustrate the KinFin workflow based on four use cases.

In Chapter 4, I describe the biology of PCNs and apply KinFin to a protein clustering dataset of Clade IV nematodes *sensu* Blaxter et al., 1998, including two PCNs: *G. pallida* and *G. rostochiensis*. I explore parameter space of the protein clustering approach and analyse effector gene families in PCNs from an evolutionary perspective.

In Chapter 5, I compare the published genome assemblies of PCNs and highlight differences and similarities. Through the use of genomic data for different species and populations, I explore patterns of variation within the genomes. I estimate rates of variation for the reference genomes of *G. pallida* and *G. rostochiensis* and investigate phylogeographic patterns of populations of *G. pallida*.

In Chapter 6, I summarise the main findings and present thoughts on future analysis inspired by the results described in this thesis.

1.2 Genomics of non-model organisms

The genomics revolutions was pioneered by model organisms such as 'the worm' (*Caenorhabditis elegans*) and 'the fly' (*Drosophila melanogaster*). Recently, decreasing costs of sequencing technologies have democratised access to these approaches, enabling the study of organisms sampled from wild populations. Analysis of the resulting data is, however, non-trivial since few solutions exists for addressing the challenges associated with these samples.

1.2.1 Low-complexity metagenomes

Advances in next generation sequencing technologies have generated vast amounts of data and knowledge (Goodwin, McPherson, and McCombie, 2016). The decrease in cost per nucleotide lead to an increased application of these technologies to non-model organisms, life forms which have so far not been intensively studied by the research community. Genome-enabled science on these species can then illuminate novel processes and reveal the patterns of evolution. For non-model species, the luxury of large amounts of material from cultured isolates is often not possible, and research must progress from organisms sourced from the wild or from complex mixtures of species. DNA extracted from a sample may therefore contain genomes from multiple organisms — food sources, host material, symbionts, pathogens, commensals and external contaminants — in addition to the target organism. In some cases, the associated genomes can be considered 'contaminants', while in others, they can provide insights into the biology of the target organism. In all cases they should be identified, isolated, and investigated with care.

Hence, genome datasets should be viewed as low-complexity metagenomes until a assessment of the taxonomic composition has been made. Several solutions

for taxonomic screening of genomic datasets exist and are discussed in Chapter 2. However, as none of these met the needs of the datasets I intended to analyse for this thesis I developed BlobTools, a modular toolkit for taxonomic interrogation of genome assemblies. The software is based on ideas from Kumar et al. (2013), which I implemented and expanded upon.

1.2.2 Definition of gene families

The field of comparative genomics is concerned with the study of similarities and differences between the information encoded in the genomes of organisms. Fitch (1970) classified homologous sequences into two groups: orthologues, where homology is a result of speciation events, and paralogues, where homology is a consequence of gene duplication.

A standard approach in comparative genomics, often referred to as the 'orthologue conjecture', is based on the assumption that between orthologues functional conservation is more likely than between paralogues. This is because gene duplication events are viewed as an important source of functional innovation (Ohno, 1970). Several studies have been aimed at testing the 'orthologue conjecture' (Nehrt et al., 2011; Altenhoff et al., 2012; Chen and Zhang, 2012) and obtained mixed results, which has led to international collaborations such as the 'Quest of Orthologs' project, targeted at benchmarking orthology-inference methods with standardised datasets (Altenhoff et al., 2016).

Exploitation of orthologue definitions across species, the study of gene family evolution, genome evolution, species phylogenetics, and as loci for population genetics and ecological genetics, is demanding. Many research projects aim to identify orthologues of interest that have a specific distribution across species,

for example identifying gene families that are synapomorphic for — or that have been specifically lost from — a particular clade. Exploring the effects of assuming different underlying phylogenies on the analysis of the origins of orthologues may assist in discriminating between competing hypotheses. Grouping species by non-phylogenetic classifiers — such as habitat, mating system or life history — may also identify protein families uniquely present/absent or exhibiting differential copy-number.

While orthologues and paralogues are readily distinguished by phylogenetic analysis, such approaches are too computationally expensive for the identification of 'clusters' of homologous sequences — containing both orthologues and paralogues — across many taxa. Hence, the standard approach is to rely on sequence similarity searches and subsequent post-processing of the results to identify putative homologues. The simplest form of orthology inference through sequence similarity searches is referred to as reciprocal-best-BLAST-hit (RBBH) approach (Bork et al., 1998; Tatusov, Koonin, and Lipman, 1997), where two sequences originating from different genomes are considered orthologous if they are recovered as each other's best hit in sequence similarity searches. While RBBH analysis has been found to be robust compared to other orthology inference methods (Salichos and Rokas, 2011; Altenhoff and Dessimoz, 2009; Hulsen et al., 2006) and is effective at recovering orthologous group seeds for further analysis (Dalquen and Dessimoz, 2013), it has certain limitations. RBBH analysis *sensu stricto* is only capable of identifying 1-to-1 orthology, and therefore suffers from high false negative rates if paralogues, uncollapsed allelic copies (arising from the genome assembly process if loci are sufficiently diverged), or very similar sequences in the case of custom sequence collections are present in the set of query or subject sequences. Another problem is the issue of transitivity (Johnson, 2007), a property of orthologues which implies that, if the proteins 'A' and 'B' and 'B' and 'C' are orthologues, it follows that the proteins

'A' and 'C' are also orthologues. This assumption is often not fulfilled in the case of RBBH analysis.

Candidate orthologues and paralogues between taxa are commonly identified through clustering of protein sequence data using tools such as OrthoFinder (Emms and Kelly, 2015), OrthoMCL (Li, Stoeckert, and Roos, 2003), and others. OrthoFinder can be viewed as an improvement on the OrthoMCL pipeline, as it is more user friendly — as each step in the pipeline can be run independently — and accounts for gene length bias and phylogenetic distance between proteomes. Both clustering pipelines construct graphs from the results of 'all vs. all' sequence similarity searches, where proteins are represented as nodes and edges between them are weighted by the search results. These graphs are then processed using the MCL (Markov Clustering) algorithm (Van Dongen, 2001). The MCL algorithm is used to deconstruct these often highly connected graphs based on random walks between nodes in the graph. This is based on the assumption that densely connected regions in a graph will be visited more frequently than sparsely connected regions. The random walk is a Markov process which assumes independence of past states and transitions between states based on a probability distribution. One of the parameters controlling the MCL algorithm is the MCL inflation value, which affects the 'granularity' of the resulting clusters: lower inflation values result in fewer clusters containing more members, while higher inflation values lead to more clusters containing fewer members. The implications of this parameter for the identification of protein families is that at lower inflation values protein families might be erroneously clustered, while at higher inflation values genuine protein families might be split into several clusters. In the original OrthoMCL paper (Li, Stoeckert, and Roos, 2003), the authors evaluated the influence of the parameter (ranging from 1.1 to 4.0) on the clustering of seven proteomes (*Arabidopsis thaliana*, *C. elegans*, *D. melanogaster*, *Homo sapiens*, *Plasmodium falciparum*, *Sacharomyces cerevisiae*, and

Escherichia coli) based on the consistency of enzyme commission (EC) numbers associated with sequences in the resulting clusters. For clusters containing two or more sequences annotated with EC numbers they calculated the EC consistency. EC consistency varied from 80% to 88% with increasing inflation value (with a pronounced difference between 1.1 and 1.5), based on which they concluded that an inflation value of 1.5 (EC consistency of 86%) balances sensitivity and specificity. The inflation value of 1.5 has become the standard for clustering analysis using the MCL algorithm.

While established pipelines for the inference of orthology exist, only a few solutions are available for the comparative analysis of their output. Most solutions are either geared towards taxonomically restricted groups of organisms or require substantial effort to implement on local computing infrastructure. In order to carry out the analysis of effector protein families in *Globodera* species, I developed KinFin which is a software solution for taxon-aware analysis of protein clustering data. The software is described in detail in Chapter 3 and applied to protein data of *Globodera* species and other nematodes of Clade IV in Chapter 4.

1.3 Plant parasitism within the phylum Nematoda

Nematoda is a phylum of vermiform, ecdysozoan animals (Dunn et al., 2008). To date over 23,000 species have been described (Hallan, 2008) and the estimated diversity ranges from 100,000 to 100,000,000 species (Lambshead, 1993). Members of this phylum display a wide range of trophic behaviours, ranging from free-living microbivors to obligate parasites of multicellular eukaryotes.

Research in the field of nematology is traditionally focussed on parasitic nematodes affecting human health. Apart from the devastating and direct effects of animal parasitic nematodes, plant parasitic nematodes (PPNs) constitute an indirect burden on human health by substantially decreasing crop yield and contributing to famine in developing nations. The estimated annual cost of PPNs on human agriculture exceeds £58 billion (Nicol et al., 2011). Over 4,100 PPN species have been described (Decraemer et al., 2006) and several are considered a significant burden on global food safety due to their role as pathogens and vectors of plant viruses.

The term plant parasitism describes a broad array of feeding modes ranging from free-living predatory behaviour over migratory ecto-/endoparasitism to facultative/obligate sedentary endoparasitism (Baldwin, Nadler, and Adams, 2004). Plant parasitism within the phylum Nematoda is estimated to have arisen independently at least four times within three of the five phylogenetic clades *sensu* Blaxter et al., 1998, namely Clade I (order Dorylaimida), Clade II (order Triplonchida) and Clade IV (order Tylenchida) (Blaxter et al., 1998; Blaxter and Koutsovoulos, 2015). Tylenchida includes the majority of economically relevant PPNs such as root-knot nematodes (genus *Meloidogyne*) and cyst nematodes (genera *Heterodera* and *Globodera*) (Jones et al., 2013). It is noteworthy that all nematodes participating in plant parasitic interactions have evolved within clades which include free-living nematodes (Blaxter et al., 1998; Megen et al., 2009).

The traditional hypothesis for the emergence of plant-parasitism within Tylenchida states that this feeding mode has evolved gradually from fungal-feeding over facultative parasitism of peripheral plant tissue into more complex interactions; eventually culminating in the development of sedentary endoparasitism (Luc et al., 1987). This hypothesis is partially supported by feeding type analyses, which

suggest a gradual evolution from simple forms of ectoparasitism and migratory endoparasitism towards complex forms of sedentary endoparasitism (Bert et al., 2008; Holterman et al., 2009). However, depending on the reconstruction method (unordered parsimony, step-matrix based parsimony and likelihood based approaches), the ancestral feeding mode of Tylenchids ranges from fungal-feeding over predatory plant-feeding to bacteriovore-feeding and is currently not resolved (Bert, Karssen, and Helder, 2011). Adaptations to plant parasitism by nematodes include the development of the stylet, a protrusible hollow mouth spear located at the anterior end of PPNs, and specialised secretory gland cells within their oesophagus (Hussey, 1989; Baldwin, Nadler, and Adams, 2004). The stylet serves both as an instrument for the penetration of host cell walls and as a structure for the delivery of effector gene products expressed in the oesophageal gland cells (Mitchum et al., 2013).

1.3.1 Comparative genomics of plant parasitic nematodes

In 1998 the genome sequence of the free-living nematode *Caenorhabditis elegans* — the first animal genome to be sequenced — was published (*C. elegans* Sequencing Consortium, 1998). Ten years later the genomes of the first plant-parasitic nematodes, the root-knot nematodes *Meloidogyne incognita* (Abad et al., 2008) and *Meloidogyne hapla* (Opperman et al., 2008) were published. More recently three other tylenchid genomes from the pine-wood nematode *Bursaphelenchus xylophilus* (Kikuchi et al., 2011), the pale potato cyst nematode *Globodera pallida* (Cotton et al., 2014) and the peach root-knot nematode *Meloidogyne floridensis* (Lunt et al., 2014) have been published. During the duration of my PhD project, two additional *Globodera* genome assemblies were published which are discussed in detail in Chapter 5. However, it should be noted, that all genome projects published to

date concern tylenchid plant parasites, while dorylaimid and triplonchid parasites are currently neglected.

The available PPN genomes have been screened for common genomic features underpinning the plant-parasitic lifestyle (Bird et al., 2015; Zarowiecki and Berriman, 2015). A common theme in PPN genomes are genes coding for cell wall modifying proteins which have been obtained via horizontal gene transfer (HGT) from bacterial and fungal donors (Danchin et al., 2010). However, the vast majority of effector genes associated with parasitism in a given PPN species are often poorly conserved between species and appear to be synapomorphies of different lineages (Bird et al., 2015; Kikuchi et al., 2011; Cotton et al., 2014). Both aspects of PPN genomes — acquisition of metabolic genes through HGT and lineage-specificity of effector genes — have also been observed in plant-parasitic fungi (Oliva et al., 2010) and oomycetes (Judelson, 2012), suggesting that these features might be general traits within genomes of plant parasites (Bird et al., 2015).

1.3.2 Nematode effector proteins

In order to minimise the impact of PPNs on agricultural crop production, it is crucial to understand how these parasites establish and maintain infection within their hosts. Gene products involved in the interaction between PPNs and their hosts are often referred to as 'effectors'. Hogenhout et al. (2009) defined effectors as 'all pathogen proteins and small molecules that alter host-cell structure and function'. Since the comparative genomic analyses carried out in Chapter 4 are guided by previously published effector sequences (see Section 4.2.1), the definition of the term — within the scope of this thesis — can therefore be broadened to include 'all gene products previously described as effectors'. Candidate effectors in PPNs are often identified among excretory/secretory proteins of the parasite, since export into the host is a necessary requirement. Expression of the underlying genes occurs primarily in three specialised secretory gland cells, one dorsal and two subventral. Effector proteins are then secreted through the stylet opening as indicated by increased plant immune response around this structure in susceptible hosts (Jones, 1981; Williamson and Kumar, 2006). Effector proteins are synthesised in the cell body of the gland cells and N-terminal signal peptides facilitate their transport through the secretory pathway after which they are packaged into secretory vesicles, released from the cell through exocytosis and eventually injected into the host (Mitchum et al., 2013). A high number of secreted nematode proteins have been identified through bioinformatic mining of transcriptome and genome data from PPNs (Rosso and Grenier, 2011; Mitchum et al., 2013; Kikuchi, Eves-van den Akker, and Jones, 2017; and references therein). The functional diversity of these effectors can be divided into three main groups: cell wall modifying enzymes, gene products altering plant development and plant defence suppressing effectors.

Cell wall modifying effectors

The plant cell wall, composed of a variety of oligo- and polysaccharides, acts as a physical barrier for PPNs which has to be overcome in order allow migration through and feeding on the host. Cell wall modifying effectors (CWMs) comprise the largest group described for PPNs and include several families of cellulases, xylanases, polygalacturonases, pectate lyases, invertases, arabinases and expansin-like proteins. This is largely due to the fact that these proteins contain well defined domains and are easily identified in metazoan genomes since they are absent in non-plant-parasitic organisms. The majority of these genes have highest identity to bacterial and fungal genes, suggesting acquisition through HGT (Danchin et al., 2010; Haegeman, Jones, and Danchin, 2011).

Plant defence manipulating effectors

Plants have evolved multiple layers of defence to sense and resist infections including effector-triggered immunity, a form of innate immune response which elicits apoptosis in infected cells. One nematode effector that has been shown to directly target the immune response in plants is SPRYSEC-19 in *G. rostochiensis*. This effector interacts with a nucleotide-binding-leucine-rich repeat (NB-LRR) protein in the host without triggering programmed cell death (Postma et al., 2012). Another example is the *G. pallida Gp-RBP-1* SPRYSEC effector which elicits Gpa2- and RanGAP2-dependent plant cell death (Sacco et al., 2009). Both effectors are members of a family of SPRY domain proteins which are found in all nematode genomes but are typically not secreted. However, in *G. pallida* this gene family experienced an enormous expansion (299 *G. pallida* proteins are predicted to have one or more SPRY domains) (Cotton et al., 2014). Akin to CWEs, some plant defence manipulating effectors show signatures of horizontal gene transfer. One example is a group of

secreted chorismate mutases which are found in both root-knot and cyst nematodes (Lambert, Allen, and Sussex, 1999; Bekal, Niblack, and Lambert, 2003). In the host these enzymes are part of the shikimate pathway. The presence of underlying genes in the parasite genomes suggests at a role during establishment of the feeding site through interference with salicylic acid production and defence signalling (Jones et al., 2003). Chorismate mutase genes are usually absent from metazoan genomes, which makes them strong candidates for horizontal gene transfer (Jones, Furlanetto, and Kikuchi, 2005). Other secreted proteins such as superoxide dismutases and glutathione peroxidases are thought to neutralize host defences involving reactive oxygen species and anti-microbial molecules (Bellafiore et al., 2008; Dubreuil et al., 2007) and the underlying gene families appear to be expanded in some nematode species (Cotton et al., 2014).

Plant development altering effectors

Sedentary PPNs cause profound changes to host cell structure and physiology during establishment and maintenance of the feeding site. One example of this group in cyst nematodes are small, secreted proteins with high identity to CLAVATA3/ESR-related (CLE) signalling peptides (Lu et al., 2009; Olsen and Skriver, 2003). In plants, CLE signalling proteins are involved in shoot, floral and root meristem maintenance and vascular development (Jun, Fiume, and Fletcher, 2008). Nematode-encoded CLEs enable the formation of feeding sites in host roots through the mimicking of plant CLE ligands (Wang et al., 2011a; Guo et al., 2011).

Evolutionary signatures of effectors

The search for proteins bearing signal peptides but lacking transmembrane domains in PPN proteomes, predicted from genomes and transcriptomes, has revealed a large collection of secreted proteins of unknown function and little homology to those in other organisms (Davis, Hussey, and Baum, 2004). Lineage-specific expansions of protein families have been shown to correlate with the emergence of novel functions and stress response (Rubin et al., 2000), and the size of gene families has been suggested to provide information on their adaptive significance (Lespinet et al., 2002).

The fact that effector screening in PPNs is primarily based on the presence of a N-terminal signal sequence may hinder the discovery of effectors excreted by non-classical secretion pathways. One such example are MIF (macrophage migration inhibitory factors) orthologues found in animal parasitic nematodes (Vermeire et al., 2008). Study of this type of effectors has so far been neglected in PPNs although there is experimental evidence of their existence. Two examples are a peroxiredoxin in *Globodera rostochiensis* (Robertson et al., 2000) and an annexin gene in *G. pallida* (Fioretti et al., 2001). Another problem concerning effector screens of PPN genomes based on signal peptides is associated with erroneous gene predictions, such as the absence of the segment coding for the signal peptide at the start of the gene (Zarowiecki and Berriman, 2015). Within this thesis, I tried to ameliorate these effects by guiding the analyses in Chapter 4 based on orthology to published effectors.

Bioinformatic classification of effector proteins can also be achieved by analysing the life-stage specific expression patterns of a PPN and comparing free-living and parasitic stages (Zarowiecki and Berriman, 2015). Eves-van den Akker

et al. (2016b) used this approach for the analysis of the *G. rostochiensis* genome and the resulting effectors were included in the analyses in Chapter 4.

Within the following chapters, I describe the development and illustrate use cases of two software solutions — Chapter 2 and 3 — which allowed me to conduct comparative genomics analyses on PCN genomes, outlined in Chapter 4 and 5. Chapter 6 includes a summary of the most important aspects of the thesis and an outlook for further research is presented.

Chapter 2

BlobTools: software for interrogation of genome assemblies

"There is a computer disease that anybody who works with computers knows about. It's a very serious disease and it interferes completely with the work. The trouble with computers is that you 'play' with them!"

- Richard P. Feynman, *Surely You're Joking, Mr. Feynman!*

2.1 Introduction

Interrogation of genome assemblies to guarantee single-taxon origin is a fundamental step in the genome assembly process. Failure to identify non-target sequence can lead to false conclusions regarding the biology of the target organism, such as its metabolic pathway complement or events of horizontal gene transfer (HGT) between species. Several reports of HGT into eukaryotic genomes have later been

shown to be based on undetected contamination in assemblies. Identification of contamination can radically change the conclusions of a study, as shown for the starlet sea anemone *Nematostella vectensis* (Artamonova and Mushegian, 2013) and the tardigrade *Hypsibius dujardini* (Koutsovoulos et al., 2016). Importantly, undetected non-target sequence contamination of published genomes will pollute public sequence databases and promote propagation of annotation errors (Merchant, Wood, and Salzberg, 2014; Kryukov and Imanishi, 2016).

Reliable assignment of a DNA sequence from a new assembly to its species-of-origin, *i. e.* the association of the sequence ID to an unique, numerical identifier (TaxID) of the NCBI Taxonomy database (Federhen, 2012), is a non-trivial problem (Bridge et al., 2003). Current contaminant screening pipelines are based on sequence similarity to sequences of known origin, sequence composition signatures such as *k*-mers, and/or shared coverage profiles across different datasets. Few are readily applicable to datasets of eukaryotic genomes of any size (Kumar et al., 2013; Eren et al., 2015; Tennessen et al., 2016; Mallet et al., 2017).

`Anvi'o` (Eren et al., 2015) can partition assemblies by clustering sequences based on the output of `CONCOCT` (Alneberg et al., 2014). `CONCOCT` uses Gaussian mixture models to predict the cluster membership of sequences by considering sequence composition and coverage profiles. `PhylOligo` (Mallet et al., 2017) relies exclusively on sequence composition and performs iterative, partially supervised clustering of sequences based on sequence composition profiles. `ProDeGe` (Tennessen et al., 2016) uses a fully unsupervised method based on sequence similarity to databases and sequence composition to partition assemblies using principal components analysis (PCA). It should be noted that while taxonomic assignment based on higher order sequence composition (such as *k*-mers of length four or greater) is highly effective for bacterial sequences, its success has been limited for eukaryotic genomes, as the information content (represented by the number of coding bases) is

lower and sequence composition spectra often show multimodal distributions (Chor et al., 2009).

Existing contaminant screening pipelines also differ in the way results are presented. `Anvi'o` depicts assemblies through interactive plots with rich annotations of sequence composition features, coverages across datasets and taxonomic/binning results. `PhylOligo` offers heatmaps of hierarchical clusterings of sequences, tree visualisations, and t-SNE (t-Distributed Stochastic Neighbour Embedding) plots, where sequence composition clusterings have been reduced to two dimensions. `ProDeGe` displays sequences in an interactive, three-dimensional *k*-mer PCA plot.

BlobPlots, or Taxon-Annotated Gc-Coverage (TAGC) plots, are another contamination detection and data partitioning methodology. Kumar et al. (2013) visualised genome assemblies as two-dimensional scatter plots (see Figure 2.1.1), in which sequences are represented by dots and coloured by taxonomic affiliation based on sequence similarity search results. For each sequence, the position on the Y-axis is determined by the base coverage of the sequence in the coverage library, a proxy for molarity of input DNA. The position on the X-axis is determined by the GC content, the proportion of G and C bases in the sequence, which can differ substantially between genomes. BlobPlots have proven to be an intuitive and powerful approach for taxonomic interrogation of genome assemblies (Koutsovoulos et al., 2014; Dentinger et al., 2015).

I developed BlobTools as a modular command-line solution for the visualisation of genome assemblies and taxonomic interrogation for purposes of quality control. It is a complete reimplementation of the `blobology` pipeline (Kumar et al., 2013). BlobTools is focussed on usability and includes improved taxonomic assignment of sequences based on custom user input, support for coverage information based on multiple formats and sequencing libraries, and novel visualisations.

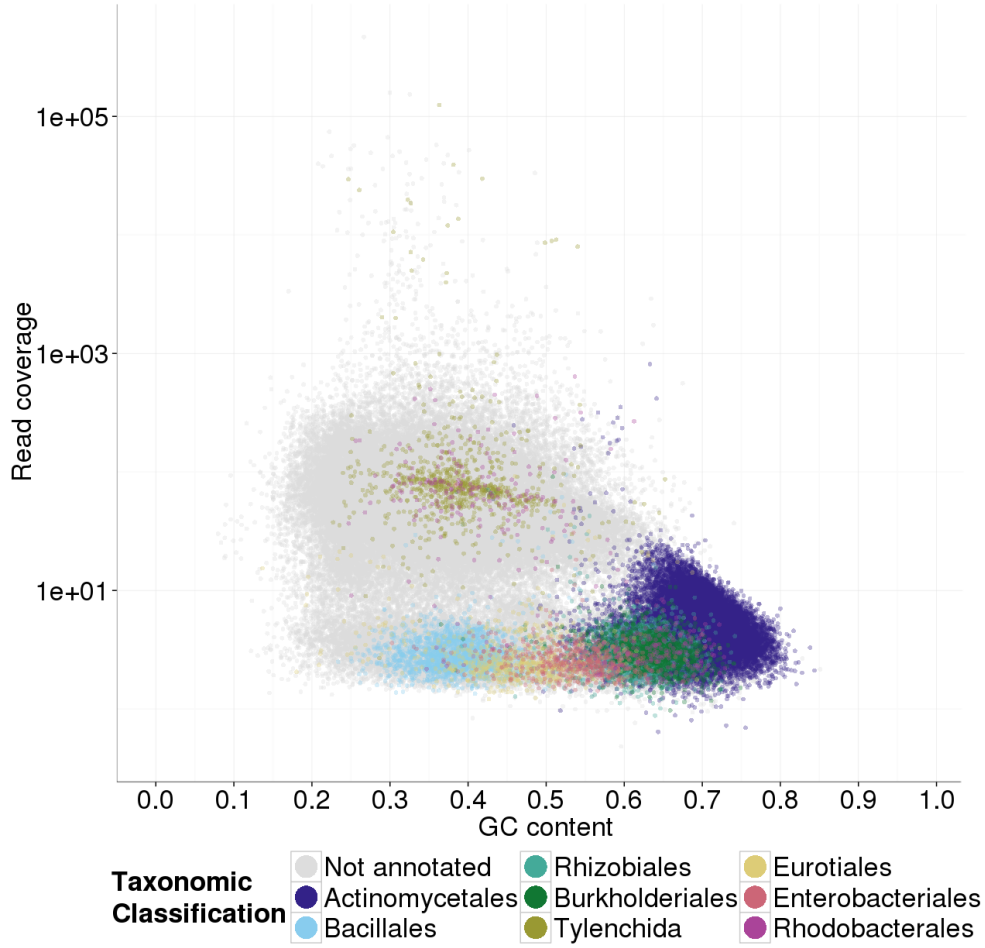


Figure 2.1.1: Taxon-Annotated Gc-Coverage (TAGC) plot. CLC assembly of *G. rostochiensis* dataset ERR123958 taxonomically annotated based on best BLAST hit against NCBI nt.

In this chapter, I describe the implementation of BlobTools and list three use cases which highlight different features of the toolkit. Parts of Sections 1.2.1, 2.1, 2.2, and 2.3 have been submitted as a 'Software Tool Article' to the Open Research publishing platform *F1000Research* (Laetsch and Blaxter, 2017a), which is currently under public peer-review (DOI: 10.12688/f1000research.12232.1). Section 2.4 was published as part of Koutsovoulos et al., 2016 in *Proceedings of the National Academy of Sciences* (DOI: 10.1073/pnas.1600338113) and Section 2.5 was published as part of Eves-van den Akker et al., 2016b in *BMC Genome Biology* (DOI: 10.1186/s13059-016-0985-1).

2.2 Implementation

BlobTools is written in the programming language Python and consists of a main executable that allows the user to interact with the implemented commands (see Table 2.2.1). It offers a simple, modular command line interface — 'samtools style' *sensu* Seemann, 2013 — which can easily be adapted to process multiple datasets simultaneously using `GNU parallel` (Tange, 2011). Inputs for BlobTools are standard file formats commonly created during the course of genome assembly projects. The primary processing in BlobTools constructs a BlobDB data structure based on user input. From this data structure, BlobTools generates easily interpretable, two-dimensional visualisations ready for publication, in addition to tabular output, which allows the user to partition sequences and sequencing reads contributing to them for separate downstream processing.

2.2.1 Taxonomy assignment approach

Taxonomy assignment in BlobTools is based on user-supplied, tab-separated-value (TSV) files composed of three columns: the input sequence ID, a NCBI TaxID, and a numerical score. I refer to these TSV files as 'hits' files below. They can be generated from the output of sequence similarity searches, such as `BLAST` (Camacho et al., 2009) or `Diamond` (Buchfink, Xie, and Huson, 2015) searches against public or custom databases, or the output of other contaminant identification tools. The BlobTools command `taxify` allows easy conversion of tabular file formats to BlobTools compatible input, in addition to annotation of similarity search results based on NCBI TaxID mapping files, as available from UniProt and NCBI.

BlobTools assigns a single NCBI taxonomy for each sequence in the assembly, based on the highest scoring NCBI TaxID in the input provided by the user at

Table 2.2.1: Tasks performed by BlobTools command.

BlobTools command	Task
`create`	Parsing of input files and creation of BlobTools data structure, *i. e.* BlobDB
`view`	Generation of tabular output for manual inspection and subsequent partitioning of sequences in the assembly, input files for `CONCOCT`, and/or COV files based on a BlobDB
`plot`	Plotting of BlobPlots based on a BlobDB
`covplot`	Plotting of CovPlots based on a BlobDB and a COV file
`seqfilter`	Partitioning of sequences from a FASTA file based on a list of sequence IDs
`bamfilter`	Partitioning of PE reads from a BAM file based on a list of sequence IDs and their mapping behaviour
`map2cov`	Generation of a COV file (containing base and read coverage) based on a BAM/CAS file
`taxify`	Annotation of tabular sequence similarity search output (*e. g.* `BLAST`/`Diamond` output) with TaxIDs from a mapping file or generation of a BlobTools 'hits' file based on custom user input

taxonomic ranks of species, genus, family, order, phylum, and superkingdom. Score calculation can be controlled through a minimal score threshold (`--min_score`) and a minimal difference in scores (`--min_diff`) between the best and second-best scoring taxonomy. In addition, three non-canonical taxonomic annotations are possible: 'no-hit', the suffix '-undef' and 'unresolved'. Sequences not assigned to any taxonomic group, or not passing the `--min_score` threshold, are labelled 'no-hit'. If a NCBI TaxID has no explicit parent at a taxonomic rank, the suffix '-undef' is appended to the next upper taxonomic rank for which one does exist. For instance,

the taxonomic family Suidae (pigs) has no order assigned to it, which results in taxonomic assignment of 'Chordata-undef' at the rank of order. In cases where the score difference between the best and second-best hits is smaller than `--min_diff`, sequences are labelled as 'unresolved'.

Multiple 'hits' files can be provided as input. In this case, the behaviour of the taxonomy assignment process can be controlled further through 'taxrules'. The highest scoring taxonomy can either be inferred across all files ('bestsum') or successively ('bestsumorder') in the order in which the files were supplied as input. In the latter case, only sequences that received no hits from one file are considered for taxonomic annotation in the next file which allows leveraging reliability of scores of different input file sources.

The original `blobology` pipeline by Kumar et al. (2013) recommended the use of a single, best BLAST hit per sequence for taxonomy assignment. However, taxonomically mis-annotated sequences in databases — often derived from inclusion of un-screened genome assemblies — can lead to erroneous taxonomic annotation. BlobTools mitigates this issue by accepting multiple hits per sequence and allocating taxonomy based on the highest sum of scores.

It should be noted that a definitive taxonomic placement for every sequence in the assembly is not required for successful taxonomic partitioning of sequences, since differential coverage and sequence composition profiles between the genomes are often sufficient.

2.2.2 Visualisation options

In BlobTools, sequences are depicted as circles in BlobPlots — as opposed to dots in the `blobology` pipeline — and the diameter of circles is scaled proportionally to

sequence length. The scatter-plot is decorated with coverage and GC histograms for each taxonomic group, which are weighted by the total span (cumulative length) of sequences occupying each bin. A legend reflects the taxonomic affiliation of sequences and lists count, total span and N50 by taxonomic group. Taxonomic groups can be plotted at any taxonomic rank and colours are selected dynamically from a colour map. The number of taxonomic groups to be plotted can be controlled (`--plotgroups`, default is '7') and remaining groups are binned into the category 'others'. An example is shown in Figure 2.3.2A.

The power of differential coverage profiles across multiple sequencing libraries for partitioning sequences in an assembly prompted the development of CovPlots (Figure 2.3.3). CovPlots are analogous to BlobPlots except that the GC-axis is replaced by a coverage-axis of a second sequencing library. CovPlots can be used for the visualisation of patterns of differential coverage signatures between taxonomic groups in an assembly.

The commands for generating BlobPlots and CovPlots support additional input parameters controlling visualisation behaviour. These include cumulative addition (`--cumulative`) or separate plotting (`--multiplot`) for each taxonomic group, exclusion (`--exclude`) or relabelling (`--relabel`) of taxonomic groups, assignment of specific HEX colours to groups (`--colour`) or labelling sequences based on arbitrary, user defined categories (`--catcolour`). The latter could be, for instance, binned categories of RNAseq mappings to sequences in the assembly as shown in Section 2.4 and Koutsovoulos et al., 2016.

ReadCovPlots (Figure 2.3.2B and 2.3.2C) visualise the proportion of reads of a library that are unmapped or mapped, showing the percentage of mapped reads by taxonomic group, as barcharts. These can be of use for rapid taxonomic screening of multiple sequencing libraries within a single project. The underlying data of

ReadCovPlots and additional metrics are written to tabular text files for custom analyses by the user.

2.2.3 Support of multiple coverage libraries

BlobTools supports coverage input (BAM/CAS format) from multiple sequencing libraries. As these data formats contain more information than needed, BlobTools parses coverage information of sequences — normalised base coverage and read coverage — into COV files in TSV format. These files can be generated through the command `map2cov` prior to construction of a BlobDB.

Within the BlobDB data structure, base and read coverage information is stored for each sequence in the assembly. If more than one coverage file is supplied, BlobTools constructs an additional coverage library ('covsum') internally, containing the sum of coverages for each sequence across all coverage files. This internal coverage library is considered when extracting views or plotting visualisations.

2.2.4 Operation

BlobTools is freely available under GNU General Public License v3.0 at `https://github.com/DRL/blobtools`. System requirements for BlobTools include a `UNIX` based operating system, `Python 2.7`, and `pip`. An installation script is provided, which installs Python dependencies, downloads and processes a copy of the NCBI TaxDump, and downloads and compiles a copy of `samtools` (Li et al., 2009). Instructions for installation and execution of BlobTools can be found at the GitHub repository and detailed documentation is available at `https://blobtools.readme.io`.

BlobTools workflows

Two recommended workflows for BlobTools have been developed: Workflow A is targeted at *de novo* genome assembly projects in the absence of a reference genome and workflow B is aimed at projects where a reference genome is available.

Workflow A (Figure 2.2.1A) proceeds through construction of a BlobDB data structure based on input files (step A1), visualisation of assembly and generation of tabular output (A2), partitioning of sequence IDs based on user-defined parameters informed by the visualisations (A3) and partitioning of PE reads based on sequence IDs (A4). It should be noted that while the BlobTools command `create` (step A1) supports multiple mapping formats, it is recommended that these are processed in advance using the command `map2cov`. Generation of tabular 'hits' files is simplified through the command `taxify`, which allows annotation of similarity search results based on TaxID mapping files or based on custom user input in tabular format.

BlobTools can process both PE and SE read files. The command `bamfilter` in step A4 is only of relevance if PE read data is used, since partitioning of SE read data is trivial and can easily be achieved via `GNU grep`. The command `bamfilter` can be controlled with a list of sequence IDs to include or to exclude. Use of an exclusion list causes all sequence IDs, except those specified, to be included. In both cases it will output up to four interleaved FASTQ files depending on the actual mapping behaviour of the read pairs and whether the parameter `--include_unmapped` is provided. Possible mapping behaviours of read pairs are: both reads mapping to included sequences (included-included: 'InIn'), one read mapping to an included sequence and the other being unmapped ('InUn'), and one read mapping to an excluded sequence and the other mapping to an included sequence ('ExIn'). If the `--include_unmapped` parameter is specified, read pairs where neither read maps to the assembly ('UnUn') are also reported. The latter case can occur if the

assembler used for generating the sequences did not make use of all reads in the dataset. The resulting PE read files can then be assembled separately and the workflow is repeated. Decisions concerning which PE read files to use is left to the discretion of the user. However, as a general rule, if target taxa have been sequenced at low coverages it might be preferable to be inclusive (using 'InIn', 'InEx', 'InUn' and 'UnUn' reads for assembly) and risking including non-target reads, than being exclusive (using only 'InIn' and 'InUn' reads for assembly) and thereby risking losing significant proportions of reads from the target genome(s).

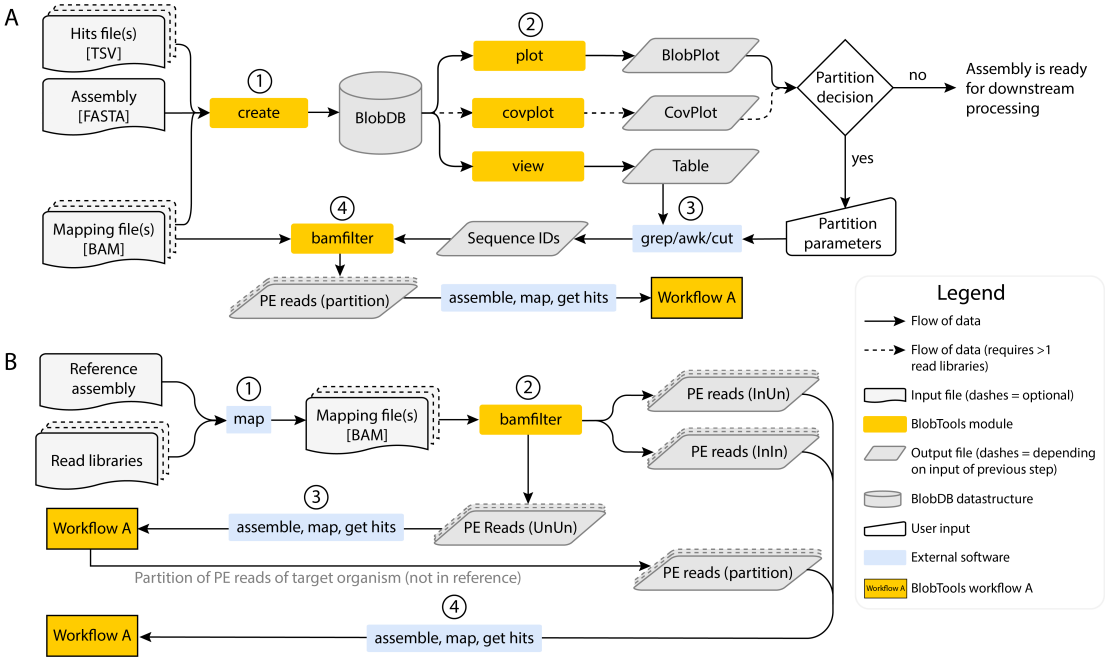


Figure 2.2.1: BlobTools workflows. **A** Workflow A. 1: Creation of a BlobDB data structure based on input files. 2: Generation of visualisations and tabular output. 3: Partitioning of sequence IDs in assembly, based on user-defined parameters informed by the visualisations. 4: Partitioning of PE reads based on sequence IDs. **B** Workflow B. 1: Reads are mapped against the reference genome. 2: BAM file is processed to generate FASTQ files based on read mapping behaviour. 3: FASTQ file of read pairs where neither read maps to the reference genome ('UnUn') are assembled *de novo* and used in workflow A. 4: partition of read pairs of target taxon recovered from workflow A are assembled together with the other target taxon read pairs from step 2 and used in workflow A.

Workflow B (Figure 2.2.1B) should be applied when a reference genome is

available. Reads are mapped against the reference genome (B1) and the resulting BAM file is processed with the command `bamfilter` (B2) using the parameter `--include_unmapped` and without providing a list of sequences. This will result in three FASTQ files: 'InIn', 'InUn', and 'UnUn'. Since taxonomic origin of the 'InIn' and 'InUn' reads has been established through the mapping step, only the 'UnUn' reads are assembled *de novo* (B3) and processed via workflow A. This decreases computational requirements substantially. If workflow A yields a PE read partition of the target organism — consisting of regions in the organism's genome not present in the reference — these reads can then be used together with the 'InIn' and 'InUn' reads from step B2 to generate a new assembly (B4) which should be screened again via Workflow A. This iterative workflow can easily be applied to projects studying highly variable species, where segmental presence/absence is common and a reference genome is expanded to form a pangenome as new samples are sequenced, or holobiomes, where reference genomes of multiple taxa are expanded as new samples are added.

2.3 Use case 1: BlobTools analysis of simulated datasets

2.3.1 Introduction

Assessment of a novel computational tool is simplified when using simulated data, since the 'truth' is known and computational outcomes can thus be evaluated empirically. I simulated two read libraries for the nematode *Caenorhabditis elegans* contaminated with other organisms. Library A contains *C. elegans* reads contaminated with reads from *Escherichia coli*, *Homo sapiens* chromosome 19 and *H. sapiens* mitochondrial (mtDNA) genome, mimicking a dataset where the target genome is

contaminated with DNA from food (*E. coli*) and operator (*H. sapiens*). Library B is composed of *C. elegans* reads contaminated with *Pseudomonas aeruginosa*, mimicking a project where the metazoan target species is heavily colonised by a prokaryotic organism.

The read datasets were assembled and processed using BlobTools (workflow A) with the goal of separating the four different genomes. To simulate phylogenetic distance between sequences in the assemblies and those in the databases used for sequence similarity searches, sequences in the databases originating from the relevant taxa were removed at different taxonomic ranks, excluding sequences from hominids, *Escherichia*, *Pseudomonas*, and *Caenorhabditis elegans*. The resulting assemblies were evaluated against the 'truth', *i. e.* based on the genome-of-origin of the reads contributing to them. In addition, the influence of sequence similarity search parameters on BlobTools taxonomic annotation was evaluated.

2.3.2 Methods

`BlobTools v1.0` (Laetsch et al., 2017) was used for all analyses.

Data

Reference genomes were retrieved from the ENSEMBL website and read datasets were simulated as Illumina HiSeq2500 PE reads with mean coverage as listed in Table 2.3.1 using `ART v2.5.8` (Huang et al., 2012) (`-l 150 -m 500 -s 10`). Reads were concatenated into the two libraries and shuffled using `BBmap shuffle v37.02` (`https://sourceforge.net/projects/bbmap/`).

Genome assembly

All assemblies were performed using `CLC assembler v5.0.0` (QIAGEN) by specifying read libraries as PE with an insert size ranging from 300 to 700b (`-p fb ss 300 700`). To assess the number of conserved genes recovered, assemblies were evaluated using `BUSCO v2.0.1` (Simão et al., 2015) against the databases: `nematoda_odb9`, `mammalia_odb9`, `enterobacterales_odb9`, and `gammaproteobacteria_odb9`.

Table 2.3.1: Simulated read libraries.

Dataset	Reference genome	INSDC Accession	Coverage (X)
Library A	*C. elegans* N2	GCA_000002985.3	50
	E. coli str. K-12 substr. MG1655	GCA_000801205	25
	H. sapiens chr19 GRCh38.p10	GCA_000001405.25	10
	H. sapiens mtDNA GRCh38.p10	GCA_000001405.25	250
Library B	*C. elegans* N2	GCA_000002985.3	25
	P. aeruginosa PAO1	GCA_000006765.1	100

Read mapping

All mapping files were created by mapping read libraries against assemblies using `BWA mem v0.7.15-r1140` (Li, 2013) and `samtools v1.5` (Li et al., 2009). BAM files were converted to COV format using BlobTools `map2cov`.

Sequence similarity searches

Sequence similarity searches were performed against NCBI nt (retrieved 2017-06-13) using `BLASTn megablast v2.6.0+` (Camacho et al., 2009) and against Uniprot Reference Proteomes (retrieved 2017-07-07) using `Diamond blastx v0.9.5` (Buchfink, Xie, and Huson, 2015). Parameters for `BLASTn` were `-evalue 1e-25` and `-outfmt '6 qseqid staxids bitscore std'` for all searches. `Diamond` was run with the parameters `--sensitive --evalue 1e-25 --outfmt 6` and results were annotated with NCBI TaxIDs using the BlobTools command `taxify` and a UniProt ID mapping file, filtered to only include NCBI TaxIDs.

For the evaluation of the impact of parameters of sequence similarity searches on BlobTools taxonomic assignment, additional parameters were used for `BLAST` and `Diamond` searches. The search parameter IDs and parameters are as follow:

- `MTS1`: `-max-target-seqs 1`
- `MTS10`: `-max-target-seqs 10`
- `HSP1`: `-max_hsps 1`
- `CUL10`: `-culling_limit 10` (only `BLAST`)

Sequence similarity searches were taxonomically restricted to simulate phylogenetic distance between query and subject sequences in the database. `BLASTn` searches were taxonomically restricted by retrieving GI lists from the NCBI nucleotide portal for the TaxIDs 9604 ('Hominidae'), 561 ('Escherichia'), 6239 ('Caenorhabditis elegans'), 28384 ('other sequences'), and 286 ('Pseudomonas') and supplying them with the parameter `-negative_gilist`. `Diamond blastx`

searches taxonomically restricted by retrieving subtree TaxIDs for the relevant groups from NCBI taxonomy portal and removing the associated sequences from Uniprot Reference Proteomes.

Evaluation of influence of sequence similarity search parameters on BlobTools taxonomy assignment

The 'true' taxonomy of sequences in the combined assembly of both simulated read libraries was compared to the taxonomy inferred by BlobTools based on sequence similarity searches. The 'true' taxonomy of each sequence was determined by mapping the simulated read libraries against the assembly. Unambiguous taxonomy — cases where only reads originating from one reference genome map to a sequence in the assembly — could be assigned to 98.07% of sequences (16,289 out of 16,610) in the assembly, totalling 99.89% (158,001,623 out of 158,178,224 b) of assembled span.

Similarity search results were supplied to BlobTools to construct BlobDBs and tabular output was generated using BlobTools `view` (`--hits -r superkingdom -r phylum -r order`). Taxrule 'bestsumorder' (`-x bestsumorder`) was specified in cases where results from both `BLASTn` searches against NCBI nt and `Diamond blastx` searches against Uniprot Reference Proteomes were used as input (always in this order). For each BlobDB, accuracy was calculated at the taxonomic ranks of order (Rhabditida, Primates, Pseudomonadales, and Enterobacterales). Tables were evaluated using the script `analyse_blobtools_tables.py` (available at `https://github.com/DRL/blobtools_manuscript`).

For each taxon, counts of bases in the assembly were classified as true/false positives/negatives as follows:

- True positives (TP): sum of bases in sequences where BlobTools taxonomic annotation of taxon is equal to 'true' taxonomy.

- False positives (FP): sum of bases in sequences where BlobTools taxonomic annotation indicated the taxon, but the 'true' taxonomy differed.

- True negatives (TN): sum of bases in sequences where both BlobTools taxonomic annotation and 'true' taxonomy was different to taxon.

- False negatives (FN): sum of bases in sequences where BlobTools taxonomic annotation of taxon failed to reflect 'true' taxonomy.

Precision and recall was calculated using the formulae:

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

The F-score (Rijsbergen, 1979) was calculated as the harmonic mean of precision and recall, based on the formula:

$$\text{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Taxonomic partitioning of read libraries using BlobTools

Both simulated read datasets were assembled together and each library was mapped individually against the assembly. The assembly was supplied to BlobTools, together with coverage information extracted from BAM files and the results of sequence similarity searches. The search results provided to BlobTools were from `BLASTn megablast` (`MTS1` and `HSP1`) and `Diamond blastx` (`MTS1`) searches against taxonomically restricted versions of NCBI nt and UniProt Reference Proteomes.

A BlobPlot, ReadCovPlots and a CovPlot were generated at the taxonomic rank of 'order'. A tabular view of the BlobDB was generated using the command `view` under the taxrule 'bestsumorder' and for the taxonomic ranks of 'superkingdom', 'phylum', and 'order'. Sequences were partitioned based on differential coverage and taxonomy annotation (see Figure 2.3.3) using the tabular view and the UNIX tools `GNU grep`, `GNU cut`, and `GNU awk`. Subsequently, read pairs were partitioned based on mapping behaviour to these sequence partitions using the command `bamfilter` and only read pairs where both reads mapped to included sequences (*i. e.* the 'InIn' set) were assembled by taxonomic group.

BlobPlots for the four assemblies based on partitioned read sets, 'RH-BT' (Rhabditida), 'PR-BT' (Primates), 'PS-BT' (Pseudomonadales), and 'EN-BT' (Enterobacterales) were generated. Coverage information was based on mapping of both simulated sequencing libraries against all four assemblies and sequences were coloured based on the genome-of-origin of the simulated reads mapping to them.

Evaluation of taxonomic partitioning of read libraries

To account for assembly and mapping biases, the original simulated read sets were also assembled separately by taxon, yielding the assemblies CELEG-SIM (reads simulated from the *C. elegans* genome), HSAPI-SIM (*H. sapiens* chromosome 19 and mtDNA), PAERU-SIM (*P. aeruginosa*), and ECOLI-SIM (*E. coli*).

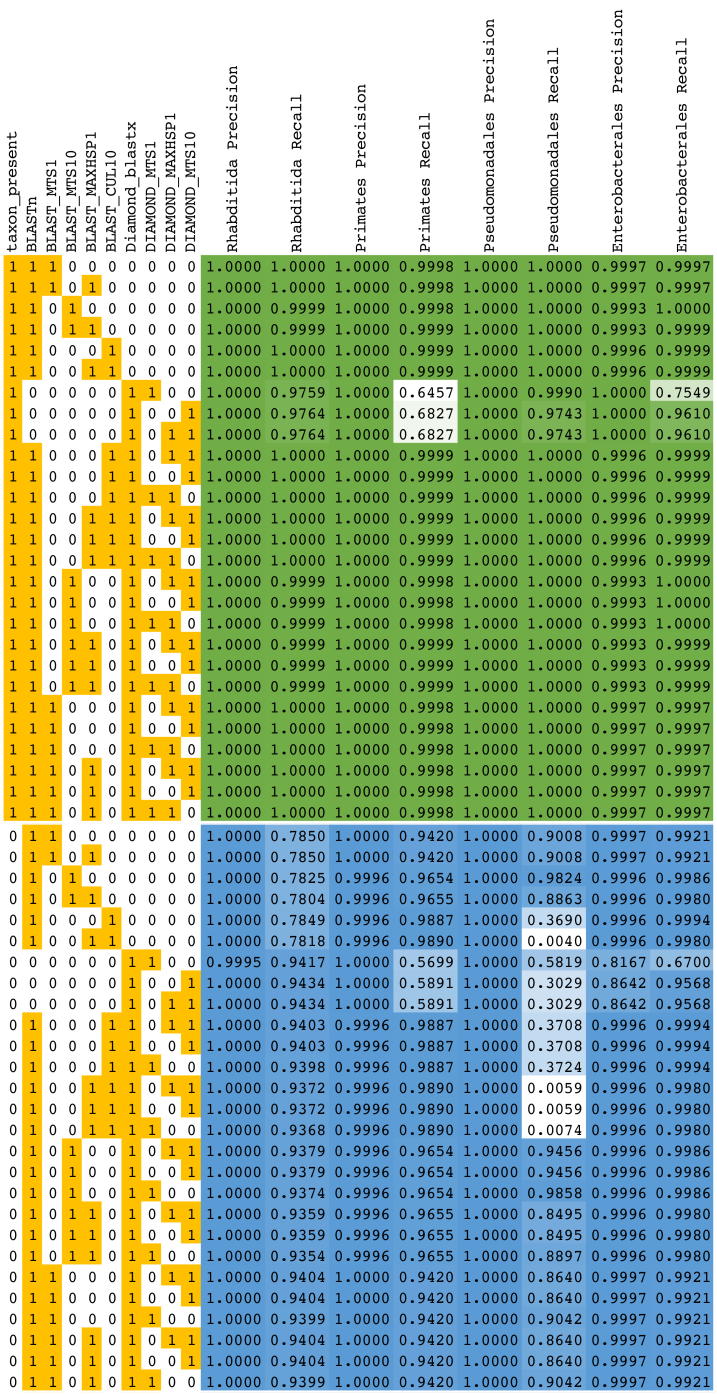
Cleaned assemblies were evaluated based on the count of simulated reads by genome-of-origin mapping to them (Table 2.3.2) and based on standard assembly metrics (Table 2.3.3).

2.3.3 Results

Influence of sequence similarity search parameters on BlobTools taxonomy assignment

Since exhaustive searches against large databases require time and computing power, I focussed on parameters that limit resource usage and control the number of returned results. In both `BLASTn` and `Diamond blastx`, the options `-max-target-seq` and `-max-hsps` are implemented. The former is an early filter applied during primary search and excludes initial hits from later examination. The latter controls the number of high-scoring pairs (HSPs) reported between a query and a subject in the search. The `BLASTn` specific parameter `-culling-limit` controls the number of hits that can be allocated to a given region on the query.

In the context of sequence similarity searches for taxonomic annotations, FNs arise mainly through phylogenetic distance of the target sequence to the ones in the database, resulting in the lack of results (*i. e.* 'no-hit'). However, both FNs and FPs



taxon_present	BLASTn	BLAST_MTS1	BLAST_MTS10	BLAST_MAXHSP1	BLAST_CUL10	Diamond_blastx	DIAMOND_MTS1	DIAMOND_MAXHSP1	DIAMOND_MTS10	Rhabditida Precision	Rhabditida Recall	Primates Precision	Primates Recall	Pseudomonadales Precision	Pseudomonadales Recall	Enterobacterales Precision	Enterobacterales Recall
1	1	1	0	0	0	0	0	0	0	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	1	0	1	0	0	0	0	0	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	0	1	0	0	0	0	0	0	1.0000	0.9999	1.0000	0.9998	1.0000	1.0000	0.9993	1.0000
1	1	0	1	1	0	0	0	0	0	1.0000	0.9999	1.0000	0.9999	1.0000	1.0000	0.9993	0.9999
1	1	0	0	0	1	0	0	0	0	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	0	1	1	0	0	0	0	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	0	0	0	0	0	1	1	0	0	1.0000	0.9759	1.0000	0.6457	1.0000	0.9990	1.0000	0.7549
1	0	0	0	0	0	1	0	0	1	1.0000	0.9764	1.0000	0.6827	1.0000	0.9743	1.0000	0.9610
1	0	0	0	0	0	1	0	1	1	1.0000	0.9764	1.0000	0.6827	1.0000	0.9743	1.0000	0.9610
1	1	0	0	0	1	1	0	1	1	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	0	0	1	1	0	0	1	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	0	0	1	1	1	1	0	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	0	1	1	1	0	1	1	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	0	1	1	1	0	0	1	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	0	1	1	1	1	1	0	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000	0.9996	0.9999
1	1	0	1	0	0	1	0	1	1	1.0000	0.9999	1.0000	0.9998	1.0000	1.0000	0.9993	1.0000
1	1	0	1	0	0	1	0	0	1	1.0000	0.9999	1.0000	0.9998	1.0000	1.0000	0.9993	1.0000
1	1	0	1	0	0	1	1	1	0	1.0000	0.9999	1.0000	0.9998	1.0000	1.0000	0.9993	1.0000
1	1	0	1	1	0	1	0	1	1	1.0000	0.9999	1.0000	0.9999	1.0000	1.0000	0.9993	0.9999
1	1	0	1	1	0	1	0	0	1	1.0000	0.9999	1.0000	0.9999	1.0000	1.0000	0.9993	0.9999
1	1	0	1	1	0	1	1	1	0	1.0000	0.9999	1.0000	0.9999	1.0000	1.0000	0.9993	0.9999
1	1	1	0	0	0	1	0	1	1	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	1	0	0	0	1	0	0	1	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	1	0	0	0	1	1	1	0	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	1	0	1	0	1	0	1	1	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	1	0	1	0	1	0	0	1	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
1	1	1	0	1	0	1	1	1	0	1.0000	1.0000	1.0000	0.9998	1.0000	1.0000	0.9997	0.9997
0	1	1	0	0	0	0	0	0	0	1.0000	0.7850	1.0000	0.9420	1.0000	0.9008	0.9997	0.9921
0	1	1	0	1	0	0	0	0	0	1.0000	0.7850	1.0000	0.9420	1.0000	0.9008	0.9997	0.9921
0	1	0	1	0	0	0	0	0	0	1.0000	0.7825	0.9996	0.9654	1.0000	0.9824	0.9996	0.9986
0	1	0	1	1	0	0	0	0	0	1.0000	0.7804	0.9996	0.9655	1.0000	0.8863	0.9996	0.9980
0	1	0	0	0	1	0	0	0	0	1.0000	0.7849	0.9996	0.9887	1.0000	0.3690	0.9996	0.9994
0	1	0	0	1	1	0	0	0	0	1.0000	0.7818	0.9996	0.9890	1.0000	0.0040	0.9996	0.9980
0	0	0	0	0	0	1	1	0	0	0.9995	0.9417	1.0000	0.5699	1.0000	0.5819	0.8167	0.6700
0	0	0	0	0	0	1	0	0	1	1.0000	0.9434	1.0000	0.5891	1.0000	0.3029	0.8642	0.9568
0	0	0	0	0	0	1	0	1	1	1.0000	0.9434	1.0000	0.5891	1.0000	0.3029	0.8642	0.9568
0	1	0	0	0	1	1	0	1	1	1.0000	0.9403	0.9996	0.9887	1.0000	0.3708	0.9996	0.9994
0	1	0	0	0	1	1	0	0	1	1.0000	0.9403	0.9996	0.9887	1.0000	0.3708	0.9996	0.9994
0	1	0	0	0	1	1	1	0	0	1.0000	0.9398	0.9996	0.9887	1.0000	0.3724	0.9996	0.9994
0	1	0	0	1	1	1	0	1	1	1.0000	0.9372	0.9996	0.9890	1.0000	0.0059	0.9996	0.9980
0	1	0	0	1	1	1	0	0	1	1.0000	0.9372	0.9996	0.9890	1.0000	0.0059	0.9996	0.9980
0	1	0	0	1	1	1	1	0	0	1.0000	0.9368	0.9996	0.9890	1.0000	0.0074	0.9996	0.9980
0	1	0	1	0	0	1	0	1	1	1.0000	0.9379	0.9996	0.9654	1.0000	0.9456	0.9996	0.9986
0	1	0	1	0	0	1	0	0	1	1.0000	0.9379	0.9996	0.9654	1.0000	0.9456	0.9996	0.9986
0	1	0	1	0	0	1	1	0	0	1.0000	0.9374	0.9996	0.9654	1.0000	0.9858	0.9996	0.9986
0	1	0	1	1	0	1	0	1	1	1.0000	0.9359	0.9996	0.9655	1.0000	0.8495	0.9996	0.9980
0	1	0	1	1	0	1	0	0	1	1.0000	0.9359	0.9996	0.9655	1.0000	0.8495	0.9996	0.9980
0	1	0	1	1	0	1	1	0	0	1.0000	0.9354	0.9996	0.9655	1.0000	0.8897	0.9996	0.9980
0	1	1	0	0	0	1	0	1	1	1.0000	0.9404	1.0000	0.9420	1.0000	0.8640	0.9997	0.9921
0	1	1	0	0	0	1	0	0	1	1.0000	0.9404	1.0000	0.9420	1.0000	0.8640	0.9997	0.9921
0	1	1	0	0	0	1	1	0	0	1.0000	0.9399	1.0000	0.9420	1.0000	0.9042	0.9997	0.9921
0	1	1	0	1	0	1	0	1	1	1.0000	0.9404	1.0000	0.9420	1.0000	0.8640	0.9997	0.9921
0	1	1	0	1	0	1	0	0	1	1.0000	0.9404	1.0000	0.9420	1.0000	0.8640	0.9997	0.9921
0	1	1	0	1	0	1	1	0	0	1.0000	0.9399	1.0000	0.9420	1.0000	0.9042	0.9997	0.9921

Figure 2.3.1: Values for precision and recall of taxonomic assignment via BlobTools based on combinations of sequence similarity search parameters. **Top**: Sequence similarity searches when query organism is present in the database. **Bottom**: Sequence similarity searches when query organism is not present in the database.

can also arise through taxonomically mis-annotated sequences in the database, *i. e.* when a genome sequence is contaminated with sequences from another genome. For instance, if the target organism is *E. coli* a FN could originate from a *C. elegans* genome contaminated with bacterial DNA, as the bacterial sequence in the assembly would be tagged with the TaxID of *C. elegans*. Simultaneously, this result would count as a FP if the target organism is *C. elegans*, since a non-nematode sequence is tagged with the TaxID of *C. elegans*. The values for precision and recall are shown in Figure 2.3.1. Detailed values for F-scores and the number of bases annotated as TP, FP, TN, FN for each of (the combinations of) search parameters are listed in Table S1 and S3 in Laetsch and Blaxter, 2017a.

Sequence similarity searches against databases without taxonomic masking yielded results mostly congruent with taxonomic annotation through read mapping. FP taxonomic annotations in `BLASTn` searches were only found for a small amount of *E. coli* sequences. One such sequence is 'contig_8499' in the combined assembly of both simulated libraries, which was assembled from read pairs originating from the X chromosome of the *C. elegans* reference (position 5,909,163 to 5,911,151). However, `BLASTn` searches identified this sequence as a Tn10 transposon with the taxonomic annotation of Enterobacterales. A likely explanation for this case is the transposition of this sequence into a fosmid clone during the *C. elegans* genome project, as has been noted by the WormBase database (see `http://www.wormbase.org/species/c_elegans/feature/WBsf977957#0123--10`). No FPs were observed for `Diamond` searches, as these are restricted to protein coding sequences. However, variation in the number of FNs was greater and caused mainly by the absence of hits. `Diamond` searches exhibited the highest number of FNs, especially for *H. sapiens* and *E. coli* sequences. This was more pronounced when using `MTS1`. If more hits are supplied, the number of FNs decreases. If `BLASTn` searches were provided in addition to `Diamond` searches (using taxrule 'bestsumorder'), recall ranges between 0.9993 and 1.0.

Taxonomic masking of sequence databases to simulate phylogenetic distance between query and subject sequences, revealed more complex patterns of interaction between search parameters which varied between the taxa. This variation is not surprising since taxonomic masking was carried out at different, non-analogous taxonomic ranks. Similar to unmasked searches, FPs are not a major concern. Lowest values of precision (ranging from 0.8167 to 0.8642) were observed for *E. coli* sequences when using `Diamond` alone. Number of FNs bases are also most extreme in taxonomic annotations based on `Diamond` alone (for *H. sapiens* and *P. aeruginosa* sequences) or when using the `BLASTn` parameter `CUL10` (for *P. aeruginosa* sequences). Hence, I discourage the use of `-culling_limit` for similarity searches used for taxonomic annotation in BlobTools.

For this dataset, the best trade-off between false positive and false negative taxonomic annotations was achieved by combining a `BLASTn` search (`-max-target-seqs 10 -evalue 1e-25`) against NCBI nt with a `Diamond` search (`--evalue 1e-25 --max-target-seqs 1`) against UniProt Reference Proteomes, in this order, using BlobTools taxrule 'bestsumorder' (lowest precision and recall, 0.9996 and 0.9374, respectively). However, a much faster search with acceptable outcome was achieved by changing the `BLASTn` parameters to `-max-target-seqs 1 -max_hsps 1` (lowest precision and recall, 0.9997 and 0.9042, respectively).

Although phylogenetic distance between query sequences and those in the databases was simulated through taxonomic masking, it should be noted that the results discussed here are by no means universal. If an organism has not been sequenced previously, the number of FNs will be high and if it was, but the data is contaminated with other taxa or the organism is a contaminant itself, the number of FPs and FNs will be greater.

Taxonomic partitioning of read libraries using BlobTools

The BlobPlot (Figure 2.3.2A) and ReadCovPlots (Figure 2.3.2B and C) of the initial assembly of both read libraries revealed the taxonomic composition of the datasets. In the BlobPlot, Rhabditida and Pseudomonadales sequences form two distinct 'blobs' at high coverage separated by their GC content. Primate and Enterobacterales sequences are visible at lower coverages (with the exception of the human mtDNA sequence near the Rhabditida 'blob'). Although taxonomic annotation is inaccurate for some sequences in each 'blob', patterns of coverage and GC defining the 'blobs' allow manual imputation of taxonomic membership.

The CovPlot (Figure 2.3.3) yields an even clearer picture by separating the 'blobs' based on the coverage received in each read library. Rhabditida sequences appear on a diagonal line as reads from both libraries contribute to their coverages. Pseudomonadales, Rhabditida and Enterobacterales sequences are clearly separated, with Primates and Enterobacterales overlapping. Partitioning of sequence IDs was achieved using parameters based on differential coverage and taxonomy annotation.

BlobPlots for the four assemblies 'RH-BT', 'PR-BT', 'PS-BT', and 'EN-BT' based on partitioned reads are shown in Figure 2.3.4. Taxonomic evaluation of cleaned assemblies (based on the count of simulated reads by genome-of-origin mapping to them) is shown in Table 2.3.2, and standard assembly metrics are listed in Table 2.3.3.

Table 2.3.2: Evaluation of read mappings. Percentages of reads (partitioned by taxonomic origin) mapped to sequences in each of the BlobTools-processed assemblies (suffix '-BT'). Reads that did not map to any sequence are listed under 'Not Mapped'. Bold: Zero reads mapped.

	Reads mapping (in %) to				
Taxonomic origin of simulated reads	'RH-BT'	'PR-BT'	'PS-BT'	'EN-BT'	Unmapped reads (%)
C. elegans	99.99	0.00	**0.00**	**0.00**	0.01
H. sapiens	0.02	99.33	**0.00**	**0.00**	0.66
P. aeruginosa	0.29	0.00	99.66	0.03	0.02
E. coli	0.72	0.22	0.06	98.64	0.35

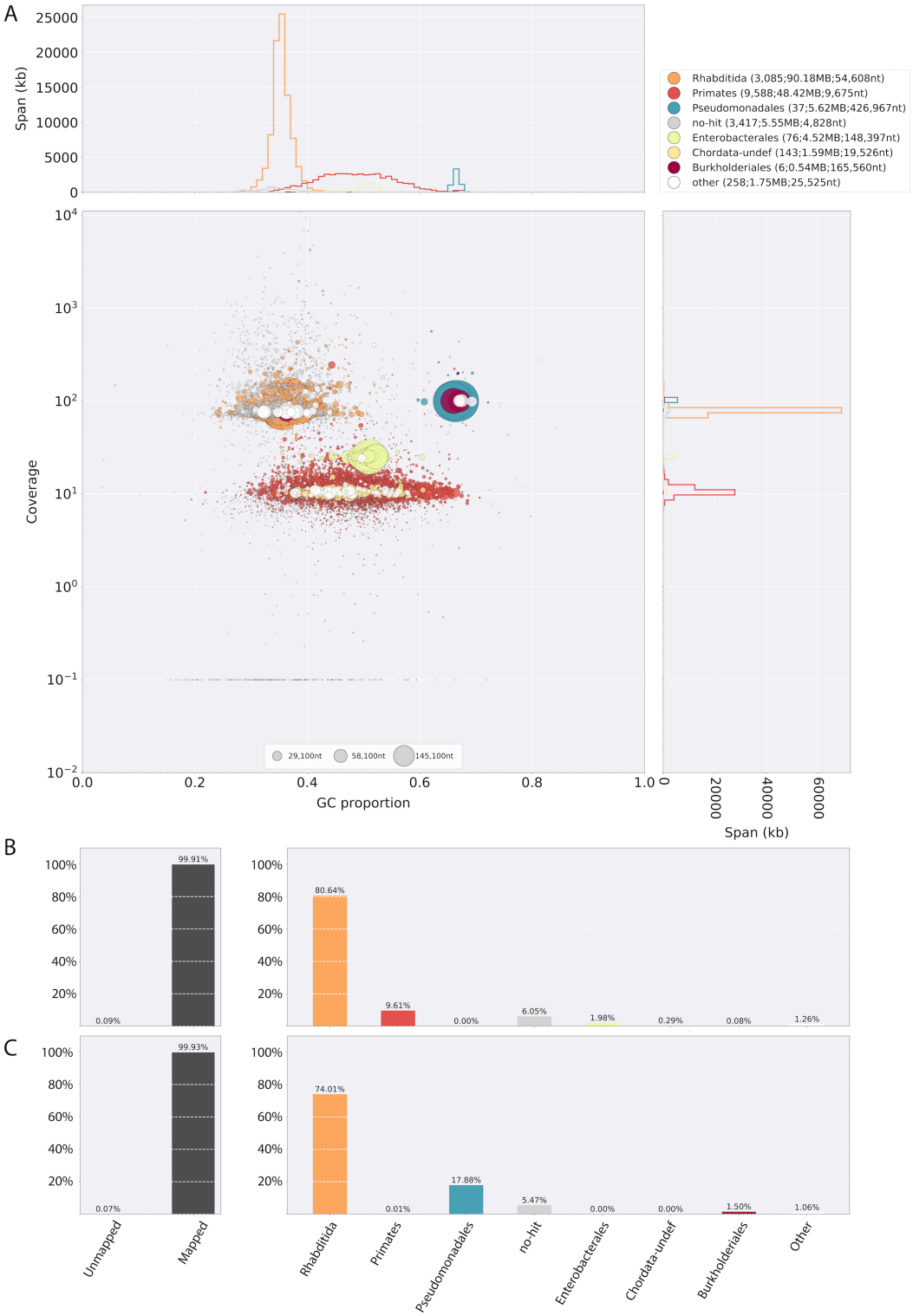


Figure 2.3.2: Visualisations of assembly of simulated sequencing libraries. **A** BlobPlot of the assembly. Sequences of the four taxa are recovered as distinct 'blobs'. **B** ReadCovPlot of library A. **C** ReadCovPlot of library B.

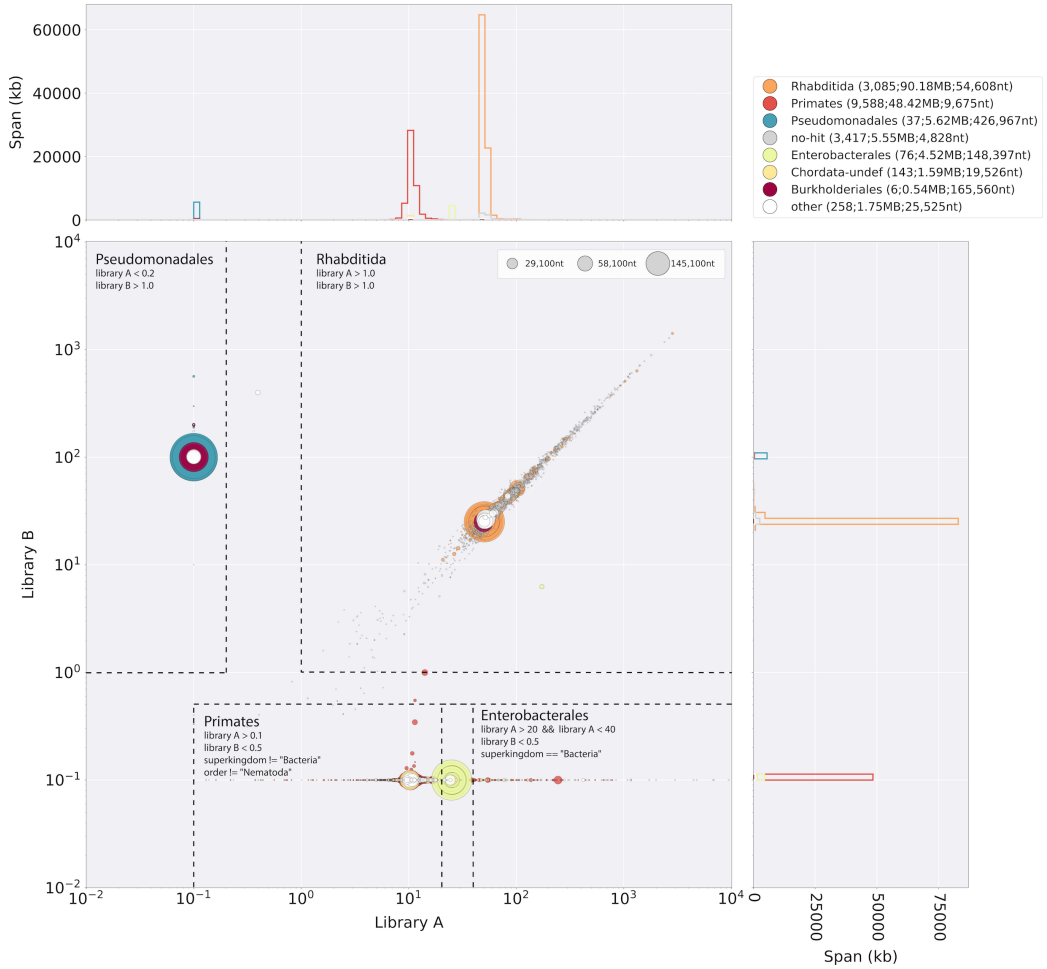


Figure 2.3.3: CovPlot of assembly of simulated sequencing libraries. Parameters for partitioning the sequences in the assembly which were applied to the tabular representation of the BlobDB are indicated as dotted lines and text annotations in the scatter plot.

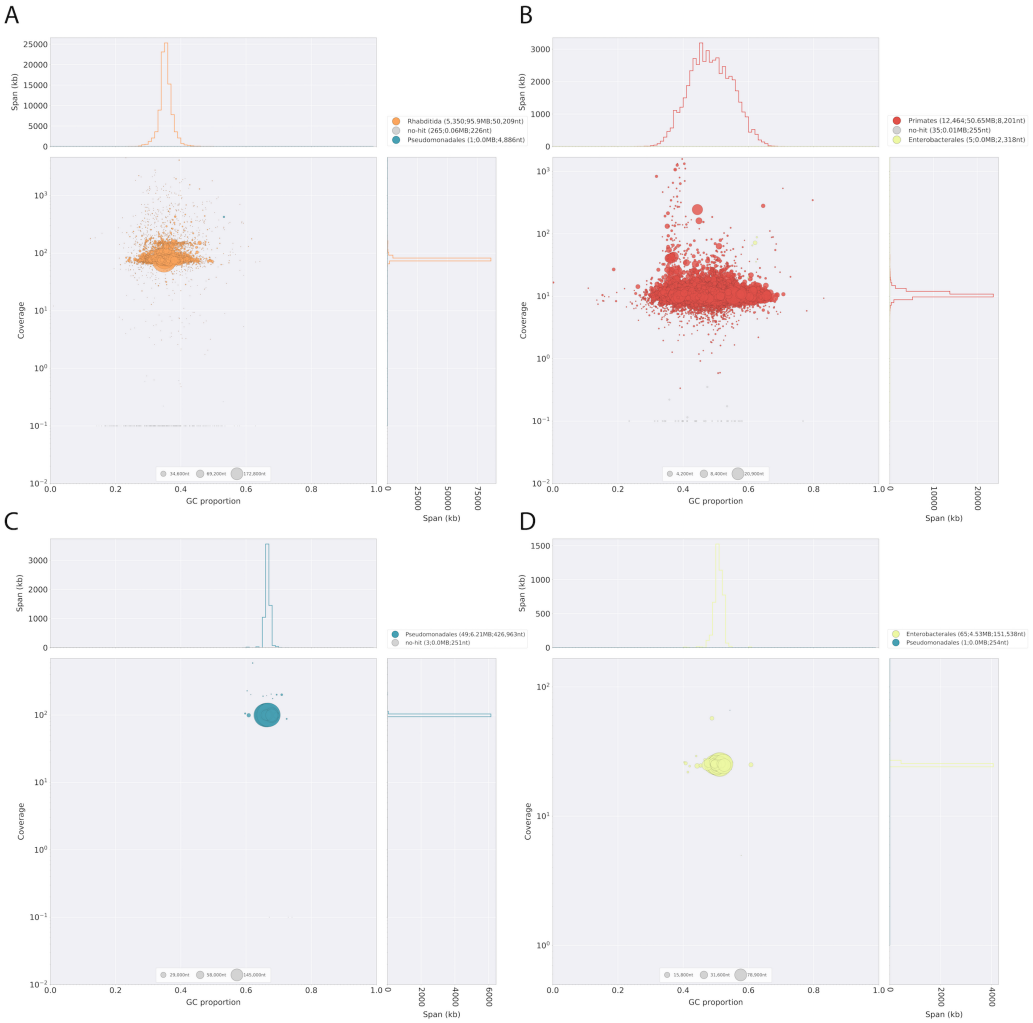


Figure 2.3.4: BlobPlots of assemblies by taxon after read partitioning. Coverage was obtained by mapping simulated reads to assemblies. Sequences are taxonomically annotated with 'true' taxonomy based on origin of simulated reads mapping to them. Sequences labelled as 'no-hit' did not receive any reads mapping to them. **A** Assembly of partition of Rhabditida reads ('RH-BT'). One *P. aeruginosa* sequence (span 4.886 b) remains. **B** Assembly of partition of Primates reads ('PR-BT'). Five *E. coli* sequences (total span 3.838 b) remain. **C** Assembly of partition of Pseudomonadales reads ('PS-BT'). **D** Assembly of partition of Enterobacterales reads ('EN-BT'). One sequence of *P. aeruginosa* (span 254 b) remains.

Table 2.3.3: Metrics of genome assemblies. Metrics of reference genomes (suffix '-REF'), assemblies generated from simulated reads by taxon (suffix '-SIM') and assemblies generated from reads partitioned using BlobTools pipeline (suffix '-BT').

Metric	CELEG-REF	CELEG-SIM	'RH-BT'	HSAPI-REF	HSAPI-SIM	'PR-BT'	PAERU-REF	PAERU-SIM	'PS-BT'	ECOLI-REF	ECOLI-SIM	'EN-BT'
Span (b)	100,286,401	95,970,640	95,964,660	58,634,185	50,765,888	50,660,776	6,264,404	6,221,846	6,215,193	4,636,831	4,561,104	4,534,517
Count	7	5536	5616	2	12,700	12,504	1	58	52	1	87	66
N50 (b)	17,493,829	51,178	50,209	58,617,616	8186	8200	6264,404	333,929	426,963	4,636,831	148,391	151,538
GC (%)	35.4	35.4	35.4	47.9	48.4	48.4	66.6	66.6	66.6	50.8	50.7	50.7
BUSCO (Complete, single copy in %)	97.8	92.7	92.8	3.1	1.5	1.6	98.2	98.2	98.2	99.5	99.5	99.2
BUSCO (Complete, duplicated in %)	0.6	0.4	0.4	0.1	0	0	0.2	0.4	0.4	0	0	0
BUSCO (Fragmented in %)	0.8	5.5	4.6	0.6	1	1	0	0	0	0.4	0.4	0.6
BUSCO (Missing in %)	0.8	1.4	2.2	96.2	97.5	97.4	1.6	1.4	1.4	0.1	0.1	0.2

2.3.4 Conclusion

The use of simulated read datasets allowed evaluation of the influence of different sequence similarity search strategies on the taxonomic annotation and revealed an optimal combination of search parameters for BlobTools taxonomic annotation: combining `BLASTn megablast(-evalue 1e-25 -max-target-seqs 1 -max_hsps 1)` searches against NCBI nt and `Diamond blastx (--evalue 1e-25 --max-target-seqs 1)` searches against UniProt ReferenceProteomes. While not generating the 'best' taxonomic annotation, search time is reduced substantially with a minor sacrifice in recall.

Taxonomic partitioning of the simulated read libraries proved successful. Only minor proportions of the reads were erroneously binned as revealed by the evaluation of read mappings by genome-of-origin to the final assemblies (Table 2.3.2). The highest percentage of erroneously partitioned reads (0.72% of reads mapping to the 'RH-BT' assembly) originated from the *E. coli* genome and the number is most likely inflated due to *E. coli* genome fragments contained in the *C. elegans* genome. The highest proportion of unbinned reads (0.66% of reads not mapping to any assembly) originated from *H. sapiens*, and could have been prevented by being more inclusive during the partitioning step of reads (by also including 'InUn' read pairs) followed by a second iteration of BlobTools workflow A. However, the small fraction of read pairs that received an erroneous taxonomic assignment or were left out during the partitioning step had little effect on the overall assembly success for each taxon (Table 2.3.3), as metrics of assemblies from simulated reads by taxon ('-SIM') are very similar to those of assemblies generated from reads partitioned using the BlobTools pipeline ('-BT').

2.4 Use case 2: BlobTools analysis of *Hypsibius dujardini* assemblies

2.4.1 Introduction

Tardigrades are meiofaunal animals within the superphylum Ecdysozoa. They have attracted the interest of the scientific community due to their unresolved phylogenetic position within Ecdysozoa and the ability of some species to withstand extreme conditions by transitioning into an ametabolic state, a process refered to as cryptobiosis (see Yoshida et al., 2017a and references therein).

Boothby et al., 2015 published a study claiming that 17.5% of the genes within the genome of the tardigrade *Hypsibius dujardini* originated from horizontal gene transfer (HGT) from multiple metazoan and bacterial taxa. The magnitude of the proportion of genes originating from HGT events was unprecedented and is almost double the proportion found in the most extreme case reported within the animal kingdom, the bdelloid rotifer *Adineta vaga*, which has acquired 9.6% of the genes in its genome through HGT (Boschetti et al., 2012).

Within eight days, collaborators and I published a rebuttal to Boothby et al., 2015 on *BioRxiv* (Koutsovoulos et al., 2015), followed by a peer-reviewed article (Koutsovoulos et al., 2016) in the same journal as the original study. Based on an independent genome assembly from a subculture of the same strain of *H. dujardini* we could attribute the inflated estimate of HGTs to non-tardigrade sequences in the assembly of Boothby et al., 2015. The original claim was also robustly challenged by other research groups using independent approaches and sequencing data (Delmont and Eren, 2016; Arakawa, 2016; Arakawa, Yoshida, and Tomita, 2016; Bemm et al., 2016). A correction to the original article, stating

that an outdated version of the genome assembly was provided in error, and a reply to Arakawa, 2016 and Bemm et al., 2016, reducing the rate of HGT to 3.8 – 7.1%, were published by the authors (Boothby et al., 2016; Boothby and Goldstein, 2016). The original paper was not retracted. The question of the proportion of genes originating from HGTs in the *H. dujardini* genome was eventually settled through a comparative genomics study using an improved assembly of *H. dujardini* (Yoshida et al., 2017a) which suggests that less than 2.3% of genes originate from HGT events.

Here, I illustrate how BlobTools allowed visualisation of the different assemblies of *H. dujardini* through BlobPlots, RNAseq-based coloured BlobPlots, and CovPlots.

2.4.2 Methods

Data

A preliminary assembly of *H. dujardini* was cleaned using `BlobTools v0.9.4` by G. Koutsovoulos and myself, as described in Koutsovoulos et al., 2016, resulting in the final assembly 'nHd.2.3'. The Boothby et al., 2015 assembly ('UNC', University of North Carolina) was obtained from the authors, together with the short insert size WGS read libraries used in the assembly. Additional read libraries used in Boothby et al., 2015 were not considered. RNAseq reads (poly(A)-selected) were obtained from Levin et al., 2016.

BlobTools analysis

I performed sequence similarity searches and read mappings of WGS and RNAseq reads, as described in Koutsovoulos et al., 2016. Coverage information of reads used for assembly 'nHd.2.3' is referred to as 'nHd reads' and coverage information of all three short read libraries used in the 'UNC' assembly by Boothby et al., 2015 is referred to as 'UNC reads'. I generated BlobDBs, coverage information, and category colour ('catcolour') files as described in Koutsovoulos et al., 2016. In 'catcolour' (CSV) files, sequences of an assembly are grouped into user-defined categories which can be used for colouring BlobPlots. For the analysis in Koutsovoulos et al., 2016, sequences were grouped into four categories, '>100 base cov', '10-99 base cov', '1-9 base cov' or '0 base cov', based on the normalised base coverage they received from the RNAseq reads mapping to them. BlobPlots, CovPlots and RNAseq-based coloured BlobPlots were generated for both assemblies using `BlobTools v1.0` based on the data generated for Koutsovoulos et al., 2016. BlobPlots and CovPlots were generated at the taxonomic ranks of phylum and superkingdom, respectively.

2.4.3 Results

The BlobPlots for both assemblies are depicted in Figure 2.4.1. It should be noted that, at the time, few genomic sequences for Tardigrada were available in public databases and therefore taxonomic annotation assigned tardigrade sequences to their phylogenetically closest and best sampled phyla, *i. e.* Nematoda, Chordata, Arthropoda and Mollusca. While minor contamination remains in the 'nHd.2.3' assembly (Figure 2.4.1A), the level of contamination in the 'UNC' assembly (Figure 2.4.1B) is substantial: roughly one third of the 'UNC' assembly is derived from

non-tardigrade genomes, mainly Proteobacteria and Bacteriodetes which also account for the largest scaffolds in the assembly.

RNAseq-based coloured BlobBlots (Figure 2.4.2) revealed further evidence for a non-tardigrade origin of many scaffolds in the 'UNC' assembly. Since poly(A)-selected RNAseq libraries fail to capture bacterial and archaeal mRNA, non-eukaryotic sequences in an assembly receive little to no base coverage. Out of the 135 Mb in the 'nHd.2.3' assembly (Figure 2.4.2A), 8.7 Mb received a base coverage of zero, which can be due to either remaining non-tardigrade sequences or non-coding regions of the *H. dujardini* genome. The 'UNC' assembly displays 94.2 Mb (out of 252.5 Mb) which receive no base coverage from RNAseq data and overlap with the sequences labelled as contaminants in Figure 2.4.1B.

CovPlots for both assemblies are shown in Figure 2.4.3. Patterns of differential coverage between the read sets ('nHd reads' and 'UNC reads') allow identification of sequences unique to each library, which are unlikely to be part of the *H. dujardini* genome. Remaining bacterial sequences in the 'nHd' assembly (Figure 2.4.3A) could thus be identified easily, as they did not receive any coverage from 'UNC reads'. However, a proportion of bacterial sequences occur at the same coverage as eukaryotic sequences, which suggests that these could be sequences harbouring HGT genes which were incorrectly taxonomically annotated. The CovPlot of the 'UNC' assembly (Figure 2.4.3B) allows a glimpse into the contamination landscape of the read datasets. Substantial proportions of the assembly received coverage from either 'UNC reads' or 'nHd reads', suggesting they are private to each dataset and possible lab-specific contaminants. Sequences which only received coverage from 'nHd reads' have most likely been assembled from other read libraries not considered here. Some sequences in the 'UNC' assembly are found at low coverages in both datasets, which suggests that the underlying organisms are biologically associated with *H. dujardini*, such as food sources, commensals, or pathogens.

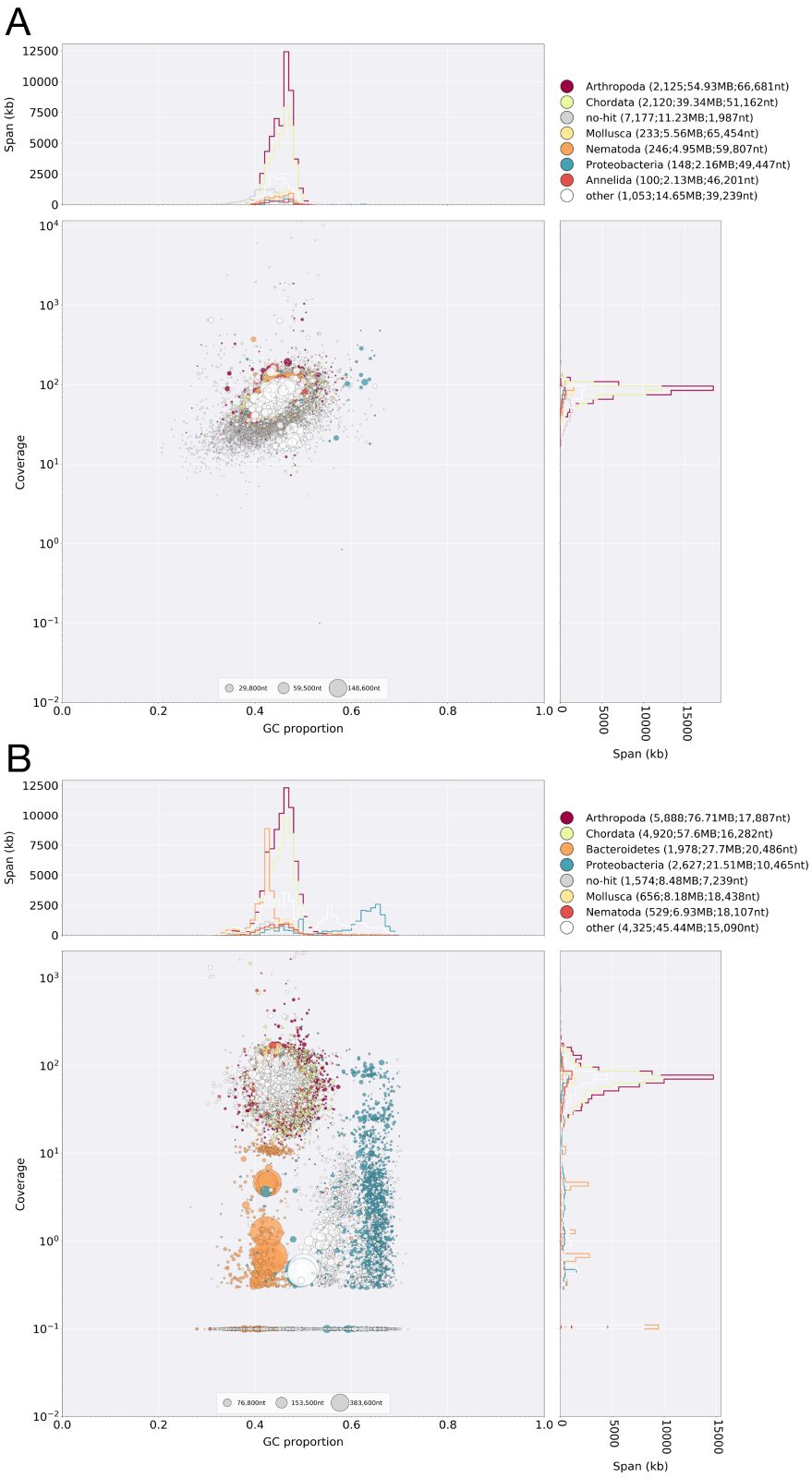


Figure 2.4.1: BlobPlots of tardigrade assemblies. **A** BlobPlot of the 'nHd.2.3' assembly with coverage information from 'nHd reads'. **B** BlobPlot of the 'UNC' assembly with coverage information from 'UNC reads'. The BlobPlot displays high amounts bacterial sequences.

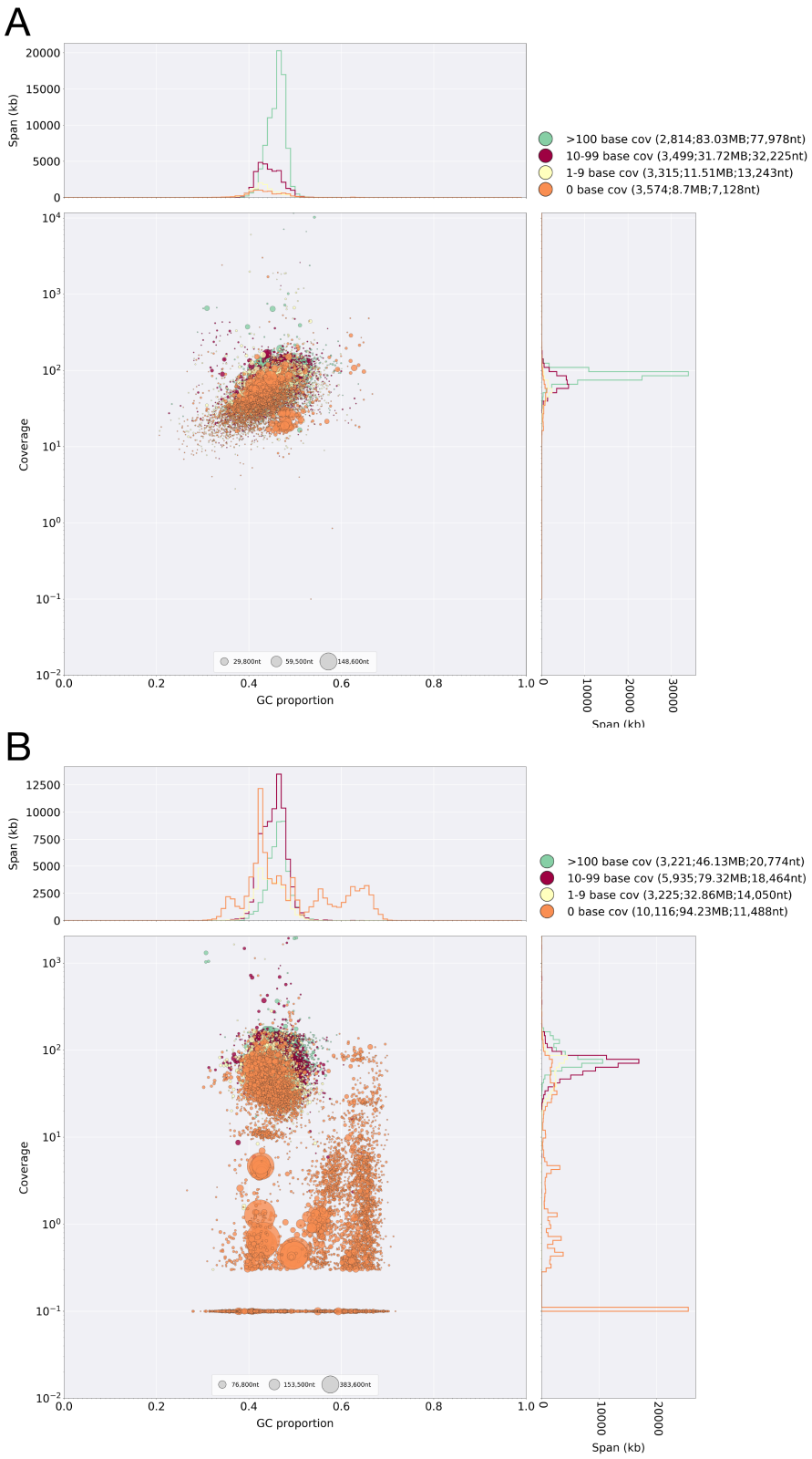


Figure 2.4.2: BlobPlots of tardigrade assemblies coloured by categories of normalised RNAseq base coverage. **A** BlobPlot of the 'nHd.2.3'. **B** BlobPlot of the 'UNC' assembly, showing low RNAseq base coverage for sequences labelled as bacterial in Figure 2.4.1.

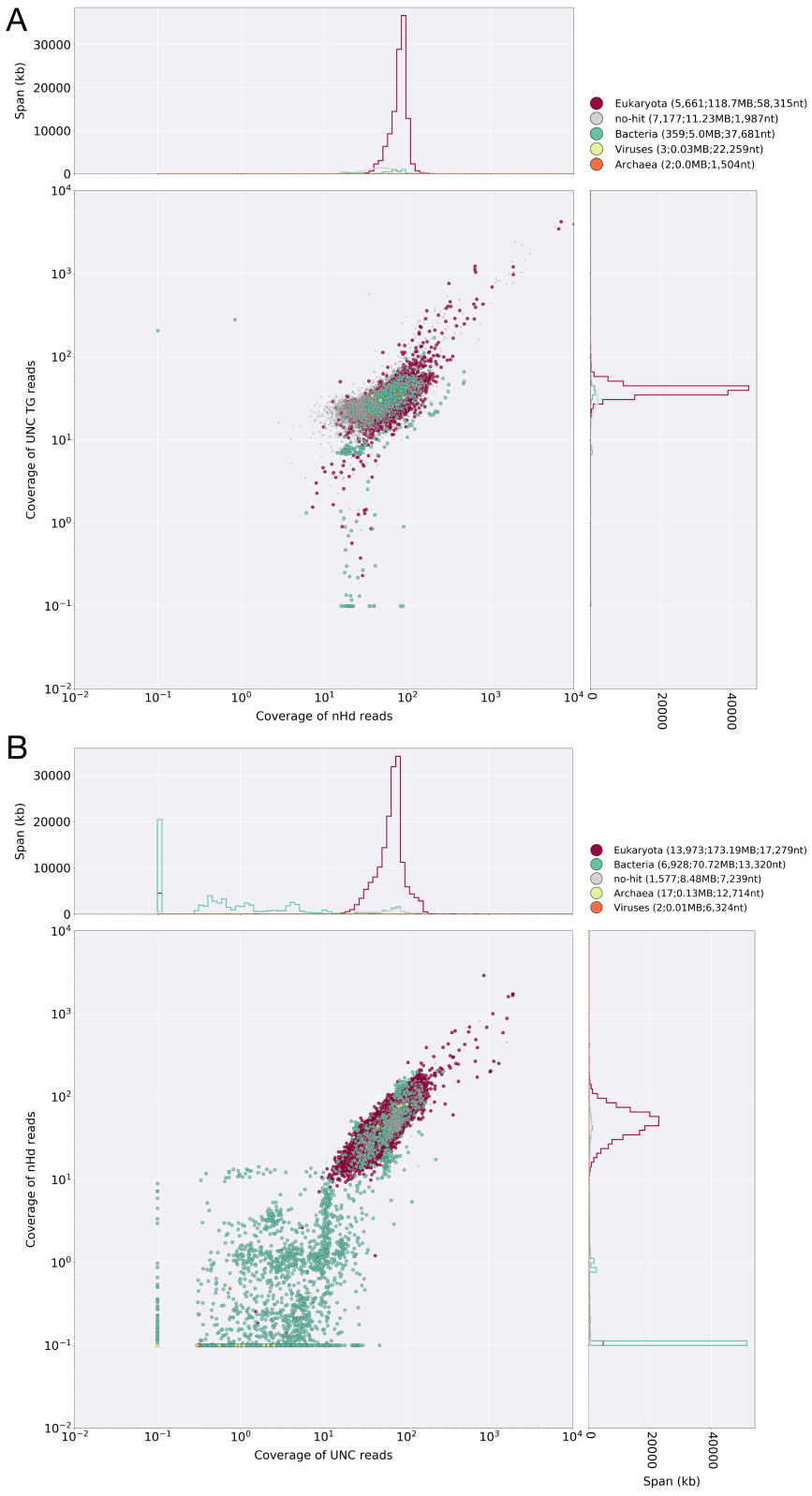


Figure 2.4.3: CovPlots of tardigrade assemblies. Scaling of diameters of circles based on sequence length was set to 'False'. **A** CovPlot of the 'nHd.2.3' assembly using coverage information of 'nHd reads' on the x-axis and of 'UNC reads' on the y-axis. **B** CovPlot of the 'UNC' assembly using coverage information of 'UNC reads' on the x-axis and of 'nHd reads' on the y-axis.

2.4.4 Conclusion

The comparative analysis of the two alternative hypotheses concerning the *H. dujardini* genome (Koutsovoulos et al., 2016), highlighted the strength of the modular approach of BlobTools as it allowed me to quickly incorporate new functions as the need for them arose: CovPlots and the `--catcolour` option of plotting functions were developed during this study. The `--catcolour` option is a flexible feature of BlobTools as it allows direct control over the 'colour' dimension of plots based on any grouping defined by the user, *e. g.* expression data, counts of predicted genes (with introns), etc. Furthermore, the controversy surrounding the genome of *H. dujardini* emphasised the need of assembly interrogation tools focussed on usability to ease adoption by the research community.

2.5 Use case 3: BlobTools analysis of *Globodera rostochiensis* assembly

2.5.1 Introduction

The yellow potato cyst nematode *Globodera rostochiensis* is an important pathogen of potato crops (Hockland et al., 2012). A genome project was initiated in order to understand the genomic differences and similarities between *G. rostochiensis* and its sister species *G. pallida*. I was involved in the analysis of the genome (Eves-van den Akker et al., 2016b), which is discussed in detail in Chapter 5.

Here, I describe how the assembly of *G. rostochiensis* was screened for contamination using BlobTools prior to further analysis by collaborators and myself. Sequence similarity searches for taxonomic annotation were performed against both public and custom sequence databases to assure that non-nematode sequences were removed prior to downstream analysis.

2.5.2 Methods

Data

The genome was assembled by the Wellcome Trust Sanger Institute as described in (Eves-van den Akker et al., 2016b), based on three WGS short read libraries: one PE and two MP read datasets.

Read mapping

I mapped read libraries to the assembly using `CLC mapper v4.21`, requiring an alignment identity of 80% along 80% of the length of reads (`-s 0.8 -l 0.8`) which resulted in three files in CAS format.

Sequence similarity searches for taxonomic annotation

I performed three sequence similarity searches of the assembly: one search against NCBI nt using `BLASTn megablast v2.3.0+` (Camacho et al., 2009) (`-evalue 1e-65` and `-max_target_seqs 1`), one search against UniProt Reference Clusters 90 (Uniref90) (Suzek et al., 2015) using `Diamond blastx v0.7.12` (Buchfink, Xie, and Huson, 2015) (`--sensitive` and `--max-target-seqs 25`), and one search against a custom sequence database of the genome assembly of *G. pallida* (Cotton et al., 2014) using `BLASTn megablast v2.3.0+` (`-evalue 1e-65` and `-max_target_seqs 1`).

BlobTools analysis

Using `Blobtools v0.9.9`, I constructed a BlobDB using the assembly, the CAS mapping files of the three read datasets, and the similarity search results provided in the order as listed above under the taxrule 'bestsum'. Diamond results were annotated with NCBI TaxIDs based on the UniProtID mapping file (retrieved from the UniProt website) using the BlobTools command `taxify`. BlobPlots were drawn at the taxonomic rank of phylum using the cumulative coverage of all three read libraries. Using tabular output of BlobTools `view`, taxonomically annotated non-nematode scaffolds with a bitscore $\geq$200 were inspected manually using the NCBI

BLAST web service and removed if evidence for non-nematode origin was found. A second BlobPlot, after removal of contaminant scaffolds and remapping of read datasets, was generated.

Screening for rDNA sequences in filtered assembly

I interrogated the filtered assembly for small subunit (SSU) and large subunit (LSU) rDNA sequences, by carrying out sequence similarity searches using `BLASTn megablast v2.3.0+ (--evalue 1e-65)` against the SILVA SSUParc/LSUParc databases (Quast et al., 2013) composed of quality checked and aligned ribosomal RNA sequence, which were translated into rDNA sequences prior to searches.

2.5.3 Results

The BlobPlots of the assembly of *G. rostochiensis* are shown in Figure 2.5.1. The unprocessed assembly (Figure 2.5.1A) already showed minimal levels of contamination, suggesting that the initial assembly was already processed to remove non-nematode sequences. Nevertheless, 23 bacterial and fungal scaffolds (cumulative length of 98.2 kb) were removed from the assembly. The filtered assembly (Figure 2.5.1B) displays minor differences in coverage compared to the unprocessed assembly, due to stochasticity of the read mapping process. Screening of the assembly for rDNA sequences revealed no evidence for additional contaminants, as hits were only observed against SSU and LSU sequences of *G. rostochiensis*.

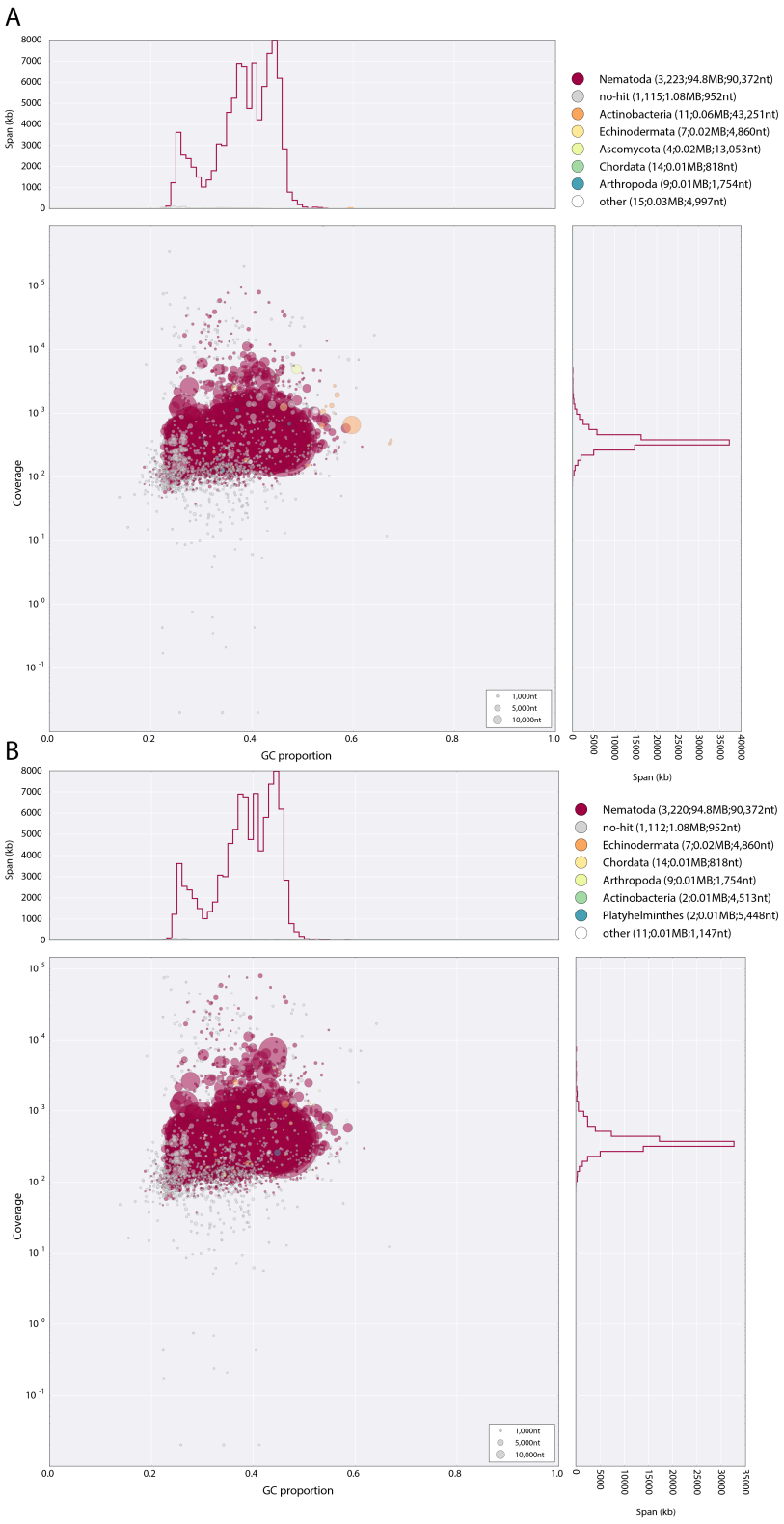


Figure 2.5.1: BlobPlots of the *G. rostochiensis* assembly. **A** BlobPlot of the unfiltered assembly. **B** BlobPlot of the assembly after filtering.

2.5.4 Conclusion

This use case illustrates how BlobTools can be used for rapid screening of genome assemblies and contaminant removal, even if read partitioning and subsequent reassembly is not carried out. The minor levels of contamination in the initial *G. rostochiensis* assembly did not warrant read partitioning, as the small amount of non-nematode sequences are unlikely to have affected the assembly process. The use of multiple sequence similarity search results for taxonomic annotation allows the incorporation of evidence from multiple sources and lets the user, through the tabular output of BlobTools 'view', inspect the scores for each taxonomy by database. This feature was exploited here to manually inspect possible contaminants.

In retrospect, the use the Uniref90 database for taxonomic annotation was a suboptimal choice, as it contains entries derived from non-redundant sequences in the UniProt Knowledgebase (The UniProt Consortium, 2017) and selected UniParc sequences (a non-redundant archive of most protein sequences from public databases) clustered using the CD-HIT algorithm (Li and Godzik, 2006) at 90% sequence identity over 80% of the length of the longest sequence. A Uniref90 cluster is given the NCBI TaxID of the common ancestor of all sequences it contains. As an example, entry 'UniRef90_O17915' (GTP-binding nuclear protein ran-1) is composed of sequences from 38 nematode taxa (of which 28 originate from animal parasites) in addition to one partially predicted protein from the genome of the alpaca, *Vicugna pacos*. The TaxID associated with the entry is Bilateria (33213), although the alpaca sequence most likely originated from contamination. In this analysis, the low value of sequence similarity searches against Uniref90 entries for taxonomic annotation was compensated by the other sequence similarity search results, but its use for BlobTools taxonomic annotation is not encouraged.

2.6 BlobTools improves the genome assembly process

In this chapter, I have presented the BlobTools pipeline and discussed its implementation. BlobTools workflows and features were highlighted based on three use cases. By analysing simulated read libraries of mixtures of bacterial and metazoan taxa (Section 2.3), the performance of the BlobTools pipeline for taxonomic annotation of sequences and subsequent partitioning of PE read datasets could be evaluated empirically. In addition, evaluation of parameters for sequence similarity searches used in BlobTools taxonomic annotation revealed optimal combinations of parameters, leveraging computational costs and accuracy. Comparative analysis of genome assemblies of the same taxon were presented in Section 2.4, based on visualisation of alternative genome assemblies of the tardigrade *H. dujardini* using BlobPlots, RNAseq-based coloured BlobPlots and CovPlots. In Section 2.5, I described a simple BlobTools analysis for identification of contaminants in a draft genome assembly.

The ease of interpretation of BlobPlots has favoured adoption by users, and the current implementation has been applied successfully to genome projects involving tardigrades (Koutsovoulos et al., 2016; Yoshida et al., 2017a), mealybugs and their endosymbionts (Husnik and McCutcheon, 2016), ectoparasitic mites (Dong et al., 2017), diptera (Dikow et al., 2017), honeybees and their metagenomes (Gerth and Hurst, 2017), nematodes (Eves-van den Akker et al., 2016b; Gawryluk et al., 2016; Slos et al., 2017; Szitenberg et al., 2017), bacteria (Mellbye et al., 2017; Samad et al., 2016; Wang and Chandler, 2016; Fuller et al., 2017), butterflies (Nowell et al., 2017), a fungal pathogen of barley (McGrann et al., 2016), and fungi (Compant et al., 2017).

Its modular interface and reliance on standard bioinformatic input formats

has lead to the integration of BlobTools into the Edinburgh Genomics QC pipeline and an assembly/QC pipeline at the University of Exeter (Leonard, 2017). Furthermore, my work on BlobTools contributed to the award of the BBSRC Research grant 'BlobToolKit: Identification and analysis of non-target data in all Eukaryotic genome projects' (Project reference BB/P024238/1), aimed at improving BlobTools and offering access through a free web-service (for more information, see `http://blobtoolkit.genomehubs.org/`). The BlobTools code base will be developed further under the umbrella of the BlobToolKit project and improvements are planned for the underlying BlobDB data structure and the process of taxonomic annotation.

Currently, BlobTools stores the information parsed from input files in Python classes which are subsequently translated (serialised) into a JSON (JavaScript Object Notation) file, termed BlobDB. Subsequent access to the data by other commands, *e. g.* for generating a BlobPlot, requires deserialisation of the entire BlobDB file which can takes several minutes if large amounts of sequence similarity searches were provided. Redesign of the BlobTools data structure and refactoring of the code would allow the use of a SQL (Structured Query Language) database, which would improve runtime and minimise file size. Python libraries for interaction with SQL databases are freely available, such as SQLAlchemy (see `https://www.sqlalchemy.org/`). Interaction of legacy BlobDBs would be guaranteed through appropriate conversion functions. This novel BlobDB would streamline interaction with the planned BlobToolKit web service, since portions of the data can be accessed rapidly for both export to plain text files and visualisation. It would improve the user experience through reduced runtime and simplify development of novel computations on the data as new types of data can readily be added with minor changes to the code base.

Several improvements to the process of taxonomic annotation are planned. For

one, an important aspect of sequence similarity search results is currently ignored: the position of hits across the length of a query. At present, summation of scores of hits is carried out across the whole length of the query sequence. By making BlobTools aware of 'regions' on the sequences in an assembly, new taxrules could be developed which compute scores of hits to competing taxonomies differently depending on the region on the query. For instance, a taxrule could be developed to consider taxonomic 'homogeneity' of hits across the length of the sequence. A locus on a sequence which receives many contradictory taxonomic annotations, due to being a highly conserved region sequenced in many taxa, could thus be assigned less weight during score computation compared to other regions on that sequence which receive fewer but more taxonomically homogenous hits. This could also be developed further to address issues of HGT and chimeric contigs/scaffolds due to errors in the assembly process. Secondly, taxonomic annotation is currently performed independently at each taxonomic rank. This can lead to edge case scenarios where, based on the sum of scores, taxonomy for the same sequence varies greatly between ranks. This could be mitigated by developing a 'root-tip' taxrule which is aware of the taxonomic hierarchy of ranks. Lastly and in sync with the goals set for the BlobToolKit project, BlobTools will be distributed with a 'black list' of sequence IDs in public databases for which taxonomic annotation has been deemed dubious in past studies.

In summary, BlobTools is a user-friendly and reliable solution for visualisation, quality control and taxonomic partitioning of genome datasets. Wider adoption of BlobTools screening by the research community will help control the influx of taxonomically mis-annotated sequences into public sequence databases and prevent inaccurate biological conclusions based on contaminated genome assemblies. Planned developments within the BlobToolKit project will improve user experience and the process of taxonomic annotation.

Chapter 3

KinFin: software for the analysis of protein families

"Science is what we understand well enough to explain to a computer.

Art is everything else we do."

- Donald E. Knuth, *Foreword to the book 'A=B' (1996)*

3.1 Introduction

In comparative genomics it is now a common approach to define gene families by clustering protein sequences — *i. e.* all proteins of the proteomes of the organisms under analysis — based on sequence similarity, and to analyse protein cluster presence and absence in different species groups as a guide to biology. Due to the high dimensionality of these data, downstream analysis of protein clusters inferred

from large numbers of species or from species with many genes is non-trivial and few solutions exist for transparent, reproducible and customisable analyses.

Several high-quality solutions to orthology analysis have been proposed. OrthoDB is a high-quality curated orthology resource (Zdobnov et al., 2017). The current release (2015) includes 3600 bacterial and 590 eukaryotic taxa, and is accessed through a responsive web interface for direct download and interrogation of clusters. OrthoDB includes rich functional annotation of sequences. While the main database includes only published genomes and is centrally managed (*i. e.* users cannot submit datasets for analysis), the OrthoDb software toolkit is available for local installation and deployment. PhylomeDB is a database of defined orthology groups, built with manual curation (Huerta-Cepas et al., 2014), but was last updated in 2014, and is, again, managed centrally and focussed on published genomes. In the ENSEMBL databases, the Compara toolkit is used to parse gene homology sets, and infers orthology and paralogy based on a given species tree (Herrero et al., 2016). Updating of Compara analyses is not trivial, and requires the ENSEMBL web toolkit for display and interrogation. For ongoing research programs, few tools for orthology analysis are available. For bacterial data, several tools for pan-genome analysis have been developed (Vinuesa and Contreras-Moreira, 2015; Chaudhari, Gupta, and Dutta, 2016; Xiao et al., 2015) but solutions that cope well with the data richness of eukaryotic species are often tailored to defined taxonomic groups (Song et al., 2015) or expect closely related taxa. `EUPAN` is a pipeline for pan-genome analysis of closely related eukaryotic genomes developed within the scope of the '3000 Rice Genomes Project' (Hu et al., 2017). The approach is based on mapping of raw reads to reference genomes, followed by coordinated assembly and lift-over of gene annotations for inferring presence/absence of gene models.

In the absence of toolkits that allow local implementation of clustering analyses, custom taxon grouping and dynamic analysis, I have developed KinFin. KinFin takes a protein clustering output by tools such as `OrthoFinder` or `OrthoMCL`, alongside functional annotation data, and user-defined species taxonomies, and derives rich aggregative annotation of orthology groups. KinFin reads from standard file formats commonly produced along genome sequencing and annotation projects and can therefore easily be integrated in comparative genomics projects for the identification of protein clusters of interest in user-defined, taxon-aware contexts.

Within this chapter, four use cases for KinFin are presented in which gene families of different taxa are investigated. Parts of this chapter have been published as an article on the bioRxiv pre-print server (Laetsch and Blaxter, 2017b) (DOI: 10.1101/159145) and have been accepted for publication to the journal 'G3: Genes, Genomes, Genetics' (Laetsch and Blaxter, 2017c). Furthermore, KinFin was used to analyse patterns of gene family evolution across ecdysozoan phyla focussing on tardigrades which has been published in the journal 'PLOS Biology' (Yoshida et al., 2017b) (DOI: 10.1371/journal.pbio.2002266).

3.2 Implementation

KinFin is a standalone Python 2.7 application. A detailed description of the functionality of KinFin can be found at `https://kinfin.readme.io/`. Required input for KinFin is an orthology clustering (format defined by `OrthoMCL/OrthoFinder`), a file linking protein sequences to taxa (SequenceIDs defined by `OrthoFinder`), and a config file. The config file guides analyses by grouping taxa into user-defined sets under arbitrary attributes. These attributes could include, for instance, standard taxonomy (as embodied in the NCBI Taxonomy TaxIDs), alternative systematic arrangements of the taxa involved, lifestyle,

geographical source or any other aspect of phenotype or other metadata. KinFin dynamically constructs sets based on the config file and computes metrics, statistics and visualisations which allow identification of clusters that are diagnostic for, or expanded/contracted in, each taxon set. Optional input files include proteome FASTA files (to extract length statistics for clusters, taxa and taxon sets), functional annotations of proteins in `InterProScan` (Jones et al., 2014) format, and a phylogenetic tree topology in Newick format.

3.2.1 Visualisation of orthologue clustering

In KinFin, global analysis of the clustering of protein sequences can be performed from the point of view of the clusters themselves (their properties and patterns) or of the constituent proteomes. The distribution of cluster size, *i. e.* the number of proteins contained in a cluster, is an important feature of analyses, and KinFin simplifies the comparison of alternative clusterings, *e. g.* clusterings originating from different MCL (Markov Clustering) inflation parameters, or with overlapping but distinct taxon composition) by generating frequency histograms of cluster size. These can then be interrogated for deviations from the expected power-law-like distribution. To aid understanding of the distribution, the user can generate a more detailed frequency histogram which considers the number of taxa contributing to each cluster (for an example see Figure 3.3.1). The behaviour of individual proteomes can be explored by creating a network representation of the clustering. KinFin can produce a graph file with nodes representing proteomes and edges connecting nodes weighted by the number of times two proteomes co-occur in clusters. Optionally, universal clusters containing proteins from all proteomes can be excluded. The graph can be interrogated using graph analysis and visualisation tools such as Gephi (Bastian, Heymann, and Jacomy, 2009).

3.2.2 Analysis based on arbitrary sets of input proteomes

Through the config file, the user can instruct KinFin to analyse the clustering under arbitrary taxon sets, *i. e.* sets of proteomes. For taxonomy-based analyses, KinFin derives analyses at different taxonomic ranks (by default phylum, order, and genus; can be modified by the user) by parsing the NCBI TaxIDs given for each proteome. Any other classification of the input proteomes can be given, and nested taxonomies specified by use of multiple, ranked attribute types. This allows, for example, the testing of congruence of clustering data with competing phylogenetic hypotheses regarding relationships of the taxa from which the input proteomes were derived.

3.2.3 Classification of clusters

KinFin builds a series of matrices associating clusters and proteomes, and clusters and user defined taxon sets. Each cluster is classified as absent or present for each proteome or taxon set, and is assigned a cluster type:

- **singleton**: composed of a single protein
- **specific**: composed of multiple proteins from a single taxon set
- **shared**: composed of multiple proteins from multiple taxon sets

3.2.4 Single-copy orthologue definition

Clusters composed of a single protein from each proteome (*i. e.* putative single-copy orthologues) are useful for downstream phylogenetic analyses. However, due to the intrinsic difficulties of genome assembly and annotation, the number of single-copy orthologues decreases as more proteomes are included in the clustering. To compensate for this, KinFin can identify 'fuzzy' single-copy orthologue clusters using the parameters `--target_count` (target number of copies per proteome, default '1'), `--target_fraction` (proportion of proteomes at `--target_count`), and lower/upper counts for proteomes outside of `--target_fraction` (`--min` and `--max`).

3.2.5 Rarefaction curves

The concept of the pan-genome is frequently used in microbial genomics to delimit the core genome — composed of regions shared by all taxa — and the accessory genome — composed of regions shared by only some taxa — that are found in the varied genomes of a species. The size of the pan-genome can be visualised using rarefaction curves, and KinFin deploys this framework to visualise the size of the pan-proteome of the different arbitrary sets defined by the user. Curves are calculated by repeated, random sampling of the proteomes in each arbitrary set and cumulative summation of novel non-singleton clusters.

3.2.6 Pairwise protein count representation tests

For user-defined attributes involving two (or more) taxon sets, pairwise representation tests of protein counts are computed for clusters containing proteins from each

taxon set using either a two-sided Mann-Whitney U test (default), Welch's t-tests, Kolmogorov-Smirnov statistic, Kruskal-Wallis H-test, or a Student's t-test. From this, clusters 'enriched' or 'depleted' in count in one set compared to another can be identified. It should be noted that the statistical tests test for non-homogeneity of the distributions of protein counts between the sets and, due to limited 'sample size' (the number of proteins of different taxon sets within a cluster), might not achieve statistical significance even when counts differ substantially between sets. In addition to text outputs, volcano plots ($\log_2$-fold change in means versus test p-value) are drawn. As a visual aid, horizontal lines are drawn at p-values 0.05 and 0.01 and vertical lines at $|\log_2\text{-fc(means)}| = 1$ and 95%-percentile of $\log_2$-fc(means).

3.2.7 Functional annotation and protein length analysis

KinFin integrates functional annotation and protein length data into analyses. If the necessary input files are provided, KinFin generates output files tabulating mean and standard deviation of sequence lengths, domain and Gene Ontology (GO) term entropy within clusters, and the fraction of proteins per cluster which are putatively secreted, based on `SignalP_Euk` (Petersen et al., 2011) annotation. Additionally, for each cluster all matching domains and inferred GO terms are listed with description and information regarding their frequency within both proteins and proteomes in the cluster.

While inference of functional annotation of a protein is relatively straightforward, no defined standards exist for inferring representative functional annotation (RFA) of clusters of proteins. KinFin is distributed with a script that infers RFAs of clusters through the parameters `--domain_taxon_cov` (minimum fraction of taxa in cluster that have at least one protein annotated with a specific domain) and `--domain_protein_cov` (minimum fraction of proteins in cluster annotated

with a specific domain), to grant users fine control over cluster functional annotation.

3.2.8 Analysis based on phylogeny

Analysis of clusters in a phylogenetic context allows the identification and quantification of clusters that are unique innovations of certain monophyletic groups (*i. e.* synapomorphies). Based on a user-defined tree topology, KinFin identifies synapomorphic clusters at nodes using Dollo parsimony. The Dollo parsimony method (Farris, 1977) assumes that while multiple, independent losses of a gene in different lineages are common, multiple independent gains of the same gene are improbable. My implementation of Dollo parsimony for the identification of synapomorphic clusters requires that only the proteomes under a given node are members of the cluster and that at least one taxon from each child node is a member. Since Dollo parsimony does not penalise multiple losses, KinFin classifies synapomorphies into 'complete presence' and 'partial absence' subgroups. The output includes lists of synapomorphies and apomorphies ('singleton' and 'non-singleton' proteome-specific clusters) and detailed description of synapomorphic clusters at each node. Prominent or consistent functional annotation can be mapped onto synapomorphic clusters, filtered by the parameter `--node_taxon_cov` (minimum presence of proteomes as fraction of total proteomes under the node), and the parameters `--domain_taxon_cov` and `--domain_protein_cov` detailed in the previous Section.

3.2.9 Analyses of clusters containing genes of interest

Output of protein clustering analysis often serves as substrate for the identification of homologues of genes of interest from a model species in the target species. KinFin is distributed with a script which takes as input a list of protein IDs or gene IDs (to obtain all isoforms or only the isoforms included in the clustering) and writes tables indicating the counts of proteins in each cluster and their representative functional annotations.

3.2.10 Output

KinFin generates output folders for each relevant column in the config file and writes overall metrics for all taxon sets, detailed metrics for each cluster and results of pairwise representation test, draws the rarefaction curve and volcano plots, and lists clusters classified as 'true' and 'fuzzy' single-copy orthologues. Resulting text files can easily be interrogated using common UNIX command line tools or spreadsheet software.

3.2.11 Operation

KinFin is freely available under GNU General Public License v3.0 at `https://github.com/DRL/kinfin`. System requirements for KinFin include a `UNIX` based operating system, `Python 2.7`, and `pip`. An installation script is provided, which installs Python dependencies and downloads mapping files for Pfam, Interpro and GO IDs from European Bioinformatics Institute (EBI) website. Instructions for installation and execution of KinFin can be found on the GitHub repository and detailed documentation is available at `https://kinfin.readme.io`.

3.3 Use case 1: Analysis of gene families in filarial nematodes

3.3.1 Introduction

In order to illustrate some of the main functionalities of KinFin, I chose to address questions regarding the biology of filarial nematodes. Filarial nematodes (Onchocercidae) include many species of medical and veterinary interest and the phylogenetic relationships among them remain under debate (Park et al., 2011; Nadler et al., 2007), with the current NCBI reference taxonomy likely to be incorrect. I analysed the proteomes of 16 species: 11 filarial nematodes, three related spirurid nematodes and two *Caenorhabditis* species. *Caenorhabditis* species were included because of the quality of available structural and functional annotations. For three species, two independent assemblies and proteome predictions were included. I used KinFin to generate a robust multi-locus alignment and phylogeny, and then incorporated this tree into KinFin analyses of synapomorphies and other features of groups of filaria. Furthermore, using sets of *Caenorhabditis elegans* genes implicated in pathways of interest, I investigated orthology and paralogy within the filarial nematodes.

3.3.2 Methods

Protein clustering

Proteomes listed in Table 3.3.1 were downloaded from WormBase parasite (WBPS8) (Howe et al., 2016; Howe et al., 2017) and `https://ngenomes.org`. Protein files

were filtered by excluding sequences shorter than 30 residues or containing non-terminal stops (`filter_fastas_before_clustering.py`) and only the representative isoform for each gene was kept (`filter_isoforms_based_on_gff3.py`). Proteins were functionally annotated through `InterProScan v5.22-61.0` (Jones et al., 2014) using the `Pfam-30.0` database (Finn et al., 2016) and output was converted to compatible input format for KinFin (`iprs_to_table.py`). `OrthoFinder v1.1.4` (Emms and Kelly, 2015) was used to generate the commands for `BLASTp` analyses. `BLASTp` commands were further modified by adding the following options `-seg yes`, `-soft_masking true` and `-use_sw_tback` as suggested by Moreno-Hagelsieb and Latimer, 2008. `BLASTp` analyses were run on the EDDIE supercomputing cluster at the University of Edinburgh using `BLASTp v2.3.0+` (Camacho et al., 2009). Proteome clustering was carried out at default MCL inflation value of 1.5.

Table 3.3.1: Protein datasets used in clustering. Taxon ID: Identifiers used in KinFin analysis. Species: Taxonomic species name. Proteins: Number of representative isoforms included in the KinFin analysis.

TAXON ID	Species	Source	ID	Proteins
AVITE	*Acanthocheilonema viteae*	WBPS8	PRJEB4306	10,123
BMALA	*Brugia malayi*	WBPS8	PRJNA10729	11,008
BPAHA	*Brugia pahangi*	WBPS8	PRJEB497	14,664
CBRIG	*Caenorhabditis briggsae*	WBPS8	PRJNA10731	22,305
CELEG	*Caenorhabditis elegans*	WBPS8	PRJNA13758	20,219
DIMMI	*Dirofilaria immitis*	WBPS8	PRJEB1797	12,423
DMEDI	*Dracunculus medinensis*	WBPS8	PRJEB500	10,919
EELAP	*Elaeophora elaphi*	WBPS8	PRJEB502	10,409
LOA1	*Loa loa*	WBPS8	PRJNA246086	12,473

Table 3.3.1 Continued from previous page

TAXON ID	Species	Source	ID	Proteins
LOA2	*Loa loa*	WBPS8	PRJNA60051	14,908
LSIGM	*Litomosomoides sigmodontis*	WBPS8	PRJEB3075	10,001
OFLEX	*Onchocerca flexuosa*	WBPS8	PRJEB512	16,094
OOCHE1	*Onchocerca ochengi*	WBPS8	PRJEB1204	12,807
OOCHE2	*Onchocerca ochengi*	WBPS8	PRJEB1809	13,580
OVOLV	*Onchocerca volvulus*	WBPS8	PRJEB513	12,110
SLABI	*Setaria labiatopapillosa*	ngenomes.org	nSl.1.1	9687
TCALL	*Thelazia callipaeda*	WBPS8	PRJEB1205	10,911
WBANC1	*Wuchereria bancrofti*	WBPS8	PRJEB536	13,056
WBANC2	*Wuchereria bancrofti*	WBPS8	PRJNA275548	11,053

Phylogenetic analysis

An initial KinFin analysis identified 781 single-copy orthologues. Sequences for these 781 clusters were extracted (`get_protein_ids_from_cluster.py` and `GNU grep`) and aligned using `mafft v7.267` (E-INS-i algorithm) (Katoh and Standley, 2013). Alignments were trimmed using `trimal v1.4` (Capella-Gutiérrez, Silla-Martínez, and Gabaldón, 2009), concatenated using `FASconCAT v1.0` (Kück and Meusemann, 2010), and analysed using `RAxML v8.1.20` (Stamatakis, 2014) under the PROTGAMMAGTR model of sequence evolution and 20 alternative runs on distinct starting trees. Non-parametric bootstrap analysis was carried out for 100 replicates.

KinFin analysis

KinFin was then rerun, providing additional classification in the config file and functional annotation data. In the config file, taxon sets were defined for the taxonomic rank of 'order' by supplying NCBI TaxIDs for each proteome, for the attribute 'clade' by grouping taxa into taxon-sets for the major filarial clades, and for the attribute 'host' by separating human parasites from those of other animals and outgroups. For the attribute of 'clade', only one proteome per species was allocated to its respective taxon set (*i. e.* LOA2, OOCHE1, and WBANC2) and unique labels were specified for the remaining taxa. The topology of the tree inferred through phylogenetic analysis was provided in Newick format and the two *Caenorhabditis* species were specified as outgroups in the config file. The Mann-Whitney-U test was selected for pairwise protein count representation tests and the required number of proteomes in a taxon-set to be used in rarefaction/representation-test computations was set to '2'.

Representative functional annotation of clusters

Using `get_protein_ids_from_cluster.py`, representative functional annotation (RFA) was inferred for all clusters (`-x 0.75 -p 0.75`, requiring that 75% of proteins in a cluster share a domain and that 75% of proteomes have at least one protein with that domain) and for synapomorphic clusters (`-n 0.75 -x 0.75 -p 0.75`, with the additional requirement that also 75% of taxa at a node are present in the cluster).

Analysis of genes of interest

Genes involved in haem biosynthesis and homeostasis were identified based on representatives from *C. elegans*, and absence of missing genes was confirmed through `TBLASTn` (Camacho et al., 2009) against the respective genomes. The presence of paralogues was confirmed by manual inspection of gene models on WormBase ParaSite.

3.3.3 Results

The 19 proteomes (derived from 16 species) included 248,750 protein sequences (with a total length of 95,162,557 aa) after filtering. `OrthoFinder`, at the MCL inflation value of 1.5, placed these into 42,691 clusters, of which 57.97% were singletons (containing 9.95% of protein sequences). The clusters displayed a power-law like frequency distribution, but with a marked deviation from this expectation at a cluster size of 19, matching the number of proteomes in the analysis (Figure 3.3.1). This pattern, although less pronounced, has been observed before for protein databases such as COG (Clusters of Orthologous Groups of proteins) (Unger, Uliel, and Havlin, 2003) and TRIBES (Enright, Kunin, and Ouzounis, 2003), and has been seen in other datasets analysed with KinFin. These clusters contain a large number of strict ('true') single-copy orthologues, and many 'fuzzy' single-copy orthologues.

KinFin can assist in deciding which of several alternative proteome predictions is more likely to be correct. Examination of the distribution of clusters within species can highlight outlier datasets. Both *C. elegans* and *C. briggsae* have higher total protein counts than any of the filarial species and display the highest proportion of singletons (CELEG: 15.7% and CBRIG: 24.4%) (Figure 3.3.2). For the species for

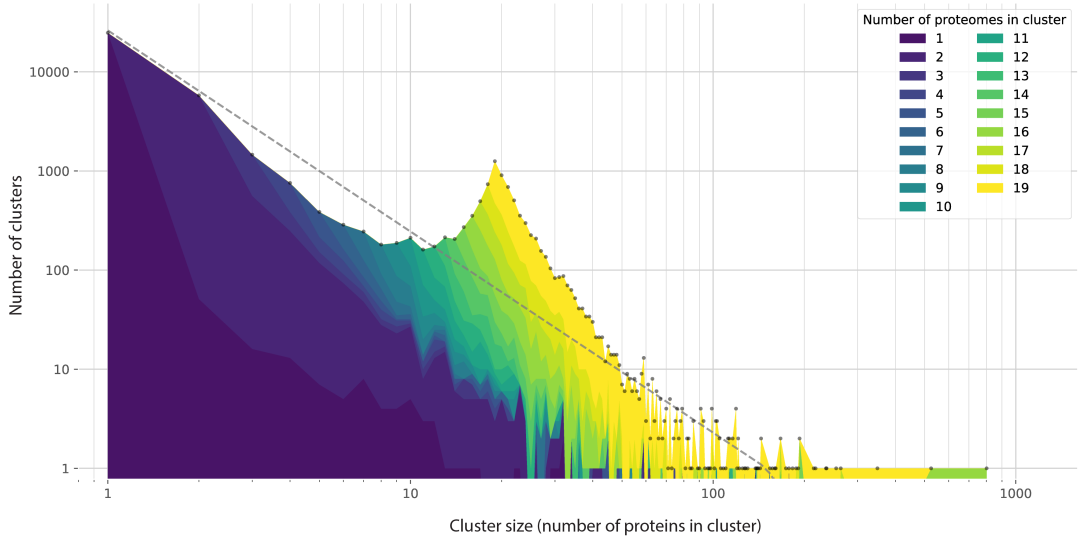


Figure 3.3.1: Distribution of cluster sizes. The distribution of counts of proteins for each cluster is coloured based on the number of proteomes present in each cluster. Total values of counts of each cluster size are indicated with grey dots. A fitted power-law curve (grey) is drawn for reference.

which two assemblies were analysed, variation in proportion of singletons is largest for *L. loa* (LOA1: 14.6% vs. LOA2: 15.1%).

The 781 single-copy orthology clusters identified in the initial KinFin analysis yielded a robustly supported phylogeny (Figure 3.3.3A). By rooting the tree with the common ancestor of *Caenorhabditis* species, the three non-filarial taxa are recovered in expected positions, with *S. labiatopapillosa* most closely related to the onchocercids, followed by *T. callipedia* and *D. medinensis*. The relationships between the onchocercid taxa is not congruent with the reference NCBI taxonomy, but with a previous analysis using a smaller number of loci (Lefoulon et al., 2016). *D. immitis* is recovered as sister to *Onchocerca* spp. (the clade defined by node 'n11' in Figure 3.3.3A), and *W. bancrofti*, *Brugia* spp. and *L. loa* (node 'n15') form a clade distinct from *L. sigmodontis*, *A. viteae* and *E. elaphi* (node 'n16').

The additional 3887 'fuzzy' single-copy orthologues identified by KinFin were

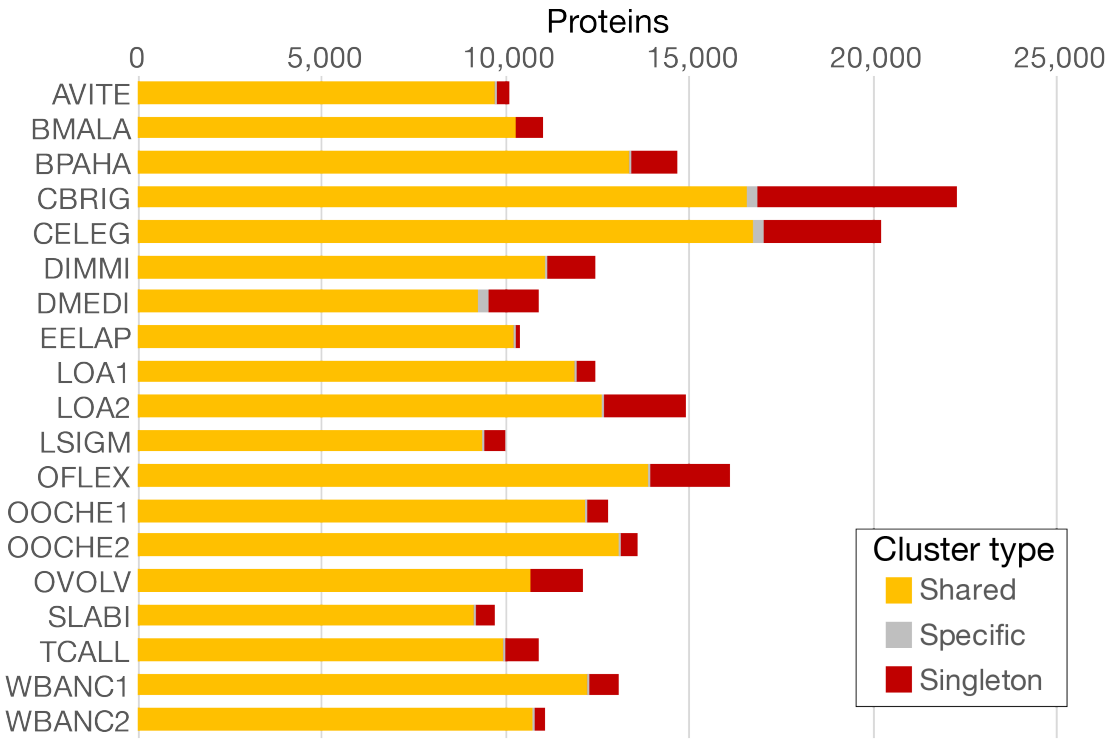
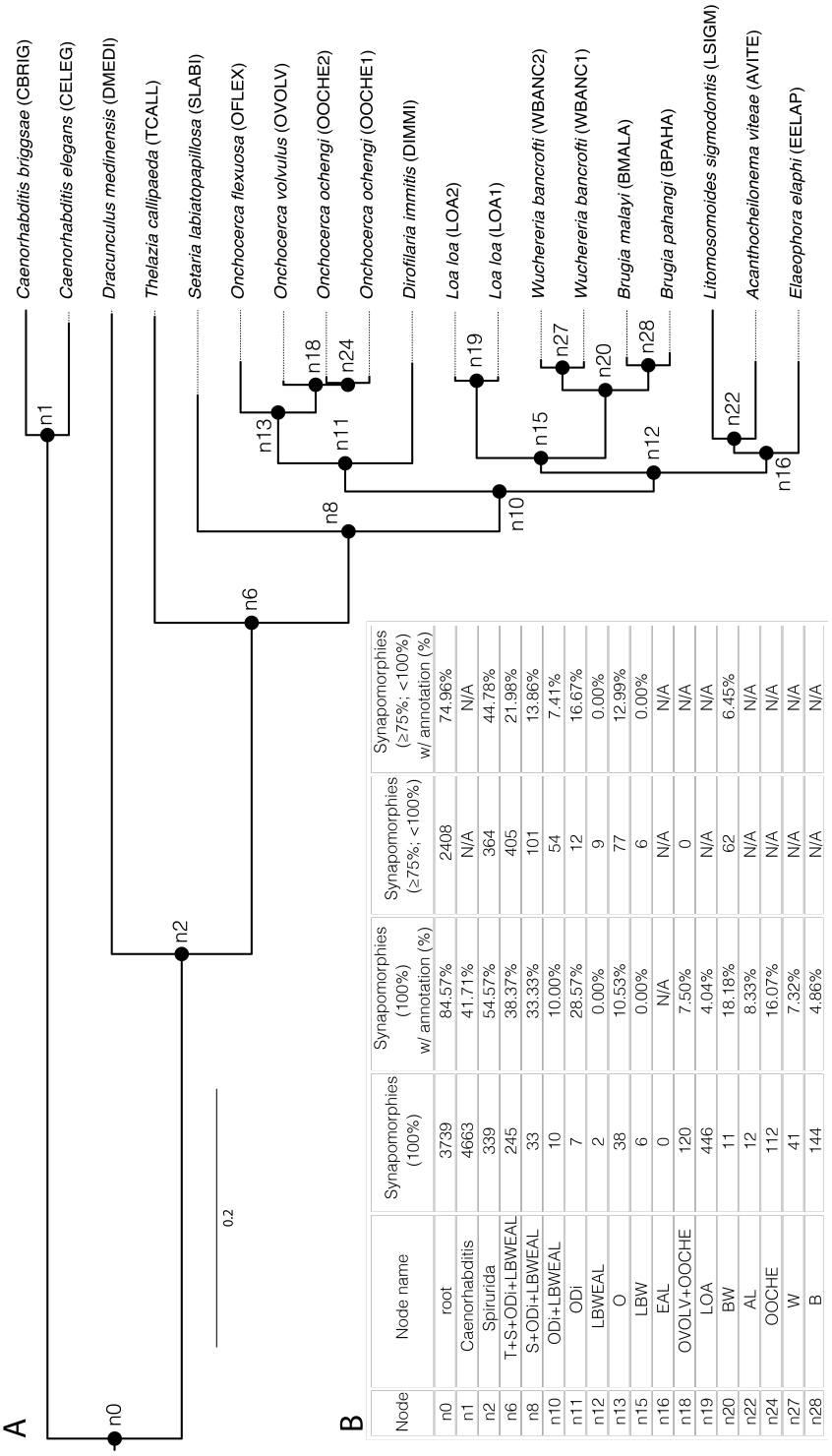


Figure 3.3.2: Count of proteins by type of cluster. 'Shared': clusters containing proteins from multiple taxa. 'Specific': clusters containing two or more proteins from a single proteome. 'Singleton': clusters containing a single protein.

not used in the analysis. However, 'fuzzy' single-copy orthologues can be useful for analysis of proteomes from more distantly related taxa, where stochastic absence and duplication can severely limit the number of single-copy loci recovered for phylogenetic analyses. 'Fuzzy' orthologues can be used in combination with tools such as PhyloTreePruner (Kocot et al., 2013) which filters out-paralogues and selects appropriate in-paralogues.

I explored the proteomic diversity represented by the three clades within Onchocercidae (Figure 3.3.3A, at nodes 'n11', 'n15', 'n16') by defining taxon sets for each of the nodes and used KinFin to generate rarefaction curves for each set (Figure 3.3.4). Curves for node 'n11' (*D. immitis* and *Onchocerca* species) and node 'n15' (*W. bancrofti*, *Brugia* species and *L. loa*) show comparable slopes and the number of non-singleton clusters recovered in both is very similar (11,084



Node	Node name	Synapomorphies (100%)	Synapomorphies (100%) w/ annotation (%)	Synapomorphies (≥75%; <100%)	Synapomorphies (≥75%; <100%) w/ annotation (%)
n0	root	3739	84.57%	2408	74.96%
n1	Caenorhabditis	4663	41.71%	N/A	N/A
n2	Spirurida	339	54.57%	364	44.78%
n6	T+S+ODi+LBWEAL	245	38.37%	405	21.98%
n8	S+ODi+LBWEAL	33	33.33%	101	13.86%
n10	ODi+LBWEAL	10	10.00%	54	7.41%
n11	ODi	7	28.57%	12	16.67%
n12	LBWEAL	2	0.00%	9	0.00%
n13	O	38	10.53%	77	12.99%
n15	LBW	6	0.00%	6	0.00%
n16	EAL	0	N/A	N/A	N/A
n18	OVOLV+OOCHE	120	7.50%	0	N/A
n19	LOA	446	4.04%	N/A	N/A
n20	BW	11	18.18%	62	6.45%
n22	AL	12	8.33%	N/A	N/A
n24	OOCHE	112	16.07%	N/A	N/A
n27	W	41	7.32%	N/A	N/A
n28	B	144	4.86%	N/A	N/A

Figure 3.3.3: Phylogenetic tree of nematodes in the analysis and functional annotation of synapomorphies. **A**: Phylogenetic tree based on 781 single-copy orthologues. Non-parametric bootstrap support for all branches is 100. Internal nodes are labelled. **B**: Table summarising 'complete presence' synapomorphic clusters (100% taxon coverage) and 'partial absence' synapomorphic clusters (75% ≤ taxon-coverage < 100%) and the percentage for which a RFA could be inferred. 'N/A' is used for cases in which nodes are ancestors of less than four taxa or when percentage of RFA could not be calculated due to lack of clusters.

clusters for ‘n11’ and 10,989 for ‘n15’). Fewer unique protein clusters (9393) were recovered when sampling node ‘n16’ (*L. sigmodontis*, *A. viteae* and *E. elaphi*). The fact that none of the curves reaches a plateau suggests that their protein space has not been sampled exhaustively.

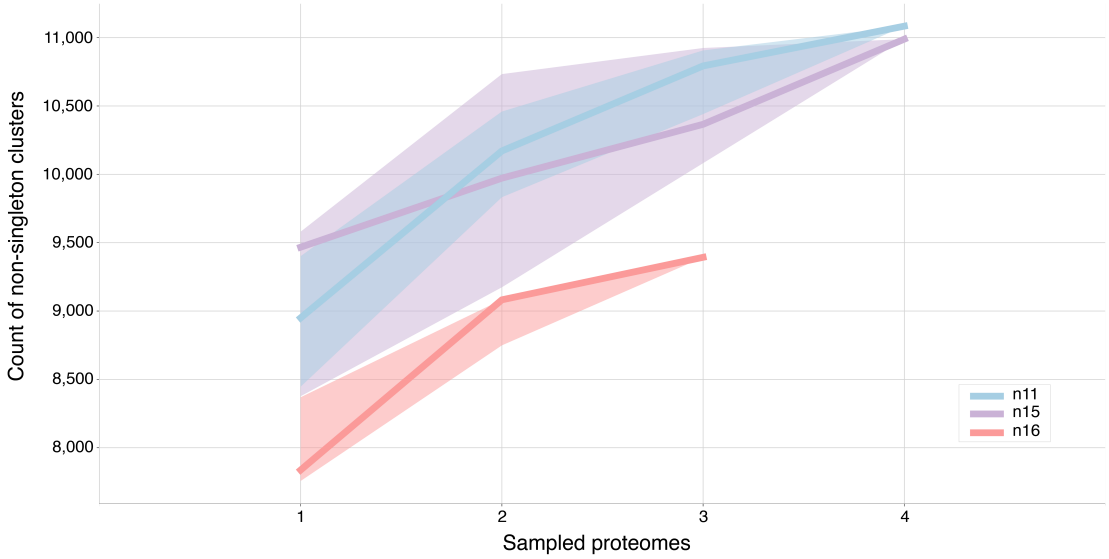


Figure 3.3.4: Rarefaction curves for taxon sets. Rarefaction curves for proteomes within taxon sets defined by major clades within the onchocercid nematodes: ‘n11’ = *D. immitis* and *Onchocerca* species. ‘n15’ : *W. bancrofti*, *Brugia* species and *L. loa*. ‘n16’: *L. sigmodontis*, *A. viteae* and *E. elaphi*. The envelope around rarefaction curves was computed based on 25 iterations of random sampling.

Of all proteins used in the analysis, 157,873 (63.47%) were annotated with InterPro (IPR) domains. RFA of clusters yielded 12,026 (28.17%) clusters where 75% of contained proteins shared an IPR domain and 75% of proteomes had at least one protein with that domain. Using the phylogeny based on single-copy orthologues (Figure 3.3.3A), I identified synapomorphies at each node and investigated their RFA (Figure 3.3.3B). While many clusters are synapomorphies of deeper nodes, Onchocercidae and the three groups within Onchocercidae have few synapomorphic gene family births (ten at ‘n10’, seven at ‘n11’, six at ‘n15’ and zero at ‘n16’). Of those, only two clusters at ‘n11’ received a RFA: a ‘Chromadorea ALT’ cluster (‘OG0007060’) and a ‘SOCS box domain’ cluster (‘OG0009843’). The

Chromadorea ALT domain is found across Nematoda and is also found in several other clusters. *B. malayi* ALT-1, the first described Chromadorea ALT protein (contained in cluster 'OG0000082'), has been proposed as a candidate vaccine target for human lymphatic filariasis (Gregory et al., 2000). The synapomorphic 'Chromadorea ALT' cluster is specific to *Onchocerca* spp. and *D. immitis* and might harbour the same potential for onchocerciasis. SOCS box domains were first identified in proteins involved in suppression of cytokine signalling, and are key regulators of both innate and adaptive immunity (Alexander, 2002). Proteins in 'OG0009843' do not contain any of the other domains usually associated with SOCS, such as SH2 (a combination found in 'OG0000874' and 'OG0007539') or Ankyrin repeat-containing domains (a combination found in 'OG0001559' and 'OG0015826'). However, they may still play an immunomodulatory role during infection as has been suggested for SOCS box proteins in *L. sigmodontis* (Godel et al., 2012).

Definition of taxon sets based on host species ('human' vs. 'other' vs. 'outgroup') recovered 628 clusters specific to filarial nematodes, but none had proteins of more than four out of seven proteomes. Hence, I found no evidence of systematic convergent adaptation to human hosts in the analysed proteomes of filarial nematodes.

KinFin permits rapid assessment of differences in copy number between species and taxon sets using protein count representation tests. Analysis of clusters shared between taxa at either side of the basal split in Onchocercinae ('n11': *D. immitis* and *Onchocerca* spp., and 'n15': other filaria) (Figure 3.3.5) identified 10 clusters with extreme differences (see Table 3.3.2). Among these was cluster 'OG0000051', which includes prolyl 4-hydroxylase orthologues, including Bm-PHY-1 and Bm-PHY-2 which are essential for development and cuticle formation, and have been suggested as potential targets for parasite control (Winter et al., 2013). While all

'n15' taxa have exactly two paralogues ('WBANC2' contained only Wb-PHY-1, but Wb-PHY-2 was located through a TBLASTN search and was present in 'WBANC1'), counts in 'n11' taxa ranged from five ('OFLEX') to 14 ('OVOLV'). Three additional singleton prolyl 4-hydroxylase clusters were identified, composed only of 'n15' taxa. The number of paralogous prolyl 4-hydroxylases in *D. immitis* and *Onchocerca* spp. could have negative implications in control measures against this locus.

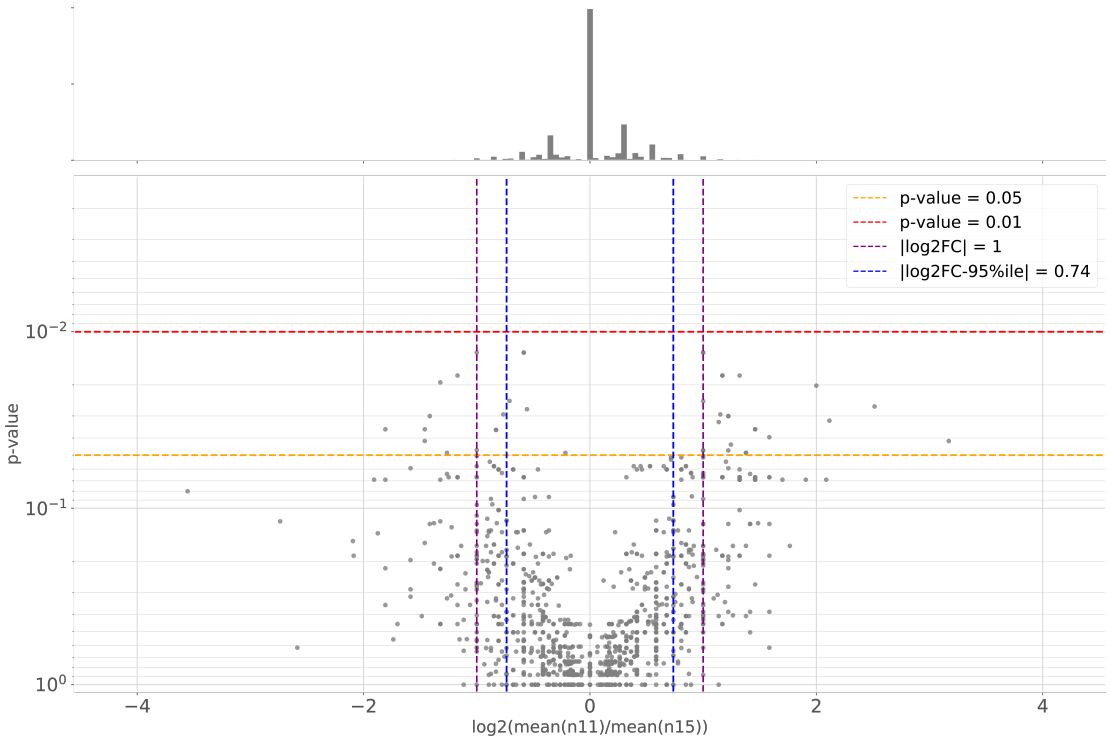
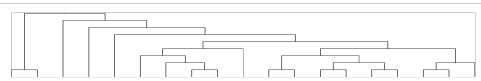


Figure 3.3.5: Volcano plot of protein count representation tests. Mann-Whitney-U tests were carried out for clusters shared between taxa at 'n11' (*D. immitis* and *Onchocerca* spp.) and 'n15' (*W. bancrofti*, *Brugia* spp. and *L. loa*). The histogram (top) shows density of data points by location on the x-axis.

Table 3.3.2: RFA of ten clusters of interest. Clusters exhibiting most extreme values for $\log_2$-fold change of means among 'n11' (*D. immitis* and *Onchocerca* spp.) and 'n15' (*W. bancrofti*, *L. loa* and *Brugia* spp.). μ_{n11}: Mean count of proteins at 'n11'. μ_{n15}: Mean count of proteins at 'n15'. Proteins: count of proteins in cluster. Proteomes: count of proteomes in cluster.

Cluster ID	$\log_2$-fc	μ_{n11}	μ_{n15}	Proteins	Proteomes	IPR IDs	IPR description
OG0000202	3.17	9.00	1.00	54	9	None	None
OG0000051	2.51	10.00	1.75	105	19	IPR013547	Prolyl 4-hydroxylase alpha-subunit, N-terminal
						IPR005123	Oxoglutarate/iron-dependent dioxygenase
OG0001059	2.12	4.33	1.00	37	19	IPR000219	Dbl homology (DH) domain
OG0000611	2.09	4.25	1.00	34	17	IPR000571	Zinc finger, CCCH-type
OG0000494	2.00	4.00	1.00	29	17	IPR008914	Phosphatidylethanolamine-binding protein
OG0000042	−3.55	1.00	11.75	116	12	None	None
OG0000062	−2.74	1.75	11.67	99	12	None	None
OG0001116	−2.58	1.00	6.00	28	10	None	None
OG0000024	−2.09	4.33	18.5	143	14	None	None
OG0000857	−2.08	1.00	4.25	31	16	None	None

To demonstrate the utility of KinFin for targeted analysis of clustered protein sequences (and their underlying genes and gene families), I focussed on the biology of haem synthesis and transport in the Onchocercidae. This pathway is a target of active investigation for drug development. While most organisms can synthesise haem, a complete haem biosynthetic pathway is lacking in all nematodes studied to date (Rao et al., 2005), and proteins of only two of the 12 catabolic steps have been described in *C. elegans* (Ce-COX-10 and Ce-COX-15). In *C. elegans*, multiple haem responsive genes (HRGs) have been characterised (Rajagopal et al., 2008; Chen et al., 2011; Sinclair and Hamza, 2015) and orthologues have been identified in *B. malayi* (Bm-HRG-1 and Bm-HRG-2) and *D. immitis* (Luck et al., 2016). In *C. elegans*, HRGs are involved in haem trafficking within the epidermis (HRG-2), to oocytes (HRG-3) and within the intestine (HRG-1/4/5/6). Other ABC transporters in *C. elegans* have been implicated in haem homeostasis (MRP-5, F22B5.4, and ABTM-1) (Severance et al., 2010; González-Cabo et al., 2011; Antonicka et al., 2003). An orthologue of MRP-5 has been described in *B. malayi* (Luck et al., 2016). Several animal parasitic nematodes (including *B. malayi*, *D. immitis*, and *O. volvulus*) have been shown to harbour a functional ferrochelatase (FeCH) acquired through horizontal gene transfer from an alphaproteobacterium (Elsworth, Wasmuth, and Blaxter, 2011; Nagayasu et al., 2013; Wu et al., 2013). Other nematodes have distinct ferrochelatase-like (FeCL) homologues which lack the active site. I catalogued homologues of these proteins in the clustering analysis (Figure 3.3.6).



Target protein	Gene name	Gene description	Cluster ID	Protein count	Taxon count	CBRIG	CELEG	DMEDI	TCALL	SLABI	OFLEX	OVOLV	OOCHE2	OOCHE1	DIMMI	LOA2	LOA1	WBANC2	WBANC1	BMALA	BPAHA	LSIGM	AVITE	EELAP	IPR IDs	IPR ID descriptions
CELEG.K07G5.6	Ce-fecl-1	Ferrochelatase-like	OG0004284	20	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	IPR001015	Ferrochelatase
BMALA.Bm1182	Be-fech-1	Ferrochelatase, HGT	OG0005607	19	17	0	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	2	IPR001015	Ferrochelatase
CELEG.T06D8.5	Ce-cox-15	Heme biosynthesis	OG0003121	21	19	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	2	1	1	1	IPR003780	COX15/CtaA family
CELEG.Y46G5A.2	Ce-cox-10		OG0006880	17	17	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	IPR000537	UbiA prenyltransferase family
CELEG.W09D6.6	Ce-hmt-1	Heme homeostasis	OG0010331	4	3	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	IPR011527;IPR003439	ABC transporter type 1, transmembrane domain;ABC transporter-like
CELEG.Y74C10AR.3a	Ce-abtm-1		OG0004667	19	19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	IPR003439;IPR011527	ABC transporter-like;ABC transporter type 1, transmembrane domain
CELEG.F22B5.4	Ce-F22B5.4		OG0002796	22	19	2	2	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	None	None
CELEG.F14F4.3	Ce-mrp-5		OG0003174	21	19	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	IPR003439;IPR011527	ABC transporter-like;ABC transporter type 1, transmembrane domain
BMALA.Bm3373a	Bm-mrp-5		OG0003174	21	19	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	1	IPR003439;IPR011527	ABC transporter-like;ABC transporter type 1, transmembrane domain
BMALA.Bm5182	Bm-hrg-1		OG0007366	16	15	1	1	1	1	0	1	1	0	1	2	1	1	1	0	1	1	1	0	1	IPR026218	Heme transporter HRG
CELEG.R02E12.6	Ce-hrg-1		OG0007366	16	15	1	1	1	1	0	1	1	0	1	2	1	1	1	0	1	1	1	0	1	IPR026218	Heme transporter HRG
BMALA.Bm2383	Bm-hrg-2		OG0004125	20	18	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	2	2	1	1	IPR033468;IPR012336	Metaxin, glutathione S-transferase domain;Thioredoxin-like fold
CELEG.K01D12.14	Ce-hrg-2		OG0011423	3	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	IPR033468;IPR012336	Metaxin, glutathione S-transferase domain;Thioredoxin-like fold
CELEG.F58E6.7	Ce-hrg-3		OG0016207	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	None	None
CELEG.F36H1.5c	Ce-hrg-4		OG0013644	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	IPR026218	Heme transporter HRG
CELEG.F36H1.9	Ce-hrg-5		OG0013643	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	None	None
CELEG.F36H1.10a	Ce-hrg-6		OG0013645	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	IPR026218	Heme transporter HRG

Figure 3.3.6: Haem biosynthesis and transport proteins. The columns showing counts of proteins per taxon are ordered based on phylogeny (cartoon above). Counts are coloured orange if they result from fragmented predictions and green if the protein was absent from the proteome but the gene was identified by directed TBLASTN searches against the respective genome assembly.

FeCL proteins were identified in all species. *B. pahangi* has two FeCL proteins while all other taxa have one, but both are located at scaffold borders and may be the result of an assembly artefact. The horizontally-acquired FeCH is absent from the *Caenorhabditis* proteomes (and genomes) but present in all the other taxa analysed. Paralogues in one of the *O. ochengi* proteomes are suggestive of misprediction. COX-10 and COX-15 are present in most taxa in the analysis; paralogues in *B. pahangi* and *O. flexuosa* are a result of fragmented assemblies. COX-10 is present in *W. bancrofti* 'WBANC1' (on scaffold 'WBA_contig0009713'), but the gene was not predicted. COX-10 was not found in *E. elaphi*, which suggests that either the corresponding genomic region was not assembled or that the gene has been lost.

Presence/absence of proteins involved in haem homeostasis showed a more complex pattern. Ce-HMT-1, an ATP-dependent phytochelatin transporter, was restricted to *Caenorhabditis* spp. and *D. medinensis*. The other ABC-transporter-like proteins (ABTM-1, MRP-5, and F22B5.4) were present across all taxa. For F22B5.4, genuine paralogues were found in both *Caenorhabditis* spp. and *O. volvulus*. Ce-MRP-5 and Bm-MRP-5 were located within the same cluster, and apparent paralogues in *O. flexuosa* and *W. bancrofti* 'WBANC1' derived from predictions located at the ends of scaffolds. While no orthologues of Ce-HRG-2/3/4/5/6 were identified, the cluster containing Ce-HRG-1 included representatives from most species. Missing orthologues of HRG-1 were identified using TBLASTN searches in *S. labiatopapillosa* (scaffold 'nSl.1.1.scaf00038'), *O. ochengi* 'OOCHE2' (scaffold 'nOo.2.0.Scaf03259'), *W. bancrofti* 'WBANC1' (scaffold 'WBA_contig0001821'), and *A. viteae* (scaffold 'nAv.1.0.scaf00129'). The two HRG-1 paralogues in *D. immitis* were identified previously (Luck et al., 2014; Luck et al., 2016). Interestingly, Bm-HRG-2 (Bm2383) is not orthologous to Ce-HRG-2 but rather to Ce-C25H3.7, an orthologue of human FAXC (failed axon connection).

3.3.4 Conclusion

I presented some of the main capabilities of KinFin through the analysis of proteomes of filarial nematodes and outgroup species. By extracting single-copy orthologues I resolved the phylogenetic relationships between filarial nematodes. I explored synapomorphic clusters and their functional annotations across the phylogeny and identified putative gene families of interest. Through definition of (phylogenetic) taxon sets, I assessed the proteomic diversity across key clades of filarial nematodes. To illustrate targeted analysis of proteins of interest, I analysed clusters containing proteins involved in haem metabolism and homeostasis using characterised orthologues from the model organism *C. elegans*.

In summary, all non-*Caenorhabditis* nematodes analysed have a functional FeCH, orthologous to the one acquired through HGT in *B. malayi*. Proteins responsible for the only two steps in haem biosynthesis described in *C. elegans* are also found in all taxa, apart from COX-10 in *E. elaphi*. The haem transporters HRG-2/3/4/5/6 are (in this analysis) restricted to *Caenorhabditis* spp., but all spirurid nematodes analysed have retained HRG-1, a FAXC-like cluster orthologous to Bm-HRG-2, and MRP-5, and these may mediate haem transport from the intestine.

3.4 Use case 2: Analysis of gene families in parasitic worms

3.4.1 Introduction

Parasitic taxa within the phyla Nematoda and Platyhelminthes are of substantial medical and veterinary interest. Estimates suggest that a quarter of the human population is infected by parasitic worms and suffers from pain, malnutrition, and disability due to the diseases they cause (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, 2016). According to the World Health Organisation, diseases associated with nematode and platyhelminth infections account for eight out of 19 of the most neglected tropical diseases (Molyneux, Savioli, and Engels, 2017). Furthermore, infections of domestic animals lead to substantial economic losses in developing countries, affecting meat and milk production (Charlier et al., 2014) as well as the livestock industry (Morgan et al., 2013).

The 50 Helminth Genome project at the Wellcome Trust Sanger Institute (WTSI) is an international collaboration with the McDonnell Genome Institute (Washington University), Edinburgh Genomics (University of Edinburgh) and several research groups around the world. The project is aimed at sequencing and understanding the genomes of those nematode and platyhelminth parasites with the greatest medical and veterinary impact. It should be noted that the word 'helminth' does not refer to a taxonomic group but rather to a phenotypic description, as both nematodes and platyhelminths belong to different superphyla (Winnepenninckx, Peer, and Backeljau, 1998). Platyhelminthes is nested within the superphylum Lophotrochozoa, while Nematoda is part of the superphylum Ecdysozoa (Dunn et al., 2008). Hence, both lineages acquired the molecular machinery necessary to infect animals independently. In Platyhelminthes animal parasitism arose most likely

once (Olson and Tkach, 2005), while in Nematoda multiple transitions from free living to animal parasitic lifestyles have occurred (Blaxter et al., 1998). Comparison of the gene family repertoire of both phyla can thus highlight similarities and differences which might one day be exploited for the development of vaccines and treatments.

Mark Blaxter and I were invited to participate in the analysis of the data generated by the 50 Helminth Genome Project. I applied KinFin to datasets generated by collaborators and analysed the clustering of 1.6 million proteins from previously published genomes and of 31 nematode and 14 platyhelminth species sequenced within the scope of the project. I surveyed synapomorphic gene families at key nodes within the phylogenetic tree of metazoans and expanded on the analysis of ferrochelatases reported in Section 3.3.

3.4.2 Methods

Data

Collaborators within the 50 Helminth Genome Project provided a phylogenetic tree, InterProScan functional annotations, and an Ensembl Compara (Herrero et al., 2016) protein clustering of 91 species, comprising 25 platyhelminths, 56 nematodes and 10 outgroup taxa from other animal phyla (Table 3.A.1).

Analysis of synapomorphic clusters

The Ensembl Compara protein clustering was analysed using `KinFin v0.8.3` (Laetsch, 2017a) by providing InterPro IDs from functional annotations and the

phylogenetic relationships of the included taxa. Synapomorphic clusters at 28 nodes of interest across the phylogenetic tree were investigated and grouped into the categories 'complete presence' (if all taxa under the node were present in the cluster) or 'partial absence' (if at least 90% of taxa under the node were present). RFA of a cluster was inferred if more than 80% of the species in the cluster contained at least one protein with that domain.

Phylogenetic analysis of ferrochelatase clusters

I screened Compara clusters for 'Ferrochelatase domain' (IPR001015) and 'Ferrochelatase active site' (IPR019772) annotation, and collaborators provided me with the protein sequences based on that list. I retrieved additional ferrochelatase protein sequences from NCBI GenBank for 17 bacterial taxa: *Rhizobium leguminosarum* (YP_002977390.1), *Sinorhizobium meliloti* (NP_386909.2), *Roseibium* sp. (ZP_07659792.1), *Ehrlichia chaffeensis* (YP_507215.1), *E. canis* (YP_303255.1), *E. ruminantium* (YP_196566.1), *Pseudomonas putida* (YP_001266120.1), *P. fluorescens* (YP_350458.1), *P. syringae* (YP_234061.1), *Leadbetterella byssophila* (YP_003998063.1), *Mucilaginibacter paludis* (EFQ76108.1), *Hydrogenobacter thermophilus* (ADO44739.1), and *Wolbachia* endosymbionts of *Brugia malayi* (YP_198549.1), *Muscidifurax uniraptor* (ZP_03788224.1), *Drosophila melanogaster* (NP_966898.1), *Culex quinquefasciatus* (YP_001975511.1), and *Dirofilaria immitis* (ABV58328.1) based on Elsworth et al. (2011). Phylogenetic analysis was carried out as described in Section 3.3.2.

3.4.3 Results

The Compara clustering comprised 1,640,269 proteins placed in 384,608 clusters, of which 275,037 were singletons (accounting for 16.77% of proteins). Results of the KinFin analysis of synapomorphic clusters at 28 nodes of interest within the tree of Metazoa are listed in Table 3.4.1. In total, 3512 'complete presence' synapomorphic clusters were found for the 28 nodes of interest, of which 16.86% were assigned a RFA. Allowing for stochastic absence of proteomes in clusters, 2369 'partial absence' synapomorphies could be identified, of which 23.98% received a RFA. The low percentage of RFAs for synapomorphic clusters could be explained by diverse gene annotation pipelines applied to each of the taxa. However, it is striking that phylogenetically deeply conserved protein clusters, *e. g.* at node 'A' (Platyhelminthes) and node 'B' (Nematoda), contained many proteins that have so far escaped functional annotation by the research community. Synapomorphic clusters which received a RFA were grouped into 24 functional categories relevant to parasitism and are visualised in Figure 3.4.1.

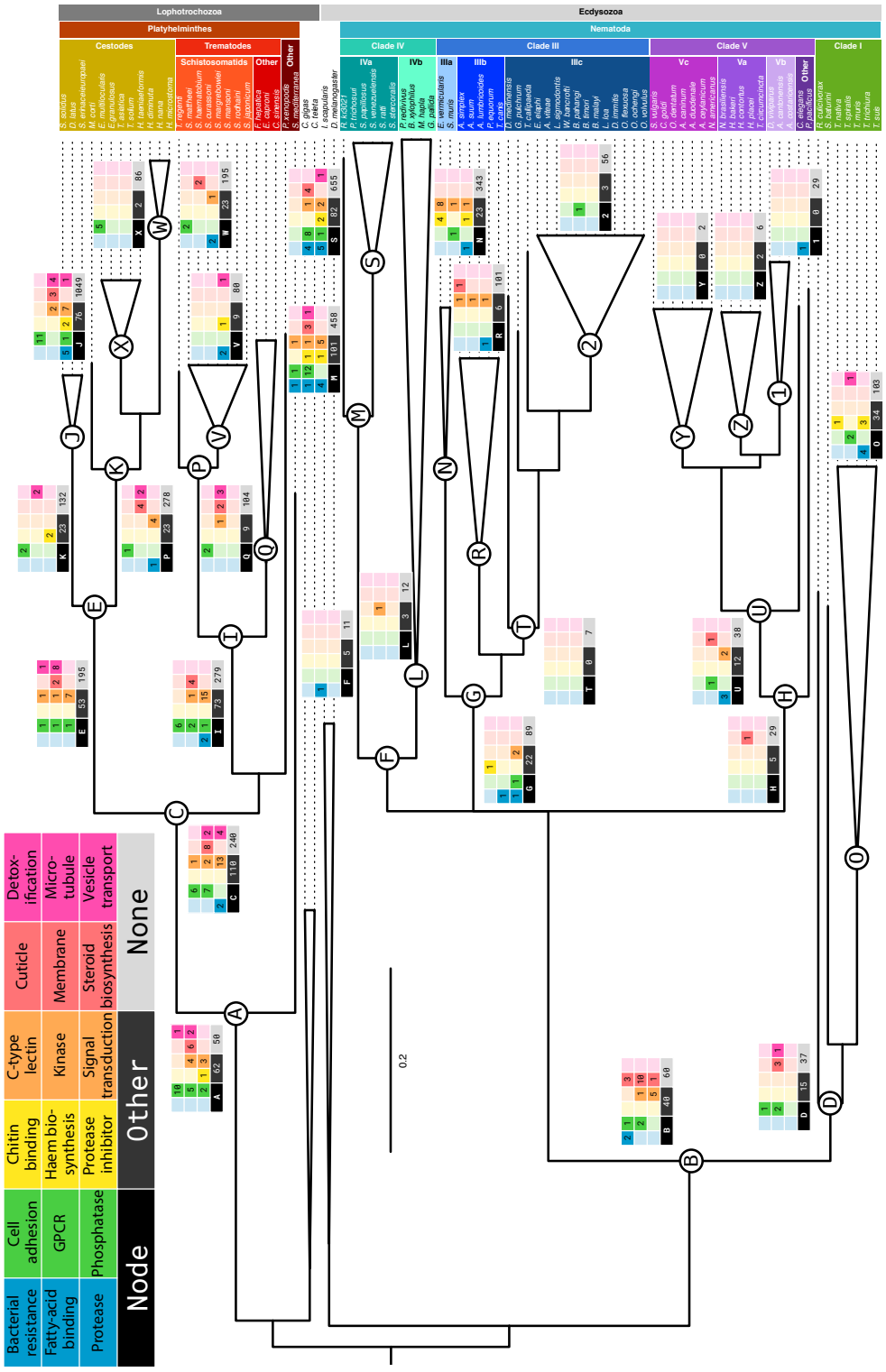

Lophotrochozoa
Ecdysozoa
Platyhelminthes
Nematoda
Cestodes
Trematodes
Schistosomatids
Other
Clade IV
Clade III
Clade V
Clade I
IVa
IVb
IIIa
IIIb
IIIc
Vc
Va
Vb
Other
S. solidus
D. ialus
H. microstomum
M. corti
E. multiloculalis
E. granulosus
T. asiatica
T. solium
H. taeniaeformis
H. diminuta
H. vana
H. microstoma
T. regenti
S. matthei
S. haematobium
S. japonicum
S. margrebowiei
S. mansoni
S. rodhaini
S. bovis
F. hepatica
E. caproni
E. sinensis
C. sinensis
O. viverrini
C. gigas
C. teleta
L. saipanensis
D. melanogaster
R. koreana
P. trichosuri
S. papillosus
S. ransomi
S. ratti
S. stercoralis
P. redivivus
B. malayi
M. hapla
G. pallida
E. vermicularis
S. mansi
A. suum
A. lumbricoides
P. equorum
T. muris
G. pulchrum
T. callipaeda
T. canis
A. viteae
L. sigmodontis
W. bancrofti
B. pahangi
B. malayi
L. loa
O. tenuis
O. flexuosa
O. ochengi
O. volvulus
S. vulgaris
C. elegans
A. caninum
A. duodenale
A. ceylanicum
N. americanus
N. brasiliensis
H. bakeri
H. contortus
H. polygyrus
T. circumcincta
D. viviparus
A. castaneosias
A. simplex
C. elegans
P. pacificus
R. culicivorax
S. ratti
T. nativa
T. spiralis
T. muris
T. trichiura
T. suis
0.2
Detox-ification
Micro-tubule
Vesicle transport
Cuticle
Membrane
Steroid biosynthesis
None
C-type lectin
Kinase
Signal transduction
Other
Chitin binding
Haem bio-synthesis
Protease inhibitor
Cell adhesion
GPCR
Phosphatase
Bacterial resistance
Fatty-acid binding
Protease
Node

Figure 3.4.1 ***(previous page)*****: Phylogenetic tree of Metazoa annotated with functional categories of synapomorphic gene families**. Rectangular panels indicate counts of synapomorphic gene families grouped by 24 functional categories, detailed in the major panel. Node: Node in phylogenetic tree to which a panel refers to. Other: synapomorphic gene families with representative functional annotation that could not be grouped into one of the 24 functional categories. None: synapomorphic gene families that had no representative functional annotation.

Functional annotation of synapomorphic clusters is diverse, but no striking signature of parasitism between nematodes and platyhelminths, or within either phylum could be observed. Some functional annotations were frequently associated with synapomorphic clusters, including several related to sensory perception (such as G-protein coupled receptors), parasite surfaces (platyhelminth tegument or nematode cuticle maintenance proteins), and protein degradation (proteases and protease inhibitors).

Within Nematoda, Clade IVa (Strongyloididae) displayed the highest number of synapomorphic clusters, including five fatty acid and retinol-binding (FAR) clusters, a novel ferrochelatase cluster, and the highest number of synapomorphic G-protein coupled receptor (GPCR) clusters within Nematoda. Fatty acid and retinol-binding proteins have previously been implicated in host-parasite interaction in both plant- and animal-parasitic nematodes (Prior et al., 2001; Rey-Burusco et al., 2015), suggesting a role in immune modulation, which might be involved in the host-parasite interaction in Strongyloididae.

Analysis of synapomorphic clusters of parasitic platyhelminths (Neodermata, composed of Protopolystoma, Trematoda and Cestoda) identified a clade-specific inositol-pentakisphosphate 2-kinase that produces inositol hexakisphosphate nano-deposits in the cyst wall of some *Echinococcus* species (Casaravilla et al., 2006). These deposits increase the surface area for protein adsorption and might play a role in host-parasite interaction (Díaz et al., 2016). Synapomorphic clusters, with

RFAs grouped under the category 'Membrane' and 'Vesicle transport', were prevalent at several platyhelminth nodes and might be involved in maintenance and function of the neoderm (double bilayer) of these taxa.

As described in Section 3.3, most nematode genomes lack genes coding for functional ferrochelatases but harbour ferrochelatase-like (FeCL) genes of unknown function which are devoid of the active site. Exceptions are animal parasites in nematode Clades III and IV that acquired a functional ferrochelatase via horizontal gene transfer from alphaproteobacteria (Elsworth, Wasmuth, and Blaxter, 2011; Wu et al., 2013; Nagayasu et al., 2013). Screening for clusters containing proteins annotated with 'Ferrochelatase domain' (IPR001015) and/or 'Ferrochelatase active site' (IPR019772) identified four clusters. Two clusters were entirely comprised of proteins from nematode taxa and lacked the active site: the synapomorphic Clade IVa cluster ('Clade IVa FeCL') and a cluster containing proteins from most nematodes ('FeCL'). The other two clusters harbour functional ferrochelatases: one consisting of non-nematode taxa ('FeCH') and one composed of 34 nematode and five non-nematode taxa ('FeCH-HGT'). Phylogenetic analysis revealed the synapomorphic 'Clade IVa FeCL' to be part of the nematode 'FeCL'-clade (Figure 3.4.2). The five non-nematode sequences in the 'FeCH-HGT' cluster were recovered inside the non-nematode FeCH clade, and were most likely an artefact of the Compara clustering method. The nematode sequences in the 'FeCH-HGT' cluster form a monophyletic clade placed next to alphaproteobacteria, consistent with previous findings. The species composition of 'FeCH-HGT' suggests that the acquisition of the functional ferrochelatase predated the split of Clades III and IV.

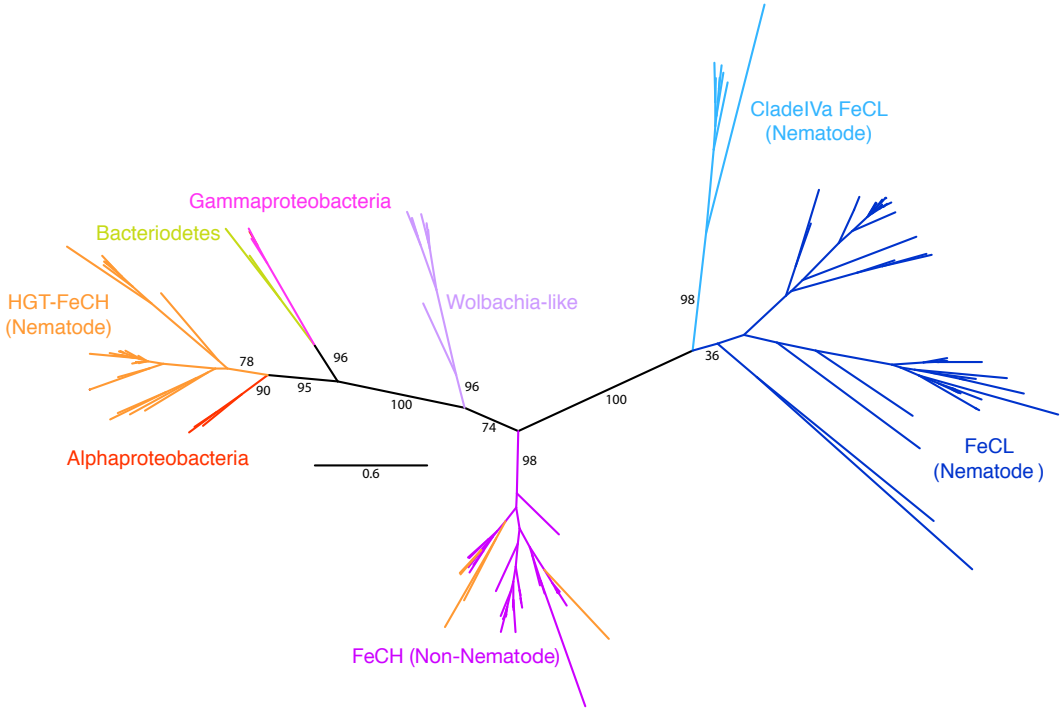


Figure 3.4.2: Phylogenetic tree of ferrochelatases. Branches are coloured by membership of protein sequences in ferrochelatase groups. Non-parametric bootstrap support is depicted for main branches only. 'FeCL' (Nematode): nematode specific FeCL proteins, devoid of active site; 'Clade IVa FeCL' (Nematode): synapomorphic FeCL cluster of Nematode Clade IVa; 'FeCH' (Non-Nematode): functional FeCH cluster composed of taxa of non-nematode phyla; 'HGT-FeCH' (Nematode): Clade III/IV specific FeCH acquired through horizontal gene transfer from alphaproteobacteria; 'Alphaproteobacteria': FeCH of non-*Wolbachia* alphaproteobacteria; 'Bacteriodetes': FeCH of *L. byssophila* and *M. paludis*; 'Gammaproteobacteria': FeCH of *Pseudomonas* spp.; 'Wolbachia-like': FeCH of *Wolbachia* spp., *E. chaffeensis* and *H. thermophilus*.

Table 3.4.1: Counts of synapomorphic clusters by node of interest.

			complete presence			partial absence			
Node ID	Name	Phylum	Count	RFA	RFA (%)	Count	RFA	RFA (%)	Proteomes
A	Plathyhelminthes	P	71	47	66.2	71	45	63.4	25
B	Nematoda	N	20	13	65.0	104	51	49.0	56
C	Neodermata	P	107	49	45.8	288	106	36.8	24
D	Clade I	N	0	0	N/A	59	22	37.3	7
E	Cestodes	P	137	40	29.2	133	35	26.3	12
F	Clade IV	N	0	0	N/A	17	6	35.3	10
G	Clade III	N	0	0	N/A	116	27	23.3	22
H	Clade V	N	0	0	N/A	35	6	17.1	17
I	Trematoda	P	177	55	31.1	205	48	23.4	11
J	Diphyllobothriidea	P	1159	110	9.5	0	0	N/A	3
K	Cyclophyllidea	P	161	29	18.0	0	0	N/A	9

Continued from previous page

			'complete presence'			'partial presence'			
Node ID	Name	Phylum	Count	RFA	RFA (%)	Count	RFA	RFA (%)	Proteomes
L	Clade IVb	N	0	0	N/A	16	4	25.0	4
M	Clade IVa	N	0	0	N/A	591	133	22.5	6
N	Clade IIIa	N	0	0	N/A	383	40	10.4	2
O	Trichocephalida	N	148	45	30.4	0	0	N/A	5
P	Schistosomatids	P	313	35	11.2	0	0	N/A	8
Q	Trematodes	P	0	0	N/A	121	17	14.0	3
R	Clade IIIb	N	0	0	N/A	112	11	9.8	5
S	Strongyloidoidae	N	762	110	14.4	0	0	N/A	5
T	Clade IIIc	N	0	0	N/A	7	0	0.0	15
U	Strongylomorpha	N	20	8	40.0	37	11	29.7	15
V	Schistosomatidae	P	93	13	14.0	0	0	N/A	7

Continued from previous page

			‘complete presence’			‘partial presence’			
Node ID	Name	Phylum	Count	RFA	RFA (%)	Count	RFA	RFA (%)	Proteomes
W	Hymenolepididae	P	225	30	13.3	0	0	N/A	3
X	Taeniidae	P	93	7	7.5	0	0	N/A	5
Y	Clade Vc	N	0	0	N/A	2	0	0.0	7
Z	Clade Va	N	0	0	N/A	8	2	25.0	5
1	Clade Vb	N	0	0	N/A	30	1	3.3	3
2	Filarioidea	N	26	1	3.8	34	3	8.8	12
	All		3512	592	16.86	2369	568	23.98	91

3.4.4 Conclusion

KinFin analysis of synapomorphic clusters was a valuable addition to other analysis performed by collaborators. For instance, it revealed that Clade IVa (Strongyloididae) displays the highest number of synapomorphic GPCR chemosensory clusters of any group. This was missed by other analyses as 'bait' sequences from *C. elegans* were too dissimilar to identify Clade IVa GPCRs (Matthew Berriman, 2017, *pers. comm.*).

Reanalysis of the origins of the HGT-derived ferrochelatase in filarial nematodes, based on the synapomorphic Clade IVa FeCL cluster, lead to a hypothesis regarding the time point of the acquisition of the functional FeCH through HGT, which must have occurred before the split of Clade III and IV. Furthermore, identification of synapomorphic clusters in Neodermata, functionally linked to membrane maintenance and vesicle transport, might reveal effective targets for control or treatment. It should be noted that each of the synapomorphic clusters at nodes comprised solely of parasitic taxa harbours the potential to be relevant for understanding their biology since these proteins are either absent from other taxa or sufficiently different to be clustered with other proteins.

3.5 Use case 3: Analysis of gene families in Ecdysozoa

3.5.1 Introduction

Previous work on the genome of the tardigrade *Hypsibius dujardini* by members of the Blaxter lab (Koutsovoulos et al., 2016) led to a collaboration with the Arakawa lab which generated genome and transcriptome data for another hypsibiid tardigrade, *Ramazzottius varieornatus*. Both species belong to the same class of tardigrades (Eutardigrada), but the terrestrial species *R. varieornatus* readily enters anhydrobiosis and is resistant to desiccation (Horikawa et al., 2008), while the limnoterrestrial tardigrade *H. dujardini* requires prolonged pre-exposure to drying conditions (Kondo, Kubo, and Kunieda, 2015). Hence, both taxa serve as complementary model organisms for the study of the molecular processes of anhydrobiosis in tardigrades.

Tardigrades are members of the superphylum Ecdysozoa together with arthropods, onychophorans, nematodes, nematomorphs, priapulids, kinorhynchs, and loriciferans (Dunn et al., 2008), but the phylogenetic relationships between these groups is under debate, as approaches based on morphological, developmental and molecular traits yield conflicting results (Campbell et al., 2011; Borner et al., 2014). Traditionally, Ecdysozoa is divided into three subgroups based on morphological and developmental traits: Panarthropoda (Arthropoda, Onychophora, and Tardigrada), Nematodida (Nematoda and Nematomorpha), and Scalidophora (Priapulida, Kynorhyncha, and Loricifera) (Nielsen, 2013; Telford et al., 2008). However, molecular phylogenies consistently recover Tardigrada as a sister phylum to nematodes and nematomorphs (Dunn et al., 2008; Campbell et al., 2011; Borner et al., 2014), thus contradicting the Panarthropoda hypothesis.

KinFin allows computation of synapomorphic clusters under a given tree topology. By supplying alternative tree topologies the effect on counts of synapomorphic clusters at key nodes can be compared. This can be seen as an analysis of gene/protein family birth, a form of rare genomic change *sensu* Rokas and Holland, 2000, and be used to evaluate competing phylogenetic hypotheses.

The collaboration between the Arakawa and the Blaxter lab was initiated in order to pool resources and data with the aim of improving genome assemblies and gene annotations for *R. varieornatus* and *H. dujardini*, by reducing the effect of heterozygous regions on the assembly and removing residual contamination. Furthermore it was aimed at investigating the molecular machinery underpinning anhydrobiosis, resolving the phylogenetic position of tardigrades within Ecdysozoa, and assessing the extent of horizontal gene transfer (HGT) in both genomes. I carried out protein clustering analysis based on the protein predictions of tardigrade genomes assembled by Yuki Yoshida and Georgios Koutsovoulos in addition to publicly available proteomes derived from genomes of other ecdysozoan taxa. I conducted KinFin analyses based on these clusterings under alternative phylogenetic hypothesis and investigated patterns of synapomorphies and representative functional annotations of clusters.

3.5.2 Methods

Data

I was supplied with proteomes derived from the genome assemblies of the tadigrades *R. varieornatus* and *H. dujardini* by Yuki Yoshida and Georgios Koutsovoulos. Protein predictions from genomes of Annelida (*Capitella teleta*), Nematoda (*Ascaris*

suum, *Brugia malayi*, *Bursaphelenchus xylophilus*, *Caenorhabditis elegans*, *Meloidogyne hapla*, *Plectus murrayi*, *Pristionchus pacificus*, *Trichuris muris*, and *Trichinella spiralis*), Arthropoda (*Anopheles gambiae*, *Apis mellifera*, *Acyrthosiphon pisum*, *Cimex lectularius*, *Dendroctonus ponderosae*, *Daphnia pulex*, *Ixodes scapularis*, *Nasonia vitripennis*, *Pediculus humanus*, *Plutella xylostella*, *Solenopsis invicta*, *Strigamia maritima*, *Tribolium castaneum*, *Tetranychus urticae*, and *Drosophila melanogaster*), Mollusca (*Octopus bimaculoides*), and Priapulida (*Priapulus caudatus*) were retrieved from public databases as described in Yoshida et al., 2017a.

KinFin analysis

Protein clustering and functional annotation was generated as described in Yoshida et al., 2017a. In brief, protein clustering was carried out at MCL inflation values of 1.1, and 1.5 – 5.0 (in increments of 0.5), to investigate robustness of conclusions based on alternative phylogenetic hypotheses. For all other analyses, the protein clustering under the MCL inflation value 1.5 was used. `OrthoFinder` clustering output was analysed using `KinFin v0.8.2` (Laetsch, 2017a) under two competing phylogenetic hypotheses: either 'Tardigrada+Arthropoda' (Panarthropoda hypothesis), where Tardigrada and Arthropoda share a concestor, or 'Tardigrada+Nematoda' (Triradiata hypothesis, due to shared pharynx morphology between the taxa), where Tardigrada and Nematoda share a concestor. Single-copy orthologues between *H. dujardini* and *R. varieornatus*, 'true' and 'fuzzy' 1-to-1 clusters between all species, and synapomorphic clusters under the two competing phylogenetic hypotheses were identified using KinFin output.

A network representation of the `OrthoFinder` clustering at MCL inflation value 1.5 was generated using the `generate_network.py` script distributed with

KinFin. The nodes in the graph were positioned using the ForceAtlas2 layout algorithm (Jacomy et al., 2014) implemented in `Gephi v0.9.1` (Bastian, Heymann, and Jacomy, 2009) (`Scaling = 10000.0`, `Stronger gravity = True`, `Gravity = 1.0`, `Dissuade hubs = False`, `LinLog mode = True`, `Prevent overlap = False`, `Edge Weight Influence = 1.0`).

3.5.3 Results

Analysis of protein clustering using KinFin

The clustering at MCL inflation value 1.5 of the 537,608 proteins in the dataset yielded 144,610 clusters, of which 125,951 were singletons, accounting for 23.43% of proteins and 11.6% of amino acid span (cumulative amino acid length). Clusters shared by two or more species accounted for the majority of amino acid span (87.9%), while comprising 12.1% of clusters. *H. dujardini* displayed more species-specific clusters than *R. varieornatus* and contained more paralogues in clusters shared with *R. varieornatus*. *H. dujardini* was also found in more clusters shared with non-tardigrade species, suggesting gene loss in *R. varieornatus*. I found 1486 tardigrade-specific clusters, of which 365 (24.56%) received a RFA, including 53 peptidase clusters, 27 kinase clusters, and 29 clusters associated with signalling function of which 18 were GPCRs. These annotations are commonly found in clade-specific protein families and suggest innovation in these classes of function is a general feature in metazoan evolution. However, certain tardigrade-specific clusters were part of the Wnt signalling pathway, including homologues to Wnt, Frizzled, and chibby proteins. 21 tardigrade-specific clusters linked to cryptobiosis

were found containing domain annotations connected to genome repair and maintenance, including molecular chaperones (2), histone/chromatin maintenance proteins (11), genome repair systems (4), nucleases (2), and chromosome cohesion components (2).

I supplied Georgios Koutsovoulos with 21 'true' and 2144 'fuzzy' 1-to-1 clusters between all species. He screened 'fuzzy' 1-to-1 clusters to eliminate outparalogues and generated a phylogeny for ecdysozoan taxa (Figure 3.5.1).

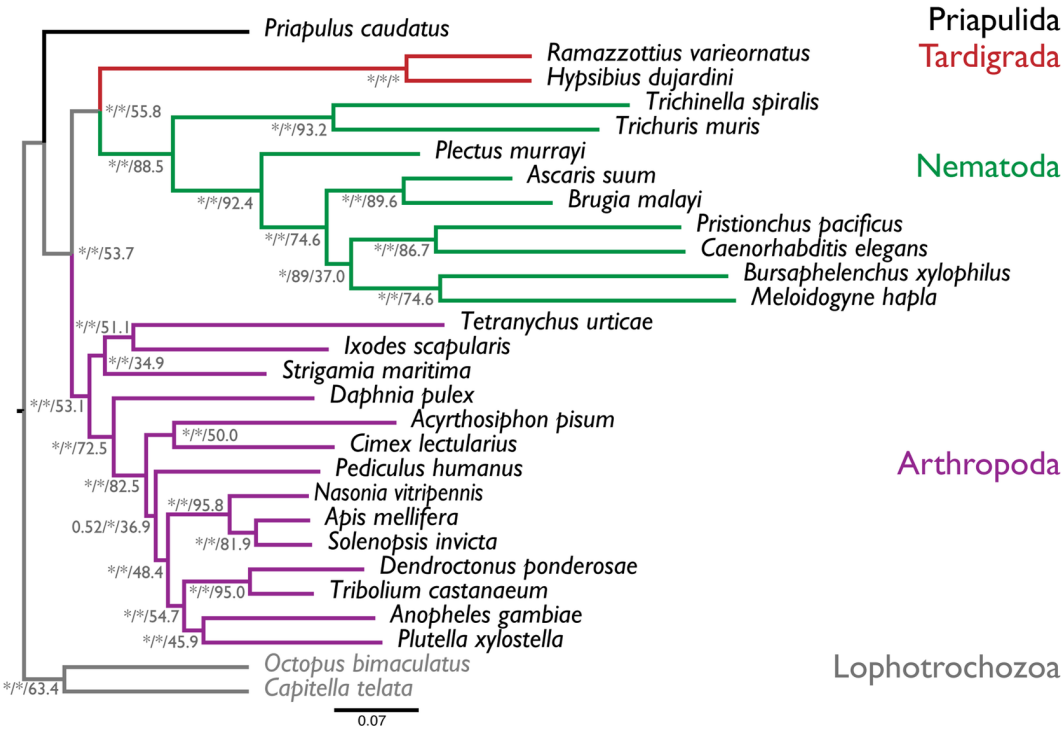


Figure 3.5.1: Phylogenetic tree of ecdysozoan phyla. Phylogeny of 28 species from 5 phyla, based on 322 loci derived from whole genome sequences, and rooted with the lophotrochozoan outgroup. The labels on the nodes are Bayes proportions from PhyloBayes analysis / bootstrap proportions from Randomized Axelerated Maximum Likelihood (RAxML) maximum likelihood bootstraps / proportion of trees of individual loci supporting each bipartition. Note that different numbers of trees were assessed at each node, depending on the representation of the taxa at each locus. * indicates maximal support (Bayes proportion of 1.0 or RAxML bootstrap of 1.0). From Yoshida et al., 2017a.

The supermatrix phylogeny strongly supported Tardigrada as a sister phylum to Nematoda. Within Nematoda and Arthropoda, the relationships of species were congruent with previous analyses, and the earliest branching taxon in Ecdysozoa was the priapulid. Support was high across the phylogeny, with only two internal nodes in Nematoda and Arthropoda receiving less-than-maximal support. Developmental and anatomical data do not, in general, support a tree linking Tardigrada with Nematoda. Tardigrades are segmented, have appendages, and have a central and peripheral nervous system anatomy that can be homologised with those of Onychophora and Arthropoda (Gross and Mayer, 2015; Martin et al., 2017). In contrast, nematodes are unsegmented, have no lateral appendages, and have a simple nervous system. The triradiate pharynx — found in Nematoda, Nematomorpha, and Tardigrada — is one possible morphological link, but Nielsen, 2013 has argued that the structures of this organ in nematodes and tardigrades (and other taxa) are not homologous and have evolved independently.

Network representation of the clustering

A network representation of the clustering where nodes represent proteomes and edges are weighted by the number of shared occurrences of proteomes in clusters, is shown in Figure 3.5.2. This type of visualisation allows exploration of protein clusterings from the point of view of proteomes as opposed to clusters. Proteins from tardigrade genomes occupy more clusters shared with arthropod taxa than with nematode taxa, which is visible from the thickness of edges which is scaled by edge weight.

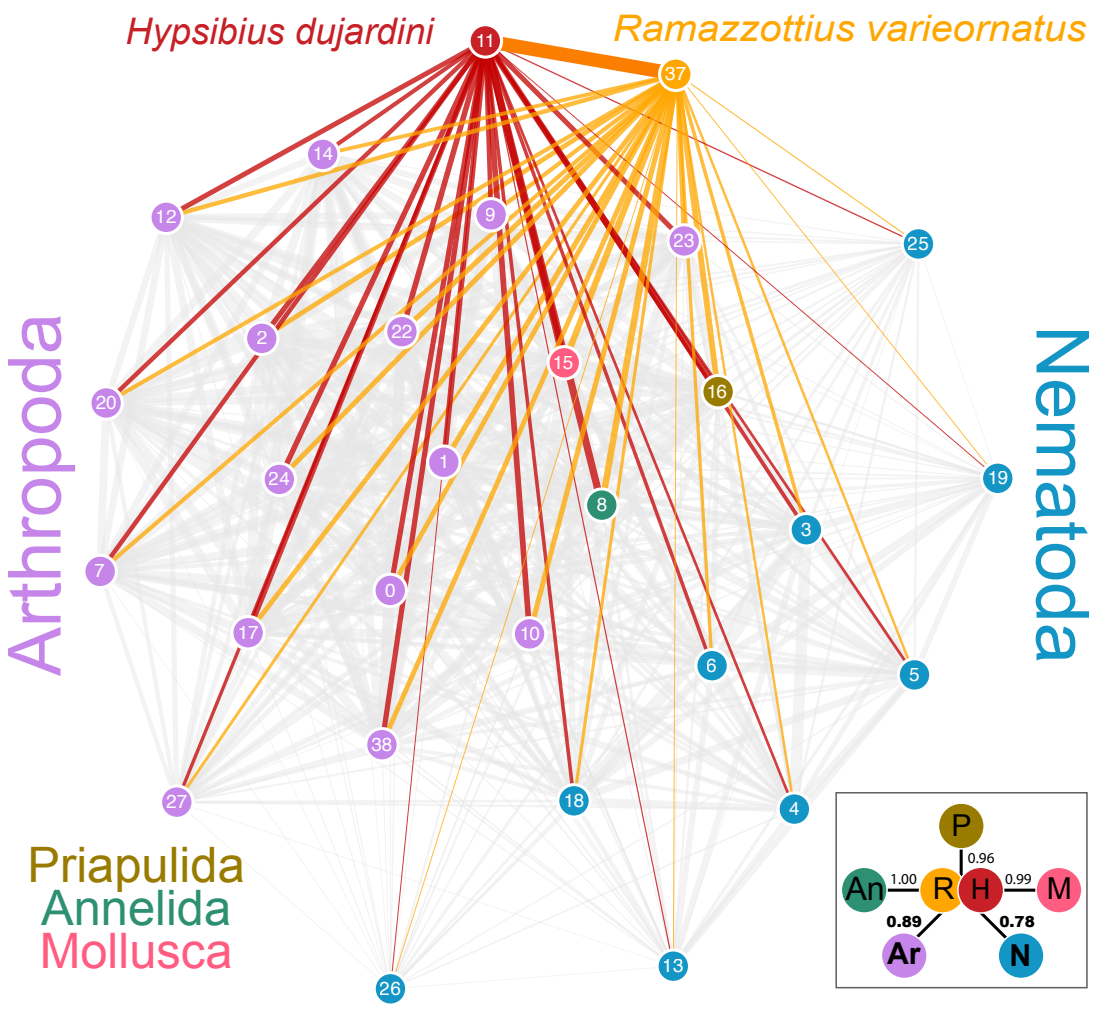


Figure 3.5.2: Network representation of the clustering. Nodes represent proteomes. The thickness of an edge connecting two nodes is weighted by the count of shared occurrences of proteomes in clusters. Links involving *H. dujardini* (red) and *R. varieornatus* (orange) are coloured. The inset box on the lower right shows the average weight of edges between each phylum and both Tardigrades, normalised by the maximum weight (*i. e.* count of co-occurrences of Tardigrades and the annelid *C. teleta*). 0: *A. gambiae*, 1: *A. mellifera*, 2: *A. pisum*, 3: *A. suum*, 4: *B. malayi*, 5: *B. xylophilus*, 6: *C. elegans*, 7: *C. lectularius*, 8: *C. teleta*, 9: *D. ponderosae*, 10: *D. pulex*, 11: *H. dujardini*, 12: *I. scapularis*, 13: *M. hapla*, 14: *N. vitripennis*, 15: *O. bimaculoides*, 16: *P. caudatus*, 17: *P. humanus*, 18: *P. murrayi*, 19: *P. pacificus*, 20: *P. xylostella*, 37: *R. varieornatus*, 22: *S. invicta*, 23: *S. maritima*, 24: *T. castaneum*, 25: *T. muris*, 26: *T. spiralis*, 27: *T. urticae*, 38: *D. melanogaster*.

Investigation of synapomorphic clusters under competing phylogenetic hypotheses

Synapomorphies were evaluated across clusterings based on nine MCL inflation values and the two competing tree topologies (Panarthropoda vs. Triradiata) and counts are visualised in Figure 3.5.3. Allowing for partial absence, I found 154 families to be synapomorphic congruent with Panarthropoda, and 99 congruent with Triradiata. Of those, 20 under Panarthropoda and five under Triradiata contained proteins from both tardigrades and at least 5 other taxa (out of 9 nematodes and 15 arthropods). This trend was observed for clusterings across all MCL inflation values tested.

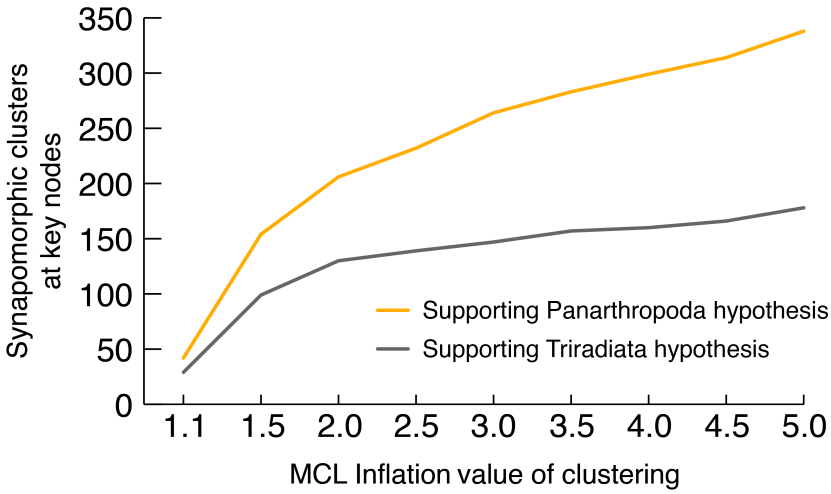


Figure 3.5.3: Count of synapomorphies under alternative phylogenetic hypotheses. Count of synapomorphies (allowing for partial absence) at key nodes supporting Panarthropoda hypothesis (orange) and Triradiata hypothesis (grey) for clusterings performed at different MCL inflation parameters.

At inflation value 1.5, I found six synapomorphic clusters where all taxa were present congruent with Panarthropoda, while not a single such cluster was found under the Triradiata hypothesis. The six loci identified as universally retained protein families in Panarthropoda included spaetzle, a cysteine-knot/cytokine-like family that is known to interact with the Toll receptor pathway in *D. melanogaster*,

where it is involved in dorso-ventral patterning as well as immune response. Other clusters were functionally annotated as having serine-type endopeptidase activity or harbouring a thioredoxin domain and thus being involved in cell redox homeostasis. However the remainder of the clusters had no informative annotation other than the presence of domains of unknown function (DUFs). Again, it is surprising that such deeply conserved loci have escaped functional, genetic and biochemical annotation.

3.5.4 Conclusion

Protein clusterings and their analyses using KinFin allowed investigation of synapomorphic clusters under competing phylogenetic hypotheses, which subsequently served as substrate for further analyses by collaborators. The visualisation of the clustering as a network, where nodes represent proteomes and edges are weighted by the number of shared occurrences in clusters, enables a novel, proteome-focussed view on clustering results which needs to be explored further.

Assessment of the extent of synapomorphic clusters at key nodes under the competing phylogenetic hypotheses, a form of rare genomic change, lent support to the Panarthropoda hypothesis (Tardigrada and Arthropoda sharing a common ancestor), but the support was not strong. Analyses under the assumption of the Tardigrada hypothesis (Tardigrada and Nematoda sharing a concestor) identified synapomorphic clusters at about half the rate than when Panarthropoda was assumed. However, it should be noted that recognition of synapomorphic protein families may be compromised by the same long branch attraction issues that plague phylogenetic analyses (Dunn et al., 2008; Campbell et al., 2011; Borner et al., 2014), and also that any taxon where gene loss is common, which has been suggested for Nematoda (Wasmuth et al., 2008), may score worse in protein family membership

metrics. Results of phylogenetic analysis conflict with the findings based on synapomorphies as they recover a topology congruent with the Triradiata hypothesis. This is supported by other rare genomic changes, such as the loss of the same three HOX genes in nematodes and tardigrades (see HOX protein analysis by Mark Blaxter in Yoshida et al., 2017a). HOX genes are involved in the anterior-posterior patterning across Metazoa.

Hence, the position of Tardigrada within Ecdysozoa remains an open question. Clearly, more genomic data is needed, especially from representatives of Onychophora, Heterotardigrada (the sister group to Eutardigrada), Nematomorpha, and enoplian (basal) Nematoda. This will hopefully allow construction of a robust phylogenetic tree.

3.6 Use case 4: Analysis of gene families in Nematoda

3.6.1 Introduction

Development of KinFin was initially sparked by the idea of analysing protein families across all available nematode proteomes derived from both genomes and transcriptomes. The underlying code has been designed with this magnitude of data in mind. WormBase ParaSite (version WBPS8) contains 100 nematode proteomes and the Blaxter Lab has generated protein predictions from 25 additional genomes and five transcriptomes, totalling 125 nematode proteomes from 107 species.

Here, I present a clustering analysis of these nematode proteomes and 26 outgroup species which was used to assess performance of KinFin and will serve as substrate for further analysis by colleagues. Furthermore, I briefly explore the effect of inclusion of isoforms on results of protein clusterings, assess the protein space uncovered by these proteomes for taxonomic clades within Nematoda, and present a network representation of the protein clustering.

3.6.2 Methods

Data

Proteomes and annotation files (GFF3) were retrieved by Duncan Berger and myself, for the 151 species listed in Table 3.A.2. Proteomes were filtered using the script `filter_fastas_before_clustering.py` distributed with `KinFin`

`v0.8.3` (Laetsch, 2017a), to exclude sequences shorter than 30 amino acids or containing internal stops.

Protein clustering

`BLAST` commands were generated using using `GNU parallel` (Tange, 2011) based on recommendations by Moreno-Hagelsieb and Latimer, 2008 (`-evalue 1e-5 -outfmt '6' -seg yes -soft_masking true -use_sw_tback`). The 22,801 sequence similarity searches between proteomes were run using `BLAST v2.4.0+` (Camacho et al., 2009) on the EDDIE supercomputing cluster at the University of Edinburgh.

Sequence IDs of non-representative isoforms for each proteome were determined using the KinFin script `filter_isoforms_based_on_gff3.py`. The sequence similarity search results were filtered to exclude hits between and within non-representative isoforms using the script `filter_sequences_from_blast.py` to generate a second set of results from which non-representative isoforms were excluded.

Protein clustering was carried out using `OrthoFinder v1.1.4` (Emms and Kelly, 2015), at 9 different MCL inflation values (1.5 – 5.0 in increments of 0.5) for both sets of BLAST results, including all isoforms ('AI') and only including representative isoforms ('RI'). The resulting clusterings based on BLAST results containing only the representative isoforms (where all non-representative isoforms are found in singleton clusters due to the lack of BLAST results), were filtered with the KinFin script `filter_sequences_from_clustering.py`.

Functional annotation of proteins was generated based on `InterProScan`

v5.22-61.0 (Jones et al., 2014) results against PFAM v30.0 (Finn et al., 2016) and SignalP-Euk v4.1 (Petersen et al., 2011).

KinFin analysis

KinFin analysis was carried out for each of the protein clusterings using KinFin v0.8.3 (Laetsch, 2017a) by providing a functional annotation file created using the script functional_annotation_of_clusters.py based on the InterProScan output. A rarefaction curve was drawn by defining custom taxonomic groups for Nematoda based on Blaxter et al., 1998 and Blaxter and Koutsovoulos, 2015 within the config file. Rarefaction curves were calculated for taxon sets of size two or greater and with 30 repetitions of random sampling of the proteomes.

KinFin output for the protein clustering at MCL inflation value 3.0 and based on the 'RI' proteome set was used to generate a network representation of the clustering using the script generate_network.py. The network was visualised using Gephi v0.9.1 (Bastian, Heymann, and Jacomy, 2009). Nodes were positioned based on the ForceAtlas2 layout algorithm (Jacomy et al., 2014) by starting from a random layout (Scaling = 10000.0, Stronger gravity = True, Gravity = 1.2, Dissuade hubs = True, LinLog mode = True, Prevent overlap = True, Edge Weight Influence = 1.0). Nodes were scaled based on proteome size and coloured based on taxonomic groupings.

3.6.3 Results

Composition of the proteomes

The set of proteins including isoforms contained 3,162,746 proteins with a cumulative length of 1,154,475,022 amino acids. Exclusion of non-representative isoforms lead to a dataset composed of 2,835,046 proteins with a total span of 982,530,736 amino acids. The number of predicted isoforms varies between proteomes depending on the quality of gene annotations of the underlying genomes and the data used to infer them, since proteomes based on transcriptomes always include isoforms. In this dataset, 83 out of 151 proteomes contained no isoforms. Percentages of non-representative isoforms above 50% were encountered in the proteomes of *H. sapiens* (80.50%), *Propanagrolaimus* sp. JU765 (61.60%, based on an unpublished version of the genome), and *D. melanogaster* (54.20%). While *H. sapiens* and *D. melanogaster* are well established model organisms the result for the panagrolaimid nematode was puzzling. However, the recent publication of the genome revisited the gene predictions, which revised the amount predicted proteins from 32,914 to 27,350 (Schiffer et al., 2017).

Protein clustering

Clustering of the proteomes including all isoforms ('AI') yielded between 1,590,254 (MCL inflation value 1.5) and 1,927,885 (5.0) clusters. If isoforms were excluded ('RI'), the number of clusters varied between 1,540,385 (1.5) and 1,870,356 (5.0). In both clusterings, no 'true' 1-to-1 clusters could be identified which is most likely an effect of the reduced quality of some of the included proteomes.

The number and span of proteins was evaluated based on the type of cluster

they were placed in: 'singleton' (containing only one protein), 'specific' (composed of multiple proteins from one proteome), and 'shared' (consisting of multiple proteins from multiple proteomes). A clustering becomes more granular with increasing MCL inflation value, *i. e.* more proteins end up in 'singletons' or 'specifics' than in clusters shared with other proteomes. The extent of this effect is summarised in Table 3.6.1 for the clustering of two sets of proteomes across the nine MCL inflation values. The proportion of proteins and amino acids placed in different cluster types is fairly consistent between the two sets of proteomes and across MCL inflation values: most proteins and amino acids are placed in shared clusters. A smaller proportion is contained in singleton clusters, while less than three percent of proteins and amino acids are grouped in proteome specific clusters. Unsurprisingly, the 'RI' proteome set has slightly higher mean percentages for both proteins and amino acids placed in singleton clusters, caused by the absence of isoforms which otherwise would place them in proteome specific clusters. The 'AI' proteome set exhibits slightly higher SD values which suggest a greater effect of MCL inflation value on the clustering. It should be noted that the effect of phylogenetic distance between proteomes probably has a great influence on the resulting clusterings in this dataset. Since for some species multiple proteomes were included, a significant proportion of clusters will appear shared although they are composed of proteomes of the same species.

KinFin analysis

KinFin analysis of the protein clusterings was carried out on the Blaxter lab computing cluster (using one cpu thread). Parsing of the input files took less than 10 minutes in all cases, but computing of metrics and writing of output files took on

Table 3.6.1: Effect of inclusion of isoforms on protein clusterings. Population mean (μ) and SD (σ) of percentage of proteins and amino acids placed in the three cluster types based on sets of proteomes. 'AI': including all isoforms. 'RI': including only representative isoforms.

		Cluster type					
		Singleton		Specific		Shared	
	Proteome set	μ	σ	μ	σ	μ	σ
Proteins (%)	'AI'	13.37	0.10	2.78	0.88	83.85	0.97
Amino acids (%)	'AI'	5.85	0.08	2.45	0.78	91.70	0.85
Proteins (%)	'RI'	13.70	0.11	2.56	0.79	83.74	0.90
Amino acids (%)	'RI'	6.17	0.10	2.33	0.72	91.50	0.82

average 15.6 hours per clustering and required 95.8 GB of memory. While this is acceptable for a dataset of this size, improvements are planned to remove residual redundancy in the computational steps and to decrease memory requirements.

KinFin output of the protein clustering at MCL inflation value 3.0 based on the 'RI' set of proteomes was used to calculate rarefaction curves for custom taxonomic groups (Figure 3.6.1) and was visualised as a network, in which nodes represent proteomes and edges between nodes are weighted by the number of co-occurrences of proteomes in clusters (Figure 3.6.2).

The rarefaction curves are a common form of visualisation of sampling success. Here, they are used to display the amount of novel non-singleton clusters uncovered by successive addition of proteomes for each taxonomic group. Clade V is the best sampled clade within Nematoda, containing model organisms such as *C. elegans* and

many other free-living and animal-parasitic species. Clade IV, containing species from both free-living species and parasites of animals, plants and invertebrates, exhibits the second steepest growth in the plot, despite containing less sampled proteomes than the entirely animal-parasitic Clades III and I, which is currently composed of animal parasites and one free-living congener. This is suggestive of an enlarged protein space comprised by Clade IV proteomes linked to the varied life styles. While both Clades III and I appear to approach a plateau, this is most likely an artefact based on prominent sampling of species of medical and veterinary importance. Inclusion of free-living taxa in Clade I and invertebrate parasitic or basal taxa in Clade III, would most likely disturb this apparent trend. Neither Clade V nor Clade IV show signs of reaching a plateau, while no data is available for Clade II nematodes.

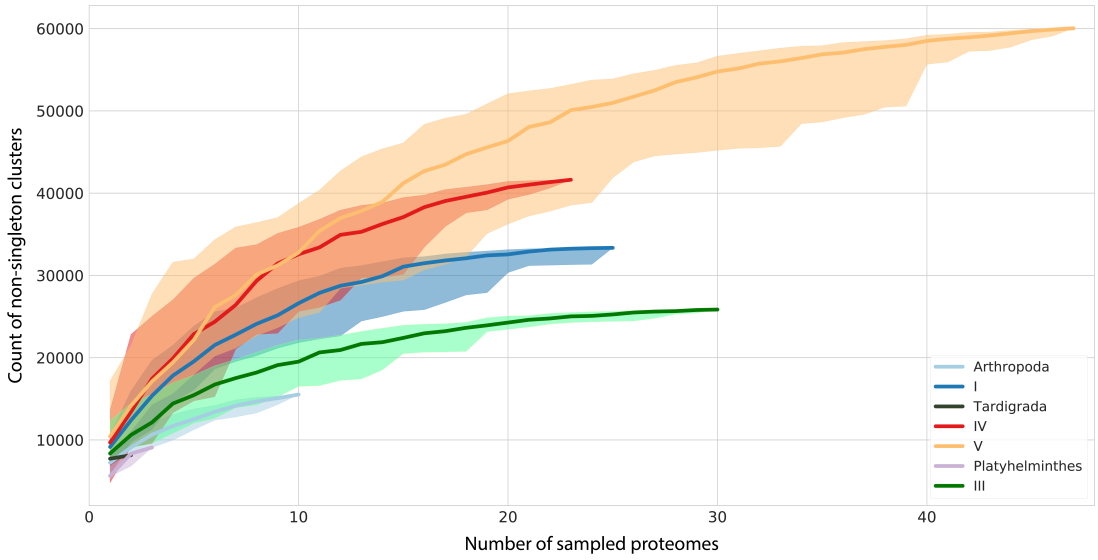


Figure 3.6.1: Rarefaction curves for taxonomic groupings proteomes. Taxonomic groupings of Nematoda was based on Blaxter et al., 1998 and Blaxter and Koutsovoulos, 2015.

The two proteomes in the analysis containing the greatest number of proteins are the Clade I nematode *R. culicivorax* (48,179) and the Clade IV nematode *M. floridensis* (47,957). In both cases, this is most likely a result of the low quality of the underlying genomes. While the network representation appears to recover some of the general taxonomic groupings, this is not a stable configuration and

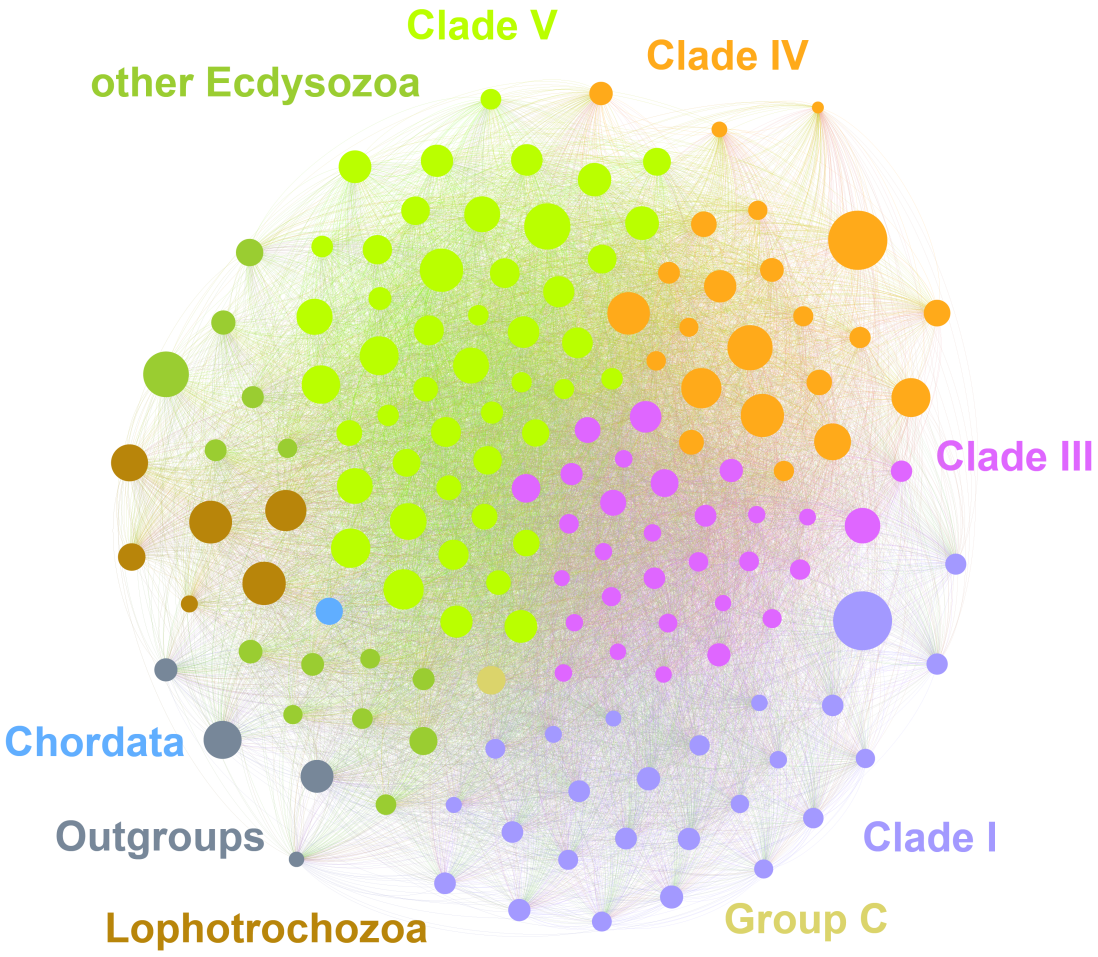


Figure 3.6.2: Network representation of a protein clustering of proteomes. Nodes represent proteomes. The thickness of an edge connecting two nodes is weighted by the count of shared occurrences of the proteomes in clusters. Nodes are scaled based on proteome size and coloured based on taxonomic groupings. For nematode species, colouring was based on clade/group *sensu* Blaxter et al., 1998 and Blaxter and Koutsovoulos, 2015.

it is difficult to draw conclusions based on it. The extreme connectivity between nodes hinders the layout algorithm from arriving at a stable topology and clearly, more work is needed to address this. However, the network representation is a useful visualisation for rapid visual identification of outliers in the clustering, since well connected proteomes are drawn to the centre of the network and low quality proteomes often are tend to remain in the periphery.

3.6.4 Conclusion

Based on the modularity of the `OrthoFinder` clustering pipeline, the BLAST results I generated can serve as a resource for future studies, since addition of a new proteome only requires $n-1$ novel BLAST searches, as opposed to n^2-1, where n is the number of non-novel proteomes to be included in the clustering.

This use case showed that KinFin is able to process large scale datasets efficiently and revealed several issues with the quality of some of the proteomes used. Based on the clusterings I generated, Flo McLean and Duncan Berger were able to identify major problems with the proteome of the Clade V nematode *H. bacteriophora*, which prompted a re-annotation project and subsequently led to an improved annotation for this species. Visualisation of the rarefaction curves for the four clades in Nematoda revealed differences in their contribution to nematode protein space. Addition of new proteomes of Clade V and Clade IV species reveals more 'novelty' than those of Clade III and Clade I, which is most likely a product of sampling strategies of nematode genome projects.

However, much more work is needed to identify outliers in the dataset that are based on low quality proteomes and assess the protein family space across Nematoda.

3.7 Kinfin facilitates large scale analysis of proteome data

As ever more genomes are sequenced and our understanding of the diversity of protein space increases, it concomitantly becomes more difficult to see the patterns in complex orthology data. To ease this bottleneck I have developed KinFin, a tool that takes the output of standard orthology inference pipelines and provides a user-friendly but rich analytical toolkit to review and interrogate orthology clustering. By permitting user definition of custom taxon sets, KinFin can be used to highlight changes in presence or membership of orthologue groups associated with either taxonomy or phenotypes of interest. Its reliance on standard input file formats and explicit parameters makes integration in comparative genomics projects easier, and thus promotes transparent and reproducible analysis of clustered protein data.

It should be noted that up until now, due to the lack of software solutions capable of analysing protein clustering data using custom user input, results of protein clustering data have essentially been 'lost to science' since very few researchers deposit their clustering data and the subsequent analysis workflow in a way that allows reproducibility. Formalisation of this type of analysis within KinFin now guarantees reproducibility given that the config file of the KinFin run, the parameters used, and the necessary input files are published.

KinFin readily processes large datasets, as shown in Section 3.4 and 3.6 and the speed of execution promotes hypothesis exploration, such as comparing alternative phylogenetic topologies (Section 3.5), or contrasting taxon sets specified in the config files. Visualisations of rarefaction curves (Section 3.3) and network representations (Section 3.6) allow identification of outlier datasets.

In this chapter, several insights were gained into the biology of the analysed taxa. In Section 3.3, I investigated protein families involved in haem metabolism in filarial nematodes, on which I expanded in Section 3.4 by analysing proteomes of nematodes and platyhelminths. I could form a hypothesis regarding the acquisition of a functional FeCH gene through horizontal gene transfer in Clade III and Clade IV nematodes and discovered a novel, synapomorphic FeCL protein family in Strongylodidae. In addition, I uncovered many synapomorphic protein families in both nematodes and platyhelminths which might be linked to parasitism and await further analysis. In Section 3.5, I showed how KinFin can be used for exploration of alternative evolutionary hypotheses, which recovered evidence for the Panarthropoda hypothesis in the phylogeny of ecdysozoans, suggesting that Tardigrada and Arthropoda are sister clades. This is based on the number of recovered synapomorphic clusters shared between both groups. However, the phylogenetic position of tardigrades remains under debate, as this result conflicts with other conclusions derived from the same data, such as phylogenetic analysis and HOX gene complement which both support the Triraditata hypothesis in which Nematoda and Tardigrada share a concestor. More data from under-sampled taxonomic groups are needed to address this question. In Section 3.6, I present preliminary work on a protein family analysis of 128 nematode and 23 outgroup proteomes, where I assess effect of inclusion of isoforms in clustering analysis and analyse contribution of clades *sensu* Blaxter et al., 1998 to the nematode protein space. This dataset has and will be a valuable resource for further research.

While KinFin is a useful addition to the comparative genomics toolbox, plenty of room for improvement exists. Mark Blaxter has recently submitted a NERC proposal to tackle unsolved questions regarding the phylogenetic relationships within the superphylum Ecdysozoa, on which I am a named researcher and in which KinFin would feature prominently. This would allow me to continue work on the

KinFin code base and develop new analyses and visualisations. Some of the features I plan to implement are as follows:

- Refactoring of code to accommodate interaction with a SQL database for storage of results and reduction of overall runtime, since often users do not need all results currently being calculated by KinFin. This would also decrease memory requirements substantially, as previously calculated results can be accessed if needed but are not necessarily kept in memory at all times.

- Refactoring of code to create a 'Read-eval-print loop' (REPL) user interface for interaction with the underlying SQL database, as illustrated by Ramanujam, 2017.

- Improved visualisations of cluster memberships of sets of taxa based on the UpSet suite of visualisation methods by Lex et al., 2014.

- Integration of benchmarking functions for clusterings based on user defined priors, *e. g.* sets of proteins which are known to be *bona fide* orthologues and which can be used to assess performance of the clustering method. The same code base would also allow standardised comparison between output of different clustering pipelines.

- Integration of algorithmic infrastructure to allow exploration of additional dimensions associated with proteins and proteomes, such as arbitrary numerical data (*e. g.* expression patterns of mRNAs associated with proteins), results of sequence similarity searches against public databases (*e. g.* BLAST results for searches of eukaryotic proteomes against bacterial proteome databases), feature track data (*e. g.* genomic location of the underlying loci, which would

allow KinFin to understand the concept of isoforms and would enable analysis concerning synteny conservation between taxa and flagging of potential contaminants/HGTs within genomes), and phylogenetic data (*e. g.* tree topologies inferred for clusters or multiple alternative phylogenies for proteomes). These features would enable analysis of protein clusterings that currently are tedious to carry out for custom datasets and would allow the user to ask complex questions such as 'what are the functional annotations of clusters yielding a certain tree topology?', 'for proteins of genes that are co-localised in the underlying genomes, which domains co-occur more often than expected?', or 'which proteins shown signs of contamination or acquisition through HGT based on the genomic location, expression and sequence similarity search data?'

- Further development of network based analysis and visualisations of the proteomes, *i. e.* 'clustering of clustering data'. The current process for the network representation is purely visual and based solely on co-occurrence of proteomes and clusters which determines the weight of edges connecting the nodes. This yields highly connected networks which are currently only useful for identification of clear outliers. Using the information supplied through additional data about the proteins, network representations could be generated for these additional dimensions which would yield more granular clusters of proteomes.

- Improvements on protein family expansion analysis by expanding on the current algorithmic infrastructure for pairwise protein count representation tests to include taxon set 'specific' expansions.

- Development of an additional graphical user interface for simple interaction with analysis and visualisation of the data.

The magnitude and multi-dimensionality of the data generated by the field of comparative genomics calls for new approaches capable of handling the amount of data and being aware of the biological connections between the different dimensions. KinFin is a robust starting point for the development of a versatile toolkit which could one day serve as an 'operating system' for comparative genomics.

Appendix

3.A Tables

Table 3.A.1: Data used in the Ensembl Compara clustering. Proteomes used by the WTSI to generate the Ensembl Compara clustering which was provided to me and used in Section 3.4. 'Proteins in clustering': count of proteins which were allocated to Compara families. Percentage of total proteins for each species is listed in brackets. Citation: Reference for the genome assemblies, if not sequenced within the 50 Helminth Genome Project.

Species	Assembly version	Proteins in clustering	Citation
Amphimedon queenslandica	1.0	24,399 (82%)	Fernandez-Valverde, Calcino, and Degnan, 2015
Capitella teleta	1.0	27,552 (86%)	Simakov et al., 2013
Ciona intestinalis	1.0	10,848 (65%)	Satou et al., 2008
Crassostrea gigas	oyster_v9	21,032 (81%)	Zhang et al., 2012
Danio rerio	9.0	25,214 (95%)	Howe et al., 2013
Drosophila melanogaster	5.0	10,862 (78%)	Adams et al., 2000
Homo sapiens	GRCh37	18,736 (90%)	Lander et al., 2001
Ixodes scapularis	1.0	13,111 (64%)	Pagel Van Zee et al., 2007
Nematostella vectensis	ASM20922v1	20,628 (83%)	Putnam et al., 2007
Trichoplax adhaerens	1.0	9577 (83%)	Srivastava et al., 2008

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Schistosoma curassoni	1.0.4	20,722 (88%)	-
Schistosoma haematobium	3.0	12,538 (96%)	Young et al., 2012
Schistosoma japonicum	1.0	11,672 (92%)	SGSFA Consortium, 2009
Schistosoma mansoni	5.2	10,290 (95%)	Berriman et al., 2009
Schistosoma margrebowiei	1.5.4	23,827 (91%)	-
Schistosoma mattheei	1.0.4	20,170 (88%)	-
Schistosoma rodhaini	1.0.4	20,380 (85%)	-
Trichobilharzia regenti	1.0.4	17,248 (78%)	-
Clonorchis sinensis	3.5	11,485 (84%)	Wang et al., 2011b
Echinostoma caproni	1.0	15,047 (81%)	-
Fasciola hepatica	1.0.4	13,258 (84%)	McNulty et al., 2017
Diphyllobothrium latum	1.0	16,335 (82%)	-
Echinococcus granulosus	1.0	9576 (93%)	Tsai et al., 2013

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Echinococcus multilocularis	1.0.4	9867 (97%)	Tsai et al., 2013
Hydatigera taeniaeformis	1.0.4	10,307 (88%)	-
Hymenolepis diminuta	1.0	10,252 (91%)	-
Hymenolepis microstoma	1.5.4	9090 (90%)	Tsai et al., 2013
Hymenolepis nana	1.0.4	12,272 (89%)	-
Mesocestoides corti	1.5.4	9129 (86%)	-
Schistocephalus solidus	1.5.4	17,749 (88%)	-
Spirometra erinaceieuropaei	1.0.4	33,229 (84%)	Bennett et al., 2014
Taenia asiatica	1.0.4	9446 (91%)	-
Taenia solium	1.0	11,168 (89%)	Tsai et al., 2013
Protopolystoma xenopodis	1.0.4	12,857 (34%)	-
Schmidtea mediterranea	3.1	24,228 (81%)	Robb, Ross, and Sánchez Alvarado, 2008
Romanomermis culicivorax	2.0	25,446 (53%)	Schiffer et al., 2013

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Soboliphyme baturini	1.0.4	8645 (66%)	-
Trichinella nativa	1.0	9730 (96%)	-
Trichinella spiralis	Tspiralis1	13,277 (81%)	Mitreva et al., 2011
Trichuris muris	2.0	8920 (81%)	Foth et al., 2014
Trichuris suis	1.0	8419 (86%)	-
Trichuris trichiura	2.0	8323 (94%)	Foth et al., 2014
Parastrongyloides trichosuri	2.0.4	13,196 (88%)	Hunt et al., 2016
Rhabditophanes sp. KR3021	2.0.4	11,729 (87%)	Hunt et al., 2016
Strongyloides papillosus	2.1.4	17,422 (94%)	Hunt et al., 2016
Strongyloides ratti	5.0.4	12,085 (97%)	Hunt et al., 2016
Strongyloides stercoralis	2.0.4	12,573 (96%)	Hunt et al., 2016
Strongyloides venezuelensis	2.0.4	15,640 (93%)	Hunt et al., 2016
Panagrellus redivivus	Pred3	17,845 (74%)	Srinivasan et al., 2013

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Bursaphelenchus xylophilus	1.2	13,453 (76%)	Kikuchi et al., 2011
Globodera pallida	1.0	13,585 (83%)	Cotton et al., 2014
Meloidogyne hapla	2.0	10,696 (74%)	Opperman et al., 2008
Acanthocheilonema viteae	1.0	9928 (95%)	-
Brugia malayi	3.0	12,255 (87%)	Ghedin et al., 2007
Brugia pahangi	1.5.4	13,468 (92%)	-
Brugia timori	1.0.4	13,131 (81%)	-
Dirofilaria immitis	2.2	11,098 (86%)	Godel et al., 2012
Dracunculus medinensis	2.0.4	9790 (89%)	-
Elaeophora elaphi	1.0.4	10,147 (97%)	-
Gongylonema pulchrum	1.0.4	21,421 (79%)	-
Litomosoides sigmodontis	2.1	9530 (93%)	-
Loa loa	3.0	12,036 (81%)	Desjardins et al., 2013

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Onchocerca flexuosa	1.0.4	13,881 (86%)	-
Onchocerca ochengi	2.0	13,193 (94%)	-
Onchocerca volvulus	3.0.4	10,655 (85%)	Cotton et al., 2016
Thelazia callipedia	1.0.4	9863 (90%)	-
Wuchereria bancrofti	2.0.4	12,117 (93%)	-
Anisakis simplex	1.5.4	17,305 (83%)	-
Ascaris lumbricoides	1.5.4	18,632 (79%)	-
Ascaris suum	3.0	14,073 (92%)	Wang et al., 2012
Parascaris equorum	1.0.4	9544 (66%)	-
Toxocara canis	1.5.4	16,929 (84%)	-
Enterobius vermicularis	1.0.4	10,523 (82%)	-
Syphacia muris	1.0.4	9442 (85%)	-
Ancylostoma caninum	1.0	27,463 (91%)	-

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Ancylostoma ceylanicum	1.0	15,051 (95%)	-
Ancylostoma duodenale	1.0	25,127 (91%)	-
Cylicostephanus goldi	1.0.4	11,339 (82%)	-
Necator americanus	1.0	16,082 (84%)	Tang et al., 2014
Oesophagostomum dentatum	1.0	22,949 (91%)	Tyagi et al., 2015
Strongylus vulgaris	1.0.4	17,610 (84%)	-
Angiostrongylus cantonensis	1.5.4	13,100 (90%)	-
Angiostrongylus costaricensis	1.5.4	12,271 (91%)	-
Dictyocaulus viviparus	1.0	12,200 (90%)	McNulty et al., 2016
Haemonchus contortus	1.0	20,925 (96%)	Laing et al., 2013
Haemonchus placei	1.5.4	17,693 (81%)	-
Heligmosomoides bakeri	1.5.4	23,019 (84%)	-
Nippostrongylus brasiliensis	1.5.4	19,692 (86%)	-

Table 3.A.1 Continued from previous page

Species	Assembly version	Proteins in clustering	Citation
Teladorsagia circumcincta	1.0	23,424 (92%)	-
Caenorhabditis elegans	WBcel235	16,964 (83%)	*C. elegans* Sequencing Consortium, 1998
Pristionchus pacificus	5.0	16,288 (67%)	Dieterich et al., 2008

Table 3.A.2: Proteomes used in Section 3.5. Taxon ID: ID used in the clustering. Phylum: taxonomic phylum of the species. In brackets, clade membership *sensu* (Blaxter et al., 1998). Version: Version/Accession number of the proteome. Source: Database from which proteome was retrieved. (BLAXTERLAB: proteomes not yet released publicly). Basis: whether proteome was derived from a genome or a transcriptome.

Taxon ID	Species name	Phylum	Version	Source	Basis
AAEGY	*Aedes aegypti*	Arthropoda	AaegL3	ENSEMBL-Metazoa 34	genome
ACANI	*Ancylostoma caninum*	Nematoda (Clade V)	PRJNA72585	WBPS8	genome
ACANT	*Angiostrongylus cantonensis*	Nematoda (Clade V)	PRJEB493	WBPS8	genome
ACEYL1	*Ancylostoma ceylanicum*	Nematoda (Clade V)	PRJNA231479	WBPS8	genome
ACEYL2	*Ancylostoma ceylanicum*	Nematoda (Clade V)	PRJNA72583	WBPS8	genome
ACOST	*Angiostrongylus costaricensis*	Nematoda (Clade V)	PRJEB494	WBPS8	genome
ADUOD	*Ancylostoma duodenale*	Nematoda (Clade V)	PRJNA72581	WBPS8	genome
AGAMB	*Anopheles gambiae*	Arthropoda	AgamP4	ENSEMBL-Metazoa 34	genome
ALUMB	*Ascaris lumbricoides*	Nematoda (Clade III)	PRJEB4950	WBPS8	genome
AMELL	*Apis mellifera*	Arthropoda	GCA_000002195	ENSEMBL-Metazoa 34	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
ANANU	*Acroboloides nanus*	Nematoda (Clade IV)	v1	BLAXTERLAB	transcriptome
APISU	*Acyrthosiphon pisum*	Arthropoda	GCA_000142985	ENSEMBL-Metazoa 34	genome
AQUEE	*Amphimedon queenslandica*	Porifera	Aqu1	ENSEMBL-Metazoa 34	genome
ASIMP	*Anisakis simplex*	Nematoda (Clade III)	PRJEB496	WBPS8	genome
ASUUM1	*Ascaris suum*	Nematoda (Clade III)	PRJNA62057	WBPS8	genome
ASUUM2	*Ascaris suum*	Nematoda (Clade III)	PRJNA80881	WBPS8	genome
AVITE	*Acanthocheilonema viteae*	Nematoda (Clade III)	PRJEB4306	WBPS8	genome
BMALA	*Brugia malayi*	Nematoda (Clade III)	PRJNA10729	WBPS8	genome
BPAHA	*Brugia pahangi*	Nematoda (Clade III)	PRJEB497	WBPS8	genome
BTIMO	*Brugia timori*	Nematoda (Clade III)	PRJEB4663	WBPS8	genome
BXYLO	*Bursaphelenchus xylophilus*	Nematoda (Clade IV)	PRJEA64437	WBPS8	genome
CAFRA	*Caenorhabditis afra*	Nematoda (Clade V)	JU1286	caenorhabditis.org (V2)	genome
CANGA	*Caenorhabditis angaria*	Nematoda (Clade V)	PRJNA51225	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
CBREN	*Caenorhabditis brenneri*	Nematoda (Clade V)	PRJNA20035	WBPS8	genome
CBRIG	*Caenorhabditis briggsae*	Nematoda (Clade V)	PRJNA10731	WBPS8	genome
CCAST	*Caenorhabditis castelli*	Nematoda (Clade V)	JU1956	caenorhabditis.org (V2)	genome
CDOUG	*Caenorhabditis doughertyi*	Nematoda (Clade V)	JU1771	caenorhabditis.org (V2)	genome
CELEG	*Caenorhabditis elegans*	Nematoda (Clade V)	PRJNA13758	WBPS8	genome
CGOLD	*Cylicostephanus goldi*	Nematoda (Clade V)	PRJEB498	WBPS8	genome
CJAPO	*Caenorhabditis japonica*	Nematoda (Clade V)	PRJNA12591	WBPS8	genome
CREMA	*Caenorhabditis remanei*	Nematoda (Clade V)	PRJNA53967	WBPS8	genome
CSINI	*Caenorhabditis sinica*	Nematoda (Clade V)	PRJNA194557	WBPS8	genome
CSP1	*Caenorhabditis* sp. 1	Nematoda (Clade V)	JU1667	caenorhabditis.org (V2)	genome
CSP21	*Caenorhabditis* sp. 21	Nematoda (Clade V)	NIC534	caenorhabditis.org (V2)	genome
CSP26	*Caenorhabditis* sp. 26	Nematoda (Clade V)	JU2190	caenorhabditis.org (V2)	genome
CSP31	*Caenorhabditis* sp. 31	Nematoda (Clade V)	JU2585	caenorhabditis.org (V2)	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
CSP32	*Caenorhabditis* sp. 32	Nematoda (Clade V)	JU2788	caenorhabditis.org (V2)	genome
CSP38	*Caenorhabditis* sp. 38	Nematoda (Clade V)	JU2809	caenorhabditis.org (V2)	genome
CSP39	*Caenorhabditis* sp. 39	Nematoda (Clade V)	NIC564	caenorhabditis.org (V2)	genome
CSP40	*Caenorhabditis* sp. 40	Nematoda (Clade V)	JU2818	caenorhabditis.org (V2)	genome
CTELE	*Capitella teleta*	Annelida	GCA_000328365	ENSEMBL-Metazoa 34	genome
CTROP	*Caenorhabditis tropicalis*	Nematoda (Clade V)	PRJNA53597	WBPS8	genome
CVIRI	*Caenorhabditis virilis*	Nematoda (Clade V)	JU1968	caenorhabditis.org (V2)	genome
DIMMI	*Dirofilaria immitis*	Nematoda (Clade III)	PRJEB1797	WBPS8	genome
DLATU	*Diphyllobothrium latum*	Platyhelminthes	PRJEB1206	WBPS8	genome
DMEDI	*Dracunculus medinensis*	Nematoda (Clade III)	PRJEB500	WBPS8	genome
DMELA	*Drosophila melanogaster*	Arthropoda	BDGP6	ENSEMBL-Metazoa 34	genome
DVIVI1	*Dictyocaulus viviparus*	Nematoda (Clade V)	PRJEB5116	WBPS8	genome
DVIVI2	*Dictyocaulus viviparus*	Nematoda (Clade V)	PRJNA72587	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
EELAP	*Elaeophora elaphi*	Nematoda (Clade III)	PRJEB502	WBPS8	genome
EVERM	*Enterobius vermicularis*	Nematoda (Clade III)	PRJEB503	WBPS8	genome
GPALL	*Globodera pallida*	Nematoda (Clade IV)	PRJEB123	WBPS8	genome
GPULC	*Gongylonema pulchrum*	Nematoda (Clade III)	PRJEB505	WBPS8	genome
GROST	*Globodera rostochiensis*	Nematoda (Clade IV)	PRJEB13504	WBPS8	genome
HBACT	*Heterorhabditis bacteriophora*	Nematoda (Clade V)	PRJNA13977	WBPS8	genome
HCONT1	*Haemonchus contortus*	Nematoda (Clade V)	PRJEB506	WBPS8	genome
HCONT2	*Haemonchus contortus*	Nematoda (Clade V)	PRJNA205202	WBPS8	genome
HDUJA	*Hypsibius dujardini*	Tardigrada	nHd3	tardigrades.org	genome
HMELP	*Heliconius melpomene*	Arthropoda	Hmel1	ENSEMBL-Metazoa 34	genome
HPLAC	*Haemonchus placei*	Nematoda (Clade V)	PRJEB509	WBPS8	genome
HPOLY1	*Heligmosomoides polygyrus*	Nematoda (Clade V)	PRJEB1203	WBPS8	genome
HPOLY2	*Heligmosomoides polygyrus*	Nematoda (Clade V)	PRJEB15396	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
HPOLY3	*Heligmosomoides polygyrus*	Nematoda (Clade V)	EdinPacBio	BLAXTERLAB	genome
HSAPI	*Homo sapiens*	Chordata	GRCh38	ENSEMBL87	genome
ISCAP	*Ixodes scapularis*	Arthropoda	scaW1	ENSEMBL-Metazoa 34	genome
LANAT	*Lingula anatina*	Brachiopoda	GCA_001039355	ENSEMBL-Metazoa 34	genome
LOA1	*Loa loa*	Nematoda (Clade III)	PRJNA246086	WBPS8	genome
LOA2	*Loa loa*	Nematoda (Clade III)	PRJNA60051	WBPS8	genome
LSIGM	*Litomosoides sigmodontis*	Nematoda (Clade III)	PRJEB3075	WBPS8	genome
MAREN	*Meloidogyne arenaria*	Nematoda (Clade IV)	HarA	BLAXTERLAB	genome
MBELA	*Mesorhabditis belari*	Nematoda (Clade V)	v1	BLAXTERLAB	genome
MBREV	*Monosiga brevicollis*	Choanoflagellates	GCF_000002865_3_V1	NCBI	genome
MFLOR	*Meloidogyne floridensis*	Nematoda (Clade IV)	PRJEB6016	WBPS8	genome
MHAPL	*Meloidogyne hapla*	Nematoda (Clade IV)	PRJNA29083	WBPS8	genome
MINCO	*Meloidogyne incognita*	Nematoda (Clade IV)	PRJEA28837	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
MLEID	*Mnemiopsis leidyi*	Ctenophora	GCA_000226015	ENSEMBL-Metazoa 34	genome
MLONG	*Mesorhabditis longespiculosa*	Nematoda (Clade V)	v1	BLAXTERLAB	transcriptome
NAMER	*Necator americanus*	Nematoda (Clade V)	PRJNA72135	WBPS8	genome
NBRAS	*Nippostrongylus brasiliensis*	Nematoda (Clade V)	PRJEB511	WBPS8	genome
NVECT	*Nematostella vectensis*	Cnidaria	GCA_000209225	ENSEMBL-Metazoa 34	genome
NVITR	*Nasonia vitripennis*	Arthropoda	GCA_000002325	ENSEMBL-Metazoa 34	genome
OBIMA	*Octopus bimaculoides*	Mollusca	PRJNA270931	ENSEMBL-Metazoa 34	genome
ODENT	*Oesophagostomum dentatum*	Nematoda (Clade V)	PRJNA72579	WBPS8	genome
OFLEX	*Onchocerca flexuosa*	Nematoda (Clade III)	PRJEB512	WBPS8	genome
OGUTT	*Onchocerca gutturosa*	Nematoda (Clade III)	nOg.1.1	BLAXTERLAB	genome
OOCHE1	*Onchocerca ochengi*	Nematoda (Clade III)	PRJEB1204	WBPS8	genome
OOCHE2	*Onchocerca ochengi*	Nematoda (Clade III)	PRJEB1809	WBPS8	genome
OTIPU	*Oscheius tipulae*	Nematoda (Clade V)	nOt.2.0	caenorhabditis.org(V2)	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
OVOLV	*Onchocerca volvulus*	Nematoda (Clade III)	PRJEB513	WBPS8	genome
PAES5	*Panagrolaimus es5*	Nematoda (Clade IV)	v1	BLAXTERLAB	genome
PCAUD	*Priapulus caudatus*	Priapulida	GCF_000485595_5	NCBI	genome
PEQUO	*Parascaris equorum*	Nematoda (Clade III)	PRJEB514	WBPS8	genome
PEXSP	*Pristionchus exspectatus*	Nematoda (Clade V)	PRJEB6009	WBPS8	genome
PJUL7	*Propanagrolaimus sp.* JU765	Nematoda (Clade IV)	v1	BLAXTERLAB	genome
PMURR	*Plectus murrayi*	C	nPm.2.0	BLAXTERLAB	genome
PPACI	*Pristionchus pacificus*	Nematoda (Clade V)	PRJNA12644	WBPS8	genome
PREDI	*Panagrellus redivivus*	Nematoda (Clade IV)	PRJNA186477	WBPS8	genome
PTRIC	*Parastrongyloides trichosuri*	Nematoda (Clade IV)	PRJEB515	WBPS8	genome
PVARI	*Paragordius varius*	Nematomorpha	v1.0	BLAXTERLAB	transcriptome
PVIND	*Pseudaphelenchus vindai*	Nematoda (Clade IV)	v1.0	BLAXTERLAB	transcriptome
RCULI	*Romanomermis culicivorax*	Nematoda (Clade I)	PRJEB1358	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
RKR3021	*Rhabditophanes* sp. KR 3021	Nematoda (Clade IV)	PRJEB1297	WBPS8	genome
RSB34	*Rhabditis* sp. SB347	Nematoda (Clade V)	v1	BLAXTERLAB	genome
RVARI	*Ramazzottius varieornatus*	Tardigrada	v101	tardigrades.org	genome
SBATU	*Soboliphyme baturini*	Nematoda (Clade I)	PRJEB516	WBPS8	genome
SCARP	*Steinernema carpocapsae*	Nematoda (Clade IV)	PRJNA202318	WBPS8	genome
SFELT	*Steinernema feltiae*	Nematoda (Clade IV)	PRJNA204661	WBPS8	genome
SGLAS	*Steinernema glaseri*	Nematoda (Clade IV)	PRJNA204943	WBPS8	genome
SLABI	*Setaria labiatopapillosa*	Nematoda (Clade III)	nSl.1.1	BLAXTERLAB	genome
SMANS	*Schistosoma mansoni*	Platyhelminthes	PRJEA36577	WBPS8	genome
SMARI	*Strigamia maritima*	Arthropoda	Smar1	ENSEMBL-Metazoa 34	genome
SMEDI	*Schmidtea mediterranea*	Platyhelminthes	PRJNA12585	WBPS8	genome
SMONT	*Steinernema monticolum*	Nematoda (Clade IV)	PRJNA205067	WBPS8	genome
SMURI	*Syphacia muris*	Nematoda (Clade III)	PRJEB524	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
SOBLE	*Syphacia oblevata*	Nematoda (Clade III)	v1	BLAXTERLAB	genome
SPAPI	*Strongyloides papillosus*	Nematoda (Clade IV)	PRJEB525	WBPS8	genome
SRATT	*Strongyloides ratti*	Nematoda (Clade IV)	PRJEB125	WBPS8	genome
SSCAP	*Steinernema scapterisci*	Nematoda (Clade IV)	PRJNA204942	WBPS8	genome
SSTER	*Strongyloides stercoralis*	Nematoda (Clade IV)	PRJEB528	WBPS8	genome
SVENE	*Strongyloides venezuelensis*	Nematoda (Clade IV)	PRJEB530	WBPS8	genome
SVULG	*Strongylus vulgaris*	Nematoda (Clade V)	PRJEB531	WBPS8	genome
TBRIT	*Trichinella britovi*	Nematoda (Clade I)	PRJNA257433 (ISS120)	WBPS8	genome
TCALL	*Thelazia callipaeda*	Nematoda (Clade III)	PRJEB1205	WBPS8	genome
TCANI1	*Toxocara canis*	Nematoda (Clade III)	PRJEB533	WBPS8	genome
TCANI2	*Toxocara canis*	Nematoda (Clade III)	PRJNA248777	WBPS8	genome
TCAST	*Tribolium castaneum*	Arthropoda	Tcas3	ENSEMBL-Metazoa 34	genome
TCIRC	*Teladorsagia circumcincta*	Nematoda (Clade V)	PRJNA72569	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
TMURI	*Trichuris muris*	Nematoda (Clade I)	PRJEB126	WBPS8	genome
TMURR	*Trichinella murrelli*	Nematoda (Clade I)	PRJNA257433 (ISS417)	WBPS8	genome
TNATI1	*Trichinella nativa*	Nematoda (Clade I)	PRJNA179527	WBPS8	genome
TNATI2	*Trichinella nativa*	Nematoda (Clade I)	PRJNA257433 (ISS10)	WBPS8	genome
TNELS	*Trichinella nelsoni*	Nematoda (Clade I)	PRJNA257433 (ISS37)	WBPS8	genome
TPAPU	*Trichinella papuae*	Nematoda (Clade I)	PRJNA257433 (ISS1980)	WBPS8	genome
TPATA	*Trichinella patagoniensis*	Nematoda (Clade I)	PRJNA257433 (ISS2496)	WBPS8	genome
TPSEU1	*Trichinella pseudospiralis*	Nematoda (Clade I)	PRJNA257433 (ISS13)	WBPS8	genome
TPSEU2	*Trichinella pseudospiralis*	Nematoda (Clade I)	PRJNA257433 (ISS141)	WBPS8	genome
TRIT6	*Trichinella* t6	Nematoda (Clade I)	PRJNA257433 (ISS34)	WBPS8	genome
TRIT8	*Trichinella* t8	Nematoda (Clade I)	PRJNA257433 (ISS272)	WBPS8	genome
TRIT9	*Trichinella* t9	Nematoda (Clade I)	PRJNA257433 (ISS409)	WBPS8	genome
TSPEU3	*Trichinella pseudospiralis*	Nematoda (Clade I)	PRJNA257433 (ISS176)	WBPS8	genome

Table 3.A.2 Continued from previous page

Taxon ID	Species name	Phylum	Version	Source	Basis
TSPEU4	*Trichinella pseudospiralis*	Nematoda (Clade I)	PRJNA257433 (ISS470)	WBPS8	genome
TSPEU5	*Trichinella pseudospiralis*	Nematoda (Clade I)	PRJNA257433 (ISS588)	WBPS8	genome
TSPIR1	*Trichinella spiralis*	Nematoda (Clade I)	PRJNA12603	WBPS8	genome
TSPIR2	*Trichinella spiralis*	Nematoda (Clade I)	PRJNA257433 (ISS3)	WBPS8	genome
TSUIS1	*Trichuris suis*	Nematoda (Clade I)	PRJNA179528	WBPS8	genome
TSUIS2	*Trichuris suis*	Nematoda (Clade I)	PRJNA208415	WBPS8	genome
TSUIS3	*Trichuris suis*	Nematoda (Clade I)	PRJNA208416	WBPS8	genome
TTRIC	*Trichuris trichiura*	Nematoda (Clade I)	PRJEB535	WBPS8	genome
TZIMB	*trichinella zimbabwensis*	Nematoda (Clade I)	PRJNA257433 (ISS1029)	WBPS8	genome
WBANC1	*Wuchereria bancrofti*	Nematoda (Clade III)	PRJEB536	WBPS8	genome
WBANC2	*Wuchereria bancrofti*	Nematoda (Clade III)	PRJNA275548	WBPS8	genome

Chapter 4

Effector gene families in *Globodera* species

Paul: *"Stilgar, do we have wormsign?"*
Stilgar: *"Usul, we have wormsign the likes of which even God has never seen."*

- Paul & Stilgar, *Dune (1984)*

4.1 Introduction

Within Nematoda, Clade IV *sensu* Blaxter et al., 1998 harbours taxa with a wide diversity of life styles including plant-parasites (Families Tylenchidae, Parasitaphelenchidae, Aphelenchoididae), animal-parasites (Strongyloididae), pathogens of invertebrates (*e. g.* Steinernematidae) and free-living bacteriovores (*e. g.* Panagrolaimidae) (Blaxter et al., 1998; Megen et al., 2009; Blaxter and Koutsovoulos, 2015; Bird et al., 2015). The family Tylenchidae includes the majority of economically relevant PPNs (Jones et al., 2013), such as root-knot nematodes (genus *Meloidogyne*), cyst nematodes (genera *Heterodera* and *Globodera*), and the pine wood nematode *Bursaphelenchus xylophilus*. A detailed description of the biology of *Globodera* species can be found in the introduction of Chapter 5. Several phylogenetic analyses of Clade IV nematodes based on established phylogenetic loci have been published (Subbotin et al., 2000; Holterman et al., 2009; Scholl and Bird, 2005; Megen et al., 2009) and general agreement exists on the phylogenetic relationships among the taxa. However, no comprehensive multi-locus phylogeny has been inferred based on the published Clade IV nematode genomes.

Here, I present a KinFin clustering analysis of proteomes derived from 19 Clade IV and two *Caenorhabditis* species genomes. First, I carry out a RBBH analysis to identify putative effector proteins within the proteomes of *Globodera* species, based on sequences deposited on public databases. The resulting *Globodera* proteins are used as seeds in subsequent analyses. I explore the effect of MCL inflation values on the clustering based on KinFin output files and perform analyses to test the influence of this parameter on the clustering and the representative functional annotations of clusters. Based on the results I decide on an optimal MCL inflation value. The resulting protein cluster set is used to infer a robust phylogenetic tree for Clade IV nematodes which serves as a basis for the investigation of synapomorphic

clusters and protein family expansions within the genus *Globodera*, with emphasis on effector proteins.

4.2 Methods

4.2.1 Compilation of a target effector protein list

A set of 226 plant-parasitic nematode effectors was compiled from the literature by querying the NCBI website with the string 'effector' and restricting the search taxonomically to tylenchid nematodes and their protein sequences retrieved (see Appendix 4.A.1). Using a RBBH approach, the sequences were used for identifying *G. pallida* and *G. rostochiensis* orthologues used in the clustering. I implemented the RBBH approach in the program `rbbh.py` (`https://github.com/DRL/thesis`), which takes FASTA protein files and reciprocal BLAST results as input. Reciprocal BLAST searches were carried out between the effector sequences and the PCN proteomes as well as between the PCN proteomes, using `BLASTp v2.6.0+` (Camacho et al., 2009) and following recommendations by Moreno-Hagelsieb and Latimer, 2008 regarding RBBH analysis (`-max_target_seqs 1 -max_hsps 1 -outfmt '6 std qlen slen qcovs qcovhsp' -evalue 1e-3 -seg yes -soft_masking true -use_sw_tback`). RBBHs were established considering hits with E-values $\leq 1e^{-5}$ and a query coverage $\geq 25\%$.

The three RBBH files (*G. pallida* vs. effectors, *G. rostochiensis* vs. effectors, and *G. pallida* vs. *G. rostochiensis*) were compared to the sets of published PCN effectors, containing 574 proteins from *G. pallida* (Thorpe et al., 2014) and 54 proteins of *G. rostochiensis* (Eves-van den Akker et al., 2016b). A protein in the *G. pallida* or *G. rostochiensis* genome was labeled 'effector' if a) it was present in

one of the published PCN effectors sets, b) it was a RBBH to one of the plant-parasitic nematode effectors mined from the literature, or c) its RBBH in the other PCN genome fell under a) or b). An effector was declared 'novel' if it had not been labeled as an effector by Thorpe et al., 2014 or Eves-van den Akker et al., 2016b.

4.2.2 Protein clustering

Proteomes were retrieved from WormBase ParaSite (WBPS8) for the taxa listed in Table 4.2.1, with the exception of *M. arenaria* for which genome assembly and annotation was carried out in the Blaxter Lab by Laura Salazar-Jaramillo. For each gene, only the representative isoform was kept and protein sequences below a length of 30 residues and containing more than one non-terminal stop codon were removed. Sequence similarity searches were performed using `BLAST v2.4.0+` (Camacho et al., 2009) (`-evalue 1e-5 -outfmt '6' -seg yes -soft_masking true -use_sw_tback`) on the EDDIE supercomputing cluster at the University of Edinburgh. Protein clustering was carried out using `OrthoFinder v1.1.4` (Emms and Kelly, 2015) across 19 different MCL inflation values (1.1, and 1.5 – 10.0 in increments of 0.5).

Table 4.2.1: Proteomes used in protein clustering analysis. Taxon ID: ID used in KinFin analysis. Family: Taxonomic family *sensu* Megen et al., 2009. Clade: phylogenetic clade *sensu* Blaxter et al., 1998. Proteins: Number of proteins used in the clustering analysis, after excluding non-representative isoforms. Life style: approximate life style of taxon, AP: animal-parasitic, IP: invertebrate-pathogenic, FL: free-living, PP: plant-parasitic.

Taxon ID	Species	Family	Clade	Proteins	Life style
BXYLO	*Bursaphelenchus xylophilus*	Parasitaphelenchidae	IV	17,516	PP
CBRIG	*Caenorhabditis briggsae*	Rhabditidae	V	22,305	FL
CELEG	*Caenorhabditis elegans*	Rhabditidae	V	20,219	FL
GPALL	*Globodera pallida*	Heteroderida	IV	16,332	PP
GROST	*Globodera rostochiensis*	Heteroderidae	IV	13,502	PP
MAREN	*Meloidogyne arenaria*	Meloidogynidae	IV	30,149	PP
MFLOR	*Meloidogyne floridensis*	Meloidogynidae	IV	48,179	PP
MHAPL	*Meloidogyne hapla*	Meloidogynidae	IV	14,395	PP
MINCO	*Meloidogyne incognita*	Meloidogynidae	IV	19,212	PP
PREDI	*Panagrellus redivivus*	Panagrolaimidae	IV	24,235	FL

Table 4.2.1 Continued from previous page

Taxon ID	Species	Family	Clade	Proteins	Life style
PTRIC	*Parastrongyloides trichosuri*	Strongyloididae+Alloionematidae	IV	14,915	AP
RKR3021	*Rhabditophanes* sp. KR3021	Strongyloididae+Alloionematidae	IV	13,387	FL
SCARP	*Steinernema carpocapsae*	Steinernematidae	IV	28,293	IP
SFELT	*Steinernema feltiae*	Steinernematidae	IV	33,355	IP
SGLAS	*Steinernema glaseri*	Steinernematidae	IV	34,021	IP
SMONT	*Steinernema monticolum*	Steinernematidae	IV	35,783	IP
SPAPI	*Strongyloides papillosus*	Strongyloididae+Alloionematidae	IV	18,281	AP
SRATT	*Strongyloides ratti*	Strongyloididae+Alloionematidae	IV	12,448	AP
SSTER	*Strongyloides stercoralis*	Strongyloididae+Alloionematidae	IV	12,850	AP
SSCAP	*Steinernema scapterisci*	Steinernematidae	IV	31,343	IP
SVENE	*Strongyloides venezuelensis*	Strongyloididae+Alloionematidae	IV	16,625	AP

4.2.3 KinFin analysis

For all analyses, `KinFin v1.0.3` (Laetsch, 2017b) was used with default parameters, unless specified otherwise.

An initial KinFin analysis, targeted at identifying single-copy orthologues suitable for phylogenetic analysis, was carried out for each of the 19 clusterings and revealed 4 to 28 'true' single-copy clusters present in all species depending on inflation value. The clustering based on the MCL inflation value 4.0 was chosen (see Section 4.3.2) for which 28 'true' single-copy clusters were found. The clustering at 4.0 was subsequently screened for clusters where at least two thirds of species are present with a protein count of 1, while the remaining species are absent, which resulted in 399 additional clusters. It should be noted that this is different from the concept of 'fuzzy' single-copy clusters, of which 3045 were found using default parameters ('`-n 1 -x 0.75 --min 0 --max 20`'). Phylogenetic analysis based on the 427 single-copy clusters was carried out as described in Section 3.3.2. The resulting tree topology was supplied in a second KinFin run for each of the 19 inflation values, together with `InterProScan v5.22-61.0` (Jones et al., 2014) results against `PFAM v30.0` (Finn et al., 2016) and `SignalP-Euk v4.1` (Petersen et al., 2011). In the KinFin config file, the concestor of *Caenorhabditis* spp. was defined as outgroup. Proteomes were grouped based on 'life style' and NCBI TaxIDs were supplied for each species as listed in Table 4.2.1.

Representative functional annotation (RFA) was inferred using InterProScan annotation (IPR) for all clusters in each of the 19 clusterings using `functional_annotation_of_clusters.py` with six parameter combinations concerning '`-p`' (minimum protein coverage of domain in cluster) and '`-x`' (minimum taxon coverage by proteins with domain in cluster): '`p=0.0 x=0.50`', '`p=0.0 x=0.75`', '`p=0.0 x=0.95`', '`p=0.25 x=0.0`', '`p=0.25 x=0.0`', '`p=0.50 x=0.0`',

and 'p=0.75 x=0.0'. For clustering at MCL inflation value 4.0, RFA was also inferred for SignalP-Euk (SignalP) annotations ('p=0.0 x=0.75') and for IPR annotations of synapomorphic clusters ('p=0.0 x=0.75 n=0.75', where n refers to the parameter '--node_taxon_cov' which specifies the minimum taxon coverage of taxa under a node). Cluster size distribution for clusterings at MCL inflation values 1.5, 4.0 and 9.0 were visualised using plot_cluster_size_distribution.py using the colour map 'viridis'.

4.2.4 Effect of MCL inflation parameter on clustering

The effect of MCL inflation value on the 19 clusterings and taxon occupancy for certain sets of taxa and inflation values was visualised based on the output KinFin using analysis_of_inflation_values.R (https://github.com/DRL/thesis) which uses the UpSetR library (Conway, Lex, and Gehlenborg, 2017).

4.2.5 Phylogenetic analysis of poly-γ-glutamate synthase cluster

Cluster 'OG0039632' (MCL inflation value 4.0) was identified based on the effectors list and contained one sequence of each PCN ('GPLIN_000553400' and 'GROS_g07961'). The *G. pallida* sequence displayed sequence similarity to a *Meloidogyne artiellia* poly-γ-glutamate synthase protein. Based on sequence similarity searches against NCBI and WormBase ParaSite (WBPS9), together with information from a previous study on genes coding for polyglutamate synthesis genes (Denker et al., 2008), a set of sequences was compiled. Two nematode sequences were retrieved from WBPS9 based on results of BLAST searches

of the two PCN proteins: 'nRc.2.0.1.t47789-RA' (*Romanomermis culicivorax*) and 'Dd_13536' (*Dytilenchus destructor*). From NCBI, 37 sequences were retrieved: 'CAC84452.1' (*M. artiellia*), 'XP_016947101.1' (*Drosophila biarmipes*), 'XP_020613253.1' (*Orbicella faveolata*), 'EDO44155.1' (*Nematostella vectensis*) 'XP_012557422.1' (*Hydra vulgaris*), 'ELT99805.1' (*Capitella teleta*), 'KXJ11887.1' (*Exaiptasia pallida*), 'XP_013397781.1' (*Lingula anatina*), 'EFN59253.1' (*Chlorella variabilis*), 'GAQ82684.1' (*Klebsormidium nitens*), 'CEL95394.1' (*Vitrella brassicaformis*), 'ABO99165.1' (*Ostreococcus lucimarinus*), 'KIW80809.1' (*Fonsecaea pedrosoi*), 'OAL34971.1' (*F. nubica*), 'EXJ67640.1' (*Cladophialophora psammophila*), 'KIW93775.1' (*C. bantiana*), 'CCO20520.1' (*Bathycoccus prasinos*), 'OQE31531.1' (*Penicillium steckii*), 'KOS41241.1' (*P. nordicum*), 'KIW02604.1' (*Verruconis gallopava*), 'EBA27436.1' (*Aspergillus fumigatus*), 'XP_001822130.2' (*A. oryzae*), 'OGM40173.1' (*A. bombycis*), 'AIO71030.1' (*Burkholderia oklahomensis*), 'AJX35602.1' (*B. oklahomensis*), 'WP_066571474.1' (*Burkholderia* sp. ABCPW 14), 'WP_066491221.1' (*Burkholderia* sp. BDU8), 'WP_076890427.1' (*B. pseudomallei*), 'WP_060364544.1' (*B. stagnalis*), 'WP_060362796.1' (*B. stagnalis*), 'WP_035533562.1' (*Paraburkholderia sacchari*), 'WP_043285452.1' (*P. oxyphila*), 'WP_051391089.1' (*P. mimosarum*), 'WP_090882466.1' (*Nitrosovibrio sp.*), 'NP_215088.1' (*Mycobacterium tuberculosis*), 'YP_001086643.1' (*Clostridioides difficile*), and 'NP_625239.1' (*Streptomyces coelicolor*). These sequences were used, together with the two PCN sequences, for a phylogenetic analysis as described in Section 3.3.2.

4.2.6 Phylogenetic analysis of NodL-like acetyltransferase cluster

Cluster 'OG0011331' (at MCL inflation value 4.0) was identified based on the effector list and contains sequences from *G. pallida* ('GPLIN_000026100'), *G. rostochiensis* ('GROS_g11033') *M. arenaria* ('MAREN.g14695', 'MAREN.g14696', and 'MAREN.g6955') *M. floridensis* ('augustus_masked-nMf.1.1.scaf05753-processed-gene-0.1-mRNA-1' and 'augustus_masked-nMf.1.1.scaf20924-processed-gene-0.0-mRNA-1'), and *M. hapla* ('MhA1_Contig222.frz3.gene26'). From NCBI, 42 NodL O-acetyltransferase protein sequences from bacterial taxa were retrieved: 'OJU42973.1' (Alphaproteobacteria bacterium 65-37), 'AMN39228.1' (*Rhodoplanes* sp. Z2-YC6860), 'AOO82925.1' and 'WP_083269893.1' (*Bosea vaviloviae*), 'SCW70859.1' (*Ancylobacter rudongensis*), 'SHG67911.1' (*Bradyrhizobium erythrophlei*), 'WP_065731432.1' (*B. icense*), 'WP_028163926.1' and 'WP_028350716.1' (*B. elkanii*), 'WP_065753467.1' (*B. paxllaeri*), 'WP_050402763.1' (*B. embrapense*), 'WP_057849985.1' (*B. valentinum*), 'WP_057862021.1' (*B. lablabi*), 'WP_057842264.1' (*B. retamae*), 'WP_057835876.1' (*B. jicamae*), 'WP_050628818.1' (*B. viridifuturi*), 'WP_081914325.1' (Rhizobiales bacterium YIM 77505), 'WP_020699280.1' (*Reyranella massiliensis*), 'WP_072294822.1' (*Nitrosovibrio* sp. Nv17), 'WP_002729555.1' (*Phaeospirillum molischianum*), 'WP_012561774.1' (*Oligotropha carboxidovorans*), 'WP_092745830.1' (*Acidovorax valerianellae*), 'WP_041799518.1' and 'WP_044406715.1' (*Rhodopseudomonas palustris*), 'WP_088432277.1' and 'WP_017357490.1' (*Stenotrophomonas maltophilia*), 'WP_088587077.1' (*Achromobacter marplatensis*), 'WP_045163657.1' (*Pseudomonas stutzeri*), 'WP_083237752.1' (*P. xanthomarina*), 'WP_091942489.1' (*Methylobacterium salsuginis*), 'WP_048445899.1' (*M. variabile*), 'WP_048463064.1' and 'WP_060848034.1' (*M. aquaticum*), 'WP_021246960.1' (*Sphingobium baderi*), 'WP_066721732.1' (*Sphingomonas pituitosa*), 'WP_058755772.1' (*S. endophytica*), 'WP_039515417.1' (*Xanthomonas arboricola*), 'WP_058362501.1' (*X. translucens*),

'WP_057673590.1' (*X. campestris*), 'WP_043093799.1' and 'WP_029562053.1' (*X. sacchari*), and 'WP_029218016.1' (*X. cassavae*). In addition, 12 bacterial and one eukaryotic acetyltransferase sequences, and two NodL-like acetyltransferases from *M. incognita* and *M. javanica* were retrieved from the supplementary information of Scholl et al., 2003. Alignment and phylogenetic analysis was performed as described in Section 3.3.2.

4.2.7 Analysis of lineage-specific protein family expansions

Clusters were visualised based on KinFin output using the script `effector_cluster_annotation.R` (`https://github.com/DRL/thesis`).

4.3 Results

4.3.1 PCN effectors identified through RBBH analyses

Results of RBBH analysis between PCN proteomes and effectors reported in the literature ('literature effectors') were used for creating a target list of PCN protein IDs for subsequent analysis of the protein clustering (see Table 4.3.1). In total, 230 proteins in *G. pallida* and 226 in *G. rostochiensis* were labeled as effectors through RBBH analysis. Of these, 17.8% (for *G. pallida*) and 85.8% (for *G. rostochiensis*) are labeled 'novel', since they were not included in the lists compiled by Thorpe et al., 2014 and Eves-van den Akker et al., 2016b.

Orthology to 'literature effectors' could be established for 50 proteins from *G.*

pallida and 49 proteins from *G. rostochiensis* and was restricted to proteins originating from Heteroderidae (*G. pallida*, *G. rostochiensis*, *H. glycines*, *H. schachtii*), Meloidogynidae (*M. incognita*), and Pratylenchidae (*P. goodeyi*, subfamily Pratylenchinae) *sensu* Megen et al., 2009. Unsurprisingly, for both PCN species most of the RBBHs to 'literature effectors' stem from Heteroderidae species: 94.0% for *G. pallida* and 91.8% for *G. rostochiensis*. 'Literature effectors' labeled 'novel', are listed in Table 4.3.2.

Due to the non-transitive nature of RBBH results, 16 cases were encountered in which a protein in a PCN species displayed a RBBH to a 'literature effector' but its RBBH in the other PCN species did not. One example is the HYP effector 'AIT18706.1' (see Table 4.3.2) which was the RBBH of 'GPLIN_001025300'. The PCN-RBBH of 'GPLIN_001025300' is 'GROS_g08893', which is RBBH of another HYP effector 'AIT18707.1'. This is expected as RBBH analyses suffer from high false negative rates if duplicated sequences are present in the query and subject sets, but does not pose an issue since the results are used for identification of putative effector families in the clustering. For 21 'literature effectors', RBBH results were transitive.

The four 'novel' effectors identified based on '*G. pallida* literature effectors' are three HYP effectors (previously reported in Eves-van den Akker et al., 2014), involved in biotrophy in early parasitic stages, and a nematode-specific fatty-acid- and retinol-binding (FAR) protein 'CAA70477.2', involved in evasion of host defences through interference with host-lipid signalling (Prior et al., 2001). Based on '*G. rostochiensis* literature effectors', six 'novel' effectors were found in *G. pallida* and 14 in *G. rostochiensis*. Comparisons against '*H. glycines* literature effectors' revealed, 13 and 19 'novel' effectors. One recently described 'tyrosinase-like' effector from *H. schachtii* (Habash et al., 2017), expressed in the oesophageal gland and involved in interference with plant hormone homeostasis, was found in

both PCN species. Two '*M. incognita* literature effectors' received RBBHs from both PCNs: MSP21 ('AAN08587.1'), an acid phosphatase with a signal peptide expressed in subventral gland cells (Huang et al., 2003), and a dual oxidase ('AAY84711') involved in cuticle biosynthesis (Bakhetia et al., 2005). The *M. incognita* protein Mi-MSP2 (AAQ10016.1), expressed in the subventral gland of parasitic J2 with unknown function (Huang et al., 2003), was only recovered as orthologous to a *G. rostochiensis* protein, but not to *G. pallida*. Orthologues to a calreticulin effector in *P. goodeyi* ('AIW66697.1') (Pestana, Abrantes, and Gouveia, 2015), which in *M. incognita* has been proposed to modulate plant defences (Jaouannet et al., 2013), was also found in both PCNs.

Table 4.3.1: Proteins labeled as 'effectors' by RBBH analysis. Species: species name associated with 'literature effector'. Count: number of sequences identified by RBBH. PCN RBBH: number of sequences identified by RBBH which also have a RBBH to the other PCN species. Novel: number of sequences identified by RBBH which were not labeled as 'effectors' by Thorpe et al., 2014 or Eves-van den Akker et al., 2016b.

	Globodera pallida			*Globodera rostochiensis*		
Species	Count	PCN RBBH	Novel	Count	PCN RBBH	Novel
G. pallida	4	4	3	2	2	2
G. rostochiensis	15	11	6	19	12	14
H. glycines	27	16	13	23	19	19
H. schachtii	1	0	1	1	0	1
M. incognita	2	2	2	3	2	3
P. goodeyi	1	1	1	1	0	1
Literature (Σ)	50	34	26	49	35	40
PCN RBBH (Σ)	180	180	15	177	177	154
All (Σ)	230	214	41	226	212	194

Table 4.3.2: Proteins labeled as 'novel' effectors by RBBH analysis. Accession: Accession number of 'literature effector'. EGLS: oesophageal gland-localized secretory protein

Accession	species	description	PCN protein	PCN species	PCN RBBH
AAK55116.1	*H. glycines*	VAP	GROS_g08878	*G. rostochiensis*	GPLIN_000167500
			GPLIN_001250900	*G. pallida*	None
AAL78229.1	*H. glycines*	Hgg18	GROS_g01822	*G. rostochiensis*	None
AAM18623.1	*H. glycines*	C-type lectin	GROS_g00665	*G. rostochiensis*	GPLIN_001121000
			GPLIN_001121000	*G. pallida*	GROS_g00665
AAM95699.2	*H. glycines*	33A09	GROS_g03474	*G. rostochiensis*	GPLIN_001227700
AAN08587.1	*M. incognita*	msp21	GROS_g09671	*G. rostochiensis*	GPLIN_000096000
			GPLIN_000096000	*G. pallida*	GROS_g09671
AAN14978.1	*H. glycines*	chitinase	GROS_g06362	*G. rostochiensis*	GPLIN_000707700
			GPLIN_000562100	*G. pallida*	None
AAN32888.1	*H. glycines*	annexin 4C10	GROS_g01784	*G. rostochiensis*	GPLIN_000241100
			GPLIN_000241100	*G. pallida*	GROS_g01784

Table 4.3.2 Continued from previous page

Accession	species	description	PCN protein	PCN species	PCN RBBH
AAN32889.1	*H. glycines*	ubiquitin-extension	GROS_g08879	*G. rostochiensis*	GPLIN_000167700
AAO33475.1	*H. glycines*	G5D08	GPLIN_000847600	*G. pallida*	None
AAO85457.1	*H. glycines*	G19B10	GROS_g03338	*G. rostochiensis*	GPLIN_001278400
AAO85458.2	*H. glycines*	G19C07	GROS_g00017	*G. rostochiensis*	GPLIN_000780600
AAO85459.1	*H. glycines*	G20E03	GROS_g05682	*G. rostochiensis*	GPLIN_000962200
AAP30756.1	*H. glycines*	10C02	GROS_g02358	*G. rostochiensis*	GPLIN_001203000
AAP30757.1	*H. glycines*	30G12;30G15	GROS_g06322	*G. rostochiensis*	GPLIN_000668600
AAP30762.1	*H. glycines*	G7E05	GROS_g10874	*G. rostochiensis*	None
AAQ10016.1	*M. incognita*	msp2	GROS_g02633	*G. rostochiensis*	None
AAX68678.1	*H. glycines*	aminopeptidase	GROS_g06227	*G. rostochiensis*	None
			GPLIN_000332300	*G. pallida*	GROS_g08036
AAY84711.2	*M. incognita*	dual oxidase	GROS_g06412	*G. rostochiensis*	GPLIN_000198600
			GPLIN_000198600	*G. pallida*	GROS_g06412

Table 4.3.2 Continued from previous page

Accession	species	description	PCN protein	PCN species	PCN RBBH
ACY70448.1	*G. rostochiensis*	CLE	GROS_g08166	*G. rostochiensis*	None
ACY70450.1	*G. rostochiensis*	CLE	GROS_g10924	*G. rostochiensis*	GPLIN_000697600
ADF28634.1	*H. glycines*	chaperonin-like	GPLIN_001517100	*G. pallida*	None
AFH68236.1	*G. rostochiensis*	1106	GROS_g14309	*G. rostochiensis*	GPLIN_000235400
AFN86180.1	*G. rostochiensis*	SPRYSEC-19	GROS_g14220	*G. rostochiensis*	None
AHW98763.1	*G. rostochiensis*	VAP	GPLIN_001139400	*G. pallida*	None
			GROS_g06952	*G. rostochiensis*	GPLIN_001449500
AHW98766.1	*G. rostochiensis*	metalloprotease	GPLIN_000521800	*G. pallida*	GROS_g03550
			GROS_g03550	*G. rostochiensis*	GPLIN_000521800
AHW98768.1	*G. rostochiensis*	amphid protein	GPLIN_000496800	*G. pallida*	GROS_g10653
			GROS_g06775	*G. rostochiensis*	None
AHW98769.1	*G. rostochiensis*	glutathione peroxidase	GPLIN_001153000	*G. pallida*	GROS_g01811
			GROS_g04155	*G. rostochiensis*	None

Table 4.3.2 Continued from previous page

Accession	species	description	PCN protein	PCN species	PCN RBBH
AHW98770.1	*G. rostochiensis*	SKIP1	GPLIN_001036200	*G. pallida*	GROS_g03245
			GROS_g01174	*G. rostochiensis*	None
AHW98771.1	*G. rostochiensis*	peroxiredoxin	GPLIN_000672400	*G. pallida*	None
			GROS_g01193	*G. rostochiensis*	None
AHW98772.1	*G. rostochiensis*	D406;4D08	GROS_g08634	*G. rostochiensis*	GPLIN_000203300
AHZ59334.1	*G. rostochiensis*	SPRY	GROS_g14234	*G. rostochiensis*	GPLIN_000776700
AIT18683.1	*G. pallida*	HYP	GPLIN_001208400	*G. pallida*	GROS_g12827
AIT18706.1	*G. pallida*	HYP	GPLIN_001025300	*G. pallida*	GROS_g08893
AIT18707.1	*G. pallida*	HYP	GROS_g08893	*G. rostochiensis*	GPLIN_001025300
AIW66697.1	*P. goodeyi*	Calreticulin	GROS_g02225	*G. rostochiensis*	None
			GPLIN_000040200	*G. pallida*	GROS_g02222
AJR19769.1	*H. glycines*	EGLS 1	GPLIN_001324100	*G. pallida*	GROS_g06371
			GROS_g09018	*G. rostochiensis*	None

Table 4.3.2 Continued from previous page

Accession	species	description	PCN protein	PCN species	PCN RBBH
AJR19771.1	*H. glycines*	G12H04;EGLS 3	GROS_g07677	*G. rostochiensis*	GPLIN_000996800
AJR19778.1	*H. glycines*	EGLS 10	GPLIN_001159200	*G. pallida*	None
AJR19779.1	*H. glycines*	EGLS 11	GPLIN_000714100	*G. pallida*	GROS_g02394
			GROS_g02394	*G. rostochiensis*	GPLIN_000714100
AJR19780.1	*H. glycines*	EGLS 12	GPLIN_000378400	*G. pallida*	GROS_g00601
			GROS_g00601	*G. rostochiensis*	GPLIN_000378400
AJR19781.1	*H. glycines*	EGLS 13;CWDE	GPLIN_000949800	*G. pallida*	GROS_g11374
AJR19783.1	*H. glycines*	novel;EGLS 15	GROS_g02470	*G. rostochiensis*	GPLIN_000763000
AJR19785.1	*H. glycines*	EGLS 17	GPLIN_000661200	*G. pallida*	None
ANB41563.1	*H. schachtii*	tyrosinase-like	GPLIN_000659000	*G. pallida*	None
			GROS_g10854	*G. rostochiensis*	None
CAA70477.2	*G. pallida*	FAR	GPLIN_001416800	*G. pallida*	GROS_g14202
			GROS_g14202	*G. rostochiensis*	GPLIN_001416800

Table 4.3.2 Continued from previous page

Accession	species	description	PCN protein	PCN species	PCN RBBH
CAC21847.1	*G. rostochiensis*	dgl-1; A4	GROS_g04623	*G. rostochiensis*	GPLIN_001317500
CAD60977.1	*G. rostochiensis*	E9 protein	GROS_g14123	*G. rostochiensis*	GPLIN_000187800

4.3.2 Assessment of effect of MCL inflation value on clustering

For the 19 clusterings at different MCL inflation values, the number of resulting clusters increased asymptotically from 115,990 (MCL inflation value 1.5) to 169,716 (10.0) with increasing MCL parameter. The proportion of singleton clusters decreased from 86.49% to 62.83% in the same direction, but absolute numbers of singletons increased from 100,315 to 106,627. This is as expected since increasing MCL inflation controls the granularity of the clustering, reducing the number of resulting clusters and thereby reducing the proportion of singletons. The proportion of clusters receiving RFA varied depending on MCL parameter and RFA parameters. The strictest RFA parameter tested was `--domain_protein_cov 0.75 --domain_taxon_cov 0.0` ('p=0.75 x=0.0', requiring that 75% of proteins in a cluster share an IPR ID), which resulted in representative functional annotation of 9.81% (1.1) to 21.57% (10.0) of clusters. The most lenient RFA parameter was `--domain_protein_cov 0.0 --domain_protein_cov 0.5` ('p=0.0 x=0.5', requiring that 50% of taxa present in a cluster have at least one protein sharing an IPR ID), yielded RFA percentages between 11.58% (1.1) and 26.25% (10.0). To further investigate the contribution of MCL and RFA parameters on the 'RFA landscape', clusters which received an RFA were visualised by the number of contributing taxa across MCL and RFA parameters (see Figure 4.3.1). Differences between RFA parameters are minor, while the MCL inflation value has a larger effect. In all panels, 1.1 appears as a clear outlier which is most likely due to the fact that its inclusivity leads to high heterogeneity of functional annotations within clusters which prevents RFA. Inflation values above 1.1 yield more consistent numbers across a wide range of taxon counts, with higher inflation values generating higher numbers of RFAs. This trend is however reversed for clusters containing 19 or more taxa, where lower inflation values generate a higher count of RFAs. This effect of higher granularity

on deeply conserved clusters across the taxa is most likely caused by several factors, since an analysis of GO terms associated with IPR IDs did not show a clear pattern linked to losses of RFAs at higher inflation values.

For all further analysis, representative functional annotations of clusters based on the RFA parameters 'p=0.0 x=0.75' are used as it results in an intermediate distribution of percentage of clusters with RFAs, ranging from 11.00% (1.1) to 23.48% (10.0).

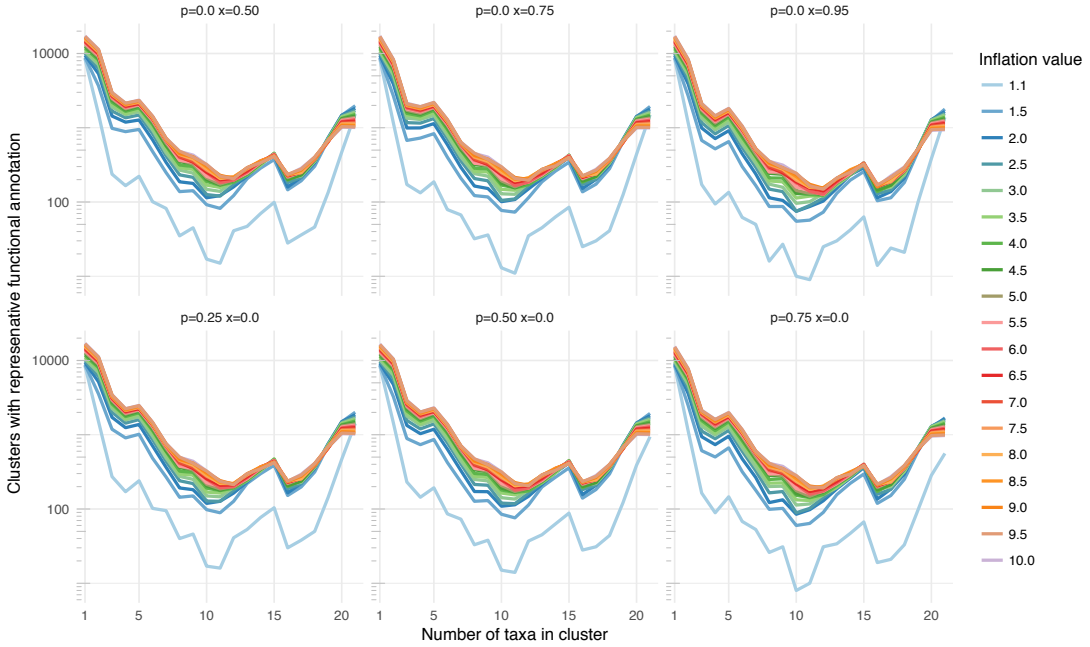


Figure 4.3.1: Visualisation of RFA of clusterings at different MCL inflation values. Each of the six parameter combinations of the RFA is visualised as a panel, titled with the parameters. Y-axes display the count of clusters which contain at least IPR based RFA in the given analysis. X-axes show the count of taxa present in the cluster. Lines are drawn for each clustering based on MCL inflation value. p: minimum proportion of proteins in a cluster sharing an IPR annotation. x: minimum proportion of taxa in a cluster having at least one protein sharing an IPR annotation.

Systematic RFA peaks at taxon counts of five and 15 were observed across all MCL and RFA parameters. Taxon composition of clusters with RFAs was therefore visualised for clusters containing five and 15 taxa at the inflation values 1.5, 4.0, and 9.0 (see Figure 4.3.2). For clusters containing five taxa the three

most frequent sets of taxa are identical between all three inflation values and account for 76.08, 76.49, and 71.82% of clusters of size five, respectively. The observed RFA peak is therefore due to closely related taxa (*Steinernema* spp., Strongyloididae and subsets of Tylenchida) and ultimately an effect of sampling. Clusters containing 15 taxa, are composed of two main configurations: all taxa except Heteroderidae and all taxa except Strongyloididae+Alloionematidae, which again is due to phylogenetic distance and sampling. The two taxa most frequently absent from clusters containing 20 or 19 taxa are *M. incognita* and *G. pallida*, suggesting that the two proteomes are lacking core proteins present in all other nematodes in this analysis.

Another view of the effect of MCL inflation values can be achieved by visualising the entire distribution of cluster sizes across clusterings. Distributions were generated for MCL inflation values 1.5, 4.0, and 9.0 and are shown in Figure 4.3.3. Deviations from the expected power-law can be observed at taxon counts of five, 15 and 21, similar to those seen for cluster RFA. While higher MCL inflation values generally seem to perform better concerning number of RFA clusters, a high granularity of clusters risks partition of genuine protein families.

I chose the MCL inflation value of 4.0 for all subsequent analysis of this dataset. This is based on two reasons: this value yielded the highest count of 'true' 1-to-1 clusters (where every taxon is present and has exactly one protein) and displayed an intermediate distribution of counts of RFA clusters along the spectrum of taxon counts. Concerning 'fuzzy' 1-to-1 clusters, this inflation value displayed the second highest count (3045), only surpassed by the value of 3.5 (3045).

4.3.3 Analysis of protein clustering

The `OrthoFinder` clustering at MCL inflation value of 4.0 placed the 477,345 proteins in 153,457 clusters of which 66.23% were singletons (accounting for 21.29% of proteins). Phylogenetic analysis based on 427 single-copy clusters yielded a robust phylogenetic tree (Figure 4.3.4).

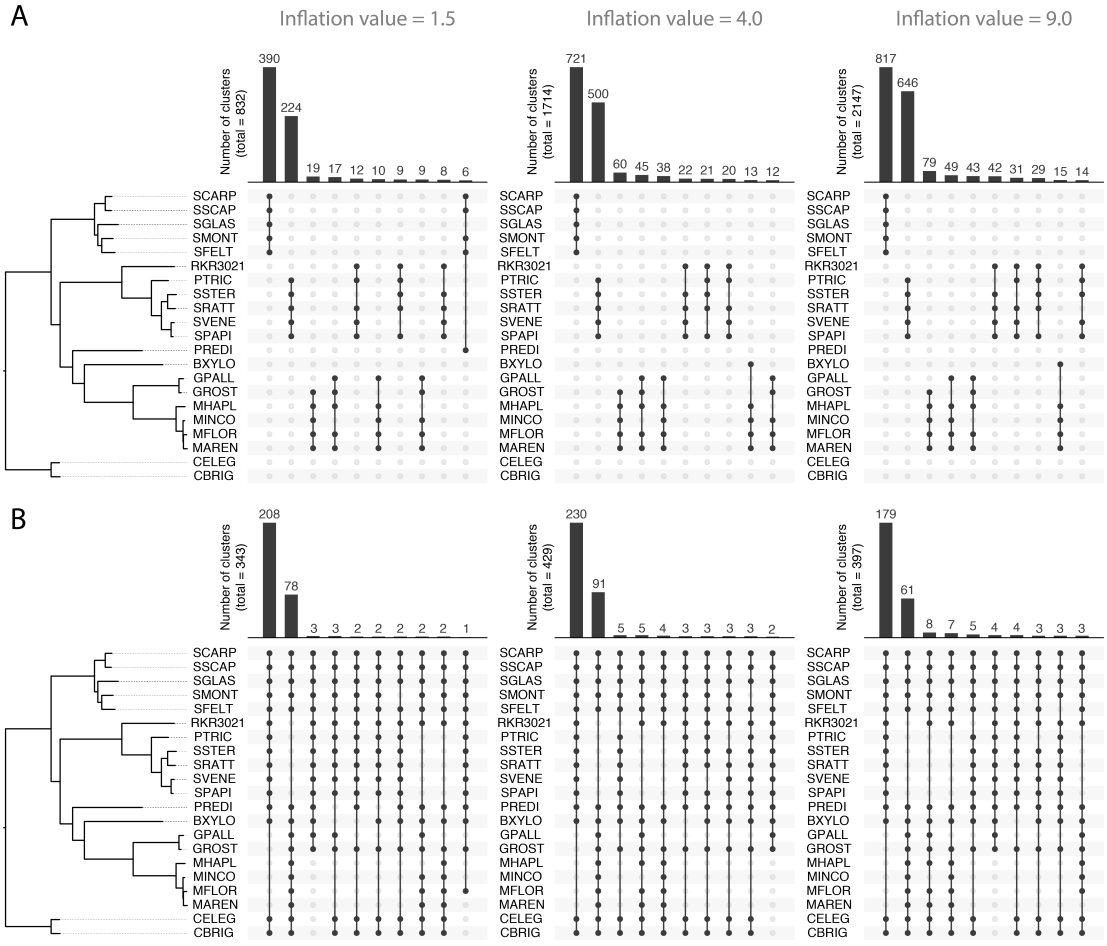


Figure 4.3.2: Analysis of taxon composition of clusters. **A**: clusters with taxon count of five. **B**: clusters with taxon count of 15. Panels with are drawn for each inflation value (columns). Taxon membership is visualised for the ten most frequent sets, with bar plots indicating count of clusters for a given set. Phylogenetic relationships for taxa are indicated using the phylogenetic tree discussed in Section 4.3.3.

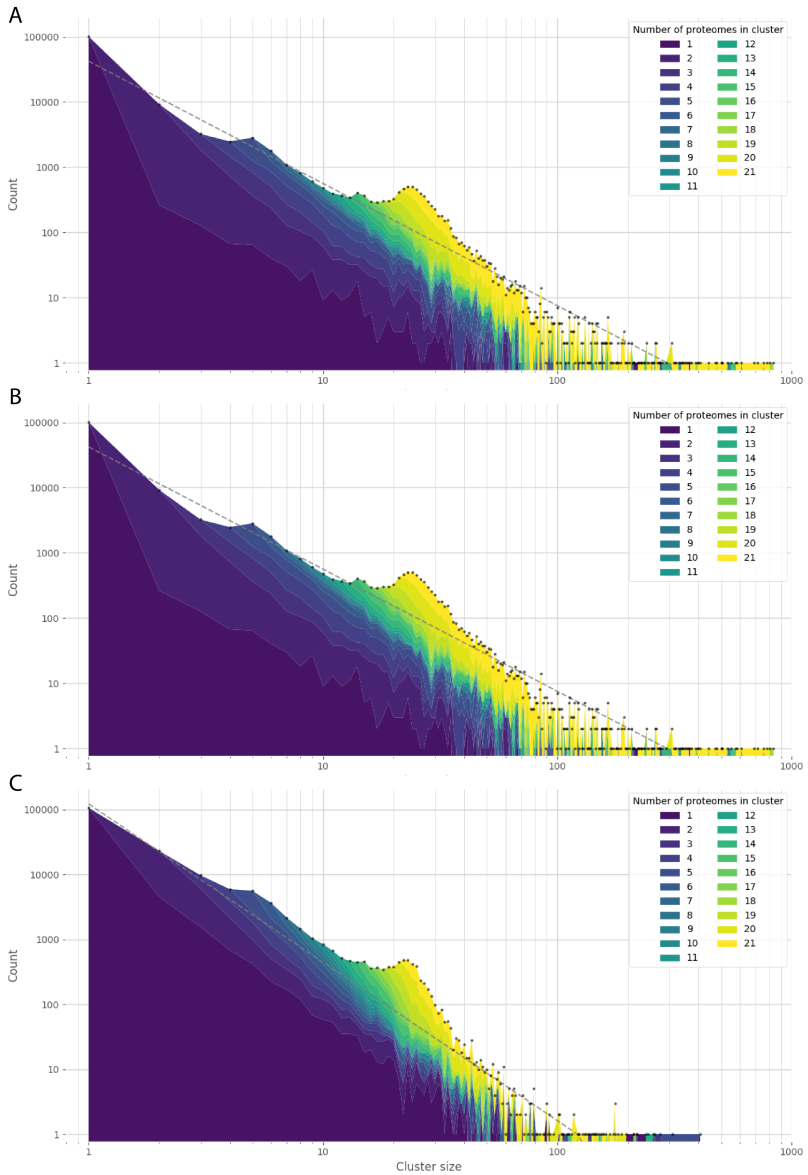


Figure 4.3.3: Cluster size distribution for three MCL inflation values. The distribution of cluster sizes is coloured based on the number of proteomes present in each cluster. Total values of counts of each cluster size are indicated with grey dots. A fitted power-law curve (grey) is drawn for reference. **A**: Clustering based on MCL inflation value of 1.5, **B**: Clustering based on MCL inflation value of 4.0. C: Clustering based on MCL inflation value of 9.0.

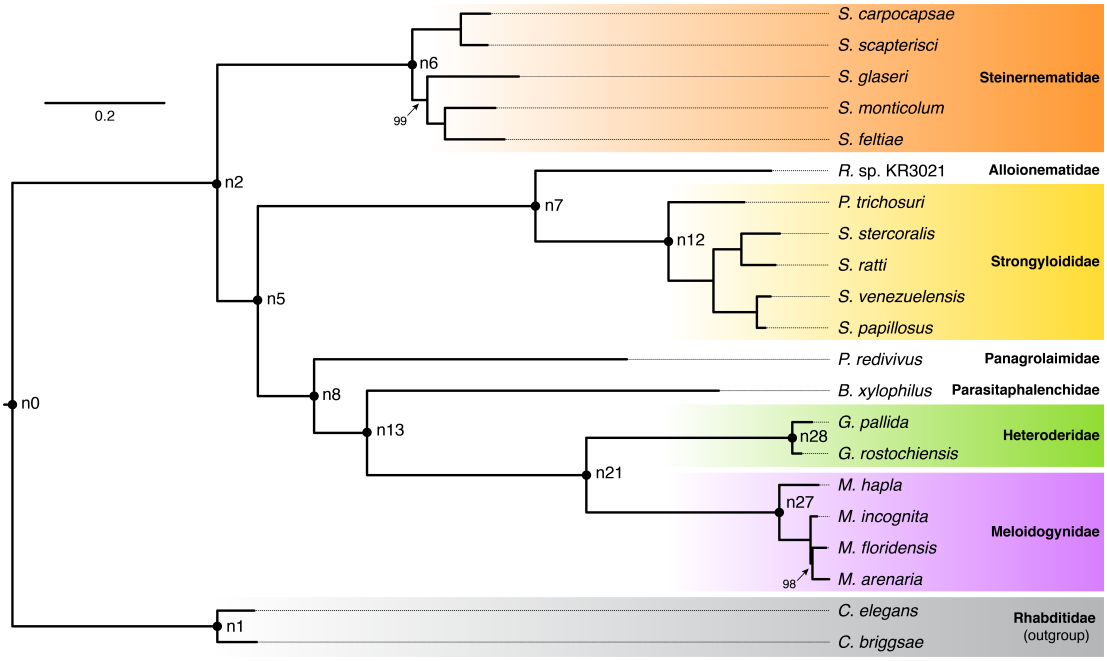


Figure 4.3.4: Phylogenetic tree of Clade IV nematodes. Tree rooted based on concestor of *Caenorhabditis* species. All branches have non-parametric bootstrap support of 100, except where indicated. Taxonomic families *sensu* Megen et al., 2009 are listed in bold. Nodes of interest have been labeled with names.

Traditional taxonomy has grouped Steinernematidae, Alloionematidae, Panagrolaimidae, and Strongyloididae in the order Rhabditida, while the plant parasitic taxa (Heteroderidae, Meloidogynidae and Parasitaphelenchidae) were unified under the separate order Tylenchida (Andrássy, 1976). In the light of molecular data, De Ley and Blaxter, 2002 revisited the taxonomy of nematodes and grouped members of these taxa within the suborder Tylenchina. Subsequent studies have confirmed this grouping (Bert et al., 2008; Megen et al., 2009; Holterman et al., 2009), with only one study (Nadler, Bolotin, and Stock, 2006) (based on 28S rDNA and the mitochondrial markers 12S rDNA and COX1) rejecting the inclusion of Steinernematidae in Tylenchina. To my knowledge, the present phylogenetic analysis presents the most comprehensive analysis to date, composed of 427 loci, totalling

107,842 amino acid sites (proportion of gaps in alignment: 25.03%). Steinernematidae is recovered as sister to all other Tylenchina, analogous to previous analyses based on 18S rDNA. Node 'n5' separates two main clades: the clade under node 'n7', composed of Strongyloididae and *Rhabditophanes* sp. KR3021, and the clade under node 'n8' comprised by the free living panagrolaimid and the plant parasitic families. Due to the limited number of taxa included in the analysis, conclusions regarding systematics of Clade IV taxa cannot be drawn with confidence. However, this tree is a robust basis for further analysis. It should be noted that, although in this analysis Heteroderidae and Meloidogynidae are sister clades, plant-parasitic endoparasitism in these groups has evolved independently (Bert et al., 2008).

4.3.4 Synapomorphic clusters

Synapomorphic clusters at 12 nodes of interest and their RFAs were investigated and counts are listed in Table 4.3.3. A very low count of synapomorphic clusters is recovered at node 'n5' (ancestor of non-*Steinernema* Clade IV nematodes).

The 456 PCN effector proteins were placed in 259 clusters. Of those, 181 were found in synapomorphic clusters of which 85 received an RFA based on IPR IDs. Among those, 18 'complete presence' clusters are synapomorphic to the ancestral node of all taxa of which six appear to be secreted based on SignalP annotations. These include a 'Nematode fatty acid retinoid binding' (IPR008632) cluster, containing the PCN orthologues to Gp-SEC-2 (Gp-FAR-1, 'CAA70477.2'), a Hsp90 cluster (IPR001404) containing the Ce-ENPL-1, a calreticulin/calnexin (IPR001580) cluster harbouring PCN orthologues to the calreticulin 'literature effector' from *P. goodeyi* (Li et al., 2015) and Ce-CRT-1, which is required for normal sperm and oocyte development (Park et al., 2001). In addition, three 'protein disulfide isomerase' clusters were found. These contain the *C. elegans* homologues

Table 4.3.3: Synapomorphic clusters. Counts, RFA (in %) and number of clusters containing at least one effector protein from PCN identified through RBBH analysis for 'complete presence' (100% of taxa under node are present) and 'partial absence' (at least 75% and less than 100% of taxa under node are present) synapomorphic clusters for each of the 12 nodes of interest. Nodes in bold are ancestors of only PPN. 'N/A' indicates cases in which nodes are ancestors of less than four taxa or when percentage of functional annotation could not be calculated due to lack of clusters.

	Complete presence			Partial absence		
Node	Count	RFA (%)	Effectors	Count	RFA (%)	Effectors
n0	1630	90.7	18	3170	84.3	20
n1	4570	40.0	0	N/A	N/A	N/A
n2	29	72.4	2	110	69.1	1
n5	1	100.0	0	18	50.0	0
n6	1964	36.7	0	1121	36.3	0
n7	1116	41.4	0	168	42.3	0
n8	48	56.3	0	120	38.3	1
n12	1234	40.5	0	300	34.3	0
n13	41	43.9	1	51	37.3	0
n21	217	41.5	2	375	44.8	7
n27	1228	28.6	0	1435	14.3	0
n28	2068	22.4	129	N/A	N/A	N/A

Ce-PDI-2 (cluster 'OG0000954'), Ce-PDI-3 ('OG0002675'), Ce-PDI-6 and Ce-PDI-6A ('OG0002540') which are part of the prolyl 4-hydroxylase beta-subunit, essential for collagen biogenesis during larval development (Winter, McCormack, and Page, 2007; Eletto et al., 2014).

Synapomorphic cluster at nodes 'n13', 'n21' and 'n28' were screened for PCN effectors based on the effector sequences previously identified (Table 4.3.1). However, all synapomorphic clusters at these nodes are of potential interest in relation to plant parasitism, since they contain only plant parasitic taxa. At node 'n13', only a single synapomorphic cluster ('OG0000615') was identified which contained a PCN effector. The cluster received no IPR RFA, but the *G. rostochiensis* orthologue was identified by Eves-van den Akker et al., 2016b as an expansin-like cell-wall degrading enzyme, acquired via HGT. Hence, this suggests that acquisition might have occurred prior to the split of Parasitaphelenchidae and Heteroderida/Meloidogynidae. Synapomorphies at this node also include a 'Glycosyl transferase family 31' cluster ('OG0005813') and a chemosensory 'Serpentine type 7TM GPCR chemoreceptor Srsx' cluster ('OG0000130') which are not part of the known effectorome of PCN. PCN proteins in both clusters carry 'SignalP-noTM' annotation, suggestive of secretion.

Node 'n21' harbours two 'complete presence' ('OG0011347' and 'OG0009436') and seven 'partial absence' synapomorphic clusters containing PCN effectors. Cluster 'OG0011347' contains PCN orthologues to a literature effector from *H. glycines*, oesophageal gland-localized secretory protein 12 ('AJR19780.1'). The cluster is composed of one protein from each species with the exception of *M. floridensis* which has three paralogues. However, two of the underlying gene predictions might be a result of the fragmented assembly since they are located at the ends of scaffolds. Cluster 'OG0009436' contains a *G. rostochiensis* protein which was identified as an effector through an elevated expression level in the 14dpi stage by Eves-van den Akker et al., 2016b. The seven 'partial absence' clusters include two composed of orthologues of literature effectors: cluster 'OG0014476' contains PCN proteins recovered as RBBH to Hg-10C02, a secreted protein of unknown function expressed in subventral gland cells in *H. glycines* (Gao et al., 2003), and to Mi-MSP2 (Huang et al., 2003). The latter cluster contains no orthologue of *G.*

pallida, analogous to the results of RBBH analysis. The five remaining clusters all contain proteins from *G. rostochiensis* acquired by HGT as described in Eves-van den Akker et al., 2016b. These include three GH53 arabinogalactan endo-1,4-beta-galactosidase clusters, one L-threonine aldolase cluster of unknown function, and one NodL-like acetyltransferase cluster ('OG0011331') containing *M. hapla* and *G. rostochiensis* proteins, reported to have been acquired through HGT from rhizobial bacteria and involved in feeding site induction (Paganini et al., 2012; Eves-van den Akker et al., 2016b). The latter is an interesting case since structure of feeding sites differ substantially between Heteroderidae and Meloidogynidae. Nevertheless both PCN and three RKN (*M. arenaria*, *M. floridensis*, and *M. hapla*) are members of this cluster. Although rhizobial taxa are common contaminants of sequencing datasets (Laurence, Hatzis, and Brash, 2014), both PCN proteins originate from contigs flanked by sequences of clear metazoan origin. In rhizobial bacteria this protein is involved in the biosynthesis of Nod factors, a family of signalling molecules which trigger root-hair deformation, as one of the early steps of nodule formation during the legume-*Rhizobium* symbiosis (Göttfert, 1993). Comparative genomic analysis of rhizobial species revealed that *nod* genes, in addition to other loci involved in establishment of symbiosis, tend to be organised on plasmids or islands within the genome (González et al., 2003). Low conservation of synteny between species in those regions suggests that they are shaped by rearrangements and horizontal transfer, lending plausibility to the hypothesis of rhizobial bacteria acting as donors to nematode taxa. Scholl et al., 2003 described two RKN NodL-like acetyltransferase, derived from EST data from *M. incognita* and *M. javanica*. Phylogenetic analysis suggested acquisition from rhizobial donors after the separation of Meloidogynidae and Heteroderidae, since PCR amplification failed to recover this gene in cyst nematodes. In order to verify monophyly of the putative NodL-like acetyltransferase from PCN and RKN found in cluster 'OG0011331', a phylogenetic tree was inferred based on NodL (O-acetyltransferases) sequences retrieved from NCBI and sequences analysed in Scholl et al., 2003 (Figure 4.3.5).

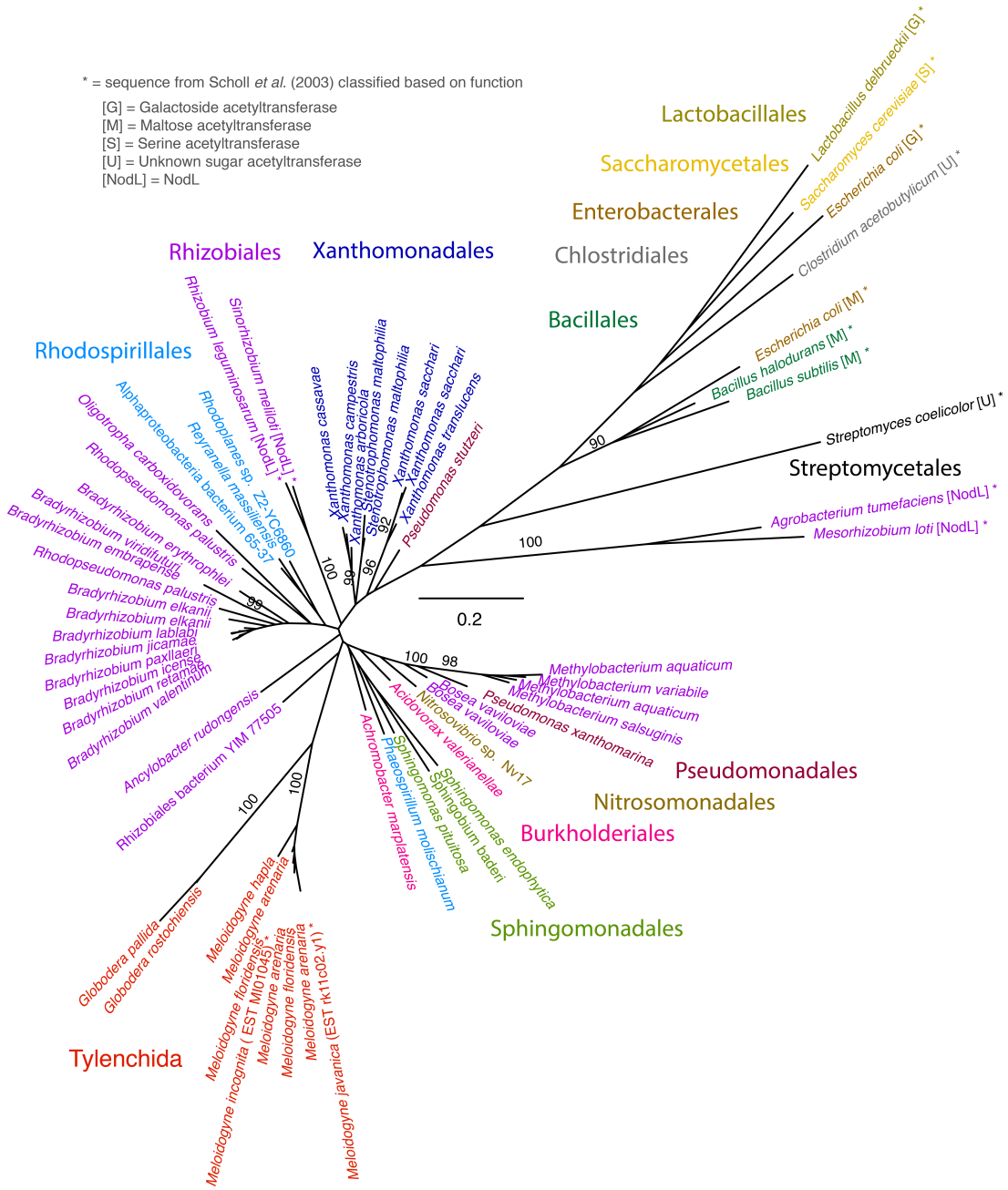


Figure 4.3.5: Phylogenetic tree of NodL-like acetyltransferase proteins. Non-parametric bootstrap support is only indicated for branches with support above 90. Species are coloured by taxonomic family (based on NCBI taxonomy). Sequences from Scholl et al., 2003 are indicated with an asterisk and their functional classification is indicated in brackets. All sequences retrieved from NCBI are O-acetyltransferases (*i. e.* NodL). Nematode (Tylenchida) sequences form a monophyletic clade.

Genes coding for NodL acetyltransferases are found sporadically across a wide range of bacterial families, which is compatible with the hypothesis of increased HGT of this locus. In the tree, these are clearly separated from other acetyltransferase sequences (from Lactobacillales, Saccharomycetales, Enterobacterales, Clostridiales, Bacillales, and Streptomycetales) by long branch lengths. Nematode NodL-like proteins form a monophyletic clade divided into PCN and RKN sequences. It should be noted that the *G. pallida* protein, one of the three *M. arenaria* proteins, and the two *M. floridensis* proteins are derived from fragmented gene models, which will have contributed to their long branch lengths. The NodL sequence most closely related to nematodes was predicted from the rhizobial taxon 'Rhizobiales bacterium YIM 77505', which was sampled within the scope of an environmental study of thermophilic bacteria of the Tengchong hot spring sediment from the Yunnan province in China (NCBI BioSample ID 'SAMN02745209') and sequenced using the PacBio RS platform. Based on the nature of this sample and the long branch lengths — posibly caused by the high error rate of the PacBio technology — no clear hypothesis regarding the HGT donor taxon can be formulated. Failure of previous studies to detect genes coding for NodL-like acetyltransferases in cyst nematode genomes might be explained by the low homology at the nucleotide level between the *Meloidogyne* and *Globodera* sequences, which could be a result of assimilation of the bacterial sequence to the background genome after acquisition through HGT. Furthermore, no clear role has been assigned yet to NodL-like acetyltransferases in PPN. While previous studies have suggested involvement in feeding site induction (Paganini et al., 2012; Eves-van den Akker et al., 2016b), this has not been proven. However, secretions of *Meloidogyne incognita*, termed NemF (RKN factor), have been shown to elicit root-hair deformations in wild-type *Lotus japonicus* similar to those caused by Nod factors (Weerasinghe, Bird, and Allen, 2005). This response was altered or absent in *L. japonicus* mutants at genes involved in Nod factor reception, suggesting that NemF interacts with the same receptor pathway. This effect was also observed in tomato which implies that the response to NemF is not a specific

feature of legumes. NodL-like acetyltransferases in RKN might be involved in the biosynthesis of NemF, analogous to NodL orthologues in rhizobial Nod factor synthesis.

Unsurprisingly, node 'n28' (ancestor of PCN) displayed the highest number of synapomorphic clusters containing effectors. The PCN proteins contained within these clusters account for 270 (59.21%) of PCN proteins labeled as effectors through RBBH analysis. Of these, 19 displayed SignalP annotation congruent with secretion. This set of 270 proteins were collated into a list ('PCN-synapomorphic effectors') for subsequent analysis in Chapter 5. One of the clusters synapomorphic to PCN ('OG0039632', annotated as 'Capsule synthesis protein, CapA') contained a *G. pallida* protein labeled as effector by Thorpe et al., 2014 due to sequence similarity to a protein to a *M. artiellia* sequence (Veronico et al., 2001), another putative HGT from bacteria to nematode. However, neither PCN sequence was recovered as a RBBH to this literature effector ('CAC84452.1'). The result of phylogenetic analysis of a selection of bacterial and eukaryotic poly-γ-glutamate synthase (PGA) proteins is shown in Figure 4.3.6. Nematode proteins form a clade in which PCN and RKN sequences are clearly separated by long branches. Poly-γ-glutamate synthase (PGA) is usually found in bacteria and cnidarian nematocytes, where it was acquired through HGT from a firmicute donor and is critical to nematocyte discharge. A phylogenetic study by Denker et al., 2008 suggests that PGA genes where acquired independently through HGT in multiple eukaryotic lineages (including Nematoda, Arthropoda, Viridiplantae, Fungi, Annelida and Choanoflagellates), in addition to be acquired once at the root of metazoans where it was subsequently lost in all lineages except Porifera and Cnidaria. While an in depth analysis of these claims would go beyond this section, the sequences originating from nematode taxa, for which a genome is available, were investigated. Nematode sequences are positioned in relative vicinity in the tree in Figure 4.3.6, but fail to form a clade. The *D. destructor* sequence originates from a scaffold ('scaffold443') which only contains

one other gene ('Dd_13535') which has no domains annotated with the exception of a transmembrane domain and no hits against NCBI nt. The *R. culicivorax* sequence resides alone on scaffold 'nRc.2.0.scaf11901'. Both of these cases suggest that these are more likely to be artefacts of the sequencing projects rather than genuine HGTs. *Burkholderia* is the most commonly encountered contaminant in sequencing datasets (Laurence, Hatzis, and Brash, 2014). The picture is different for the PCN species: the *G. rostochiensis* sequence originates from a scaffold containing 29 genes ('GROS_00231') and the *G. pallida* sequence is located on a scaffold ('pathogens_Gpal_scaffold_148') together with 47 other genes. The neighbouring genes display no sign of bacterial origin (both PCN genes have a Fibronectin type III gene on one side). Hence, it appears that *Globodera* species acquired a PGA gene via HGT from Proteobacteria, as did *M. artiellia*. However, no other proteome in the clustering contained a protein annotated as 'Capsule synthesis protein, CapA' (IPR019079), suggesting that this acquisition took place either independently in *Globodera* spp. and *M. artiellia* or occurred in the common ancestor of both groups and was subsequently lost in all four *Meloidogyne* species in this analysis.

4.3.5 Protein family expansions

KinFin facilitates assessment of protein family expansions through pairwise representation tests of mean counts of proteins in 'shared' clusters between groupings and the background (all taxa not in a grouping). Results for these tests are visualised as volcano plots in Figure 4.3.7. Differences in protein counts in 'shared' clusters for groups of interest, results of pairwise protein count representation tests of plant-parasitic nematodes vs. the background, and the genus *Globodera* vs. the other genera are depicted in Figure 4.3.7. One example for a consistent family expansion in PPNs (Figure 4.3.7A) is cluster 'OG0000346', a Serine-threonine/tyrosine-protein kinase, present in all taxa with a single copy, except in PPNs where counts range

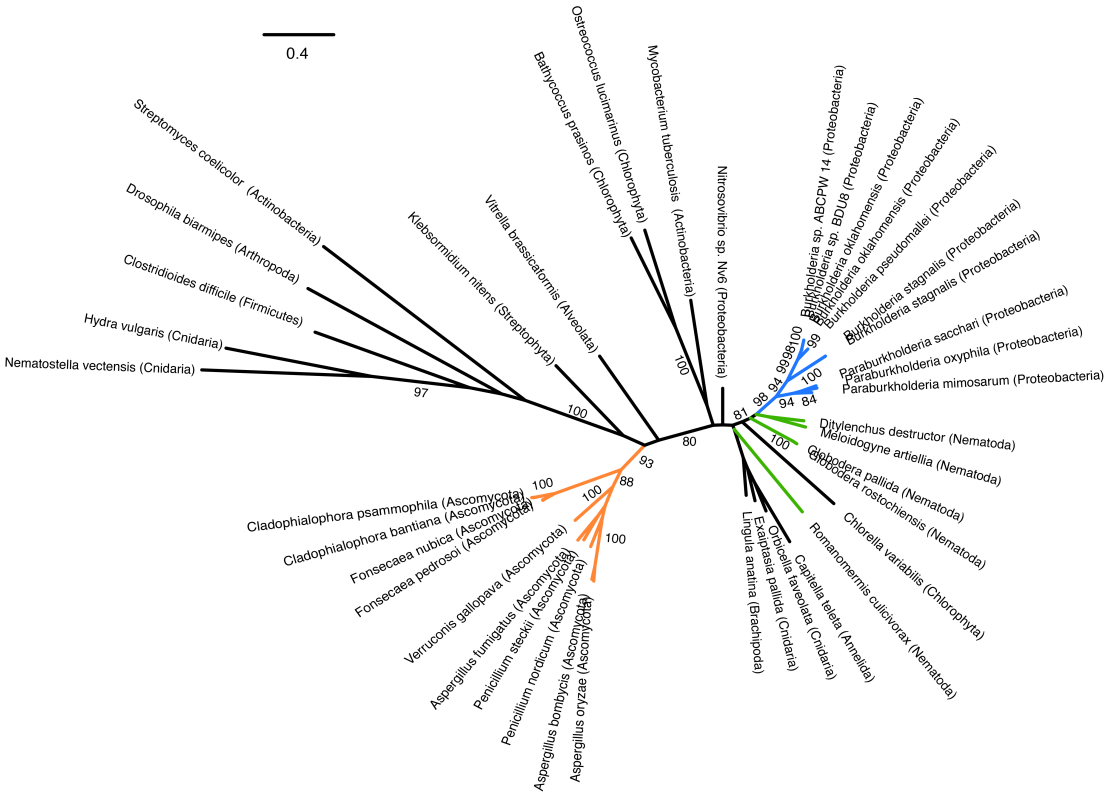


Figure 4.3.6: Phylogenetic tree of polyglutamate synthesis proteins. Non-parametric bootstrap support is only indicated for branches with support above 80. Blue: Monophyletic Proteobacteria clade (with the exception of a *Nitrosovibrio* sp. NV6 sequence). Orange: Monophyletic Ascomycota clade. Green: Nematode sequences.

from two (*B. xylophilus*) to ten (*M. arenaria*). Concerning representation count differences between PCN and other genera (Figure 4.3.7B), an extreme expansion was observed PCN within a BTB/POZ cluster 'OG0000116' synapomorphic to 'n21': while all *Meloidogyne* spp. contain one copy, *G. pallida* harbours 54 paralogues and *G. rostochiensis* displays 21. BTB/POZ domains are highly conserved structural motifs involved in protein-protein interactions (Stogios et al., 2005).

Pairwise protein count representation test implemented in KinFin require the clusters to be shared between the groups that are to be compared. In order to explore protein family expansion further, all clusters (as well as the subset of clusters containing PCN effectors) were visualised based on the number of proteins they

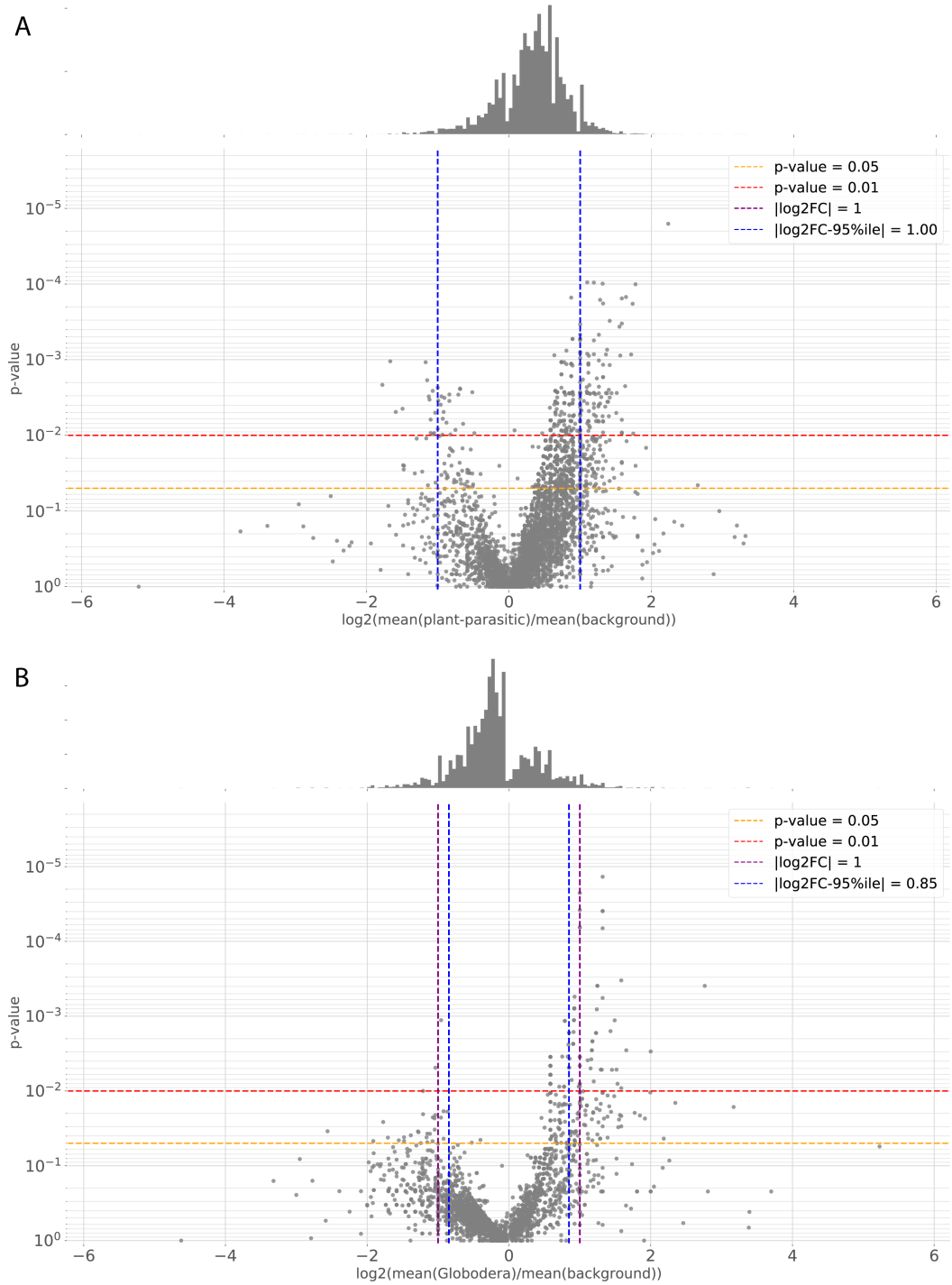


Figure 4.3.7: Volcano plots for results of pairwise protein count representation tests. The histogram (top) shows density of data points by location on the x-axis. **A**: results for tests between plant-parasitic taxa and all other taxa. **B**: results for tests between members of the genus *Globodera* and all other genera.

contain and the number of taxa that contributed to them, based on IPR and SignalP RFAs. The results are shown in Figure 4.3.8 and 4.3.9.

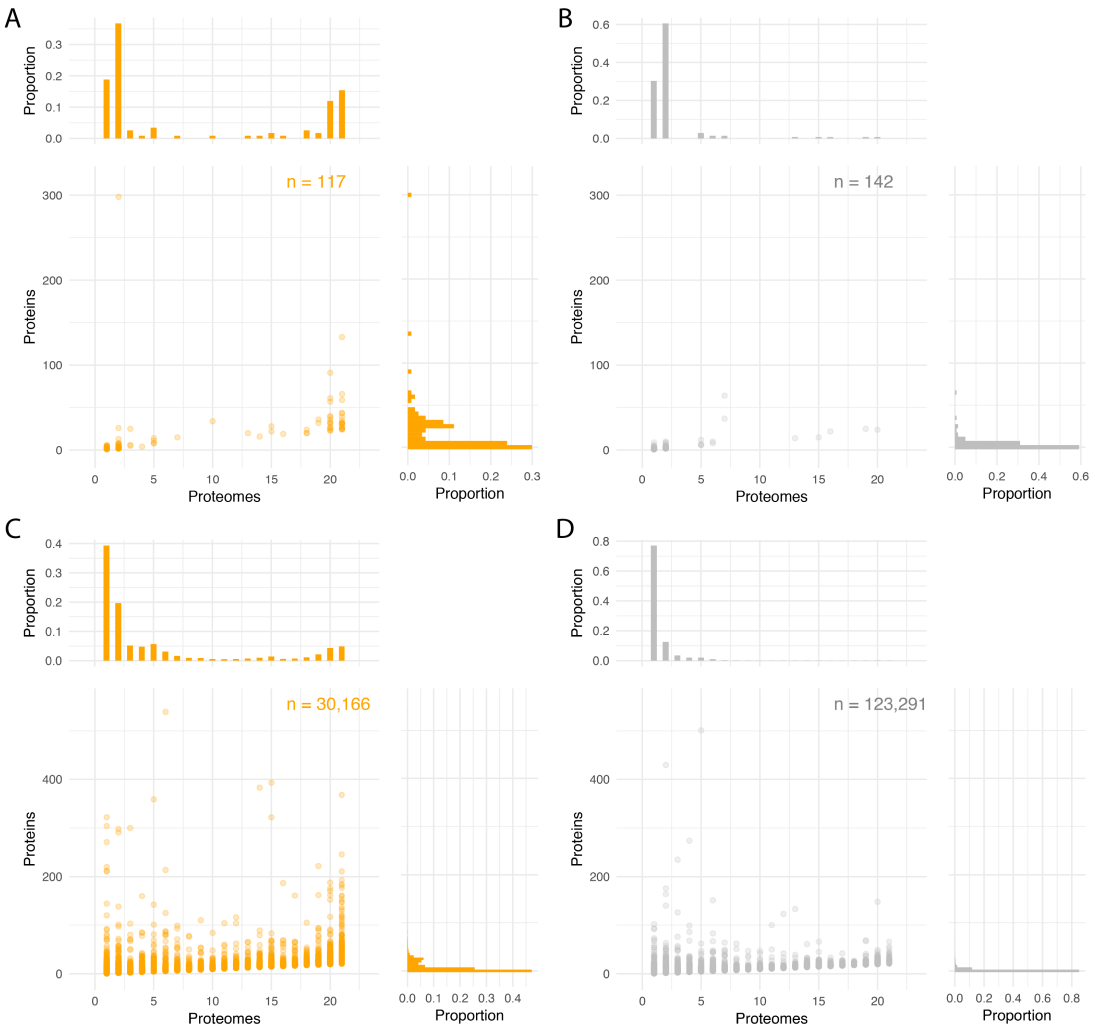


Figure 4.3.8: Visualisation of clusters based on IPR RFA. Clusters that received at least one IPR ID through RFA are visualised in **A** and **C**. Clusters that received no IPR ID through RFA are depicted in **B** and **D**. Scatter plots are decorated with histograms depicting the proportion of clusters in each bin. **A**: Effector clusters which received an IPR RFA. **B**: Effector clusters that which did not receive a IPR RFA. **C**: All clusters which received a IPR RFA. **D**: All clusters which did not receive a IPR RFA.

As expected, the proportion of effector clusters that received IPR RFAs (Figure 4.3.8A and B, 45.2%) is higher than for the complete clustering (Figure 4.3.8C

and D, 19.7%). The same is true for SignalP RFAs suggestive of secretion (Figure 4.3.9A and B, 53.3%, vs. Figure 4.3.9C and D, 10.8%). The distribution of effector clusters with IPR RFAs by number of containing taxa (Figure 4.3.8A, top histogram) display four clear peaks. The two peaks at taxon count one and two are due to singletons and PCN specific clusters, while the two peaks at 20 and 21 are caused by the synapomorphic effector clusters at node 'n0' in the tree. This effect is not as pronounced for effector clusters with SignalP RFAs suggestive of secretion (Figure 4.3.9A, top histogram), as very few clusters containing more than two taxa are annotated as such. This could be caused by increased domain shuffling involving signal peptides in PCN taxa, leading to neo-functionalisation of conserved proteins, but this remains to be tested.

There were three outliers in effector clusters with IPR RFAs which may be protein family expansions (Figure 4.3.8A, scatter plot): a cluster containing 298 proteins from both PCN taxa ('OG0000011'), one containing 133 proteins from all taxa ('OG0000050'), and one composed of 91 proteins from all taxa except *G. pallida* ('OG0000087'). 'OG0000011' is a SPRY domain cluster which contains the Gr-SPRYSEC-19 effector ('GROS_g14234') involved in host immunity suppression (Postma et al., 2012), the highly variable Gp-RBP-1 ('GPLIN_000437400') avirulence factor targeted by the host immune system (Sacco et al., 2009), as well as 276 other SPRY proteins from *G. pallida* and 20 from *G. rostochiensis*. However, within this cluster only 8.7% of proteins have SignalP annotation suggestive of secretion. This protein family expansion has been reported before (Cotton et al., 2014). 'OG0000050' is a 'CAP domain' cluster (IPR014044) involved in a multitude of cellular processes. The cluster was labeled as effector because it contains two *G. pallida* proteins which were recovered as orthologues to *H. glycines* ('AAK55116.1') and *G. rostochiensis* ('AHW98763.1') venom-allergen proteins, which have been shown to be involved in suppression of host immunity in certain PPNs (Lozano-Torres et al., 2014). It contains seven proteins from each PCN, while counts for

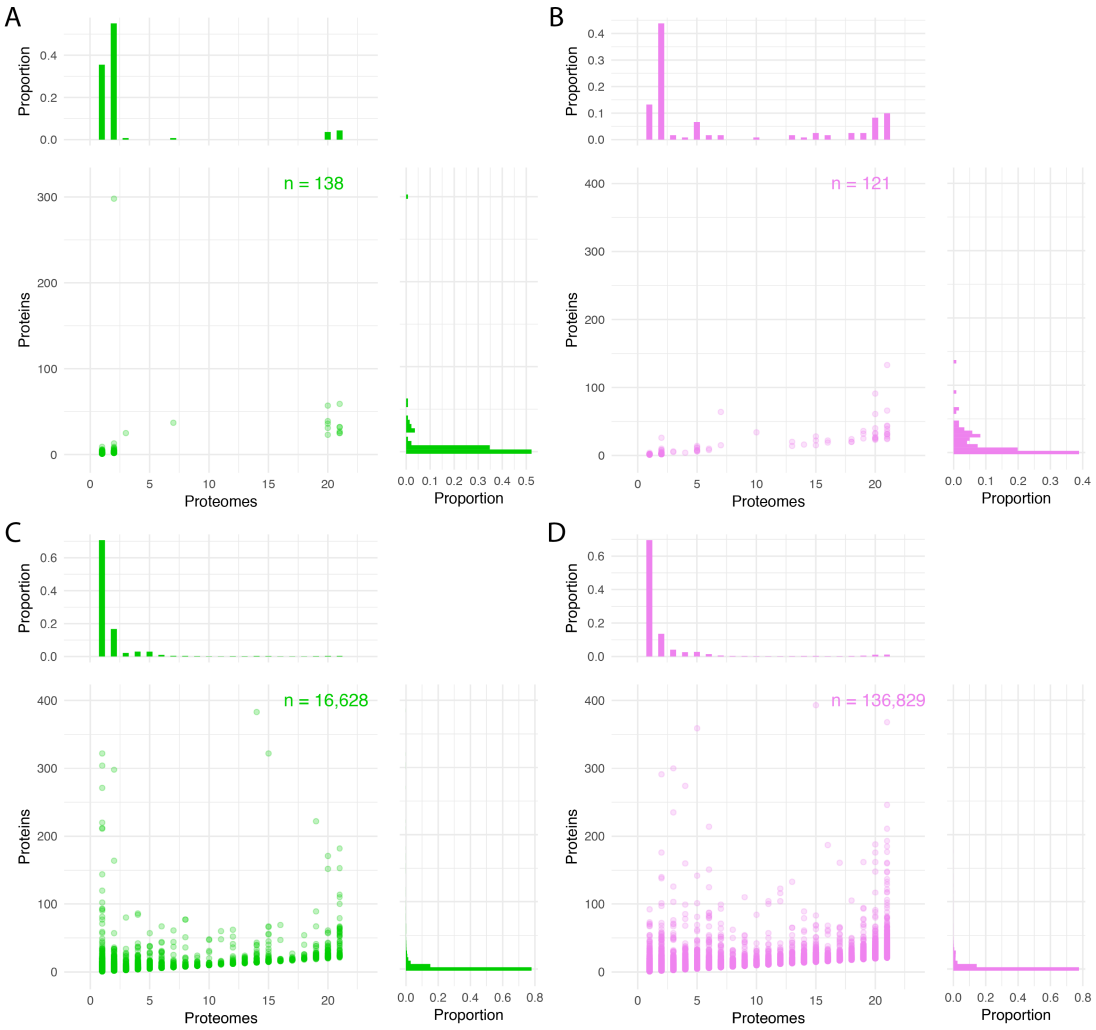


Figure 4.3.9: Clusters with SignalP RFA. Clusters that received a SignalP RFA of 'SignalP-noTM' or 'SignalP-noTM' in combination with 'SignalP-TM' are shown in **A** and **C**. Clusters that received a SignalP RFA of 'SignalP-TM' or received no RFA are shown in **B** and **D**. Scatter plots are decorated with histograms depicting the proportion of clusters in each bin. **A**: Effector clusters which received an SignalP RFA. **B**: Effector clusters that which did not receive a SignalP RFA. **C**: All clusters which received a SignalP RFA. **D**: All clusters which did not receive a SignalP RFA.

other taxa vary from one to 17 with no discernible pattern that could be linked to lifestyle or phylogeny. The cluster 'OG0000087' was annotated with the domains 'SKP1 component dimerisation' (IPR016072) and 'SKP1 component POZ domain' (IPR016073). Both domains are found in S-phase kinase-associated (SKP) proteins, which are involved in ubiquitin-mediated degradation of proteins involved in

core developmental processes (Nayak et al., 2002). The cluster only contains one protein from *G. rostochiensis* which was labeled as effector based on orthology to a SKP1 protein reported from the same species ('AHW98770.1'). This literature effector was shown to be expressed in key parasitic stages and might be involved in the fundamental developmental changes caused to the host cell during formation of the syncytium (Ali et al., 2015). Screening of the *G. pallida* proteome revealed three proteins annotated with domain IPR016072 and one with domain IPR016073, but no protein exhibiting both. The biggest cluster in this analysis is a BTB/POZ protein family composed of *C. briggsae*, and the five *Steinernema* species.

4.4 Conclusion

I analysed proteomes of Clade IV nematodes which contain taxa of medical, economical and agricultural importance. Evaluation of the effect of MCL inflation values on representative functional annotation of the resulting clusterings, revealed small differences across the parameter interval from 1.5 to 10.0. Unsurprisingly, taxonomic composition of the underlying proteome set has a much greater influence on representative functional annotation. Hence, estimation of optimal MCL inflation value for clusterings should be done for each dataset using, for example, the number of 'true' or 'fuzzy' 1-to-1 clusters. In this dataset, the distribution of 'true' and 'fuzzy' 1-to-1 clusters peaked around MCL Inflation values which also generated an intermediate distributions of RFA clusters.

KinFin output was used to infer a robust phylogeny for Clade IV nematodes based on 427 loci, which subsequently was used in a second KinFin run to investigate synapomorphic clusters at key nodes with respect to PCN effector proteins. Clusters of interest were identified for the basal node — shared by all nematodes in the analysis — suggesting that these effector proteins have either been repurposed for plant parasitism in PCNs or that these carry out deeply conserved functions that have simply been labeled as effector due to expression pattern in parasitic stages or their phenotype when disrupted. Most of the PCN effector proteins were, however, restricted to the PCN taxa. Analysis of synapomorphic clusters at the node representing the common ancestor of *Globodera* and *Meloidogyne* revealed a NodL-like acetyltransferase previously reported to be restricted to *Meloidogyne* species (Scholl et al., 2003). Phylogenetic analysis suggest monophyly of all nematode NodL-like acetyltransferases and could be explained by acquisition of the underlying genetic locus from a rhizobial bacterium prior to the split of Heteroderidae and Meloidogynidae. Additional genomes of other Heteroderidae are urgently needed to fully assess the patterns of effector protein family evolution in PCN, especially those

putatively acquired through horizontal gene transfer. During this analysis it also became apparent that several proteomes (*e. g. M. hapla*, *M. floridensis*, and *M. arenaria*) still suffer from high levels of contamination. This manifested itself mainly during exploration of synapomorphic clusters in Meloidogynidae, which uncovered clusters composed of proteins with high similarity to bacterial sequences originating from short scaffolds with no proximity to loci of obvious eukaryotic origin.

I generated a list of putative effector proteins in PCN proteomes based on RBBH results to published effector sequences. The use of RBBH analysis for identification of PCN effector seeds was preferred over the alternative of adding the effector proteins to the respective proteomes, as some species for which many effectors have been sequenced do not have proteomes predicted from genomes. A third option would have been the use of EST datasets, but would also have affected the clustering due to the amount of missing data. The RBBH analysis revealed a list of 456 effectors identified in the proteomes of PCNs, of which 235 have not been labeled as effectors by Thorpe et al., 2014 and Eves-van den Akker et al., 2016b. This list will be used in further analysis in Chapter 5.

Appendix

4.A Tables

Table 4.A.1: Literature effector proteins used in RBBH analysis.

ID	Product	Source	Species	Publication
ANB41563.1	tyrosinase-like	NCBI	*Heterodera schachtii*	Habash et al., 2017
AAC05133.1	cellulase	NCBI	*Meloidogyne incognita*	Ding et al., 1998
AAC48325.1	GH5 cellulase	NCBI	*Globodera rostochiensis*	Smant et al., 1998
AAC48326.1	GH5 cellulase	NCBI	*Heterodera glycines*	Smant et al., 1998
AAC48327.1	GH5 cellulase	NCBI	*Heterodera glycines*	Smant et al., 1998
AAC48341.1	GH5 cellulase	NCBI	*Globodera rostochiensis*	Smant et al., 1998
AAD27559.1	VAP	NCBI	*Caenorhabditis elegans*	unpublished
AAF80747.1	pectate lyase	NCBI	*Globodera rostochiensis*	Vanholme et al., 2007
AAK08974.1	pectate lyase	NCBI	*Heterodera glycines*	De Boer et al., 2002
AAK21961.1	VAP	NCBI	*Heterodera glycines*	Gao et al., 2003
AAK55116.1	VAP	NCBI	*Heterodera glycines*	Gao et al., 2001b

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AAK60209.1	VAP	NCBI	*Heterodera glycines*	Gao et al., 2001b
AAK85303.1	GH5 cellulase	NCBI	*Heterodera glycines*	Gao et al., 2002a
CAC84452.1	polyglutamate synthase	NCBI	*Meloidogyne artiellia*	Veronico et al., 2001
AAL40720.1	calreticulin	NCBI	*Meloidogyne incognita*	Li et al., 2015
AAL78229.1	Hgg18	NCBI	*Heterodera glycines*	Gao et al., 2001a
AAM18623.1	C-type lectin	NCBI	*Heterodera glycines*	Rehman, Gupta, and Goyal, 2016
AAM21970.1	pectate lyase	NCBI	*Globodera rostochiensis*	Vanholme et al., 2007
AAM28240.1	polygalacturonase	NCBI	*Meloidogyne incognita*	Rehman, Gupta, and Goyal, 2016
AAM50038.1	G27D09	NCBI	*Heterodera glycines*	Gao et al., 2003
AAM50039.1	GH5 cellulase	NCBI	*Heterodera glycines*	Gao et al., 2003
AAM74953.1	chorismate mutase	NCBI	*Heterodera glycines*	Gao et al., 2003
AAM74954.1	pectate lyase	NCBI	*Heterodera glycines*	Gao et al., 2003
AAM95699.2	33A09	NCBI	*Heterodera glycines*	Gao et al., 2003

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AAN08583.1	msp17	NCBI	*Meloidogyne incognita*	Huang et al., 2003
AAN08587.1	msp21	NCBI	*Meloidogyne incognita*	Huang et al., 2003
AAN14978.1	chitinase	NCBI	*Heterodera glycines*	Gao et al., 2002b
AAN32884.1	GH5 cellulase	NCBI	*Heterodera glycines*	Gao et al., 2003
AAN32886.1	G2D01	NCBI	*Heterodera glycines*	Gao et al., 2003
AAN32887.1	cellulose binding protein	NCBI	*Heterodera glycines*	Gao et al., 2003
AAN32888.1	4C10	NCBI	*Heterodera glycines*	Gao et al., 2003
AAN32889.1	ubiquitin-extension	NCBI	Heterodera glycines	Gao et al., 2003
AAN32890.1	4D09	NCBI	*Heterodera glycines*	Gao et al., 2003
AAN32891.1	5D06	NCBI	*Heterodera glycines*	Gao et al., 2003
AAN32892.1	4D06	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO33473.1	G4E02	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO33475.1	G5D08	NCBI	*Heterodera glycines*	Gao et al., 2003

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AAO33476.1	G6E07	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO33477.1	G4G05	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO33478.1	ubiquitin-extension	NCBI	Heterodera glycines	Gao et al., 2003
AAO85452.1	G12H04	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO85454.1	G16B09	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO85455.1	G17G06	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO85456.1	G18H08	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO85457.1	G19B10	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO85458.2	G19C07	NCBI	*Heterodera glycines*	Gao et al., 2003
AAO85459.1	G20E03	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30754.1	G11A06	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30755.1	29D09	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30756.1	10C02	NCBI	*Heterodera glycines*	Gao et al., 2003

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AAP30757.1	30G12	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30758.1	25A01	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30759.1	G13A06	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30760.1	G10A07	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30761.1	G20G04	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30762.1	G7E05	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30763.1	G8H07	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30764.1	G28B03	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30766.1	G30D08	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30767.1	G21E12	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30768.1	G22C12	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30772.1	G23G12	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30773.1	G24A12	NCBI	*Heterodera glycines*	Gao et al., 2003

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AAP30774.1	G30E03	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30775.1	G32E03	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30776.1	G34B08	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30834.1	G10A06	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30835.2	G33E05	NCBI	*Heterodera glycines*	Gao et al., 2003
AAP30836.1	G30C02	NCBI	*Heterodera glycines*	Gao et al., 2003
AAQ09004.1	pectate lyase	NCBI	*Meloidogyne incognita*	Huang et al., 2003
AAQ10016.1	msp2	NCBI	*Meloidogyne incognita*	Huang et al., 2003
AAQ10025.1	msp10	NCBI	*Meloidogyne incognita*	Huang et al., 2003
AAQ97032.1	pectate lyase	NCBI	*Meloidogyne incognita*	Huang et al., 2004
AAR37371.1	msp40	NCBI	*Meloidogyne incognita*	Niu et al., 2016
AAX68678.1	aminopeptidase (M1)	NCBI	*Heterodera glycines*	Rehman, Gupta, and Goyal, 2016
AAY84711.2	dual oxidase	NCBI	*Meloidogyne incognita*	Rehman, Gupta, and Goyal, 2016

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
ABI33933.1	16D10	NCBI	*Meloidogyne hapla*	Rehman, Gupta, and Goyal, 2016
ABL61274.1	VAP	NCBI	*Meloidogyne arenaria*	unpublished
ABN14272.1	pectate lyase	NCBI	*Heterodera schachtii*	Peng et al., 2016
ABN14273.1	pectate lyase	NCBI	*Heterodera schachtii*	Peng et al., 2016
ABO38109.1	VAP	NCBI	*Meloidogyne incognita*	Wang et al., 2007
ACU64826.1	pectate lyase	NCBI	*Globodera pallida*	Peng et al., 2016
ACV33082.1	calreticulin	NCBI	*Ditylenchus destructor*	Li et al., 2015
ACY70448.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ACY70450.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ACY70451.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ACY70452.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ACY70453.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ACY70454.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
ACY70455.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ACY70456.1	CLE	NCBI	*Globodera rostochiensis*	Lu et al., 2009
ADC35399.1	VAP	NCBI	*Ditylenchus destructor*	unpublished
ADD82420.1	calreticulin	NCBI	*Bursaphelenchus xylophilus*	Li et al., 2011
ADF28634.1	chaperonin	NCBI	*Heterodera glycines*	Rehman, Gupta, and Goyal, 2016
ADG86237.1	VAP	NCBI	*Bursaphelenchus xylophilus*	Lin et al., 2011
ADG86238.1	VAP	NCBI	*Bursaphelenchus xylophilus*	Lin et al., 2011
ADG86239.1	VAP	NCBI	*Bursaphelenchus xylophilus*	Lin et al., 2011
ADV57652.1	VAP	NCBI	*Bursaphelenchus mucronatus*	Lin et al., 2011
ADV57653.1	VAP	NCBI	*Bursaphelenchus doui*	Lin et al., 2011
ADW77532.1	pectate lyase	NCBI	*Heterodera glycines*	Peng et al., 2016
ADW77534.1	pectate lyase	NCBI	*Heterodera glycines*	Peng et al., 2016
ADW77536.1	pectate lyase	NCBI	*Heterodera glycines*	Peng et al., 2016

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
ADW77537.1	rbp-2	NCBI	*Heterodera glycines*	Maier et al., 2013
AFH68236.1	1106	NCBI	*Globodera rostochiensis*	unpublished
AFI80890.1	FAR	NCBI	*Radopholus similis*	unpublished
AFK76483.1	calreticulin	NCBI	*Radopholus similis*	Li et al., 2015
AFN86180.1	SPRYSEC-19	NCBI	*Globodera rostochiensis*	Qin et al., 2000
AFQ55440.1	VAP	NCBI	*Ditylenchus destructor*	unpublished
AFZ77091.1	FAR	NCBI	*Meloidogyne javanica*	Iberkleid et al., 2013
AGA60308.1	FAR	NCBI	*Aphelenchoides besseyi*	Cheng et al., 2013
AHW83206.1	VAP	NCBI	*Meloidogyne incognita*	unpublished
AHW98758.1	SPRYSEC-4	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98760.1	GH5 cellulase	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98761.1	GH5 cellulase	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98762.1	GH5 cellulase	NCBI	*Globodera rostochiensis*	Ali et al., 2015

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AHW98763.1	VAP	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98765.1	pectate lyase	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98766.1	metalloprotease	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98767.1	E9 protein	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98768.1	putative amphid protein	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98769.1	glutathione peroxidase	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98770.1	Skip-1	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98771.1	peroxiredoxin	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHW98772.1	D406	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AHZ59334.1	SPRY-15	NCBI	*Globodera rostochiensis*	Ali et al., 2015
AIT18660.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18661.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18662.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AIT18663.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18664.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18666.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18667.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18668.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18669.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18670.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18671.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18672.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18674.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18675.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18676.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18677.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AIT18678.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18679.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18680.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18681.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18682.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18683.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18684.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18685.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18686.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18687.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18688.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18689.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18692.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AIT18696.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18697.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18698.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18699.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18701.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18702.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18703.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18706.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18707.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18708.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18709.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18710.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18711.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AIT18712.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18713.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18714.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18715.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18716.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18717.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18718.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18719.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18720.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18721.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18722.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18723.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014
AIT18724.1	HYP	NCBI	*Globodera pallida*	Eves-van den Akker et al., 2014

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AIW66697.1	calreticulin	NCBI	*Pratylenchus goodeyi*	Li et al., 2015
AJR19769.1	Hg-GLAND-1	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19770.1	Hg-GLAND-2	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19771.1	Hg-GLAND-3	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19772.1	Hg-GLAND-4	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19773.1	Hg-GLAND-5	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19774.1	Hg-GLAND-6	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19775.1	Hg-GLAND-7	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19776.1	Hg-GLAND-8	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19777.1	Hg-GLAND-9	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19778.1	Hg-GLAND-10	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19779.1	Hg-GLAND-11	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19780.1	Hg-GLAND-12	NCBI	*Heterodera glycines*	Noon et al., 2015

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
AJR19781.1	Hg-GLAND-13	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19782.1	Hg-GLAND-14	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19783.1	Hg-GLAND-15	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19784.1	Hg-GLAND-16	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19785.1	Hg-GLAND-17	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR19786.1	Hg-GLAND-18	NCBI	*Heterodera glycines*	Noon et al., 2015
AJR22224.1	10A07	NCBI	*Heterodera schachtii*	Hewezi et al., 2015
AKR17057.1	CLE	NCBI	*Rotylenchus reniformis*	Wubben et al., 2015
AKR17058.1	CLE	NCBI	*Rotylenchus reniformis*	Wubben et al., 2015
AKR17059.1	CLE	NCBI	*Rotylenchus reniformis*	Wubben et al., 2015
ALX34942.1	FAR	NCBI	*Heterodera avenae*	unpublished
BAE48369.1	pectate lyase	NCBI	*Bursaphelenchus xylophilus*	Kikuchi et al., 2006
BAE48370.1	pectate lyase	NCBI	*Bursaphelenchus xylophilus*	Kikuchi et al., 2006

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
BAE48373.1	pectate lyase	NCBI	*Bursaphelenchus mucronatus*	Kikuchi et al., 2006
BAE48375.1	pectate lyase	NCBI	*Bursaphelenchus mucronatus*	Kikuchi et al., 2006
CAA70477.2	FAR	NCBI	*Globodera pallida*	unpublished
CAB66341.1	putative amphid protein	NCBI	*Globodera rostochiensis*	Rehman, Gupta, and Goyal, 2016
CAC21847.1	A4	NCBI	*Globodera rostochiensis*	Qin et al., 2000
CAD60975.1	A42	NCBI	*Globodera rostochiensis*	unpublished
CAD60977.1	E9 protein	NCBI	*Globodera rostochiensis*	unpublished
Q9BN21.1	CLE	NCBI	*Heterodera glycines*	Gao et al., 2003
Q86RQ1.1	CLE	NCBI	*Heterodera glycines*	Gao et al., 2003
RrCEP1	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP2	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP3	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP4	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a

Table 4.A.1 Continued from previous page

ID	Product	Source	Species	Publication
RrCEP5	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP6	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP7	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP8	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP9	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP10	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP11	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a
RrCEP12	CEP	publication	*Rotylenchus reniformis*	Eves-van den Akker et al., 2016a

Chapter 5

Comparative genomics of the *Globodera* species complex

"An old adage assures us there is no royal road to knowledge. There is certainly no royal road to a knowledge of nematodes. The traffic in this direction has not justified the installation of through trains and sleeping cars; so he who takes this route must be prepared to put up with inconveniences, and to make the best of certain disgusting passages."

- Nathan A. Cobb, *Nematodes and Their Relationships, Yearbook of the US Department of Agriculture, 1914*

5.1 Introduction

5.1.1 Potato cyst nematodes

Potato cyst nematodes (PCNs) are sedentary tylenchid endoparasites of the genus *Globodera* that parasitise *Solanum tuberosum* (potato) and several other solenaceous hosts. PCNs originated and co-evolved with their hosts in South America (Plantard et al., 2008; Evans and Stone, 1977) and were introduced into Europe in the 19th century as a result of resistance breeding programs against potato blight (Evans, Franco, and De Scurrah, 1975). From Europe they have spread to other regions in Asia, Africa and the US, and are considered a major pest in temperate regions (Franklin, 1951; Hockland et al., 2012; Mimee et al., 2015). The implications of the independent introduction events on population structure, effector diversity and phylogeography are only poorly understood. Early attempts to characterise different populations of morphologically indistinguishable strains, which exhibit distinct patterns of virulence, resulted in a pathotype naming scheme based on multiplication rates on host plants (Kort et al., 1977).

The losses in crop yield for UK potato farmers caused by the PCNs *G. pallida* and *G. rostochiensis* are estimated to reach £50 million per year or 9% of production (DEFRA, 2010). The discovery of a single resistance gene in potato acting against the 'Ro1' pathotype of *G. rostochiensis* in the 1970s has selected for *G. pallida* in mixed populations and therefore allowed *G. pallida* to become the prevalent metazoan parasite of potatoes in Northern Europe (Minnis et al., 2002). For *G. pallida* no comparable single, dominant resistance gene in the host is available and research has focussed on quantitative trait loci which confer a certain degree of resistance but are difficult to breed and can be overcome by virulent pathotypes. Furthermore, the ability of PCNs to remain dormant in the soil for many years makes

crop rotation unfeasible (Trudgill, Phillips, and Elliott, 2014) and recent legislation has limited the use of nematicides to control infestation (Clayton et al., 2008). As a consequence there is great need to understand the biology of these parasites and their interactions with the host in order to develop novel approaches to guarantee sustainable and competitive potato production within the UK. One effort in this direction started in 2008 in the form of the *Globodera* genome sequencing project and culminated in a reference genome of *G. pallida* and the analysis of gene expression profiles throughout the parasite's life cycle (Cotton et al., 2014). In addition to the reference strain *G. pallida* 'Lindley' (pathotype Pa2/3), genome data for several other populations of *G. pallida* from the UK and South America as well as for the sister species *G. rostochiensis* were generated during the project.

5.1.2 *Globodera pallida*

The reference genome and life-stage specific analysis of gene expression for *G. pallida* by Cotton et al., 2014 offers a comprehensive and detailed view of the biology of this parasite. Potato cyst nematodes begin their life cycle as second stage juveniles (J2) as they hatch from cysts in the soil. This occurs in response to an environmental cue in the form of host root secretions. The emergence from dormancy is associated with large-scale up-regulation of transcriptional activity. High levels of expression were observed for genes whose products are involved in carbohydrate metabolism, defence responses against pathogens and the plant immune system, and poly-A transferase activity (Cotton et al., 2014). The J2 larva penetrates the host cell wall using its stylet and migrates intercellularily to the vascular cylinder of the root in order to establish the initial feeding site (syncytium), a feeding cell composed of fused host cells which provides the nematode with nutrients. Establishment of the feeding site requires substantial changes to the host cell structure, including localised cell wall degradation and protoplast fusion to progressively enlarge the feeding cell until eventually up to 200 neighbouring cells are incorporated (Lilley, Atkinson, and Urwin, 2005). Genes up-regulated in J2 were enriched for products with signal peptides, suggesting secretion of proteins involved in this step. During the three to six week long development into an adult male or an egg-laying female, the J2 undergoes three moulting steps (Sobczak and Golinowski, 2011). The transitions through these early parasitic stages correlate with the largest changes in gene expression during the life cycle of the nematode. Down-regulated gene classes include those coding for proteins involved in signal transduction, chemotaxis and neurotransmission, as expected during the transition of a free-living organism to a sedentary parasite (Cotton et al., 2014). Up-regulation is observed for genes coding for products involved in lipid metabolism and protein degradation, as well as a large group of glutathione synthesase proteins, suggested to be involved in neutralising plant defences — break-down of cytotoxic hydrogen

peroxide released by the plant upon cell damage — and alteration of signalling or regulation of plant development in which glutathione has important roles. The sex of the adult nematode is determined by its success in establishing the feeding site, to be specific by the size of the syncytium and its proximity to vascular tissue which guarantees abundance of nutrients. This environmental mode of sex determination results in a greater proportion of males as population density increases (Phillips, Forrest, and Farrer, 1982; Sobczak and Golinowski, 2011). The sexual fate is fixed shortly before the moult to J3. Males feed until the end of J3, emerge as J4 from the feeding site and migrate to find females. Although both pre-parasitic J2 larva and males migrate through host cells they display only minor similarities in differentially expressed genes, limited to genes related to neuromuscular functions. Up-regulation in non-feeding, migratory males is observed primarily for genes associated with storage mobilisation, protein and lipid metabolism, and sperm production. In contrast, female worms begin to enlarge and adopt a spherical shape upon reaching adulthood. Once the female is fertilised, the embryos develop into J2 larva inside the body of the female and enter dormancy. Subsequently, the cuticle of the female transforms into a robust cyst which becomes detached from the root after the death of the host plant. The J2 larva inside the cyst are able to survive in the soil for decades in the absence of a host (Spears, 1968).

Populations of *G. pallida*

During the *Globodera* genome project, genome data were generated for seven populations of *G. pallida*, including the reference population 'Lindley'. The populations are listed in Table 5.1.1. The samples comprise five populations from the UK and two from South America.

The two South American populations have been reported to be distinct from

Table 5.1.1: Genomic reads of *G. pallida* populations. All reads were generated on the Illumina HiSeq2000 platform from PE libraries with an insert size of 475 b. Reads are available on ENA.

ENA Run ID	Population	Pathotype	Origin	Reads
ERR114517	'Lindley'	Pa2/3	England (UK)	78,227,699
ERR123952	'Bedale'	Pa2/3	England (UK)	143,620,834
ERR123953	'Luffness'	Pa2/3	Scotland (UK)	135,215,151
ERR123954	'Newton'	Pa2/3	England (UK)	123,555,048
ERR123955	'Pa1'	Pa1	Scotland (UK)	129,324,633
ERR123956	'P5A'	P5A	Peru	134,270,353
ERR123957	'P4A'	P4A	Peru	126,694,929

each other and from the European populations based on molecular markers (Blok and Phillips, 1995; Blok, Phillips, and Harrower, 1997; Blok et al., 1998; Subbotin et al., 2000). The population 'P4A' appears to be more closely related to European populations. The 'Lindley' population is a representative of the virulent pathotype 'Pa2/3' which is able to overcome the H2 resistance from *Solanum multidissectum*, as are 'Bedale', 'Newton', and the Scottish 'Luffness' population. The latter population has been suggested to be a result of a separate introduction to Europe as it has been shown to be distinct from other European populations, and to share an ancestral relationship with the 'P4A' population from Peru (Pylypenko et al., 2005; Madani et al., 2010). The other UK 'Pa2/3' populations have been shown to constitute a monophyletic group, based on 250 random amplified polymorphic DNA (RAPD) markers (Blok, Phillips, and Harrower, 1997). The population 'Pa1' has morphotypes not present in other populations of *G. pallida* — females are 'cream' coloured as opposed to 'white' — and lacks the virulence gene necessary to

overcome the H2 resistance (Phillips et al., 1992). However, previous studies have been based on limited number of loci and the heterogeneity of some populations of *G. pallida* — paired with the lack of suitable marker loci for population membership delimitation — have resulted in a complex picture of population structure.

5.1.3 *Globodera rostochiensis*

The yellow potato cyst nematode *G. rostochiensis* comprises a lesser threat to UK potato industry compared to *G. pallida*, since the dominant pathotype 'Ro1' in the UK can be controlled by a single major resistance locus in potato crops (H1). A single introductory event of *G. rostochiensis* into Europe has been suggested (Phillips and Trudgill, 1998; Hockland et al., 2012). In contrast, introduction to the USA and Canada appears to have occurred within the past century and was initially confined to a few localised regions, due to enforcement of strong quarantine measures (Olsen and Mulvey, 1962; Orchard, 1965). However, despite these efforts an increase in outbreaks has been observed since 2006 (Sun et al., 2007; Mahran et al., 2010) and research into detection and management of *G. rostochiensis* are of major interest. High-throughput genotype-by-sequencing (GBS) approaches have been applied to populations of *G. rostochiensis* in order to resolve the population structure of 'Ro1' in Canada (Mimee et al., 2015). This approach is based on sequencing of DNA of pooled samples of cysts, which has been digested using restriction enzymes. Reads are subsequently processed and *de novo* assembled through the `UNEAK` (Universal Network Enabled Analysis Kit) pipeline (Lu et al., 2013), based on which variants are called.

A reference genome assembly for the 'Ro1' population of *G. rostochiensis* was generated within the scope of the *Globodera* genome project by staff at the Wellcome Trust Sanger Institute. I was invited to coordinate the genome annotation. Together

with Mark Blaxter and Sebastian Eves van-den-Acker, I organised a collaborative manual genome curation event at the University of Edinburgh, during which 15 researchers from Canada, the Netherlands, USA, and UK revised and improved gene models I previously predicted on the assembly. This was made possible through the use of a Badger genome exploration environment (Elsworth, Jones, and Blaxter, 2013) — to query functional annotations of the *G. pallida* and *G. rostochiensis* genomes — and a WebApollo (Lee et al., 2013) instance for the collaborative curation of structural gene predictions. Both web services were configured by Michael Clarke and myself. During and after the curation event, approximately one-eighth (1566) of the predicted gene models were inspected and, if necessary, refined. I used the resulting set of curated gene models to re-predict gene models on the assembly, which served as basis for subsequent comparative genomics analyses. The results were published by Sebastian Eves van-den-Acker and myself as joint first authors in the journal 'BMC Genome Biology' as Eves-van den Akker et al. (2016b) (DOI: 10.1186/s13059-016-0985-1). Furthermore, I coordinated the deposition of the *G. rostochiensis* genome assembly and its final annotation on WormBase ParaSite.

5.1.4 *Globodera ellingtonae*

Based on morphological and molecular differences to known *Globodera* species, in 2012 a new PCN species, *G. ellingtonae*, was described from samples collected from potato fields in Oregon, USA (Handoo et al., 2012). Marker sequences of ITS1 and 28S rDNA suggested that *G. ellingtonae* is more similar to the tobacco cyst nematode *G. tabacum* and to *G. rostochiensis*, than to *G. pallida*. A high-quality draft genome assembly of *G. ellingtonae*, generated from long (PacBio) and short (Illumina MiSeq and HiSeq) read data, was published recently (Phillips et al., 2017), but no gene annotation is available yet.

5.1.5 Comparative genomics of potato cyst nematodes

In this chapter, I describe analyses I performed on the genome assemblies of *G. pallida*, *G. rostochiensis*, and other nematodes. These include some analysis published in Eves-van den Akker et al. (2016b) (DOI: 10.1186/s13059-016-0985-1), such as comparison of quality metrics between the genome assemblies of *G. pallida* and *G. rostochiensis*, lack of conservation of synteny between both genomes, and patterns of non-canonical splice-sites across Nematoda. Here, I estimate the rate of variation for the reference populations used to generate the genome assemblies of *G. pallida* ('Lindley') and *G. rostochiensis* ('Ro1') and present new results concerning the phylogeography of *G. pallida* populations and assess their genomic variation.

5.2 Methods

5.2.1 Data

Files related to the genome assemblies of *G. pallida* and *G. rostochiensis* were retrieved from WBPS8: assemblies, repeat-masked assemblies (soft-masked), and annotations in GFF3. In addition, assemblies and annotations in GFF3 were downloaded for all 100 nematodes species on WBPS8. The genome assembly of *G. ellingtonae* ('ASM172322v1') was retrieved from NCBI (BioSample ID 'SAMN04393202'). A novel genome annotation of *Heterorhabditis bacteriophora* in GFF3 was provided by Flo McLean. Read datasets for the *G. pallida* populations listed in Table 5.1.1 and for the 'Ro1' population of *G. rostochiensis* ('ERR123958') were downloaded from ENA. Intronic features were added to annotation files using `GenomeTools v1.5.9` (Gremme, Steinbiss, and Kurtz, 2013) (`gt gff3 -sort -tidy -retainids -fixregionboundaries -addintrons`) and the resulting output was converted to BED format using `GNU awk`, `GNU sed` and `Perl`. For subsequent processing of BED files, `BedTools v2.26.0` (Quinlan and Hall, 2010) was used.

5.2.2 Assessment of regions in PCN genomes

For *G. pallida*, *G. rostochiensis* and *G. ellingtonae*, feature track files in BED format were created using the script `masked_fasta2bed` (`https://github.com/DRL/thesis`) based on repeat-masked assemblies. The resulting BED files delimit regions composed of canonical nucleotides ('AGCT'), unknown regions ('N'), IUPAC ambiguity codes for nucleotides ('MRWSYKVHDB'), and repeats (nucleotides in lower case). For the assemblies of *G. pallida* and *G. rostochiensis*, the BED files

were filtered to only include runs of N's shorter than 10 b. These BED files delimit the regions in each assembly on which further analyses are performed.

5.2.3 Synteny analysis

Synteny analysis between the *G. pallida* and *G. rostochiensis* genomes was performed as described in Eves-van den Akker et al. (2016b). In brief, synteny between scaffolds was assessed based on a `OrthoMCL` clustering (Li, Stoeckert, and Roos, 2003) (at MCL inflation value of 1.5) using `i-adhore-3.0.01` (Simillion et al., 2008). The *G. rostochiensis* scaffold 'GROS_00007' was visualised together with the four largest homologous *G. pallida* scaffolds using `circos v0.67-7` (Krzywinski et al., 2009), including GC-content and `BLASTn` results (E-value cutoff of 10^{-65}) between the scaffolds.

5.2.4 Splice sites

During the collaborative curation event for the *G. rostochiensis* genome assembly, several attendees reported a high incidence of non-canonical splice sites (GC/AG as opposed to GT/AG). Using the script `extractRegionFromCoordinates.py` (available at `https://github.com/DRL/GenomeBiology2016_globodera_rostochiensis`), I extracted splice donor and acceptor sites from GFF3 files of the two PCN genomes and genomes of representative species across the phylogeny of Nematoda for subsequent analysis by collaborators (see Eves-van den Akker et al., 2016b). I repeated this analysis for the 100 nematode genomes deposited on WBPS8 and contrasted levels of non-canonical GC/AG splice sites with the N50 metric, intron count and number of unique proteins to each genome, defined as the

sum of proteins present in singletons and proteome-specific clusters based on the KinFin analysis in Section 3.6.

5.2.5 Sequencing of additional *G. pallida* populations

Cysts derived from a single mother cyst from each of the *G. pallida* populations 'Pa1' and 'Luffness' were reared in root-trainers on the susceptible potato cultivar 'Desiree' by Vivian Blok at the James Hutton Institute in Dundee. Two months after inoculation, I sampled 29 ('Pa1_A') and 21 ('Pa1_B') adult females from two 'Pa1' cohorts and 24 adult females from one 'Luffness' cohort. Sample 'Pa1_B' was collected from a single root which is equivalent to sampling 'sister' worms, while 'Pa1_A' and 'Luffness' were sampled from multiple roots and are therefore 'cousins'. I carried out DNA extractions based on a protocol developed by AurÃľlien Richard for *C. elegans* which I modified for low-input samples. The protocol is available at `https://github.com/DRL/thesis`. The resulting DNA concentrations varied between 4.5 and 5.4 ng/μl, measured on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific). From each sample, 1 μl was used to carry out whole genome amplification reactions using the REPLI-g UltraFast Kit (Quiagen). This yielded DNA concentrations of 106.7 ng/μl ('Pa1_A'), 32.0 ng/μl ('Pa1_B'), and 67.2 ng/μl ('Luffness'). From each sample, two Illumina NexteraXT sequencing libraries — based on un-amplified ('WGS_NX') and amplified ('WGA_NX') DNA — and one Illumina Nextera sequencing library — based on amplified DNA ('WGA_N') — were generated by Edinburgh Genomics and sequenced on the Illumina HiSeq4000 platform. These datasets are referred to as 'bottlenecked' populations hereinafter.

5.2.6 Quality and adapter trimming of reads

Read datasets of the *G. rostochiensis* 'Ro1' population, seven *G. pallida* populations (listed in Table 5.1.1), and the nine samples from the 'bottlenecked' populations were adapter and quality trimmed via `trimmomatic v0.36` (Bolger, Lohse, and Usadel, 2014) using default parameters, except that minimum length of trimmed reads was set to 50. Only paired reads were kept for subsequent analysis.

5.2.7 Read mapping

Trimmed reads were mapped using `bwa mem v0.7.15-r1140` (Li, 2013) and `samtools v1.5` (Li et al., 2009). Only read pairs for which both reads mapped as 'proper pairs' — reads mapping in forward-reverse orientation with an insert size falling within a distribution based on 256,000 reads pairs — were kept. Duplicated reads were flagged using `Picard Tools v2.9.0 MarkDuplicates` (`http://broadinstitute.github.io/picard/`) and BAM files were sorted using `samtools v1.5`.

5.2.8 Coverage analysis of PCN datasets

Coverage information was extracted from BAM files via `BedTools genomecov` (`-bga`) to create BED files containing base coverage for each region in the assembly. These regions were filtered to only contain regions with five or more reads mapping to them. These files delimit known regions in the assembly for which each read dataset provides sufficient coverage for further analysis. UpsetR plots were created using the script `coverage_upsetr.R` based on data extracted via the script `generate_upsetr_expression.py` (both are available at `https://github.`

com/DRL/thesis). In the case of the 'Lindley' and 'Ro1' coverage BED files, additional BED files were created by excluding regions containing more than twice the median coverage to exclude regions of high coverage due to repetitive regions. Coverage decay plots were generated for CDS regions of the *G. pallida* assembly using each of the *G. pallida* BAM files via the script `bamCov.py` (`https://github.com/DRL/thesis`).

5.2.9 Variant calling

Variant calling was performed via `FreeBayes v1.1.0` (Garrison and Marth, 2012) with parameters geared towards analysis of pooled samples of diploid organisms. Only reads with a mapping quality of 20 or more were considered for variant calling, and the minimum read depth and number of reads supporting an alternate allele was set to five. The maximum read depth per sample was set to 500 (`-n 4 --strict-vcf -p 2 --haplotype-length 0 -m 1 -q 20 -Q 20 -J -K -C 5 --min-coverage 5 --max-coverage 500 -=`). Three variant callings were performed: one for each of the reference populations 'Lindley' and 'Ro1', and one joint calling of all *G. pallida* population datasets. Filtering, calculation of metrics, and subsetting of VCF files was performed using `bcftools v1.5` (`https://github.com/samtools/BCFtools`). Parameters for hard filtering of variants consisted in standard quality filters of variants: a minimum read depth of five reads, at least one read mapping on each strand and at least one read 'balanced' on each side of the variant site (`'DP>=5 & QUAL > 1 & QUAL/AO > 10 & FORMAT/GQ >= 10 & RPL >=1 & RPR>=1 & SAF>=1 & SAR>=1'`). Estimation of heterozygosity, *i. e.* the inbreeding coefficient $F_{is} = \frac{\text{SNPs}_{Obs,Hom} - \text{SNPs}_{Exp,Hom}}{\text{SNPs}_{All} - \text{SNPs}_{Exp,Hom}}$, based on biallelic SNPs and calculation of 'missingness' was carried out for each sample using `VCTtools v0.1.15` (Danecek et al., 2011).

5.2.10 Estimation of SNP frequency in reference populations

In order to assess SNP frequencies in the *G. pallida* 'Lindley' and *G. rostochiensis* 'Ro1' reference populations — which were used to construct the reference assemblies — the rates of variants per base were calculated based on the VCF files described in the previous section. Two sets of BED files were used to delimit regions based on which variant rates were calculated:

- 'ALL': all continuous regions of length ≥500 b with a read coverage ≥ five
- 'COV': all continuous regions of length ≥500 b with a read coverage ≥ five and a maximum read coverage of twice the median coverage of the dataset, *i. e.* 128 for 'Lindley' and 142 for 'Ro1'

The 'ALL' BED file delimits regions for which sufficient coverage for variant calling was observed, while the 'COV' BED files exclude regions for which excessive coverage was observed which might be due to paralogous genes or low-complexity/repeat regions. Variants in VCF files were subsetted by type — SNPs, Indels, and MNPs — using `bcftools v1.5` (`https://github.com/samtools/BCFtools`), and subsequently processed using `BedTools coverage` (option: `-counts`) and `GNU awk` to count biallelic variants for each type, in regions delimited by the 'ALL' and 'COV' BED files. The variant rates were calculated by dividing the number of variants of each type by the total length of the sampled region.

5.2.11 Phylogenetic analysis of SNP data

The hard filtered VCF file containing variants of all *G. pallida* datasets was processed to convert sequence IDs to numerical identifiers and sample names were shortened to one letter codes using `GNU sed`. Phylogenetic analysis was carried out on the resulting VCF file using `SNPhylo v20140701` (Lee et al., 2014) with 0.02 as minimum minor allele frequency, 0.1 as maximum percent of missing data, and setting the number of autosomal sequences to the number of scaffolds for which variants were called. The program's internal quality filtering removed 3,019,960 'low quality' sites and 1,613,734 biallelic and polymorphic SNPs were used. `SNPhylo` uses `SNPRelate` (Zheng et al., 2012) for linkage disequilibrium (LD) pruning of SNPs based on the pairwise genotypic correlation within sliding windows of 500,000 b to reduce the influence of clusters of SNPs on the phylogenetic analysis. Since little is known concerning patterns of LD across the genome of *G. pallida*, analyses were conducted for values for LD thresholds ranging from 0.1 to 0.9. `SNPhylo` converts the resulting SNPs to FASTA format, which are aligned using `MUSCLE v3.8.31` (Edgar, 2004). Phylogenetic trees were inferred using the maximum likelihood method implemented in `IQ-TREE v1.5.5` (Nguyen et al., 2015) using its automatic model selection (Kalyaanamoorthy et al., 2017) and performing 100 non-parametric bootstraps. Automatic model selection converged on the models 'TVM' (transversion model: variable base frequencies, variable transversion rates, transition rates equal), 'TVM+I' (transversion model with proportion of invariable sites), and 'GTR' (general time reversible model: variable base frequencies and symmetrical substitution matrix) depending on the dataset, based on bayesian information criterion.

5.2.12 Assessment of signatures of selection in coding regions

The hard filtered VCF file of all *G. pallida* datasets was filtered further to include only biallelic SNPs without any missing data. The resulting VCF file, the FASTA file of the *G. pallida* assembly and the GFF3 file of the annotations were partitioned into separate files for each scaffold that contained a variant in coding regions. The script `generate_popgenome_calls.py` (`https://github.com/DRL/thesis`) was used to generate R scripts which execute the McDonald-Kreitman test (MK-test) (McDonald and Kreitman, 1991) implemented in `PopGenome v2.2.4` (Pfeifer et al., 2014) on coding regions of genes for each scaffold and test significance using a Fisher's exact test. The South American 'P5A' population was selected as outgroup in the analysis based on the results described in Section 5.3.7. This approximate version of the MK-test allows computation of the neutrality-index (NI) based on SNPs in VCF files, but assumes that probability of co-occurrence of SNPs in the same codon is small and hence only examines codons with a single SNP. The null hypothesis in the MK-test is that the ratio of non-synonymous to synonymous substitutions between populations (fixed sites, F) is equal to the equivalent ratio within populations (polymorphic sites, P). Assuming that synonymous substitutions are neutral, departures from this equality are attributed to selection. NI ($\frac{P_n/P_s}{F_n/F_s}$) measures the direction and magnitude of the departure from neutrality. $NI > 1$ indicates an excess in non-synonymous polymorphic sites which is interpreted as a sign of negative selection. $NI < 1$ occurs if there is an excess of non-synonymous sites fixed between populations which suggests positive selection.

5.3 Results

5.3.1 Comparison of PCN assemblies

Standard genomic metrics for the three published PCN assemblies (Table 5.3.1) highlight the superiority of the assembly of *G. ellingtonae*. The use of long read (PacBio) data in conjunction with short read (Illumina HiSeq and MiSeq) data yielded a highly contiguous assembly with CEGMA (Parra, Bradnam, and Korf, 2007) completeness values similar to those of *G. rostochiensis*.

The assembly size of the *G. ellingtonae* assembly is in agreement with that of *G. rostochiensis*, suggesting that the 20 Mb of N's in the *G. pallida* genome might be an artefact of the assembly process. For the two PCN assemblies for which gene predictions are available, the cumulative span of genic regions is identical. However, the number of predicted genes in *G. pallida* is greater than in *G. rostochiensis*, suggesting a higher incidence of fragmented gene models. Furthermore, CEGMA metrics indicate a lower completeness concerning eukaryotic core genes in *G. pallida*. The differences in contiguity between the assemblies are visualised as cumulative length plots at the level of scaffolds, contigs and regions containing N's in Figure 5.3.1.

The high contiguity of the *G. ellingtonae* assembly is evident from the cumulative length curve at the level of scaffold. In contrast, the assemblies of *G. pallida* and *G. rostochiensis* contain high numbers of short scaffolds as a result of the short read assembly process. The cumulative length curve of regions containing N's hints not only at differences between *G. pallida* and the other two assemblies in the number of regions but also in the length distribution of these regions. The distributions for the three assemblies are visualised as histograms in Figure 5.3.2. The distribution for the *G. pallida* assembly shows a high number of continuous regions of N's

Table 5.3.1: Metrics of PCN assemblies. †: Values taken from Phillips et al. (2017). N/A: metrics not available.

Metric	*G. pallida*	*G. rostochiensis*	*G. ellingtonae*
Assembly size (Mb)	123.6	95.9	106.0
Scaffolds (n)	6873	4281	2246
Scaffold N50 (bp)	120,481	88,688	327,189
Longest scaffold (bp)	599,721	688,384	2,517,252
Contig N50 (bp)	11,611	11,372	13,178
Longest contig (bp)	93,564	111,501	173,609
Span of N's (bp)	19,976,929	4,399,212	899,007
GC (%)	36.7	38.1	36.7
CEGMA (Complete/Partial %)	74.19/80.65	96.4/95.56	92.3/96.3 (†)
Mean CEGMA (Complete/Partial)	1.23/1.29	1.15/1.24	N/A
Genes (n)	16,403	14,308	N/A
Span of genic regions (Mb)	39.57	39.57	N/A
Mean gene length (b)	2765.6	2412.3	N/A
Span of exonic regions (Mb)	17.76	18.21	N/A
Mean exon length (b)	135.1	145.5	N/A
Span of intronic regions (Mb)	21.81	21.36	N/A
Mean intron length (b)	189.6	192.7	N/A
Proteins (n)	16,403	14,309	N/A
Proteins w/ start and stop (n)	14,598	13,495	N/A

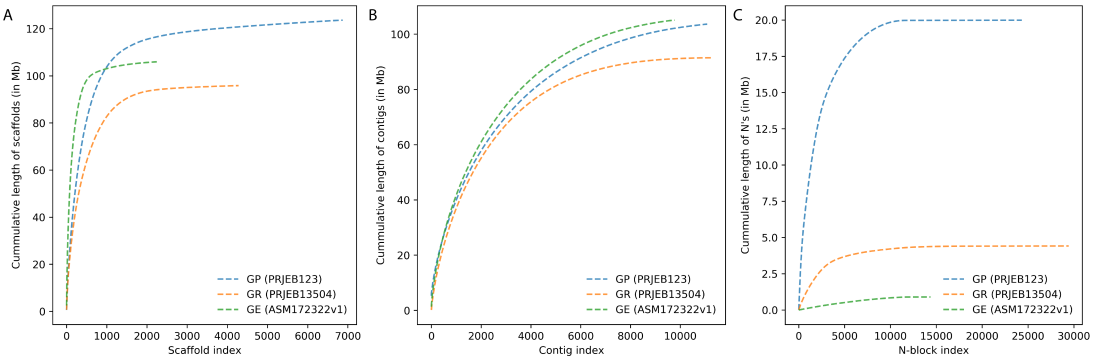


Figure 5.3.1: Cumulative length plots for PCN genomes. **A**: cumulative length plot for scaffolds. B: cumulative length plot of contigs. **C**: cumulative length plot of regions containing N's. GP: *G. pallida*. GR: *G. rostochiensis*. GE: *G. ellingtonae*. Differences between assemblies are most pronounced at the level of scaffold and regions containing N's.

ranging from 2500 to 15000 b, which are absent from the other assemblies. For the assembly of *G. pallida*, 454 PE libraries of insert sizes of 3 kb, 8 kb, and 20 kb were used (Cotton et al., 2014), which most likely are accountable for the high number of contiguous regions of N's.

Table 5.3.2: Span of PCN reference assemblies. 'AGCT': regions in the assemblies neither annotated as low-complexity/repeat regions nor containing ≥10 N's. N's: regions composed of ≥10 N's. dust: low-complexity regions annotated by dust masker. tandem: tandem repeats annotated by TRF. RM: Repeat regions annotated by RepeatMasker. Values for low complexity and repeat regions are not additive since regions were annotated independently and do overlap.

Assembly	Span						
	Total (b)	'AGCT' (%)	'agctn' (%)				
			Total	N's	dust	tandem	RM
G. pallida	123,625,196	55.81	44.19	16.16	28.06	5.19	19.36
G. rostochiensis	95,876,286	79.24	20.76	4.59	17.03	7.72	0.07

Based on the soft-masked assembly files and the annotations retrieved from

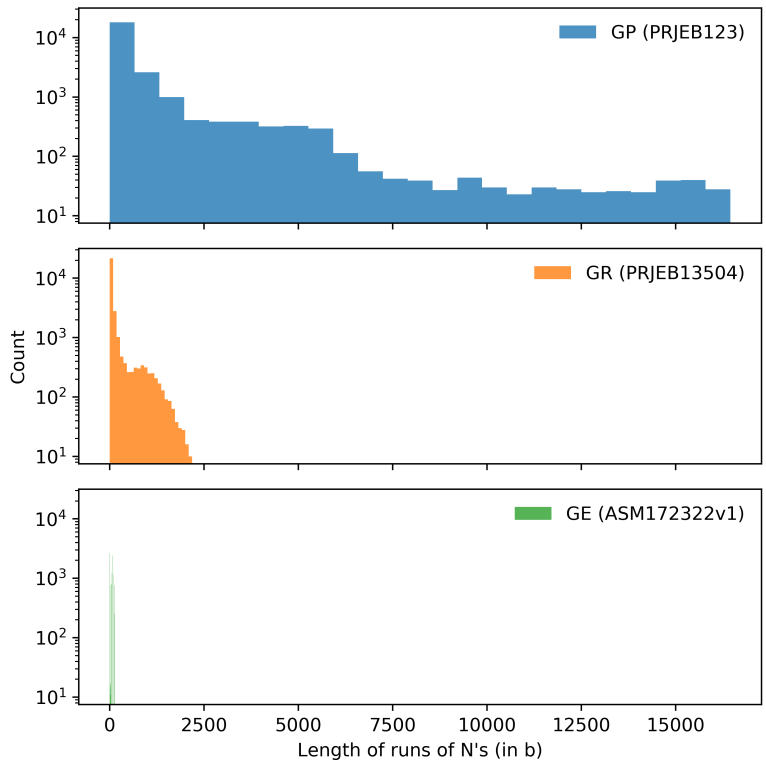


Figure 5.3.2: Histogram of length distribution of regions containing N's. Runs of N's were visualised in 25 bins for each assembly. GP: *G. pallida*. GR: *G. rostochiensis*. GE: *G. ellingtonae*. The *G. pallida* assembly displays more and longer stretches of unknown regions than the other two assemblies.

WBPS8, annotations concerning repeat and low-complexity regions were investigated. The length of regions composed of 'AGCT' and 'agctn' (*i. e.* low-complexity/repeat region or continuous regions of ≥10 N's) are listed in Table 5.3.2. The length of 'AGCT' regions in the *G. pallida* assembly is 69.0 Mb (55.81%) compared to 76.0 Mb (79.24%) of the *G. rostochiensis* assembly. The difference in proportion of total assembly length is due to the length of regions composed of N's (16.16%) in the *G. pallida* assembly, in addition to the number of regions annotated by the RepeatMasker algorithm. Due to the nature of the repeat finding algorithms, sites within coding regions can also been annotated as low-complexity/repeat regions. In *G. pallida*, 2.14 Mb (12.03%) of exonic regions overlap with annotations indicating low-complexity/repeat regions, compared to 0.66 Mb (3.62%) of exonic

regions in *G. rostochiensis*. The span for further analysis was estimated as the total number of bases in the assembly not containing runs of N's of length ten or longer, yielding 103.65 Mb for *G. pallida* and 91.48 Mb for *G. rostochiensis*.

5.3.2 Synteny between *G. pallida* and *G. rostochiensis*

Based on orthology inferred through an OrthoMCL clustering of proteins, 109 syntenic clusters of scaffolds were identified which contained at least five consecutive syntenic protein coding loci. In total, 38.2 Mb (36.9%) of the *G. pallida* assembly were partially syntenic to 31.1 Mb (34.0%) of the *G. rostochiensis* assembly (ignoring N's). Breaks in synteny between two scaffolds were observed in 20 pairs, seven of which involved inversions. The low proportion of syntenic regions most likely reflects the draft nature of both assemblies. However, high rates of intrachromosomal rearrangements have been observed for other nematodes, such as *C. elegans* and *C. briggsae* (Coghlan and Wolfe, 2002; Kent and Zahler, 2000). A subset of the largest syntenic cluster is shown in Figure 5.3.3. Synteny breakpoints which primarily co-occur with large insertions in the *G. pallida* assembly may suggest either large-scale genomic rearrangements or over-scaffolding of the *G. pallida* assembly.

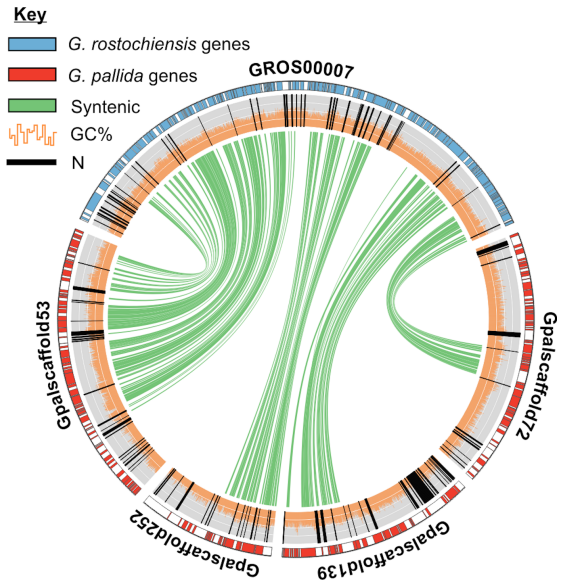


Figure 5.3.3: Conservation of synteny between *G. pallida* and *G. rostochiensis*. *G. rostochiensis* genes (blue) in scaffold 'GROS_00007' (500 kb) are syntenic (green arcs) with *G. pallida* genes (red) on four scaffolds. Synteny breakpoints primarily co-occur with large insertions in the *G. pallida* assembly. GC content and regions of undetermined sequence are represented by orange and black bars, respectively

5.3.3 Analysis of GC/AG splice sites

The result of the analysis of non-canonical GC/AG splice site frequencies in genomes of representative species of the phylum Nematoda, published in Eves-van den Akker et al. (2016b), is shown in Figure 5.3.4. These results indicate that the two genomes of PCN nematodes exhibit the highest proportion of non-canonical GC/AG splice sites among the surveyed species and that certain taxa do not display any gene with those splice sites, which might be an artefact of certain gene annotation pipelines.

In order to test whether the frequency of GC/AG splice sites in PCNs represent an extreme case within the phylum Nematode, I surveyed all 100 species of nematodes available on WBPS8. The percentage of GC/AG splice sites plotted against the N50 metric for each of the 100 nematode genomes is displayed in Figure 5.3.5. Across the nematode genomes surveyed, percentages ranged from 0% (in nine species including all three *Meloidogyne* species) to 8.9% in *H. bacteriophora*. However, an improved genome annotation of *H. bacteriophora* generated recently by Flo McLean only contains GC/AG splice sites in 0.80% of the introns, suggesting that this high rate is an artefact. A similar case is *H. polygyrus* for which alternative assemblies yield different rates of GC/AG splice sites (1.1 vs. 4.3%). Two *Angyostrongylus* species exhibit consistent and even higher proportions than those in *Globodera* species. In general, non-canonical GC/AG splice sites appear to be associated with mid-range N50s. The percentage of proteins unique to each genome (Figure 5.3.6) revealed further outlier proteomes, such as that of *Romanomermis culicivorax* (63.2% of proteins do not cluster with other species), but no clear connection exists between number of proteins unique to a proteome and GC/AG splice sites.

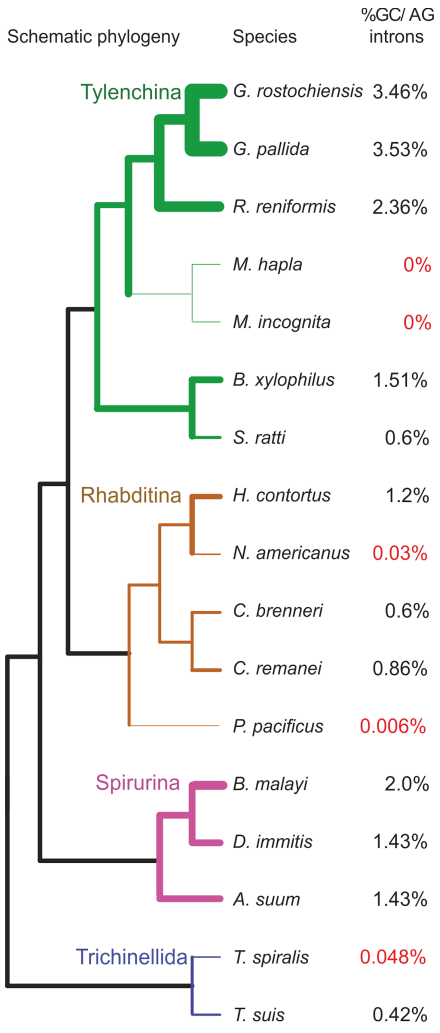


Figure 5.3.4: Percentages of GC/AG splice sites across selected nematode species Percentages of GC/AG splices sites with associated consensus sequences are shown for 17 species against a schematic phylogeny of the phylum Nematoda (adapted from Blaxter and Koutsovoulos, 2015). Thickness of branches is scaled by percentages of GC/AG splices sites. Red numbers indicate those which likely represent under reporting due to over-strict parameter settings during gene prediction.

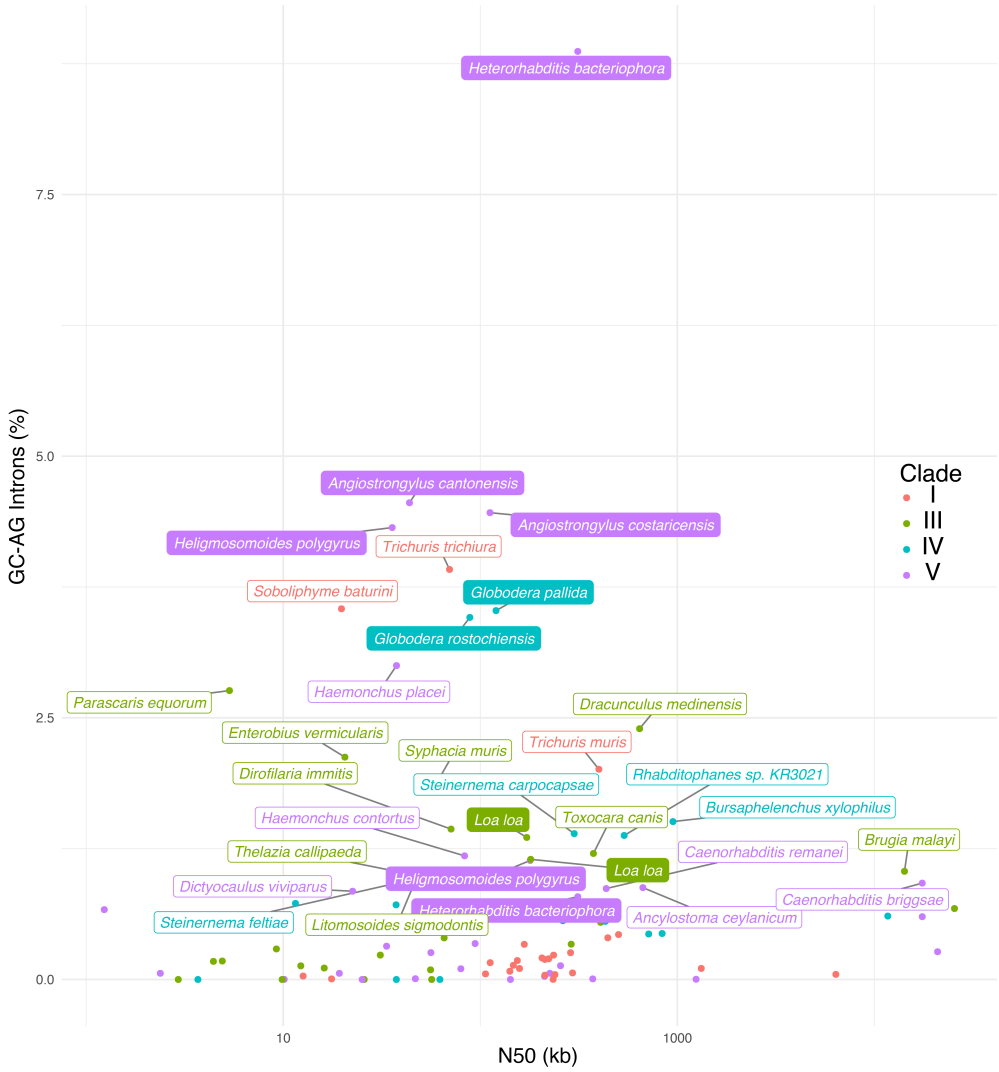


Figure 5.3.5: Percentages of GC/AG splice sites and N50 of nematode genomes. Genomes with a GC/AG splice site percentage ≥0.75% are labelled. Labels are filled for species for which either two alternative genomes/annotations exist (*Heterorhabditis bacteriophora*, *Heligmosomoides polygyrus*, *Loa loa*) or for which closely related species exhibiting high rates (*Globodera* species and *Angiostrongylus* species). High percentages of GC/AG splice site appear to be associated with mid-range N50s, and no clear phylogenetic pattern is apparent.

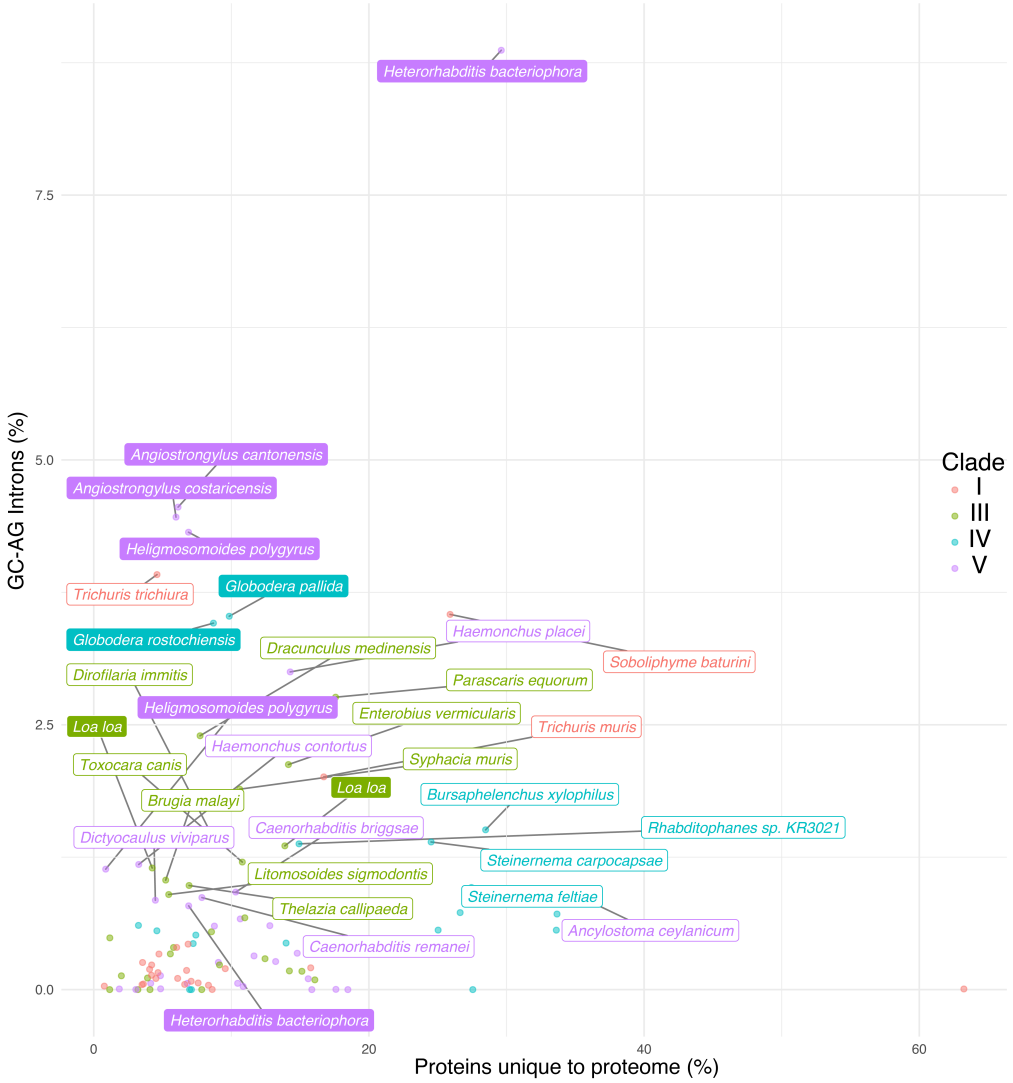


Figure 5.3.6: Percentages of GC/AG splice sites and unique proteins of nematode genomes. Percentage of unique proteins is the number of proteins in singleton and proteome-specific clusters in the KinFin analysis of the OrthoFinder clustering at MCL 3.0 ('RI') in Section 3.6.

In order to investigate the high proportion of GC/AG sites in the *Angiostrongylus* species, I analysed the counts of splice sites by gene in the species indicated by filled labels in Figure 5.3.5. These consist of two pairs of sister species with high and congruent GC/AG percentages (*Globodera* and *Angiostrongylus* species), one pair of splice site congruent annotations (*Loa loa*) and two pairs of alternative gene predictions which exhibit different numbers of GC/AG splice sites (*H. polygyrus* and *H. bacteriophora*). Histograms of the counts of GC/AG splice sites per gene of these taxa are shown in Figure 5.3.7. Counts per gene in *Globodera* and *Angiostrongylus* species exhibit similar distributions, as opposed to *Heligmosomoides* and *Heterorhabditis*, which could indicate that the high number of GC/AG splice sites in the gene annotations of *Angiostrongylus* species is a genuine feature of the underlying genomes. However, consistent and coordinated re-annotation of many nematode genomes is warranted, and parameters such as non-canonical splice-sites or proportion of unique proteins might be useful metrics for quality control of genome annotations.

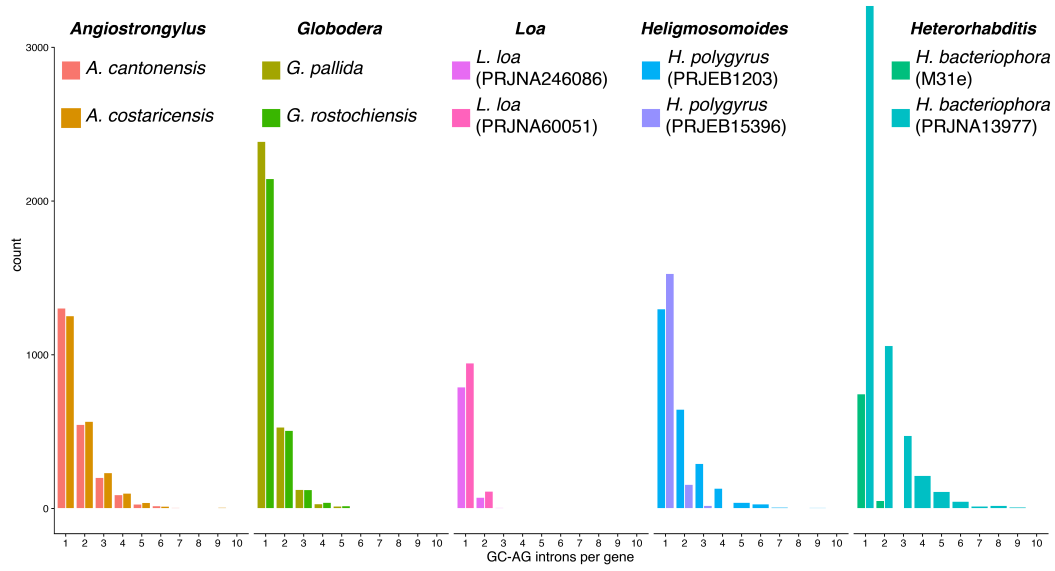


Figure 5.3.7: Distribution of GC/AG splice sites across genes. Distributions of counts of GC/AG splice sites per gene.*Angiostrongylus*, *Globodera*, and *Loa* species display similar distribution between the respective genomes, while alternative gene predictions *H. polygyrus* and *H. bacteriophora* differ.

5.3.4 SNP frequency in PCN reference populations

Assessment of coverage of both reference assemblies — by their respective population datasets — revealed distinct patterns of coverage. For the assembly of *G. pallida*, regions covered by five or more reads account for 92.27 Mb (87.64%) of the total span of the assembly, compared to 90.84 Mb (99.30%) for the assembly of *G. rostochiensis*. Of these, 84.76 Mb for *G. pallida* and 89.07 Mb for *G. rostochiensis* are composed of continuous regions of length ≥ 500 b ('ALL' regions). To account for repetitive regions, further filtering was performed based on a maximum read coverage of twice the median coverage for each dataset ('COV' regions). This reduced the span of continuous regions of length ≥ 500 b to 60.41 Mb in *G. pallida* and 82.12 Mb in *G. rostochiensis*. While the 'ALL' regions are almost identical in length between both species, filtering based on twice the median coverage removed a much longer span in the *G. pallida* assembly. This is consistent with the high proportion of bases in the assembly annotated as repeats/low-complexity regions listed in Table 5.3.2.

Variants were called on both PCN reference assemblies using the respective population datasets based on which they were assembled. Rates of variants for both types of regions — 'ALL' and 'COV' — were calculated and are listed in Table 5.3.3. *G. rostochiensis* displays lower rates of SNPs, MNPs, and indels than *G. pallida*, although a much greater proportion of the assembly is covered by reads. Hence it appears that the *G. pallida* assembly harbours a greater proportion of repeat/low-complexity regions and that — even if these are excluded — the numbers of variant sites are higher than in *G. rostochiensis*. Decrease and increase of the minimum length of continuous regions had little effect on the variant rate in both species, suggesting that these are stable estimates. The lower proportion of variant rates in *G. rostochiensis* are in agreement with the hypothesis that a single introductory event led to the establishment of the 'Ro1' population in Europe leading to a lower

genetic diversity due to a founder effect. Higher genetic diversity in *G. pallida* also explains challenges encountered during the assembly process. Cotton et al. (2014) reported that the read datasets of the 'Lindley' population used for the assembly of the reference genome displayed a high rate of polymorphisms — with at least 1.2% of sites being variable — which is roughly twice the estimated rate of this analysis. However, no details were given on how this estimate was calculated. The present analysis should be seen as a conservative estimate, as I tried to not overestimate the rate of variation due to neither the mapping nor the variant calling process, as only reads in proper pairs were used and variants were stringently quality filtered.

Table 5.3.3: Estimates of variant rates in PCN reference populations. Regions: type of regions on which rates of variants were estimated. COV: continuous regions (≥500 b) for which read coverage ranged between five and twice the median coverage of the dataset. ALL: all continuous regions (≥500 b) with read coverage ≥ five. Span: sampled length of the genome.

				1×10^{-3} Variants/b		
Species	Population	Regions	Span (Mb)	SNPs	MNPs	Indels
G. pallida	Lindley	COV	60.41	5.55	1.29	0.22
		ALL	84.76	6.69	1.54	0.26
G. rostochiensis	Ro1	COV	82.12	1.72	0.49	0.05
		ALL	89.07	1.90	0.52	0.06

5.3.5 Coverage in *G. pallida* population datatsets

Coverage of the *G. pallida* genome was assessed individually for each dataset as well as together. Coverage decay plots (Figure 5.3.8) were used to visualise the proportional coverage of coding regions in the genome for each read datasets at different read coverage thresholds. The UK and South America populations (Figure 5.3.8A) cover a higher proportion of coding regions at a higher read depth than the 'bottlenecked' populations (Figure 5.3.8B). This is to be expected since average read depth is smaller due to the limited amount of genomic material used for sequencing. Among the datasets in Figure 5.3.8A, the 'Lindley' population appears as an outlier. This dataset contains fewer reads than the other population read sets and the shape of the coverage decay curve is distinct from all the others exhibiting a steeper decrease at lower read coverages.

A possible explanation for this is that the 'Lindley' dataset was grown in culture as opposed to soil (Vivian Blok, 2015, *pers. comm.*) and contains much less contamination from non-nematode organisms than the other population datasets. The assembly and scaffolding process of the *G. pallida* genome was a complex procedure, due to the number of different sequencing technologies used — see supplementary information in Cotton et al., 2014 — and the inherent level of contaminant material associated with wild isolates. This could have lead to erroneous assemblies such as chimeric scaffolds composed of both non-nematode and nematode contigs. One example of this is a ribosomal operon sequence of the ascomycote *Fusarium* sp. integrated into scaffold 'pathogens_Gpal_scaffold_190' in the region 155,768 – 157,563). While this sequence in itself is non-coding, the fungal 18S sequence overlaps with the gene model 'GPLIN_000641700', a seven exon gene with no functional annotation. If more of these cases exist — and these contaminants are not part of the 'Lindley' dataset — a pattern like the one in Figure 5.3.8A would be observed. This sort of contamination due to chimeric sequences is hard to address

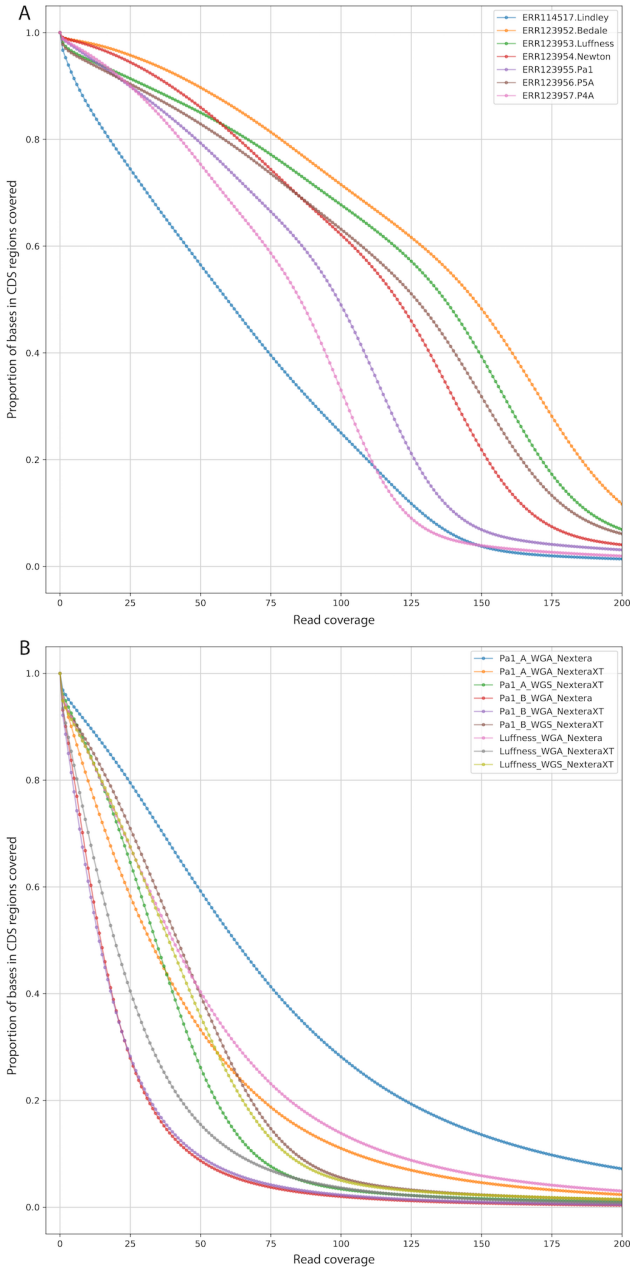


Figure 5.3.8: Coverage decay plots of population datasets for CDS regions. **A**: *G. pallida* populations from Table 5.1.1. **B**: 'bottlenecked' *G. pallida* populations sequenced as described in Section 5.2.5.

with currently available contamination screening tools. However, future versions of BlobTools will be able to address this problem, as discussed in Section 2.6.

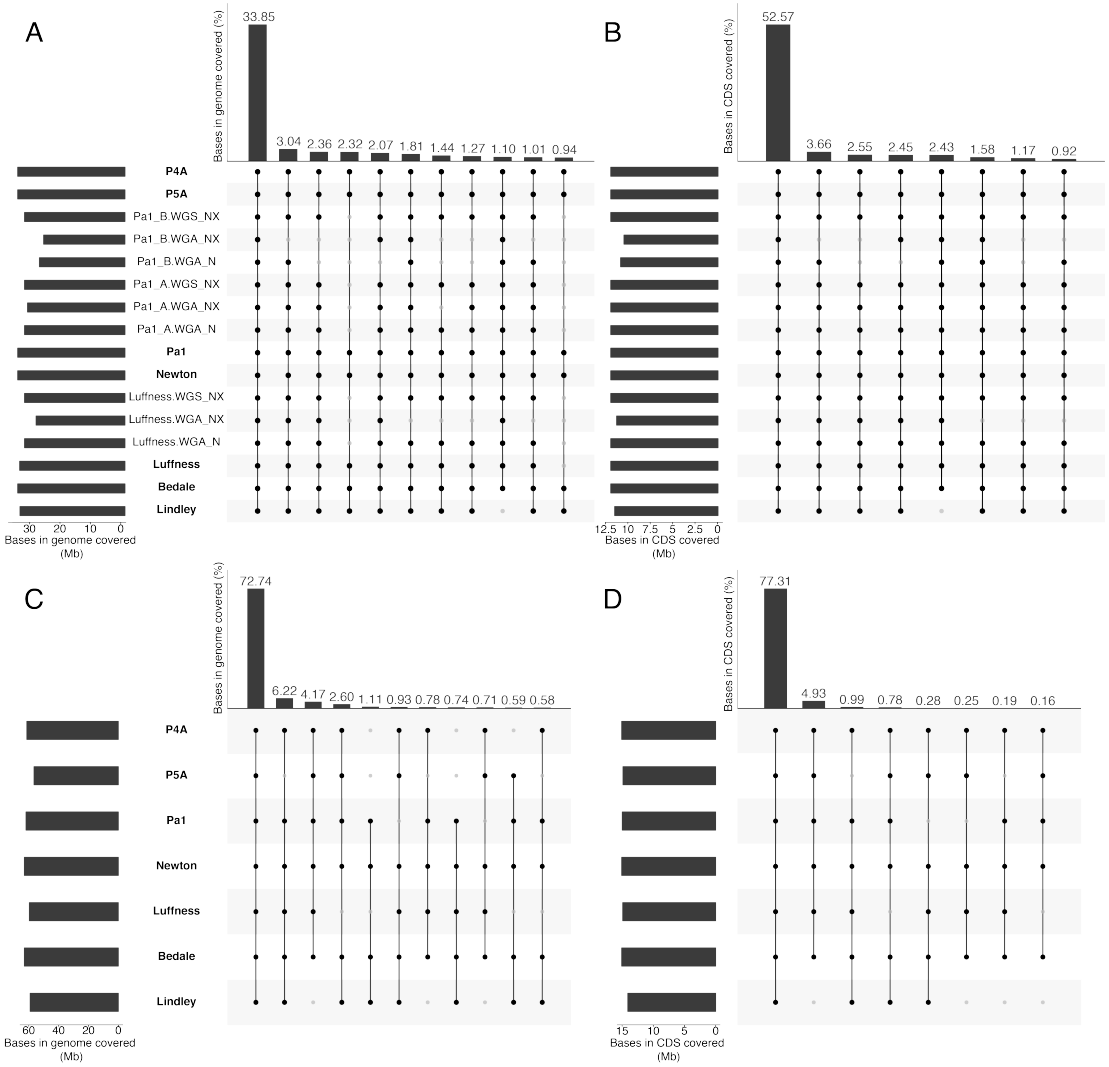


Figure 5.3.9: Coverage of *G. pallida* assembly regions by sets of read sets. **A** and **B**: Genomic (A) and coding (B) regions covered by sets all *G. pallida* read sets. **C** and **D**: Genomic (C) and coding (D) regions covered by sets of *G. pallida* population read sets from Table 5.1.1.

In order to assess how many bases in the genome of *G. pallida* are covered by all datasets at a read coverage of five or more, I visualised the proportion of genomic and coding regions covered by sets of *G. pallida* datasets. The result is depicted in Figure 5.3.9. Mandatory coverage of five or more reads in all datasets

reduces the genomic region covered to 33.85% (Figure 5.3.9A). Subsequent set-dropout is lead by read sets derived from the 'bottlenecked' populations followed by 'Lindley' and 'Luffness'. Of the bases in coding regions (Figure 5.3.9B), 52.57% are covered by all read datasets and dropout of 'Lindley' is observed in the fifth most frequent set, despite the fact that other 'bottleneck' populations display much lower general coverage. This is in agreement with the theory that non-nematode regions contribute to the lack of coverage from the relatively uncontaminated 'Lindley' dataset. Exclusion of 'bottlenecked' datasets recovers a greater proportion of covered span in both genomic (Figure 5.3.9C) and coding (Figure 5.3.9D) regions. Set-dropout for genomic regions is lead by the distantly related South American 'P5A' population followed by 'Lindley'. For coding regions, 'Lindley' is again the first population to leave the set, which could suggests that up to 5 Mb of coding regions in the current assembly might not be part of the true *G. pallida* genome.

5.3.6 Variation across *G. pallida* populations

After hard filtering of the VCF file the variant calling of all *G. pallida* samples yielded 4,633,694 variants. Of these 78.86% were SNPs and 22.67% indels. A high proportion of sites are multiallelic (15.90%), but only 2.27% are multiallelic SNPs. Hence, many variant sites are composed of multiple categories of variants, which inconveniences subsequent analyses which are based on biallelic variants and is likely to underestimate the variation in the populations. This is undoubtedly caused by the nature of the samples which are derived from pooled specimens of highly polymorphic organisms. Results of the estimation of heterozygosity as the inbreeding coefficient F_{is} for each of the samples are listed in Table 5.3.4. The highest values for F_{is} were observed for read datasets of the 'bottlenecked' sample 'Pa1_B', which is composed of females derived from the offspring of a single cyst. This suggests that even minor 'bottlenecking' of PCN can reduce the level of heterozygosity substantially, since the number of homozygous sites is lower for 'Pa1_A' which derives from the same population but contains 'cousins' as opposed to 'sisters'. The 'Pa1' sample exhibits much higher heterozygosity since it was generated by pooling hundreds of cysts. Analogously, read datasets based on the 'Luffness' sample derived from 'inbred' females displays lower heterozygosity than the pooled sample 'Luffness' from the same population. Unsurprisingly, highest values for heterozygosity were estimated in the South American population 'P4A'. However, the other South American population displayed only intermediate signs of outbreeding, exceeded by several UK populations, such as 'Newton', 'Bedale', and 'Pa1'. The 'Lindley' population appears to be largely in Hardy-Weinberg-Equilibrium as no major excess or depletion of homozygotes is observed. Although missing sites were excluded for the estimation of heterozygosity, it should be noted that a correlation ($r^2 = 0.6385$, $p = 1.24 * 10^{-4}$) exists between the fraction of missing genotypes for a sample in the VCF file and its inbreeding coefficient F_{is} (see Figure 5.3.10). The population 'P4A' is the only population which deviates from

this trend as it display a low number of missing sites in the VCF file but appears to be highly heterozygous. Since correlation does not necessarily imply causation, the pattern could simply be an artefact caused by the inherent lower coverage of the samples derived from 'bottlenecked' populations which are based on less input DNA.

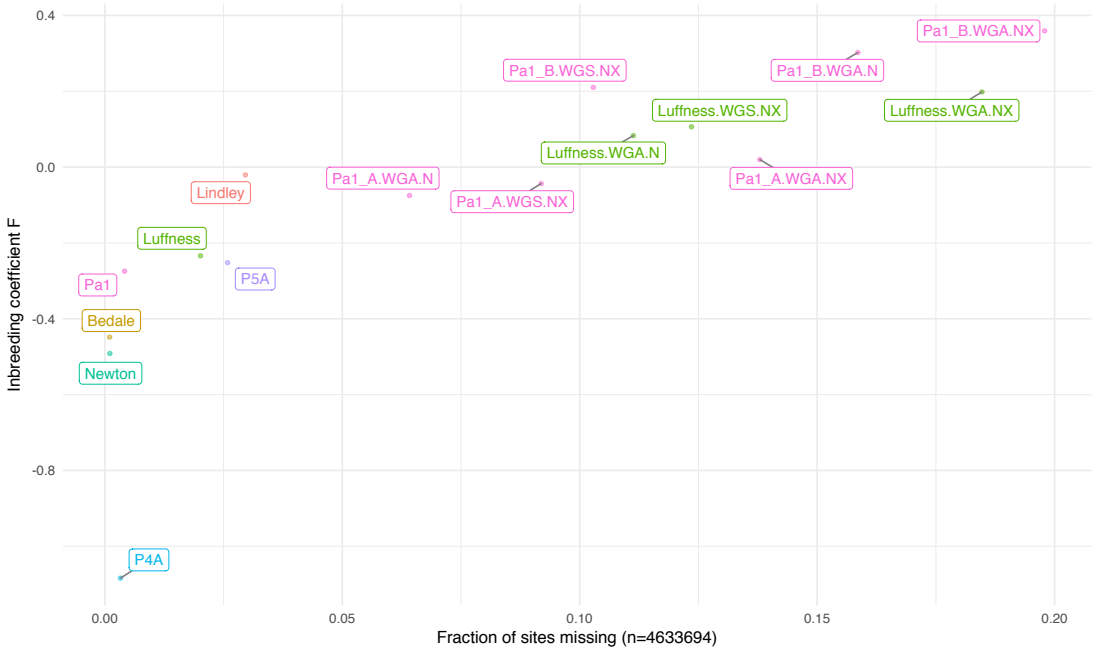


Figure 5.3.10: Correlation between results of heterozygosity estimates and missing data

Table 5.3.4: Heterozygosity of *G. pallida* datasets. Heterozygosity, *i. e.* the inbreeding coefficient F_{is}, was calculated based on 2,486,754 biallelic SNPs and ignoring missing sites. Expected number of homozygous sites ($E(HOM)$) was estimated as 1,749,352.1. $O(HOM)$: observed number of homozygous sites. F: inbreeding coefficient (positive values in bold)

Dataset	Population	$O(HOM)$	F
Lindley	Lindley	1,734,103	-0.02068
P4A	P4A	950,046	-1.08395
P5A	P5A	1,563,462	-0.25209
Newton	Newton	1,386,896	-0.49153
Bedale	Bedale	1,418,699	-0.44840
Luffness	Luffness	1,576,756	-0.23406
Luffness.WGS.NX	Luffness	1,827,750	**0.10632**
Luffness.WGA.N	Luffness	1,810,571	**0.08302**
Luffness.WGA.NX	Luffness	1,895,438	**0.19811**
Pa1	Pa1	1,546,994	-0.27442
Pa1_A.WGS.NX	Pa1	1,717,305	-0.04346
Pa1_A.WGA.N	Pa1	1,694,161	-0.07485
Pa1_A.WGA.NX	Pa1	1,763,969	**0.01982**
Pa1_B.WGS.NX	Pa1	1,904,404	**0.21027**
Pa1_B.WGA.N	Pa1	1,972,048	**0.30200**
Pa1_B.WGA.NX	Pa1	2,014,214	**0.35918**

5.3.7 Phylogenetic analysis of *G. pallida* populations

The phylogenetic analysis of biallelic SNPs in all *G. pallida* populations was based on output of `SNPhylo` across nine different LD-thresholds used for SNP pruning. The number of resulting sites used in tree construction are listed in Table 5.3.5.

Phylogenetic analyses yielded three distinct tree topologies (Figure 5.3.11A, B, and C), depending on the linkage disequilibrium (LD) threshold used for SNP pruning. Exclusion of both South American populations from the dataset recovered a stable tree topology across all LD-thresholds (Figure 5.3.11D), indicating that the signal contained in SNP data separates European populations reliably. Exclusion of either 'P4A' or 'P5A' also resulted in stable tree topologies across the sampled parameters space, suggesting that presence of the two South American populations in combination with the LD-pruning algorithm for SNP in `SNPrelate` yielded subsets of SNPs which indicate different evolutionary relationships. The number of biallelic SNP sites in the alignments used for inferring trees in Figure 5.3.11D, E and F were 33,178, 37,442, and 39,243, and the number of phylogenetically informative sites were 1491, 2398, and 1592, respectively.

Taken together, the results of the phylogenetic analyses suggest, that the 'Luffness' and 'Pa1' populations form distinct monophyletic clades, as do the three English populations 'Lindley', 'Bedale' and 'Newton' (Figure 5.3.11D). This is in agreement with previous phylogenetic studies (Blok and Phillips, 1995; Blok, Phillips, and Harrower, 1997; Blok et al., 1998; Subbotin et al., 2000). Inclusion of both South American populations in the dataset interferes with the LD-pruning algorithm and results in biallelic sites which yield contradicting tree topologies (Figure 5.3.11A, B, and C). This can be caused by several factors such as the high rate of heterozygosity observed for 'P4A' or the high amount of multiallelic non-SNP variation observed in the dataset. Further analysis of the patterns of multiallelic

Table 5.3.5: Phylogenetic analysis based on SNPs in *G. pallida* populations. Sites: sites in the alignment created by SNPhylo. All/Informative/Constant: classification of sites in alignment by `IQ-TREE`. Phylogenetic model: model inferred by `IQ-TREE`. Topology: Topology of the resulting trees as show in Figure 5.3.11.

	Sites				
LD-threshold	All	Informative	Constant	Phylogenetic model	Topology
0.1	6970	852	4998	TVM	A
0.2	8901	835	6878	GTR	A
0.3	12,209	905	9970	TVM	A
0.4	17,026	1127	14,307	TVM	A
0.5	26,379	1625	22,110	TVM+I	B
0.6	43,874	2421	35,259	TVM	C
0.7	72,560	4693	55,979	TVM	A
0.8	130,176	12,910	93,854	GTR	B
0.9	232,628	36,126	154,240	TVM	A

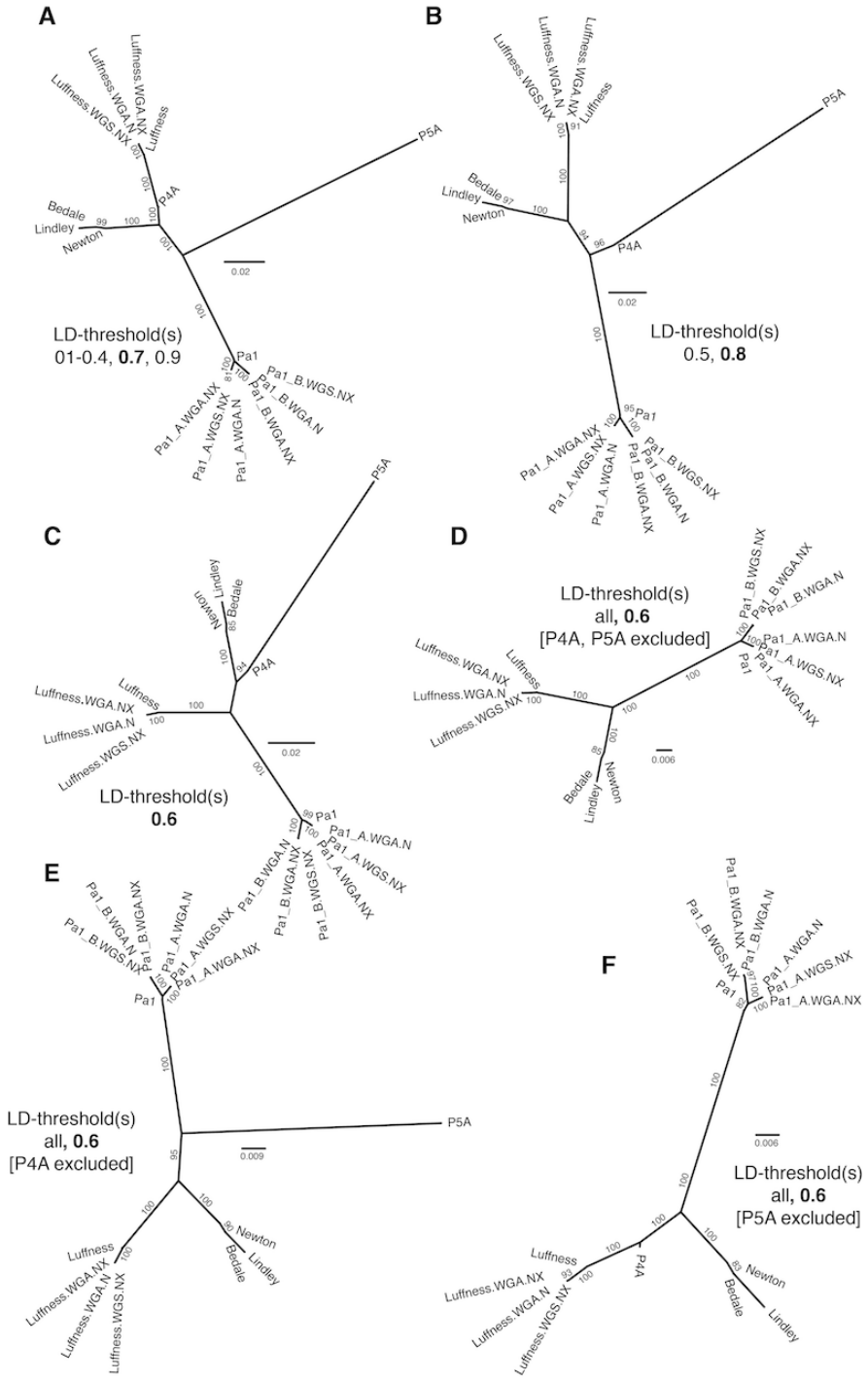


Figure 5.3.11: Phylogenetic trees based on biallelic SNP data. **A–C**: Tree topologies listed in Table 5.3.5. **D–F**: Tree topologies recovered when excluding samples from the datasets. Annotations indicate the range of LD thresholds for which a given topology was recovered (LD-threshold of the depicted tree in bold). Modifications to taxon composition of the analysed SNP data are shown in squared brackets. Non-parametric bootstrap support are indicated on branches.

variation in the dataset could shed more light on this issue but are complex to carry out. Separate analysis of the European populations together with either one of the South American populations, recovers 'P5A' as being distantly related to the European populations (Figure 5.3.11E) and suggest that 'P4A' might have an ancestral relationship with 'Luffness'. The English populations 'Lindley', 'Newton' and 'Bedale', are more closely related to 'Luffness' than to 'Pa1'. This could be caused by a separate introductory event to Europe of the 'Pa1' population through a South American population not included in this dataset. The other European populations could be the result of an introduction of a 'P4A'-like ancestor. Phylogenetic studies based on *Cytb* sequences recovered the same pattern and suggested that a P4A-like or a P4A-hybrid population gave rise to the UK populations closely related to 'Luffness' (Pylypenko, Phillips, and Blok, 2008; Madani et al., 2010; Hoolahan et al., 2012). The conflicting phylogenetic pattern which arises when both South American populations are included in the analysis, might be the result of 'P4A' being a hybrid of an unknown species, closer related to to 'P5A', and the ancestor of the European population 'Luffness'. Testing of this hypothesis would require assemblies of the populations which is not achievable with the available sequencing data due to the levels of heterozygosity. It should be noted that LD pruning — as implemented in the `SNPrelate` package used by `SNPhylo` — is performed for each scaffold individually. In fragmented assemblies, such as the one of *G. pallida*, this can lead to under-estimation of linkage between SNPs, adding an additional caveat to this analysis.

5.3.8 Signatures of selection in coding regions

Of the 2930 scaffolds containing biallelic SNPs in coding regions after quality filtering, only 2222 scaffolds contained SNPs in coding regions with no missing genotypes for which `PopGenome` could calculate the MK-test statistic. The MK-test was applied to the 121,284 CDS regions (92.26% of all CDS regions) on these scaffolds. However, the MK-test failed to produce results for 98.83% of the CDS regions. Of these, CDS regions without any SNP account for 24.20%, while 74.62% are due to lack of fixed sites synonymous or non-synonymous between 'P5A' and the other populations. For 1423 CDS regions `PopGenome` generated a result for the MK-test, but 718 of these yield values for NI approaching ∞, due to lack of fixed sites between the two sets of populations or fixed sites within the ingroup. The 705 CDSs for which the MK-test could be calculated successfully are located on 375 scaffolds and are part of 599 genes. These included seven of the effectors identified through RBBH analysis in Chapter 4 and all display NI values of zero, indicating neutrality. *G. pallida* proteins recovered from synapomorphic clusters that contain at least one of the effector proteins identified through RBBH analysis, yielded one protein from a synapomorphic cluster in all Clade IV nematodes (a NUDIX hydrolase containing a SPRY domain, NI=0) and 19 proteins from clusters synapomorphic to *Globodera* species for which the the MK-test returned a result for at least one of the underlying coding sequences. However, none of these received a significant p-value (Fisher's-Exact test, p-value<0.05). In the whole analysis, only eight CDS regions received a significant p-value, with seven displaying NI values of zero indicating neutral evolution, which belong to the genes 'GPLIN_000313300', 'GPLIN_000362400', 'GPLIN_000727500', 'GPLIN_000917100', 'GPLIN_000971800', 'GPLIN_001104400', and 'GPLIN_00139 5700'. One CDS in the gene 'GPLIN_000284900' displayed a NI value of 0.30, located on scaffold 'pathogens_Gpal_scaffold_57'. It is located on a contig with no other gene surrounding it, contains no introns, and its protein sequence is annotated

with three 'Zinc finger, CCHC-type' domains, one 'aspartic peptidase' domain and one 'domain of unknown function' (DUF1759). A BLAST search of its 1062 amino acid protein sequence against NCBI reveals 27% identity to a Pao retrotransposon in the genome of *Brugia malayi*.

5.4 Conclusion

The results presented in this chapter highlight some of the problems which persist in the genomic resources available to date for potato cyst nematodes, such as possible remaining contaminants in the reference genome of *G. pallida* revealed by coverage analysis, issues with the amount of unknown and repeat regions in the assembly, and extreme and complex variation within the sequencing datasets of populations of *G. pallida*. Datasets based on 'bottlenecked' samples — in combination with WGA approaches and/or low input sequencing library preparation methods — display lower levels of heterozygosity and may serve as a viable technique for the study of variation of *G. pallida* populations. However, the low coverage achieved in the presented datasets and the lack of replication prevented a formal assessment of the success of this approach. Other promising approaches for the study of *G. pallida* populations are the use of GBS datasets (Mimee et al., 2015) or the selective capture of target regions of the genome, which are carried out at the James Hutton Institute by John Jones and Vivian Blok (John Jones, 2016, *pers. comm.*). The latter approach consists in designing capture probes based on genic regions which are used to enrich the proportion of these regions in the sequencing data. It should however be noted that due to the issues I highlighted here — concerning chimeric sequences and the high proportion of repetitive/low-complexity regions in the *G. pallida* assembly — care should be taken to only include validated gene models in the probe set, since repeat elements and non-*Globodera* regions could limit the success of this approach.

Comparison of the genome assemblies of PCNs revealed the superior contiguity of the recently published draft genome of *G. ellingtonae* (Phillips et al., 2017). Once gene annotations are available, these could be exploited to expand on the study the evolution of gene families across plant parasitic nematodes and serve as a basis for further analysis on the level of synteny between PCN genomes. Analysis of

the repeat/low-complexity and unknown regions in the genome assemblies of *G. pallida* and *G. rostochiensis* revealed differences in the structure of these assemblies and are likely to be a result of the assembly process of the *G. pallida* genome. Estimation of rates of variation in the reference populations 'Lindley' and 'Ro1', revealed higher rates for *G. pallida* which is in agreement with the theory that the *G. rostochiensis* 'Ro1' population is the result of a single introductory event causing a founder effect.

Analysis of splice sites revealed possible problems with several datasets currently available on WormBase parasite. The metrics presented here might serve as novel benchmarks for the assessment of quality of gene predictions in draft genomes. Re-annotation of several nematode genomes is warranted and efforts should be coordinated and assessed using comparative genomics approaches.

Phylogenetic analysis of biallelic SNPs in the datasets of *G. pallida* populations revealed robust patterns for the European populations which support hypotheses formulated in previous phylogeographic studies using smaller number of loci. The South American populations 'P4A' and 'P5A' display varying phylogeographic patterns depending on the parameters used. When analysed separately, 'P5A' is more distantly related to the European populations than 'P4A'. The patterns observed for 'P4A' are suggestive of a hybrid origin of this population, which has already been proposed by other researchers (Pylypenko, Phillips, and Blok, 2008; Madani et al., 2010; Hoolahan et al., 2012).

Analysis of signatures of selection on coding regions in the genome of *G. pallida* recovered few regions for which an assessment could be made. The eight CDS regions that received a significant *p*-value in the MK-test displayed no sign of positive or negative selection and it can be concluded that due to the nature of the populations from which datasets have been generated, assessments regarding selective processes on coding regions of the genome of *G. pallida* can not be inferred.

Alternatives to this approach, such as the generation of FASTA files from SNP data is inconvenienced by the same problems that impede success of approaches based on variant data alone. Furthermore, as far as I am aware, assumptions made by available software for the analysis of variation data are violated by the complex structure of variation in the highly heterozygous populations of *G. pallida*. Future use of long read sequencing data for the generation of PCN reference genomes, paired with short read data from 'bottlenecked' populations will reveal phylogeographic patterns and shed light on the selective processes acting on the coding regions of these organisms.

Chapter 6

Outlook

"Nothing is built on stone;
All is built on sand, but we must build as if the sand were stone."

- Jorge Luis Borges, *In Praise of Darkness, (1974)*

In this thesis, I presented two software solutions that I developed in order to address two common challenges associated with genomics of non-model organisms: taxonomic interrogation of genome assemblies and custom taxon analysis of clustered protein data for the purpose of protein family analysis. I showed examples of their functionality — based on several use cases involving a wide range of taxonomic groups of organisms — thereby illustrating their general suitability for the field of genomics.

BlobTools, described in Chapter 2, has been well received by the research community and several publications by collaborators, other researchers, and myself, proved that it is a valuable addition to the bioinformatic toolbox. Shortcomings

of the software solution — such as its current inability to distinguish chimeric sequences from *bona fide* contamination — have been pointed out and planned improvements to the software are discussed in Section 2.6.

KinFin, discussed in Chapter 3 and applied to potato cyst nematodes and other clade IV nematodes in Chapter 4, is a software solution which integrates structural and functional genomic data of organisms and allows the user to analyse these high-dimensional data in the light of evolutionary, ecological and taxonomic hypotheses. The use cases presented in Chapter 3 involved analyses of groups of taxa of medical, veterinary and evolutionary interest and I was able to replicate previously reported findings by other researchers as well as to formulate new hypotheses regarding timing of horizontal gene transfer and to identify proteins of interest based on evolutionary patterns.

Through the analysis of publicly available genomic data from multiple organisms and sources with both BlobTools and KinFin, I was able identify problems concerning taxonomic composition and structural gene predictions in several genomes. One example is the number of single-copy orthologues encountered in different analyses of KinFin. The analysis of ecdysozoan protein families (Section 3.5) — which included 28 species from five phyla — yielded 21 'true' single-copy orthologues. In comparison, the analysis of protein families of Clade IV nematodes in Chapter 4 returned 28 single-copy orthologues, and only after exploration of clustering parameter space. Hence, either Clade IV nematodes have a higher turnover of genes than the taxa across five phyla or, and this is more likely, quality of gene predictions in Clade IV nematodes is suboptimal due to fragmented assemblies, contamination, and uncollapsed haplotypes in the case of polyploid taxa in the genus *Meloidogyne*. I think it lies within the obligations of the genomics community to develop standardised infrastructure and procedures to limit the influx of questionable data into public databases. The BlobToolKit grant awarded to Mark Blaxter is a first

step into this direction. However, since all of biology is inherently interconnected — as all organisms are related through a common ancestor — we have also to start thinking about methods to validate existing data in the public databases, since misannotated sequences have the potential to propagate false conclusions across many areas of research. I have highlighted some simple metrics which can be used to assess gene predictions in genome assemblies, but coordinated efforts by the research community to address this problem are needed urgently.

Analysis of the evolutionary patterns of protein families of potato cyst nematodes (described in Chapter 4) was focussed on effector proteins based on sequences in the literature. Previously reported findings could be replicated and two cases of putative horizontal gene transfers from bacteria to *Globodera* species were investigated which lead to formulation of hypotheses about the time point of their acquisition. A set of effector proteins was compiled for PCNs based on orthology to effector sequences published in the literature. Future analyses of the patterns of protein family evolution in PCNs could be improved significantly through genomes data from other Heteroderidae such as the soybean cyst nematode, *Heterodera glycines*, or the sugarbeet nematode, *H. schachtii*.

In Chapter 5, I analysed published genome assemblies of PCNs and assessed their quality. Through the use of genomic data for different species and populations of PCNs, I explored patterns of variation within their genomes and estimated rates of variation for the reference genomes of *G. pallida* and *G. rostochiensis*. I investigated phylogeographic patterns of populations of *G. pallida* which agree with previously published results. Analysis of patterns of selection across coding regions within the *G. pallida* species complex highlighted currently unsolved problems due to the level and nature of variation observed in the different populations. I presented a possible solution for the amelioration of this effect through the application of whole genome amplification approaches applied to 'bottlenecked' samples.

In summary, this thesis provides novel solutions to common challenges in the field of comparative genomics and — by applying them to the study evolutionary patterns in effector gene families in potato cyst nematodes — identifies current obstacles in the analysis of highly complex populations of plant parasitic nematodes.

Bibliography

Abad, Pierre, Jerome Gouzy, Jean-Marc Aury, et al. (2008). "Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*". In: *Nature Biotechnology* 26.8, pp. 909–915.

Adams, Mark D, Susan E Celniker, Robert A Holt, et al. (2000). "The genome sequence of *Drosophila melanogaster*". In: *Science* 287.5461, pp. 2185–2195.

Alexander, Warren S (2002). "Suppressors of cytokine signalling (SOCS) in the immune system". In: *Nature Reviews Immunology* 2.6, pp. 410–416.

Ali, Shawkat, Maxime Magne, Shiyan Chen, et al. (2015). "Analysis of putative apoplastic effectors from the nematode, *Globodera rostochiensis*, and identification of an expansin-like protein that can induce and suppress host defenses". In: *PLoS One* 10.1.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, et al. (2014). "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* 11.11, pp. 1144–1146.

Altenhoff, Adrian M and Christophe Dessimoz (2009). "Phylogenetic and functional assessment of orthologs inference projects and methods". In: *PLoS Computational Biology* 5.1.

Altenhoff, Adrian M, Romain A Studer, Marc Robinson-Rechavi, et al. (2012). "Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs". In: *PLoS Computational Biology* 8.5.

Altenhoff, Adrian M, Brigitte Boeckmann, Salvador Capella-Gutierrez, et al. (2016). "Standardized benchmarking in the quest for orthologs". In: *Nature Methods* 13.5, pp. 425–430.

Andrássy, I (1976). *Evolution as a basis for the systematization of nematodes*. Pitman Publishing Limited.

Antonicka, Hana, Andre Mattman, Christopher G Carlson, et al. (2003). "Mutations in COX15 produce a defect in the mitochondrial heme biosynthetic pathway, causing early-onset fatal hypertrophic cardiomyopathy". In: *American Journal of Human Genetics* 72.1, pp. 101–114.

Arakawa, Kazuharu (2016). "No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade". In: *PNAS* 113.22, E3057.

Arakawa, Kazuharu, Yuki Yoshida, and Masaru Tomita (2016). "Genome sequencing of a single tardigrade *Hypsibius dujardini* individual". In: *Scientific Data* 3.

Artamonova, Irena I and Arcady R Mushegian (2013). "Genome sequence analysis indicates that the model eukaryote *Nematostella vectensis* harbors bacterial consorts". In: *Applied and Environmental Microbiology* 79.22, pp. 6868–6873.

Bakhetia, Manjula, Wayne Charlton, Howard J Atkinson, et al. (2005). "RNA interference of dual oxidase in the plant nematode *Meloidogyne incognita*". In: *Molecular Plant-Microbe Interactions* 18.10, pp. 1099–1106.

Baldwin, James G, Steven A Nadler, and Byron J Adams (2004). "Evolution of plant parasitism among nematodes". In: *Annual Review of Phytopathology* 42, pp. 83–105.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). "Gephi: an open source software for exploring and manipulating networks". In: *International AAAI Conference on Weblogs and Social Media*.

Bekal, Sadia, Terry L Niblack, and Kris N Lambert (2003). "A chorismate mutase from the soybean cyst nematode *Heterodera glycines* shows polymorphisms that correlate with virulence". In: *Molecular Plant-Microbe Interactions* 16.5, pp. 439–446.

Bellafiore, Stéphane, Zhouxin Shen, Marie-Noelle Rosso, et al. (2008). "Direct identification of the *Meloidogyne incognita* secretome reveals proteins with host cell reprogramming potential". In: *PLoS Pathogens* 4.10.

Bemm, Felix, Clemens L Weiß, Jörg Schultz, et al. (2016). "Genome of a tardigrade: Horizontal gene transfer or bacterial contamination?" In: *PNAS* 113.22.

Bennett, Hayley M, Hoi Ping Mok, Effrossyni Gkrania-Klotsas, et al. (2014). "The genome of the sparganosis tapeworm *Spirometra erinaceieuropaei* isolated from the biopsy of a migrating brain lesion". In: *Genome Biology* 15.11, p. 510.

Berriman, Matthew, Brian J Haas, Philip T LoVerde, et al. (2009). "The genome of the blood fluke *Schistosoma mansoni*". In: *Nature* 460.7253, pp. 352–358.

Bert, Wim, Gerrit Karssen, and Johannes Helder (2011). "Phylogeny and evolution of nematodes". In: *Genomics and molecular genetics of plant-nematode interactions*. Springer, pp. 45–59.

Bert, Wim, Frederik Leliaert, Andy R Vierstraete, et al. (2008). "Molecular phylogeny of the Tylenchina and evolution of the female gonoduct (Nematoda: Rhabditida)". In: *Molecular Phylogenetics and Evolution* 48.2, pp. 728–744.

Bird, David McK, John T Jones, Charles H Opperman, et al. (2015). "Signatures of adaptation to plant parasitism in nematode genomes". In: *Parasitology* 142, pp. 1–14.

Blaxter, Mark L and Georgios Koutsovoulos (2015). "The evolution of parasitism in Nematoda". In: *Parasitology* 142 Suppl 1, S26–39.

Blaxter, Mark L, Paul De Ley, James R Garey, et al. (1998). "A molecular evolutionary framework for the phylum Nematoda". In: *Nature* 392.6671, pp. 71–75.

Blok, Vivian C and Mark S Phillips (1995). “The use of repeat sequence primers for investigating genetic diversity between populations of potato cyst nematodes with differing virulence”. In: *Fundamental and Applied Nematology* 18.6, pp. 575–582.

Blok, Vivian C, Mark S Phillips, and Brian E Harrower (1997). “Comparison of British populations of potato cyst nematodes with populations from continental Europe and South America using RAPDs”. In: *Genome* 40.3, pp. 286–293.

Blok, Vivian C, Gaynor Malloch, Brain Harrower, et al. (1998). “Intraspecific variation in ribosomal DNA in populations of the potato cyst nematode *Globodera pallida*”. In: *Journal of Nematology* 30.2, pp. 262–274.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel (2014). “Trimmomatic: a flexible trimmer for Illumina sequence data.” In: *Bioinformatics* 30.15, pp. 2114–2120.

Boothby, Thomas C and Bob Goldstein (2016). “Reply to Bemm et al. and Arakawa: Identifying foreign genes in independent *Hypsibius dujardini* genome assemblies”. In: *PNAS* 113.22, E3058–3061.

Boothby, Thomas C, Jennifer R Tenlen, Frank W Smith, et al. (2015). “Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade”. In: *PNAS* 112.52, pp. 15976–15981.

— (2016). “Correction for Boothby et al., Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade”. In: *PNAS* 113.36, E5364.

Bork, Peer, Thomas Dandekar, Yolande Diaz-Lazcoz, et al. (1998). “Predicting function: from genes to genomes and back”. In: *Journal of Molecular Biology* 283.4, pp. 707–725.

Borner, Janus, Peter Rehm, Ralph O Schill, et al. (2014). “A transcriptome approach to ecdysozoan phylogeny”. In: *Molecular Phylogenetics and Evolution* 80, pp. 79–87.

Boschetti, Chiara, Adrian Carr, Alastair Crisp, et al. (2012). “Biochemical diversification through foreign gene expression in bdelloid rotifers”. In: *PLoS Genetics* 8.11.

Bridge, Paul D, Peter J Roberts, Brian M Spooner, et al. (2003). “On the unreliability of published DNA sequences”. In: *New Phytologist* 160.1, pp. 43–48.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson (2015). “Fast and sensitive protein alignment using DIAMOND”. In: *Nature Methods* 12.1, pp. 59–60.

Camacho, Christiam, George Coulouris, Vahram Avagyan, et al. (2009). “BLAST+: architecture and applications”. In: *BMC Bioinformatics* 10, p. 421.

Campbell, Lahcen I, Omar Rota-Stabelli, Gregory D Edgecombe, et al. (2011). “MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda”. In: *PNAS* 108.38, pp. 15920–15924.

Capella-Gutiérrez, Salvador, José M Silla-Martínez, and Toni Gabaldón (2009). “trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.” In: *Bioinformatics* 25.15, pp. 1972–1973.

Casaravilla, Cecilia, Charles Brearley, Silvia Soulé, et al. (2006). "Characterization of myo-inositol hexakisphosphate deposits from larval *Echinococcus granulosus*". In: *FEBS Journal* 273.14, pp. 3192–3203.

C. elegans Sequencing Consortium (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology". In: *Science* 282.5396, pp. 2012–2018.

Charlier, Johannes, Mariska van der Voort, Fiona Kenyon, et al. (2014). "Chasing helminths and their economic impact on farmed ruminants". In: *Trends in Parasitology* 30.7, pp. 361–367.

Chaudhari, Narendrakumar M, Vinod Kumar Gupta, and Chitra Dutta (2016). "BPGA – an ultra-fast pan-genome analysis pipeline". In: *Scientific Reports* 6, p. 24373.

Chen, Caiyong, Tamika K Samuel, Jason Sinclair, et al. (2011). "An intercellular heme-trafficking protein delivers maternal heme to the embryo during development in *C. elegans*". In: *Cell* 145.5, pp. 720–731.

Chen, Xiaoshu and Jianzhi Zhang (2012). "The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data". In: *PLoS Computational Biology* 8.11.

Cheng, Xi, Yu Xiang, Hui Xie, et al. (2013). "Molecular characterization and functions of fatty acid and retinoid binding protein gene (*Ab-far-1*) in *Aphelenchoides besseyi*". In: *PLoS One* 8.6.

Chor, Benny, David Horn, Nick Goldman, et al. (2009). "Genomic DNA *k*-mer spectra: models and modalities". In: *Genome Biology* 10.10.

Clayton, Rob, Mike Storey, Bill Parker, et al. (2008). *Impact of reduced pesticide availability on control of potato cyst nematodes and weeds in potato crops*. Tech. rep. Potato Council Ltd, Oxford, UK.

Coghlan, Avril and Kenneth H Wolfe (2002). "Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*". In: *Genome Research* 12.6, pp. 857–867.

Compant, Stéphane, Jonathan Gerbore, Livio Antonielli, et al. (2017). "Draft genome sequence of the root-colonizing fungus *Trichoderma harzianum* B97". In: *Genome Announcements* 5.13.

Conway, Jake R, Alexander Lex, and Nils Gehlenborg (2017). "UpSetR: An R package for the visualization of intersecting sets and their properties". In: *Bioinformatics* 33.18, pp. 2938–2940.

Cotton, James A, Catherine J Lilley, Laura M Jones, et al. (2014). "The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode". In: *Genome Biology* 15.3.

Cotton, James A, Sasisekhar Bennuru, Alexandra Grote, et al. (2016). "The genome of *Onchocerca volvulus*, agent of river blindness". In: *Nature Microbiology* 2, p. 16216.

Dalquen, Daniel A and Christophe Dessimoz (2013). "Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals". In: *Genome Biology and Evolution* 5.10, pp. 1800–1806.

Danchin, Etienne GJ, Marie-Noëlle Rosso, Paulo Vieira, et al. (2010). "Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes". In: *PNAS* 107.41, pp. 17651–17656.

Danecek, Petr, Adam Auton, Goncalo Abecasis, et al. (2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–2158.

Davis, Eric L, Richard S Hussey, and Thomas J Baum (2004). "Getting to the roots of parasitism by nematodes". In: *Trends in Parasitology* 20.3, pp. 134–141.

De Boer, Jan M, Eric L Davis, Richard S Hussey, et al. (2002). "Cloning of a putative pectate lyase gene expressed in the subventral esophageal glands of *Heterodera glycines*". In: *Journal of Nematology* 34.1, pp. 9–11.

De Ley, P and ML Blaxter (2002). "Systematic position and phylogeny". In: *The Biology of Nematodes*. Taylor & Francis, pp. 1–30.

Decraemer, Wilfrida, David J Hunt, Roland N Perry, et al. (2006). "Structure and classification". In: *Plant Nematology*. CABI, pp. 3–32.

DEFRA (2010). *Science search on* `http://randd.defra.gov.uk/`.

Delmont, Tom O and A Murat Eren (2016). "Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies". In: *PeerJ* 4.

Denker, Elsa, Eric Bapteste, Hervé Le Guyader, et al. (2008). "Horizontal gene transfer and the evolution of cnidarian stinging cells". In: *Current Biology* 18.18, pp. 858–859.

Dentinger, Bryn TM, Ester Gaya, Heath O'Brien, et al. (2015). "Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life". In: *Biological Journal of the Linnean Society* 117.1, pp. 11–32.

Desjardins, Christopher A, Gustavo C Cerqueira, Jonathan M Goldberg, et al. (2013). "Genomics of *Loa loa*, a *Wolbachia*-free filarial parasite of humans." In: *Nature Genetics* 45.5, pp. 495–500.

Díaz, Alvaro, Cecilia Casaravilla, Anabella A Barrios, et al. (2016). "Parasite molecules and host responses in cystic echinococcosis". In: *Parasite Immunology* 38.3, pp. 193–205.

Dieterich, Christoph, Sandra W Clifton, Lisa N Schuster, et al. (2008). "The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism". In: *Nature Genetics* 40.10, pp. 1193–1198.

Dikow, Rebecca B, Paul B Frandsen, Mauren Turcatel, et al. (2017). "Genomic and transcriptomic resources for assassin flies including the complete genome sequence of *Proctacanthus coquilletti* (Insecta: Diptera: Asilidae) and 16 representative transcriptomes". In: *PeerJ* 5.

Ding, Xinhua, John Shields, Rex Allen, et al. (1998). "A secretory cellulose-binding protein cDNA cloned from the root-knot nematode (*Meloidogyne incognita*)". In: *Molecular Plant-Microbe Interactions* 11.10, pp. 952–959.

Dong, Xiaofeng, Stuart D Armstrong, Dong Xia, et al. (2017). "Draft genome of the honey bee ectoparasitic mite, *Tropilaelaps mercedesae*, is shaped by the parasitic life history". In: *GigaScience* 6.3, pp. 1–17.

Dubreuil, Géraldine, Marc Magliano, Emeline Deleury, et al. (2007). "Transcriptome analysis of root-knot nematode functions induced in the early stages of parasitism". In: *New Phytologist* 176.2, pp. 426–436.

Dunn, Casey W, Andreas Hejnol, David Q Matus, et al. (2008). "Broad phylogenomic sampling improves resolution of the animal tree of life". In: *Nature* 452, pp. 745–749.

Edgar, Robert C (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic Acids Research* 32.5, pp. 1792–1797.

Eletto, Davide, Daniela Eletto, Devin Dersh, et al. (2014). "Protein disulfide isomerase A6 controls the decay of IRE1α signaling via disulfide-dependent association". In: *Molecular Cell* 53.4, pp. 562–576.

Elsworth, Ben, Martin Jones, and Mark L Blaxter (2013). "Badger – an accessible genome exploration environment". In: *Bioinformatics* 29.21, pp. 2788–2789.

Elsworth, Benjamin, James Wasmuth, and Mark L Blaxter (2011). "NEMBASE4: the nematode transcriptome resource". In: *International Journal for Parasitology* 41.8, pp. 881–894.

Emms, David M and Steven Kelly (2015). "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy". In: *Genome Biology* 16.1, p. 157.

Enright, Anton J, Victor Kunin, and Christos A Ouzounis (2003). "Protein families and TRIBES in genome sequence space". In: *Nucleic Acids Research* 31.15, pp. 4632–4638.

Eren, A Murat, Özcan C Esen, Christopher Quince, et al. (2015). "Anvi'o: an advanced analysis and visualization platform for 'omics data." In: *PeerJ* 3.

Evans, K. and A. R. Stone (1977). "A Review of the Distribution and Biology of the Potato Cyst-Nematodes *Globodera rostochiensis* and *G. pallida*". In: *PANS* 23.2, pp. 178–189.

Evans, Ken, Javier Franco, and Maria M De Scurrah (1975). "Distribution of species of potato cyst-nematodes in South America". In: *Nematologica* 21.3, pp. 365–369.

Eves-van den Akker, Sebastian, Catherine J Lilley, John T Jones, et al. (2014). "Identification and characterisation of a hyper-variable apoplastic effector gene family of the potato cyst nematodes". In: *PLoS Pathogens* 10.9.

Eves-van den Akker, Sebastian, Catherine J Lilley, Hazijah B Yusup, et al. (2016a). "Functional C-terminally encoded peptide (CEP) plant hormone domains evolved *de novo* in the plant parasite *Rotylenchulus reniformis*". In: *Molecular Plant Pathology* 17.8, pp. 1265–1275.

Eves-van den Akker, Sebastian, Dominik R Laetsch, Peter Thorpe, et al. (2016b). "The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence". In: *Genome Biology* 17.124.

Farris, James S (1977). "Phylogenetic analysis under Dollo's Law". In: *Systematic Biology* 26.1, pp. 77–88.

Federhen, Scott (2012). "The NCBI Taxonomy database". In: *Nucleic Acids Research* 40, pp. 136–143.

Fernandez-Valverde, Selene L, Andrew D Calcino, and Bernard M Degnan (2015). "Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*". In: *BMC Genomics* 16, p. 387.

Finn, Robert D, Penelope Coggill, Ruth Y Eberhardt, et al. (2016). "The Pfam protein families database: towards a more sustainable future". In: *Nucleic Acids Research* 44, pp. 279–285.

Fioretti, Luca, Andrew Warry, Andrew Porter, et al. (2001). "Isolation and localisation of an annexin gene (*gp-nex*) from the potato cyst nematode, *Globodera pallida*". In: *Nematology* 3.1, pp. 45–54.

Fitch, Walter M (1970). "Distinguishing homologous from analogous proteins". In: *Systematic Biology* 19.2, pp. 99–113.

Foth, Bernardo J, Isheng J Tsai, Adam J Reid, et al. (2014). "Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction". In: *Nature Genetics* 46.7, pp. 693–700.

Franklin, Mary T (1951). "The cyst-forming species of *Heterodera*". In: *Commonwealth Agricultural Bureaux*. Farnham Royal.

Fuller, Skylar L, Elizabeth Savory, Alexandra J Weisberg, et al. (2017). "Isothermal amplification and lateral flow assay for detecting crown gall-causing *Agrobacterium* spp." In: *Phytopathology* 107.9, pp. 1062–1068.

Gao, B, R Allen, T Maier, et al. (2002a). "Identification of a new ss-1,4-endoglucanase gene expressed in the esophageal subventral gland cells of *Heterodera glycines*". In: *Journal of Nematology* 34.1, pp. 12–5.

Gao, Bingli, Rex Allen, Thomas R Maier, et al. (2001a). "Identification of putative parasitism genes expressed in the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*". In: *Molecular Plant-Microbe Interactions* 14.10, pp. 1247–1254.

— (2001b). "Molecular characterisation and expression of two venom allergen-like protein genes in *Heterodera glycines*". In: *International Journal for Parasitology* 31.14, pp. 1617–1625.

Gao, Bingli, Rex Allen, Thomas R Maier, et al. (2002b). "Characterisation and developmental expression of a chitinase gene in *Heterodera glycines*". In: *International Journal for Parasitology* 32.10, pp. 1293–1300.

Gao, Bingli, Rex Allen, Thomas R Maier, et al. (2003). "The parasitome of the phytonematode *Heterodera glycines*". In: *Molecular Plant-Microbe Interactions* 16.8, pp. 720–726.

Garrison, Erik and Gabor T Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: *ArXiv preprints* 1207.3907.

Gawryluk, Ryan M R, Javier Del Campo, Noriko Okamoto, et al. (2016). "Morphological identification and single-cell genomics of marine diplonemids". In: *Current Biology* 26.22, pp. 3053–3059. ISSN: 09609822.

GBD 2015 Disease and Injury Incidence and Prevalence Collaborators (2016). "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015". In: *Lancet* 388.10053, pp. 1545–1602.

Gerth, Michael and Gregory D D Hurst (2017). "Short reads from honey bee (*Apis* sp.) sequencing projects reflect microbial associate diversity". In: *PeerJ* 5.

Ghedin, Elodie, Shiliang Wang, David Spiro, et al. (2007). "Draft genome of the filarial nematode parasite *Brugia malayi*". In: *Science* 317.5845, pp. 1756–1760.

Godel, Christelle, Sujai Kumar, Georgios Koutsovoulos, et al. (2012). "The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets". In: *FASEB Journal* 26.11, pp. 4650–4661.

González, Víctor, Patricia Bustos, Miguel A Ramírez-Romero, et al. (2003). "The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments". In: *Genome Biology* 4.6.

González-Cabo, Pilar, Arantxa Bolinches-Amorós, Juan Cabello, et al. (2011). "Disruption of the ATP-binding cassette B7 (ABTM-1/ABCB7) induces oxidative stress and premature cell death in *Caenorhabditis elegans*". In: *Journal of Biological Chemistry* 286.24, pp. 21304–21314.

Goodwin, Sara, John D McPherson, and W Richard McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies". In: *Nature Review Genetics* 17.6, pp. 333–351.

Göttfert, Michael (1993). "Regulation and function of rhizobial nodulation genes". In: *FEMS Microbiology Reviews* 10.1-2, pp. 39–63.

Gregory, William F, Agnes K Atmadja, Judith E Allen, et al. (2000). "The abundant larval transcript-1 and -2 genes of *Brugia malayi* encode stage-specific candidate vaccine antigens for filariasis". In: *Infection and Immunity* 68.7, pp. 4174–4179.

Gremme, Gordon, Sascha Steinbiss, and Stefan Kurtz (2013). "GenomeTools: a comprehensive software library for efficient processing of structured genome annotations". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10.3, pp. 645–656.

Gross, Vladimir and Georg Mayer (2015). "Neural development in the tardigrade *Hypsibius dujardini* based on anti-acetylated α-tubulin immunolabeling". In: *EvoDevo* 6.12.

Guo, Yongfeng, Jun Ni, Robert Denver, et al. (2011). "Mechanisms of molecular mimicry of plant CLE peptide ligands by the parasitic nematode *Globodera rostochiensis*". In: *Plant Physiology* 157.1, pp. 476–484.

Habash, Samer S, Zoran S Radakovic, Radomira Vankova, et al. (2017). "*Heterodera schachtii* tyrosinase-like protein - a novel nematode effector modulating plant hormone homeostasis". In: *Scientific Reports* 7.1.

Haegeman, Annelies, John T Jones, and Etienne GJ Danchin (2011). "Horizontal gene transfer in nematodes: a catalyst for plant parasitism?" In: *Molecular Plant-Microbe Interactions* 24.8, pp. 879–887.

Hallan, J (2008). "Biology catalog". `https://insects.tamu.edu/research/collection/hallan/`.

Handoo, Zafar A, Lynn K Carta, Andrea M Skantar, et al. (2012). "Description of *Globodera ellingtonae* n. sp. (Nematoda: Heteroderidae) from Oregon". In: *Journal of Nematology* 44.1, pp. 40–57.

Herrero, Javier, Matthieu Muffato, Kathryn Beal, et al. (2016). "Ensembl comparative genomics resources". In: *Database*.

Hewezi, Tarek, Parijat S Juvale, Sarbottam Piya, et al. (2015). "The cyst nematode effector protein 10A07 targets and recruits host posttranslational machinery to mediate its nuclear trafficking and to promote parasitism in *Arabidopsis*". In: *Plant Cell* 27.3, pp. 891–907.

Hockland, Sue, Bjoern Niere, Eric Grenier, et al. (2012). "An evaluation of the implications of virulence in non-European populations of *Globodera pallida* and *G. rostochiensis* for potato cultivation in Europe". In: *Journal of Nematology* 14.1, pp. 1–13.

Hogenhout, Saskia A, Renier AL Van der Hoorn, Ryohei Terauchi, et al. (2009). "Emerging concepts in effector biology of plant-associated organisms". In: *Molecular Plant-Microbe Interactions* 22.2, pp. 115–122.

Holterman, Martijn, Gerrit Karssen, Sven van den Elsen, et al. (2009). "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". In: *Phytopathology* 99.3, pp. 227–235.

Hoolahan, Angelique H, Vivian C Blok, Tracey Gibson, et al. (2012). "A comparison of three molecular markers for the identification of populations of *Globodera pallida*". In: *Journal of Nematology* 44.1, pp. 7–17.

Horikawa, Daiki D, Takekazu Kunieda, Wataru Abe, et al. (2008). "Establishment of a rearing system of the extremotolerant tardigrade *Ramazzottius varieornatus*: a new model animal for astrobiology". In: *Astrobiology* 8.3, pp. 549–556.

Howe, Kerstin, Matthew D Clark, Carlos F Torroja, et al. (2013). "The zebrafish reference genome sequence and its relationship to the human genome". In: *Nature* 496.7446, pp. 498–503.

Howe, Kevin L, Bruce J Bolt, Scott Cain, et al. (2016). "WormBase 2016: expanding to enable helminth genomic research". In: *Nucleic Acids Research* 44.D1, pp. D774–780.

Howe, Kevin L, Bruce J Bolt, Myriam Shafie, et al. (2017). "WormBase ParaSite - a comprehensive resource for helminth genomics." In: *Molecular and Biochemical Parasitology* 215, pp. 2–10.

Hu, Zhiqiang, Chen Sun, Kuang-chen Lu, et al. (2017). "EUPAN enables pan-genome studies of a large number of eukaryotic genomes". In: *Bioinformatics* 33.15, pp. 2408–2409.

Huang, Guozhong, Bingli Gao, Tom Maier, et al. (2003). "A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita*". In: *Molecular Plant-Microbe Interactions* 16.5, pp. 376–381.

Huang, Guozhong, Ruihua Dong, Tom Maier, et al. (2004). "Use of solid-phase subtractive hybridization for the identification of parasitism gene candidates from the root-knot nematode *Meloidogyne incognita*". In: *Molecular Plant Pathology* 5.3, pp. 217–222.

Huang, Weichun, Leping Li, Jason R Myers, et al. (2012). "ART: a next-generation sequencing read simulator". In: *Bioinformatics* 28.4, pp. 593–594.

Huerta-Cepas, Jaime, Salvador Capella-Gutiérrez, Leszek P Pryszcz, et al. (2014). "PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome". In: *Nucleic Acids Research* 42.Database issue, pp. D897–902.

Hulsen, Tim, Martijn A Huynen, Jacob de Vlieg, et al. (2006). "Benchmarking ortholog identification methods using functional genomics data". In: *Genome Biology* 7.4, R31.

Hunt, Vicky L, Isheng J Tsai, Avril Coghlan, et al. (2016). "The genomic basis of parasitism in the *Strongyloides* clade of nematodes". In: *Nature Genetics* 48.3, pp. 299–307.

Husnik, Filip and John P McCutcheon (2016). "Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis". In: *PNAS* 113.37, pp. 5416–5424.

Hussey, Richard S (1989). "Disease-inducing secretions of plant-parasitic nematodes". In: *Annual Review of Phytopathology* 27.1, pp. 123–141.

Iberkleid, Ionit, Paulo Vieira, Janice de Almeida Engler, et al. (2013). "Fatty acid- and retinol-binding protein, Mj-FAR-1 induces tomato host susceptibility to root-knot nematodes". In: *PLoS One* 8.5.

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, et al. (2014). "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software". In: *PLoS One* 9.6.

Jaouannet, Maëlle, Marc Magliano, Marie-Jeanne Arguel, et al. (2013). "The root-knot nematode calreticulin Mi-CRT is a key effector in plant defense suppression". In: *Molecular Plant-Microbe Interactions* 26.1, pp. 97–105.

Johnson, Toby (2007). "Reciprocal best hits are not a logically sufficient condition for orthology". In: *ArXiv preprints* 0706.0117.

Jones, J. T., Cleber Furlanetto, and Taisei Kikuchi (2005). "Horizontal gene transfer from bacteria and fungi as a driving force in the evolution of plant parasitism in nematodes". In: *Nematology* 7.5, pp. 641–646.

Jones, John T., Cleber Furlanetto, Erin Bakker, et al. (2003). "Characterization of a chorismate mutase from the potato cyst nematode *Globodera pallida*". In: *Molecular Plant Pathology* 4.1, pp. 43–50.

Jones, John T., Annelies Haegeman, Etienne G. J. Danchin, et al. (2013). "Top 10 plant-parasitic nematodes in molecular plant pathology". In: *Molecular Plant Pathology* 14.9, pp. 946–961.

Jones, Michael GK (1981). "Host cell responses to endoparasitic nematode attack: structure and function of giant cells and syncytia". In: *Annals of Applied Biology* 97.3, pp. 353–372.

Jones, Philip, David Binns, Hsin-Yu Chang, et al. (2014). "InterProScan 5: genome-scale protein function classification". In: *Bioinformatics* 30.9, pp. 1236–1240.

Judelson, Howard S (2012). "Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution". In: *Eukaryotic Cell* 11.11, pp. 1304–1312.

Jun, Jihyung, Elisa Fiume, and Jennifer C Fletcher (2008). "The CLE family of plant polypeptide signaling molecules". In: *Cellular and Molecular Life Sciences* 65.5, pp. 743–755.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K F Wong, et al. (2017). "ModelFinder: fast model selection for accurate phylogenetic estimates". In: *Nature Methods* 14.6, pp. 587–589.

Katoh, Kazutaka and Daron M Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." In: *Molecular Biology and Evolution* 30.4, pp. 772–780.

Kent, W James and Alan M Zahler (2000). "Conservation, regulation, synteny, and introns in a large-scale *C. briggsae – C. elegans* genomic alignment". In: *Genome Research* 10.8, pp. 1115–1125.

Kikuchi, Taisei, Sebastian Eves-van den Akker, and John T. Jones (2017). "Genome evolution of plant-parasitic nematodes". In: *Annual Review of Phytopathology* 55.1, pp. 333–354.

Kikuchi, Taisei, Hajime Shibuya, Takuya Aikawa, et al. (2006). "Cloning and characterization of pectate lyases expressed in the esophageal gland of the pine wood nematode *Bursaphelenchus xylophilus*". In: *Molecular Plant-Microbe Interactions* 19.3, pp. 280–287.

Kikuchi, Taisei, James A Cotton, Jonathan J Dalzell, et al. (2011). "Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*". In: *PLoS Pathogens* 7.9.

Kocot, Kevin M, Mathew R Citarella, Leonid L Moroz, et al. (2013). "PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics". In: *Evolutionary Bioinformatics* 9, pp. 429–435.

Kondo, Koyuki, Takeo Kubo, and Takekazu Kunieda (2015). "Suggested involvement of PP1/PP2A activity and *de novo* gene expression in anhydrobiotic survival in a tardigrade, *Hypsibius dujardini*, by chemical genetic approach". In: *PLoS One* 10.12.

Kort, John, Hans Ross, Jürgen Rumpenhorst, et al. (1977). "An international scheme for identifying and classifying pathotypes of potato cyst-nematodes *Globodera rostochiensis* and *G. pallida*". In: *Nematologica* 23.3, pp. 333–339.

Koutsovoulos, Georgios, Benjamin Makepeace, Vincent N Tanya, et al. (2014). "Palaeosymbiosis revealed by genomic fossils of *Wolbachia* in a strongyloidean nematode". In: *PLoS genetics* 10.6.

Koutsovoulos, Georgios, Sujai Kumar, Dominik R Laetsch, et al. (2015). "The genome of the tardigrade *Hypsibius dujardini*". In: *bioRxiv*.

Koutsovoulos, Georgios, Sujai Kumar, Dominik R Laetsch, et al. (2016). "No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*". In: *PNAS* 113.18, pp. 5053–5058.

Kryukov, Kirill and Tadashi Imanishi (2016). "Human contamination in public genome assemblies". In: *PLoS One* 11.9.

Krzywinski, Martin, Jacqueline Schein, Inanc Birol, et al. (2009). "Circos: an information aesthetic for comparative genomics". In: *Genome Research* 19.9, pp. 1639–1645.

Kück, Patrick and Karen Meusemann (2010). "FASconCAT: Convenient handling of data matrices". In: *Molecular Phylogenetics and Evolution* 56.3, pp. 1115–1118.

Kumar, Sujai, Martin Jones, Georgios Koutsovoulos, et al. (2013). "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots". In: *Frontiers in Genetics* 4.237.

Laetsch, Dominik R (2017a). "DRL/kinfin: KinFin v0.8.3". In: `https://doi.org/10.5281/zenodo.377001`.

— (2017b). "DRL/kinfin: KinFin v1.0.3". In: `https://doi.org/10.5281/zenodo.834857`.

Laetsch, Dominik R and Mark L Blaxter (2017a). "BlobTools: Interrogation of genome assemblies [version 1; referees: awaiting peer review]". In: *F1000Research* 6.1287.

— (2017b). "KinFin: Software for taxon-aware analysis of clustered protein sequences". In: *bioRxiv*.

— (2017c). "KinFin: Software for taxon-aware analysis of clustered protein sequences". In: *G3: Genes, Genomes, Genetics* 7.10, pp. 3349–3357.

Laetsch, Dominik R, Georgios Koutsovoulos, Tim Booth, et al. (2017). "DRL/blobtools: BlobTools v1.0". In: `https://doi.org/10.5281/zenodo.833879`.

Laing, Roz, Taisei Kikuchi, Axel Martinelli, et al. (2013). "The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery". In: *Genome Biology* 14.8.

Lambert, Kris N, Keith D Allen, and Ian M Sussex (1999). "Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic

nematode *Meloidogyne javanica*". In: *Molecular Plant-Microbe Interactions* 12.4, pp. 328–336.

Lambshead, P John D (1993). "Recent developments in marine benthic biodiversity research". In: *Oceanis* 19.6, pp. 5–24.

Lander, Eric S, Lauren M Linton, Bruce Birren, et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921.

Laurence, Martin, Christos Hatzis, and Douglas E Brash (2014). "Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes". In: *PLoS One* 9.5.

Lee, Eduardo, Gregg A Helt, Justin T Reese, et al. (2013). "Web Apollo: a web-based genomic annotation editing platform". In: *Genome Biology* 14.8.

Lee, Tae-Ho, Hui Guo, Xiyin Wang, et al. (2014). "SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data". In: *BMC Genomics* 15.

Lefoulon, Emilie, Odile Bain, Benjamin L Makepeace, et al. (2016). "Breakdown of coevolution between symbiotic bacteria *Wolbachia* and their filarial hosts". In: *PeerJ* 4.

Leonard, Guy (2017). "guyleonard/single_cell_workflow: Updated workflow, introducing normalisation and CLI". In: `https://doi.org/10.5281/zenodo.438690`.

Lespinet, Olivier, Yuri I Wolf, Eugene V Koonin, et al. (2002). "The role of lineage-specific gene family expansion in the evolution of eukaryotes". In: *Genome Research* 12.7, pp. 1048–1059.

Levin, Michal, Leon Anavy, Alison G Cole, et al. (2016). "The mid-developmental transition and the evolution of animal body plans". In: *Nature* 531.7596, pp. 637–641.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, et al. (2014). "UpSet: Visualization of Intersecting Sets". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12, pp. 1983–1992.

Li, Heng (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv:1303.3997*.

Li, Heng, Bob Handsaker, Alec Wysoker, et al. (2009). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079.

Li, Li, Christian J Stoeckert, and David S Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." In: *Genome Research* 13.9, pp. 2178–2189.

Li, Weizhong and Adam Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". In: *Bioinformatics* 22.13, pp. 1658–1659.

Li, Xundong, Kan Zhuo, Mei Luo, et al. (2011). "Molecular cloning and characterization of a calreticulin cDNA from the pinewood nematode *Bursaphelenchus xylophilus*". In: *Experimental Parasitology* 128.2, pp. 121–126.

Li, Yu, Ke Wang, Hui Xie, et al. (2015). "A nematode calreticulin, Rs-CRT, is a key effector in reproduction and pathogenicity of *Radopholus similis*". In: *PLoS One* 10.6.

Lilley, Catherine J, Howard J Atkinson, and Peter E Urwin (2005). "Molecular aspects of cyst nematodes". In: *Molecular Plant Pathology* 6.6, pp. 577–588.

Lin, Shifeng, Heng Jian, Haijuan Zhao, et al. (2011). "Cloning and characterization of a venom allergen-like protein gene cluster from the pinewood nematode *Bursaphelenchus xylophilus*". In: *Experimental Parasitology* 127.2, pp. 440–447.

Lozano-Torres, Jose L, Ruud H P Wilbers, Sonja Warmerdam, et al. (2014). "Apoplastic venom allergen-like proteins of cyst nematodes modulate the activation of basal plant innate immunity by cell surface receptors". In: *PLoS Pathogens* 10.12.

Lu, Fei, Alexander E Lipka, Jeff Glaubitz, et al. (2013). "Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol". In: *PLoS Genetics* 9.1.

Lu, Shun-Wen, Shiyan Chen, Jianying Wang, et al. (2009). "Structural and functional diversity of CLAVATA3/ESR (CLE)-like genes from the potato cyst nematode *Globodera rostochiensis*". In: *Molecular Plant-Microbe Interactions* 22.9, pp. 1128–1142.

Luc, Michel, Armand R Maggenti, Renaud Fortuner, et al. (1987). "A reappraisal of Tylenchina (Nemata) 1. For a new approach to the taxonomy of Tylenchina". In: *Revue de Nématologie*.

Luck, Ashley N, Christopher C Evans, Molly D Riggs, et al. (2014). "Concurrent transcriptional profiling of *Dirofilaria immitis* and its *Wolbachia* endosymbiont throughout the nematode life cycle reveals coordinated gene expression". In: *BMC Genomics* 15.

Luck, Ashley N, Xiaojing Yuan, Denis Voronin, et al. (2016). "Heme acquisition in the parasitic filarial nematode *Brugia malayi*". In: *FASEB Journal* 30.10, pp. 3501–3514.

Lunt, David H, Sujai Kumar, Georgios Koutsovoulos, et al. (2014). "The complex hybrid origins of the root knot nematodes revealed through comparative genomics". In: *PeerJ* 2.

Madani, Mehrdad, Sergei A Subbotin, Leonard J Ward, et al. (2010). "Molecular characterization of Canadian populations of potato cyst nematodes, *Globodera rostochiensis* and *G. pallida* using ribosomal nuclear RNA and cytochrome *b* genes". In: *Canadian Journal of Plant Pathology* 32.2, pp. 252–263.

Mahran, Amro, Susan J Turner, Trevor Martin, et al. (2010). "The golden potato cyst nematode *Globodera rostochiensis* pathotype Ro1 in the Saint-Amable regulated area in Quebec, Canada". In: *Plant Disease* 94.12, pp. 1510–1510.

Maier, Thomas R, Tarek Hewezi, Jiqing Peng, et al. (2013). "Isolation of whole esophageal gland cells from plant-parasitic nematodes for transcriptome analyses and effector identification". In: *Molecular Plant-Microbe Interactions* 26.1, pp. 31–35.

Mallet, Ludovic, Tristan Bitard-Feildel, Franck Cerutti, et al. (2017). "PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies". In: *Bioinformatics* 33.20, pp. 3283–3285.

Martin, Christine, Vladimir Gross, Hans-Joachim Pflüger, et al. (2017). "Assessing segmental versus non-segmental features in the ventral nervous system of onychophorans (velvet worms)". In: *BMC Evolutionary Biology* 17.1, p. 3.

McDonald, John H and Martin Kreitman (1991). "Adaptive protein evolution at the *Adh* locus in *Drosophila*". In: *Nature* 351.6328, pp. 652–654.

McGrann, Graham R D, Ambrose Andongabo, Elisabet Sjökvist, et al. (2016). "The genome of the emerging barley pathogen *Ramularia collo-cygni*". In: *BMC Genomics* 17.

McNulty, Samantha N, Christina Strübe, Bruce A Rosa, et al. (2016). "*Dictyocaulus viviparus* genome, variome and transcriptome elucidate lungworm biology and support future intervention". In: *Scientific Reports* 6.

McNulty, Samantha N, Jose F Tort, Gabriel Rinaldi, et al. (2017). "Genomes of *Fasciola hepatica* from the Americas reveal colonization with *Neorickettsia* endobacteria related to the agents of Potomac horse and human Sennetsu fevers". In: *PLoS Genetics* 13.1.

Megen, Hanny van, Sven van den Elsen, Martijn Holterman, et al. (2009). "A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences". In: *Nematology* 11.6, pp. 927–950.

Mellbye, Brett L, Edward W Davis, Eva Spieck, et al. (2017). "Draft genome sequence of *Nitrobacter vulgaris* strain Ab1, a nitrite-oxidizing bacterium". In: *Genome Announcements* 5.18.

Merchant, Samier, Derrick E Wood, and Steven L Salzberg (2014). "Unexpected cross-species contamination in genome sequencing projects". In: *PeerJ* 2.

Mimee, Benjamin, Marc-Olivier Duceppe, Pierre-Yves Véronneau, et al. (2015). "A new method for studying population genetics of cyst nematodes based on Pool-Seq and genomewide allele frequency analysis". In: *Molecular Ecology Resources* 15.6, pp. 1356–1365.

Minnis, Stephen T, Patrick P J Haydock, Said K Ibrahim, et al. (2002). "Potato cyst nematodes in England and Wales - occurrence and distribution". In: *Annals of Applied Biology* 140.2, pp. 187–195.

Mitchum, Melissa G, Richard S Hussey, Thomas J Baum, et al. (2013). "Nematode effector proteins: an emerging paradigm of parasitism". In: *New Phytologist* 199.4, pp. 879–894.

Mitreva, Makedonka, Douglas P Jasmer, Dante S Zarlenga, et al. (2011). "The draft genome of the parasitic nematode *Trichinella spiralis*". In: *Nature Genetics* 43.3, pp. 228–235.

Molyneux, David H, Lorenzo Savioli, and Dirk Engels (2017). "Neglected tropical diseases: progress towards addressing the chronic pandemic". In: *Lancet* 389.10066, pp. 312–325.

Moreno-Hagelsieb, Gabriel and Kristen Latimer (2008). "Choosing BLAST options for better detection of orthologs as reciprocal best hits". In: *Bioinformatics* 24.3, pp. 319–324.

Morgan, Eric R, Johannes Charlier, Guy Hendrickx, et al. (2013). "Global change and helminth infections in grazing ruminants in Europe: impacts, trends and sustainable solutions". In: *Agriculture* 3.3, pp. 484–502.

Nadler, Steven A, Eugene Bolotin, and S Patricia Stock (2006). "Phylogenetic relationships of *Steinernema* Travassos, 1927 (Nematoda: Cephalobina: Steinernematidae) based on nuclear, mitochondrial and morphological data". In: *Systematic Parasitology* 63.3, pp. 161–181.

Nadler, Steven A, Ramon A Carreno, Hugo H Mejía-Madrid, et al. (2007). "Molecular phylogeny of clade III nematodes reveals multiple origins of tissue parasitism". In: *Parasitology* 134.10, pp. 1421–1442.

Nagayasu, Eiji, Sohta A Ishikawa, Shigeru Taketani, et al. (2013). "Identification of a bacteria-like ferrochelatase in *Strongyloides venezuelensis*, an animal parasitic nematode". In: *PLoS One* 8.3.

Nayak, Sudhir, Fernando E Santiago, Hui Jin, et al. (2002). "The *Caenorhabditis elegans* Skp1-related gene family: diverse functions in cell proliferation, morphogenesis, and meiosis". In: *Current Biology* 12.4, pp. 277–287.

Nehrt, Nathan L, Wyatt T Clark, Predrag Radivojac, et al. (2011). "Testing the ortholog conjecture with comparative functional genomic data from mammals". In: *PLoS Computational Biology* 7.6.

Nguyen, Lam-Tung, Heiko A Schmidt, Arndt von Haeseler, et al. (2015). "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies". In: *Molecular Biology and Evolution* 32.1, pp. 268–274.

Nicol, Julie M, S J Turner, Danny L Coyne, et al. (2011). "Current nematode threats to world agriculture". In: *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Springer, pp. 21–43.

Nielsen, Claus (2013). "The triradiate sucking pharynx in animal phylogeny". In: *Invertebrate Biology* 132.1, pp. 1–13.

Niu, Junhai, Pei Liu, Qian Liu, et al. (2016). "Msp40 effector of root-knot nematode manipulates plant immunity to facilitate parasitism". In: *Scientific Reports* 6.

Noon, Jason B, Tarek Hewezi, Thomas R Maier, et al. (2015). "Eighteen new candidate effectors of the phytonematode *Heterodera glycines* produced specifically in the secretory esophageal gland cells during parasitism". In: *Phytopathology* 105.10, pp. 1362–1372.

Nowell, Reuben W, Ben Elsworth, Vicencio Oostra, et al. (2017). "A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*". In: *GigaScience* 6.7, pp. 1–7.

Ohno, S. (1970). *Evolution by gene duplication*. Allen and Unwin.

Oliva, Ricardo, Joe Win, Sylvain Raffaele, et al. (2010). "Recent developments in effector biology of filamentous plant pathogens". In: *Cellular Microbiology* 12.6, pp. 705–715.

Olsen, Addie N and Karen Skriver (2003). “Ligand mimicry? Plant-parasitic nematode polypeptide with similarity to CLAVATA3”. In: *Trends in Plant Science* 8.2, pp. 55–57.

Olsen, Orvil A and Roland H Mulvey (1962). “The discovery of golden nematode in Newfoundland”. In: *Canadian Plant Disease Survey* 42.253.

Olson, Peter D and Vasyl V Tkach (2005). “Advances and trends in the molecular systematics of the parasitic Platyhelminthes”. In: *Advances in Parasitology* 60, pp. 165–243.

Opperman, Charles H, David M Bird, Valerie M Williamson, et al. (2008). “Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism”. In: *PNAS* 105.39, pp. 14802–14807.

Orchard, W. R. (1965). *Occurrence of the golden nematode on Vancouver Island, British Columbia*. Canadian Deptartment of Agriculture.

Paganini, Julien, Amandine Campan-Fournier, Martine Da Rocha, et al. (2012). “Contribution of lateral gene transfers to the genome composition and parasitic ability of root-knot nematodes”. In: *PLoS One* 7.11.

Pagel Van Zee, J, N S Geraci, F D Guerrero, et al. (2007). “Tick genomics: the *Ixodes* genome project and beyond”. In: *International Journal for Parasitology* 37.12, pp. 1297–1305.

Park, Byung-Jae, Duk-Gyu Lee, Jae-Ran Yu, et al. (2001). “Calreticulin, a calcium-binding molecular chaperone, is required for stress response and fertility in *Caenorhabditis elegans*”. In: *Molecular Biology of the Cell* 12.9, pp. 2835–2845.

Park, Joong-Ki, Tahera Sultana, Sang-Hwa Lee, et al. (2011). “Monophyly of clade III nematodes is not supported by phylogenetic analysis of complete mitochondrial genome sequences.” In: *BMC Genomics* 12.392.

Parra, Genis, Keith Bradnam, and Ian Korf (2007). “CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes”. In: *Bioinformatics* 23.9, pp. 1061–1067.

Peng, Huan, Jiangkuan Cui, Haibo Long, et al. (2016). “Novel pectate lyase genes of *Heterodera glycines* play key roles in the early stage of parasitism”. In: *PLoS One* 11.3.

Pestana, Margarida, Isabel Abrantes, and Manuela Gouveia (2015). “Effect of chemical stress imposed by *Solanum nigrum* in calreticulin and beta-1,4-endoglucanase genes and in infectivity of *Pratylenchus goodeyi*”. In: *European Journal of Plant Pathology* 141.4, pp. 747–759.

Petersen, Thomas Nordahl, Soren Brunak, Gunnar von Heijne, et al. (2011). “SignalP 4.0: discriminating signal peptides from transmembrane regions”. In: *Nature Methods* 8.10, pp. 785–786.

Pfeifer, Bastian, Ulrich Wittelsbürger, Sebastian E Ramos-Onsins, et al. (2014). “PopGenome: an efficient Swiss army knife for population genomic analyses in R”. In: *Molecular Biology and Evolution* 31.7, pp. 1929–1936.

Phillips, Mark S, J M S Forrest, and Linda A Farrer (1982). "Invasion and development of juveniles of *Globodera pallida* in hybrids of *Solanum vernei* x *S. tuberosum*". In: *Annals of Applied Biology* 100.2, pp. 337–344.

Phillips, Mark S and David L Trudgill (1998). "Variation of virulence, in terms of quantitative reproduction of *Globodera pallida* populations, from Europe and South America, in relation to resistance from *Solanum vernei* and *S. tuberosum* ssp. *andigena* CPC 2802". In: *Nematologica* 44.4, pp. 409–423.

Phillips, Mark S, Brian E Harrower, David L Trudgill, et al. (1992). "Genetic variation in British populations of *Globodera pallida* as revealed by isozyme and DNA analyses". In: *Nematologica* 38.1, pp. 304–319.

Phillips, Wendy S, Dana K Howe, Amanda M V Brown, et al. (2017). "The draft genome of *Globodera ellingtonae*". In: *Journal of Nematology* 49.2, pp. 127–128.

Plantard, Olivier, Damien Picard, Sylvie Valette, et al. (2008). "Origin and genetic diversity of Western European populations of the potato cyst nematode (*Globodera pallida*) inferred from mitochondrial sequences and microsatellite loci". In: *Molecular Ecology* 17.9, pp. 2208–2218.

Postma, Wiebe J, Erik J Slootweg, Sajid Rehman, et al. (2012). "The effector SPRYSEC-19 of *Globodera rostochiensis* suppresses CC-NB-LRR-mediated disease resistance in plants". In: *Plant Physiology* 160.2, pp. 944–954.

Prior, Alison, John T Jones, Vivian C Blok, et al. (2001). "A surface-associated retinol- and fatty acid-binding protein (Gp-FAR-1) from the potato cyst nematode *Globodera pallida*: lipid binding activities, structural analysis and expression pattern". In: *Biochemical Journal* 356.Pt 2, pp. 387–394.

Putnam, Nicholas H, Mansi Srivastava, Uffe Hellsten, et al. (2007). "Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization". In: *Science* 317.5834, pp. 86–94.

Pylypenko, Liliya A, Mark S Phillips, and Vivian C Blok (2008). "Characterisation of two Ukrainian populations of *Globodera pallida* in terms of their virulence and mtDNA, and the biological assessment of a new resistant cultivar Vales Everest". In: *Nematology* 10.4, pp. 585–590.

Pylypenko, Liliya A, Taketo Uehara, Mark S Phillips, et al. (2005). "Identification of *Globodera rostochiensis* and *G. pallida* in the Ukraine by PCR". In: *European Journal of Plant Pathology* 111.1, pp. 39–46.

Qin, Ling, Hein Overmars, Johannes Helder, et al. (2000). "An efficient cDNA-AFLP-based strategy for the identification of putative pathogenicity factors from the potato cyst nematode *Globodera rostochiensis*". In: *Molecular Plant-Microbe Interactions* 13.8, pp. 830–836.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, et al. (2013). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". In: *Nucleic Acids Research* 41.Database issue, pp. D590–596.

Quinlan, Aaron R and Ira M Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842.

Rajagopal, Abbhirami, Anita U Rao, Julio Amigo, et al. (2008). "Haem homeostasis is regulated by the conserved and concerted functions of HRG-1 proteins". In: *Nature* 453.7198, pp. 1127–1131.

Ramanujam, A (2017). "Awesome Command Line Tools". In: *PyCon 2017*. Python Software Foundation.

Rao, Anita U, Lynn K Carta, Emmanuel Lesuisse, et al. (2005). "Lack of heme synthesis in a free-living eukaryote". In: *PNAS* 102.12, pp. 4270–4275.

Rehman, Sajid, Vijai K Gupta, and Aakash K Goyal (2016). "Identification and functional analysis of secreted effectors from phytoparasitic nematodes". In: *BMC Microbiology* 16.48.

Rey-Burusco, M Florencia, Marina Ibáñez-Shimabukuro, Mads Gabrielsen, et al. (2015). "Diversity in the structures and ligand-binding sites of nematode fatty acid and retinol-binding proteins revealed by Na-FAR-1 from *Necator americanus*". In: *Biochemical Journal* 471.3, pp. 403–414.

Rijsbergen, CJ Van (1979). *Information Retrieval*. Butterworths.

Robb, Sofia M C, Eric Ross, and Alejandro Sánchez Alvarado (2008). "SmedGD: the *Schmidtea mediterranea* genome database". In: *Nucleic Acids Research* 36.Database issue, pp. D599–606.

Robertson, Lee, Walter M Robertson, Miroslaw Sobczak, et al. (2000). "Cloning, expression and functional characterisation of a peroxiredoxin from the potato cyst nematode *Globodera rostochiensis*". In: *Molecular and Biochemical Parasitology* 111.1, pp. 41–49.

Rokas, Antonis and Peter WH Holland (2000). "Rare genomic changes as a tool for phylogenetics". In: *Trends in Ecology and Evolution* 15.11, pp. 454–459.

Rosso, Marie-Noëlle and Eric Grenier (2011). "A wide range of effectors are secreted during parasitism". In: *Genomics and molecular genetics of plant-nematode interactions*. Ed. by J. Jones, G. Gheysen, and C. Fenoll. 2nd. Springer, pp. 287–307.

Rubin, Gerald M, Mark D Yandell, Jennifer R Wortman, et al. (2000). "Comparative genomics of the eukaryotes". In: *Science* 287.5461, pp. 2204–2215.

Sacco, Melanie Ann, Kamila Koropacka, Eric Grenier, et al. (2009). "The cyst nematode SPRYSEC protein RBP-1 elicits Gpa2- and RanGAP2-dependent plant cell death". In: *PLoS Pathogens* 5.8.

Salichos, Leonidas and Antonis Rokas (2011). "Evaluating ortholog prediction algorithms in a yeast model clade". In: *PLoS One* 6.4.

Samad, Abdul, Friederike Trognitz, Livio Antonielli, et al. (2016). "High-quality draft genome sequence of an endophytic *Pseudomonas viridiflava* strain with herbicidal properties against its host, the weed *Lepidium draba* L." In: *Genome Announcements* 4.5.

Satou, Yutaka, Katsuhiko Mineta, Michio Ogasawara, et al. (2008). "Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations". In: *Genome Biology* 9.10.

Schiffer, Philipp H, Michael Kroiher, Christopher Kraus, et al. (2013). "The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda". In: *BMC Genomics* 14.923.

Schiffer, Philipp H, Etienne Danchin, Ann M Burnell, et al. (2017). "Signatures of the evolution of parthenogenesis and cryptobiosis in the genomes of panagrolaimid nematodes". In: *bioRxiv*.

Scholl, Elizabeth H and David McK Bird (2005). "Resolving tylenchid evolutionary relationships through multiple gene analysis derived from EST data". In: *Molecular phylogenetics and evolution* 36.3, pp. 536–545.

Scholl, Elizabeth H, Jeffrey L Thorne, James P McCarter, et al. (2003). "Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach". In: *Genome Biology* 4.6.

Seemann, Torsten (2013). "Ten recommendations for creating usable bioinformatics command line software". In: *GigaScience* 2.1.

Severance, Scott, Abbhirami Rajagopal, Anita U Rao, et al. (2010). "Genome-wide analysis reveals novel genes essential for heme homeostasis in *Caenorhabditis elegans*". In: *PLoS Genetics* 6.7.

SGSFA Consortium (2009). "The *Schistosoma japonicum* genome reveals features of host-parasite interplay". In: *Nature* 460.7253, pp. 345–351.

Simakov, Oleg, Ferdinand Marletaz, Sung-Jin Cho, et al. (2013). "Insights into bilaterian evolution from three spiralian genomes". In: *Nature* 493.7433, pp. 526–531.

Simão, Felipe A, Robert M Waterhouse, Panagiotis Ioannidis, et al. (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics* 31.19, pp. 3210–3212.

Simillion, Cedric, Koen Janssens, Lieven Sterck, et al. (2008). "i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles". In: *Bioinformatics* 24.1, pp. 127–128.

Sinclair, Jason and Iqbal Hamza (2015). "Lessons from bloodless worms: heme homeostasis in *C. elegans*". In: *Biometals* 28.3, pp. 481–489.

Slos, Dieter, Walter Sudhaus, Lewis Stevens, et al. (2017). "*Caenorhabditis monodelphis* sp. n.: defining the stem morphology and genomics of the genus *Caenorhabditis*". In: *BMC Zoology* 2.1.

Smant, Geert, Jack P Stokkermans, Yitang Yan, et al. (1998). "Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes". In: *PNAs* 95.9, pp. 4906–4911.

Sobczak, Miroslaw and Wladyslaw Golinowski (2011). "Cyst nematodes and syncytia". In: *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Springer, pp. 61–82.

Song, Giltae, Benjamin J A Dickins, Janos Demeter, et al. (2015). "AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of *Saccharomyces cerevisiae*". In: *PLoS One* 10.3.

Spears, Joseph F (1968). *The golden nematode handbook*. Tech. rep. USDA.

Srinivasan, Jagan, Adler R Dillman, Marissa G Macchietto, et al. (2013). "The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle". In: *Genetics* 193.4, pp. 1279–1295.

Srivastava, Mansi, Emina Begovic, Jarrod Chapman, et al. (2008). "The *Trichoplax* genome and the nature of placozoans". In: *Nature* 454.7207, pp. 955–960.

Stamatakis, Alexandros (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." In: *Bioinformatics* 30.9, pp. 1312–1313.

Stogios, Peter J, Gregory S Downs, Jimmy J S Jauhal, et al. (2005). "Sequence and structural analysis of BTB domain proteins". In: *Genome Biology* 6.10.

Subbotin, Sergei A, Paul Halford, Andrew Warry, et al. (2000). "Variations in ribosomal DNA sequences and phylogeny of *Globodera* parasitising solanaceous plants". In: *Nematology* 2.6, pp. 591–604.

Sun, F, S Miller, S Wood, et al. (2007). "Occurrence of Potato Cyst Nematode, Globodera rostochiensis, on Potato in the Saint-Amable Region, Quebec, Canada". In: *Plant Disease* 91.7, pp. 908–908.

Suzek, Baris E, Yuqi Wang, Hongzhan Huang, et al. (2015). "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". In: *Bioinformatics* 31.6, pp. 926–932.

Szitenberg, Amir, Laura Salazar-Jaramillo, Vivian C. Blok, et al. (2017). "Comparative genomics of apomictic root-knot nematodes: hybridization, ploidy, and dynamic genome change". In: *Genome Biology and Evolution* 9.10, pp. 2844–2861.

Tang, Yat T, Xin Gao, Bruce A Rosa, et al. (2014). "Genome of the human hookworm *Necator americanus*". In: *Nature Genetics* 46.3, pp. 261–269.

Tange, Ole (2011). "GNU Parallel - The Command-Line Power Tool". In: *The USENIX Magazine* 36.1, pp. 42–47.

Tatusov, Roman L, Eugene V Koonin, and David J Lipman (1997). "A genomic perspective on protein families". In: *Science* 278.5338, pp. 631–637.

Telford, Maximilian J, Sarah J Bourlat, Andrew Economou, et al. (2008). "The evolution of the Ecdysozoa". In: *Philosophical transactions of the Royal Society of London, Series B, Biological sciences* 363.1496, pp. 1529–1537.

Tennessen, Kristin, Evan Andersen, Scott Clingenpeel, et al. (2016). "ProDeGe: a computational protocol for fully automated decontamination of genomes". In: *ISME Journal* 10.1, pp. 269–272.

The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 45.D1, pp. D158–169.

Thorpe, Peter, Sophie Mantelin, Peter J Cock, et al. (2014). "Genomic characterisation of the effector complement of the potato cyst nematode *Globodera pallida*". In: *BMC Genomics* 15.923.

Trudgill, David L, Mark S Phillips, and MJ Elliott (2014). "Dynamics and management of the white potato cyst nematode *Globodera pallida* in commercial potato crops". In: *Annals of Applied Biology* 164.1, pp. 18–34.

Tsai, Isheng J, Magdalena Zarowiecki, Nancy Holroyd, et al. (2013). "The genomes of four tapeworm species reveal adaptations to parasitism". In: *Nature* 496.7443, pp. 57–63.

Tyagi, Rahul, Anja Joachim, Bärbel Ruttkowski, et al. (2015). "Cracking the nodule worm code advances knowledge of parasite biology and biotechnology to tackle major diseases of livestock". In: *Biotechnological Advances* 33.6, pp. 980–991.

Unger, Ron, Shai Uliel, and Shlomo Havlin (2003). "Scaling law in sizes of protein sequence families: from super-families to orphan genes". In: *Proteins* 51.4, pp. 569–576.

Van Dongen, Stijn Marinus (2001). "Graph clustering by flow simulation". PhD thesis. Utrecht University.

Vanholme, Bartel, Wouter van Thuyne, Katrien Vanhouteghem, et al. (2007). "Molecular characterization and functional importance of pectate lyase secreted by the cyst nematode *Heterodera schachtii*". In: *Molecular Plant Pathology* 8.3, pp. 267–278.

Vermeire, Jon J, Yoonsang Cho, Elias Lolis, et al. (2008). "Orthologs of macrophage migration inhibitory factor from parasitic nematodes". In: *Trends in Parasitology* 24.8, pp. 355–363.

Veronico, Pasqua, John T Jones, Mauro Di Vito, et al. (2001). "Horizontal transfer of a bacterial gene involved in polyglutamate biosynthesis to the plant-parasitic nematode *Meloidogyne artiellia*". In: *FEBS Letters* 508.3, pp. 470–474.

Vinuesa, Pablo and Bruno Contreras-Moreira (2015). "Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids". In: *Methods in Molecular Biology* 1231, pp. 203–232.

Wang, Jianbin, Makedonka Mitreva, Matthew Berriman, et al. (2012). "Silencing of germline-expressed genes by DNA elimination in somatic cells". In: *Developmental Cell* 23.5, pp. 1072–1080.

Wang, Jianying, Amy Replogle, Richard Hussey, et al. (2011a). "Identification of potential host plant mimics of CLAVATA3/ESR (CLE)-like peptides from the plant-parasitic nematode *Heterodera schachtii*". In: *Molecular Plant Pathology* 12.2, pp. 177–186.

Wang, Xiaoyun, Wenjun Chen, Yan Huang, et al. (2011b). "The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*". In: *Genome Biology* 12.10.

Wang, Xuan, Hongmei Li, Yongjian Hu, et al. (2007). "Molecular cloning and analysis of a new venom allergen-like protein gene from the root-knot nematode *Meloidogyne incognita*". In: *Experimental Parasitology* 117.2, pp. 133–140.

Wang, YaDong and Christopher Chandler (2016). "Candidate pathogenicity islands in the genome of 'Candidatus *Rickettsiella isopodorum*', an intracellular bacterium infecting terrestrial isopod crustaceans". In: *PeerJ* 4.

Wasmuth, James, Ralf Schmid, Ann Hedley, et al. (2008). "On the extent and origins of genic novelty in the phylum Nematoda". In: *PLoS Neglected Tropical Diseases* 2.7.

Weerasinghe, Ravisha R, David McK Bird, and Nina S Allen (2005). "Root-knot nematodes and bacterial Nod factors elicit common signal transduction events in *Lotus japonicus*". In: *PNAS* 102.8, pp. 3147–3152.

Williamson, Valerie M and Amar Kumar (2006). "Nematode resistance in plants: the battle underground". In: *Trends in Genetics* 22.7, pp. 396–403.

Winnepenninckx, Birgitta MH, Yves Van de Peer, and Thierry Backeljau (1998). "Metazoan relationships on the basis of 18S rRNA sequences: a few years later". In: *American Zoologist* 38.6, pp. 888–906.

Winter, Alan D, Gillian McCormack, and Antony P Page (2007). "Protein disulfide isomerase activity is essential for viability and extracellular matrix formation in the nematode *Caenorhabditis elegans*". In: *Developmental Biology* 308.2, pp. 449–461.

Winter, Alan D, Gillian McCormack, Johanna Myllyharju, et al. (2013). "Prolyl 4-hydroxlase activity is essential for development and cuticle formation in the human infective parasitic nematode *Brugia malayi*". In: *Journal of Biological Chemistry* 288.3, pp. 1750–1761.

Wu, Bo, Jacopo Novelli, Daojun Jiang, et al. (2013). "Interdomain lateral gene transfer of an essential ferrochelatase gene in human parasitic nematodes". In: *PNAS* 110.19, pp. 7748–7753.

Wubben, Martin J, Lily Gavilano, Thomas J Baum, et al. (2015). "Sequence and spatiotemporal expression analysis of CLE-motif containing genes from the reniform nematode *Rotylenchulus reniformis* (Linford & Oliveira)". In: *Journal of Nematology* 47.2, pp. 159–165.

Xiao, Jingfa, Zhewen Zhang, Jiayan Wu, et al. (2015). "A brief review of software tools for pangenomics". In: *Genomics Proteomics Bioinformatics* 13.1, pp. 73–76.

Yoshida, Yuki, Georgios Koutsovoulos, Dominik R Laetsch, et al. (2017a). "Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*". In: *bioRxiv*.

Yoshida, Yuki, Georgios Koutsovoulos, Dominik R Laetsch, et al. (2017b). "Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*". In: *PLoS Biology* 15.7, pp. 1–40.

Young, Neil D, Aaron R Jex, Bo Li, et al. (2012). "Whole-genome sequence of *Schistosoma haematobium*". In: *Nature Genetics* 44.2, pp. 221–225.

Zarowiecki, Magdalena and Matt Berriman (2015). "What helminth genomes have taught us about parasite evolution". In: *Parasitology* 142.S1, S85–S97.

Zdobnov, Evgeny M, Fredrik Tegenfeldt, Dmitry Kuznetsov, et al. (2017). "OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs." In: *Nucleic Acids Research* 45.D1, pp. D744–749.

Zhang, Guofan, Xiaodong Fang, Ximing Guo, et al. (2012). "The oyster genome reveals stress adaptation and complexity of shell formation". In: *Nature* 490.7418, pp. 49–54.

Zheng, Xiuwen, David Levine, Jess Shen, et al. (2012). “A high-performance computing toolset for relatedness and principal component analysis of SNP data”. In: *Bioinformatics* 28.24, pp. 3326–3328.