UNIVERSITY OF HULL

Comparative genomics of *Meloidogyne haplanaria*

being a Thesis submitted for the Degree of

Masters of Research in Biological Sciences

in the University of Hull

by

Michael Robert Winter BSc

September 2020

Dedication

For Lucas.

Acknowledgements

Many people have helped me throughout the performance of this study, some of which I would like to thank here.

I would like to say thank you to my family particularly my fiancée, Pelina Santos, for your uncompromising support and reassurance throughout this process. I could not have done it without you.

I would also like to thank my supervisors and colleagues from the EvoHull Evolutionary Biology Group at the University of Hull for supplying ideas and support throughout.

Special thanks also go to Amir Szitenberg, Chris Collins, Laura Salazar-Jaramillo, and Kamil Jaron for their insightful correspondence and technical assistance.

Finally, I would like to thank the James Reckitt Charity, Hull, for grants awarded.

Publications and Conferences

Early concepts of the methods used in Chapter 2 were presented as a poster at Population Genetics Group 53. University of Leicester, 2020.

Contents

Abstract

Root-knot nematodes are a scientifically and agriculturally important group of plant parasites. Genomic investigations into this group have proven difficult due to a complex genomic arrangement and recent inter-species hybridisations. Here we design and employ novel bioinformatic workflows to assemble the genomes of *Meloidogyne* species and perform phylogenomic analyses on them. We use *Meloidogyne haplanaria* - an emerging crop pest recently shown to be capable of breaking cultivated resistance - as a test organism. We assemble and annotate its genome for the first time and infer its position in the *Meloidogyne* phylogeny. This will inform future investigations into diagnostic and control methods, as well as investigations into the evolutionary history of the genus. The workflows themselves will provide accessible bioinformatic tools for the reproducible assembly and phylogenomic analysis of *Meloidogyne* genomes to the wider scientific community. Greater elucidation of the complex genomics of the *Meloidogyne* genus can grant insight into many biological processes, including hybridisation, parasitic adaptation and evolution in the absence of recombination, and the effect that different parthenogenetic sexual systems can have on genomic architecture.

List of Figures

List of Tables

Chapter 1: Introduction

Root-knot nematodes (RKNs) are a group of obligate parasitic nematodes that infect virtually all species of flowering plant Trudgill and Blok, 2001; Eves-van den Akker and Jones, 2018). The biology of this group is remarkably interesting as well as economically important, but its genomic and evolutionary complexity make investigation difficult (Szitenberg *et al.*, 2017). To ease this process we developed reproducible bioinformatics workflows (Köster and Rahmann, 2012a) specifically for analysis of *Meloidogyne* species, and used them to perform a genomic analysis on the emerging parasite *Meloidogyne haplanaria* (Eisenback et al., 2003; Joseph et al., 2016).

1.1 A genomic approach

1.1.1 Comparative genomics

Comparative genomics is the process of gathering genomic information from a group of organisms to infer homology, structural variation, and evolutionary relationships. This information can range from raw sequencing reads to assembled and annotated genomes. By comparing characteristics, we can detect biological similarities or differences between species and groups (Hardison, 2003; Haubold and Wiehe, 2004; International Helminth Genomes Consortium, 2019). At a basic level these analyses encompass comparison of summary statistics including genome length, ploidy state, number of protein coding genes, and others, though deeper comparison of sequences at a nucleotide level through a phylogenomic analysis can provide high resolution inference of evolutionary history and homology (Hardison, 2003; Haubold and Wiehe, 2004). Before a thorough genomic comparison can be performed, the genomes of each study organism must be sequenced, assembled, and annotated (International Helminth Genomes Consortium, 2019).

1.1.2 Genome assembly

Several methods of genome assembly exist, each designed for data produced using different sequencing chemistries (Mardis, 2011; Lu, Giordano and Ning, 2016; Jayakumar and Sakakibara, 2019). Only short-read assembly is pertinent to this study; long-read assembly, and other third generation methods, will only be discussed in passing. Short-read assembly is most often performed using reads generated through Illumina chemistry

(Chen et al., 2014). The DNA of an organism is extracted, prepared and run through a sequencing machine. The DNA is fragmented into short sequences - reads - of around 100-300 base pairs which are individually read, converted to a nucleotide text sequence, and outputted in a *.fastq* file. Depending on the chemistry used, one sequencing run can generate tens of millions of reads and tens of gigabytes of raw data. To be used in downstream analyses, the sequenced reads must be organised and reconnected as accurately as possible into a resemblance of the original genome. This process is called genome assembly (Figure 1.1) (Kitts, 2002; Young and Gillung, 2019).

Raw reads are assessed and trimmed for quality, before being run into a genome assembly software. Modern genome assemblers such as *SPAdes* (Bankevich et al., 2012), *Velvet* (Zerbino and Birney, 2008), and *platanus* (Kajitani et al., 2014), employ a de Bruijn graph (DBG) method to order reads, as they are considered faster and more random-access-memory (RAM) efficient than other available methods (Li et al., 2012; Khan et al., 2018). Under the DBG method reads are fragmented into smaller sequences of length *k*, called *k*-mers. These *k*-mers are aligned on a graph. If multiple *k*-mers corroborate a position, a directed edge is drawn. A Eulerian path is then found through the graph including all directed edges. This path is an assembled sequence, referred to as a contig (Pevzner, Tang and Waterman, 2001; Kang et al., 2013). Genome assemblies typically consist of many separate contigs due to inability of the assembler to determine a connecting path between them, leading to fragmented assemblies. Assemblies that are contiguous to a chromosomal level are the ultimate goal and ideal, but this is rarely the case, particularly with non-model organisms and complex genomes, and especially when using short-read sequence data (Chakraborty et al., 2016). Several supplementary methods exist to extend contigs and close gaps between them, creating scaffolds, but unconstrained and over-aggressive gap closing can impair the quality of the assembly (Pop et al., 2004; Scheibye-Alsing et al., 2009; Pryszcz and Gabaldón, 2016).

Difficulties accompanying short-read assembly include an inability to correctly reconstitute regions of the genome with a high amount of repeat content - microsatellites, tandem repeats - due to a multitude of possible sequence overlaps lowering confidence (Du and Liang, 2019). The likelihood that a base called in a sequence or assembly is assessed on coverage; how many reads fall in that position and corroborate the individual base (Desai et al., 2013). With highly repetitive regions, reads could be multiples of the same genomic sequence, thereby increasing its coverage, or more likely, contiguous

sequences with low coverage. This ambiguity lowers confidence in the assembly of these regions, as well as the capability of the assembler to assemble them. Short-read assemblies can be improved through the application of long-read sequences to bridge gaps between contigs and scaffolds. In particular, regions with high repeat content which cannot be disentangled with short-reads alone can be sequenced in their entirety by long reads, vastly increasing the continuity of the assembly (Miller et al., 2017; De Maio et al., 2019; Du and Liang, 2019).

1.1.3 Genome annotation

Once a genome assembly of acceptable quality has been created, it requires annotating. Annotation is the process of detecting or predicting the number and locations of genes in the assembly, marking their location, assigning ontology, and extracting the desired genetic features (Koonin and Galperin, 2003a; Campbell and Yandell, 2015). Numerous methods of gene prediction are available, the two most common being prediction through sequence similarity, and prediction through *ab initio* methods and machine learning (Campbell and Yandell, 2015). In sequence similarity prediction, an annotation package or software is provided with coding sequences (CDS) of the same or closely related species. Sequence similarity searches are then performed, often by *BLAST*, and hits above a predefined threshold within the assembly are pulled and written to an output file (Harrison, 2014). In *ab initio* prediction, CDS of the same or closely related species are again provided, but instead of sequence similarity, the *ab initio* software examines the provided CDS for genetic markers, such as start/stop codons and promoter regions. This step is called “training” the software. The format of the genetic structure inferred from the training CDS is then applied to the genome assembly needing annotation. Genetic structures and sequences conforming to the rules created through training are identified as genes, highlighted, and extracted (Kitts, 2002; Campbell and Yandell, 2015). Examples of *ab initio* gene predictors are *SNAP* and *Augustus* (Korf, 2004; Hoff and Stanke, 2019). Annotation of a genome is typically done using a combination of these software and methods, using the predictions of previous steps to train the next, with each step providing more accurate predictions (Salzberg, 2019).

1.1.4 Phylogenomics

In comparative genomics, phylogenomics is defined as the process of using data from several related species to infer evolutionary history (Delsuc, Brinkmann and Philippe, 2005; Williams et al., 2019; Young and Gillung, 2019). Like phylogenetics, the mechanism of detecting evolutionary proximity is sequence similarity, but unlike phylogenetics, which uses data limited to at most a few genes, phylogenomics collates the evolutionary history of as many genes as possible, not only giving greater statistical power to any conclusions, but allowing phenomena such as contrasting evolutionary history between genes or gene copies within the same genome or group to be detected (Young and Gillung, 2019).

Once orthology has been inferred and sequences cleaned, species are organised into dendrograms, with more closely related species connected at nodes further right on the x-axis, with the x-axis being evolutionary time, or number of substitutions per site. An example of this can be seen in Figure 1.4. Trees are usually rooted at an outgroup; a species known to be more evolutionarily distant to all species in the analysis than any could be to each other (Wilberg, 2015).

1.2 Root knot nematodes

Root-knot nematodes (RKN) (*Meloidogyne*) are a genus of microscopic phytoparasitic roundworms that infect the roots of many plant species, in which they cause the formation of galls within the host root tissue (Perry, Moens and Starr, 2009). RKNs are obligate endoparasites and can severely affect the fitness and yield of their host. The genus *Meloidogyne* includes over 100 species of nematode geographically distributed across all continents barring Antarctica (Jones et al, 2013), and with the collective ability to infect tens of thousands, if not virtually all, plant species (Elling, 2013; Ralmi et al., 2016).

Many species of RKN infect agricultural crops, causing reduced yield and costing tens of billions of pounds in crop losses each year (Elling, 2013; Bernard, Egnin and Bonsi, 2017a), and it is estimated that RKNs account for between five and ten percent of annual worldwide agricultural losses (Sasser and Freckman, 1987; Nicol *et al.*, 2011; Bernard, Egnin and Bonsi, 2017a), with some localities losing as much as 30% (Collange et al., 2011).

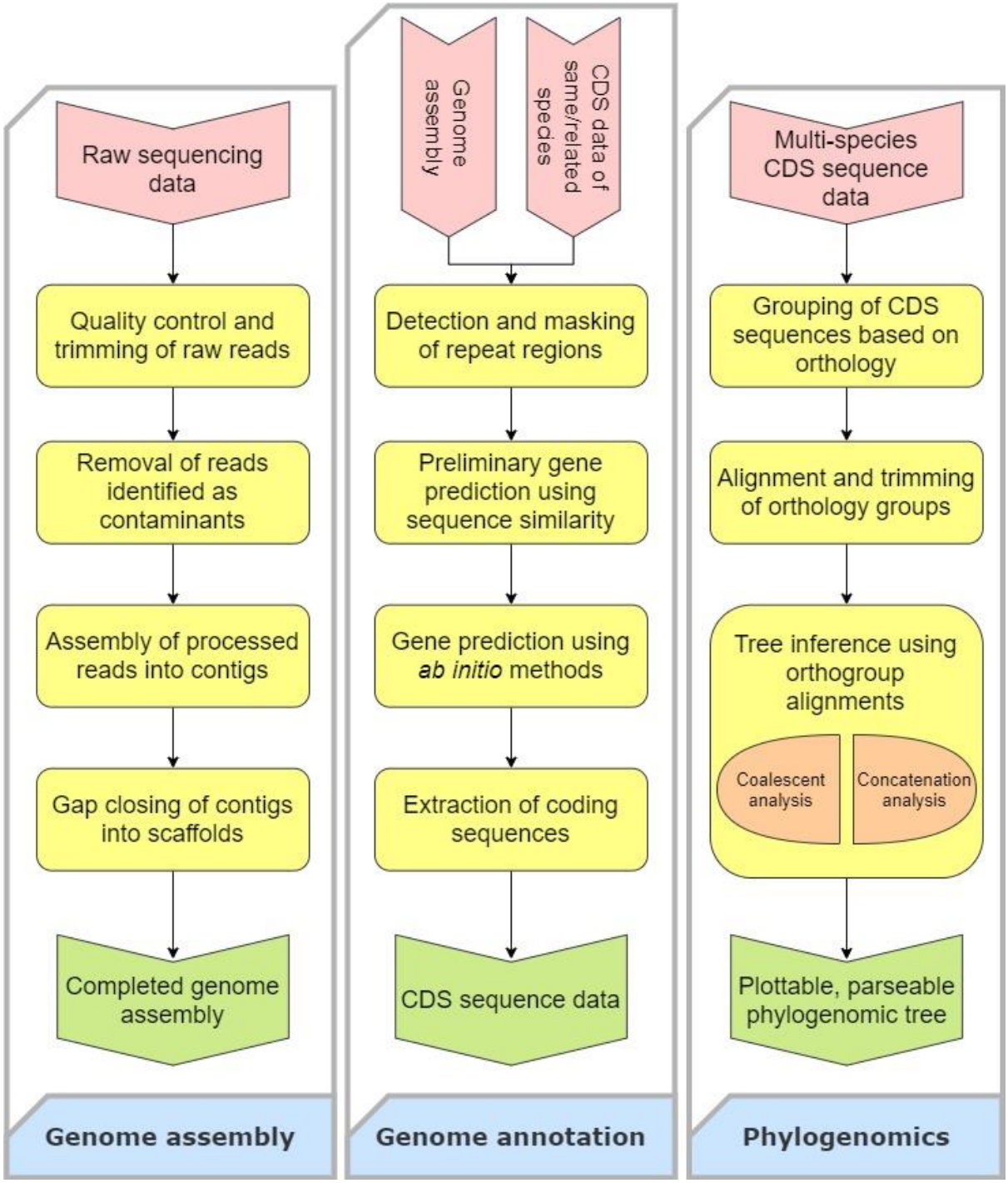

Figure 1.1: Flowcharts showing the basic stages of different bioinformatic processes. Left, the genome assembly process. Centre, the genome annotation process. Right, the stages of phylogenomic analysis.

The morphological characteristics of *Meloidogyne* nematodes change as they develop. Embryogenesis generates a first stage juvenile (J1) which moults while still inside the egg to become a second stage juvenile (J2). The J2 hatches from the egg and rapidly establishes a parasitic relationship with the host plant. Once this relationship is

established, the J2 remains in a preliminary feeding stage for up to eight weeks before moulting three times into an adult RKN (Eisenback, 1985; Wesemael, Viaene and Moens, 2011). Adult males are slender, worm-like, and motile, abstaining from feeding and leaving the host plant, travelling through the soil matrix. They generally measure up to 1400 µm. Adult females remain sedentary and stay with the host plant after their fourth moult, continuing to feed while producing eggs. Females are typically swollen with eggs, appearing teardrop shaped and measuring around 700um long by 400um wide (Figure 1.2a-c) (Eisenback, 1985).

Gall formation in the root of the host plant is induced through a combination of influences. The swelling of the egg mass is a contributor and can be seen as brown spots on the plant (Figure 1.3a-c), but the main catalyst is the formation of hypertrophied ‘giant’ cells. Giant cells are formed through injection of a cocktail of effector proteins through the stylet into the host root cell. This cocktail includes a cellulose binding protein, to degrade the host cell wall (Menezes et al., 2019). Most notably, these proteins hijack the CLE (CLV3/endosperm surrounding region) signalling pathway, a series of intercellular signalling molecules that control physiological and developmental processes such as maintenance of meristematic function and cell homeostasis (Betsuyaku, Sawa and Yamada, 2011; Rutter et al., 2014), Disruption of this pathway and the uncoupling of cell division from mitotic endoreduplication causes increased cytoplasmic content and hypertrophy of the cell (Caillaud et al., 2008). Giant cells continue to grow whilst being parasitised by the sedentary nematode, leading to gall formation (Rutter et al., 2014).

1.2.1 Agricultural impact

RKNs, and in particular the Clade I tropical RKNs (Figure 1.4) are devastating pathogenic crop pests (Bird et al., 2009; Elling, 2013). Over a fifth of worldwide agricultural losses can be attributed to them, with crop losses ranging from 5% in fields treated with nematode controls, to 100% in untreated lab settings (Sasser and Freckman, 1987; Nicol et al., 2011; Wesemael, Viaene and Moens, 2011; Bernard, Egnin and Bonsi, 2017a). These losses amount to billions of dollars annually, and severely affect developing countries where control measures are unavailable or unattainable (Bebber, Holmes and Gurr, 2014; Koutsovoulos, Poullet, et al., 2019). As a result of these factors, the *Meloidogyne* genus was recently declared to be the most scientifically and economically important of all plant-parasitic nematodes (Jones et al, 2013).

Current methods for treatment of nematode infected areas are varied in both application and effectiveness. Chemical methods primarily involve the application of nematicides; powerful pesticides engineered to prevent and remove RKNs. Despite being reasonably effective, many of these treatments have been phased out and withdrawn due to rising concerns over levels of toxicity to humans (Burns et al., 2017).

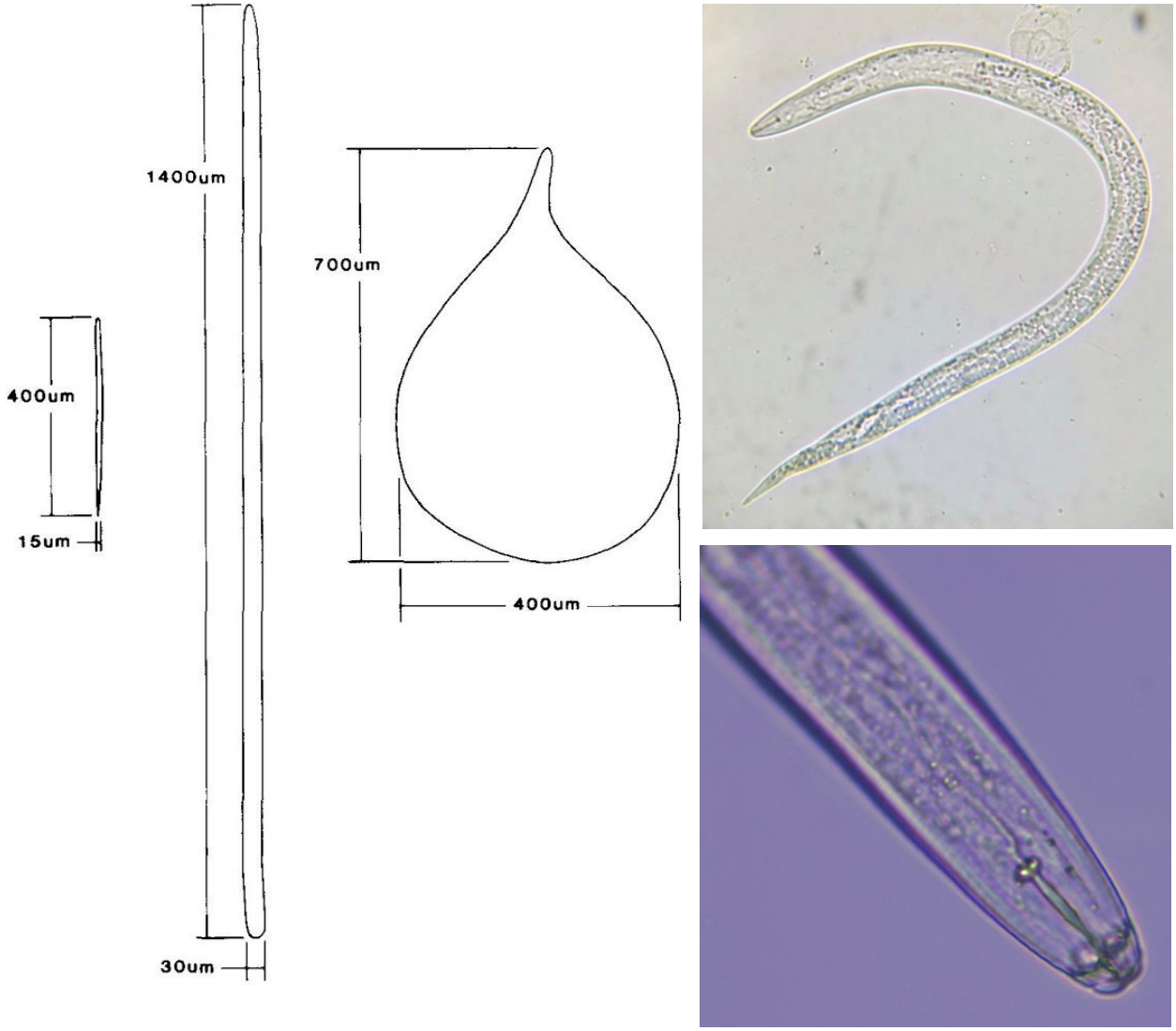

Figure 1.2a-c: Root-knot nematode morphology. A, Biological drawing showing general shape and dimensions of *Meloidogyne* nematodes (left, J2 stage juvenile, centre, adult male, right, adult female swollen with eggs) (Eisenback, 1985). B, second stage juvenile (J2) RKN (unknown species) (Nelson, 2015). C, Head of an adult male *Meloidogyne incognita*. Stylet can be clearly seen (Nelson, 2018).

Non-chemical methods of control include; soil solarisation - soil is covered and heated to kill nematode eggs - crop rotation to non-host crops, trap cropping, and introducing bacterial and fungal nematode control agents into the soil (Ralmi et al., 2016). RKN management also employs molecular and genetic techniques, partly in preliminary species identification of infestation, but predominantly in the development of resistant plant cultivars through selective breeding and genetic engineering (Powers et al., 2005; Jongman, Carmichael and Bill, 2020). Some success has been had with breeding and using resistant cultivars, but RKN populations quickly respond to this evolutionary pressure, breaking resistance relatively quickly and continuing to infect the crop (Nicol et al., 2011; Wesemael, Viaene and Moens, 2011; Joseph et al., 2016).

Figure 1.3a-c: Symptoms of root-knot nematode infection. A, top left, Brown egg masses resulting from *Meloidogyne* infection (Nelson, 2018). B, top right, Tomato (*Solanum lycopersicum*) infected with *Meloidogyne incognita.* Roots are clearly galled. Brown coloured egg masses can be seen on some galls (Nelson, 2017). C, bottom left, Carrot (*Daucus carota*) infected with *Meloidogyne* spp. Effect of RKN pathogenicity can be clearly seen (Nelson, 2007).

1.2.2 Clade I and the *Meloidogyne incognita* group

Tropical root-knot nematodes

Using molecular methods, the *Meloidogyne* genus has been divided into several well defined clades (Figure 1.4), of which Clade I, containing the tropical RKNs, is the most impactful and destructive in terms of global crop losses (Mwageni et al., 2000; Trudgill and Blok, 2001; Bebber, Holmes and Gurr, 2014).

Within Clade I is the *Meloidogyne incognita* group (MIG) (Figure 1.4), which includes its namesake, *M. incognita,* as well as other notable species; *M. javanica* and *M. arenaria.* All of these species are incredibly damaging crop pests, which alongside *M. hapla* - a Clade II apomict - are responsible for up to 95% of all RKN incurred damage (Wesemael, Viaene and Moens, 2011). Though not a member of the MIG, it is important to also introduce *Meloidogyne enterolobii. M. enterolobii* is a Clade I apomictic RKN that can infect many plant cultivars resistant to members of the MIG, and is arguably as agriculturally damaging (Yang and Eisenback, 1983; Wesemael, Viaene and Moens, 2011; Rashidifard et al., 2019).

Figure 1.4: Diagram of clades within the *Meloidogyne* genus and explanation of tree. Orange, *Meloidogyne incognita* group (MIG). Green, Clade I of *Meloidogyne.* Blue, Clade II of *Meloidogyne.* Black lines are 'branches' and meeting points of branches are 'nodes'. The left-most branch/node is the 'root'. Nodes closer to the root are 'basal', and features to the right of these are 'distal'. Species in a phylogenomic tree are arranged to represent evolutionary distance inferred through sequence similarity.

All MIG species are parthenogenic apomicts, except for *M. floridensis*, an automict. Apomixis is a sexual system of mitotic parthenogenesis, completely absent of meiosis and recombination. Mitotic sister chromatids split into daughter ova, each containing a full genomic complement, which go on to mature into adult females (Butlin, Schön and Griffiths, 1998). This is an advantageous evolutionary strategy in a system such as soil where movement and signalling are limited, and also enables the evolution of a sedentary lifestyle. Apomictic nematodes have particularly high reproductive rates and short generation times, increasing their potential success as infectious colonists (Trudgill and Blok, 2001). Potential deleterious effects of apomixis include reduced genetic diversity and increased susceptibility to Muller's ratchet; the accumulation of deleterious alleles in a population that does not undergo recombination (Butlin, Schön and Griffiths, 1998). Despite this, apomictic RKNs exhibit great capacity for adaptation (Castagnone-Sereno et al., 2013).a

Figure 1.5: Diagram of hypotriploidy and divergent genomic copies. Genomes of species in the MIG contain two divergent heterozygous copies, *A,* and *B,* each inherited from a different ancestral progenitor, parent A (blue) or parent B (red). Alongside this, some species are hypotriploid, containing another copy, A^1 (*A* prime), that is homozygous to *A*. The proportion of copy *A* that is also present in A^1 differs between *Meloidogyne* species.

Genomics of Clade I and the MIG

Many species of Clade I are thought to be hypotriploid; they have a portion of one haploid genotype present in a second copy, referred to as prime (^{1}) (Figure 1.5). The proportion of this prime copy differs across species (Eisenback and Triantaphyllou, 1991; Lunt et al., 2014), and was confirmed to present in the MIG by a read coverage analysis (Szitenberg

et al. 2017). Each copy was inherited from a different ancestral progenitor before divergence into a paraphyletic group.

Alongside hypotriploidy, species in the MIG contain two divergent genomic copies, *A* and *B* (Figure 1.5) Lunt et al. (2014), and later Szitenberg et al (2017), tested for divergent copies across much of Clade I using a novel intra-genomic sequence similarity analysis, wherein the number of genes with two highly similar, presumably orthologous copies was counted to estimate the presence and proportion of the prime copy. Here it was found that all members of Clade I tested (*M. enterolobii, M. floridensis, M. arenaria, M. javanica, M. incognita)* exhibited divergent gene pairs, *M. floridensis* slightly less so than others. This was confirmed by a Robinson-Foulds distance matrix, which confirmed two different evolutionary topologies within the MIG. Heterozygosity between these copies is on average 3%, but as high as 12% in some regions. Several hypotheses of the origin of this genomic constitution were suggested, including allelic sequence divergence from lack of recombination, and whole genome duplication followed by sequence divergence and segmental loss, but neither explanation accounts for the hypotriploid portion. It is now primarily thought that the divergent genomic copies are a relic from a past hybridisation event, where two allopatric diploid parental genomes hybridise and gradually decline into hypotriploidy through gene conversion and segmental loss (Szitenberg et al., 2017).

1.2.3 Biological interest

RKNs are a prime example of biochemical not morphological adaptation driving evolution. Species and clades that are separated evolutionarily by significant structural and functional genomic distinctions are incredibly similar morphologically, and identifying RKNs to a species level based on morphology remains challenging, requiring specialised knowledge and mature specimens (Oliveira, Monteiro and Blok, 2011), resulting in molecular identification methods have been developed to ease this process (Wesemael, Viaene and Moens, 2011). In contrast, biochemical adaptations are more distinctive and prevalent (Castagnone-Sereno and Danchin, 2014; Koutsovoulos et al., 2018). In the presence of apomixis and the absence of genetic recombination, genetic diversity within a population is expected to be limited, curtailing propensity for adaptation (Trudgill and Blok, 2001). However, MIG species have displayed a capacity toward breaking resistance of engineered cultivars, revealing the presence of a rapid mechanism of adaptation (Koutsovoulos, Marques, et al., 2019). This could be an effect of a relatively higher number of transposable elements within the genomes of the MIG facilitating a form of

recombination (Castagnone-Sereno and Danchin, 2014; Blanc-Mathieu et al., 2017; Kozlowski et al., 2020). Though most often seen following segmentation or duplication events, the dichotomy within *Meloidogyne* genomes - combined with copies present in the hypotriploid portion - increases the likelihood of novel adaptations through neo- and sub-functionalization (Andersson, Jerlström-Hultqvist and Näsvall, 2015). The most apt method to investigate evolution within *Meloidogyne* is through the application of genomic methods, but these require a continuing effort to generate genomic resources such as genome assemblies to contribute to the emerging genomic dataset in the literature. The sequencing, assembly and analysis of an increasing number of species adds to this dataset, extending the scope and power of future studies. The identification of genes important to parasitism or circumvention of host defence, particularly genes that allow the RKN to infect resistant cultivars is also a priority. With better genomic resources of this genus informing more detailed analysis of genomes, evolutionary history and functional adaptations, better methods can be developed to identify, treat, and prevent infection.

1.3 *Meloidogyne haplanaria*

1.3.1 Description

Meloidogyne haplanaria was first isolated in Texas, USA, from peanut plants suspected to be infected with RKNs (Eisenback et al, 2003). Since then, it has been isolated from ash and elm trees, common beans, Indian hawthorn, okra, pea, radish, soybean, and tomato (Bendezu et al, 2004; Ye et al, 2019) and found in Arkansas and Florida (Joseph et al, 2016; Weimin et al, 2019). *Meloidogyne haplanaria* has never been detected or identified outside of the USA.

1.3.2 Agricultural impact

Several cultivars of crop have been engineered to contain genes bestowing resistance to some of the most damaging clade I nematodes. One such gene is the *Mi* gene which confers resistance to cultivars of peanut - ‘nemaTAM’ - and tomato (Simpson et al., 2003). Though ‘nemaTAM’ peanuts have been shown to be resistant to *M. haplanaria*, the same cannot be said of tomatoes, which have been found to be susceptible to *M haplanaria*, revealing this RKN as an emerging and potentially prolific crop pest (Joseph et al., 2016).

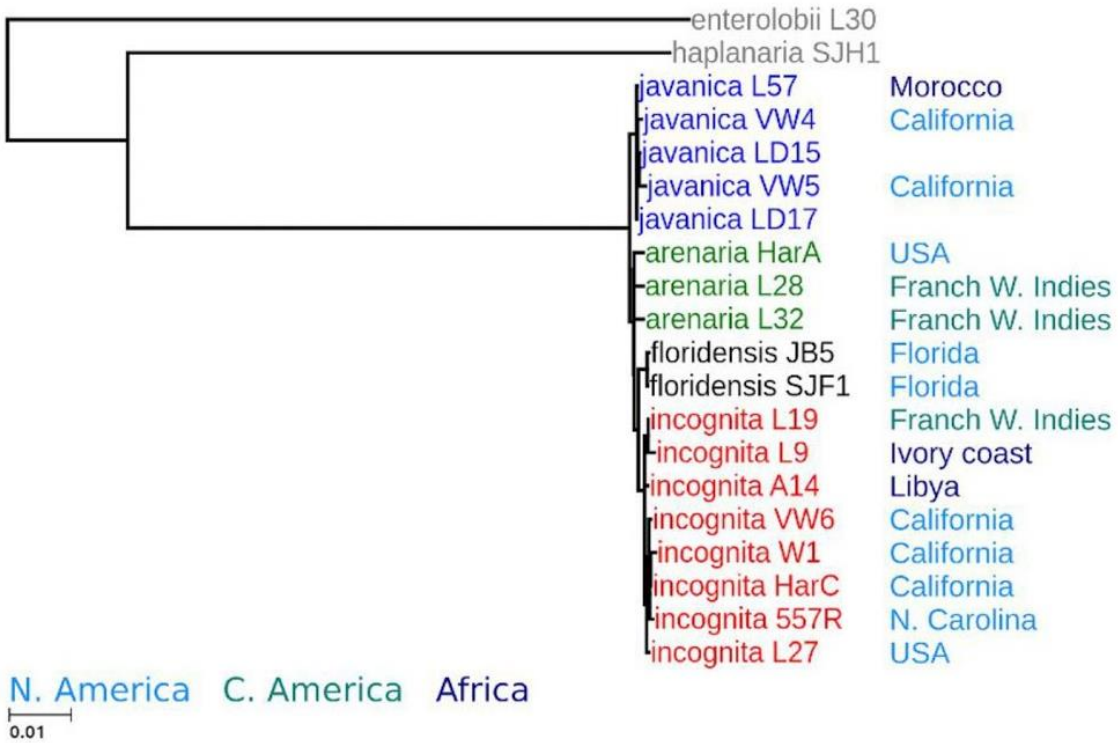

Figure 1.6: Mitochondrial phylogeny of a selection of clade I species (Szitenberg et al. 2017). Tree was created using a concatenation of mitochondrial genes. *M. haplanaria* is positioned distal to *M. enterolobii* and basal to the MIG.

1.3.3 Genomic profile and evolutionary history

M. haplanaria's position in the evolutionary tree of the *Meloidogyne* genus is contested, with some studies using mitochondrial data declaring it a sister taxa to *M. enterolobii* (Joseph et al., 2016; Álvarez-Ortega, Brito and Subbotin, 2019), and other studies using mitochondrial ribosomal data claiming it is sister taxa to the MIG, distal to *M. enterolobii* (Ye, Robbins and Kirkpatrick, 2019). A mitochondrial phylogeny constructed by Szitenberg et al (2017) including concatenated mitochondrial genes of the *M. haplanaria* isolate used here (SJH1) is shown in Figure 1.6. As yet no analysis of the position of *M. haplanaria* has used a phylogenomic approach, nor considered the potential presence of highly divergent homeologs within the genome, a state common across the MIG.

1.4 Research questions

1.4.1 What is the genomic profile of *M. haplanaria?*

Some previous multigene and mitochondrial analyses have placed *M. haplanaria* phylogenetically between *M. enterolobii* and the MIG (Ye, Robbins and Kirkpatrick, 2019). Given this estimation, some proportion of hypotriploidy is to be expected. Does *M. haplanaria* contain a second copy of one homeolog in a hypotriploid arrangement, similar to other Clade I RKNs? Based on its estimated position in Clade I, *M. haplanaria* may also be a hybrid like species in the MIG. This could be possible if the hybridisation event giving rise to the MIG occurred before the divergence of the MIG and *M. haplanaria*, or less likely but still possible, multiple hybridisations occurred in the history of clade I.

1.4.2 Phylogenetic position of *Meloidogyne haplanaria*

Multigene and mitochondrial methods are insufficient to determine the evolutionary origin of potentially divergent homeologs. If both copies form a monophyly in the tree it would indicate that *M. haplanaria* is a hybrid. If the homeologs fall para- or polyphyletically, with one or both homeologs positioned within the MIG, several hypotheses become viable, including that *M. haplanaria* and the MIG could share a one or both parents.

1.4.3 Unknown ancestor of either genomic copy

Previous analyses predicted *M. floridensis*, a clade I automict, as one parent of the MIG (Lunt et al., 2014). This was disputed by Szitenberg (2017) who placed *M. floridensis* within the MIG, obscuring the group's parental origins. *M. haplanaria*'s position in mitochondrial analyses supports it as a potential genomic parent (García and Sánchez-Puerta, 2015; Szitenberg et al., 2017). Detection of a subgenome within a group of the MIG would also indicate this. Or alternatively, the subgenomes of *M. haplanaria* could be paraphyletic, but sharing neither copy with the MIG, instead indicating a *C-D* hybrid system alongside the MIG's *A-B* system.

All of these questions can be investigated or tested using genomic and bioinformatic techniques.

1.5 Reproducible workflows, method rationale, and data origin

1.5.1 Reproducibility

In recent years, a “crisis of reproducibility” has emerged in scientific research, where many past studies have been found to be irreproducible, and in some cases irreplicable (Peng, 2011; McNutt, 2014). This crisis has also affected the computation fields of biology. Efforts have since been made to define a conceptual framework of “research reproducibility”. The key tenets of reproducible research are transparency, replication and corroboration. Transparency in particular can be attained through abiding by the FAIR principles; a set of guiding principles to improve reproducibility (Wilkinson et al., 2016). Transparency involves ensuring the availability of all data used in the study, whether raw or processed, as well as comprehensive lists of all required software, scripts and packages. This can be divided into each principle of FAIR; data must be findable, accessible, interoperable, and reusable (Wilkinson *et al.*, 2016). Replication entails the ability of a second investigator attempting to reproduce the study to perform all necessary analyses with relative ease using code and data provided alongside the study. Corroboration is attained if the replicated analyses produce the same result produced in the study. The gold standard of reproducibility in computational science is said to be provision of executable code, sufficient to replicate all analyses performed within the study, to attain the same results (Peng, 2011). One way to ensure reproducibility is to house computational methods within a workflow or notebook, reducing the effort of the replicating investigator from many hours at a command line to as little as entering a single command, and therefore reducing the potential for operator error to impact reproducibility (Cohen-Boulakia et al., 2017).

Studies investigating the genomics and evolution of the *Meloidogyne* genus often aim for reproducibility, performing analyses at command line or within Jupyter notebooks and providing dependency lists, code, and data. Despite this, studies remain difficult to reproduce; partly due to a lack of automated workflows and standardised bioinformatic resources to work with *Meloidogyne* data in a thoroughly reproducible way, and partly due to code ageing and spawning dependency issues .

Though many protocols for assembly, annotation and phylogenomic analysis exist in the literature, the disconnect between protocol and practise in the absence of an automated workflow manager can introduce ambiguity and inaccuracy, leading to difficulty in

reproduction of methods, and ultimately inability to replicate results. All attempts have been made to ensure the reproducibility of this study with the use of several computational reproducibility practices.

Environment/dependency management

In the past, studies striving for reproducibility have provided written lists of dependencies - programs or software required for the analysis - with version numbers, for replication to be accurate. In cases like this, dependencies must be installed individually one by one, with sub-dependencies also requiring install. Even when records are perfect, this process leads to many lost hours and introduces potential conflicts. To prevent these issues for future reproduction attempts, dependency and version control in this study is managed using *conda* environments (Anaconda Software Distribution, 2016). *Conda* creates compartmentalised job-specific environments wherein all packages and modules are version-controlled and recorded, ensuring that the workflow is resistant to dependency breakdowns and issues arising from incompatibility. As well as this, when *conda* installs a package it automatically installs all dependencies of that package, increasing efficiency and preventing dependency 'breadcrumb trails' (Pflüger, 2019). Using environments in this way increases the transparency and replicability of the methods and the study.

Version control/script hosting

All scripts and code required to perform the methods and analyses done here are hosted on Github (https://github.com/mrmrwinter/), as well as plots and figures created throughout this study, in line with common best practices (Ram, 2013). Reproducible methods such as this greatly increase the transparency of the analysis, while also making replication easier through code availability.

Workflow management

Workflows can automate analyses and increase their replicability. For this study, we used the *snakemake* workflow manager to house and automate methods and analyses *snakemake* - detailed below - allows the assembly and phylogenomic analysis of *Meloidogyne* genomes to be streamlined and monitored transparently, as well as condensing the required command line work to just a few entries. Using a workflow manager such as this ensures the replicability of the study, while also allowing configuration for iterative analyses.

1.5.2 Snakemake workflows

Snakemake is a Python-based workflow manager that allows the automated execution of several bioinformatic processes in succession (Köster and Rahmann, 2012b). A *snakemake* workflow is broken into 'rules', with each rule performing a specific function. The complexity of a workflow can range from a single rule executing a single shell command, to a complex multi-channel pipeline with hundreds of rules, executing processes ranging from shell commands or standalone scripts to interactive html viewers, nested workflows, and bioinformatic packages (Köster and Rahmann, 2012a). This format lends itself well to procedures such as genome assembly, annotation, and phylogenomics, where analyses consist of many steps, using many methods and software.

Reproducibility of analyses is increased greatly by workflow managers. As well as removing room for operator error through automation, the transparency of the method is increased due to localisation of all processes, standardisation of their format, and logging. Corroboration is also more easily attainable when using *snakemake* or other workflow managers due to workflow-wide logging. As well as printing the command line output of each executed rule to file for error reporting and analysis, the housing of all configurable parameters in an external file ensures that configuration for each run or analysis is accurately tracked. *Snakemake* increases the accessibility of bioinformatic analyses like those performed here through removing the skill threshold usually present with computational biology. Very little experience of command line, code, or other abilities formerly vital to the process, is needed, and a workflow can be run by anyone with minimal computer literacy, given a brief set of instructions from the documentation. This democratises the method and makes it much more accessible to non-bioinformaticians, in turn making the replication of studies much easier, particularly when the workflow is version controlled using conda or a similar environment manager. *Snakemake* can call individual environments for individual rules, enabling several versions of a language to be used simultaneously. For more advanced users, the modularity of a *snakemake* workflow enables the addition of custom rules and processes. If the researcher wishes to perform another step not already present in the workflow it can be added easily, after which it will be iterated with each run alongside the original set of rules.

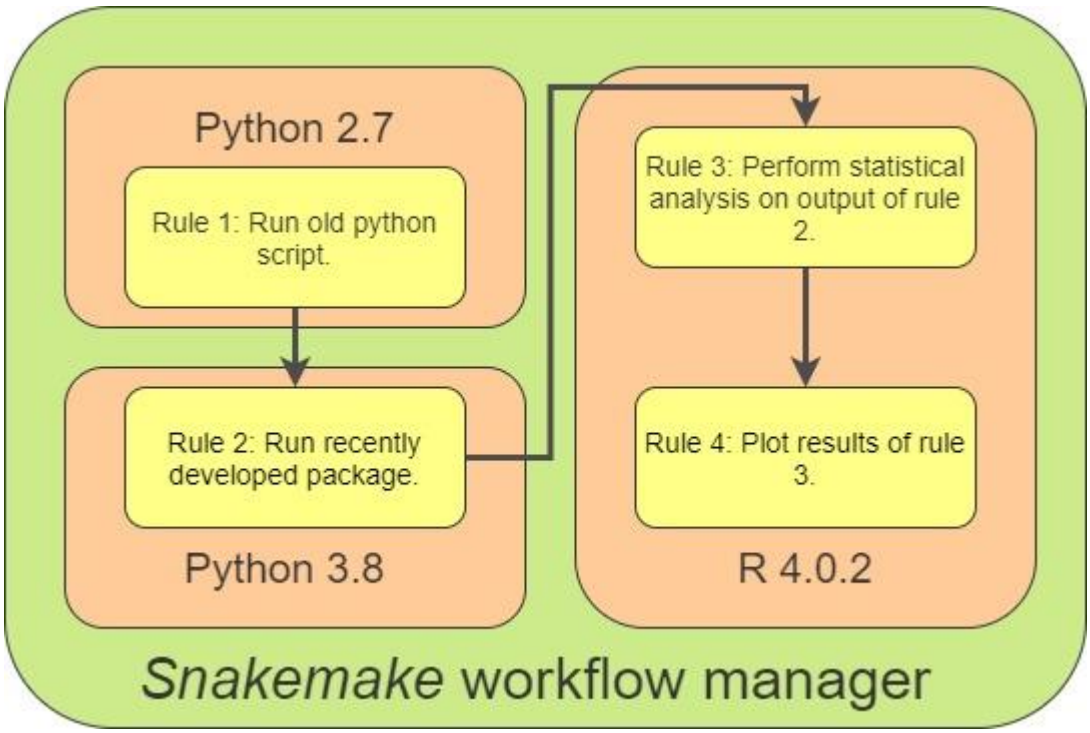

Figure 1.7: Diagram of reproducibility framework. Diagram showing the organisational levels of a reproducible *snakemake* workflow implementing *conda. Snakemake* (green) houses and performs all processes and methods, which are broken down into rules (yellow). Rules are housed in *conda* environments (orange) each running a programming language specific to the requirements of the rule. For example, rule 1 is performed in Python 2, *snakemake* passes the output to rule 2, which runs in a Python 3 environment. The output of rule 2 is passed by *snakemake* into a *conda* environment running *R,* and so forth.

Assembly

A *snakemake* workflow was written to perform the genome assembly methods used in this study. Genome assembly is an iterative process. Different assembly algorithms have varying success depending on the taxa assembled, meaning that an experimental approach is often needed to achieve a relatively high quality assembly. This is easily performed in a *snakemake* workflow, where the output of one rule - i.e., trimmed reads for assembly - can be fed into several rules simultaneously, in this case several different assemblers. This automation combined with ***snakemake'***s parallelisation functionality can greatly accelerate run times and researcher efficiency. Though many genome assembly workflows already exist, none have yet been built with *Meloidogyne* assembly in mind, and creation of a novel workflow ensures transparency and understanding of all stages.

Phylogenomics

A *snakemake* workflow was written to perform phylogenomic analyses, specifically designed to handle hybridised *Meloidogyne* genomes. Phylogenomic analysis is considered a difficult process, even by phylogeneticists and bioinformaticians. Housing the workflow in a manager such as *snakemake* democratises the analysis, increasing the accessibility of the method and the field to non-bioinformaticians (Perkel, 2017; Young and Gillung, 2019; Lawlor and Sleator, 2020).

The workflow runs a complete phylogenomic analysis from start - CDS input data - to finish - plotted figures and summary statistics. As with the assembly workflow, all parameters of all packages and scripts within the workflow are configurable, many of which are managed through a single configuration file. As a minimum, configuration of this file and provision of input data is all required to proceed from CDS genome annotations to a finished phylogenomic analysis. Several plotted trees are outputted at varying stages of the workflow to allow visualisation of the effect of configuration at each stage.

1.5.3 Rationale of chosen methods

The chosen assembly method closely follows the methods used by Szitenberg et al. (2017) with some minor updates and changes depending on availability of more up to date software and specialist requirements of this study. This adapted workflow is detailed in Chapter 2, Section 2.2.2, and automated in a *snakemake* workflow (Chapter 1, Section 1.5.2).

The prediction based process of genome annotation lends itself to errors, and as such no single annotation software can reliably predict all of the genes present in a genome. For this reason, annotation workflows, or suites, are often used (Salzberg, 2019). Suites combine several annotation software and methods into a single workflow that annotates the genome iteratively, using previous annotations as templates for other software. One such annotation suite is *MAKER2,* which will be used in this study (Holt and Yandell, 2011). A detailed description of the methods can be found in Chapter 2, Section 2.2.3. *MAKER2* is widely employed within the literature as the preferred method of annotation, as has been previously used to annotate *Meloidogyne* genomes (Szitenberg et al., 2017; International Helminth Genomes Consortium, 2019).

Similarly to the genome assembly rationale and method, this study follows roughly the same orthology and phylogenomic method applied in Szitenberg et al. (2017) with some modular changes based on software availability and post-publication releases. For instance, standalone python scripts and packages were updated or standardised to run in a Python 3 environment. Full details of orthology definition and phylogenomics methods can be found in Chapter 3, Section 3.2.

1.5.4 Origins of data

Other than sequence data of *Meloidogyne haplanaria*, all data used in this study is publicly accessible and available. *Meloidogyne haplanaria* data (Isolate: SBF1) (Joseph et al. 2016) was produced from DNA isolated from a specimen infecting a tomato plant with the *Mi*-1 resistance gene, cultivated in captivity in Florida. DNA was extracted from an egg mass and sequenced using Illumina HiSeq chemistry, producing two read libraries. MIG and *M. enterolobii* genome assemblies, as well as *M. incognita* CDS data, were retrieved from GenBank (Supplementary table 3). The *M. hapla* assembly was retrieved from WormBase.

1.6 Statement of intent

Using these newly created workflows, alongside Jupyter workbooks and a proprietary annotation workflow, we will assemble and annotate the genome of *Meloidogyne haplanaria*, collect comparative statistics and compare it against other genomes of Clade I RKNs. The genome will be profiled, and an inter-species sequence similarity analysis performed to detect signals of hybridisation and ploidy. Following this, we will use the resulting annotations to create a phylogenomic tree of the *Meloidogyne* genus, treating each divergent homeolog as an individual taxonomic unit. This analysis will allow inference of the phylogenetic position of each homeolog of *M. haplanaria*.

Alongside providing a novel annotated assembly for use by the wider community and providing insight into *Meloidogyne*'s evolutionary history, we will produce and provide *snakemake* workflows for assembly and phylogenomic analysis of *Meloidogyne* genomes, making repeating and performing these types of analyses easier and more accessible.

Chapter 2: Genome assembly and annotation of Meloidogyne haplanaria

2.1 Introduction

Here we assemble and annotate the genome of *Meloidogyne haplanaria* and perform an intra-genomic sequence similarity analysis to detect signatures of divergent genomic copies. The resources generated here will enable comparative and phylogenomic analysis of the species.

Advances in modern sequencing technologies have enabled researchers to analyse and observe biological characteristics of organisms at a whole-genome scale (Bickel et al., 2009; Green, Rubin and Olson, 2017; Mardis, 2017). This has spurred a huge increase in the scope and statistical power of many genomic analyses techniques, and created new fields in the form of, among others, comparative genomics, phylogenomics, and bioinformatics (Horner et al., 2010; Young and Gillung, 2019). However, all downstream analyses rely on the prerequisite steps of genome assembly and annotation (Dominguez Del Angel et al., 2018).

Genome assembly is performed within a novel *snakemake* workflow to enable easy replicability and iteration of methods.

2.1.1 Genome assembly and annotation

Genome assembly

Genome assembly is the process of taking sequenced reads of your target organism(s) and using bioinformatic methods to arrange them into contiguous fragments, attempting to approach as high a level of completeness as possible. Different sequencing technologies yield greater assembly success, short-read technologies such as Illumina make fragmented assemblies with a low error rate, whereas long reads such as Hi-C, NanoPore, and PacBio yield the most complete assemblies, though with a higher rate of error (Bradnam et al., 2013; Richards, 2018). Assembly generally consists of several necessary steps; quality control of the raw sequencing data, cleaning and trimming of that data, assembly into sequences and contigs, and appraisal and validation of the final

assembly. Quality control is performed on the raw sequence data to remove reads that are deemed low quality. Summary statistics of the raw data are usually collected at this time, providing quantitative information on read count, length, insert size, GC content, and coverage, among others. This stage usually includes the trimming of sequencing adapters, and filters for lingering primer sequences. Short reads sequenced with paired end chemistry are often merged at this stage. Following quality control, the data is trimmed and screened for contaminants. If not already performed, adapter and primer sequences are trimmed from the reads. Reads may be trimmed here based on preference, often to a specified length or sliding window coverage threshold. This is followed by contaminant screening. Reads are assigned taxonomy based on sequence similarity to a database and reads that fall clearly outside of expectation are designated as contaminants and removed. Surviving reads are then passed into an assembly algorithm or software (Kitts, 2002; Pop, 2009). Many such assemblers exist, using different methods or algorithms and performing differently with varying taxa, but the central process is the same. Reads are broken into *k*-mers; sequences of bases of length *k*. *K*-mers are then arranged using one of several methods into increasingly long sequences (Chikhi and Medvedev, 2014).

Often, many different quality control and trimming measures, as well as several different assemblers, will be used to produce several assemblies of varying quality and characteristics. These final assemblies are appraised for quantitative statistics and validated against each other, after which a single assembly is usually chosen to progress to annotation (Bradnam et al., 2013; Gurevich et al., 2013). One such statistic is N50, a measure of the shortest contig length required to cover half of the genome. This measurement is often used as an indication of completeness and quality of assembly prior to gene annotation, though modern approaches use other metrics alongside it (Castro and Ng, 2017).

Genome annotation

Genome annotation is the process of detecting gene content and location in a genome assembly based on supplied structural or sequence information of those genes and assigning ontology to them (Campbell and Yandell, 2015). This is an oversimplified explanation, as genome annotation has been, and continues to be, a difficult process (Salzberg, 2019). An annotation workflow typically runs through several iterations of gene prediction, applying a different software or algorithm at each stage. One method of

prediction is through sequence similarity. Coding sequences (CDS) or mRNA from the same or a closely related species are used as query and hits scoring above a predefined threshold are annotated as genes (Koonin and Galperin, 2003b; Harrison, 2014).

Another method is *ab initio* gene prediction, wherein genes are discovered through detection of markers inferred from supplied CDS data. These markers can be structural identifiers of protein-coding genes, such as start/stop codons, or stretches of sequence that do not conform to predicted values. While relatively straightforward in prokaryotic organisms, differential splicing present in eukaryotes makes *ab initio* gene prediction challenging, and resulting annotations are much more sensitive to erroneous predictions. Repeats in the genome - microsatellites or transposable elements - can have a strong impact on the success of gene prediction algorithms (Tarailo-Graovac and Chen, 2009). It is therefore also necessary to screen the assembly prior to annotation to detect and mask these regions. Once an organism's genome has been assembled and annotated it can provide many insights into its biology. It can be compared against its contemporaries and used to infer structural variation, transposable element activity, evolutionary history, and more (Salzberg, 2019).

2.1.2 *Meloidogyne* genomics

Root-knot nematodes (RKN) (Genus: *Meloidogyne*) are a genus of plant parasitic nematodes that infect the roots of plants, causing gall formation and reduced host fitness and yield. Many billions of dollars are spent on the control of *Meloidogyne* infection each year, and the genus has been labelled as the most scientifically and economically important species of plant parasitic nematode (Jones et al, 2013; Bernard, Egnin and Bonsi, 2017b). Within the *Meloidogyne* genus there are well defined clades (Chapter 1, Figure 1.4), of which Clade I will be most pertinent to this study and will be discussed primarily here. Clade I contains the tropical root-knot nematodes, including the infamous *Meloidogyne incognita* group (MIG). The MIG includes several notoriously damaging crop pests; *M. incognita, M. javanica,* and *M. arenaria.*

Genomic resources of the *Meloidogyne* genus, such as transcriptomes, assemblies, and annotations, are being generated and made available at an accelerating rate, but this has not always been the case and there is still a scarcity limiting the scope of comparative and phylogenomic analyses.

There are 17 *Meloidogyne* genome assemblies currently hosted on Genbank, all but two of which are species from clade I. Clade I *Meloidogyne* assemblies generated with short read sequencing typically consist of tens of thousands of contigs, with an N50 around the same number. while long read assemblies, such as the recently published *Meloidogyne luci* genome, contain a few hundred contigs, with an N50 over a million. Novel *Meloidogyne* genomes are improving with the advancement of sequencing technologies. For example, a short read *M. enterolobii* assembly from 2017 (Szitenberg et al. 2017) contained 42,008 and an N50 of 9279. A long read assembly of the same species from 2020 (INRAE) contains 4437 contigs and the N50 equals 143,330.

Several characteristics of this genus's biology complicate otherwise relatively straightforward processes. Species of the MIG contain two divergent genomic copies or subgenomes, *A* and *B* (Chapter 1, Figure 1.5). First supported by Lunt et al. (2014) and confirmed by Szitenberg et al. (2017), a recent hybridisation event in their shared history, presumably between two sexually reproducing species, bestowed them with two divergent genomic copies. Szitenberg et al (2017) generated phylogenomic trees of the Clade I *Meloidogyne* species wherein each divergent genomic copy is treated as an individual operational taxonomic unit (OTU). The result of this was a tree where the MIG diverges into two clear groups, each containing a representative from each species, *A* or *B.* These copies exhibit average heterozygosity of around 3%, making contiguous assemblies difficult (Pryszcz and Gabaldón, 2016).

Evidence suggests that as well as being hybrids, some Clade I *Meloidogyne* species are also hypotriploid; they contain a portion of one divergent genomic copy (*A*) present in a second copy (A^1) (Chapter 1, Figure 1.5). This proportion differs in size across the MIG (Triantaphyllou, 1991; Lunt et al., 2014; Szitenberg et al., 2017). It is thought to be a remnant of past hybridisation events between diploid species, that has gradually decomposed due to segmental loss and gene conversion.

Genome assemblies typically comprise one, potentially fragmented, haploid strand which is representative of the consensus of bases across the genome. In most genome assembly algorithms, highly heterozygous regions are appended to the assembly as if it were a different region of the same strand, rather than being collapsed based on consensus base calling and coverage. This is a problem common to *Meloidogyne* assemblies, leading to overestimation of genome size. Algorithms that predict genome size based on *k*-mer coverage of reads provide estimates up to a third smaller than

resulting assemblies produced with traditional assembly methods (Ranallo-Benavidez, Jaron and Schatz, 2020). While this can be detrimental to many comparative analyses and can be collapsed purposely if needed, analyses relying on downstream gene annotation can benefit from the included gene content.

All of the above means that despite this genus being very important (Jones et al, 2013), as well as biologically interesting regarding genome composition, studies into their evolutionary history or structural and genetic variation are only now beginning to sufficiently resolve true inter-species relationships and adaptive mechanisms thanks to the increasing availability of *Meloidogyne* genome assemblies (Lunt et al., 2014; Szitenberg et al., 2017).

2.1.3 *Meloidogyne haplanaria*

Meloidogyne haplanaria, the Texas peanut root-knot nematode, is endemic to the USA and has been isolated from Arkansas, Texas, and Florida, where they infect, most notably, tomatoes and peanuts, as well as many other species of plant from several orders (Bendezu, Morgan and Starr, 2004; Joseph et al., 2016; Ye, Robbins and Kirkpatrick, 2019). There is currently no existing genome assembly, transcriptome, or annotation available for *M. haplanaria*. Whole genome sequence data exists, generated by Joseph et al. (2016) with Illumina HiSeq chemistry. Very little is known of the genomics of *M. haplanaria*, and it has been studied only in passing until recently (Joseph et al., 2016). Many papers investigating *Meloidogyne* genomics make mention of it, but it is often not the focus of their analyses (Szitenberg et al., 2017; Álvarez-Ortega, Brito and Subbotin, 2019). Previous analyses have focused primarily on morphology, biochemistry, and small sets of genes (Eisenback et al., 2003; Joseph et al., 2016; Ye, Robbins and Kirkpatrick, 2019).

As well as little previous insight into genomic composition, several other seemingly open gaps in the literature may also complicate the investigation. To begin, the sexual system of *M. haplanaria*, and by extension molecular rate of evolution, is yet unknown. While this should not affect the assembly process itself, it does affect the confidence of predicted outcomes. Apomictic species typically have much larger genomes and almost double the number of predicted genes than automictic species (Szitenberg et al. 2017). It has been suggested by some groups that apomictic *Meloidogyne* species have a large amount of

highly active transposable elements (Kozlowski et al., 2020). If this is true, and *M. haplanaria* is automictic, this heightened repeat content could impair the annotation process, and must be assessed at the time and dealt with carefully.

Phylogenetic mitochondrial and ribosomal analysis either places *M. haplanaria* as a sister taxa to the MIG, with *M. enterolobii* as outgroup (Ye et al. 2019), or groups *M haplanaria* and *M. enterolobii* together as a monophylum (Álvarez-Ortega, Brito and Subbotin, 2019; Ye, Robbins and Kirkpatrick, 2019).

2.1.4 Aims of the study

In this study we will assemble the genome of *M. haplanaria* using a novel *snakemake* workflow, employing several different assembly methods and aiming for the most contiguous assembly possible. All resulting assemblies will be scored using currently accepted metrics of quality, and appraised based on comparative statistics. This assembly will then be annotated, with genes being predicted using several iterative methods, and final CDS will be extracted. Comparative statistics will be gathered about the quality of each annotation stage. To detect ploidy state we apply a technique first developed by (Lunt et al., 2014), wherein coding sequences are aligned against a novel database containing all coding sequences in the assembly and the percentage identity of the second top hit is plotted, revealing the presence of multiple gene copies.

2.2 Methods

2.2.1 Reproducibility

All scripts and workflows are made available, and all attempts have been made to ensure reproducibility. See https://github.com/mrmrwinter. Genome annotation and intragenomic blast were conducted within Jupyter notebooks.

All stages of the genome assembly were performed in a novel *snakemake* workflow (Chapter 1, Section 1.5.2) written specifically for the *Meloidogyne* genus, available here: https://github.com/mrmrwinter/Meloidogyne-assembly-workflow/. The directory structure is standardised according to recommended best practice. Files to create *conda*

environments are kept in the *envs/* directory. All scripts used by the workflow are contained in the *scripts/* directory. Raw data and input data, though identical, are stored in separate directories within the *data/* directory to prevent the workflow affecting the raw data. Another distinguishment is made between results and output. Output - *output/* - contains all intermediary data from the workflow, including contamination, quality, and mapping outputs, whereas *results/* contains final workflow output, such as assemblies and tables of results or comparative statistics. The *reports/* folder contains logs of package performance and rule checking outputs.

The workflow requires only raw FASTQ data and brief configuration to reproduce assembly stages and summary analyses. Every stage is performed and controlled by the *snakemake* workflow according to configurable input parameters.

2.2.2 Assembly

Two raw, paired end libraries (here, 14 and 58) of Illumina HiSeq reads of a single *M. haplanaria* sample (Isolate: SJH1) isolated from tomato with the *Mi* resistance gene (Joseph et al., 2016). This isolate was initially sequenced in the same project as that generating the data in Szitenberg et al (2017), but the nuclear genome was not assembled or analysed. Data was provided in FASTQ format and processed in a *snakemake* assembly workflow (Chapter 1, Section 1.5.2).

Quality control and trimming

Reads were run into *FastQC* (Andrews and Others, 2010) and *fastp* (Chen et al., 2018) to generate quality assessments and summary statistics, and detect presence of adapters and primers in raw input libraries. Reads were then run into *Trimmomatic* (Bolger, Lohse and Usadel, 2014). *Trimmomatic* removes bases and reads that do not meet predefined quality threshold. *Trimmomatic* module *ILLUMINACLIP* was also used to remove Illumina TruSeq adapters from the data. For this analysis, we used a sliding window of 5bp with a quality of 20x, trimmed the leading end by 5bp and the trailing end by 10bp, and set a minimum length of 30bp. Following the trimming stage, reads were again run through *FastQC* and *fastp* to ensure adapter removal and collect post-trimming quality scores. Reports and plots of quality were collected and collated in Table 2.1 and Figure 2.3.

Genome profiling and characterisation

Following quality control and trimming, reads were analysed in the characterisation stage of the workflow. First, *KMC* (Kokot, Dlugosz and Deorowicz, 2017) is run on the data to generate *k*-mer counts. *K*-mer counts are then inputted into *Genomescope 2* (Ranallo-Benavidez, Jaron and Schatz, 2020). *Genomescope 2* uses k-mer frequencies to detect probable ploidy of the sequenced organism, as well as genome size, repeat content, and heterozygosity (Figure 2.4a-d; Table 2.2). A plot is created displaying these values. Alongside *Genomescope 2,* reads were run through *smudgeplot* (Ranallo-Benavidez, Jaron and Schatz, 2020). *Smudgeplot* extracts heterozygous k-mer pairs and uses the ratio of their coverage to frequency to predict ploidy. This process also outputs a plot for interpretation (Figure 2.5).

Pre-assembly

After quality filtering and trimming, reads are assembled into contigs through a process of pre-assembly. These pre-assembled contigs can then be used to calculate overall read depth and as flags for contamination removal. Pre-assembly was performed using *SPAdes* (Bankevich et al., 2012) with default settings. Details of assembly *spades* can be found in Table 2.3.

Mapping

The pre-assembly is then used as a reference to map trimmed reads back to. This allows calculation of coverage for individual regions. Mapping was performed using *bwa-mem* (Li and Durbin, 2009) and indexing with *samtools* (Li et al., 2009), both housed in script *mapping.smk.*

Contaminant detection and removal

Pre-assembled contigs were used as individual queries in sequence similarity searches using *BLAST* (Altschul et al., 1990). The reasoning for this is that once pre-assembled contigs have been assigned taxonomy, this can be referenced against coverage of those contigs to detect contamination, and the corresponding reads removed from the final assembly stages. Contamination detection was performed by *blobtools* (Laetsch and Blaxter, 2017). *Blobtools* removes contaminant contigs that may impair or falsely improve the final assembly. It does this by cross-referencing taxonomy, coverage and pre-assembled contigs. It then creates plots of coverage over GC% content and colours data points by phylum, as well as printing a table with this information. Using this table,

c*leaning.smk* creates a list of all unassigned reads and reads assigned to Nematoda with coverage over 20x. Reads not in this list are designated contaminants are removed from the trimmed read libraries using *samtools* (Li et al., 2009). Full process housed in *contaminants.smk* and *cleaning.smk.*

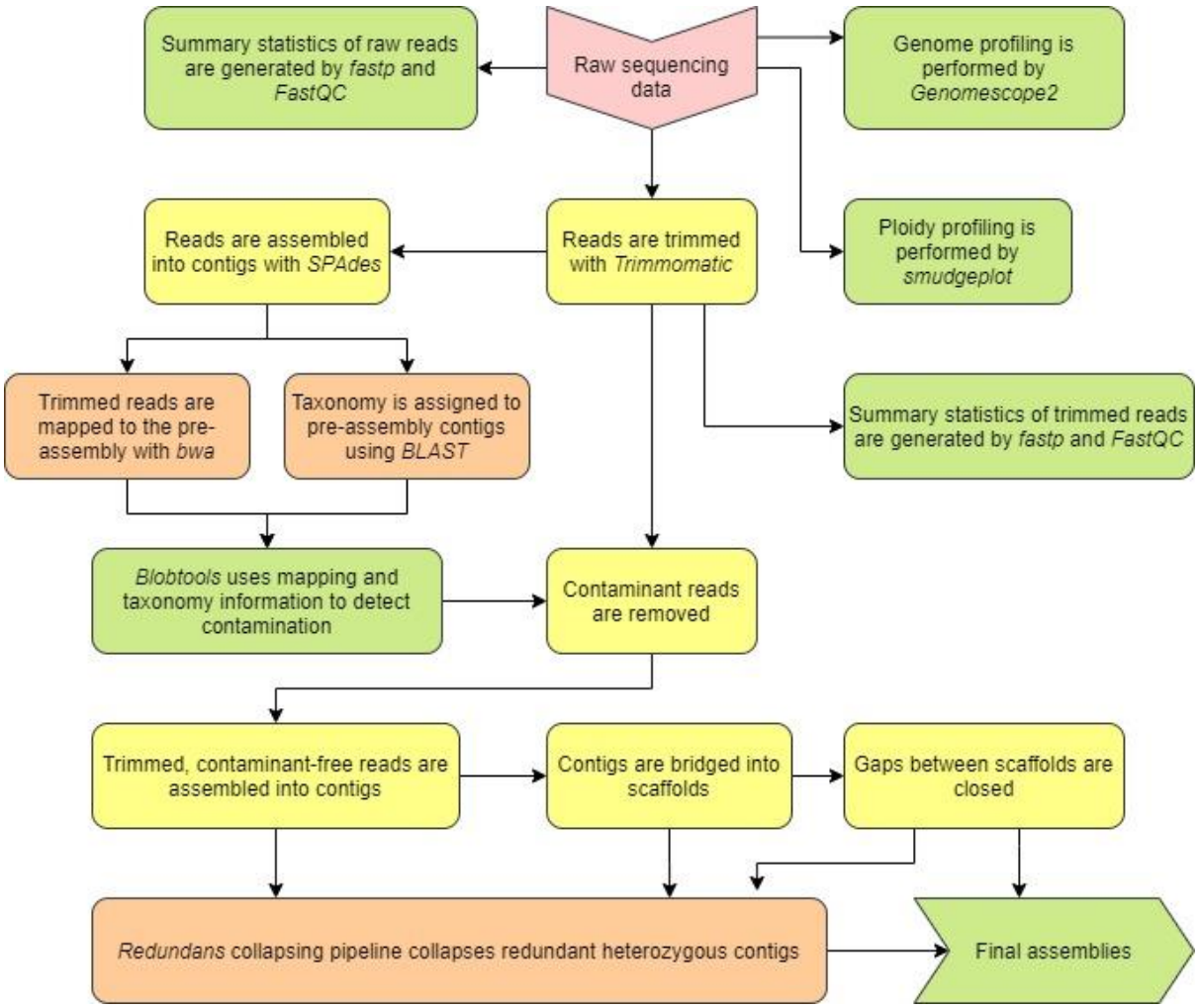

Figure 2.1: Flowchart showing the assembly process. Red is input data, yellow is a software stage, orange is a data transformation, green denotes a stage has an outputted result.

Assembly

Several assemblies were produced from these methods, each using a slightly different assembly process. First, trimmed filtered reads were run into the *platanus* workflow (Kajitani et al., 2014). Reads were assembled into contigs using *platanus assemble* with default parameters. Resulting contigs were then entered into *platanus scaffold* alongside

both read pairs of both trimmed libraries to scaffold contigs together. Outputted scaffolds were then gap closed with *platanus gap closer*, again using both read pairs from both trimmed libraries., resulting in three assemblies; *platanus_assemble, platanus_scaffold*, and *platanus_gapClosed*. Assemblies *spades*, *platanus_scaffold*, and *platanus_gapClosed* were then run through the *redundans* pipeline. *Redundans* is a gap reduction pipeline designed for highly heterozygous genomes (Pryszcz and Gabaldón, 2016). The *redundans* pipeline includes *platanus* as its first stage but it was opted for to run it separately in this study to have more control over the assembly process. Assembly scripts are housed in *platanus.smk* and *redundans.smk.*

Appraisal

To attain the highest quality of annotation and downstream phylogenomic resolution as possible it is necessary to ensure we continue with the best available assembly. All assemblies were appraised by *BUSCO* for completeness (Table 2.3) (Simão et al., 2015). *BUSCO* is a genome appraisal and gene prediction software that operates on the detection of universal single copy orthologs. Assemblies were also run through *QUAST*; an assembly comparison and validation software (Table 2.3) (Gurevich et al., 2013). A comparison of the chosen *M. haplanaria* assembly against other *Meloidogyne* assemblies can be found in Table 2.4. Appraisal scripts are all housed in *appraisal.smk.*

2.2.3 Annotation

QUAST and *BUSCO* results indicated that the assembly created and gap closed with platanus, then collapsed with redundans (*platanus gapClosed* + *redundans*), was the highest quality and most contiguous produced, and that it should be used to generate annotation. The *MAKER2* pipeline was chosen based on its wide adoption and validated methods (Holt and Yandell, 2011). *MAKER2* supports multiple sequence homology and *ab initio* prediction tools and runs the data consecutively through each. Each *MAKER2* run is configured in the control file *maker_opts.ctl*, parameters of which are flagged at each stage.

Repeat masking

RepeatModeller (Smit, Hubley and Green, 2015b) was run with the *M. haplanaria* assembly as input. This creates a database of repeats detected within the genome. This was flagged as a repeat library in *maker_opts.ctl* and used as input for *RepeatMasker* (Smit, Hubley and Green, 2015a). Using this database and an inbuilt minimal RepBase

(Bao, Kojima and Kohany, 2015) repeat database, *RepeatMasker* masks repeat content in the genome.

Sequence homology inference

The first round of annotation with *MAKER2* uses sequence homology based method, *EST2GENOME. MAKER2* is run with *est2genome* and *protein2genome* switched on and supplied with nucleotide CDS and protein sequences from *M. incognita* (Bioproject: PRJNA340324). The output was run through *GAAS* to generate summary statistics and annotation information was extracted and merged before being validated by *fathom* (Korf, 2004; Dainat et al., 2020). Erroneous predictions were removed, and collated annotations were converted into a *.hmm* file.

***Ab initio* prediction**

The *.hmm* annotation file is then passed through *MAKER2* with *SNAP* flagged on. *SNAP* is an *ab initio* gene predictor that produces predictions of gene models (Korf, 2004). After *SNAP* has finished, the resulting annotation data was again validated with *fathom,* screened for errors, and converted into a *.hmm* file*. SNAP* was run again with this new file as input. This process was performed three times, with results of the previous step becoming input for the next. In this way, *SNAP* can improve upon each prediction using its own predictions as training. Each of these annotation sets was parsed by *GAAS* for comparative statistics (Table 2.5; Supplementary table 2).

Based on summary statistics, the output of the second run of *SNAP* prediction was deemed the most successful and data rich. This output was converted into a GenBank file and split into two sets, test and training, using *randomSplit.pl*. The training set was then used to train *Augustus*, another gene predictor package (Hoff and Stanke, 2019). Once finished, *Augustus* is optimised using *optimise_augustus.pl.* Following this, *Augustus* was trained using these optimised parameters. *MAKER2* was then run with *Augustus* flagged on in *makeropts.ctl*, using the test set as input. The resulting annotation was merged and parsed by *GAAS*, to attain summary statistics (Table 2.5; Supplementary table 2). CDS annotations were identified, and their sequences written to a FASTA file using *AGAT* (Dainat and Hereñú, 2020).

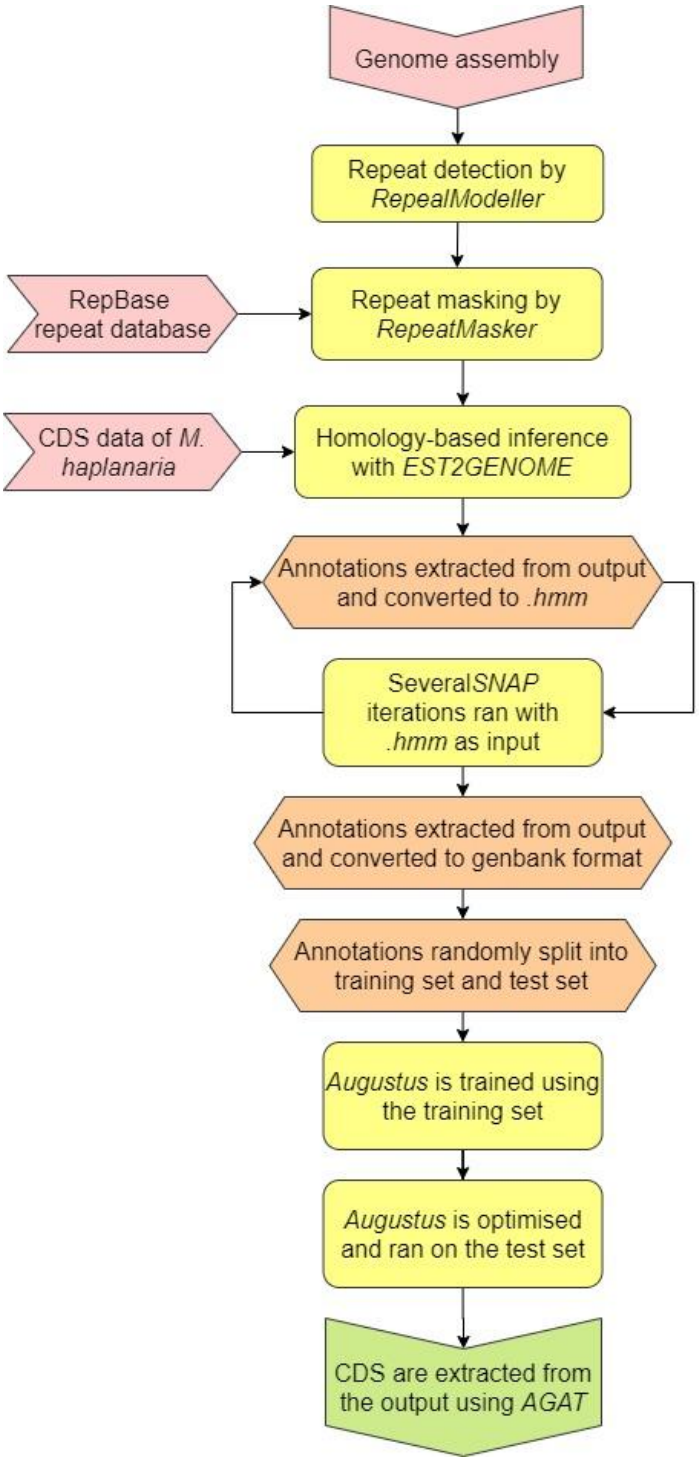

Figure 2.2: Flowchart displaying the annotation process, from genome assembly to extracted coding sequences. Red arrows are input data, yellow rectangles are software stages, orange hexagons are data transformations, green arrows are output.

2.2.4 Intragenomic sequence similarity analysis

An intragenomic sequence similarity analysis, hereafter intragenomic blast, was performed on all CDS extracted from the *M. haplanaria* assembly, as well as CDS sequences from *M. incognita* W1, *M. javanica* VW4, *M. arenaria* HarA, *M. floridensis* SJF1, *M. enterolobii* L30, and *M. hapla* PRJNA29083. For each species, a *BLAST* database is created containing all available CDS of that species. *Blastn* searches the database using a local alignment algorithm, detecting sequences that exhibit similarity to the query. These 'hits' are then extracted to a table and assigning values such as sequence similarity or length. In an intragenomic blast analysis, the second highest percent identity hits of each query are collated in a table, excluding genes where top hits were not one-hundred percent. Hits over 99.5% sequence identity were removed, and the list for each species was plotted in a histogram (Figure 2.8). A Jupyter notebook was written to perform this analysis in a controlled, reproducible way, available at: https://github.com/mrmrwinter/igb.

2.3 Results

2.3.1 Quality control and trimming

Results of trimming can be found in Table 2.1. Of note, the percentage of bases with coverage over 30x rose by over 9% in both libraries, to over 92% in both. 12.7% of reads were dropped from library 14 and 11.3% of reads were dropped from library 58.

Table 2.1: Quantitative statistics of both *Meloidogyne haplanaria* libraries before and after trimming.

Library	14 before	14 after	58 before	58 after
Insert size peak	269	150	169	150
Total reads (M)	192.192	177.345	160.364	143.890
Total bases (Gbp)	28.829	22.529	24.055	19.188
Q20 bases (Gbp)	25.804	22.033	22.239	18.781
Q30 bases (Gbp)	23.385	20.885	20.385	17.915
Percent of bases Q30 (%)	82.73	92.70	84.75	93.36
GC content (%)	31.7	31.1	31.5	30.6
Read pairs	96,096,276	62,757,549	80,182,026	48,356,913
Mean length (bp)	150, 150	136, 117	150, 150	141, 125
Duplication rate (%)	3.86	3.99	4.16	4.33
Percent of original read pairs (%)	100.00	65.31	100.00	60.31
Dropped reads	-	12,191,203	-	9,034,262
Dropped reads (%)	-	12.69	-	11.27

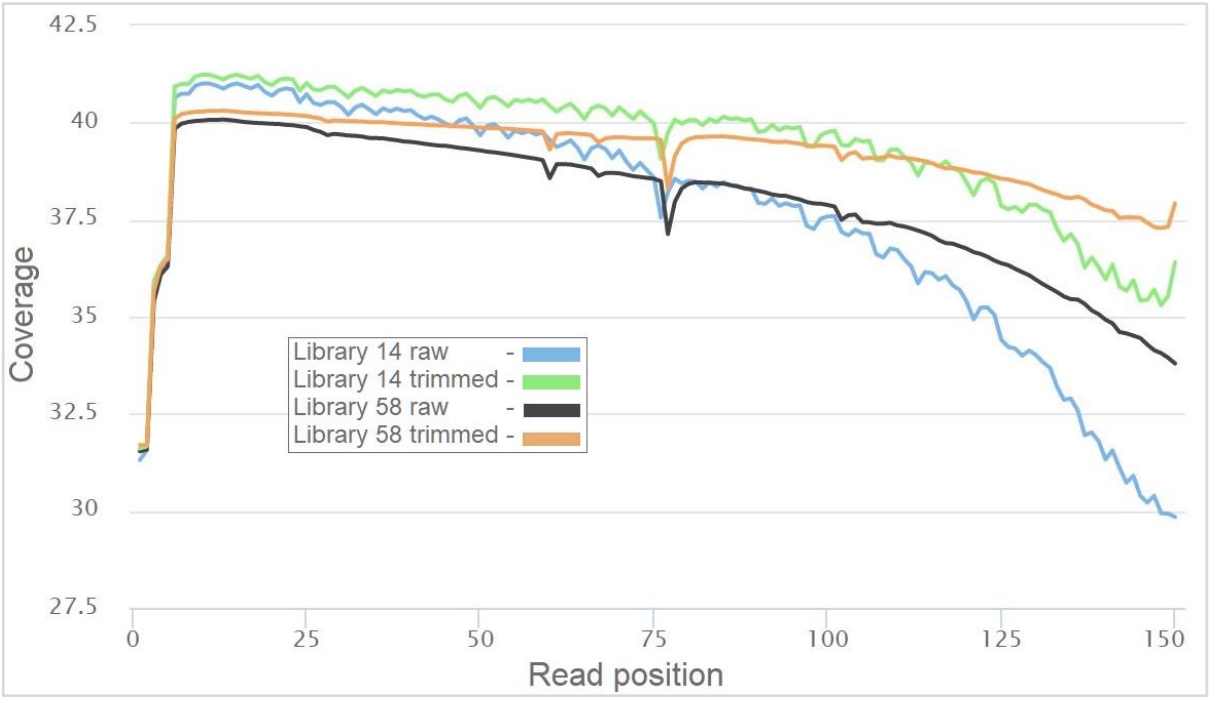

Figure 2.3: Trimming coverage against raw coverage. Blue and green are library 14 raw and trimmed, respectively. Grey and orange and library 58 raw and trimmed, respectively. Average coverage of total reads increased after trimming.

2.3.2 Characterisation and profiling

Genomescope2 successfully converged (Figure 2.4a-d). Summary statistics can be found in Table 2.2. Genome size was predicted as 72,685,528 base pairs. Heterozygosity was estimated at 2.46%. The shape of the distribution (blue) indicates that the data fits a model of triploidy (black) (Figures 2.4a & c).

Table 2.2: Results of *Genomescope2* analysis.

Ploidy model	Genome size (bp)	Hetero-zygosity (%)	Unique k-mers (%)	Ploidy in each k-mer configuration (%)			Error (%)	Duplicate percentage (%)	K-mer size (bp)
				AAA	AAB	ABC			
Triploid	72,685,528	2.46	70.3	97.5	2.46	0.0419	0.937	3.2	21

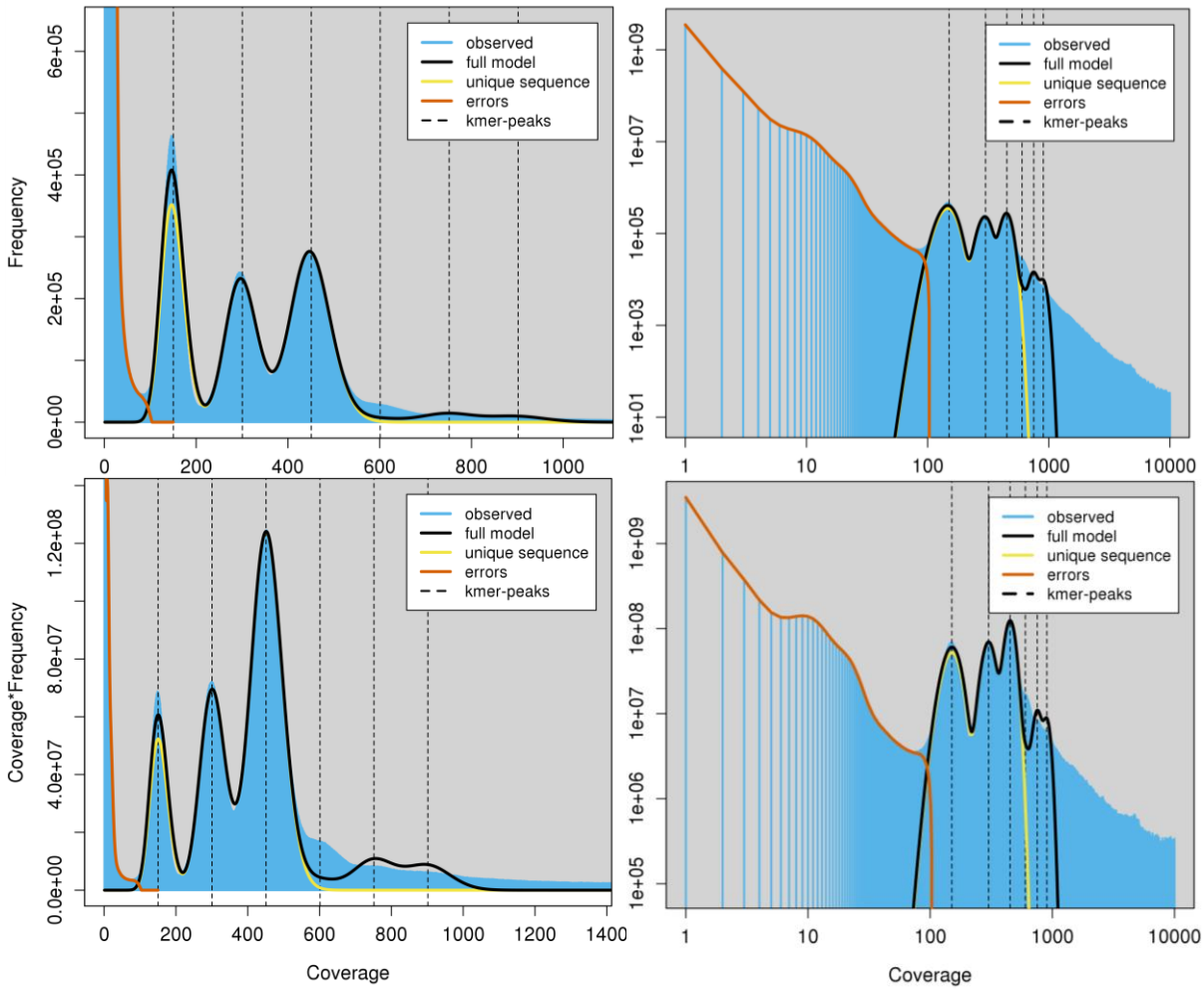

Figure 2.4a-d: *Genomescope2* plots of k-mer coverage of combined libraries. Plots showing results of *Genomoscope2* analysis. The shape of the distribution indicates that the data fits a model of triploidy.

Smudgeplot ran successfully and produced two plots (Figure 2.5a-b). From these plots *smudgeplot* predicts that *M. haplanaria* is primarily triploid; 64% of *k*-mer pairs followed a triploid distribution. However, a small peak in diploidy - 11% of *k*-mer pairs - could be a signature of hypotriploidy. Due to the low quality of short read data around 25% of *k*-mers did not fall within the boundaries of a ploidy estimation, likely due to low representation in the dataset.

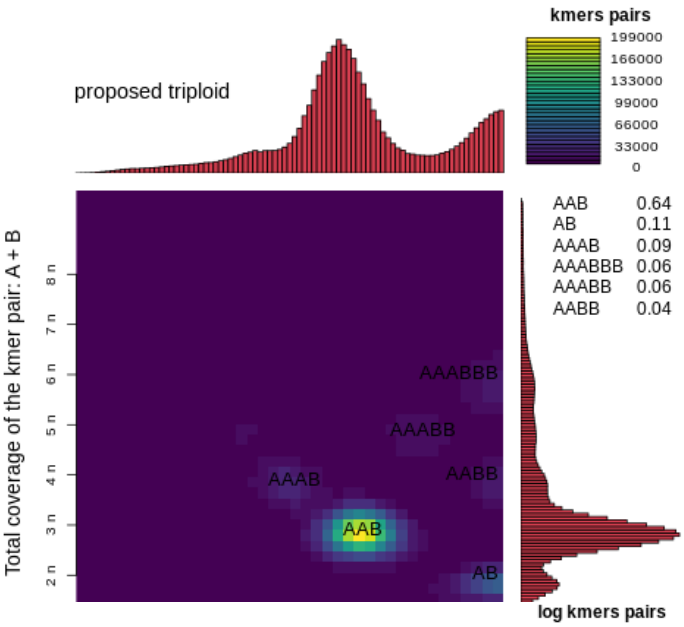

Figure 2.5a: Smudgeplot displaying *k*-mer distribution.

Smudgeplot showing ploidy estimations of combined libraries based on *k*-mer distributions. Yellow hue indicates strength of signal. The only strong signal in the heatmap is under AAB, providing strong evidence indicating triploidy (64% of *k*-mer pairs).

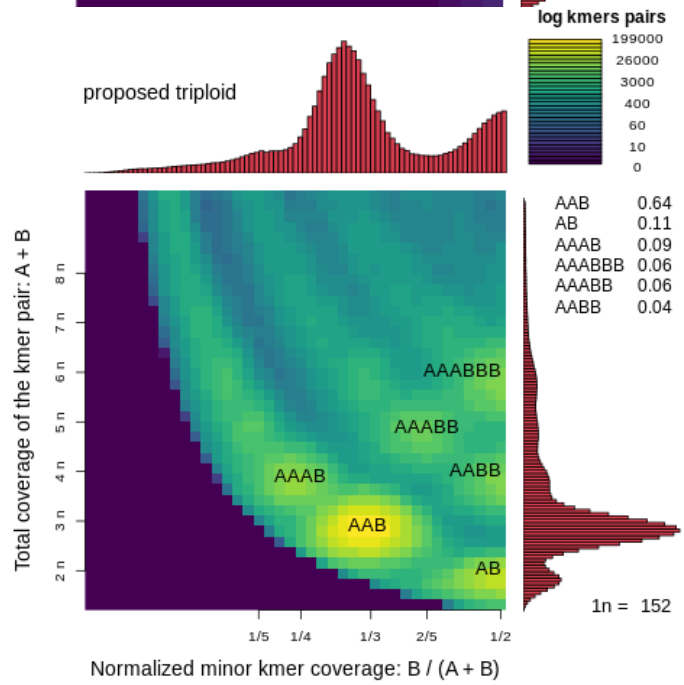

Figure 2.5b: Smudgeplot displaying logarithmic *k*-mer distribution.

Smudgeplot showing ploidy estimations of combined libraries based on logarithmic *k*-mer distributions. Yellow hue indicates strength of signal. The strongest signal is under AAB, indicating triploidy (64% of *k*-mer pairs), though a faint signal can be seen under AB (11% of *k*-mer pairs)

2.3.3 Contaminant removal

SPAdes pre-assembly created an assembly 193,498,073 base pairs in length spread over 213,221 scaffolds. *Bwa* mapped 95.06% of reads from library 14 and 94.61% of reads from library 58 to the *SPAdes* pre-assembly. For each library, *blobtools* produced a plot with two distinct clouds around 10^3 and 10^1 coverage. Within the 10^1 coverage cloud, many contigs were being assigned to Rotifera or Arthropoda. Reads below coverage of 20x, and reads not assigned to Unidentified or Nematoda, were removed.

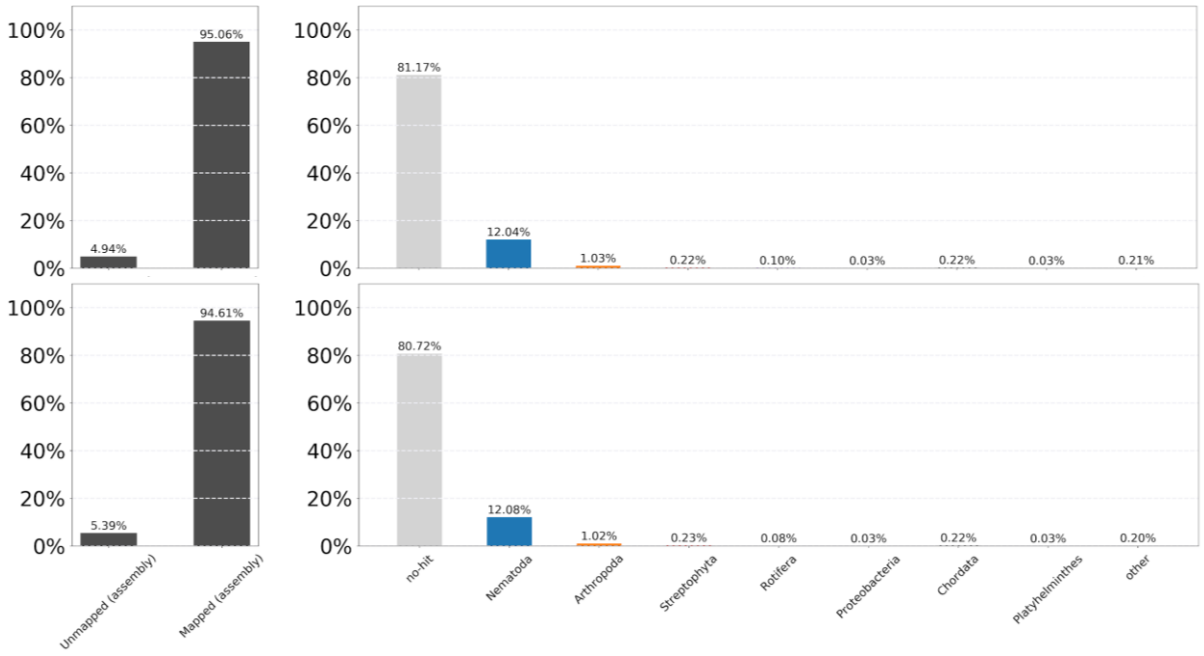

Figure 2.6a-b: Successful read mapping and detected contaminant percentages. A, top, 95.06% of reads from library 14 mapped back to pre-assembly. B, bottom, 94.61% of reads from library 58 mapped back to the pre-assembly.

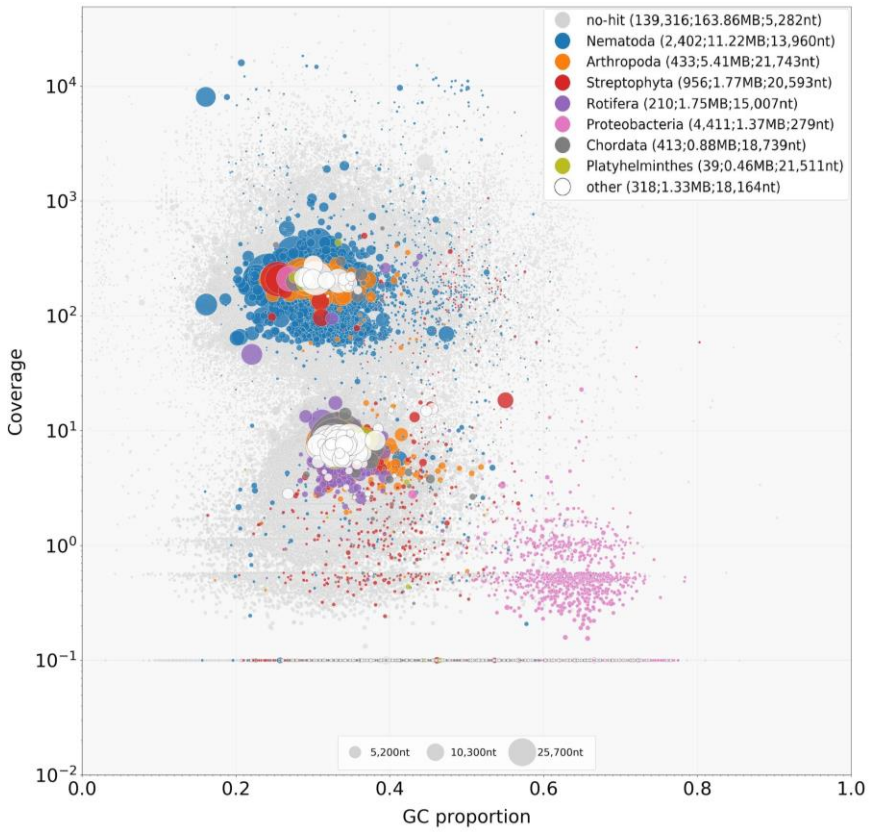

Figure 2.7a: Blobplot of library 14 showing coverage, GC content, and taxon. Plot shows a large amount of proteobacteria (pink) and unidentified reads with very low coverage. Some tomato (red) contamination is present both at low and high coverages. High coverage arthropod (orange) hits are considered false taxon assignments.

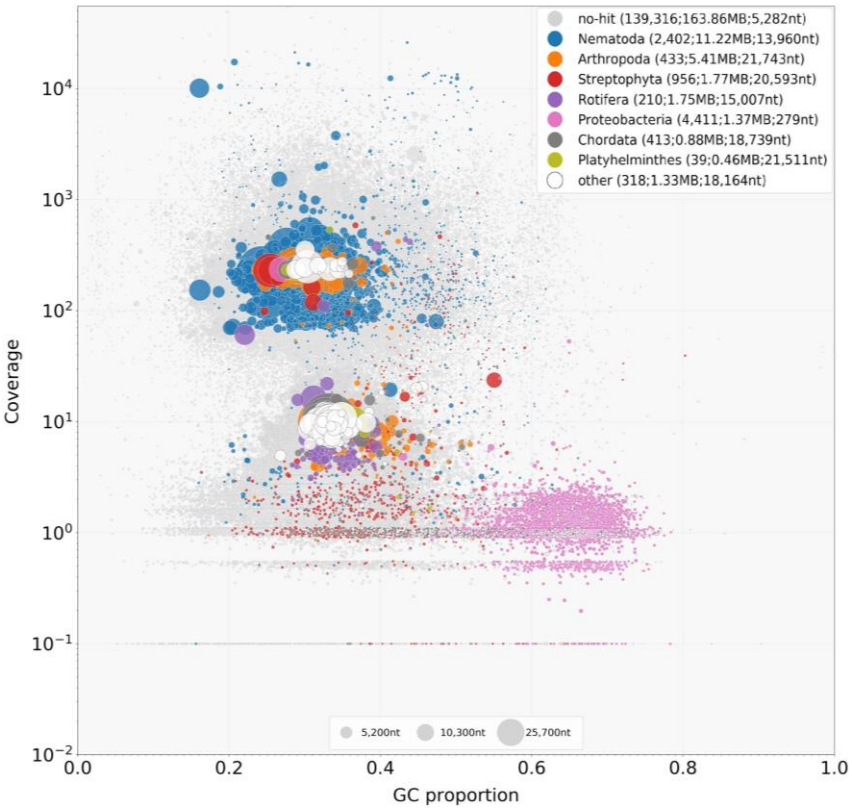

Figure 2.7b: Blobplot of library 58 showing coverage, GC content, and taxon. As Figure 2.7a.

2.3.4 Assembly

Results of different assembly methods can be found in Table 2.3. Given the results of the appraisal stages of the assembly process, assembly *PGCR,* hereafter referred to as the *M. haplanaria* assembly, was chosen to continue into the annotation stage. Reasons for this choice were a total genome size similar to profiling estimations (72 Mbp estimated, 69Mbp observed), a large N50 (24,820) in comparison to genome size and N50 of other assemblies (311 - 8052) , a higher BUSCO score than most other assemblies (79.2%), and higher scoring statistical metrics overall. A comparison of the chosen *M. haplanaria* genome against other published genomes in the literature can be found in Table 2.4.0

2.3.5 Annotation

Sequence homology prediction

EST2GENOME predicted 13,645 genes, of which all were CDS (Table 2.5).

***Ab initio* prediction**

The first iteration of *SNAP* predicted 22,697 genes and CDS, with the second *SNAP* iteration refining this estimate to 14,978 genes and CDS. Third iteration *SNAP* predicted

Table 2.3: Assembly results table. Blue signifies a good result, red signifies a bad result, as determined by *QUAST.*

Genome assembly	*SPAdes*	*Platanus assemble*	*Platanus scaffold*	*Platanus gapClosed*	*Platanus scaffold + redundans*	*Platanus gapClosed + redundans*	*SPAdes + redundans*
Reference	*spades*	*PA*	*PS*	*PGC*	*PSR*	*PGCR*	*spadesR*
Complete BUSCO (%)	90.43	28.71	65.02	70.63	64.36	76.9	93.73
Partial BUSCO (%)	7.92	27.39	19.47	15.18	21.12	11.55	4.29
# N's	456333	0	4518833	1653102	31819	87055	88047
# contigs	227621	492456	292691	292691	20219	7130	60728
# contigs (>= 1000 bp)	48600	19532	8531	8487	12510	5289	25900
# contigs (>= 5000 bp)	9696	1016	3525	3526	4171	3065	9220
# contigs (>= 10000 bp)	2617	165	2097	2086	1388	2033	3529
# contigs (>= 50000 bp)	14	0	120	119	7	214	113
Largest contig	130740	42110	227182	226635	95239	279646	154700
Total length (Mb)	216.7	140.1	113.4	113.1	65.9	69.3	163.0
Total length (>= 1000 bp)	173916399	42588431	69901102	69730810	62327815	68365113	150968133
Total length (>= 10000 bp)	39748115	2360142	48977488	48792154	21763537	55423312	68469010
Total length (>= 50000 bp)	921652	0	8684419	8626972	467427	16286264	7824102
N50	3752	311	6384	6398	6712	24820	8052
N75	1318	170	147	147	3455	12485	3722
L50	14244	72011	3055	3052	2759	750	4986
L75	38652	241191	71763	72132	6165	1730	12450
GC (%)	32.26	30.81	30.88	30.91	29.87	30.13	31.97

Table 2.4: Comparison of final *M. haplanaria* and other *Meloidogyne* assemblies from the literature.

Assembly	*M. javanica*	*M. incognita*	*M. arenaria*	*M. enterolobii*	*M. floridensis*	*M. luci*	*M. haplanaria*
Year	2017	2017	2017	2020	2017	2019	2020
Sequencing technology	Illumina HiSeq	Illumina HiSeq	Illumina HiSeq	Hybrid short and long read	Illumina HiSeq	PacBio + 10X	Illumina HiSeq
Isolate	VW4	W1	HarA		SJF1	V13	SJH1
BioProject				PRJEB36431		PRJEB27977	-
Scaffolds	34,394	33,735	46,509	4,451	9,134	327	7130
Genome Span (Mbp)	142.60	122.04	163.77	240.05	74.89	209.16	69.37
Longest scaffold (Kbp)	223.5	248.8	163.2	1466.8	88.4	6421.93	279.6
N50	14,133	16,498	10,504	143,476	13,256	1,711905	24,820
GC	29.6	29.9	29.5	30.02	30.2	30.2	30.13
Mapped reads	98.82%	99.10%	98.80%	-	-	-	-
BUSCO complete (%)	-	52.40%	56.10%	83.17	50.60	83.5	76.9
BUSCO fragmented (%)	-	9%	9.60%	6.93	8.50	7.59	11.55
Predicted Genes	26.917	24.714	30.308	-	14,144	-	20,213
Functional Annotated	17.659	15.938	20.813	-	-	-	-

Table 2.5: Annotation statistics from each stage of the *MAKER2* workflow.

Run name	EST2GENOME	SNAP1	SNAP2	SNAP3	Augustus
Run description	est2genome and protein2genome	First iteration of SNAP	Second iteration of SNAP	Third iteration of SNAP	Augustus using snap2 output as input
Number of genes	13645	22697	14978	1861	20213
Number of cdss	13645	22697	14978	1861	20213
Number of exons	89038	45127	109091	1861	149065
Number of exons in cds	89038	44953	107371	1861	147500
Number of introns in cds	75393	22256	92393	0	127287

only 1861 genes and CDS, but the same number of exons with no introns, indicating that the algorithm had become too sensitive through overtraining. As a result, the output of iteration 2 of *SNAP* was used to train *Augustus,* which itself predicted 20,213 genes and CDS (Table 2.5). The *Augustus* annotations were chosen to be used in downstream analyses.

2.3.6 Intragenomic blast

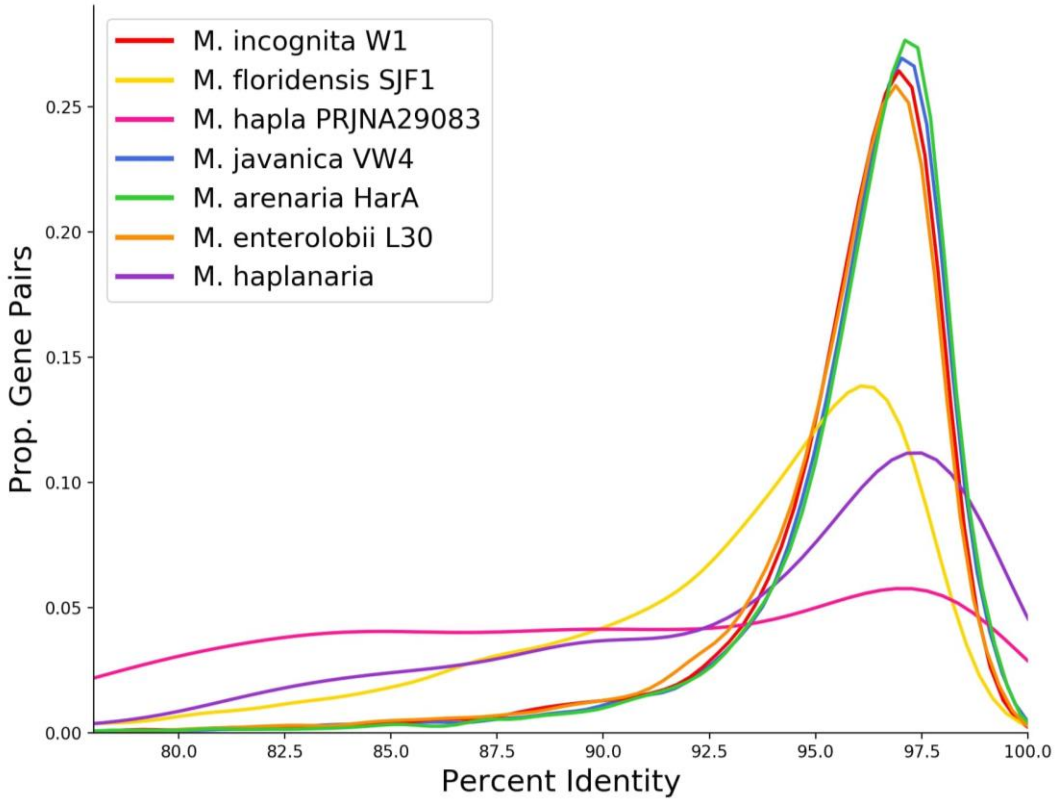

Figure 2.8: Intragenomic blast results. Percent identity on the *x* axis, proportion of genes found to be paired on the *y* axis. *M. arenaria* (green), *M. javanica* (blue), *M. incognita* (red), and *M. enterolobii* orange) can be seen grouped closely together in a large peak over 96.5-97%. *M. hapla* (pink) has almost no peak. *M. floridensis* and *M. haplanaria* peak to around the same height, with the *M. floridensis* peak over around 96% and the peak of *M. haplanaria* over 97.5%.

The intragenomic blast analysis detects the presence of divergent gene copies using a percent identity sequence similarity search of a gene against all genes in that annotation set, suggesting a hybrid origin for *M. haplanaria*. This analysis predicts that the *M. haplanaria* genome contains a large amount of paired gene copies with around 97.5% similarity (Figure 2.8). This is very short of the amount found in most MIG species and is much higher than *M. hapla,* a nearly homologous diploid species from Clade II. The count for *M. haplanaria* is however very close to the amount found in *M. floridensis.*

2.4 Discussion

We determine from these results that the genome of *M. haplanaria* is most similar to that of *M. floridensis* in length and GC content. The overall quality of the *M. haplanaria* assembly is acceptable when compared to assemblies made with similar methods and sequencing technologies, but vastly inferior to modern assemblies with more modern methods, such as *M. enterolobii* PRJEB36431 (2020) and *M. luci* V13 (Susic et al. 2019). This is almost entirely due to limitations on assembly contiguity imposed by using solely short read technology. The future availability of long read data of *M. haplanaria* would greatly increase the contiguity and quality of any assembly.

Profiling analyses performed by *GenomeScope2* and *Smudgeplot* concordantly indicated that *M. haplanaria* contains a triploid genome arrangement. Bimodal distribution of the k-mer frequency over *k*-mer coverage (Figure 2.4a & c) fit a triploid model well, though observed measures slightly peaked over the expected, indicating some departure from exact triploidy.

Several improvements can be made regarding the contaminant removal stages of the workflow. Despite pre-assembly, *blast* failed to assign taxonomy to around 81% of all contigs, leading to a large amount of unidentified reads. This means that for these reads, only GC content and coverage can be used to indicate contamination; two much less reliable metrics for detecting contaminants. This could be improved upon by constructing the pre-assembly out of more stringently trimmed reads, or also querying against a custom *Meloidogyne k*-mer database to try and assign unidentified reads (Wood, Lu and Langmead, 2019). Following its performance in this study, the snakemake assembly workflow will be changed to filter contaminants differently. In future, it may be more

accurate to drop all reads with very low coverage, automatically drop reads assigned to likely contaminants - Streptophyta, Homo, Pan, Bacteria - then blast the remaining reads against a custom database of contaminant free *Meloidogyne* genomes, thus ensuring much better accuracy of taxonomy assignment.

Choice of assembler has a significant impact on the state and quality of the resulting assembly, despite input data being identical (Dominguez Del Angel et al., 2018). We find the same as Szitenberg et al. (2017) that when provided with only short read data, *platanus* produces better assemblies than its contemporaries. *Platanus allee* was brought to our attention partway through the assembly process (Kajitani et al., 2019). Used independently from *platanus*, *platanus allee* is an assembly software designed to produce assemblies representative of both haplotypes of highly heterozygous genomes. Addition of this to the assembly would be useful for future studies alongside the *redundans* pipeline to produce both haploid and diploid representative assemblies.

It is thought that the short length of the *M. haplanaria* assembly relative to the profiling prediction and other assemblies of species in the genus is a product of the *redundans* pipeline. As well as shortening the overall length of the assembly, *redundans* may also have collapsed heterozygous gene pairs into single representative sequences, stripping the assembly of data vital to a thorough phylogenomic analysis and skewing counts of homeologous pairs. Because of this the results of the intragenomic blast can be called into question. If some homeologs from one subgenome were removed by *redundans*, then the observed peak shown by *M. haplanaria* in Figure 2.8 is an underestimate, some degree lower than the actual peak. This does not change the determination that *M. haplanaria* contains divergent genomic copies, only that the extent may be greater than the effect identified in this study.

Overall, however, the *snakemake* assembly workflow (Köster and Rahmann, 2012) performed well, running in its entirety from start to finish and successfully completed all jobs required of it. The workflow took just under two weeks to run, including assembly stages, on a local node using 28 cores and 400 gigabytes of RAM. Alongside modifications to assemblers and quality control mentioned above, a primary aim now is to parallelise the workflow and set it up to run on a high performance computer, using up to a terabyte of RAM and many nodes at once, each running 28 cores. This will cut the runtime of the workflow dramatically to a speed where, combined with the automation of

snakemake, genome assembly of *Meloidogyne* species can become a simple, routine process that can be done by anyone with basic computer literacy.

Annotation proved to be a computationally challenging part of this study, as was expected. *MAKER2* performed well on smaller assemblies, particularly the *M. haplanaria* assembly, *PGCR*, generated above, but required the resources of a high performance computer to annotate larger genomes. Though all attempts were made to parallelise *MAKER2,* time limitations prevented further annotations of assemblies. *MAKER2* successfully annotated the *M. haplanaria* assembly, *PGCR*, ultimately predicting 20,213 genes.

During the annotation process, the third iteration of *SNAP* become overtrained; a caveat to using *ab initio* approaches. This was interpreted from the heavily reduced count of genes predicted and the identical counts of genes predicted and genes with one exon. For this reason, *Augustus* was trained with the gene predictions of *SNAP'*s second iteration.

It became clear that running the assembly through the *redundans* pipeline had detrimental effects on its quality in terms of what was required for annotation and further downstream analysis. In effect, *redundans* had caused a trade-off of content for contiguity. Temporal and computational limitations prevented annotation of any other assemblies, though we intend to annotate the assembly *PGC* in the hope that many homeologous copies are still intact.

Following its use in this study, RepBase - a source of repeat databases for this study - began requiring a licence for use. This makes future annotation and repeat masking using the methods employed here more difficult and less accurate, given that without a license, any future repeat database would be limited to what could be generated by the researcher. Annotation based on sequence similarity would produce predictions of a much higher quality if the reference it was given was from the same species it was attempting to annotate. As yet, no transcriptome data of *M. haplanaria* exists, but its availability would immediately increase the quality of annotation.

Intragenomic blast analysis of *M. haplanaria* CDS extracted from *Augustus* annotations showed that *M. haplanaria* does contain multiple gene copies as was hypothesised. This

presence of multiple homeologous pairs would indicate that *M. haplanaria* is also a hybrid, along with the MIG. Further phylogenomic analyses would have to treat it as such. The percentage of gene pairs found was much closer to the percentage exhibited by *M. floridensis* than the MIG, though this may be an artifact of the *redundans* pipeline removing homeologous sequences, artificially lowering the amount of CDS detectable for use in the intragenomic blast analysis.

In conclusion, we have successfully assembled and annotated the genome of *Meloidogyne haplanaria,* while also finding evidence that *M. haplanaria* is a hypotriploid nematode, containing homeologous subgenomes that exhibit a notable degree of divergence. Phylogenomic analyses seeking to determine the position of *M. haplanaria* within *Meloidogyne* must consider this divergence, as no confident species position can be established until the position of the subgenomes is known.

Chapter 3: Phylogenomic analysis of *Meloidogyne haplanaria*

3.1 Introduction

Root-knot nematodes of the genus *Meloidogyne* cause billions of pounds in agricultural crop damage annually (Bernard et al. 2017). Knowledge of their evolutionary history is important to understand the biological processes surrounding parasitic adaptation, but investigation has been limited due to a scarcity of adequate genomic resources. In this study we apply a novel phylogenomic workflow to infer the evolutionary history of the root-knot nematode *Meloidogyne haplanaria*.

3.1.1 *Meloidogyne* phylogenetics

The phylogeny of the *Meloidogyne* genus is a convoluted one due to several hybridization events within its past (Lunt, 2008; Szitenberg et al., 2017). Phylogenetic mitochondrial and ribosomal analyses of this genus find that species within it fall into several well-defined clades, of which arguably the most well-studied is Clade I (Chapter 1, Figure 4.4). Clade I includes, among others, *Meloidogyne enterolobii* and the *Meloidogyne incognita* group (MIG) (Szitenberg et al., 2017; Álvarez-Ortega, Brito and Subbotin, 2019). Members of the MIG include its namesake *Meloidogyne incognita,* as well as *M. javanica, M. arenaria,* and *M. floridensis,* all of which are prolific crop pests (Bebber, Holmes and Gurr, 2014).

Lunt et al (2014) and Szitenberg et al. (2017) found evidence of past hybridisation events within the genus. As a result of these hybridisation events, *M. enterolobii* and members of the MIG contain divergent homeologous subgenomes, with divergent evolutionary histories. An analysis performed by Szitenberg et al (2017) accounted for multiple genomic copies of *M. enterolobii* and the MIG, treating them as individual taxonomic units within the tree. The study found that the MIG's evolutionary history diverges at its base into two distinct groups, *A* and *B*, both with a representative from each species. These subgenomes exhibit more similarity to the equivalent subgenome of another MIG species (*A*-*A*), than they do to the opposing intragenomic homeolog (*A*-*B*). Both copies of *M. enterolobii* fall as a monophyletic outgroup to the MIG, though both descend from a

different origin than those within the MIG (Abad et al. 2008; Lunt et al. 2014; Szitenberg et al. 2017).

Very few studies discussing the evolutionary history of *Meloidogyne* account for the presence of divergent subgenomes, as they are only discernible with large amounts of nuclear data. Phylogenetic investigation with genomic data of this scale is referred to as phylogenomics.

3.1.2 Phylogenomics

Phylogenomics is a field that has arisen alongside bioinformatics since the advent of high-throughput next-generation sequencing (Young and Gillung, 2019) and describes the process of using genome-scale multilocus data to infer the evolutionary history of an organism or taxonomic group (Figure 3.1) (Rodríguez-Ezpeleta and Philippe, 2009; Young and Gillung, 2019). Phylogenomics is well placed to provide finer resolution and structure to the evolutionary history of the *Meloidogyne* genus. As Szitenberg et al (2017) demonstrated, the ability to incorporate and compare large amounts of nuclear data from both subgenomes through phylogenomics allows resolution of evolutionary history to a level that was previously unattainable. Though performed infrequently in the past, the increase in available genome assemblies and genome-scale data of *Meloidogyne* species has and will continue to make these kinds of analysis more common in this field.

The phylogenomic process

Genes predicted through the genome annotation process are grouped based on sequence similarity into orthogroups containing all representatives of a locus. After some quality control and filtering, orthogroup members are aligned and converted into phylogenomic trees via one of two methods: concatenation or coalescence (Figure 3.1). Concatenation analysis involved trimming all single-locus alignments to an equal length and concatenating them end to end to form a super-alignment. This alignment is passed to phylogenetic software wherein a single phylogenomic tree is inferred (Young and Gillung, 2019). Coalescent analysis takes a different approach: Instead of concatenating alignments, a phylogenetic tree is created for each individual orthogroup. A phylogenomic tree is then inferred from the probabilistic tendency of trees of all genes (Mirarab et al., 2014). Following tree inference, the overall phylogeny is drawn as a figure for interpretation.

Though phylogenomics has yielded some convincing answers to large evolutionary questions (Misof et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019; Williams et al., 2019), technological limitations, data variability, and parameter modification can all significantly affect the result (Delsuc, Brinkmann and Philippe, 2005; Young and Gillung, 2019).

Limitations

Technological limitations include, but are not limited to low quality, data-poor, genome assembly and annotation, and the suitability of evolutionary models to explain substitution rates (Young and Gillung, 2019). Assembly quality can be degraded through the variable applicability of assembly and annotation methods to non-model organisms, overly strict filtering of contaminants and heterozygous contigs, among other things (Horner et al., 2010; Jayakumar and Sakakibara, 2019). Annotation quality can fall short due to lack of available transcriptome data to train software, exceptionally high or unsuccessfully filtered repeat content, or complexity of the study organism pushing the boundaries of the annotation packages algorithm (Campbell and Yandell, 2015; Salzberg, 2019). Ultimately, the best available technical or methodological setup will be limited by low quality sequence data (Siu-Ting et al., 2019; Young and Gillung, 2019).

Model choice

Model choice can also impact the quality and topology of the final phylogeny. To infer a tree, a model of the rate of substitution must be supplied. The choice of model for a given taxon, as well as a model's ability to account for heterotachyous rates - different evolutionary rates within a taxa - has been and continues to be discussed at length (Kolaczkowski and Thornton, 2004; Reddy et al., 2017; Prasanna et al., 2020). Apomictic reproduction renders *Meloidogyne* species incapable of homologous recombination, which raises questions regarding their mode and rate of evolution.

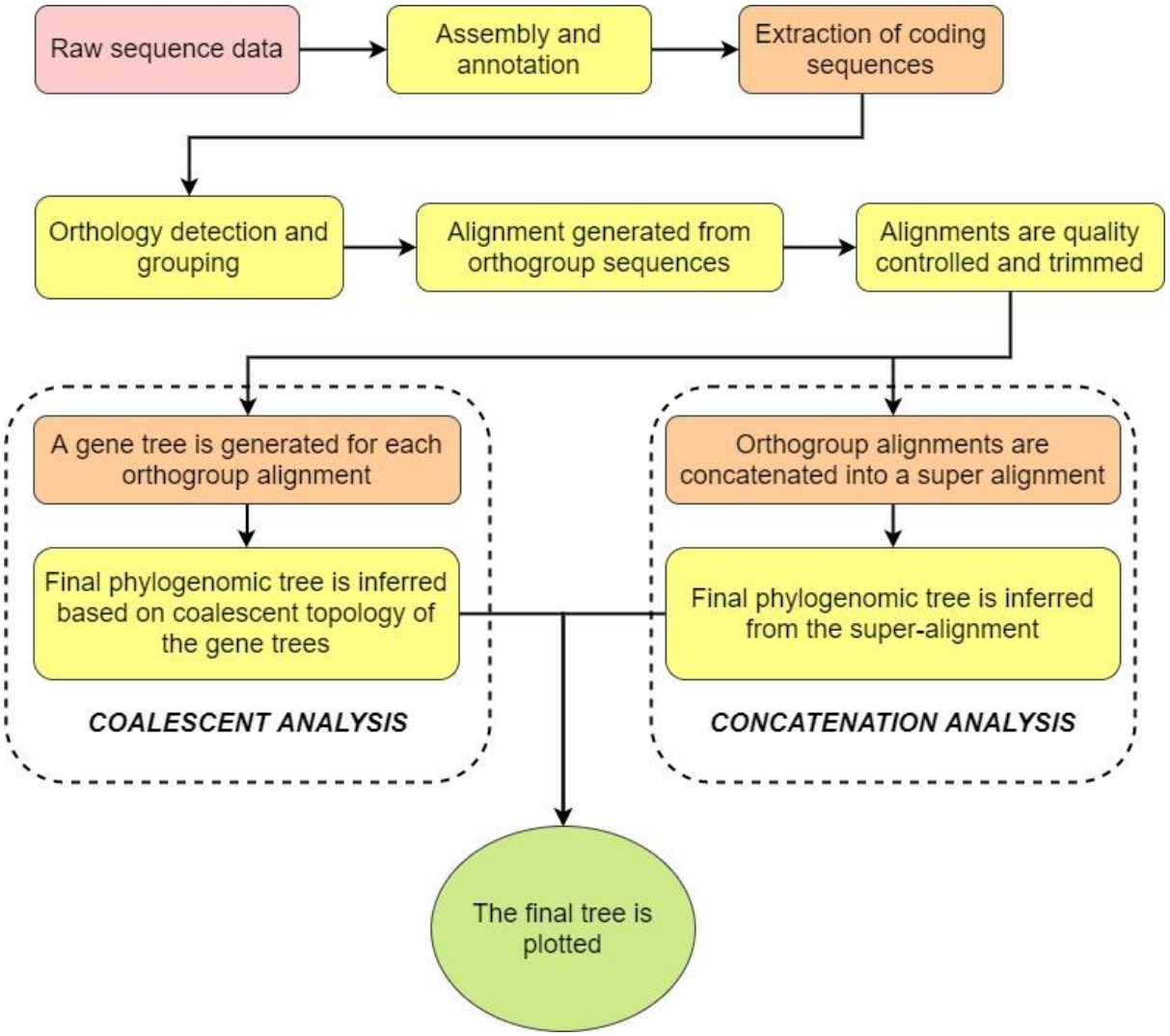

Figure 3.1: An example phylogenomics workflow. A phylogenomic study can perform either a coalescent or a concatenation analysis. In a coalescent analysis a gene tree is inferred for each orthogroup alignment, then the consensus topology of these trees is used to infer a final phylogenomic tree. In a concatenation analysis, all orthogroup alignments are concatenated end on end into a super-alignment. The final tree is then inferred from overall similarity between sequences in the super-alignment.

Though model selection packages, when run on *Meloidogyne* orthology datasets, frequently recommend the GTR model (García and Sánchez-Puerta, 2015; Janssen et al., 2017; Álvarez-Ortega, Brito and Subbotin, 2019) this can be because the GTR model is the most neutral, loosely restricted model and fits even with lower quality data (Sumner et al., 2012; Sumner, Fernández-Sánchez and Jarvis, 2012). A potential emerging alternative that accounts for the differing evolutionary rates possible between *Meloidogyne* species is the GHOST model, which has been shown to recover more

accurate topologies for heterotachyously evolved species (Crotty et al., 2020). This model works by … and has been used frequently in the literature since its publication

Metrics of confidence

Many phylogenomic studies fail to accurately portray the statistical support of their result (Kumar et al., 2012). Bootstrapping is the most commonly used metric to display confidence in a given node, however the significance of the bootstrap declines as the size of the dataset increases, causing almost all nodes to score one hundred percent, even when the resulting trees infer conflicting evolutionary histories. It has been shown possible to generate conflicting trees from the same dataset, each with maximum bootstrap values (Reddy et al., 2017). To combat this, other metrics have been developed based upon Bayesian statistics and concordance factors.

Concordance factors are metrics that describe what percentage of gene tree topologies within a multigene tree conform with the base of a clade (Ané et al., 2007). This is called gene concordance factor (gCF). Alongside this is the novel site concordance factor (sCF), developed by (Minh, Hahn and Lanfear, 2020) and implemented in *IQTREE* (Minh et al., 2020). sCF is defined as a measure of the percentage of sites that conform in a reference tree. The ability of these metrics to represent underlying incongruent gene topologies makes them an insightful metric to include alongside bootstrap values (Kumar et al., 2012).

3.1.3 *Meloidogyne haplanaria*

The Texas peanut root-knot nematode, *Meloidogyne haplanaria,* is an emerging crop pest that parasitises several species of plant and agricultural crop, most notably tomato and peanut, is endemic to the USA, and has been isolated from Arkansas, Florida, and Texas (Eisenback et al., 2003; Joseph et al., 2016; Ye, Robbins and Kirkpatrick, 2019).

The phylogenetic position of *M. haplanaria* remains unclear. Mitochondrial analysis places it either as sister taxa to *M. enterolobii* (Joseph et al., 2016); (Álvarez-Ortega, Brito and Subbotin, 2019) or branching between *M. enterolobii* and the MIG such as what is seen in Chapter 1, Figure 1.6 (Szitenberg et al. 2017; Ye et al. 2019; Santos et al. 2019). Multigene analyses using several ribosomal and mitochondrial genes place find the same dichotomy (Holterman et al., 2009; Álvarez-Ortega, Brito and Subbotin, 2019). No

genome scale phylogenomic analysis has yet included *M. haplanaria* due to the lack of an assembled genome.

In Chapter 2 we assembled and annotated the genome of *M. haplanaria,* in which we found evidence supporting the presence of divergent subgenomes, alongside some evidence of hypotriploidy. Given these similarities to the other Clade I RKNs and its position in mitochondrial phylogenies, we hypothesise that *M. haplanaria* is also the descendant of a past hybridisation event like that of *M. enterolobii* and the MIG. The studies above defining *M. haplanaria*'s phylogenetic position do not account for the presence of divergent genomic copies and therefore lack the power to elucidate the evolutionary history of the group to the necessary resolution. To test our hypothesis, only a phylogenomic analysis similar to that performed by Szitenberg et al (2017) would be sufficient to resolve the evolutionary history of either subgenome.

This study generates orthology groups of *M. haplanaria* and five other *Meloidogyne* species, detects divergent gene pairs within them and performs a phylogenomic concatenation analysis. The resulting tree is representative of both genome copies from all included species, allowing us to determine the positions of either *M. haplanaria* subgenome. Iterative parameter configurations were applied to control uninformative data and increase phylogenetic signal. In addition to this, the final alignment was filtered for several parameters, including alignment length, parsimoniously informative sites, and missing percentage of sequences. We employ a novel *snakemake* workflow (Köster and Rahmann, 2012a) (Chapter 1, Section 1.5.2) to perform these analyses and generate summary statistics, due to its enabling of parameter iteration and replicability.

3.2 Methods

3.2.1 Reproducibility

All scripts and workflows are made available, and all attempts have been made to ensure reproducibility. See https://github.com/mrmrwinter/phylogenomics_MIG.

All stages of the phylogenomic analysis were performed in a *snakemake* workflow (Chapter 1, Section 1.5.2) written specifically for the *Meloidogyne* genus, available here: https://github.com/mrmrwinter/phylogenomics_MIG. The directory structure is

standardised according to recommended best practice. Environment files are kept in the *envs/* directory. Scripts used by the phylogenomics workflow are contained in the *scripts/* directory. Python scripts to process, plot and display the data from some of these analyses were written, along with those from Szitenberg et al. (2017), and referred to in italics below. Raw and input data is stored in the *data/* directory. *Output/* contains all intermediary data from the workflow, including transformed data, alignments, and trees, *results/* contains the final plotted trees and figures, as well as summary statistics, and the *reports/* folder contains logs of outputs and package performance.

The workflow requires only input CDS data and brief configuration to reproduce all phylogenomic analyses. Replication or repetition of this analysis requires only input data and brief configuration. Every stage is performed and controlled by the *snakemake* workflow according to configurable input parameters.

3.2.2 Orthology detection

Orthologues were grouped together using *OrthoFinder2* (Emms and Kelly, 2018). Alongside amino acid CDS of *M. haplanaria* extracted from the annotation performed in a previous chapter (Chapter 2, Section 2.2), amino acid CDS of *M. incognita, M. floridensis, M. arenaria, M. enterolobii, M. javanica,* and *M. hapla* (Supplementary table 3) were passed into *OrthoFinder2* using an inflation value of 2 and otherwise default settings. An inflation value of 2 was chosen based on experiments done by Szitenberg et al (2017), wherein after several iterations of *OrthoFinder2* with inflation values ranging from 1.1 to 9 it was determined an inflation value of 2 recovered the most orthogroups exhibiting at least one representative from each of the Clade I species.

The number of orthogroups with one-to-one, two-to-two, and one-to-four representatives from each species were counted using *orthology_tabulation.py* and plotted in a table (Table 3.1) using *pandas* (Virtanen et al., 2020). Groups with one-to-one and two-to-two representatives per orthogroup were dropped due to lack of information of all included species. Histograms were plotted showing copy number per orthogroup per species (Figure 3.2a-f), and heatmaps were plotted (Figure 3.3a-c) showing the number of orthogroups containing each species and how many copies were shared between species using *matplotlib* (Hunter, 2007). *Orthology_tabulation.py* also creates a list of all

orthogroups containing one-to-four representatives. This list is passed into the tree generation stages of the *snakemake* workflow.

3.2.3 Alignment preparation and clustering of homeologs

Each orthogroup was subjected to the following process: Sequences were aligned with *mafft* using the Smith-Waterman algorithm (L-INS-i) and one thousand iterative refinement cycles (*--localpair --maxiterate 1000)* (Katoh et al., 2002). Resulting alignments were trimmed with *trimAl* (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009) using a range of different parameters, available in Table 3.3, and output a collection of trimmed alignments for each . *trimAl* was also run with relaxed settings (*-gt 0.7 -st 0.001*) to generate an initial tree from the most available data, regardless of data quality. Trimmed orthogroup alignments were used to generate gene trees with *RAxML* using a random seed and the Generalised Time Reversible (GTR) model (*-p 123 -m GTRGAMMA)* (Stamatakis, 2014). Scripts initially written by Szitenberg et al (2017) were adapted to perform collapsing of sister nodes. *Collapse_and_cluster.py* was applied to the gene tree to collapse paralogs into two distinct A and B clusters per orthogroup, representing each divergent homeolog within each species. Alignments containing two or more sequences from a single species with small overlap (<20bp) were dropped to avoid inclusion of two exons of the same ortholog as separate orthologous loci. Using the *RAxML* gene trees produced, paralogous leaves were collapsed keeping the least derived sequence each time. Collapsed paralogs were then clustered into two groups based on patristic distance. Orthogroups with more than one representative per paralog per cluster were dropped here. Trimmed orthogroup alignments were then edited based on the resulting gene tree contents of the collapsing and clustering process. Orthogroups that pass the collapsing stage are concatenated using *concatenation.py* into a super-alignment. Summary statistics of these concatenations (Table 3.3) are then generated by *AMAS* (Borowiec, 2016).

Based on summary statistics of concatenations, iteration *trimAlgt0.7st0.5* was chosen as the most accurate tree to carry forward. Histograms were plotted (Figure 3.8) of *AMAS* (Borowiec, 2016) summary statistics for all trimmed gene tree alignments using *matplotlib (Hunter, 2007)* in order to visualise data distributions and a filtering step was performed with *filtering.py* using the following parameters, dropping alignments that didn't pass the filter; alignments of length between 300-3000 base pairs, with under 200 parsimony

informative sites, a proportion of variable sites lower than 20%, an amount of missing data below 15%, and at least 8 taxa.

3.2.3 Subsampling

Forty randomly selected trimmed orthogroup alignments from iteration *gt0.7st0.001* and *gt0.7st0.5* were resampled fifteen times, concatenated and transformed into densitrees using *toytree* in *subsampling.ipynb* (Figure 3.6a-b).

3.2.4 Tree building

All concatenated trees and *OrthoFinder2* species trees were built using *IQTREE* with default settings, using a GTR model and one-thousand bootstraps (*-m GTR -B 1000)* (Minh et al., 2020). Alignments for the initial *trimAlgt0.7st0.001* tree and the final filtered *trimAlgt0.7st0.5* tree were also run through *IQTREE* using the GHOST model (Figure 3.4b and Figure 3.9). All trees were rooted, drawn and plotted using *toytree* and *toyplot* (Eaton, 2020).

3.3 Results and Discussion

3.3.1 Orthology detection and exploration

OrthoFinder2 generated 20,167 orthogroups on its first iteration, of which 3926 contained between one and four representatives from each included species. The *OrthoFinder2* species tree alignment indicated that *M. enterolobii* was positioned outside of *M. haplanaria* and the MIG, meaning it is a suitable outgroup to use in analyses to position *M. haplanaria.* Its increased phylogenetic proximity compared to *M. hapla* helps with data quality and accuracy of the analysis (Wilberg, 2015). Its concordance with previous mitochondrial analyses is encouraging, indicating that the majority of genes shared by either genome of *M. haplanaria* fall monophyletically between *M enterolobii* and the MIG. *M. hapla* was dropped from the phylogenomic analysis at this stage and was excluded from all further analyses due to its redundancy as an outgroup and its potentially detrimental effect on the number of recovered orthogroups. All trees would instead be rooted with *M. enterolobii. OrthoFinder2* was run again, excluding *M. hapla* from the input.

The second run of *OrthoFinder2* generated 22,174 orthogroups. The second species tree alignment was congruent with the previous one, placing *M. haplanaria* between *M.*

enterolobii and the MIG. As expected, exclusion of *M. hapla* increased the number of orthogroups generated by 2,007, including 413 containing between one-to-four representatives from each species (Table 3.1). Homeolog count analyses found *M. haplanaria* has a copy number of two in very few orthogroups. Though this could indicate that *M. haplanaria* has little to no hypotriploidy, it is likely the result of heterozygous contig removal by *redundans* during genome assembly. This is suggestive of problems with assembly and annotation and could limit the phylogenetic signal of either genomic copy (Figure 3.2f).

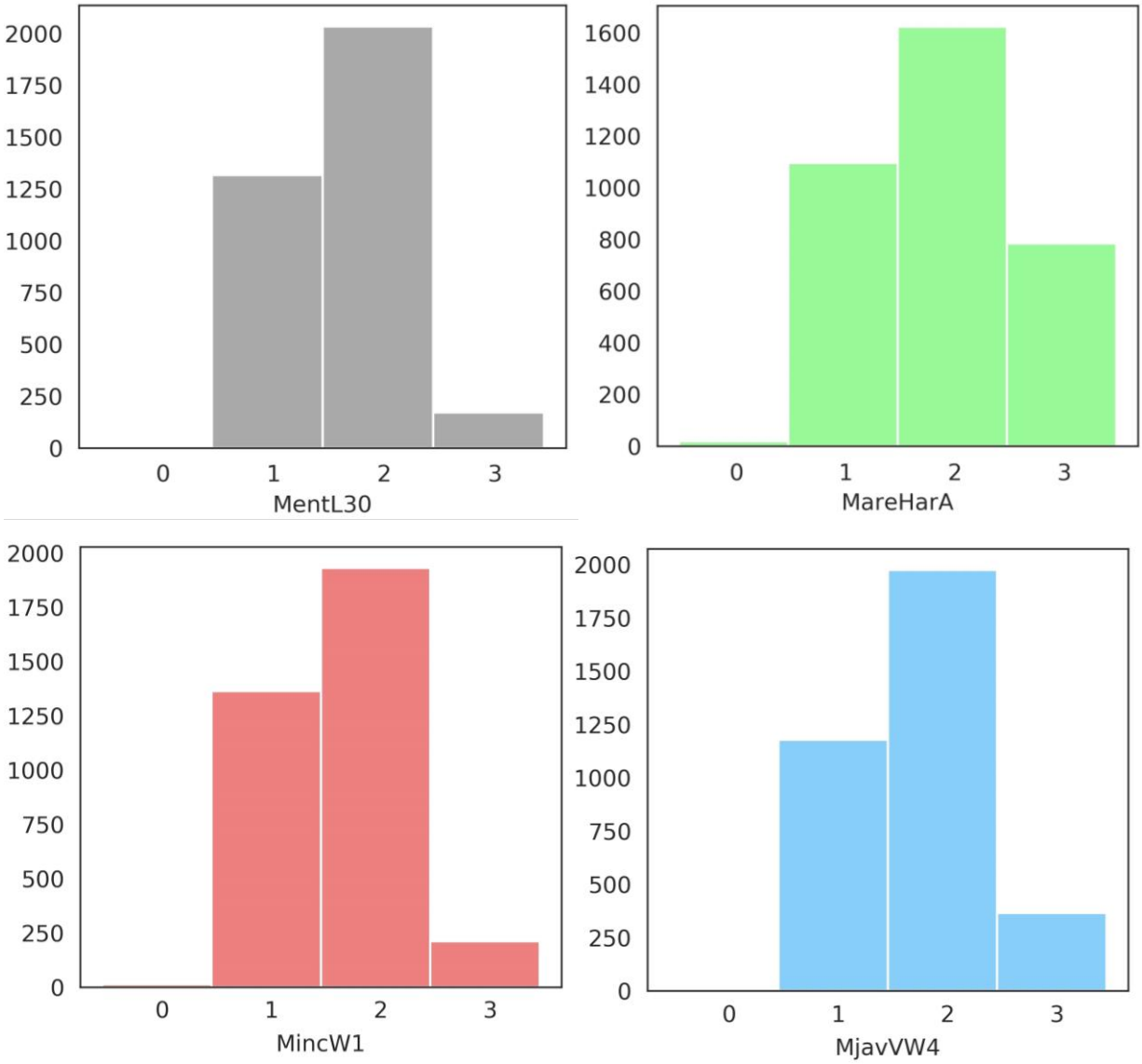

Figure 3.2a-d: Orthology copy number. Histograms showing ortholog copy number within orthogroups. A, top left, *M. enterolobii.* B, top right, *M. arenaria.* C, bottom left, *Meloidogyne incognita*. D, bottom right, *Meloidogyne javanica.* Copy number is on the x-axis, gene count is on the y-axis. A higher amount of genes with two or more copies indicates presence of divergent homeologs.

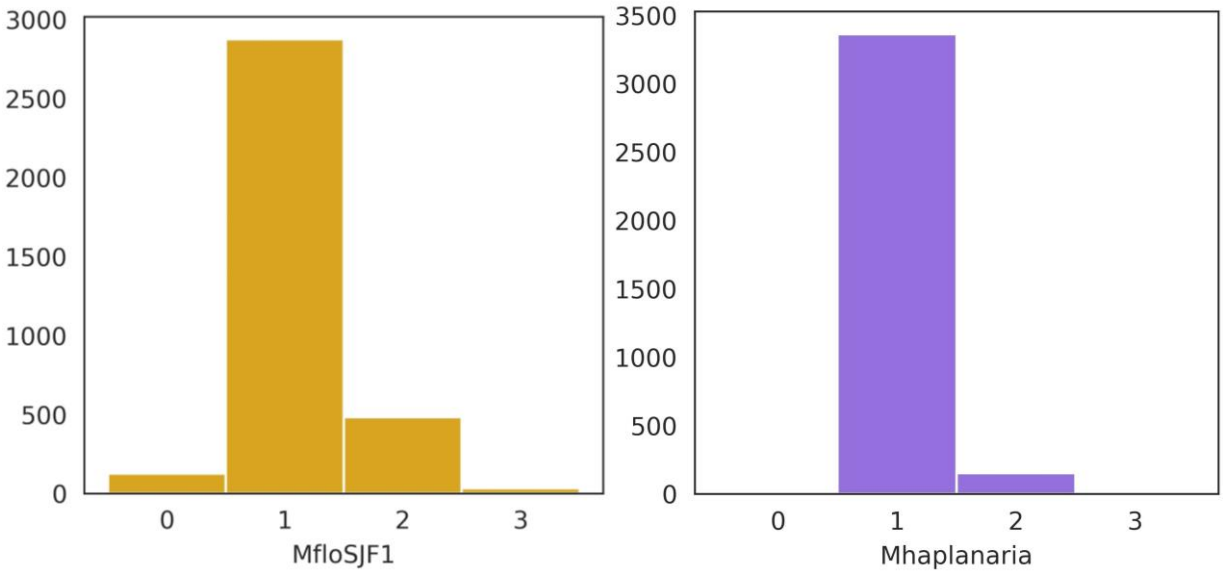

Figure 3.2e-f (continued): Orthology copy number. Histograms showing ortholog copy number within orthogroups. E, left, *Meloidogyne floridensis*. F, right, *Meloidogyne haplanaria.* Copy number is on the x-axis, gene count is on the y-axis. A higher amount of genes with two or more copies indicates presence of divergent homeologs.

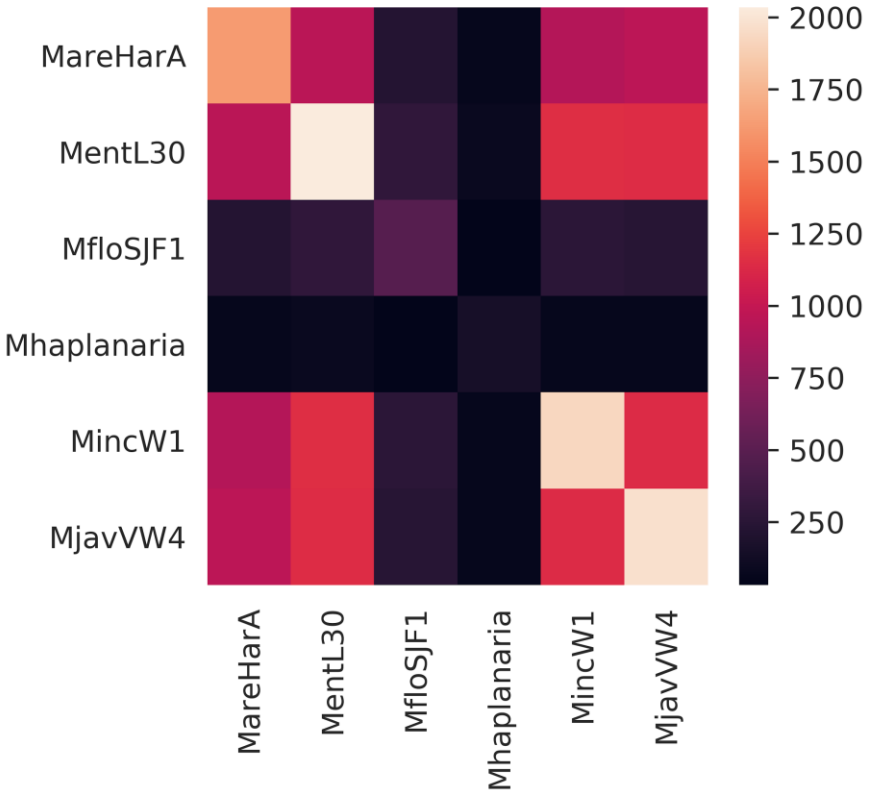

Figure 3.3a: Orthology heatmap. Heatmap shows how many orthogroups with two orthology copies are shared between each species.

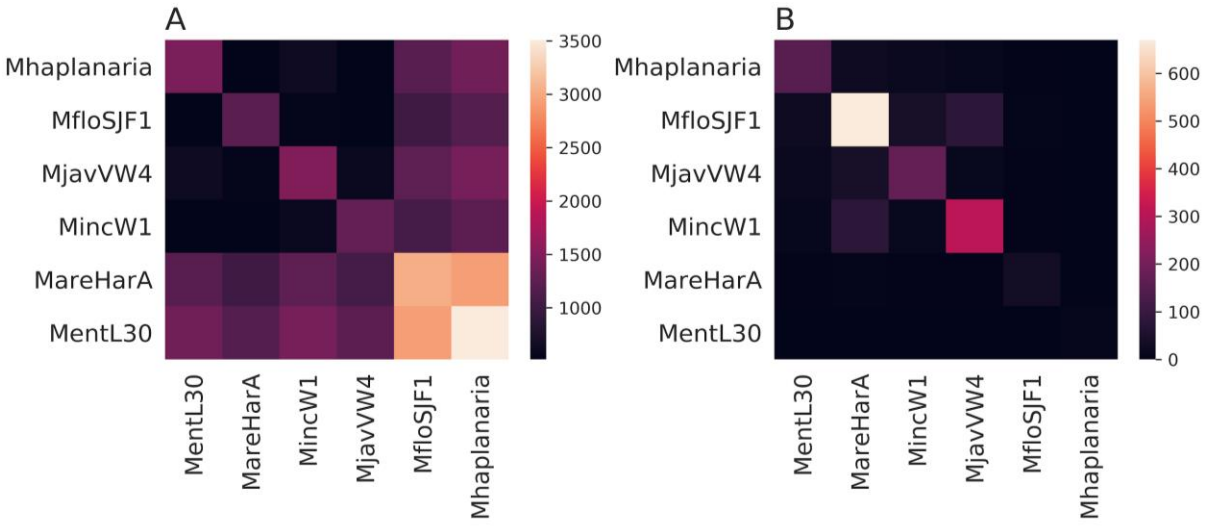

Figure 3.3b-c (continued): Orthology heatmap. Heatmap shows how many orthogroups with, A, one , and , B, three, orthology copies are shared between each species.

Table 3.1: Number of orthology groups with *n-n* representatives for all species.

	Total	One-one's	Two-to-two's	Between one and four
M. hapla included	20167	65	2	3926
M. hapla excluded	22174	93	10	4339

3.3.2 Alignment, *trimAl* parameter modification, and clustering

Counts of input and output orthogroup alignments for *mafft,* all *trimAl* iterations, and the clustering stage can be found in Table 3.2. *TrimAl* calculates scores for each sequence and column in an alignment. It can then be configured to trim relative to these scores. Several parameters were reconfigured. First, gap threshold (*-gt), which* relates to the gap score calculated by *trimAl* and removes columns that do not meet the threshold, and similarity threshold (*-st*), which relates to residue similarity score (RSS) and removes columns where the percent of residues passing the RSS is below the determined threshold. Second, the minimum residue overlap (*-resoverlap*) score for each residue, and *-seqoverlap,* the minimum percent of a sequence past the *-resoverlap* threshold needed to keep the sequence. *-gt*, *-st, -resoverlap,* and *-seqoverlap* are user defined parameters with incremental control. Boolean parameters include *gappyout,* which calculates a gap-score cut-off point and removes columns in the alignment that do not reach that value, *nogaps,* which deletes columns with at least one gap, and *noallgaps,* which removes columns containing only gaps. As expected, the number of orthogroups

produced from each iteration reflected the strictness of the parameters. The effect of RSS threshold (*st*) was not evident until over 25%.

Table 3.2: Orthogroup progress through alignment, trimming, and clustering.

Stage	Parameters	Number of orthogroups passed forward	Number of orthogroups passed out o
OrthoFinder2	Iteration value 2	4339	-
mafft	localpair maxiterate 1000	4406	-
trimAl	gt 0.7 st 0.001	3526	533
	gt 0.7 st 0.25	3526	533
	gt 0.7 st 0.5	3512	199
	Gt 0.7 st 0.75	3441	129
	resoverlap 0.5 seqoverlap 50	3966	184
	resoverlap 0.75 seqoverlap 75	1565	51
	noallgaps	3437	479
	nogaps	2925	461
	gappyout	3444	507

Table 3.3: Concatenation summary statistics.

Alignment name	No. taxa	Alignment length (bp)	Total matrix cells	Undetermined characters	Missing percent	Number of variable sites	Proportion of variable sites (%)	Parsimony informative sites	Proportion parsimony informative (%)	GC content (%)
gt 0.7 st 0.001	12	622477	7469724	2197030	29.412	68866	0.111	32586	0.052	0.361
gt 0.7 st 0.25	12	622477	7469724	2197030	29.412	68866	0.111	32586	0.052	0.361
gt 0.7 st 0.5	12	207215	2486580	739657	29.746	14811	0.071	4372	0.021	0.36
gt 0.7 st 0.75	12	94823	1137876	367529	32.3	622	0.007	0	0	0.358
resoverlap 0.5 seqoverlap 50	12	206740	2480880	717638	28.927	14948	0.072	4641	0.022	0.36
resoverlap 0.75 seqoverlap 75	12	47162	565944	146689	25.919	3267	0.069	1100	0.023	0.37
noallgaps	12	974092	11689104	5612629	48.016	95503	0.098	36480	0.037	0.358
nogaps	12	249101	2989212	680913	22.779	25718	0.103	12985	0.052	0.369
gappyout	12	717418	8609016	2958169	34.361	80545	0.112	35664	0.05	0.36

3.3.3 Tree building

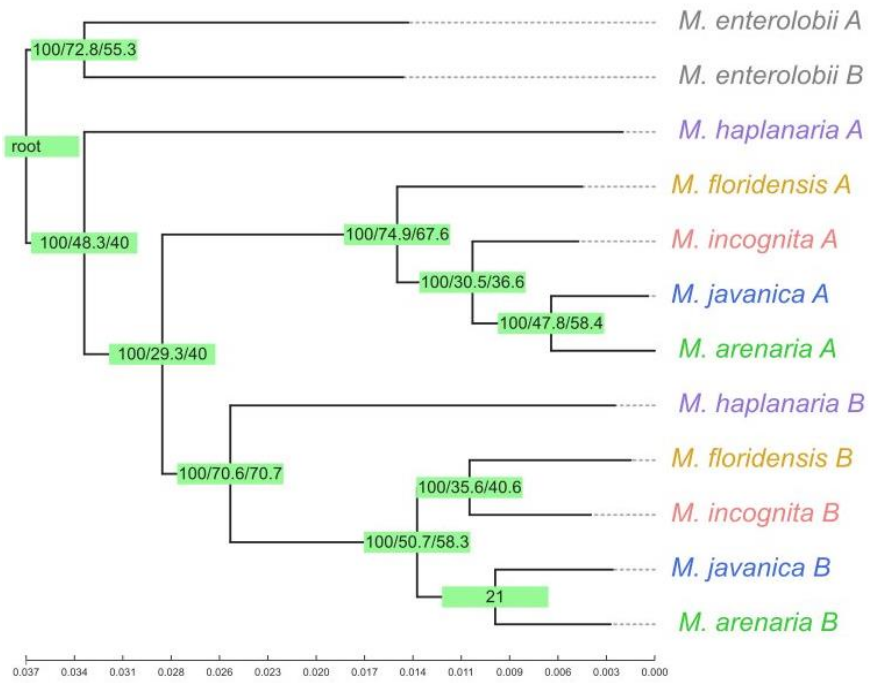

Figure 3.4a (left): Initial tree generated with relaxed trimming settings and the GTR model of substitution. *M. haplanaria B* falls as sister taxa to the MIG (70.6 gcf). *M. haplanaria A* falls outside of the MIG (48.3 gcf).

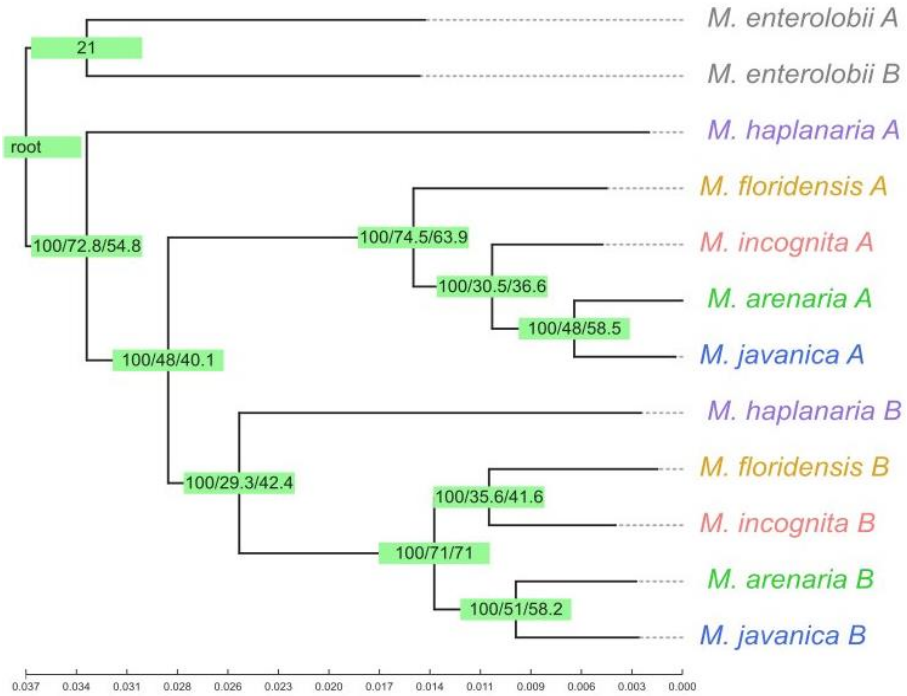

Figure 3.4b (right): Initial tree with relaxed trimming settings and the GHOST model of substitution. Topology of the tree is the same as Figure 3.4, however concordance metrics favour the position of *M. haplanaria A* (72.8 gcf) over the position of *M. haplanaria B* (29.3 gcf).

Initial trees with relaxed trimming

An initial tree was generated with the relaxed parameter configuration (*-gt 0.7 -st 0.001)* (Figure 3.4a & b). Using relaxed settings such as this allows an overview of the untrimmed data and provides a null hypothesis that can be tested for robustness. The initial tree generated with relaxed *trimAl* settings placed *M. haplanaria A* monophyletically between

M. enterolobii and the MIG and grouped *M. haplanaria B* with the *B* clade of the MIG. In order to test the robustness of this tree, the same tree building process was applied iteratively to the orthogroup alignments at several varying degrees of strictness; strictness being a varying threshold of alignment score and sequence similarity requirements to pass through trimming. Trees of these experiments can be found in Figure 3.5a-h.

The summary statistics of *gt0.7st0.25 were identical to those of gt0.7st0.001,* and the increase in *st* had little to no effect on the topology of the tree (Figure 3.6a). At *st0.5* the tree resolved *M. haplanaria A* and *B* into a monophyletic sister group to the MIG. Concatenated alignment length and number of undetermined characters was reduced by around 60%, indicating removal of low quality data (Figure 3.6b). *Gt0.7st0.75* is too strict for the data and constricts the topology and quality of the tree (Figure 3.6c). The manipulation of similarity threshold (*-st*) had a scalar effect on the topology of the trees produced. Higher levels of similarity threshold greatly reduced the proportion and total of parsimonious informative sites.

Resoverlap0.5seqoverlap50 causes *M. haplanaria B* to fall out of the MIG in comparison to the initial tree and places its branch very close to the root node shared between *M. haplanaria A* and the MIG (Figure 3.6d). *Resoverlap0.75seqoverlap75* has a similar effect, also causing *M. haplanaria A* and *M. haplanaria B* to switch positions relative to the initial tree, with *M. haplanaria B* falling as outgroup to *M. haplanaria A* and the MIG. The root of *M. haplanaria A* is very close to the root node shared between *M. haplanaria B* and the MIG (Figure 3.6e). Stricter trimming thresholds all pull *M. haplanaria B* towards a monophyletic position alongside *M. haplanaria A.* In trees where *M. haplanaria A* is either sister taxa or immediately basal to *M. haplanaria B,* the branch length of *M. haplanaria A* increases. Removal of all gaps from the alignment with *-noallgaps* produced a tree with *M. haplanaria B* positioned alongside the MIG subgroup *B,* with high support values (100/68.9/70.3) (Figure 3.6g).

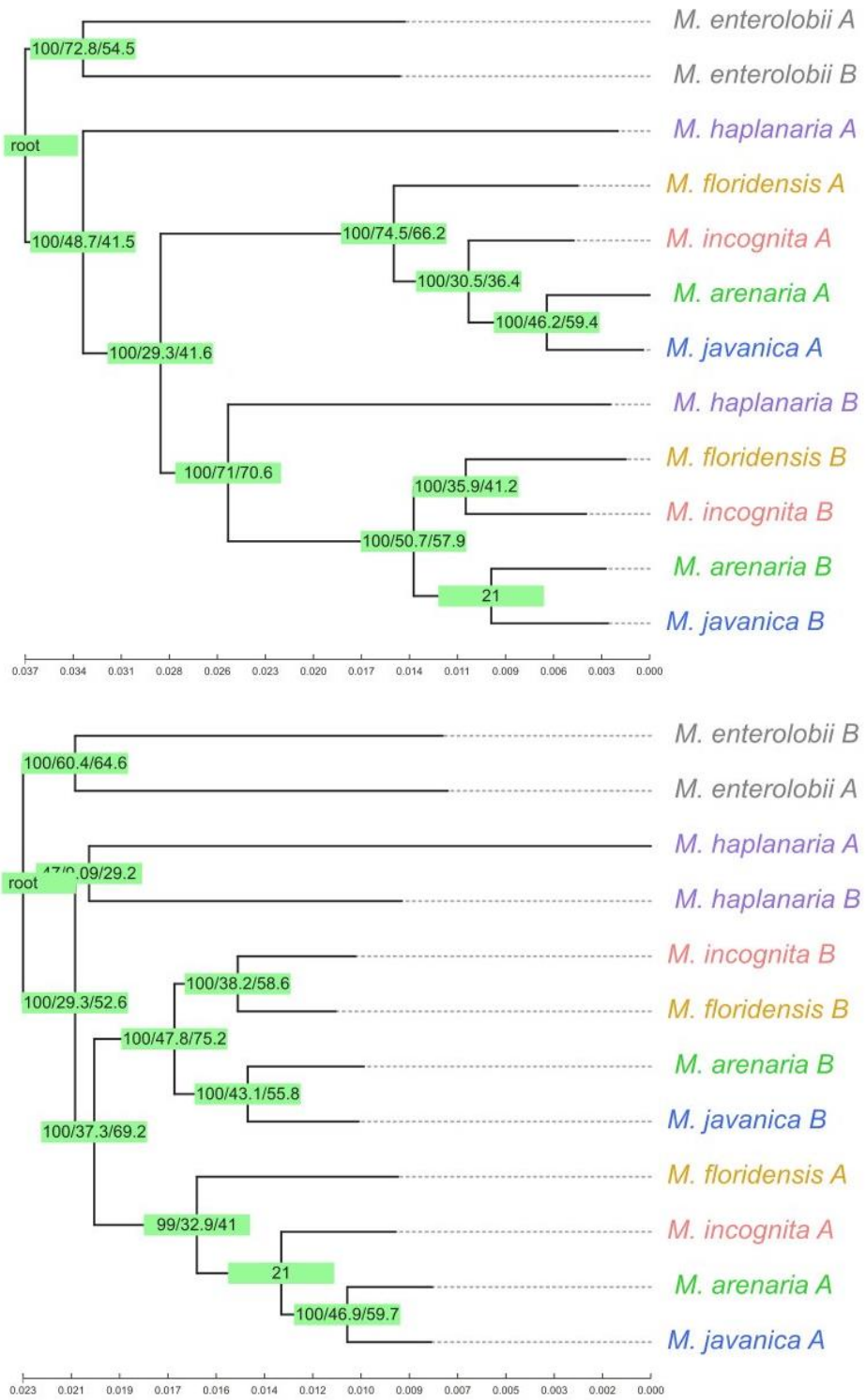

Figure 3.5a-b: Trees generated from various configurations of *trimAl* parameters. A, top, parameters *-gt 0.7 -st 0.25*. B, bottom, *-gt 0.7 -st 0.5*. Node values are bootstrap, general concordance factor, and site concordance factor. The variance of effect of *-st* (similarity threshold) is minimal below 0.5. General concordance factor (gcf) and site concordance factor (scf) of the position of *M. haplanaria B* rose with an increase in *-st*, while gcf and scf of *M. haplanaria A* fell.

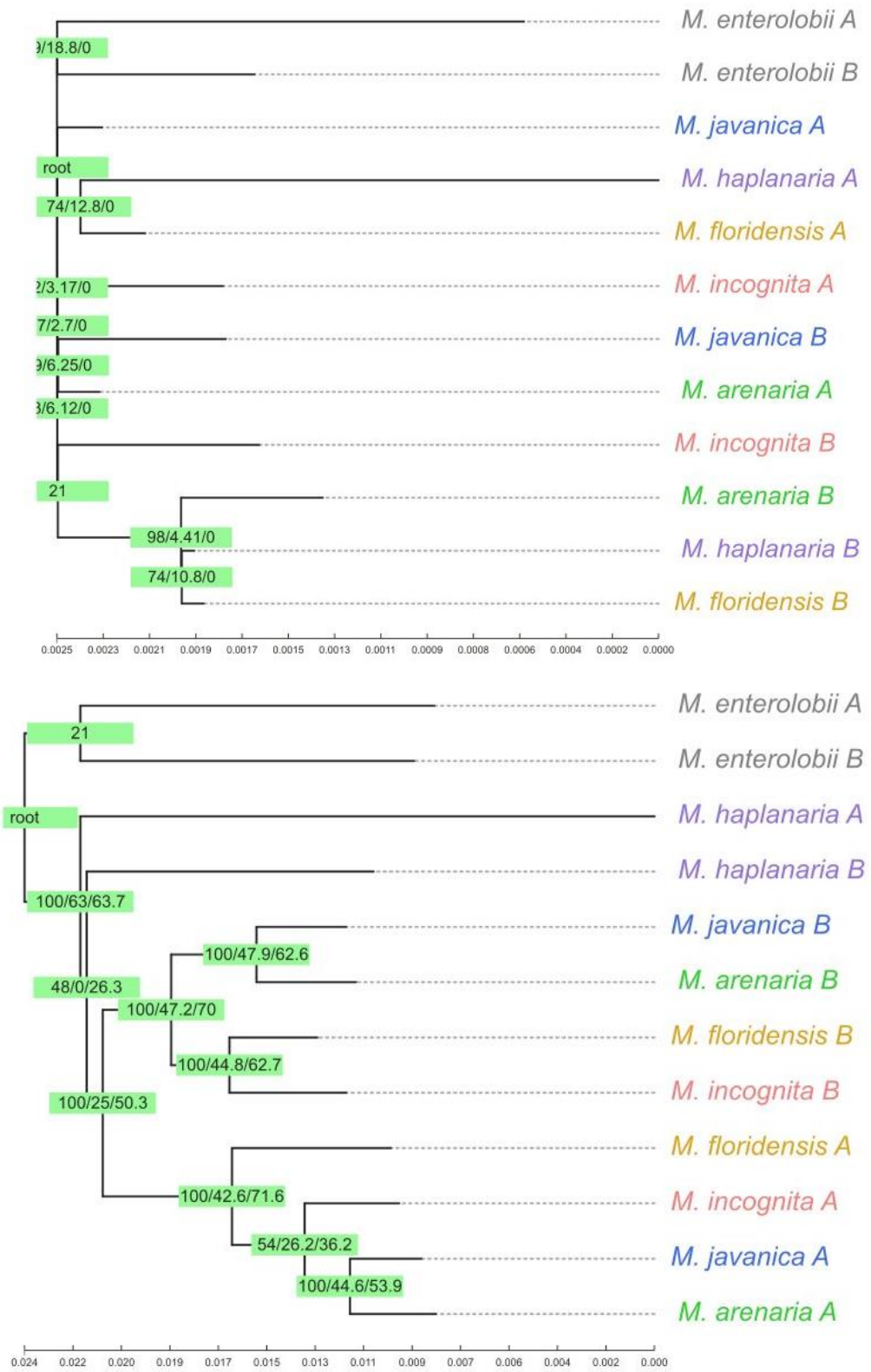

Figure 3.5c-d: Trees generated from various configurations of *trimAl* parameters. C, top, parameters *-gt 0.7 -st 0.75*. D, bottom, *-resoverlap 0.5 -seqoverlap 50.* Node values are bootstrap, general concordance factor, and site concordance factor.

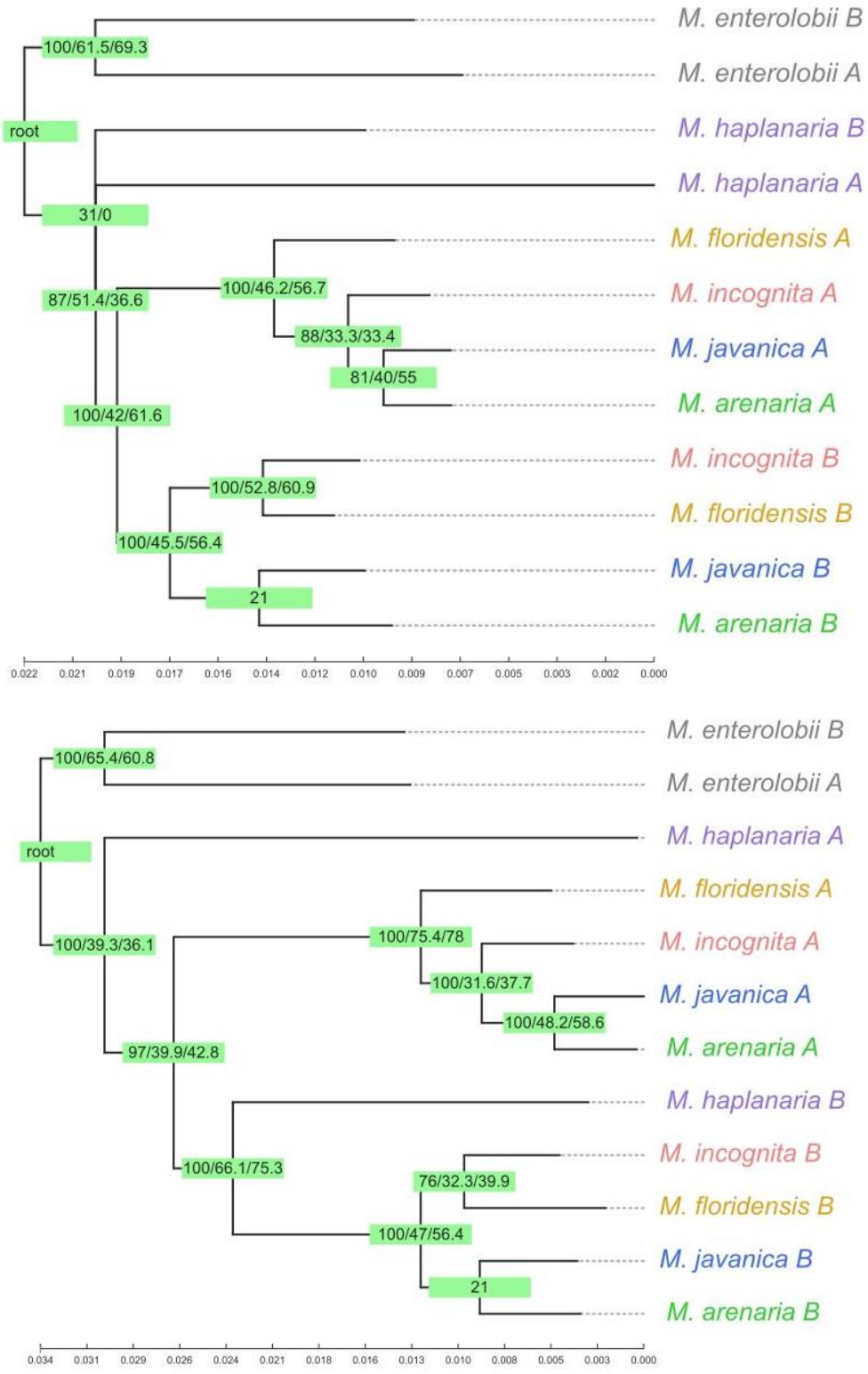

Figure 3.5e-f: Trees generated from various configurations of *trimAl* parameters. E, top, *-resoverlap 0.75 -seqoverlap 75*. F, bottom, *-nogaps*. Node values are bootstrap, general concordance factor, and site concordance factor. In F, *M. haplanaria* is positioned within the MIG with high support values.

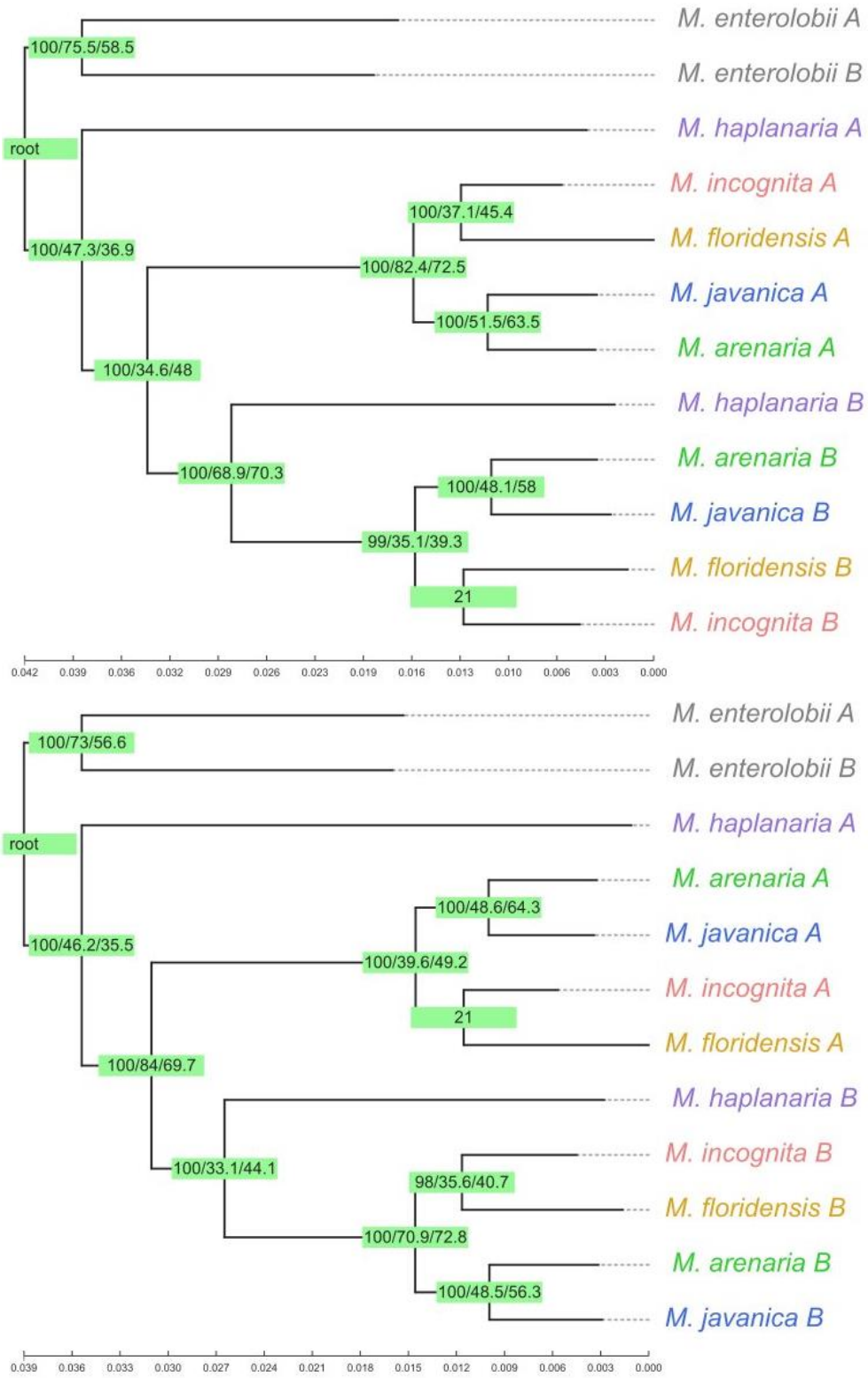

Figure 3.5g-h: Trees generated from various configurations of *trimAl* parameters. G, top, *-noallgaps*. H, bottom, *-gappyout.* Node values are bootstrap, general concordance factor, and site concordance factor. The *gappyout* tree displays strong support for *M. haplanaria B* in the MIG.

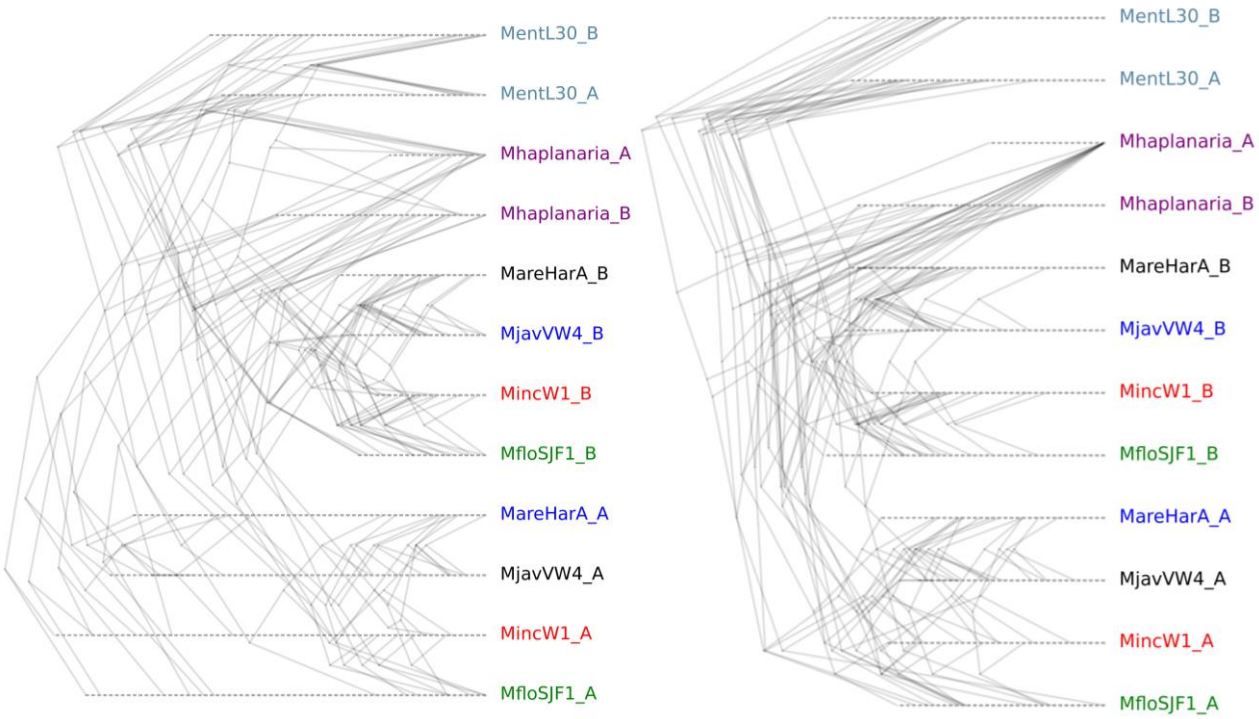

Figure 3.6a-b: Densitrees of subsampled trimmed orthogroup alignments. A, left, built using orthology alignments trimmed with relaxed *trimAl* settings *-gt 0.7 -st 0.001* and B, built using orthology alignments trimmed with stricter *trimAl* settings *-gt 0.7 -st 0.5.* Both figures are composed of 15 sub-trees, each generated with a concatenation of 40 randomly selected orthogroup alignments.

3.3.4 Quality filtering of best trimal alignment

Filtering and re-concatenating shortened the overall alignment by almost 25% (Figure 3.7a). Despite dropping orthogroup alignments with an abnormally large amount or proportion of parsimony informative sites, the overall proportion of parsimony informative sites slightly rose (Figure 3.7g). Filtering for missing percentage >15% did not significantly change the missing percent between unfiltered and filtered concatenations (Figure 3.7c). Amount of undetermined characters was not a value that was controlled for but has fallen significantly as a result to the other filters (Figure 3.7h). Filtering significantly lowered the number of undetermined characters and shortened average alignment length while maintaining proportion of parsimony informative sites.

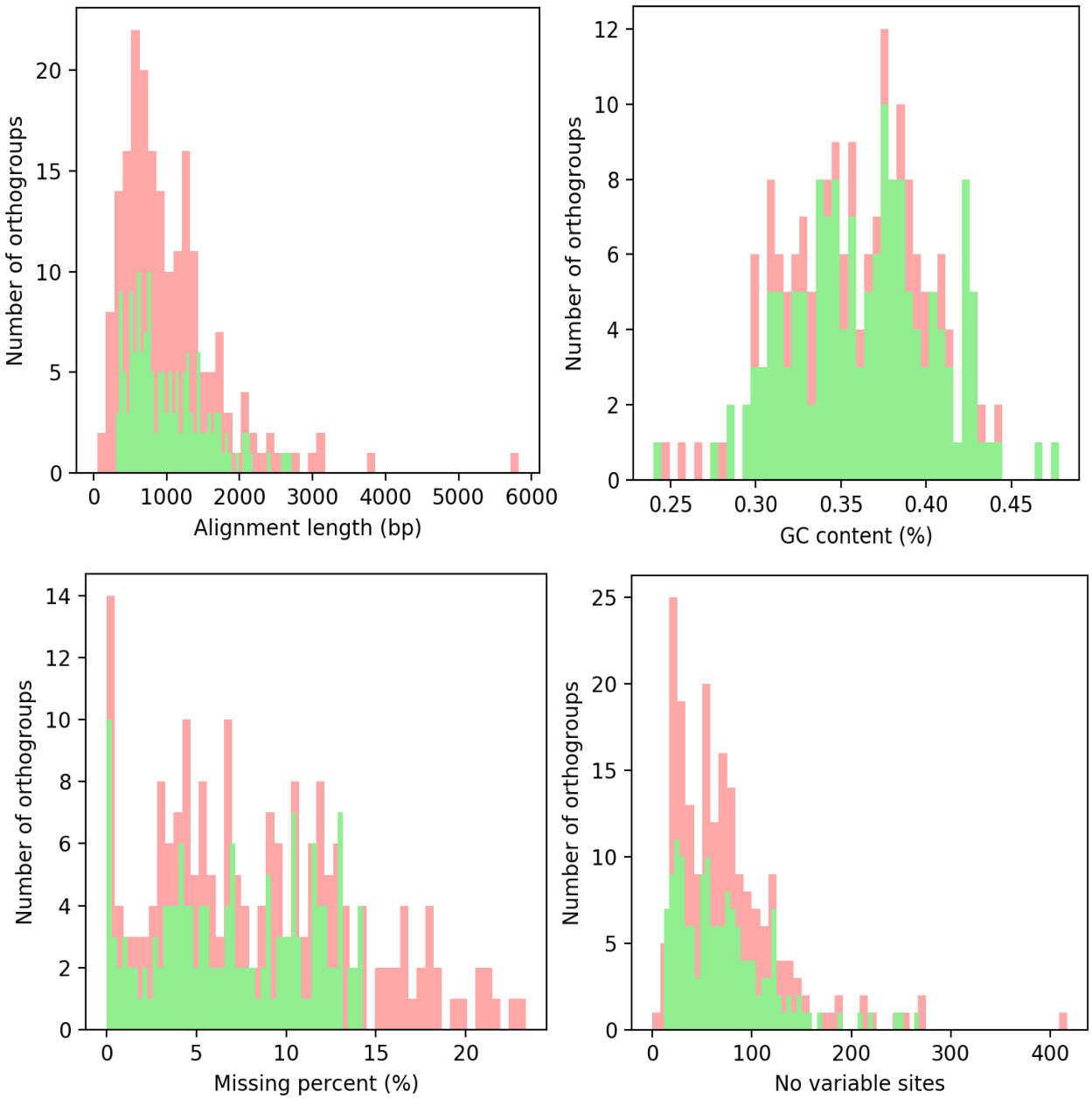

Figure 3.8a-d: Notable AMAS filtering results. A, top left, length of orthogroup alignments. B, top right, guanidine-cytosine content of orthogroup alignments. C, bottom left, shows missing percentage of the alignment. D, bottom right, shows the number of variable sites in each alignment. Red displays the results before filtering; green shows the result afterwards. GC content remained stable. Some alignments were removed due to large alignment length, but many were removed by other filters. The number of variable sites seemingly correlates to the total number of matrix cells, but removal of many through filtering has not greatly affected the number of parsimony informative sites.

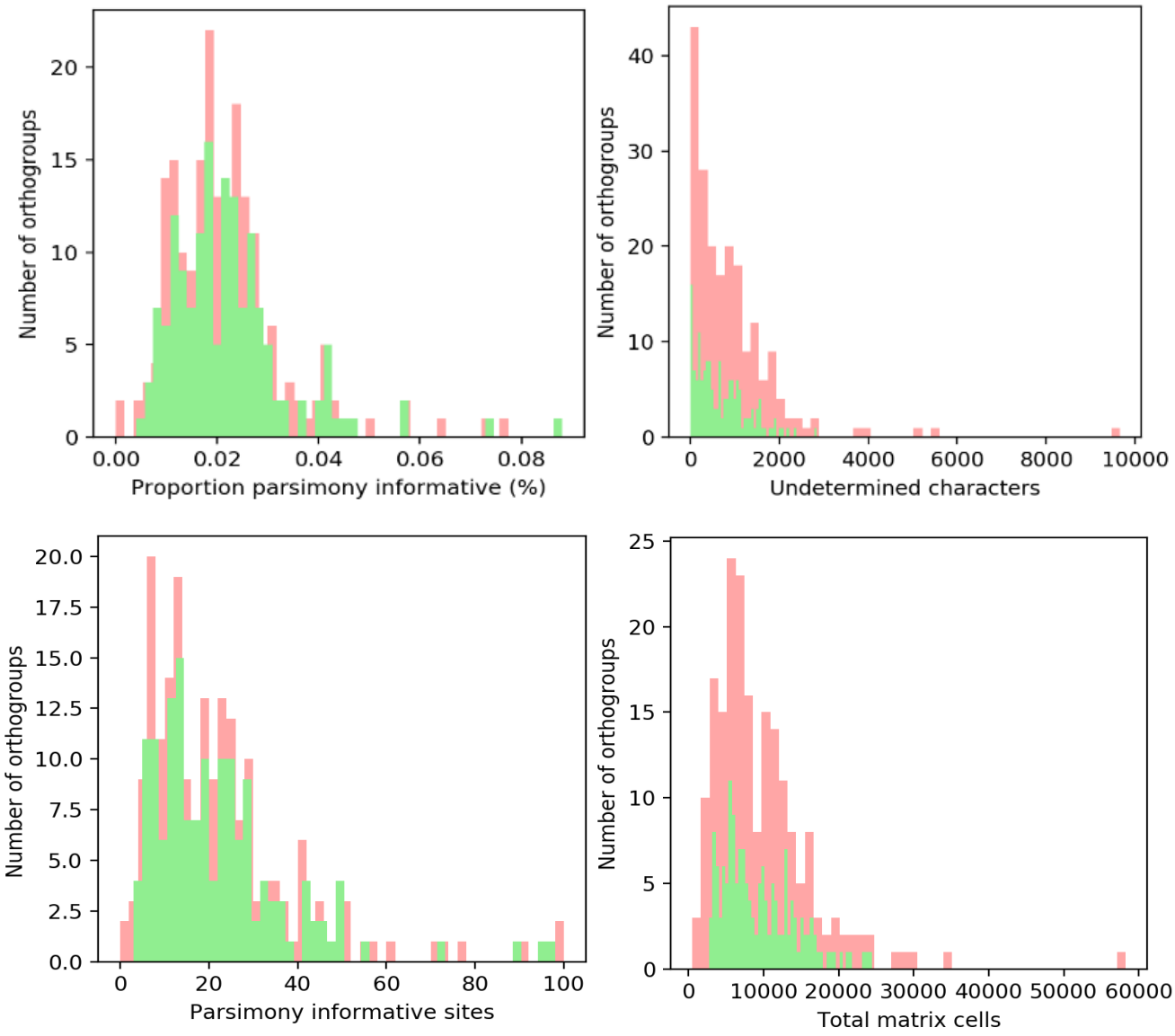

Figure 3.7e-h: Notable AMAS filtering results. G, top left, shows the proportion of alignments that is parsimony informative. H, top right, shows the number of undetermined characters in orthogroup alignments. E, bottom left, shows the number of parsimony informative sites in each orthogroup. F, bottom right, shows the total number of matrix cells in the alignment. Matrix cells represent variable base positions between taxa in the alignment. Red displays unfiltered results. Green displays filtered results. The number of undetermined characters in alignments was dramatically reduced by filtering, while the proportion of parsimony informative sites remained that same, indicating there is a lot of noise in the data, but a strong phylogenomic signal is still there.

3.3.5 Final trees

Two final trees were inferred from quality filtered alignments, using either the GTR GAMMA model (Figure 3.8) or the GHOST model (Figure 3.9) of substitution. In both trees, both *M. haplanaria* subgenomes are positioned as a monophyletic group between *M. enterolobii* and the MIG with similar confidence metrics supporting them (GTR = Bootstrap: 100, GCF: 27.9, SCF: 53.6. GHOST = Bootstrap: 100, GCF: 28.6, SCF: 51.2). These metrics are not particularly high and indicate that there is still either phylogenetic conflict or low quality data within the dataset. Filtering did raise support values of all nodes but not by a large amount.

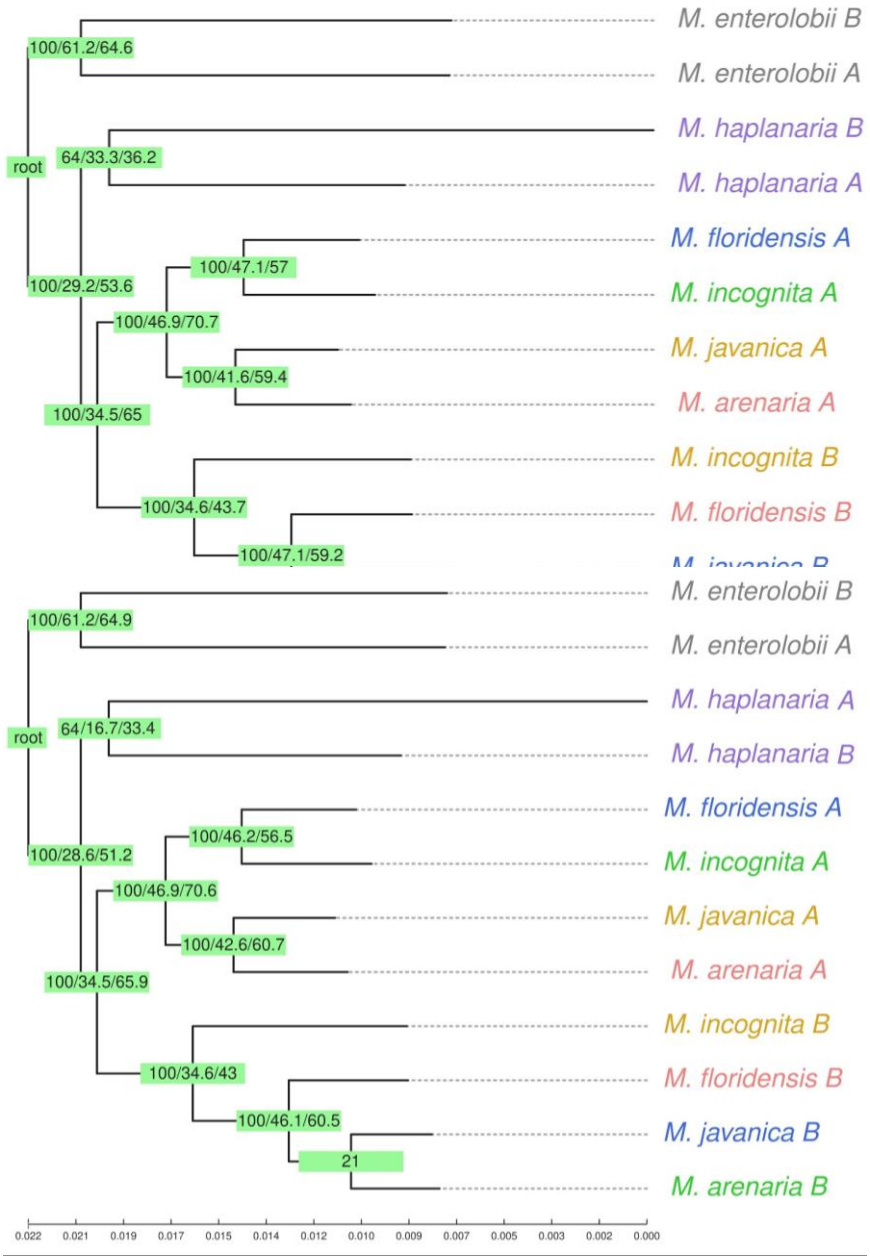

Figure 3.8: Filtered tree built with the Generalised Time Reversible (GTR) model and *trimAl* settings - *gt 0.7 -st 0.5*. Support values for monophyly are low compared to other nodes and support values in other trees.

Figure 3.9: Filtered tree built with the GHOST model. Application of the GHOST model had no effect on the topology of the tree, though support values for monophyletic *M. haplanaria* subgenomes fell.

3.3.4 Phylogenomic position of *M. haplanaria*

The orthofinder species tree placed *M. haplanaria* between *M. enterolobii* and the MIG, indicating that there is enough phylogenetic signal in each sample to correctly infer the majority of relationships concordantly with previous studies.

Relaxed trimming produces a tree with minimal quality filtering and the most available data. Though this tree cannot be considered reliable due to the inclusion of poor quality data, it provides us with a null hypothesis that can be tested for robustness; that *M. haplanaria A* is sister taxa to the MIG, between the MIG and *M. enterolobii*, and that *M. haplanaria B* falls within group *B* of the MIG (Figure 3.4a & b). This hypothesis is unlikely, despite high concordance metrics in some trees, Figure 3.5g in particular. No previous phylogenetic analysis indicates *M. haplanaria* as a member of the MIG, and factors affecting input data, as well as phylogenomic methods have not been addressed. My analysis in this chapter indicates that the poor quality and low gene representation of *M. haplanaria A* subgenome is reducing the phylogenetic similarity between *M. haplanaria A* and *B.* Subsampled trees (Figure 3.6a-b) indicate that *M. haplanaria B* is more often positioned closer to the MIG than to *M. enterolobii*, and this similarity combined with the poor quality of *M. haplanaria A* could be increasing the tendency for *M. haplanaria B* to group within with the MIG and not *M. haplanaria A*.

Another hypothesis is that the *M. haplanaria* subgenomes fall with their respective group of the MIG subgenomes; *A* with *A* and *B* with *B*. This is not strongly supported by my analyses, none of the trees produced here support these positions and increasingly strict quality constraints pull the subgenomes further away from their positions were this hypothesis correct. It is not supported by any previous species-level analyses performed here or elsewhere.

A third, more likely hypothesis is that *M. haplanaria* is a sister clade to the MIG, with monophyletic subgenomes falling between the MIG and *M. enterolobii* (Chapter 3, Figure 3.8). The results of this study seem to support this, as do some phylogenetic analyses in the literature of single locus data (Santos et al., 2019; Ye, Robbins and Kirkpatrick, 2019). Reasonably tight constraints on the quality of the orthology data, to control for the lower quality *M. haplanaria A* CDS data, cause *M. haplanaria B* to come to this position in our

analyses. However, concordance metrics for this arrangement were low, but not overly so considering support values of nodes we are confident about.

The low parsimony informative content of *M. haplanaria A* could be from issues with the *M. haplanaria* assembly. As a result of the *redundans* pipeline (Pryszcz and Gabaldón, 2016) removing some heterozygous sequences, the shortened final assembly may contain less CDS from *M. haplanaria A*. The inclusion of low quality taxa results in a high number of gaps post-alignment and can cloud the signal of genes that are actually present, impacting topology. This has the effect of lengthening the branch of that taxon and skewing the tree.

Due to absence of an *M. haplanaria* transcriptome, early sequence similarity methods of annotation using *EST2GENOME* within *MAKER2* made use of an *M. incognita* transcriptome as a reference training set. Biased CDS extraction could potentially affect tree topology. Given the position of *M. haplanaria* found in this analysis, in the continued absence of a transcriptome future studies should use a combined dataset of *M. enterolobii* and *M. incognita* as a training reference. Creating a training set from a variety of Clade I species would limit any bias in gene prediction resulting from different structural characteristics between species. This would provide a more robust and representative set of CDSs to include in phylogenomic analysis.

3.3.5 Conclusions

We have determined that the most likely position of both subgenomes of *M. haplanaria* is as a monophyletic pair between *M. enterolobii* and the MIG, and have identified that improvements in the *M. haplanaria* genome assembly are required to determine their position conclusively.

The success of an analysis like this one is limited by the quality of the input data available. Though CDS data of annotated short-read assemblies provides enough phylogenetic signal to perform a reasonably powerful analysis, the low representation of one subgenome of *M. haplanaria* prevented us from confidently positioning either homeolog. Other factors certainly exacerbated this, including indications of conflicting phylogenetic signals within the dataset.

The creation of a thorough and complete mitochondrial phylogeny of Clade I would aid in interpretation of the results of this, and future, studies. Mitochondrial phylogenies in the literature place *M. haplanaria* in conflicting positions, several of which conflict with all nuclear trees created here (Joseph et al., 2016); (Álvarez-Ortega, Brito and Subbotin, 2019), and clarification of this position would help inform hypotheses of nuclear position and evolutionary history.

Using a more distant outgroup, despite the initial loss of defined orthology groups, may be more beneficial in tree building stages. Of the over 4000 orthogroups defined early in the workflow, only between around 50 to 500 passed all stages of quality control and alignment to be used in final concatenations. Compared to this, the gain of a few hundred unaligned orthogroups is outweighed by the benefits of trees having a definitive outgroup. Given our understanding of the hybrid interactions between ancestors of Clade I and the contrasting relationships between *M. haplanaria* and *M. enterolobii* in some phylogenies from the literature, using an outgroup from Clade II may pre-emptively circumvent issues distinguishing the relevant positions of more basal members of the clade.

The inclusion of other Clade I species as and when genome assemblies are announced will undoubtedly improve the resolution of the position of *M. haplanaria*, particularly if new species are evolutionarily close. Alongside this is the inclusion of CDS data extracted from higher quality assemblies of already included species. Since the performance of this study, several such genome assemblies have been released, of *M. enterolobii, M. incognita, M. arenaria* and *M. javanica.* Thanks to the creation of the novel *snakemake* workflow used in this study, these improvements can be easily made.

Performing phylogenomic analyses from within the *snakemake* workflow proved to be efficient and easy to replicate. This lays the foundation for reproducibility of this study and accessibility to replication of this analysis. The automated and easily configurable nature of *snakemake* means that newly released *Meloidogyne* genome assemblies and annotations can be retrieved and subjected to phylogenomic analysis almost immediately, as simply as the editing of a text file and running one line of code at the command line.

Chapter 4: Discussion

In this study I set out to design and deploy novel workflows to assemble and annotate the genome of the RKN *M. haplanaria*, then test for divergent subgenomes and perform a phylogenomic analysis. This was in order to understand more about the genomic profile of *M. haplanaria* and its relationship to the phylogeny and genomic diversity of clade I of *Meloidogyne* as a whole.

4.1 Assembly and annotation of *M. haplanaria*

4.1.1 Genome assembly

The genome of *M. haplanaria* was assembled using our *snakemake* assembly workflow. The genome assembly process progressed gradually, with summary statistics reflecting successively more robust assembly methods as the workflow was developed and expanded. *Meloidogyne* genomes, and genomes of other organisms with an exceptionally complex genomic profile, are notoriously hard to assemble well, so success or failure of methods in this study must be assessed with this in mind (Richards, 2018).

Genome assembly stages produced six assemblies, for which comparative statistics were collected. As was to be expected, the quality, size, and contiguity of these assemblies varies greatly depending on the assembly method used, though the main factor affecting length and contiguity was the application of the *redundans* pipeline (Pryszcz and Gabaldón, 2016).

Contig counts range from several thousand to hundreds of thousands, all of the latter being produced by assembly methods that did not employ *redundans* to close gaps.
Assemblies of other Clade I species from the literature created using the same sequencing technology and similar methods exhibit contig counts similar to those of the *redundans* gap closed assemblies (Blanc-Mathieu et al., 2017; Szitenberg et al., 2017).

Genome length of initial assemblies ranged from 113 - 216Mb. Gap closing by *redundans* reduced assembly length by almost half, reducing the range to 63 - 163Mb. This is within the range expected based on genome length of other Clade I *Meloidogyne* species, particularly species of the MIG, whose genome lengths range from 74Mb to 164Mb (Lunt et al., 2014; Blanc-Mathieu et al., 2017; Szitenberg et al., 2017). It is important to make the distinction between a haploid and diploid assembly. A diploid assembly contains

contigs from both homeologs. Due to the way assemblies are represented as one strand, this artificially increases the length of the assembly. A haploid assembly - the kind that the *redundans* pipeline creates - does not contain both homeologs, instead removing one representative. This gives more accurate depictions of genome length and other summary statistics but reduces the power of the assembly to represent both subgenomes in orthology and phylogenomic analyses. *Genomescope 2* predicted a final haploid assembly size of 72.6Mb (Ranallo-Benavidez, Jaron and Schatz, 2020). Of the assemblies created here, *PSR* (66Mb) and *PGCR (*69Mb) most closely resemble this prediction.

Quality of assemblies was ranked on percentage of detected universal single-copy orthologs - BUSCOs (Simão et al*.*, 2015) - and metrics of completeness, N50, L50, N75, and L75 (Gurevich et al*.*, 2013; Simão et al*.*, 2015). The most complete assemblies in terms of presence of BUSCOs are *spades* (90.43%)*, spadesR* (93.73%), and *PGCR* (76.9%)*.*

The highest N50, N75, L50, and L75, were recorded for assembly *PGCR*. The lowest recorded were for assembly *PA*. This is to be expected due to the gap closing processes applied to *PGCR* connecting contigs, therefore increasing these metrics (Castro and Ng, 2017). Low scores for assembly *PA* can therefore be assumed to result from lack of gap closing. This is reflected in the total contig counts of either assembly; *PA* consists of over seventy times the amount of contigs as *PGCR.* Measures such as N50 must be interpreted carefully, as overly aggressive gap closing can artificially inflate the score (Castro and Ng, 2017). In some cases, contigs that are not contiguous *in vivo* can be forced together by overpowered gap closing packages. This artificially increases the N50 and contiguity of the assembly, leading to higher scores in quality assessment and appraisal stages for what is essentially a lower quality assembly.

Several improvements could be made to the assembly methods employed in Chapter 2, Section 2.2, the first of which is better detection and removal of contaminants. Although not commonly mentioned explicitly in published studies, all DNA extractions, including RKN DNA extractions, are contaminated with a large amount of bacterial DNA. In addition to this, RKN DNA extractions are also likely to be contaminated with human DNA from the researcher, DNA of the host plant, and fungal DNA from the soil culture. Inclusion of these contaminant reads could affect the assembly, artificially increasing the number of contigs and lowering overall coverage and contiguity. Contaminants in published

assemblies have been known to impair BLAST searches of Genbank causing incorrect taxonomic assignment of queries. We performed contaminant detection and removal on reads before assembly, but improvements to the protocol could be made. To begin, low coverage reads could be dropped prior to running *blobtools* (Laetsch and Blaxter, 2017). Whether defined as a contaminant read or not, low coverage reads add little and have the potential to reduce the quality of the assembly, so can and should be dropped during the quality control stages, rather than during contaminant removal. Secondly, many reads (>80%) were unidentified by BLAST taxonomy assignment (Altschul et al., 1990) at the thresholds used in Chapter 2. A large proportion of these reads were in or around the central cloud of blob distributions, at a good level of coverage (>10^2) and the expected proportion of GC content (~30%). Though these reads were kept in the dataset, the presence of a not insignificant amount of reads in this central distribution assigned to non-Nematode groups - Arthropoda, Rotifera, etc - that were also not expected as potential contaminants, indicates one of two things; either the taxonomy assignment stages are struggling to assign these reads correctly - leading to many Arthropoda, Rotifera, and unidentified mis-assignments within the central cloud - or there is indeed a level of Arthropoda and Rotifera contamination in the dataset.

Accurately determining taxonomy through BLAST top-hit analysis is known to be error prone (Koski and Golding, 2001). Given the unlikelihood of contamination with as high a coverage as these reads, we determined them to be mis-assignments and retained them in the dataset. Future studies could be improved to avoid these issues. Several iterations of contamination detection could be run with variable settings and algorithms, then removal based on consensus assignment. Reporting of likelihood of erroneous taxonomic assignment could also inform removal stages. An iterative approach towards contaminant removal would also have the benefit that pre-assembly of contigs would become more accurate with each iteration. In *blobtools*, it is contigs of the pre-assembly that are queried into BLAST for taxonomic assignment. With gradually lower contamination, mis-assemblies in the pre-assembly become less common, reducing the likelihood of misassignment of taxonomy. All of these improvements can be automated in *snakemake*, with feedback loops performing iterations based on previous results.

The application of *redundans* to the assembly process allowed detection of an accurate haploid genome length and reasonably better contiguity and completeness metrics than assemblies without. However, it became clear in later stages of the study that *redundans*

had caused the removal of a significant amount of homeologous sequences from the assembly, which will no doubt impair the completeness of genome annotations and the amount of phylogenetic signal detectable from each subgenome.

A possible addition to the snakemake assembly workflow is the genome assembler *platanus allee.* Unlike most genome assembly software, *platanus allee* assembles the genome whilst splitting and retaining either heterozygous haplotype, resulting in assemblies that are much more comprehensive and complete representations of the genome (Kajitani et al., 2019). Addition of this to the *snakemake* assembly workflow to produce a diploid assembly in parallel to the haploid *redundans* assembly would enable the collection of accurate genome statistics as well as generating the most informative phylogenomic data as well. *Platanus allee* seems particularly appealing for assembling *Meloidogyne* genomes, given their high heterozygosity and the desire for maximum representation of either haplotype for phylogenomic analysis of subgenomes.

4.1.2 Genome annotation

Genome annotation is a difficult process (Salzberg, 2019). Successful annotation requires the identification of the maximum number of structural characteristics possible, the quality of which can have a large impact on downstream orthology analysis, where species are compared like-with-like across genomes. This is made more difficult by the significant computational requirements of annotation, and often a lack of reference data to inform algorithms.

Annotation of the *Meloidogyne haplanaria* assembly proved difficult, with the majority of hurdles stemming from technical issues. *MAKER2* (Holt and Yandell, 2011) required the computational power of the VIPER high performance computer, meaning all software had to be installed and tested within it, a process which when combined with iteration runtimes of over a week, proved fairly intensive and continued over several months. Due to the time-consuming nature of this, subsequent iterations of annotation to increase robustness and test replicability were beyond the scope of this study, as was the creation of a *snakemake* workflow to perform each stage automatically.

The *PGCR* assembly was chosen to be used as a test set due to its shorter length reducing runtime, as well as its accurate predicted genome length and high contiguity metrics. However following annotation and clustering of divergent gene pairs it became

clear that *PGCR* lacked a large amount of data from one homeolog, likely a result of the *redundans* pipeline removing heterozygous contigs. *Redundans* (Pryszcz and Gabaldón, 2016) was initially added to the workflow to remove heterozygous contigs in order to give an accurate estimation of the genome length of *M. haplanaria* to inform comparative statistics. Following the realisation that *redundans* had limited the signal of one subgenome, an attempt was made to annotate assembly *PGC*, following the logic that if *redundans* had caused the loss of data then the *PGC* assembly and annotation would be more complete, despite its lower scoring summary metrics. Unfortunately, limitations of computational resources and time constraints prevented the success of this, and it remains an area of future investigation.

Overall, the number of genes predicted for *M. haplanaria* (20,213) falls well within the range expected for Clade I RKNs; slightly less than members of the MIG, but slightly more than *M. floridensis* (Szitenberg et al., 2017). Despite this, many divergent copies of these genes are not present due to loss during the assembly stage, which undoubtedly affected downstream phylogenomic analyses.

The creation of a snakemake workflow to perform the annotation process would be the foundational step towards future improvements. Besides just automating processes such as summary statistics collection, *MAKER2* can be built into a workflow wherein successive iterations are performed, and the output of precluding annotation runs fed to subsequent ones automatically. The parallelisation functionality of snakemake can also potentially speed up the process, an important factor given that the main constraint limiting iteration, troubleshooting, and improvement, is runtime due to limited computational power (Köster and Rahmann, 2012a).

Another factor that presumably limited the accuracy and success of the annotation stages is the lack of an available transcriptome or set of CDS data for *M. haplanaria.* In the absence of this data from the target species of annotation, sequence similarity methods of gene prediction can use mRNA data from a closely related species. In the case of this study, CDS data of *M. incognita* was used as a reference for sequence similarity gene prediction, though the data used were CDS from a short-read assembly. *MAKER2* allows this approach and is still reputably accurate with resulting predictions (Holt and Yandell, 2011), however substitution with *M. haplanaria* mRNA data or even CDS data from a higher quality *M. incognita* assembly would no doubt improve the results of these steps.

Regardless of this, informing *MAKER2* with an *M. incognita* transcriptome worked considerably well, detecting almost as many genes as MIG annotations performed with a same-species transcriptome. Future investigation could involve the creation of a reference set containing a collation of Clade I transcriptomic data, to see if this improves accuracy of gene prediction.

4.2 Genomic profiling of *M. haplanaria*

4.2.1 Testing for the presence of *M. haplanaria* subgenomes

Alongside hypotriploidy, species within the MIG contain two divergent homeologs - subgenomes. Given *M. haplanaria*'s estimated phylogenetic position close to this group, the presence of hybridity is a factor that must be investigated before an accurate phylogenomic analysis could be performed. Without this distinction, the polyphyletic subgenomes cannot be properly positioned, and an accurate history cannot be inferred.

An intragenomic sequence similarity analysis - or intragenomic blast (IGB) - found the presence of a significant number of divergent gene pairs within the *M. haplanaria* genome, indicative of the presence of divergent subgenomes, suggesting a past hybridisation event similar to that experienced by the MIG (Chapter 2, Figure 2.8) (Lunt et al., 2014; Szitenberg et al., 2017). An analysis of homeolog count within orthogroups corroborated this (Chapter 3, Figure 3.3f). We can therefore conclude that *M. haplanaria* contains divergent subgenomes akin to species of the MIG and *M. enterolobii*, and that its homeologs must be treated as individual taxonomic units throughout phylogenomic analysis.

This analysis was performed almost exactly as Lunt et al (2014) and Szitenberg et al (2017) performed it previously, however more recent algorithms and packages for sequence similarity detection have been developed (Bermúdez, 2019; Wood, Lu and Langmead, 2019). Experimentation with these packages may lead to a more accurate or informative method. The BLAST results generated as part of this analysis contain information on gene ontology. We believe that there is much information to be found through investigation of this ontology, the function of which genes are present in divergent pairs, and how the presence and heterozygosity of these genes differs across species in relation to parasitic adaptation and ecology. The precursory heterozygosity of these

homeologs could accelerate the sub- or neo-functionalisation processes, contributing to the short amount of time hybrid species require to break cultivated resistance in the host.

4.2.2 Interpretation of ploidy

We attempted to identify the genomic profile of *M. haplanaria.* Many Clade I species of *Meloidogyne* have been suggested to be hypotriploid; a portion of one copy of their genome is duplicated in a 2n + <n configuration. To investigate whether this is the case for *M. haplanaria*, we performed a ploidy analysis using several methods of genome profiling, including *Genomescope 2* and *smudgeplot* (Ranallo-Benavidez, Jaron, and Schatz, 2020). Both analyses indicated some level of triploidy, though the loose model fit of *Genomescope 2* results combined with variation of heat in the smudgeplots leads us to conclude that *M*. *haplanaria* is hypotriploid, not wholly triploid.

Given the information generated here, in comparison to assemblies of other *Meloidogyne* species in the literature, *M. haplanaria* most closely resembles *M. floridensis*. Their genome lengths, both predicted and observed, are very similar (*M. floridensis*: 74.9Mb vs *M. haplanaria*: 69.4Mb), as is the GC content percentage (*M. floridensis*: 30.2% vs *M. haplanaria*: 30.13%). Apomictic members of the MIG have GC content percentages from 29.5% to 29.9% (Chapter 2, Table 2.4). *M. haplanaria* and *M. floridensis* also show similar results from the intraspecific sequence similarity analysis, with both species peaking to around the same proportion of gene pairs - *M floridensis* showing slightly more. Estimates of gene copy number between *M. floridensis* and *M. haplanaria* are also very similar.

Some studies have proposed that the reason for a reduced number of heterozygous gene pairs in *M. floridensis* is automixis pushing the genome towards homozygosity (Handoo et al., 2004; Lunt et al., 2014; Szitenberg et al., 2017). If estimates of heterozygosity and copy number are accurate, it may be possible that the reduced number of homoeologous pairs in *M. haplanaria* is a result of a similar process, as the sexual system of *M. haplanaria* is yet to be determined.

An alternative and more likely explanation of these findings is that the removal of some heterozygous sequences through the *redundans* algorithm has removed much of one copy, *M. haplanaria A*. This would explain many of the results seen here in terms of the number of heterozygous gene pairs, meaning *M. haplanaria* potentially contains a much higher amount and has a higher level of heterozygosity overall. The strength and

accuracy of the results of these types of analysis are highly dependent on the quality of the genome assembly used as a foundation to perform them (Young and Gillung, 2019).

Further research should include a ploidy analysis similar to that performed by Szitenberg et al (2017), wherein copy number is inferred from read mapping percentages to either divergent homeolog, would add support to the determination of hypotriploidy in *M. haplanaria* made here. As well as this, optimisation of genome assembly to prevent collapsing or removal of homeologs will reveal the degree to which *M. haplanaria*'s subgenomes are heterozygous more accurately, alongside retaining as much heterozygosity as possible to ensure a strong phylogenetic signal from both homeologs. A chromosome level assembly, generated from exceedingly high quality long-read sequence data would greatly contribute to determination of ploidy, again highlighting the advantages and need for long-read data.

4.3 Phylogenomic analysis

We attempted to determine the phylogenetic position of *M. haplanaria* to test the accuracy and efficacy of our phylogenomics *snakemake* workflow. The workflow performed well without the need for a high amount of computational resources. It ran locally on several hardware configurations as low as 4 cores and 16 gigabytes of RAM, well within the range of the average user setup. Though runtime will predictably scale with the size of the dataset, the workflow is accessible to the non-power user wishing to run a quick analysis.

4.3.1 Improvements to methods and workflow

Future modules and features to improve the workflow, besides supplementary documentation, could include performance of a coalescent analysis alongside the concatenation analysis (Mirarab et al., 2014). Low support values and incongruence would show the possibility of phylogenetic conflicts within the dataset, which would lead to further testing and refinement.

The creation of a Robinson-Foulds distance matrix would allow testing for phylogenetic conflicts within the data (Reddy et al., 2017), such as that seen within subsamples densitrees (Chapter 3, Figure 3.7) and between differently trimmed trees (Chapter 3,

Figure 3.6a-h). Phylogenetic conflicts can arise when orthogroups within the dataset exhibit contrasting topology, and hybridisation is thought to be a primary contributor (Smith et al., 2015). Conflict can be controlled in the dataset through subsetting and binning of conflicting gene trees in a coalescent approach.

Features to reiterate over rules, such as repeating alignment trimming stages with different parameters, can be added to the workflow (Borowiec et al., 2015). This study performed one form of alignment trimming for each tree, whereas a combination of trimming parameters may more sufficiently control the data. Studies by Borowiec et al (2015 & 2016) perform trimming based on sequence overlap then retrim to remove all columns containing all gaps, two approaches that were employed separately here producing conflicting trees (Chapter 3, Figure 3.6b and 3.6.g).

Aside from changes to the workflow, improvements can be made to the data used. Using input CDS data of *M. haplanaria* generated from an assembly not cleaned by *redundans* would likely result in a more robust and higher confidence tree, particularly for subgenome *M. haplanaria A,* due to homeologs from this copy being removed from the assembly by *redundans.*

4.3.2 Improvements to data

Likewise, the availability of long-read sequence data to generate a more complete assembly. Several new genome assemblies have been published of different Clade I species since this study began that have employed long-reads, including a novel assembly of *M. luci* and much higher quality genomes of *M. enterolobii,* and *M. incognita* (Asamizu et al., 2020; Danchin GJ, 2020; Susič et al., 2020). Though the comparative statistics of some of these assemblies were included in Table 2.4, CDS annotations from these assemblies were not used in orthology definition stages of the phylogenomic analysis. CDS annotations generated from long-read assemblies will be potentially more complete than those of short-read assemblies due to a long-read assembly generally being more contiguous and complete themselves, leading to the inference of more orthology groups, which increases the resolution of the final tree (Amarasinghe et al., 2020).

Increased taxon sampling generally improves the support values and confidence of an inferred phylogeny. This will be especially informative in further defining the position of either *M. haplanaria* subgenome, as the genomes of species closest to it in mitochondrial and ribosomal phylogenies, such as *M. ethiopica* and *M. hispanica,* are yet to be assembled. Alongside this the growing availability of long-read assemblies of taxa not previously included in phylogenetic analyses of *Meloidogyne* will ultimately generate trees with higher node support, increasing our confidence of positions across the phylogeny.

4.3.3 Phylogenetic position of *M. haplanaria*

We performed a phylogenomic analysis to identify the phylogenetic position of *M. haplanaria* subgenomes using genome scale data. This was in order to progress beyond simple systematics and be able to determine and interpret the position of either divergent homeolog. The phylogeny of *Meloidogyne* subgenomes is convoluted and difficult to resolve due to seemingly many past hybridisation events (Lunt et al., 2014).

We found support in all trees for both subgenomes of *M. haplanaria* being more distal - further from the root of the phylogeny - than *M. enterolobii.* This is in conflict with some phylogenies generated by studies using mitochondrial and ribosomal data where *M. haplanaria* is positioned as sister taxa to *M. enterolobii* (Joseph et al., 2016; Álvarez-Ortega, Brito and Subbotin, 2019; Santos et al., 2019), but in agreement with others (Le et al., 2019; Ye, Robbins and Kirkpatrick, 2019).

The position of *M. haplanaria B* varied between being within the *B* group of the MIG, or being a sister taxa to *M. haplanaria A,* basal - nearer the root of the phylogeny - to the MIG, depending on the alignment trimming and quality control measures used (Figure 4.1a-b). Across all trees the greatest support from both general concordance factors and site concordance factors was for the former; *M. haplanaria* subgenomes as a polyphyletic group, with one subgenome positioned with group *B* of the MIG and the other subgenome distal to the MIG and basal to *M. enterolobii* (Figure 4.1a). If trees positioning *M. haplanaria B* within the MIG are accurate, it would indicate that the hybridisation events giving rise to *M. haplanaria* and the MIG involved one of the same ancestors. This also conflicts with mitochondrial and ribosomal trees, in which *M. haplanaria* is always positioned as an outgroup to the MIG, making the position found here of *M. haplanaria B* within the *B* group of the MIG surprising.

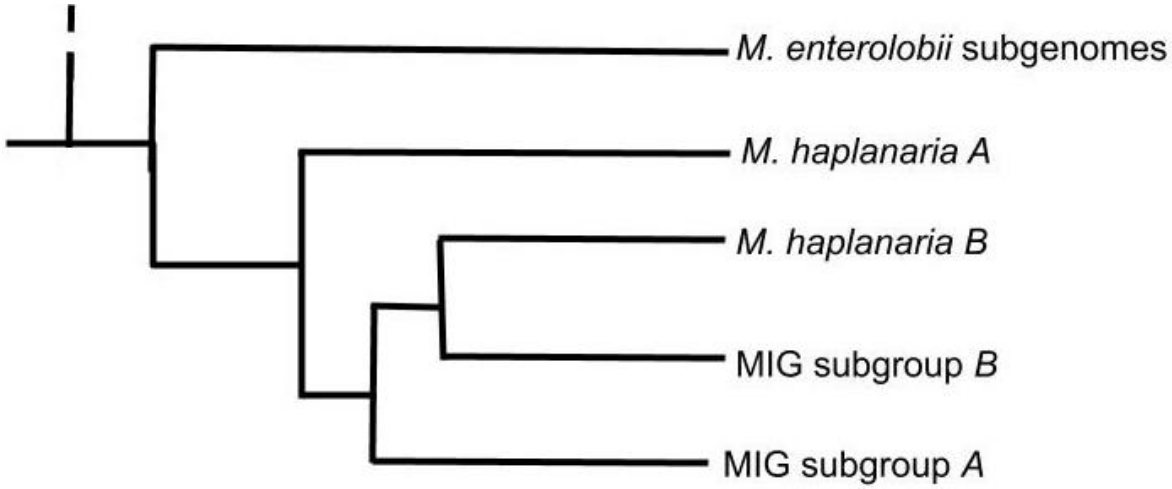

Figure 4.1a: Basic diagram of the polyphyletic *M. haplanaria* hypothesis. *M. haplanaria B* is positioned alongside group *B* of the MIG. This suggests that the ancestor of *M. haplanaria B* is also a parent of the MIG. The likely explanation for this topology is that *M. haplanaria A* is inaccurate due to the effect of long-branch attraction, and its genuine biological position is alongside group *A* of the MIG. This hypothesis is supported by higher confidence values than the other, but conflicts with mitochondrial and ribosomal trees in the literature.

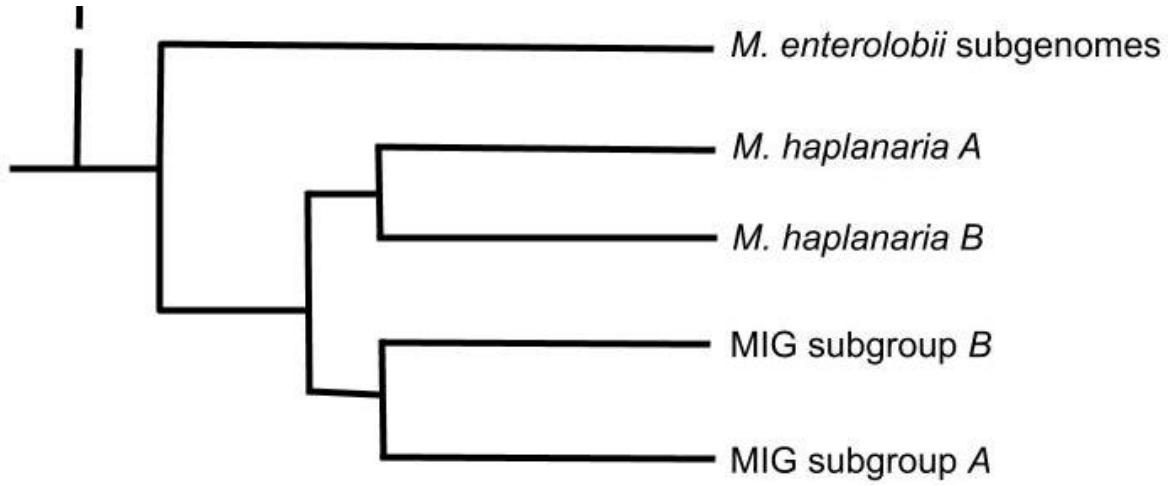

Figure 4.1b: Basic diagram of the monophyletic *M. haplanaria* hypothesis. *M. haplanaria B* and *M. haplanaria A* are positioned as a monophyly between *M. enterolobii* and the MIG, suggesting that *M. haplanaria* arose from a separate hybridisation event to either that created *M. enterolobii* or the MIG. This tree has less support from concordance metrics than the previous one but is more supported by mitochondrial and ribosomal phylogenies in the literature.

An alternative explanation for these findings is that the poor quality of *M. haplanaria A* is increasing the inferred phylogenetic distance between itself and *M. haplanaria B*, creating a situation where *M. haplanaria B* has a higher sequence similarity to the MIG. Conversely, the poor quality and long branch length of *M. haplanaria A* could be forcing it to position more basally in the tree through the effect of long branch attraction (Bergsten, 2005), when it's actual phylogenetic position is within group *A* of the MIG.

The low quality and low phylogenetic signal of *M. haplanaria A* has most likely dampened the detection of many homeologs in what is in reality a highly heterozygous genome. This is supported by the fact that even with the removal of heterozygous contigs by *redundans*, there remains enough of a divergence in the phylogenomic signal to detect a high amount of heterozygosity and multiple homeologs, separate the homeologs into clusters and find polyphyletic topologies for many orthogroups. Just as Szitenberg et al. (2017) found with the MIG, the discovery of potentially polyphyletic positions for either subgenome would not have been revealed without using the approach applied here of splitting the subgenomes into individual taxonomic units.

Thanks to the development of the phylogenomics *snakemake* workflow to perform and replicate this protocol, the analysis can be repeated with wider taxon sampling and the inclusion data from other species with very little effort. There are currently several other known *Meloidogyne* species that are positioned phylogenetically according to mitochondrial and ribosomal analyses either within the MIG - *M. hispanica, M. luci* and *M. ethiopica* - or between the MIG and *M. enterolobii,* very close to *M. haplanaria.* Genome assemblies are not yet available for any of these species except *M. luci,* but due to the design of the *snakemake* workflow, they can be included in a similar analysis with ease as soon as they become available. As well as defining the phylogenetic position of any newly assembled *Meloidogyne* species, their addition will increase the robustness and validity of the positions of other species in the tree. Increased taxon sampling can offset the effect of long branch attraction. Long branch attraction occurs when a taxon in a phylogeny is of low quality. The reduced sequence similarity between the low quality taxon and its genuine biological sister taxa force the low quality taxon basally towards the root of the tree. This is possibly the case with *M. haplanaria A* in the phylogenies produced here; increased taxon sampling would prevent this. As the wider database of *Meloidogyne* genomes expands, methods such as those performed by the phylogenomics *snakemake* workflow can only become more accurate.

Though the exclusion of *M. hapla* as an outgroup from the phylogeny increases the number of orthogroups defined, this was not by a large amount, and the inclusion of *M. hapla* may change the relationship between *M. haplanaria* and *M. enterolobii* given that some mitochondrial analyses propose they could be sister taxa (Wilberg, 2015; Joseph et al., 2016). Other than the addition of more clade I species, several things could improve resolution and robustness of the tree and increase confidence in the position of *M. haplanaria.* A more complete - albeit more fragmented - genome assembly would include more homeologous information, leading ultimately to *A* and *B* taxonomic units of *M. haplanaria* containing a greater phylogenetic signal. Higher quality sequences would also limit the effect of long branch attraction.

4.4 Increasing importance of *M. haplanaria*

M. haplanaria is considered an emerging crop pest (Joseph et al., 2016). Since its identification and classification in 2003 it has been found in several other states of the USA and has been shown to be capable of infecting resistant plant cultivars containing the *Mi1* resistance gene, something many MIG species cannot do, indicating that *M. haplanaria* has the adaptive potential to become prolific (Eisenback, 2003; Ye et al., 2019). Despite this, few studies investigating the biology and agricultural impact of *M. haplanaria* have been performed. The potential agricultural and economic impact of *M. haplanaria* can be given context by comparison to other clade I species. *M. javanica, M. incognita,* and *M. arenaria*, all RKNs of the MIG, have been isolated from several continents and has been reported as far north as Germany and the UK (CABBI and EPPO, 2002; CABBI and EPPO, 2003; Tesařová et al., 2003). Some have suggested that increasing global temperatures resulting from the climate crisis could increase the proclivity of these damaging species for temperate regions, leading to increased infection rates and populations in North America and Northern Europe (Elling, 2013). We know from the findings in this study that *M. haplanaria* is at the least very closely related to these MIG species. As of yet *M. haplanaria* has only been isolated from sub-tropical regions of North America but increasing temperatures in temperate latitudes could facilitate range expansion in a similar way.

4.5 Future directions for *Meloidogyne* genomics and phylogenomics

4.5.1 Workflow based bioinformatics

We designed and created *snakemake* workflows to perform methods for genome assembly and phylogenomics. The difficulties surrounding annotation stages of this study prevented the creation of a *snakemake* workflow to perform these steps automatically. The difference in accessibility and difficulty of executing parts of the analysis with and without *snakemake* workflows was stark. Being able to easily iterate through genome assembly methods and automation of several non-variable stages greatly reduced the amount of time each analysis took. In contrast, the annotation stages of the study proved time and manually intensive, owing to the lack of automation provided by a workflow structure. *Snakemake* provided a framework for executing methods that was much faster, reproducible, and overall easier to use and record than executing methods at the command line, as was done for the annotation methods, or through using a Jupyter notebook as in the intragenomic BLAST analysis. The protocol in the workflows is designed around best practices for either analysis, removing the need to re-establish them by end users.

Workflows were run and tested many times throughout this study as experiments with different parameter configurations were performed. The running of these iterations required very little input from the researcher; the editing of a few words in a text file and entry of a single command into the terminal. As well as this ease of use, the combination of a *snakemake* workflow with package management through *conda* ensured each replication was performed in an identical virtual environment and an identical way, other than the manually changed parameters. Replication using *snakemake* workflows is accurate, accessible, and fast.

4.5.2 Collaborative bioinformatics

There is a lack of specialised workflows and software to handle *Meloidogyne* genomes, most probably due the complexity of the genus, and in part the niche demand for such tools. We have made the workflows developed here publicly available in the hope that other researchers may find them useful, and that the community can contribute to their

future development and improvement. Hosting the workflows on Github allows collaboration and increases the transparency of their function and development (Dabbish et al., 2012).

Alongside reproducing the analyses performed here, the *snakemake* workflows we created can be used to assemble other *Meloidogyne* species and reproduce phylogenomic analysis inclusive of novel *Meloidogyne* genomes almost immediately upon them being assembled or published. Both processes require very little configuration and are accessible to researchers with little bioinformatics or programming experience, and in the case of the phylogenomics workflow few computational resources. The automated nature of *snakemake* also means that iterations can be run in parallel, providing the required computational power, further accelerating the speed of analysis (Köster and Rahmann, 2012b). We expect that many researchers in this field will find these workflows incredibly useful and efficient.

As well as inclusion of newly assembled taxa, *snakemake* supports the inclusion of additional modules and rules (Köster and Rahmann, 2012b). Further stages of analysis, ranging from simple plotting or data parsing to complex pipelines, can be easily appended to either *snakemake* workflow with little knowledge of programming. In the case of the assembly workflow, new assembly algorithms or quality control packages can be added as they are released, maintaining the workflow at the cutting edge of assembly technology. For the phylogenomics workflow, other methods of tree inference or trimming could be added according to preference, and modules can be added to plot trees automatically using personally tailored scripts.

4.5 Conclusions

In conclusion, we find several things. First, we find that *Meloidogyne haplanaria* contains divergent homeologous subgenomes. The amount of homeologous pairs detected is similar to *M. floridensis*, though the count may be erroneously low due to aspects of the assembly process removing some heterozygous regions.

This state of divergence was most likely gained as the result of a past hybridisation event, supported by our finding that *M. haplanaria* contains hypotriploid genomic architecture

like that of the MIG. Like the MIG, this presumably began as a full homozygous copy gained from a hybridisation event, followed by degradation through segmental loss and gene conversion.

We were unable to confidently position both subgenomes within the *Meloidogyne* phylogeny due to conflicting phylogenetic signals (Reddy et al., 2017). Most trees were congruent with mitochondrial analyses, placing both subgenomes as a monophyly distal to *M. enterolobii* and basal to the MIG. Evidence was found that the correct position of the *M. haplanaria B* subgenome may be within group *B* of the MIG, indicating a possible hypothesis of *M. haplanaria* being a member of the MIG and arising from the same ancestral hybridisation event. If this were the case it would mean that the position of *M. haplanaria A* in phylogenies produced in this study is incorrect, most likely due to poor quality causing long branch attraction towards *M. enterolobii,* pulling its inferred position out of the MIG (Bergsten, 2005). This hypothesis is also in conflict with mitochondrial and ribosomal phylogenies, none of which place *M. haplanaria* here, though is worthy of future investigation with higher quality data. However, if *M. haplanaria* is indeed polyphyletic, or monophyletic and basal to the MIG, the presence of divergent homeologs would imply that hybridisation between ancestors of some clade I species may be more common than previously thought.

The resolution of any phylogenomic tree produced through this study is limited by the amount of phylogenetic signal that can be extracted from genomes and assemblies produced with short-reads (Young and Gillung, 2019). Reassembly with more iterative refinement of parameters or additionally, long-read sequence data of *M. haplanaria,* would greatly increase the quality of its assembly, especially in terms of completeness and contiguity. This would allow more accurate interpretations of genome size, composition, and ploidy state. Application of the *redundans* pipeline severely limited the representation of a large percentage of homeologs CDSs from one *M. haplanaria* subgenome. This was not realised until it was too late, and as a result one *M. haplanaria* subgenome is low quality, impacting the tree and potentially causing long-branch attraction errors. *Redundans* provided us with a haploid representation of the genome, however future annotation and phylogenomic analysis of *Meloidogyne* species should be performed with an un-collapsed assembly.

Novel transcriptomic data of *M. haplanaria* would increase the accuracy of gene prediction software, especially in collaboration with a higher quality assembly, leading to an increased number of heterozygous CDS extracted, resulting in more accurate estimations of heterozygosity and frequency of divergent gene pairs. Detection of a higher quantity of heterozygous CDS would increase the quality and phylogenetic signal of the *M. haplanaria A* subgenome, resulting in a more accurate positioning within the tree and more accurate concordance metrics.

A dearth of taxa around *M. haplanaria,* is also limiting the resolution and may be allowing the effect of long-branch attraction to skew the tree, but lack of available assemblies in the literature prevents inclusion of potential additional species in an analysis this data-rich. Thankfully, with the use of workflows designed here, species can be included as soon as they become available.

The outcome of this study that we expect will provide the greatest contribution to the field is the generation of the *Meloidogyne* genome assembly and phylogenomics workflows. The creation of these workflows has laid the foundations for fast, accurate, and reproducible investigation and analysis into a taxonomic group whose notoriously convoluted genomic architecture has historically reduced accessibility to all but committed bioinformaticians. With the modular, configurable nature of *snakemake*, novel genomes can be analysed almost as soon as they are generated with very minimal effort, and further analyses can be appended as needed.

5.0 Bibliography

Abad, P. *et al.* (2008) 'Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita', *Nature biotechnology*, 26(8), pp. 909–915.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of molecular biology*, 215(3), pp. 403–410.

Álvarez-Ortega, S., Brito, J. A. and Subbotin, S. A. (2019) 'Multigene phylogeny of root-knot nematodes and molecular characterization of Meloidogyne nataliei Golden, Rose & Bird, 1981 (Nematoda: Tylenchida)', *Scientific reports*. nature.com, 9(1), p. 11788.

Amarasinghe, S. L. *et al.* (2020) 'Opportunities and challenges in long-read sequencing data analysis', *Genome biology*, 21(1), p. 30.

Anaconda Software Distribution. *Conda*. Version 2-2.4.0, Anaconda, Nov. 2016. Computer Software. *Anaconda*, www.anaconda.com.

Andersson, D. I., Jerlström-Hultqvist, J. and Näsvall, J. (2015) 'Evolution of new functions de novo and from preexisting genes', *Cold Spring Harbor perspectives in biology*, 7(6). doi: 10.1101/cshperspect.a017996.

Andrews, S. and Others (2010) 'FastQC: a quality control tool for high throughput sequence data'. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Ané, C. *et al.* (2007) 'Bayesian estimation of concordance among gene trees', *Molecular biology and evolution*, 24(2), pp. 412–426.

Asamizu, E. *et al.* (2020) 'Root-knot nematode genetic diversity associated with host compatibility to sweetpotato cultivars', *Molecular plant pathology*, 21(8), pp. 1088–1098.

Bankevich, A. *et al.* (2012) 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *Journal of computational biology: a journal of computational molecular cell biology*. liebertpub.com, 19(5), pp. 455–477.

Bao, W., Kojima, K. K. and Kohany, O. (2015) 'Repbase Update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA*, 6, p. 11.

Bebber, D. P., Holmes, T. and Gurr, S. J. (2014) 'The global spread of crop pests and pathogens', *Global ecology and biogeography: a journal of macroecology*, 23(12), pp. 1398–1407.

Bendezu, I. F., Morgan, E. and Starr, J. L. (2004) 'HOSTS FOR MELOIDOGYNE HAPLANARIA', *Nematropica*, pp. 205–210.

Bergsten, J. (2005) 'A review of long-branch attraction', *Cladistics: the international journal of the Willi Hennig Society*, 21(2), pp. 163–193.

Bermúdez, J. (2019) 'SLAST: Simple Local Alignment Search Tool'. doi: 10.1101/840546.

Bernard, G. C., Egnin, M. and Bonsi, C. (2017) 'The Impact of Plant-Parasitic Nematodes on Agriculture and Methods of Control', in Shah, M. M. and Mahamood, M. (eds) *Nematology - Concepts, Diagnosis and Control*. InTech.

Betsuyaku, S., Sawa, S. and Yamada, M. (2011) 'The Function of the CLE Peptides in Plant Development and Plant-Microbe Interactions', *The Arabidopsis book / American Society of Plant Biologists*, 9, p. e0149.

Bickel, P. J. *et al.* (2009) ‘An overview of recent developments in genomics and associated statistical methods’, *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1906), pp. 4313–4337.

Bird, D. M. *et al.* (2009) ‘The genomes of root-knot nematodes’, *Annual review of phytopathology*, 47, pp. 333–351.

Blanc-Mathieu, R. *et al.* (2017) ‘Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes’, *PLoS genetics*, 13(6), p. e1006777.

Blanc-Mathieu, R., Perfus-Babeoch, L. and Aury, J. M. (2016) ‘Peculiar hybrid genomes of devastating plant pests promote plasticity in the absence of sex and meiosis’, *BioRxiv*. biorxiv.org. Available at: https://www.biorxiv.org/content/10.1101/046805v1.abstract.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics* , 30(15), pp. 2114–2120.

Borowiec, M. L. *et al.* (2015) ‘Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa’, *BMC genomics*, 16, p. 987.

Borowiec, M. L. (2016) ‘AMAS: a fast tool for alignment manipulation and computing of summary statistics’, *PeerJ*, 4, p. e1660.

Bradnam, K. R. *et al.* (2013) ‘Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species’, *GigaScience*, 2(1), p. 10.

Burns, A. R. *et al.* (2017) ‘The novel nematicide wact-86 interacts with aldicarb to kill nematodes’, *PLoS neglected tropical diseases*, 11(4), p. e0005502.

Butlin, R. K., Schön, I. and Griffiths, H. I. (1998) ‘Introduction to reproductive modes’, *Sex and parthenogenesis: evolutionary ecology of reproductive modes in non-marine ostracods*. Blackhuys Publishers, Leiden, The Netherlands, pp. 1–24.

CABBI and EPPO (2002) *Meloidogyne javanica. [Distribution map]*. CAB International. Available at: https://www.cabi.org/isc/abstract/20066500855.

CABBI and EPPO (2003) *Meloidogyne arenaria. [Distribution map]*. CAB International. Available at: https://www.cabi.org/isc/abstract/20066500900

Caillaud, M.-C. *et al.* (2008) ‘Root-knot nematodes manipulate plant cell functions during a compatible interaction’, *Journal of plant physiology*, 165(1), pp. 104–113.

Campbell, M. S. and Yandell, M. (2015) ‘An Introduction to Genome Annotation’, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 52, pp. 4.1.1–4.1.17.

Capella-Gutiérrez, S., Silla-Martínez, J. M. and Gabaldón, T. (2009) ‘trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses’, *Bioinformatics* , 25(15), pp. 1972–1973.

Castagnone-Sereno, P. *et al.* (2013) ‘Diversity and evolution of root-knot nematodes, genus Meloidogyne: new insights from the genomic era’, *Annual review of phytopathology*, 51, pp. 203–220.

Castagnone-Sereno, P. and Danchin, E. G. J. (2014) ‘Parasitic success without sex – the nematode experience’, *Journal of evolutionary biology*, 27(7), pp. 1323–1333.

Castro, C. J. and Ng, T. F. F. (2017) ‘U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs’, *Journal of computational biology: a journal of computational molecular cell biology*, 24(11), pp. 1071–1080.

Chakraborty, M. *et al.* (2016) ‘Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage’, *Nucleic acids research*. academic.oup.com, 44(19), p. e147.

Chen, C. *et al.* (2014) ‘Software for pre-processing Illumina next-generation sequencing short read sequences’, *Source code for biology and medicine*, 9, p. 8.

Chen, S. *et al.* (2018) ‘fastp: an ultra-fast all-in-one FASTQ preprocessor’, *Bioinformatics* . Oxford University Press, 34(17), pp. i884–i890.

Chikhi, R. and Medvedev, P. (2014) ‘Informed and automated k-mer size selection for genome assembly’, *Bioinformatics* . academic.oup.com, 30(1), pp. 31–37.

Cohen-Boulakia, S. *et al.* (2017) ‘Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities’, *Future generations computer systems: FGCS*, 75, pp. 284–298.

Collange, B. *et al.* (2011) ‘Root-knot nematode (Meloidogyne) management in vegetable crop production: The challenge of an agronomic system analysis’, *Crop protection (Guildford, Surrey)*. Elsevier BV, 30(10), pp. 1251–1262.

Crotty, S. M. *et al.* (2020) ‘GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments’, *Systematic biology*. academic.oup.com, 69(2), pp. 249–264.

Dabbish, L. *et al.* (2012) ‘Social coding in GitHub: transparency and collaboration in an open software repository’, in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. New York, NY, USA: Association for Computing Machinery (CSCW ’12), pp. 1277–1286.

Dainat, J. *et al.* (2020) *NBISweden/GAAS: GAAS-v1.2.0*. doi: 10.5281/zenodo.3835504.

Dainat, J. and Hereñú, D. (2020) *NBISweden/AGAT: AGAT-v0.4.0*. doi: 10.5281/zenodo.3877441.

Danchin GJ, K. G. (2020) ‘Meloidogyne enterolobii, whole genome shotgun sequencing project’, *NCBI Genbank*. doi: CAJEWN000000000.

Delsuc, F., Brinkmann, H. and Philippe, H. (2005) ‘Phylogenomics and the reconstruction of the tree of life’, *Nature reviews. Genetics*. Nature Publishing Group, 6(5), pp. 361–375.

De Maio, N. *et al.* (2019) ‘Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes’, *Microbial genomics*, 5(9). doi: 10.1099/mgen.0.000294.

den Akker, S. E. and Jones, J. T. (2018) ‘Sex: Not all that it’s cracked up to be?’, *PLoS genetics*. Public Library of Science, 14(2), p. e1007160

Desai, A. *et al.* (2013) ‘Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data’, *PloS one*. journals.plos.org, 8(4), p. E60204.

Dominguez Del Angel, V. *et al.* (2018) ‘Ten steps to get started in Genome Assembly and Annotation’, *F1000Research*, 7. doi: 10.12688/f1000research.13598.1.

Du, H. and Liang, C. (2019) ‘Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads’, *Nature communications*, 10(1), p. 5360.

Eaton, D. A. R. (2020) ‘Toytree: A minimalist tree visualization and manipulation library for Python’, *Methods in ecology and evolution / British Ecological Society*. Edited by M. Matschiner, 11(1), pp. 187–191.

Eisenback, J. D. (1985) ‘Detailed morphology and anatomy of second-stage juveniles, males, and females of the genus Meloidogyne (root-knot nematodes)’, *An advanced treatise on Meloidogyne*. North Carolina State University Graphics Raleigh, NC, USA, 1, pp. 47–77.

Eisenback, J. D. *et al.* (2003) ‘Meloidogyne haplanaria n. sp. (Nematoda: Meloidogynidae), a Root-knot Nematode Parasitizing Peanut in Texas’, *Journal of nematology*, 35(4), pp. 395–403.

Eisenback, J. D. and Triantaphyllou, H. H. (1991) ‘Root-knot nematodes: Meloidogyne species and races’, *Manual of agricultural nematology*. Marcel Dekker New York, 1, pp. 191–274.

Elling, A. A. (2013) ‘Major emerging problems with minor meloidogyne species’, *Phytopathology*, 103(11), pp. 1092–1102.

Emms, D. M. and Kelly, S. (2018) ‘OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences’. doi: 10.1101/466201.

Erin D. Foster, A. D. (2017) ‘Open Science Framework (OSF)’, *Journal of the Medical Library Association: JMLA*. Medical Library Association, 105(2), p. 203.

García, L. E. and Sánchez-Puerta, M. V. (2015) ‘Comparative and evolutionary analyses of Meloidogyne spp. Based on mitochondrial genome sequences’, *PloS one*, 10(3), p. e0121142.

Green, E. D., Rubin, E. M. and Olson, M. V. (2017) ‘The future of DNA sequencing’, *Nature*. nature.com, 550(7675), pp. 179–181.

Gurevich, A. *et al.* (2013) ‘QUAST: quality assessment tool for genome assemblies’, *Bioinformatics* , 29(8), pp. 1072–1075.

Handoo, Z. A. *et al.* (2004) ‘Morphological, Molecular, and Differential-Host Characterization of Meloidogyne floridensis n. sp. (Nematoda: Meloidogynidae), a Root-Knot Nematode Parasitizing Peach in Florida’, *Journal of nematology*, 36(1), pp. 20–35.

Hardison, R. C. (2003) ‘Comparative genomics’, *PLoS biology*, 1(2), p. E58.

Harrison, P. M. (2014) ‘Computational Methods for Pseudogene Annotation Based on Sequence Homology’, in Poliseno, L. (ed.) *Pseudogenes: Functions and Protocols*. New York, NY: Springer New York, pp. 27–39.

Haubold, B. and Wiehe, T. (2004) ‘Comparative genomics: methods and applications’, *Die Naturwissenschaften*, 91(9), pp. 405–421.

Hoff, K. J. and Stanke, M. (2019) ‘Predicting Genes in Single Genomes with AUGUSTUS’, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis... [et al.]*, 65(1), p. e57.

Holt, C. and Yandell, M. (2011) ‘MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects’, *BMC bioinformatics*, 12, p. 491.

Holterman, M. *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’, *Phytopathology*, 99(3), pp. 227–235.

Horner, D. S. *et al.* (2010) 'Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing', *Briefings in bioinformatics*, 11(2), pp. 181–197.

Hunter, J. D. (2007) 'Matplotlib: A 2D Graphics Environment', *Computing in science & engineering*. IEEE Computer Society, 9(3), pp. 90–95.

International Helminth Genomes Consortium (2019) 'Comparative genomics of the major parasitic worms', *Nature genetics*, 51(1), pp. 163–174.

Janssen, T. *et al.* (2017) 'Integrative taxonomy of root-knot nematodes reveals multiple independent origins of mitotic parthenogenesis', *PloS one*, 12(3), p. e0172190.

Jayakumar, V. and Sakakibara, Y. (2019) 'Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data', *Briefings in bioinformatics*. academic.oup.com, 20(3), pp. 866–876.

Jones, J. T., Haegeman, A., Danchin, E. G. J., Gaur, H. S., Helder, J., Jones, M. G. K., Kikuchi, T., Manzanilla-López, R., Palomares-Rius, J. E., Wesemael, W. M. L. and Perry, R. N. (2013) 'Top 10 plant-parasitic nematodes in molecular plant pathology', *Molecular plant pathology*, 14(9), pp. 946–961.

Jongman, M., Carmichael, P. C. and Bill, M. (2020) 'Technological Advances in Phytopathogen Detection and Metagenome Profiling Techniques', *Current microbiology*. Springer. doi: 10.1007/s00284-020-01881-z.

Joseph, S. *et al.* (2016) 'First Report of Meloidogyne haplanaria Infecting Mi-Resistant Tomato Plants in Florida and Its Molecular Diagnosis Based on Mitochondrial Haplotype', *Plant disease*, 100(7), pp. 1438–1445.

Kajitani, R. *et al.* (2014) 'Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads', *Genome research*, 24(8), pp. 1384–1395.

Kajitani, R. *et al.* (2019) 'Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions', *Nature communications*, 10(1), p. 1702.

Kang, X. *et al.* (2013) 'De Bruijn Graph-Based Whole-Genomic Sequence Assembly Algorithms and Applications', in *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pp. 2094–2097.

Katoh, K. *et al.* (2002) 'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic acids research*, 30(14), pp. 3059–3066.

Khan, A. R. *et al.* (2018) 'A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective', *Evolutionary bioinformatics online*, 14, p. 1176934318758650.

Kitts, P. (2002) 'Genome assembly and annotation process', *McEntyre J, Ostell Jeditors. The NCBI Handbook. Bethesda: National Center for Biotechnology Information*. Citeseer. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.220.1987&rep=rep1&type=pdf.

Kokot, M., Dlugosz, M. and Deorowicz, S. (2017) 'KMC 3: counting and manipulating k-mer statistics', *Bioinformatics* , 33(17), pp. 2759–2761.

Kolaczkowski, B. and Thornton, J. W. (2004) 'Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous', *Nature*. nature.com, 431(7011), pp. 980–984.

Koonin, E. V. and Galperin, M. Y. (2003a) *Genome Annotation and Analysis*. Kluwer Academic.

Koonin, E. V. and Galperin, M. Y. (2003b) *Principles and Methods of Sequence Analysis*. Kluwer Academic.

Korf, I. (2004) 'Gene finding in novel genomes', *BMC bioinformatics*, 5, p. 59.

Koski, L. B. and Golding, G. B. (2001) 'The closest BLAST hit is often not the nearest neighbor', *Journal of molecular evolution*, 52(6), pp. 540–542.

Köster, J. and Rahmann, S. (2012a) 'Building and documenting workflows with python-based snakemake', in *German Conference on Bioinformatics 2012*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. Available at: https://drops.dagstuhl.de/opus/volltexte/2012/3717/.

Köster, J. and Rahmann, S. (2012b) 'Snakemake—a scalable bioinformatics workflow engine', *Bioinformatics* . Narnia, 28(19), pp. 2520–2522.

Koutsovoulos, G. D. *et al.* (2018) 'Multiple independent adaptations to different ranges of host plants indicate high adaptability despite clonal reproduction in the nematode pest Meloidogyne incognita', *bioRxiv*. doi: 10.1101/362129.

Koutsovoulos, G. D., Marques, E., *et al.* (2019) 'Population genomics supports clonal reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest', *Evolutionary applications*. Wiley Online Library, 26, p. 909.

Koutsovoulos, G. D., Poullet, M., *et al.* (2019) 'The polyploid genome of the mitotic parthenogenetic root-knot nematode Meloidogyne enterolobii', *bioRxiv*. doi: 10.1101/586818.

Kozlowski, D. K. L. *et al.* (2020) 'Transposable Elements activity and role in Meloidogyne incognita genome dynamics and adaptability'. doi: 10.1101/2020.04.30.069948.

Kumar, S. *et al.* (2012) 'Statistics and truth in phylogenomics', *Molecular biology and evolution*, 29(2), pp. 457–472.

Laetsch, D. R. and Blaxter, M. L. (2017) 'BlobTools: Interrogation of genome assemblies', *F1000Research*. F1000 Research Limited, 6(1287), p. 1287.

Lawlor, B. and Sleator, R. D. (2020) 'The democratization of bioinformatics: A software engineering perspective', *GigaScience*, 9(6). doi: 10.1093/gigascience/giaa063.

Le, T. M. L. *et al.* (2019) 'A New Root-knot Nematode, Meloidogyne Moensi n. Sp. (Nematoda: Meloidogynidae), Parasitizing Robusta Coffee from Western Highlands, Vietnam', *Helminthologia*, 56(3), pp. 229–246.

Li, H. *et al.* (2009) 'Subgroup 1000 Genome Project Data Processing. 2009. The sequence alignment/map format and SAMtools', *Bioinformatics* , 25(16), pp. 2078–2079.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics* . Narnia, 25(14), pp. 1754–1760.

Li, Z. *et al.* (2012) 'Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph', *Briefings in functional genomics*, 11(1), pp. 25–37.

Lu, H., Giordano, F. and Ning, Z. (2016) 'Oxford Nanopore MinION Sequencing and Genome Assembly', *Genomics, proteomics & bioinformatics*. Elsevier, 14(5), pp. 265–279.

Lunt, D. H. (2008) 'Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins', *BMC evolutionary biology*, 8, p. 194.

Lunt, D. H. *et al.* (2014) 'The complex hybrid origins of the root knot nematodes revealed through comparative genomics', *PeerJ*. PeerJ Inc., 2, p. e356.

Mardis, E. R. (2011) 'A decade's perspective on DNA sequencing technology', *Nature*. nature.com, 470(7333), pp. 198–203.

Mardis, E. R. (2017) 'DNA sequencing technologies: 2006–2016', *Nature protocols*. nature.com, 12(2), pp. 213–218.

McNutt, M. (2014) 'Reproducibility', *Science*, 343(6168), p. 229.

Menezes, A. M. F. *et al.* (2019) 'In Silico Characterization of Meloidogyne Genus Nematode Cellulose Binding Proteins', *Brazilian archives of biology and technology* = . Tecpar, 62. doi: 10.1590/1678-4324-2019180120.

Miller, J. R. *et al.* (2017) 'Hybrid assembly with long and short reads improves discovery of gene family expansions', *BMC genomics*, 18(1), p. 541.

Minh, B. Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular biology and evolution*, 37(5), pp. 1530–1534.

Minh, B. Q., Hahn, M. W. and Lanfear, R. (2020) 'New methods to calculate concordance factors for phylogenomic datasets', *Molecular biology and evolution*. doi: 10.1093/molbev/msaa106.

Mirarab, S. *et al.* (2014) 'ASTRAL: genome-scale coalescent-based species tree estimation', *Bioinformatics* , 30(17), pp. i541–8.

Misof, B. *et al.* (2014) 'Phylogenomics resolves the timing and pattern of insect evolution', *Science*. science.sciencemag.org, 346(6210), pp. 763–767.

Mwageni, W. *et al.* (2000) 'The importance of tropical root-knot nematodes (Meloidogyne spp.) and factors affecting the utility of Pasteuria penetrans as a biocontrol agent', *Nematology: international journal of fundamental and applied nematological research*. Brill, 2(8), pp. 823–845.

Nelson, S 2007, *Root knot nematodes infecting carrot*, photograph, viewed 20 August 2020, <https://www.flickr.com/photos/scotnelson/9807795696/>

Nelson, S 2015, *Plant parasitic nematode with stylet*, photograph, viewed 20 August 2020, <https://www.flickr.com/photos/scotnelson/27834122886/>

Nelson, S 2017, *Tomato (Solanum lycopersicum): Root-knot nematodes*, photograph, viewed 20 August 2020, <https://www.flickr.com/photos/scotnelson/38706032911>

Nelson, S 2018, *Egg masses of Meloidogyne incognita,* viewed 20 August 2020, <https://www.flickr.com/photos/scotnelson/25008358137/>

Nelson, S 2018, *Meloidogyne incognita (root-knot nematode): Adult male*, photograph, viewed 20 August 2020, <https://www.flickr.com/photos/scotnelson/28114624449/>

Nicol, J. M. *et al.* (2011) 'Current Nematode Threats to World Agriculture', in Jones, J., Gheysen, G., and Fenoll, C. (eds) *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Dordrecht: Springer Netherlands, pp. 21–43.

Oliveira, C. M. G. de, Monteiro, A. R. and Blok, V. C. (2011) 'Morphological and molecular diagnostics for plant-parasitic nematodes: working together to get the identification done', *Tropical plant pathology*. Sociedade Brasileira de Fitopatologia, 36(2), pp. 65–73.

One Thousand Plant Transcriptomes Initiative (2019) ‘One thousand plant transcriptomes and the phylogenomics of green plants’, *Nature*. nature.com, 574(7780), pp. 679–685.

Peng, R. D. (2011) ‘Reproducible research in computational science’, *Science*, 334(6060), pp. 1226–1227.

Perkel, J. M. (2017) ‘Democratizing bioinformatics’, *Nature*. Macmillan Publishers Ltd., London, England, 543(7643), pp. 137–138.

Perry, R. N., Moens, M. and Starr, J. L. (2009) *Root-knot Nematodes*. CABI.

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) ‘An Eulerian path approach to DNA fragment assembly’, *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp. 9748–9753.

Pflüger, M. (2019) ‘Reproducible Data Analysis with conda’, in *Open Science Days 2019*. pure.mpg.de. Available at: https://pure.mpg.de/rest/items/item_3027294/component/file_3027295/content.

Pop, M. *et al.* (2004) ‘Comparative genome assembly’, *Briefings in bioinformatics*. academic.oup.com, 5(3), pp. 237–248.

Pop, M. (2009) ‘Genome assembly reborn: recent computational challenges’, *Briefings in bioinformatics*. academic.oup.com, 10(4), pp. 354–366.

Powers, T. O. *et al.* (2005) ‘Incorporating Molecular Identification of Meloidogyne spp. into a Large-scale Regional Nematode Survey’, *Journal of nematology*, 37(2), pp. 226–235.

Prasanna, A. N. *et al.* (2020) ‘Model Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep Basidiomycota Relationships’, *Systematic biology*. academic.oup.com, 69(1), pp. 17–37.

Pryszcz, L. P. and Gabaldón, T. (2016) ‘Redundans: an assembly pipeline for highly heterozygous genomes’, *Nucleic acids research*, 44(12), p. e113.

Ralmi, N. H. A. A. *et al.* (2016) ‘Occurrence and control of root knot nematode in crops: a review’, *Australian journal of crop science*. Southern Cross Publishers, 11(12), p. 1649.

Ram, K. (2013) ‘Git can facilitate greater reproducibility and increased transparency in science’, *Source code for biology and medicine*. scfbm.biomedcentral.com, 8(1), p. 7.

Ranallo-Benavidez, T. R., Jaron, K. S. and Schatz, M. C. (2020) ‘GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes’, *Nature communications*, 11(1), p. 1432.

Rashidifard, M. *et al.* (2019) ‘Molecular characterisation of Meloidogyne enterolobii and other Meloidogyne spp. from South Africa’, *Tropical plant pathology*. Springer International Publishing, 44(3), pp. 213–224.

Reddy, S. *et al.* (2017) ‘Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling’, *Systematic biology*, 66(5), pp. 857–879.

Richards, S. (2018) ‘Full disclosure: Genome assembly is still hard’, *PLoS biology*, p. e2005894.

Rodríguez-Ezpeleta, N. and Philippe, H. (2009) ‘Phylogenomics’, in Schaechter, M. (ed.) *Encyclopedia of Microbiology (Third Edition)*. Oxford: Academic Press, pp. 633–643.

Rutter, W. B. *et al.* (2014) ‘Members of the Meloidogyne avirulence protein family contain multiple plant ligand-like motifs’, *Phytopathology*, 104(8), pp. 879–885.

Salzberg, S. L. (2019) ‘Next-generation genome annotation: we still struggle to get it right’, *Genome biology*, 20(1), p. 92.

Santos, D. *et al.* (2019) ‘New Hosts and Records in Portugal for the Root-Knot Nematode Meloidogyne luci’, *Journal of nematology*, 51, pp. 1–4.

Sasser, J. N. and Freckman, D. W. (1987) ‘A world perspective on nematology: the role of the society’. Society of Nematologists. Available at: http://agris.fao.org/agris-search/search.do?recordID=US8903406.

Scheibye-Alsing, K. *et al.* (2009) ‘Sequence assembly’, *Computational biology and chemistry*. Elsevier, 33(2), pp. 121–136.

Simão, F. A. *et al.* (2015) ‘BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs’, *Bioinformatics* , 31(19), pp. 3210–3212.

Simpson, C. E. *et al.* (2003) ‘Registration of “NemaTAM”peanut’, *Crop*. dl.sciencesocieties.org. Available at: https://dl.sciencesocieties.org/publications/cs/abstracts/43/4/1561.

Siu-Ting, K. *et al.* (2019) ‘Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics’, *Molecular biology and evolution*. academic.oup.com, 36(6), pp. 1344–1356.

Smit, A. F. A., Hubley, R. and Green, P. (2015a) ‘RepeatMasker Open-4.0. 2013--2015’.

Smit, A. F. A., Hubley, R. and Green, P. (2015b) ‘RepeatModeler Open-1.0. 2008--2015’, *Seattle, USA: Institute for Systems Biology. Available from: httpwww. repeatmasker. org, Last Accessed May*, 1, p. 2018.

Smith, S. A. *et al.* (2015) ‘Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants’, *BMC evolutionary biology*, 15, p. 150.

Stamatakis, A. (2014) ‘RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies’, *Bioinformatics* , 30(9), pp. 1312–1313.

Sumner, J. G. *et al.* (2012) ‘Is the general time-reversible model bad for molecular phylogenetics?’, *Systematic biology*, 61(6), pp. 1069–1074.

Sumner, J. G., Fernández-Sánchez, J. and Jarvis, P. D. (2012) ‘Lie Markov models’, *Journal of theoretical biology*. Elsevier, 298, pp. 16–31.

Susič, N. *et al.* (2020) ‘Genome sequence of the root-knot nematode Meloidogyne luci’, *Journal of nematology*, 52, pp. 1–5.

Szitenberg, A. *et al.* (2017) ‘Comparative Genomics of Apomictic Root-Knot Nematodes: Hybridization, Ploidy, and Dynamic Genome Change’, *Genome biology and evolution*, 9(10), pp. 2844–2861.

Tarailo-Graovac, M. and Chen, N. (2009) ‘Using RepeatMasker to identify repetitive elements in genomic sequences’, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 4, p. Unit 4.10.

Triantaphyllou, H. H. (1991) ‘Meloidogyne Species and Races’, *Manual of Agricultural Nematology*. CRC Press, p. 191.

Trudgill, D. L. and Blok, V. C. (2001) ‘Apomictic, polyphagous root-knot nematodes: exceptionally successful and damaging biotrophic root pathogens’, *Annual review of phytopathology*, 39, pp. 53–77.

Virtanen, P. *et al.* (2020) ‘SciPy 1.0: fundamental algorithms for scientific computing in Python’, *Nature methods*, 17(3), pp. 261–272.

Wesemael, W., Viaene, N. and Moens, M. (2011) ‘Root-knot nematodes (Meloidogyne spp.) in Europe’, *Nematology: international journal of fundamental and applied nematological research*. Brill, 13(1), pp. 3–16.

Wilberg, E. W. (2015) ‘What’s in an Outgroup? The Impact of Outgroup Choice on the Phylogenetic Position of Thalattosuchia (Crocodylomorpha) and the Origin of Crocodyliformes’, *Systematic biology*, 64(4), pp. 621–637.

Wilkinson, M. D. *et al.* (2016) ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific data*, 3, p. 160018.

Williams, T. A. *et al.* (2019) ‘Phylogenomics provides robust support for a two-domains tree of life’, *Nature Ecology & Evolution*. Nature Publishing Group, 4(1), pp. 138–147.

Wood, D. E., Lu, J. and Langmead, B. (2019) ‘Improved metagenomic analysis with Kraken 2’, *Genome biology*, 20(1), p. 257.

Yang, B. and Eisenback, J. D. (1983) ‘Meloidogyne enterolobii n. sp. (Meloidogynidae), a Root-knot Nematode Parasitizing Pacara Earpod Tree in China’, *Journal of nematology*. ncbi.nlm.nih.gov, 15(3), pp. 381–391.

Ye, W., Robbins, R. T. and Kirkpatrick, T. (2019) ‘Molecular characterization of root-knot nematodes (Meloidogyne spp.) from Arkansas, USA’, *Scientific reports*. nature.com, 9(1), p. 15680.

Young, A. D. and Gillung, J. P. (2019) ‘Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics’, *Systematic entomology*, 67, p. 215.

Zerbino, D. R. and Birney, E. (2008) ‘Velvet: algorithms for de novo short read assembly using de Bruijn graphs’, *Genome research*. genome.cshlp.org, 18(5), pp. 821–829.

6.0 Appendix

Supplementary table 1: Full output report of *QUAST* assembly appraisal software.

Statistics generated using *QUAST (Gurevich et al., 2013)*.

Assembly	*spades*	*Platanus assemble*	*Platanus scaffold*	*Platanus gap closed*	*Platanus scaffold + redundans*	*Platanus gap closed + redundans*	*spades + redundans*
Reference	*spades*	*PA*	*PS*	*PGC*	*PSR*	*PGCR*	*spadesR*
Complete BUSCO (%)	90.43	28.71	65.02	70.63	64.36	76.9	93.73
Partial BUSCO (%)	7.92	27.39	19.47	15.18	21.12	11.55	4.29
# N's	456333	0	4518833	1653102	31819	87055	88047
# contigs (>= 10000 bp)	2617	165	2097	2086	1388	2033	3529
# contigs (>= 25000 bp)	174	9	601	600	107	740	668
# contigs (>= 50000 bp)	14	0	120	119	7	214	113
Total length (>= 0 bp)	2.17E+08	1.4E+08	1.13E+08	1.13E+08	65884696	69377218	1.63E+08
Total length (>= 1000 bp)	1.74E+08	42588431	69901102	69730810	62327815	68365113	1.51E+08
Total length (>= 5000 bp)	88695240	7958324	59360953	59270186	41103491	63017623	1.08E+08
Total length (>= 10000 bp)	39748115	2360142	48977488	48792154	21763537	55423312	68469010
Total length (>= 25000 bp)	5859401	272088	25178421	25124162	3604489	34463090	26346651
Total length (>= 50000 bp)	921652	0	8684419	8626972	467427	16286264	7824102
# contigs	227621	492456	292691	292691	20219	7130	60728
Largest contig	130740	42110	227182	226635	95239	279646	154700
Total length	2.17E+08	1.4E+08	1.13E+08	1.13E+08	65884696	69377218	1.63E+08
GC (%)	32.26	30.81	30.88	30.91	29.87	30.13	31.97
N50	3752	311	6384	6398	6712	24820	8052
N75	1318	170	147	147	3455	12485	3722
L50	14244	72011	3055	3052	2759	750	4986
L75	38652	241191	71763	72132	6165	1730	12450
# N's per 100 kbp	210.56	0	3985.23	1460.7	48.29	125.48	54.01
Complete BUSCO (%)	90.43	28.71	65.02	70.63	64.36	76.9	93.73
Partial BUSCO (%)	7.92	27.39	19.47	15.18	21.12	11.55	4.29

Supplementary table 2: Counts of genomic features predicted throughout the annotation process. Counts parsed using *GAAS* (Dainat *et al.*, 2020).

Run number	1	2	3	4	5
MAKER2 ran with:	EST2GENOME	First iteration of SNAP	Second iteration of SNAP	Third iteration of SNAP	First iteration of Augustus
Number of genes	13645	22697	14978	1861	20213
Number of mrnas	13645	22697	14978	1861	20213
Number of mrnas with utr both sides	-	70	66	7	161
Number of mrnas with at least one utr	-	288	2173	9	1232
Number of cdss	13645	22697	14978	1861	20213
Number of exons	89038	45127	109091	1861	149065
Number of five_prime_utrs	-	288	2173	9	1190
Number of three_prime_utrs	-	70	66	7	203
Number of exons in cds	89038	44953	107371	1861	147500
Number of exons in five_prime_utr	-	455	3801	9	2497
Number of exons in three_prime_utr	-	70	93	7	409
Number of introns in cds	75393	22256	92393	-	127287
Number of introns in exon	75393	22430	94113	-	128852
Number of introns in five_prime_utr	-	167	1628	-	1307
Number of introns in three_prime_utr	-	-	27	-	206
Number of single exon gene	2100	8071	1373	1861	2111

Number of single exon mrna	2100	8071	1373	1861	2111
mean mrnas per gene	1	1	1	1	1
mean cdss per mrna	1	1	1	1	1
mean exons per mrna	6.5	2	7.3	1	7.4
mean five_prime_utrs per mrna	-	0	0.1	0	0.1
mean three_prime_utrs per mrna	-	0	0	0	0
mean exons per cds	6.5	2	7.2	1	7.3
mean exons per five_prime_utr	-	1.6	1.7	1	2.1
mean exons per three_prime_utr	-	1	1.4	1	2
mean introns in cdss per mrna	5.5	1	6.2	-	6.3
mean introns in exons per mrna	5.5	1	6.3	-	6.4
mean introns in five_prime_utrs per mrna	-	0	0.1	-	0.1
mean introns in three_prime_utrs per mrna	-	-	0	-	0
Total gene length	29821175	16056169	43559209	562458	60248757
Total mrna length	29821172	16056169	43559209	562458	60248757
Total cds length	13914219	7086711	13117464	561129	17641011
Total exon length	13914219	7106475	13202198	562458	17754530
Total five_prime_utr length	-	11599	72817	365	65132
Total three_prime_utr length	-	8165	11917	964	48387
Total intron length per cds	15982346	8877958	29601189	-	41848159

Total intron length per exon	15982346	8972124	30451124	-	42623079
Total intron length per five_prime_utr	-	91041	801224	-	678711
Total intron length per three_prime_utr	-	-	9454	-	63240
Mean gene length	2185	707	2908	302	2980
Mean mrna length	2185	707	2908	302	2980
Mean cds length	1019	312	875	301	872
Mean exon length	156	157	121	302	119
Mean five_prime_utr length	-	40	33	40	54
Mean three_prime_utr length	-	116	180	137	238
Mean cds piece length	156	157	122	301	119
Mean five_prime_utr piece length	-	25	19	40	26
Mean three_prime_utr piece length	-	116	128	137	118
Mean intron in cds length	211	398	320	-	328
Mean intron in exon length	211	400	323	-	330
Mean intron in five_prime_utr length	-	545	492	-	519
Mean intron in three_prime_utr length	-	-	350	-	306
Longest genes	26654	17547	43446	4758	34445
Longest mrnas	26654	17547	43446	4758	34445
Longest cdss	15495	9366	14823	4758	18729
Longest exons	4805	9366	8955	4758	8955

Longest five_prime_utrs	-	92	92	72	2799
Longest three_prime_utrs	-	426	825	191	2311
Longest cds piece	4805	9366	8955	4758	8955
Longest five_prime_utr piece	-	92	92	72	436
Longest three_prime_utr piece	-	426	667	191	723
Longest intron into cds part	9829	12314	9148	-	9148
Longest intron into exon part	9829	12314	9148	-	9148
Longest intron into five_prime_utr part	-	2761	4134	-	5427
Longest intron into three_prime_utr part	-	-	2451	-	3049
Shortest genes	9	177	178	180	75
Shortest mrnas	9	177	178	180	75
Shortest cdss	9	12	18	60	6
Shortest exons	1	4	2	180	2
Shortest five_prime_utrs	-	1	1	11	1
Shortest three_prime_utrs	-	6	10	48	10
Shortest cds piece	1	1	1	60	1
Shortest five_prime_utr piece	-	1	1	11	1
Shortest three_prime_utr piece	-	6	4	48	1
Shortest intron into cds part	31	5	5	-	5
Shortest intron into exon part	31	5	5	-	5

Shortest intron into five_prime_utr part	-	9	5	-	5
Shortest intron into three_prime_utr part	-	-	39	-	10

Supplementary table 3: Source information of genomic data.

Species	Description	Reference	Format	BioProject	Biosample or doi location	Isolate code
M. haplanaria	Two libraries of raw short-read sequence data	Joseph *et al.* (2016)	Fastq (nt)	PRJNA340324 *	-	SJH1
M. incognita	CDSs	Szitenberg *et al.* (2017)	Fasta (aa)	PRJNA340324	10.5281/zenodo.399475	W1
	Genome assembly		Fasta (nt)		SAMN05712521	
M. javanica	CDSs	Szitenberg *et al.* (2017)	Fasta (aa)		10.5281/zenodo.399475	VW4
	Genome assembly		Fasta (nt)		SAMN05712519	
M. arenaria	CDSs	Szitenberg *et al.* (2017)	Fasta (aa)		10.5281/zenodo.399475	HarA
	Genome assembly		Fasta (nt)		SAMN05712513	
M. floridensis	CDSs	Szitenberg *et al.* (2017)	Fasta (aa)		10.5281/zenodo.399475	SJF1
	Genome assembly		Fasta (nt)		SAMN05712529	
M. enterolobii	CDSs	Szitenberg *et al.* (2017)	Fasta (aa)		10.5281/zenodo.399475	L30
	Genome assembly	INRAE	Fasta (nt)	PRJEB36431	SAMEA6504944	-
M. hapla	CDSs	Opperman *et al.* (2008)	Fasta (aa)	PRJNA29083	SAMN02743742	VW9
M. luci	Genome assembly	SC	Fasta (nt)	PRJEB27977	SAMEA5770813	V13