2025-01-08

Exploring Technical and Biological Variation in Nematode Genomes

Mariene, Grace Mukiri

UNIVERSITY OF CALGARY

Exploring Technical and Biological Variation in Nematode Genomes

by

Grace Mukiri Mariene

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN VETERINARY MEDICAL SCIENCES

CALGARY, ALBERTA

JANUARY, 2025

Abstract

Parasitic species of nematodes are the causative agents of major infectious diseases of humans, livestock and crops. Mass drug administration programs have been the primary control strategy for livestock and, increasingly, human parasites, significantly contributing to the global issue of resistance to these anthelmintic drugs. The need for new anthelmintics demands a deeper understanding of nematode biology, which, in turn, would greatly benefit from high-quality assembled and annotated genomes. However, challenges in the current bioinformatics protocols that guide accurate genome assembly and annotations introduce errors and variability, which potentially compromise the preservation of genetic material and accuracy, particularly critical in drug discovery. These variations are often overlooked and unaccounted for by many researchers.

This thesis, in addition to finding the most effective protocols to generate a high-quality genome assembly and annotation for the Canadian isolate of *Heligmosomoides bakeri*, also describes an in-depth investigation of the sources and impact of genome assembly variations to be considered in genomics research. The performance of several commonly used assembly programs was compared using long-read DNA sequencing data from three nematodes: *Caenorhabditis bovis*, *Haemonchus contortus,* and *H. bakeri*. The results of these comparisons showed that each program yielded different results which could have critical impacts in gene finding. This work is described in Chapter 2. The sources and potential impacts of variations in differently scaffolded *H. bakeri* genomes was also investigated. The results, which are described in Chapter 3, indicated that most variations stemmed from the starting preliminary contig assemblies. In Chapter 4, the potential utility of *H. bakeri* as a model organism for Clade V parasitic nematodes is described using a genomics approach of orthology and phylogenomic comparisons of gene families involved in detoxification of xenobiotics.

Together, this work describes cost-effective quality control, assembly and annotation protocols that can be extended to create genomics resources for livestock parasites, such as *Ostertagia ostertagi*, *Cooperia punctata* and *Cooperia oncophora*, which are the most prevalent parasitic nematodes of Albertan cattle with rapidly emerging anthelmintic resistance, as well as other highly heterozygous invertebrate organisms. Additionally, the work creates awareness and cautions researchers, particularly those interested in comparative genomics and genic analysis, of the critical impacts of “small” variations in the starting draft genomes and the need for informed and more careful assembly generation approaches and selection.

Preface

Chapter 1 is original, unpublished work by GMM. Figure 1.1 is from (Goussarov *et al*., 2022); an open access publication distributed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). Text boxes with numbered and larger fonts for the “Overlap”, “Layout” and “Consensus” were added to the figure to enhance readability. Figure 1.2 was downloaded from https://training.galaxyproject.org/training-material/topics/assembly/images/ex1-4.png; CC-BY-4.0 (Hiltemann *et al*., 2023; Batut *et al*., 2018). The text “Example #1” was cropped out and text boxes labeled “Reads”, “Divided into k-mers” and “Reconstructed sequence” were added for clarity. Circles highlighting k-mer overlaps were included for emphasis and clarity. Figure 1.3 was downloaded from https://training.galaxyproject.org/training-material/topics/assembly/images/ex2-4.png; CC-BY-4.0 (Hiltemann *et al*., 2023; Batut *et al*., 2018). The text “Example #2” was cropped out and text boxes labeled “Reads”, “Divided into k-mers” and “Reconstructed sequence” were added for clarity. Circles highlighting k-mer overlaps were included for emphasis and clarity. Figure 1.4 is used with permission from Benjamin Langmead and was downloaded from ads1-slides/0580_asm__practice.pdf at master · BenLangmead/ads1-slides · GitHub. Text boxes labeled “Allele a”, “Allele b”, “Haplotype” and ““Bubble” on a heterozygous site” were added for emphasis and clarity. The blue arrow was added to the figure to highlight “Bubble” on a heterozygous site. Figure 1.5 is from (Smythe *et al.*, 2019) which is distributed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). The red arrow was added to the figure to highlight *Heligmosomoides bakeri*.

The work described in Chapter 2 is available as a preprint at bioRxiv (doi: https://doi.org/10.1101/2024.02.26.582167) and has, at the time of writing, been peer-reviewed, revised and resubmitted to the International Journal of Parasitology. The work included in this thesis is the revised version of the manuscript. The figure names and section headings have been formatted to follow the thesis format.

Chapter 3 is original, unpublished work by GMM. Figure 3.1 is downloaded from https://www.genome.gov/genetics-glossary/Mapping; (Mapping, 2023). Text boxes with larger fonts for the “Genetic map” and “Physical map” were added to the figure for readability and clarity. Figure 3.2 is downloaded from http://bch709.plantgenomicslab.org/fig/mate.png; CC BY 4.0 license. The text “Contigs to scaffolds” was cropped out to reduce repetition with the main text. Figure 3.3 is by Prakrutiuday - Own work, CC BY-SA 4.0,

https://commons.wikimedia.org/w/index.php?curid=115662977 and was downloaded from https://upload.wikimedia.org/wikipedia/commons/thumb/a/a4/HiCschematic.png/1000px-HiCschematic.png

Chapter 4 is original, unpublished work by GMM. Figure 4.1 is from (Pollo *et al.*, 2023) and Figure 4.2 is from (Sharma *et al.*, 2024).

Chapter 5 is original, unpublished work by GMM.

Acknowledgments

It has been a long exhausting journey of faith, hope and resilience, but God! I am forever grateful to Almighty God for seeing me through it all.

I would like to thank my supervisor Dr. James Wasmuth for the opportunity to realize my dream in research and for his immeasurable support and advice as I navigated through all the challenges.

I am grateful to my committee members Dr. John Gilleard, Dr. Graham Plastow and Dr. Frank van der Meer for their support and valuable advice throughout my graduate experience.

Very special thanks to my dad, John, my mum, Charity, my sister, Leah, my nephew, Jayden and my grandma, Grace. All of their love, care, support, encouragement, prayers, and faith were instrumental in seeing my thesis to completion. I kept hope alive and did not give up, because of you.

Many thanks to my former Wasmuth lab mates who became friends and “partners in crime”. Thanks for all the escapades and decompressing moments under our “special tree”. A big thank you to Dr. Kyle—my “African” brother—for your kindness and wicked sense of humor, Dr. Stephen—my white board decompressing partner—for making the trip and conference in San Diego so much fun, and Dr. Chenhua—my emergency contact—for your empathy, care and kindness. Special thanks to Kefa and Lucy for believing in me; I have been riding on your prayers and faith.

Table of Contents

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 Introduction to nematodes

Nematodes, or roundworms, include species with vastly different ecological roles—including free-living and parasitic lifestyles—occupying niches across marine, freshwater, and terrestrial ecosystems and includes important pathogenic groups that infect humans, animals, and plants. The phylum Nematoda has been divided into five clades—Clades I through V—which is supported by datasets from individual marker genes through to genome-wide studies involving hundreds of orthologs (Blaxter *et al.*, 1998; Smythe *et al.*, 2019). Clade V is equivalent to the order Rhabditina and includes some of the most well studied animal species. The most famous Clade V species is the free-living *Caenorhabditis elegans*, a model organism that has advanced humanity's knowledge across medical, evolutionary, and ecological disciplines (Greener, 2021; Gray and Cutter, 2014; Leung *et al.*, 2008). Clade V is also home to parasitic species with significant relevance to animal (including human) health. Many of these species infect their host's gastrointestinal tract and include the hookworms, *Ancylostoma caninum* and *Necator americanus*, and the Trichostrongylidae, *Haemonchus contortus*, *Ostertagia ostertagi*, and *Cooperia oncophora*.

Parasitic nematodes are responsible for a substantial global health and economic burden. High parasitic loads can cause significant morbidity, including diarrhea, weight loss, and anemia, especially in the case of blood-feeding species. These infections often lead to considerable economic losses due to decreased productivity, increased healthcare costs, and mortality. Soil-transmitted helminths, as parasitic worms are also known, infect an estimated 1.5 billion people, approximately 24% of the world's population, primarily affecting low-income regions where sanitation is limited (World Health Organization, 2022). Additionally, parasitic nematodes—such as the Trichostrongylidae—cause severe economic impacts in livestock. For instance, annual losses attributed to cattle nematodes are estimated at US$1.41billion in Mexico, US$7.11 billion in Brazil, and US$2 billion in the United States (Grisi *et al.*, 2014; Rodríguez-Vivas *et al.*, 2017; Stromberg and Gasbarre, 2006). These losses encompass not only treatment costs but also reduced productivity in infected animals, which is particularly relevant to industries and communities dependent on livestock.

1.2 Anthelmintic resistance

The significant health and economic challenges posed by parasitic nematodes have motivated research into nematode biology and the development of effective treatment strategies. Control of parasitic infections has traditionally relied on anthelmintic drugs administered through mass

drug administration (MDA) programs. New classes of anthelmintic drugs were launched in the market every decade from the 1950s to the 1980s; phenothiazine (1950s), followed by the benzimidazoles (1960s), the imidazothiazole-tetrahydropyrimidines (1970s), and the avermectins-milbemycins (1980s) (reviewed in (Kaplan, 2004). However, widespread and often indiscriminate use of these drugs has led to the emergence and spread of anthelmintic resistance. Anthelmintic resistance is defined as the heritable ability of helminths to survive drug doses that would normally be effective in killing them (Sangster and Dobson, 2002). This resistance has emerged in multiple nematode species affecting both humans and animals and represents a major obstacle to effective control and is part of the major global challenge and concern of antimicrobial drug resistance (reviewed in (Kaplan, 2004; Wolstenholme and Kaplan, 2012)).

The mechanisms underlying anthelmintic resistance are complex and differ between drug classes and possibly between species. While genetic mutations have been linked to resistance in several species, these mechanisms vary, necessitating species-specific research. For example, resistance mechanisms for some anthelmintics involve alterations in drug targets or metabolic pathways, but differences in resistance patterns across species demonstrate the need to independently study each parasite's biological response to drugs (Hu *et al.*, 2013). Addressing anthelmintic resistance is critical not only for effective parasite control but also for the development of sustainable strategies that minimize resistance emergence.

1.3 Genomic studies of nematodes

The detrimental impacts of parasitic nematodes on humans, animals, and crops, coupled with rising concerns over anthelmintic resistance and a growing human population dependent on livestock and agriculture for food, highlight the urgent need for effective anthelmintic drugs and vaccines. One approach to achieving these goals, which provides a deeper understanding of the nematodes' genetics, is through whole-genome sequencing. This process generates genomic resources essential for comprehensive genome-wide research on these parasites.

Genomics is the study of an organism's complete DNA sequence and has become a mainstay of biomedical, evolutionary, and ecological research. *C. elegans* was the first animal, to have its genome fully sequenced, assembled and annotated (The *C. elegans* Sequencing Consortium, 1998). Inspired by this, and motivated by the global health importance, researchers in parasitic nematodes embraced genomics and related omics technologies: transcriptomics, proteomics, and metabolomics. As of this writing, WormBase Parasite, a comprehensive public genomic

resource for parasitic worms, hosts at least 177 nematode genome assemblies for at least 135 species (WormBase Parasite (WBPS18; WS285) (Howe *et al.*, 2017).

Research on parasitic nematode genomes has significantly advanced knowledge of host-parasite interactions, including how parasites establish infections, maintain chronic infections, and exploit host resources (Zarowiecki and Berriman, 2015; Wasmuth *et al.*, 2008). This knowledge is crucial for identifying potential drug targets and understanding the genetic factors involved in drug resistance (Gilleard, 2006, 2013; Saunders *et al.*, 2013; Cotton *et al.*, 2016; Coghlan *et al.*, 2023; Laing *et al.*, 2013; International Helminth Genomes Consortium, 2019). Moreover, complete genomes are needed to investigate genome modifications to understand function of genes (Doyle, 2022).

1.3.1 Complexity of nematode genomes

Nematode genomes exhibit considerable complexity. Their sizes vary from approximately 34 Mb in *Bunonema sp.* (RGD898) to 656 Mb in *Heligmosomodies polygyrus* (Casasa *et al.*, 2021; Stevens *et al.*, 2023). This variation reflects differences in genome organization, repeat content, and gene density. Like the genomes of other animals, the genomes of nematodes include extensive intergenic and intronic regions, overlapping genes, and numerous repetitive elements, such as short tandem repeats (STRs) and transposable elements (TEs) (The *C. elegans* Sequencing Consortium, 1998; Chen and Stein, 2006). In *C. elegans*, repetitive elements constitute approximately 18% of the genome, with TEs alone contributing 12% (The *C. elegans* Sequencing Consortium, 1998; Wilson, 1999).

Additionally, nematode genomes undergo alternative splicing resulting in different isoforms of a gene. Alternative splicing sites are enriched with transcribed STRs which cause merging of complementary intronic repeats, bringing exons into close proximity (Lian and Garner, 2005). Adding to the complexity, worms are diploid, with two sets of chromosomes that can have two different alleles at a gene locus, a phenomenon known as heterozygosity. Heterozygosity introduces genetic polymorphisms which increase genetic diversity through changes in allele sequences and result from natural selection, mutation rates, large effective population sizes and host migration (Blouin *et al.*, 1995; Brasil *et al.*, 2012; Troell *et al.*, 2006; Gilleard and Beech, 2007; Allendorf, 2014). Nematodes exhibit extensive genetic diversity, complicating the identification of causal variants underlying biological mechanisms such as anthelmintic resistance (Doyle and Cotton, 2019).

1.3.2 DNA sequencing technologies

Our ability to read the nucleotides in a molecule of DNA—termed sequencing—has evolved dramatically since the advent of Sanger sequencing in 1977 and is enabling large-scale genome studies. Sanger sequencing, with its high accuracy, remains valuable for small-scale projects and specific tasks, but it has largely been superseded by high-throughput sequencing methods capable of handling large, complex genomes. The second-generation sequencing approach uses sequencing-by-synthesis technology to generate massive amounts of short-read data with high accuracy (Straiton *et al.*, 2019). Of the many companies that offered sequencing-by-synthesis, Illumina™ is the only remaining supplier of note. Illumina reads are typically 150-300 bp in length, making them well-suited for short genomic regions, but their limited read length poses challenges for assembling complex genomes. The third-generation sequencing platforms, currently led by Pacific BioSciences™(PacBio) and Oxford Nanopore Technologies™ (ONT), produce longer reads, often exceeding 10,000 bp, and have transformed genomics by enabling more contiguous genome assemblies. However, these long reads generally come with lower accuracy than short reads, though recent improvements, such as PacBio's HiFi technology, have raised long-read accuracy to approximately 99%, enabling more reliable genome assemblies (Jain *et al.*, 2017, 2022; Liu-Wei *et al.*, 2024; Ardui *et al.*, 2018; Wick *et al.*, 2019; van Dijk *et al.*, 2018; Mitsuhashi and Matsumoto, 2020). These long-read sequencing technologies have begun to deliver on the promise of telomere-to-telomere assemblies (Miga *et al.*, 2020). Notwithstanding the enthusiasm for these developments within the research community and the general public (Devlin, 2022; CBC Radio, 2022), it remains important to acknowledge that each sequencing technology brings distinct advantages and limitations.

Despite advancements in sequencing technologies, specific challenges remain in studying parasitic nematode genomes. Long-read sequencing protocols require hundreds of nanograms of high-molecular-weight DNA as input, which exceeds the amount available from most single worms. This often demands pooling of multiple worms, further exacerbating the high genetic diversity present in nematode populations (Stevens *et al.*, 2023). Furthermore, to obtain sufficient input DNA, larger adult stages of worms are often targeted, which can only be acquired by sacrificing the host.

1.3.3 Genome assembly

Genome assembly is the computational (*in silico*) reconstruction of fragmented sequence reads from the cell-bound chromosomes. Assembly is essential for understanding the organization and structure of genomes, as it generates contiguous sequences (contigs) representing large sections—the aspiration is chromosomes—of the genome. Short-read assemblies are often fragmented due to difficulties in resolving repetitive regions and allelic variation, especially in heterozygous genomes. Long-read technologies offer a solution to this fragmentation by spanning repetitive regions and complex loci, allowing for more accurate detection of structural variants, alternative haplotypes, and other genomic features.

1.3.3.1 Long-read assembly algorithms

Long-read assembly algorithms fall into two main categories: the Overlap-Layout-Consensus (OLC) and the De Bruijn Graph (DBG) models (reviewed in (Li *et al.*, 2012; Wajid and Serpedin, 2012)).

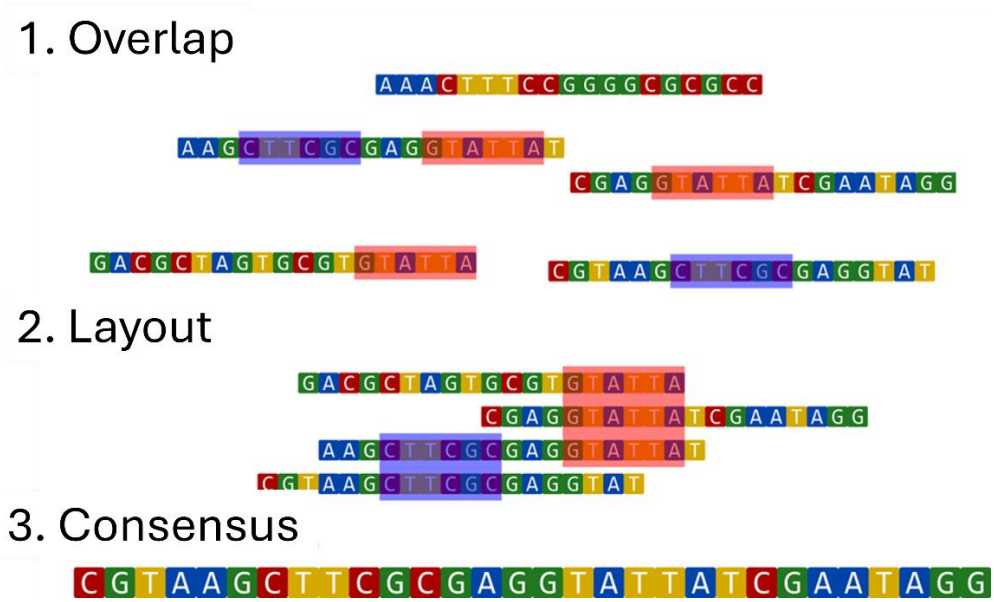

Figure 1.1. The Overlap-Layout-Consensus algorithm. This figure is from (Goussarov *et al.*, 2022); CC BY 4.0.

In the OLC approach, the first step involves a computationally intensive process to generate an overlap graph based on all pairwise alignments of the input sequence reads (overlap step) (Figure 1.1). This is followed by a layout step, where the assembler traverses the overlap graph to generate contiguous sequences, or contigs. The final step, consensus, utilizes the best alignment agreements among reads to correct errors and produce the most accurate contigs (Li *et al.*, 2012).

Conversely, the DBG approach constructs directed graphs from shorter substrings (k-mers) extracted from the long reads (Figure 1.2). The assembler then traverses these k-mer graphs to reconstruct the genome sequence (de Bruijn, 1946; Li *et al.*, 2012; Idury and Waterman, 1995).

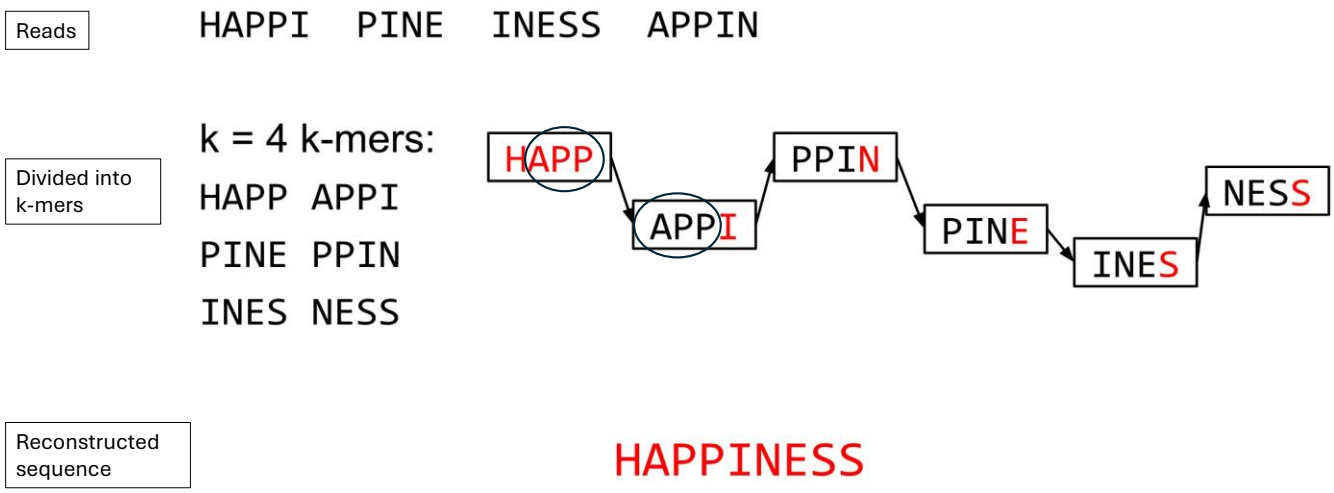

Figure 1.2. The De Bruijn Graph algorithm. This figure was downloaded from https://training.galaxyproject.org/training-material/topics/assembly/images/ex1-4.png; CC-BY-4.0 (Hiltemann *et al.*, 2023; Batut *et al.*, 2018). Each node (indicated by rectangular boxes) is a k-mer and each directed graph (indicated by arrows) is a (k-1) overlap between the suffix (end) of the source read (indicated by the first circle on the left) and the prefix (beginning) of the next read (indicated by the second circle).

The DBG algorithm encounters greater challenges in handling repeats compared to the OLC algorithm. Repeated sequences generate identical overlapping (k-1)-mers with variable nucleotides, leading to multiple branching paths and alternative reconstructions of the repeated sequences (Figure 1.3).

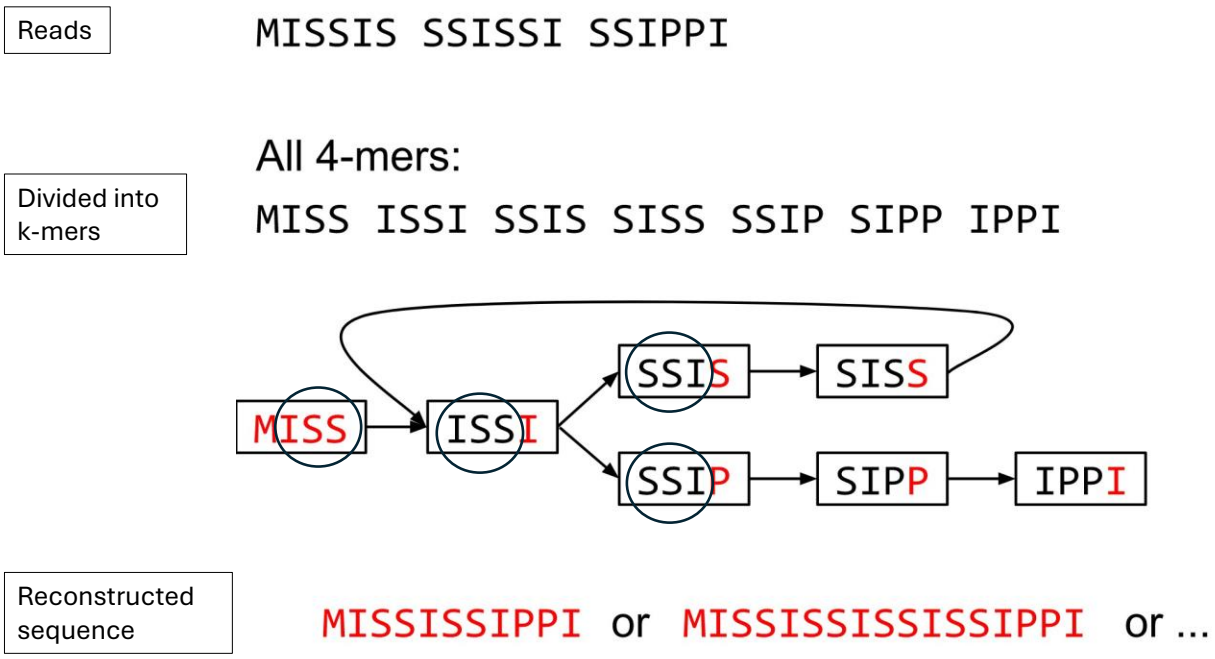

Figure 1.3. The problem with repeats in the De Bruijn Graph algorithm. This figure was downloaded from https://training.galaxyproject.org/training-material/topics/assembly/images/ex2-4.png; CC-BY-4.0 (Hiltemann *et al*., 2023; Batut *et al*., 2018). Each node (indicated by rectangular boxes) is a k-mer and each directed graph (indicated by arrows) is a (k-1) overlap (indicated by circles) between the suffix (end) of the source read and prefix (beginning) of next read.

Heterozygosity presents an additional challenge in genome assembly and confounds both the DBG and OLC assembly algorithms, by creating “bubbles” on assembly graphs if sequences from a heterozygous gene locus are too divergent to be collapsed into a single contig (Figure 1.4). As a result, the sequences are assembled into separate contigs, leading to increased haplotypic duplications and inflated genome sizes.

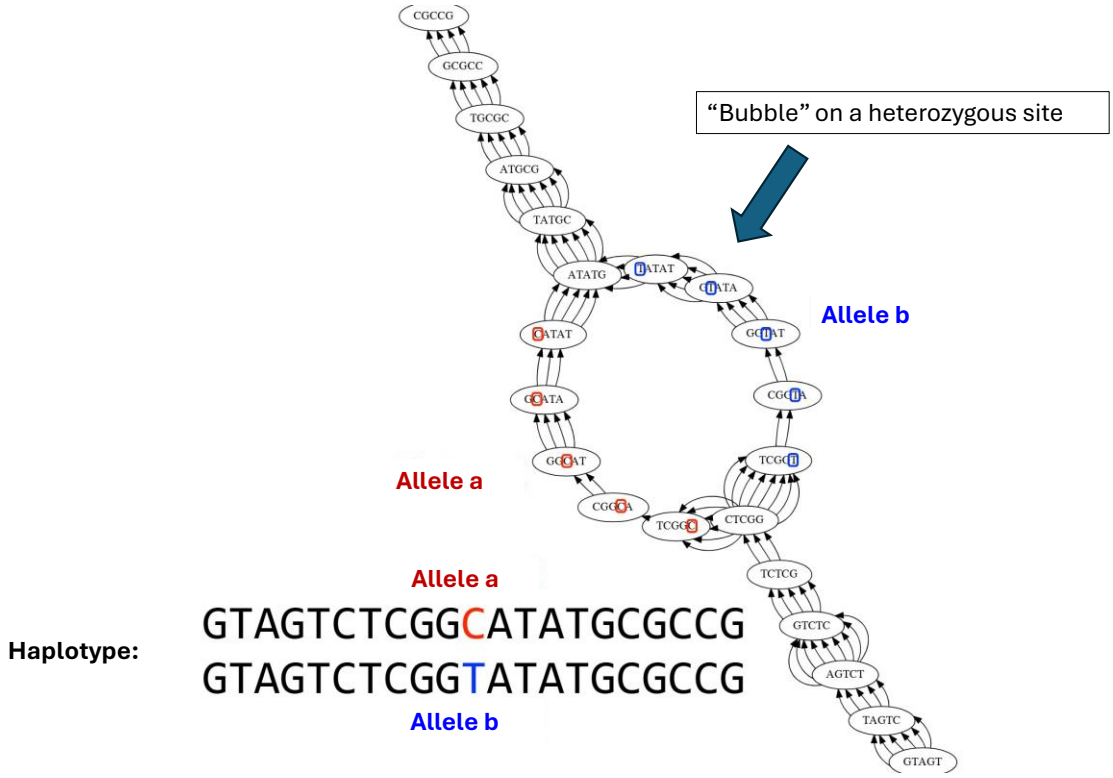

Figure 1.4. The problem with heterozygosity. Figure by Benjamin Langmead and downloaded from ads1-slides/0580_asm__practice.pdf at master · BenLangmead/ads1-slides · GitHub. The blue arrow indicates a "bubble" created on the assembly graph if sequences from a heterozygous gene locus (haplotype) are too divergent to be assembled into a single contig. An example is shown by Allele a (in red) which has a cytosine base and Allele b (in blue) which has a thymine base.

The field of genome assembly is rapidly advancing, with new assemblers continuously improving on existing methods. In this study, I evaluated eight widely used long-read assemblers: Canu, SMARTdenovo, Flye, Redbean, Falcon, Falcon-unzip, HiCanu, and Hifiasm. Selection criteria included their popularity in scientific literature, prior use in nematode genome assembly, and adoption by BioGenome projects (see https://github.com/sanger-tol/genomeassembly) (Pollo *et al.*, 2020; Jung *et al.*, 2020; Stevens *et al.*, 2020; Sun *et al.*, 2021).

Canu and SMARTdenovo use the OLC algorithm. Canu implements a three-step process: base correction, trimming, and consensus building (Koren *et al.*, 2017, 2018). In contrast, SMARTdenovo performs all-vs-all raw read alignment without error correction for consensus

building (Liu *et al.*, 2021). Flye and Redbean use the DBG algorithm. Flye, applies the Generalized DBG algorithm which incorporates repeat graphs to collapse haplotype bubbles (Kolmogorov *et al.*, 2019; Lin *et al.*, 2016). Redbean employs a Fuzzy Bruijn Graph (FBG) to simplify repeat handling by consolidating repeat nodes and connecting different nodes based on adjacency in reads (Ruan and Li, 2020).

Falcon and Falcon-unzip are diploid-aware assemblers that use the Hierarchical Genome Assembly Process (HGAP) to build pre-assemblies from the longest so-called seed reads, followed by contig alignment (Chin *et al.*, 2013, 2016). HiCanu and Hifiasm, optimized for PacBio HiFi reads, offer further advancements. HiCanu, an extension of Canu, introduces features like homopolymer compression, overlap-based error correction, and strict filtering of false overlaps (Nurk *et al.*, 2020). Hifiasm, whose latest release (0.20.0-r639), as of this writing, can also support ultralong ONT reads, performs three rounds of error correction, followed by overlap alignment and string graph construction to create a well-connected assembly, potentially resolving heterozygous bubbles (Cheng *et al.*, 2021).

1.3.3.2 Assembly quality control

Quality control (QC) is essential for producing high-quality genome assemblies, especially to address potential contamination. Foreign DNA can mix with the target species' genomic material during sequencing, introducing contaminants. Tools like BlobTools and machine learning methods, like decision trees, can help screen for contamination sources (Laetsch and Blaxter, 2017; Fierst and Murdock, 2017). ONT and PacBio long reads (Continuous Long Reads (CLR)), prone to higher error rates, allow more accurate contamination detection at the contig level since contigs contain fewer errors than raw reads. For example, bacterial contaminants often assemble into distinct contigs, enabling removal prior to other post-processing steps.

The Illumina short-read technology has standardized QC tools like FastQC (Simon, 2010). However, there is no equivalent standard for long reads; indeed, each platform requires unique QC metrics (Fukasawa *et al.*, 2020). Additionally, the Phred quality score—a widely used base pair resolution accuracy metric—lacks standardization in long-read platforms, requiring high coverages for ONT and PacBio CLR reads to ensure reliable consensus sequences (Fukasawa *et al.*, 2020).

In ONT and PacBio CLR reads, simply increasing sequencing depth does not necessarily increase assembly contiguity or accuracy, as errors accumulate and assembly improvements plateau. Post-assembly QC is crucial. Polishing, a base correction process, uses original

sequence reads to improve base accuracy by correcting insertions and deletions (indels) in the assembly. Multiple aligned reads at each genomic position contribute to base-calling consensus, enabling polishing tools to correct errors. Illumina short reads from the same organism further improve accuracy through iterative polishing, though this method has limitations in repetitive and GC-rich regions (Logsdon *et al.*, 2020). ONT reads, with higher error rates, benefit greatly from polishing. Platform-specific tools like Pilon and Racon (for Illumina) and Medaka and Arrow (for ONT and PacBio CLR, respectively) are widely used for polishing.

In genomes with high heterozygosity, divergent haplotype sequences can lead to redundant contigs, inflating assembly size. Tools like Purge_dups remove these redundant haplotypes, refining haploid assembly representation. The Purge_dups pipeline uses mapped read depth to generate primary contigs (representing the true haploid assembly) and phased contigs (containing divergent haplotigs) (Guan *et al.*, 2020).

1.3.3.3 Assembly evaluation

Several factors affect assembly accuracy including: read depth, read length, genome complexity, sequencing technology, and the assembler used (Cheng *et al.*, 2021; Dominguez Del Angel *et al.*, 2018; Jung *et al.*, 2019; Watson and Warr, 2019). Assembly accuracy is commonly evaluated using multiple metrics (Bradnam *et al.*, 2013). Typically, people assembling genomes prioritize contiguity, completeness, or synteny to a related species (Doyle *et al.*, 2020). The popular N50 metric—indicating the length of the longest contigs that cover 50% of the genome—is widely used to measure contiguity, though this is not a guarantee of accuracy (Thrash *et al.*, 2020). N50 has faced criticism for its tendency to overlook shorter contigs, potentially misrepresenting assembly contiguity (Wang and Wang, 2023). Complementary metrics like N90, fragment count, longest contig length, and the contig-to-chromosome (CC) ratio can provide a fuller picture of assembly contiguity (Wang and Wang, 2023). To assess completeness, tools like Merqury and BUSCO are commonly used (Rhie *et al.*, 2020; Simão *et al.*, 2015). Merqury, compares k-mers from sequence reads with the assembly, identifying copy-number errors (Rhie *et al.*, 2020). BUSCO (Benchmarking Universal Single-Copy Orthologs), evaluates the expected gene content and builds on principals first implemented by the CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline (Parra *et al.*, 2007). These systems use a set of single copy orthologs that are found in a broad range of taxa and the measure of accuracy assumes that these orthologs should occur once, and only once, in the new assembly. BUSCO uses OrthoDB, a database of orthologs partitioned at various taxonomic lineages (Kuznetsov *et al.*, 2022). To assess the completeness of nematode genome

assemblies, two partitions are frequently used: ‘nematoda’ which contains 3131 orthologs or ‘metazoa’ which contains 954 orthologs. Highly contiguous assemblies are expected to have high completeness (a positive correlation between N50 and BUSCO). However, it is not uncommon for assemblies to have low N50 values and high BUSCO scores; similarly, assemblies with varying contiguity can still yield comparable BUSCO scores (Panfilio *et al.*, 2019; Jauhal and Newcomb, 2021; McCartney *et al.*, 2021). There are concerns about an over-reliance on BUSCO scores; using ortholog sets that are pan-phylum or broader may miss conserved genes in species with unique ecological adaptations and may under-represent genes with true copy number variations in the assembled genome (Huelsmann *et al.*, 2019; Palfalvi *et al.*, 2020; Wang and Wang, 2023; Ma *et al.*, 2021; Rhie *et al.*, 2020). Despite these limitations, BUSCO remains a widely used tool for assessing genome completeness, though this is still an area of continual innovation. For instance, the recently developed OMArk system compares proteomes against precomputed gene families across the tree of life to assess genome completeness, gene repertoire consistency relative to closely related species, and detect potential contamination (Nevers *et al.*, 2024).

Finally, assembly accuracy can be evaluated by mapping high-quality reads to the assembly, as performed by the Inspector program (Chen *et al.*, 2021). High mapping rates demonstrate higher accuracy. Mapping RNA-seq data can additionally assess gene completeness and accuracy by evaluating read counts per gene (Liao *et al.*, 2014; Conesa *et al.*, 2016). Mapping genetic markers, such as microsatellites, to contigs or gene loci offers an additional accuracy check.

1.3.4 Genome annotation

Genome annotation is the process of identifying and characterizing genomic elements within the genome assembly including genes, regulatory elements, and repetitive regions. It involves both structural annotation, which determines the location and structure of these features, and functional annotation, which assigns biological functions to them.

1.3.4.1 Repetitive elements

The first step is to identify and mask out repetitive sequences, as these will confound efforts to find the other genomic elements (Yandell and Ence, 2012). Repeats are broadly categorized into low-complexity sequences (such as homopolymeric nucleotide runs) and transposable elements (TEs), which can occupy significant portions of eukaryotic genomes (Kapitonov and Jurka, 2008). For example, TEs cover approximately 45% of the human genome, 37% of the

mouse genome, 20% of the *Drosophila melanogaster*, and 12% of the *C. elegans* genome (The *C. elegans* Sequencing Consortium, 1998; Lander *et al.*, 2001). The sequences of TEs in the same family are often highly divergent and their detection requires specialized tools and custom repeat libraries from a species' taxonomic lineage. Databases such as Dfam and Repbase provide TE family alignments and DNA repeats, capturing intra-family sequence variations (Hubley *et al.*, 2016; Bao *et al.*, 2015; Storer *et al.*, 2021). Tools like RepeatModeler and RepeatMasker use these databases to screen genomes for repeats and generate custom repeat libraries (Tarailo-Graovac and Chen, 2009; Flynn *et al.*, 2020).

1.3.4.2 Structural annotation

Structural annotation identifies genomic features, including protein-coding genes, non-coding elements, regulatory sequences, and repeats. For many researchers the major component is finding the location and intron-exon boundaries of protein-coding genes, a process that uses intrinsic (*ab initio*), extrinsic, and hybrid methods (Dominguez Del Angel *et al.*, 2018).

The intrinsic methods, such as GeneMark, use statistical models that incorporate sequence characteristics that might discriminate between protein-coding and intergenic regions of the assembly (Hoff *et al.*, 2016; Brůna *et al.*, 2021). Examples of these characteristics include long-open reading frames (distance between putative start and stop codons), GC content (the percentage of guanine (G) and cytosine (C) bases in a region), and the codon usage biases. The extrinsic methods, such as exonerate and miniprot, directly align sequences—proteins, transcripts—from public databases to the assembly (Slater and Birney, 2005; Li, 2023). These alignments can be accepted as the appropriate gene model and/or used as inputs, along with the outputs of intrinsic methods, to train models for other gene predictor software, like AUGUSTUS (Stanke *et al.*, 2006). The hybrid methods combine several approaches and generate a consensus or rule-based annotation. The most popular hybrid method is Braker3, currently on version 3 (Gabriel *et al.*, 2024).

1.3.4.3 Functional annotation

Functional annotation assigns biological meaning to the structural features identified. Currently there are three widely used automated routes: searching for protein domains/motifs (via databases like InterPro), identifying orthologues (via the KEGG KO database), and conducting homology searches (e.g., BLASTP against UniProt) (Altschul *et al.*, 1997; Mulder and Apweiler, 2007; Camacho *et al.*, 2009; Pundir *et al.*, 2017; Kanehisa *et al.*, 2016). Functional annotation also serves as a quality control step, helping to identify suspicious genes (such as TEs or

bacterial contamination) based on the presence of unexpected protein domains or orthology assignments (Dominguez Del Angel *et al.*, 2018).

In studies between populations, annotating single nucleotide polymorphisms (SNPs) helps to explore trait associations. SNP annotation tools like SnpEff assess variants based on their genomic positions (Cingolani *et al.*, 2012). Large species-specific variant repositories are publicly available, such as dbSNP for human, the Bovine HapMap for cattle, and CeNDR for *C. elegans* (Sherry *et al.*, 1999; Cook *et al.*, 2017; The Bovine HapMap Consortium, 2009).

1.3.4.4 Annotation evaluation

Annotation accuracy often depends on the contiguity and accuracy of the underlying genome assembly, as errors frequently arise from assembly issues (Wintersinger *et al.*, 2018). Although eukaryotic long-read assemblies now offer improved continuity and fewer gaps, they may still contain indel errors that can alter or truncate protein-coding regions (Jayakumar and Sakakibara, 2019; Phillippy, 2017). Indel errors are particularly problematic for short exons—frequently found at the start of nematode genes—which may go undetected or be discarded during annotation (Stein *et al.*, 2003).

The accuracy of annotation varies across species (Yandell and Ence, 2012; Holt and Yandell, 2011). One way to assess accuracy is to calculate the percentage of predicted proteins containing known protein domains, as provided by InterPro (Blum *et al.*, 2021). Although protein domain counts vary across organisms, the percentage of proteins with at least one functionally relevant protein domain remains relatively stable, generally ranging from 57% to 75% in the proteomes of eukaryotic model organisms (Holt and Yandell, 2011).

The Annotation Edit Distance (AED) metric from the Sequence Ontology Project measures annotation accuracy by assessing concordance between predicted features and overlapping evidence, and can also compare two annotations (Eilbeck *et al.*, 2009; Yandell and Ence, 2012). The manual adjustment of intron-exon boundaries can be carried out on Artemis and Apollo web browsers leading to improved annotations (Lewis *et al.*, 2002; Rutherford *et al.*, 2000; Yandell and Ence, 2012). Finally, BUSCO is widely used to assess proteome completeness. It utilizes HMMER, a profile hidden Markov model-based search tool, to compare predicted proteins against hidden Markov model (HMM) profiles for marker genes in the BUSCO dataset (Eddy, 2011; Simão *et al.*, 2015).

1.4 Comparative genomics

Comparative genomics examines similarities and differences in the genomes of different species to understand their evolutionary relationships, functional biology, and the genetic basis of specific traits and adaptations. In the study of nematodes, comparison of genomes from parasitic and free-living species, along with integrating transcriptomic and proteomic data, can reveal the genetic basis of parasitism (Adams *et al.*, 2020; Wasmuth *et al.*, 2008; Hunt *et al.*, 2016).

1.4.1 Gene families

Comparative genomics has identified several gene families implicated in parasitism evolution in nematodes, particularly those involved in xenobiotic metabolism and immune response modulation (International Helminth Genomes Consortium, 2019; Adams *et al.*, 2020; Stevens *et al.*, 2020; Laing *et al.*, 2015; Johnston *et al.*, 2017; Osbourn *et al.*, 2017; Smyth *et al.*, 2018). Xenobiotic metabolism plays a role in drug and insecticide resistance and understanding these pathways—particularly those involving the cytochrome P450s (CYP), glutathione S-transferases (GST), UDP glucuronosyltransferases (UGT), and nuclear hormone receptors (NHR)—may shed light on the genetic mechanism of anthelmintic resistance (Sharma *et al.*, 2024; Daborn *et al.*, 2002; David *et al.*, 2013; Ménez *et al.*, 2019).

CYP genes participate in Phase I detoxification, catalyzing reactions that increase drug solubility by adding hydrophilic groups (McCarthy and Sinal, 2005). UGT and GST genes are active in Phase II detoxification, conjugating Phase I metabolites with anionic groups (e.g., glutathione) to make them inert and water-soluble, promoting excretion (Susa *et al.*, 2024). NHR genes, as ligand-gated transcription factors, regulate pathways in response to developmental, environmental, and nutritional cues, marking them as potential drug targets (Taubert *et al.*, 2010; Evans and Mangelsdorf, 2014; Wang *et al.*, 2017).

Three gene families have been discovered in *Heligmosomoides bakeri* that modulate the immune system of its host: TGF-β mimics (TGM), ARI, and BARI. TGM proteins bind to host TGF-β receptors, inducing T regulatory immune cells (Johnston *et al.*, 2017). ARI and BARI proteins act as alarmin release inhibitors, blocking interleukin-33 (IL-33), a cytokine essential for immune response initiation against *H. bakeri* in mice (Osbourn *et al.*, 2017).

1.5 Assembly variations

In comparative genomics, the search is for similarities and variation between genomes. Here, it is important to consider biological variation and technical variation. Biological variation is the naturally occurring differences—including single nucleotide variants (SNVs) and gene copy number variation (CNVs)—between individuals, populations, or species. Technical variation is the changes erroneously introduced by sampling, DNA extraction, sequencing library preparation, sequencing itself, and the assembly software (Dohm *et al.*, 2008).

1.5.1 Examples of technical variation

One major source of technical variation is sequencing bias, which can be categorized as either coverage bias or error bias. Coverage bias occurs when sequence reads are unevenly distributed across the genome, often in GC-rich or GC-poor regions, leading to under-representation or even absence of coverage in specific areas (Dohm *et al.*, 2008; Chen *et al.*, 2013; Ross *et al.*, 2013). Although long-read sequencing technologies generally avoid PCR amplification, PCR steps during library preparation can introduce biases, particularly under-representing GC-rich regions in methods such as long-range PCR (Aird *et al.*, 2011; Oyola *et al.*, 2012).

Error bias is another significant source of variation, marked by non-uniform rates of mismatches, insertions, or deletions across the genome (Pinard *et al.*, 2006; Ross *et al.*, 2013; Lee *et al.*, 2016). Platform-specific sources of error further complicate assembly accuracy, as certain sequencing technologies struggle with homopolymers and other repetitive elements. For instance, homopolymeric sequences longer than five bases often introduce indel errors in PacBio single-pass sequencing, while homopolymers, short tandem repeats (STRs), and highly methylated regions cause signal interference in ONT MinION, affecting base-calling accuracy (Audano *et al.*, 2019; Eid *et al.*, 2009; Weirather *et al.*, 2017; Rhoads and Au, 2015; Giesselmann *et al.*, 2019; Mitsuhashi *et al.*, 2019). Despite advances in the hardware and software leading to frequent improvements, it is important to recognise that these errors introduce frameshifts that can alter codons and disrupt protein-coding regions, often resulting in truncated or misassembled genes (Vallender, 2017; Watson and Warr, 2019; Sacristán-Horcajada *et al.*, 2021).

Most types of repetitive sequences are a challenge for assembly algorithms, creating bifurcations in assembly graphs, which appear as either expansions or collapses when evaluated. Interspersed repeats are often reported as expansions, while tandem repeats are

commonly collapsed into a single consensus sequence. Haplotypes with tandem and interspersed repeats across scaffolds generate additional expansions or collapses, increasing assembly variation (Chen *et al.*, 2021). Evaluation programs may also mistake heterozygous breakpoints for structural errors, like haplotype switches, if assemblers struggle to differentiate between haplotypes. This can result in sequences that merge both haplotypes, creating an erroneous assembly where one haplotype appears collapsed while the other shows an expansion (Chen *et al.*, 2021).

Furthermore, highly divergent allele sequences may cause assemblers to incorrectly place them on separate contigs, artificially inflating genome size and gene dosage estimates (Kelley and Salzberg, 2010). Assemblers that collapse homologous haplotypes risk generating false gene duplications, further distorting genomic interpretations (Kelley and Salzberg, 2010; Korlach *et al.*, 2017; Roach *et al.*, 2018; Guan *et al.*, 2020).

1.6 Thesis overview

This thesis describes my investigations into the performance of widely used long-read assembly programs, with a focus on identifying bioinformatics protocols that produce high-quality genome resources for highly heterozygous parasitic nematodes. Accurate genome assemblies are essential for understanding anthelmintic resistance and host-parasite interactions, providing a foundation for drug development and functional genomics studies.

In Chapter 2, I explore the effects of assembly method choice on genome quality and annotation accuracy, examining assemblies and annotations from *Caenorhabditis bovis*, *Haemonchus contortus*, and *Heligmosomoides bakeri*. I show that the choice of assembly software and parameters significantly impacts genome accuracy, with implications for comparative studies.

In Chapter 3, I evaluate whether scaffolding algorithms introduce technical variation that could hinder comparative analyses. I compare scaffolded assemblies of *H. bakeri* generated using different methods, addressing taxonomic misclassifications and clarifying species identity to support reproducible research in *Heligmosomoides* genomics.

In Chapter 4, I test the utility of *H. bakeri* as a model for Clade V parasitic nematodes, focusing on gene families involved in xenobiotic metabolism. By examining gene family evolution, I provide insights into resistance mechanisms and functional pathways relevant to parasitism.

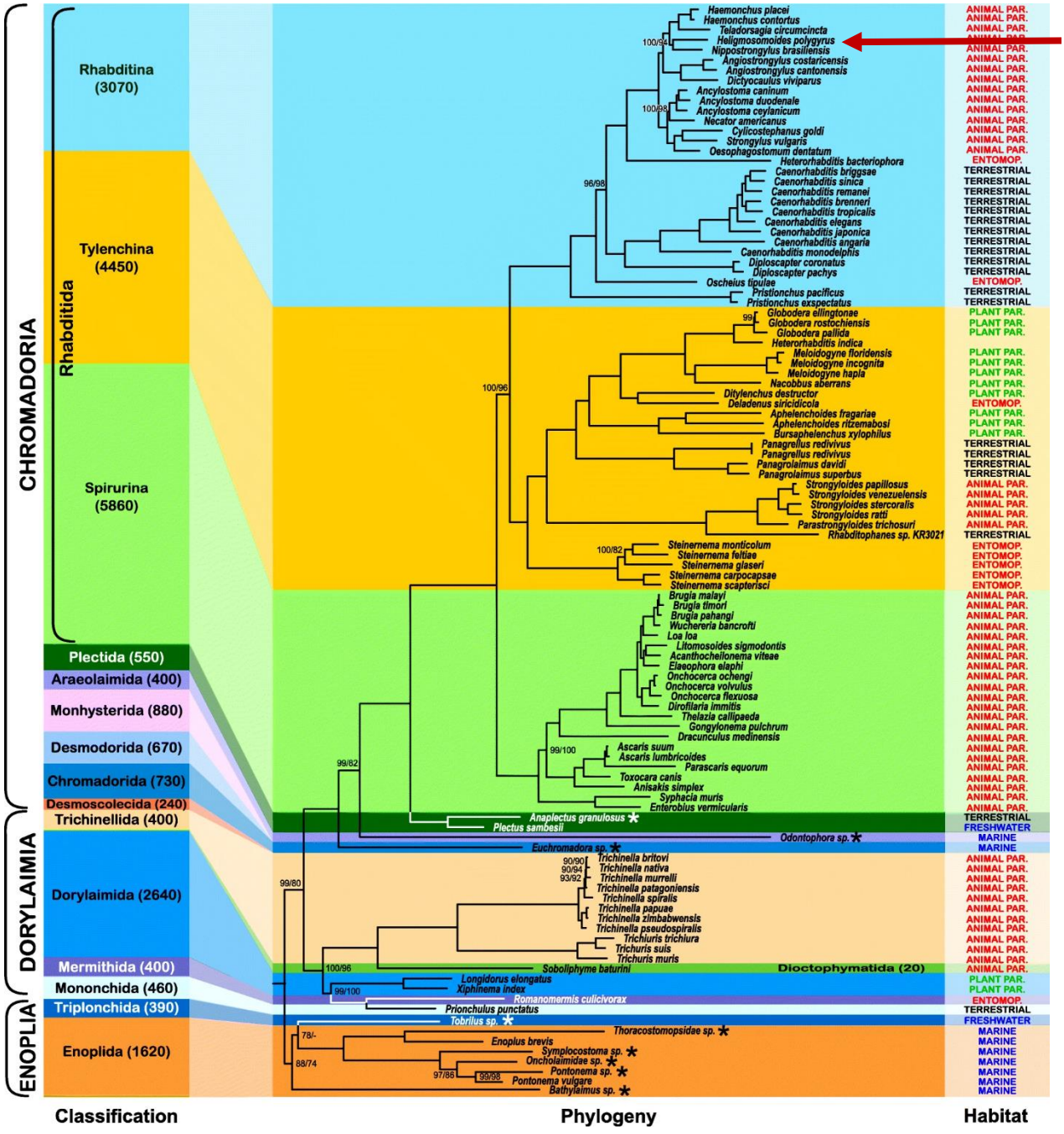

Figure 1.5. Phylogeny of nematodes. My thesis has focused on Clade V nematodes (Rhabditina suborder). This figure is from (Smythe *et al*., 2019). Here, *H. bakeri* is named *H. polygyrus* (indicated by the red arrow), but this misclassification has recently been corrected through genomic comparisons (Chow *et al*., 2019; Stevens *et al*., 2023).

Chapter 2

Genome assembly variation and its implications for gene discovery in nematodes

2.1 Abstract

Genome assemblers are a critical component of genome science, but the choice of assembly software and protocols can be daunting. Here, we investigate genome assembly variation and its implications for gene discovery across three nematode species—*Caenorhabditis bovis*, *Haemonchus contortus*, and *Heligmosomoides bakeri*—highlighting the critical interplay between assembly choice and downstream genomic analysis. Selecting commonly used genome assemblers, we generated multiple assemblies for each species, analyzing their structure, completeness, and effect on gene family analysis. Our findings demonstrate that assembly variations can significantly affect gene family composition, with notable differences in gene families important in anthelmintic discovery and immunomodulation. Despite broadly similar performance using various assembly metrics, comparisons of assemblies with a single species revealed underlying structural rearrangements and inconsistencies in gene content, which would affect downstream analyses. This emphasizes the need for continuous refinement of genome assemblies and their annotations.

2.2 Introduction

Genome assembly software serves a critical, often underappreciated role in the life sciences. Its purpose is to reconstruct the complete sequence of an organism's DNA—its genome—from shorter DNA fragments obtained through one or more sequencing technologies. A typical assembler performs various tasks, such as error correction, read alignment and overlap detection, and consensus building. The decreased cost of sequencing has empowered the launch of large BioGenome consortia which aim to sequence, assemble and annotate the genomes of as many animal and plant species as possible (Ebenezer *et al.*, 2022; Mieszkowska *et al.*, 2022). Individual research groups are also able to sequence the genomes of their favourite species. This places the evolution and refinement of genome assemblers at the core of current and future breakthroughs in health, agricultural, and environmental sciences.

Understanding accuracy of the output of a genome assembler—long strings of As, Gs, Cs, Ts and the occasional N—is helpful, as errors in the produced assembly will carry into and magnify through the annotation and analysis, often without notice. There are many sources of error. One example is high and/or inconsistent levels of heterozygosity in the organism’s genome, which leads to sequence reads being assembled as separate loci rather than collapsed in the haploid assembly. A second example is long regions of repetitive DNA and transposable elements scattered through the genome being incorrectly collapsed or rearranged, leading to gaps. While

some genome assemblers like Platanus specialise in one particular problem, for example heterozygosity (Kajitani *et al.*, 2014), most assemblers are generalists and seek to identify and resolve as many potential errors as possible. There are different approaches to solving these challenges and the result is that different assemblers—when given the same sequence reads as input—will produce different assemblies.

Evaluations of genome assembler performance show that even within a phylum there is no one optimal assembler: the Assemblathon2 competition used 102 metrics to compare assemblies from three chordate species (Bradnam *et al.*, 2013); within the molluscs genome size dictated the choice of preferred assembler (Sun *et al.*, 2021); and others have demonstrated the challenges of constructing a haploid assembly from tetraploid protozoa and plants (Jung *et al.*, 2020; Pollo *et al.*, 2020). These studies, and many others, demonstrate the need to continually evaluate assemblers for one's favourite species, which for us are the nematodes.

The phylum Nematoda is home to species that are parasites of humans, livestock, and crops, free-living contributors to environmental health, and pivotal model organisms that help in the understanding of human biology. Accordingly, nematodes have often led the way in animal genomes: the first animal species to have its genome sequenced was *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998); at least 177 genome assemblies—generated from a broad range of assemblers—are available for at least 135 species (WormBase Parasite (WBPS18; WS285) (Howe *et al.*, 2017). This availability of genomic information has led to comparative genomic studies on various scales, much of it focused on the evolution of gene families (Baker *et al.*, 2021; Wasmuth *et al.*, 2008; International Helminth Genomes Consortium, 2019). For this present study, in addition to understanding how different genome assemblers perform with different nematode species, we wanted to look deeper at the downstream consequences of the technical variation. Previous studies have demonstrated that the quality of genome assembly can critically affect downstream annotation (Florea *et al.*, 2011; Ungar *et al.*, 2024; Watson and Warr, 2019). Additionally, it has been shown that the number of apparent lineage-specific genes is inflated if datasets are generated by different gene finding algorithms (Weisman *et al.*, 2022). This motivated us to examine the effect of genome assembler heterogeneity on gene family comparisons.

We examined assemblies and annotations from three nematode species—*Caenorhabditis bovis*, *Haemonchus contortus*, and *Heligmosomoides bakeri*—which reflect different genomic

features—size and heterozygosity—and sequencing platforms—Oxford Nanopore Technologies (ONT; https://nanoporetech.com/) and Pacific Biosciences (PacBio; https://www.pacb.com/).

Caenorhabditis bovis is associated with bovine parasitic otitis disease and is the only pathogenic member of its genus (Kreis, 1964). Its genome assembly is approximately 63 Mb, making it the smallest *Caenorhabditis* and one of the smallest nematode genomes so far sequenced (Stevens *et al.*, 2020). *Haemonchus contortus* parasitizes small ruminants (sheep and goats) and the emergence of drug resistant strains motivated the study of its genome to understand mechanisms of resistance and develop new drugs and vaccines (Laing *et al.*, 2013). The *H. contortus* genome assembly has undergone multiple rounds of improvement—a long-read assembly was scaffolded with optical mapping and manually curated—to give seven scaffolds, with a total length of ~283Mb (Doyle *et al.*, 2020). Finally, *Heligmosomoides bakeri* is an intestinal parasite of mice that serves as a model for studying host-parasite interactions and testing anthelmintic efficacy (Colomb *et al.*, 2024; Hu *et al.*, 2013). We note that many earlier descriptions of the *H. bakeri* genome were described as *Heligmosomoides polygyrus*, a long-standing misclassification of the study species which has been recently corrected through genomic comparisons (Chow *et al.*, 2019; Stevens *et al.*, 2023).

We generated genome assemblies for each species using several genome assemblers, selected for their popularity in the literature, previous use to assemble nematode genomes, and adoption by BioGenome projects (see https://github.com/sanger-tol/genomeassembly) (Jung *et al.*, 2020; Pollo *et al.*, 2020; Stevens *et al.*, 2020; Sun *et al.*, 2021). These assemblies were annotated with a single, widely used pipeline (Gabriel *et al.*, 2024). We found that most programs returned assemblies with encouraging and similar assembly metrics—size, number of scaffolds, gene completeness—but intra-species comparisons revealed structural rearrangements in the scaffolds and diverse complements of conserved orthologues. We explored the effect of assembly differences on annotation more deeply by carrying out a phylogenomic comparison of large gene families—the cytochrome P450s (CYP), glutathione S-transferases (GST), UDP glucuronosyltransferases (UGT), and nuclear hormone receptors (NHR)—and a key set of immunomodulators, the transforming growth factor beta mimics (TGM), alarmin release inhibitor (ARI), and the ‘binds alarmin receptor and inhibits’ proteins (BARI). We found that despite having the same sequence reads as input, the assemblies’ annotations could vary by as much as 51%.

2.3 Methods

2.3.1 Ethics statement

All animal experiments were approved by the University of Calgary's Life and Environmental Sciences Animal Care Committee (Protocol AC18-0168). All protocols for animal use and euthanasia were in accordance with the Canadian Council for Animal Care (Canada).

2.3.2 Data availability

All genome assemblies and Braker3 gene models generated in this study are available on the Dryad Digital Repository (https://doi.org/10.5061/dryad.p2ngf1vzh). All software commands and parameters are described in https://github.com/GraceMariene/Implications_of_Genome_Assembly_Variations. Sequence reads for the *H. bakeri* genome are available on the sequence read archive (SRA): SRX23938287 and SRX23938288. Supplementary Material can be found in the attached Excel and Word files entitled "Chapter 2 Supplementary Tables.xlsx" and "Chapter 2 Supplementary Figures.docx". Supplementary Figures are referred to in text as Figure S 2.X (and as Supplementary Fig. SX in the separate Word document), while Supplementary Tables are referred to as Table SXX in the separate Excel file.

2.3.3 Sequence data

2.3.3.1 *Caenorhabditis bovis* and *Haemonchus contortus*

DNA sequence reads for *C. bovis* and *H. contortus* were retrieved from the Sequence Read Archive (SRA), and their genomes downloaded from WormBase Parasite (WBPS18; WS285) (Howe *et al.*, 2017; Leinonen *et al.*, 2011). *Caenorhabditis bovis* DNA (BioProject: PRJEB34497) was sequenced using the Oxford Nanopore MinION platform (SRA accessions: ERX3873585, ERX3873584 and ERX3873583; mean coverage: ~183-fold) and the Illumina platform (SRA accession: ERX3606318) (Stevens *et al.*, 2020). *Haemonchus contortus* DNA (BioProject PRJEB506) was sequenced using the PacBio RS and Sequel platforms (SRA accessions: ERS629376 and SRX7677207; mean coverage: ~133-fold) and the Illumina platform (SRA accession: ERS086777) (Doyle *et al.*, 2020). Already available RNA-Seq data was used for genome annotation (see Supplementary Table S1 for accession numbers).

2.3.3.2 *Heligmosomoides bakeri*

We generated *H. bakeri* DNA reads as part of a *de novo* sequencing project (paper *in prep.*). We isolated adult *H. bakeri* worms from experimentally infected C57BL/6 mice (Pollo *et al.*, 2023). We separated females and males and pooled 20 worms of each sex. DNA was extracted using MagAttract HMW DNA kit (Qiagen, Canada, Cat#67563) following the manufacturer's instructions with the following adjustments: increasing magnetic beads incubation time to 15 minutes, and an elution volume of 50 uL. DNA was subsequently purified with the Genomic DNA Clean & Concentrator kit (Zymo Research, USA, Cat# D4064) following the manufacturer's instructions in a final volume of 25 uL. A SMRTbell DNA library was constructed according to the Pacbio protocol for HiFi libraries from low DNA input using SMRTbell Express Template Prep Kit 2.0 (Pacbio, USA, 100-938-900). The female and male samples were barcoded during library preparation and pooled together. The average library size was 12 kb based on Femto Pulse System (Agilent, Canada) analysis. Sequencing was performed on one SMRT 8m cell on a Sequel IIe instrument using Sequel II Binding kit 2.2/Sequel II Sequencing kits 2.0 with 30 hours movie. The library construction and sequencing were carried out at the University of Delaware DNA Sequencing & Genotyping Center. For the assemblies in this study, we combined sequence reads from female and male worms (SRA accessions: SRX23938287 and SRX23938288; mean coverage: ~25-fold). Already available RNA-Seq data was used for genome annotation (see Supplementary Table S2 for accession numbers) (Pollo *et al.*, 2023).

2.3.4 Genome assembly and quality control

2.3.4.1 Generation of preliminary assemblies

For *C. bovis* and *H. contortus* we used six long-read assemblers: Redbean v2.5 (Ruan and Li, 2020), Canu v2.1.1 (Koren *et al.*, 2017), Flye v2.8.2 (Kolmogorov *et al.*, 2019; Lin *et al.*, 2016), SMARTdenovo v1.0.0 (Liu *et al.*, 2021), Falcon-kit v1.8.1 and Falcon-unzip v1.3.7 (Chin *et al.*, 2016). As ONT and PacBio RS reads have error rates higher than PacBio HiFi reads, we followed the protocols from the *C. bovis* reference publication for assembly, decontamination, and polishing (Stevens *et al.*, 2020). For *H. bakeri*, we used three HiFi-compatible assemblers: Flye v2.8.2 (Kolmogorov *et al.*, 2019; Lin *et al.*, 2016), Hifiasm v0.16.1-r375 (Cheng *et al.*, 2021) and HiCanu v2.1.1 (Nurk *et al.*, 2020).

2.3.4.2 Decontamination and polishing of preliminary assemblies

To identify contigs from non-nematode organisms in the assemblies of *C. bovis*, *H. contortus* and *H. bakeri*, we used Blobtools v1.1.1 (Laetsch and Blaxter, 2017). Long reads were aligned

to assembly contigs using minimap2 v2.17-r941 (Li, 2018) and the taxonomic classification determined by searching contigs against the Uniprot Proteome database using DIAMOND v2.0.6.144 (Buchfink *et al.*, 2015; Pundir *et al.*, 2017). For polishing—identification and correction of small errors—the *C. bovis* and *H. contortus* assemblies underwent multiple rounds of polishing using tools appropriate for the sequencing platforms used. Briefly:

1. Four rounds with Racon v1.4.2 using long reads for both *C. bovis* and *H. contortus* assemblies (Vaser *et al.*, 2017),
2. One round with Medaka v1.3.2 (Oxford Nanopore Technologies., 2018) using long reads for *C. bovis* assemblies (https://github.com/nanoporetech/medaka),
3. One round with Arrow v2.3.3 with long reads for *H. contortus* assemblies (installed via the GenomicConsensus package v2.3.3, http://github.com/PacificBiosciences/pbbioconda),
4. Two rounds with Racon v1.4.2 and Pilon v1.14 using Illumina reads from both *C. bovis* and *H. contortus* for the two assemblies (Vaser *et al.*, 2017; Walker *et al.*, 2014).

2.3.4.3 Removal of residual haplotype duplications

Haplotypic variation can, if large enough, lead to a single locus being represented multiple times in the haploid assembly. We identified these residual haplotypic duplications for all *H. contortus* and *H. bakeri* assemblies generated as part of this study. This was not carried out for *C. bovis* due to its low heterozygosity and low BUSCO gene duplication rate (Stevens *et al.*, 2020). We used the Purge_dups v1.2.5 tool (Guan *et al.*, 2020). As recommended in the Purge_dups documentation, we initially applied the default cut-off threshold, followed by two manually set cut-offs derived from the default cutoffs (Supplementary Table S3).

2.3.4.4 Assembly evaluation

For all three species, all the generated assemblies were evaluated for contiguity and completeness using various metrics including: number of fragments, assembly size, N50 (50% of the assembly is contained on contigs equal to or larger than this value), length of the longest contig, and BUSCO (the proportion of universal single-copy orthologs found in the assembly). Briefly, we used BUSCO v4.1.4 to evaluate the completeness of single copy orthologues from the nematoda_odb10 dataset and Inspector v1.0.2 to identify both large- and small-scale errors (Chen *et al.*, 2021; Manni *et al.*, 2021).

2.3.5 Genome synteny

We used Nucmer (from MUMmer v3.23) to create assembly-to-assembly alignments (Marçais *et al.*, 2018). We used the following assemblies as the reference for Nucmer: Redbean v2.5 for *C. bovis*; Doyle for *H. contortus*; Hifiasm for *H. bakeri*. For *C. bovis*, we used the dnadiff script (part of Nucmer) to report one-to-one alignment coordinates and viewed the alignments with Circos v0.69-8 (Krzywinski *et al.*, 2009). For all species, we used NucDiff v2.0.3 to measure large- and small scale differences between each assembly and the reference (Khelik *et al.*, 2017).

2.3.6 Visualising BUSCO content analysis

We generated Sankey diagrams using the SankeyMATIC builder (available at http://sankeymatic.com/build/) to visualize the transitions of BUSCO genes, i.e., those genes that changed categories between assemblies. Set-based BUSCO genes data was displayed using UpSet plots (Lex *et al.*, 2014). For *H. contortus*, we created a heatmap of missing BUSCO genes with the ggplot2 package in R v3.4.2 (Wickham, 2016). Most figures were edited with Adobe Illustrator (available at https://www.adobe.com). Whenever possible, a colour-blind accessible palette was chosen using (https://davidmathlogic.com/colorblind/).

2.3.7 Functional annotation of BUSCO genes

We used the BlastKOALA webserver to identify the putative functions of BUSCO genes missing from the *H. contortus* and *H. bakeri* genome assemblies (Kanehisa *et al.*, 2016).

2.3.8 Genome annotation

2.3.8.1 Soft masking assemblies

In the process of genome annotation, particularly gene finding, it is important to identify and mask out regions of the assemblies which are highly repetitive and/or contain transposable elements. We used RepeatModeler v2.0.3 and RepeatMasker v4.1.2-p1 (Flynn *et al.*, 2020; Tarailo-Graovac and Chen, 2009). For RepeatModeler, we used NCBI/RMBLAST v2.11.0+ to create a repeat library from the Dfam database of transposable elements and DNA repeats (Storer *et al.*, 2021). Next, we used the RepeatMasker util (queryRepeatDatabase.pl script) to create a custom repeat library specific to nematode species. Both repeat libraries were combined as input for the RepeatMasker program.

2.3.8.2 Gene prediction

Protein-coding genes (gene models) in the generated assemblies were predicted using Braker3 (Gabriel *et al.*, 2024) installed using the Singularity container (available at https://hub.docker.com/r/teambraker/braker3). For *H. contortus* and *H. bakeri*, we used short-read RNA-seq data from the SRA as external evidence for the gene prediction; PRJEB506 (Laing *et al.*, 2013) and PRJNA750155 (Pollo *et al.*, 2023) respectively. RNA-seq data was aligned using the STAR v2.7.10a aligner (Dobin *et al.*, 2013) and the generated alignments, in bam-format, used to produce a training set for AUGUSTUS v3.5.0, using GeneMark-ET v4.71 (Hoff *et al.*, 2016). We filtered redundant training gene structures with DIAMOND v0.9.24 (Buchfink *et al.*, 2015). For *C. bovis*—for which no RNA-Seq data is available—we used the longest isoforms from *C. elegans* gene models (BioProject: PRJNA13758) retrieved from WormBase Parasite (WBPS18; WS285) (Howe *et al.*, 2017), and the default protein sequences retrieved from UniProtKB/Swiss-Prot database as the protein evidence for gene prediction (Boutet *et al.*, 2007). Braker3 used the ProtHint pipeline to produce hints using GeneMark-EP v4.71, DIAMOND v0.9.24 and Spaln v2.3.3d tools (Brůna *et al.*, 2020; Buchfink *et al.*, 2015; Gotoh *et al.*, 2014; Iwata and Gotoh, 2012; Lomsadze *et al.*, 2005).

2.3.8.3 Identification of gene family members

Members of the Cytochrome P450 (CYP), Glutathione S-transferase (GST), UDP-glucuronosyltransferase (UGT) and nuclear hormone receptors (NHR) gene families were identified using BLASTP v2.9.0+ against the full *C. elegans* protein set (Camacho *et al.*, 2009). For each species, we clustered the proteins for each family using OrthoFinder2 ('-S blast', all other parameters were default) and included the *C. elegans* proteins to serve as assumed outgroups (Emms and Kelly, 2019).

For the TGM protein family, the protein sequences were taken from the publication (Smyth *et al.*, 2018). The latest ARI and BARI protein sequences were provided by Dr. Henry McSorley (University of Dundee) (Osbourn *et al.*, 2017). These protein sequences were mapped to the genome assemblies using the splice-aware aligners exonerate and miniprot (Slater and Birney, 2005; Li, 2023). For both, we used a scoring threshold which only accepted an alignment which scored within a specified percentage of the top score for a given query. For miniprot we used '--outs 0.9', and for exonerate we used '--percent 90' and '--percent 70'. We found that the results for 'exonerate --percent 70' were broadly similar to 'miniprot --outs 0.9' (Supplementary Table S4).

2.4 Results

We generated fifteen new assemblies—six for *C. bovis*, six for *H. contortus*, and three for *H. bakeri*—which were used, along with the published genomes for each species, for intra-species comparisons. All the assemblies are available (see Data availability section). For readability, we will refer to the published genomes by their papers' first author—Stevens for *C. bovis*, Doyle for *H. contortus*, and Chow for *H. bakeri*—and by the name of the genome assembly program for the others (Chow *et al.*, 2019; Doyle *et al.*, 2020; Stevens *et al.*, 2020). Each new assembly was subject to decontamination and polishing protocols with an additional step of purging duplicates in the highly heterozygous *H. contortus* and *H. bakeri* assemblies. Decontamination—the removal of contigs from bacteria and other heterogenous sources—removed approximately 20% of the nucleotides of any given intermediate assembly and, in *C. bovis*, up to 90% of the scaffolds (Supplementary Tables S5-S7). Polishing—using the sequence reads to identify and fix any small errors—led to notable improvements in the ONT *C. bovis* assemblies; for example, the Redbean2.5 assembly was improved from 79% to 96% of the BUSCO genes found as complete (Supplementary Table S5). However, polishing had a minor impact on the PacBio-based assemblies, *H. contortus* (RSII) and *H. bakeri* (HiFi), presumably because of differences in the error rates between PacBio and ONT reads at the time. Purging duplicates—phasing out allelic repeats in highly heterozygous genomes based on read depth—led to notable reductions in the genome sizes and duplication of BUSCO genes; for example, the BUSCO genes' duplication in the *H. contortus*' Canu assembly was reduced by 48% (Supplementary Table S6).

Below, we compare the assemblies of each species using general assembly statistics, synteny, and BUSCO gene completeness. We then use four gene families to demonstrate how differences in assembly impact downstream analysis.

2.4.1 *Caenorhabditis bovis* assemblies

The Stevens *C. bovis* assembly was generated using ONT reads with the Redbean assembler v2.3 and polished with Racon and Medaka. We generated assemblies using an updated Redbean (v2.5, hereafter Redbean2.5), Flye, Canu, SMARTdenovo, Falcon, and Falcon-Unzip (Chin *et al.*, 2016; Kolmogorov *et al.*, 2019; Koren *et al.*, 2017; Lin *et al.*, 2016; Liu *et al.*, 2021; Ruan and Li, 2020). Following decontamination and polishing (Supplementary Table S5), the overall standard assembly metrics were relatively similar (Table 2.1). Compared to the Stevens assembly, the Redbean2.5 assembly was nearly identical in size but on fewer scaffolds (21 vs

35) and the N50 was 400 kb larger. The Canu assembly size was 10% larger and on almost four times as many scaffolds. The other four assemblies were comparable to the Stevens assembly in terms of assembly size and the number of scaffolds, but the N50s were between 25% and 50% smaller.

Table 2.1. Assembly statistics for *Caenorhabditis bovis.*

Assembly Stats	Stevens	Redbean v2.5	Flye	Canu	SMARTdenovo	Falcon	Falcon-unzip
# Scaffolds	35	21	34	134	32	48	41
Assembly size (Mb)	62.73	62.72	62.93	69.16	63.39	64.93	64.40
N50 (Mb)	7.60	8.03	5.86	4.01	3.56	4.41	4.41
Longest contig (Mb)	10.86	10.08	10.12	12.16	6.29	9.90	9.79
BUSCO % n=3131	C:96.2 [S:95.6, D:0.6] F:1.3, M:2.5	C:96.0 [S:95.3, D:0.7] F:1.4, M:2.6	C:96.4 [S:95.5, D:0.7] F:1.0, M:2.6	C:96.3 [S:93.2 D:3.1] F:1.1, M:2.6	C:96.1 [S:93.2 D:0.9] F:1.1, M:2.8	C:96.1 [S:95.0 D:1.1] F:1.2, M:2.7	C:96.2 [S:95.1 D:1.1] F:1.1, M:2.7
Inspector QV (Quality Value)	25.1	25.1	25.0	25.3	25.1	25.0	25.2
Structural errors	12	8	9	10	16	21	9

BUSCO genes' categories: Complete 'C', is the sum of single copy 'S' and duplicated 'D' genes; 'F' denotes Fragmented genes, and 'M' denotes Missing genes (all % from 3131 genes). N50 indicates the length of the longest contigs that cover 50% of the genome.

Through BUSCO analysis, we found a similar percentage of genes were complete across the seven assemblies (Table 2.1 and Supplementary Tables S8-S14). A notable difference was the duplication rate in the Canu assembly, which was five times that of the Stevens assembly: 18 genes vs 97 genes. Of the 3131 BUSCO orthologues, 3040 genes (97%) were complete (single copy or duplicated) in at least one assembly and 94.9% (2972) were complete in all seven assemblies (Supplementary Fig. S1). Of these complete genes, 15 were duplicated in all assemblies. Ninety-one genes were missing or fragmented in all seven assemblies, which suggests they are undergoing pseudogenization or have been lost as a consequence of a genome size reduction in *C. bovis* (Stevens *et al.*, 2020). Ninety-one (3%) were absent (fragmented or missing) in all seven assemblies. In total, the BUSCO classification differed for 178 genes, *e.g.* from Duplicated to Fragmented (Figure 2.1A and Supplementary Tables S15 and S16). Expectedly, there was strong agreement between the Stevens (Redbean2.3) and the new Redbean2.5 assemblies; nine genes complete and single in the Stevens assembly were fragmented or missing in the Redbean2.5 assembly with six genes in the reverse situation (Figure 2.1A). Disagreements on BUSCO classification was more striking between the other

assemblies. For example, SMARTdenovo and Falcon had near identical BUSCO scores, but 36 single copy genes in the SMARTdenovo assembly were classified as duplicated (16), fragmented (13) or missing (7) in the Falcon assembly. In a direct comparison of Redbean2.5 and Canu assemblies, 18 genes considered fragmented or missing in Redbean2.5 were found complete and single copy in Canu (Figure 2.1B and Supplementary Table S17).

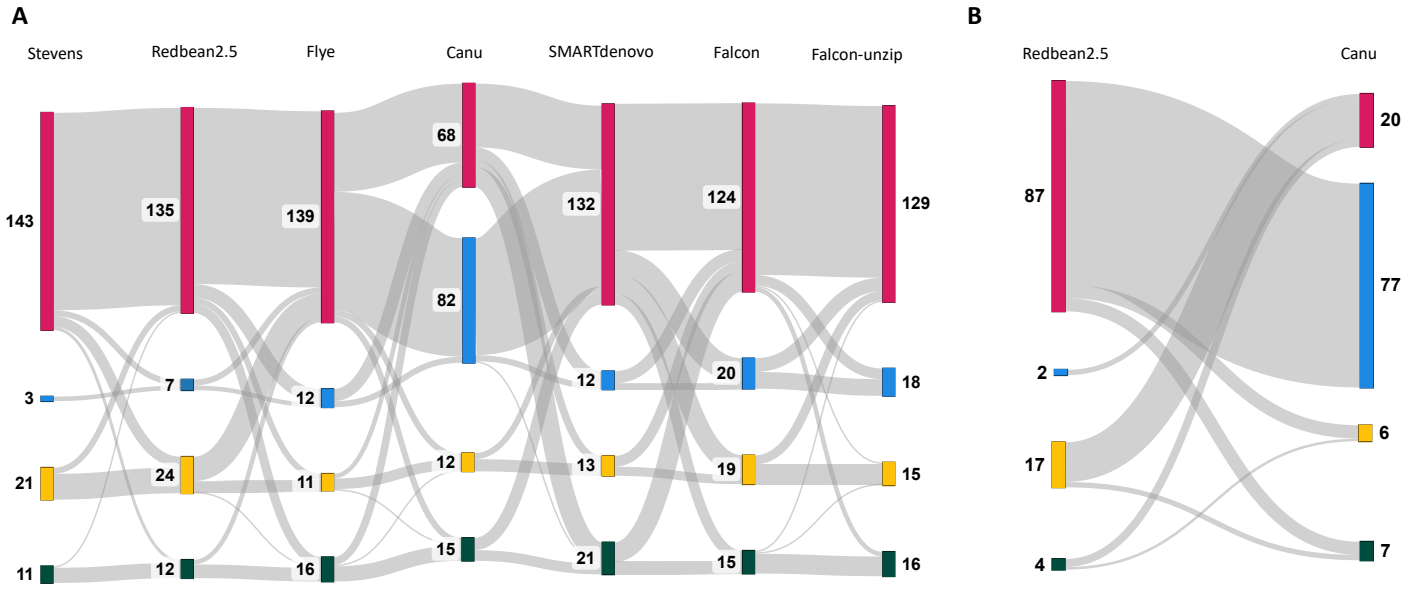

Figure 2.1.Transitions of BUSCO genes across multiple *Caenorhabditis bovis* assemblies. The Sankey diagrams show how BUSCO genes switch categories: red – complete and single copy genes; blue – complete and duplicated; yellow – fragmented; green – missing. (A) The 178 genes that changed BUSCO categories in at least one assembly. The order of the assemblies is arbitrary and follows Table 2.1. The data for this panel is in Supplementary Tables S15 and S16. (B) Comparison of the 110 genes which differed between Redbean2.5 and Canu assemblies. The data for this panel is in Supplementary Table S17.

We used the original ONT reads to evaluate the reads alignment to each of the assemblies, identifying large structural and small errors (Chen *et al.*, 2021). All the assemblies had near identical Inspector quality values (QV) as determined by the Inspector software; the Redbean2.5 assembly had the fewest large structural errors, and the Flye assembly had the fewest small-scale errors (Table 2.1 and Supplementary Table S18).

Next, we investigated the structural variation—insertions, deletions, and translocations—between assemblies. We needed a reference assembly and considered Redbean2.5 assembly to have the best accuracy metrics, albeit slightly. As expected, the Stevens and Flye assemblies—both use the De Bruijn Graph (DBG) algorithm—had the fewest number of differences (Table 2.2 and Supplementary Table S19). These were followed by the SMARTdenovo assembly, which used the Overlap-Layout-Consensus (OLC) algorithm, and the related Falcon and Falcon-Unzip assemblies, which used the Hierarchical Genome Assembly Process (HGAP). The OLC-based Canu assembly had the most differences to the Redbean2.5 assembly. For all assemblies, at least 97% of the variants were insertions or deletions of typically short regions (1-to-10bp). However, we did find longer duplications and tandem duplications which could impact accurate curation of genes. Using Braker3 to annotate the assemblies with gene models, we found five genes from the Stevens assembly—four genes in tandem duplications and one gene on a duplication—and 19 genes from the Redbean2.5 assembly—17 genes in tandem duplications and two on duplications—in these regions (Supplementary Table S20).

Table 2.2. NucDiff results for *Caenorhabditis bovis* assemblies.

	Stevens (Redbean2.3)	Flye	Canu	SMARTdenovo	Falcon	Falcon-unzip
Total number	8,215	8,527	62,971	12,895	19,319	16,540
Insertions	2,421	3,350	25,809	4,913	8,002	6,718
Deletions	5,553	4,916	36,399	7,684	10,929	9,363
Translocations	194	207	478	218	290	336
Relocations	29	18	142	38	39	57
Reshufflings	8	11	32	15	21	18
Inversions	10	25	109	27	38	48
Unaligned seq	0	0	2	0	0	0

Each assembly was aligned to the reference assembly, Redbean2.5.

Larger structural variants—translocations, relocations, reshufflings, and inversions—were present even between Redbean assemblies (Table 2.2), although the largest changes were split scaffolds; the Redbean2.5 Scaffold 3 was split across the Stevens Scaffolds 6, 7, and 11 (Figure 2.2). This suggests that assemblies might be improved through using alternative assemblies to join scaffolds which split in another assembly. For example, in the better Redbean2.5 assembly, the Scaffolds 4 and 8 are linked through Scaffold 1 in the Stevens and SMARTdenovo, and

Scaffold 4 in the Falcon-based assemblies (Figure 2.2). For comparisons against the other assemblies, translocations and relocations typically manifested as short scaffolds, typically at one end of the Redbean2.5 assembly. We noted that two scaffolds of the Canu assembly did not align to the Redbean2.5 assembly. Through a BLASTX search against the NCBI NR database (Camacho *et al.*, 2009), we found that these were most likely of bacterial origin and were not removed by the decontamination protocols.

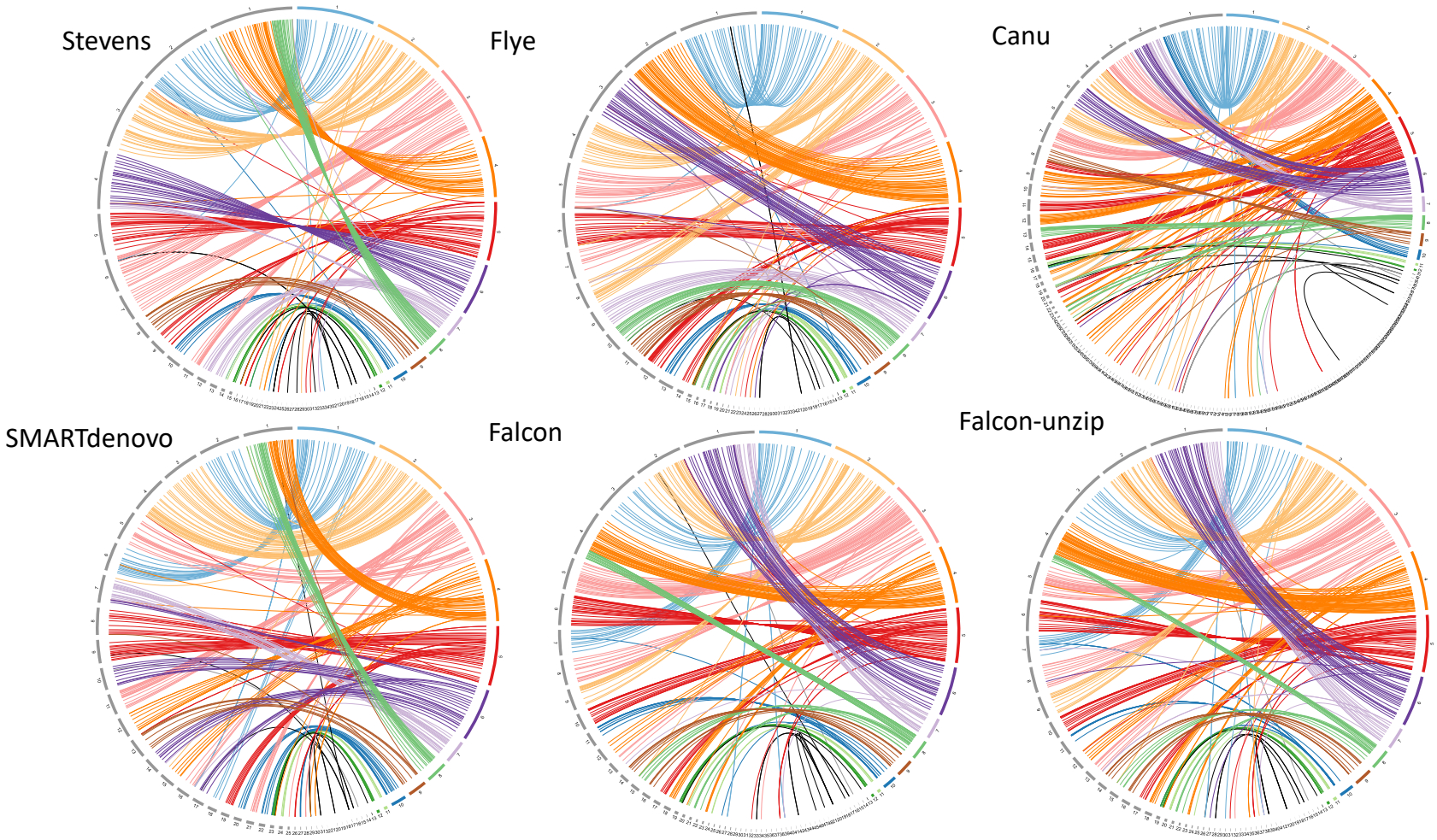

Figure 2.2. Synteny across six genome assemblies for the *Caenorhabditis bovis*. In each Circos plot, the reference assembly (Redbean2.5) is represented by the coloured scaffolds and the labeled assembly is represented by the gray scaffolds. The lines linking syntenic regions are coloured according to the reference scaffold. The scaffolds are ordered by length (longest to shortest) and for clarity we only label the first ten scaffolds.

2.4.2 *Haemonchus contortus* assemblies

The published *H. contortus* assembly is the result of an ongoing multinational consortium and involved deep sequencing using multiple platforms and labour-intensive manual identification and correction of assembly errors (Doyle *et al.*, 2020). We generated six assemblies using only the PacBio reads that were used in the Doyle assembly. These six assemblies, once decontaminated and polished, varied considerably in their length, fragmentation, and completeness (Table 2.3). There was a 26% difference between the longest (Canu 341 Mb) and shortest (Falcon-unzip 262 Mb) assemblies. The SMARTdenovo assembly was the most contiguous, with the fewest scaffolds and largest N50. The Redbean2.5 and Falcon assemblies were the closest to the Doyle assembly in terms of assembly length.

For the Inspector reports, we used both the reads and the Doyle assembly for reference-free and reference-guided metrics, respectively (Table 2.3 and Supplementary Table S21). We were initially surprised to see that the Doyle assembly had the lowest QV score. Here, it is important to note that this score is reference-free, and a function of read alignments to the assembly. The Doyle assembly has undergone extensive curation, in which data, including optical mapping, was used to identify mistakes in the preliminary long-read driven assembly. So, a higher level of disagreement with the raw sequence reads might be expected. For example, the Doyle assembly had the highest number of haplotype switches, likely a consequence of the manual improvement that sought to maximize scaffold continuity (Supplementary Table S21). For the six assemblies that we generated, the Inspector metrics were similar. The coverage for the Doyle-to-Doyle alignments was 98%, which validates the effectiveness of the Inspector software. For the alignments between the new assemblies and the Doyle assemblies the coverage was relatively low (79% - 84%), highlighting the notable differences between the new assemblies. The duplication rate—where the assembly-to-Doyle coverage was two or more—was lowest for the Redbean, Falcon and Falcon-unzip assemblies and highest in the Canu assembly.

Table 2.3. Assembly statistics for *Haemonchus contortus*.

Assembly Stats	Doyle	Redbean v2.5	Flye	Canu	SMARTdenovo	Falcon	Falcon-unzip
# Scaffolds	7	1844	3249	2364	1473	2402	1975
Assembly size (Mb)	283.44	296.37	331.35	341.72	314.72	267.14	262.37
N50 (Mb)	47.40	0.31	0.24	0.34	0.45	0.23	0.25
Longest contig (Mb)	51.80	3.45	4.85	2.76	4.47	3.22	3.24
BUSCO % n=3131 lineage=nematode	C:94.8 [S:93.7, D:1.1], F:0.7, M:4.5	C:93.9 [S:86.7, D:7.2], F:0.9, M:5.2	C:94.1 [S:85.0, D:9.1], F:1.2, M:4.7	C:95.2 [S:86.6, D:8.6], F:0.8, M:4.0	C:94.7 [S:86.5, D:8.2], F:1.2, M:4.1	C:88.2 [S:84.2, D:4.0], F:1.5, M:10.3	C:87.7 [S:84.2, D:3.5], F:1.4, M:10.9
Inspector Reference-free (mapping reads-to-assembly):							
Inspector QV (Quality Value)	22.0	42.9	39.6	44.6	39.1	31.8	42.8
Structural errors	170	53	146	67	180	826	62
Inspector results Reference-guided (mapping assemblies-to-Doyle):							
Coverage of Doyle (%)	98	84	80	84	84	79	79
Reference bases with depth 0; 1; 2+ (%)	2; 98; 0	16; 82; 2	20; 76; 4	16; 78; 6	16; 81; 3	21; 77; 2	21; 77; 2
Number of collapses	0	1197	1633	1512	1234	6370	1019
Number of expansions	0	1004	1136	1127	1113	845	888
Number of inversions	0	1	0	0	0	2	0

BUSCO genes' categories: Complete 'C', is the sum of single copy 'S' and duplicated 'D' genes; 'F' denotes Fragmented genes, and 'M' denotes Missing genes (all % from 3131 genes). N50 indicates the length of the longest contigs that cover 50% of the genome.

In considering the BUSCO results, the six new assemblies varied more than those of the other two species, with between 87.7% to 95.2% of genes found complete. The duplication rate was notably higher in the six assemblies compared to the Doyle reference. Of the 3131 pan-nematode orthologous genes, 3020 genes (96.5%) were complete in at least one assembly, with 2525 genes (80.6%) complete in all seven assemblies and 94 genes (3.0%) missing or fragmented in all seven (Supplementary Figs. S2-S4). It is striking that fewer genes were missing in the Canu and SMARTdenovo assemblies compared to the Doyle assembly. We looked more broadly and found that of the 141 genes missing in the Doyle assembly, 56 (39.7%) were complete in at least one of the new assemblies and 27 in all six assemblies (Figure 2.3 and Supplementary Table S22). It is likely that these represent errors in the current reference assembly.

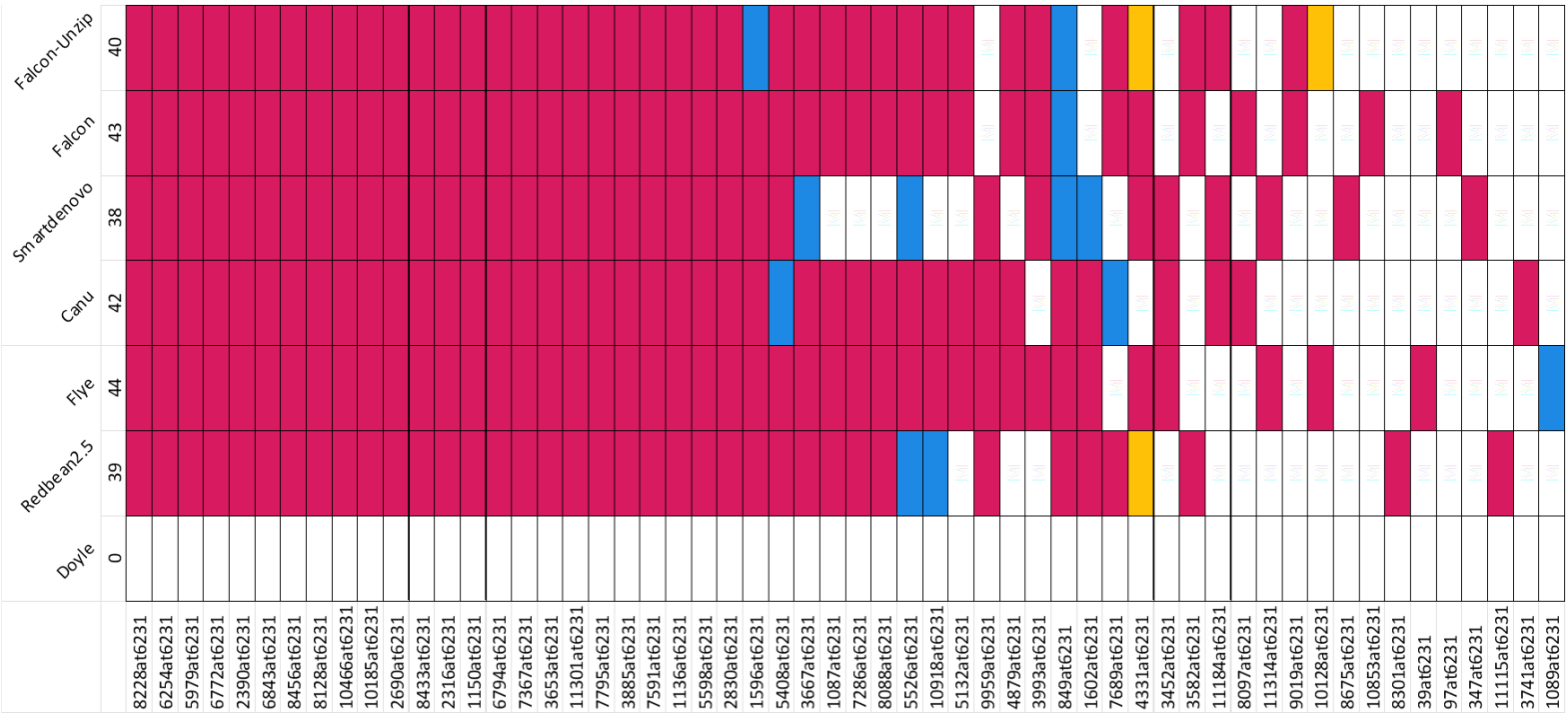

Figure 2.3. BUSCO genes missing from the *Haemonchus contortus* Doyle assembly but complete in at least one other assembly. Red – complete and single copy genes; blue – complete and duplicated; yellow – fragmented; white – missing. The data for this figure is in Supplementary Table S22.

We used BlastKOALA to annotate these 141 genes and found at least three noteworthy examples (Supplementary Table S23) (Kanehisa *et al.*, 2016). One example is *let-756* which is involved in the fibroblast growth factor receptor signaling pathway essential in nematode larval development. The knock-down of *let-756* in *C. elegans* by RNA interference (RNAi) induces larval arrest in worms (Roubin *et al.*, 1999). A second example is *nck-1*, which enables kinase activity required for neuronal guidance. The knock-down of *nck-1* in *C. elegans* negatively disrupts axon guidance, neuronal cell positioning, and causes abnormalities in the excretory canal cell, gonad, and male mating (Mohamed and Chin-Sang, 2011). A third example is *gcn-2*, a kinase involved in the nutrient-sensing and, while knock out is not fatal in *C. elegans*, it does severely impact *C. elegans'* survival during nutrient stress, specifically amino acid limitation (Rousakis *et al.*, 2013).

Of the 35 BUSCO genes duplicated in the Doyle reference assembly, 29 were also duplicated in most of the new assemblies and three were single copy in all six new assemblies (Supplementary Fig. S2). There were large exchanges between single copy and duplicated categories, suggesting that assembly algorithms successfully collapsed divergent haplotypes of different genes at different rates (Figure S 2.1 and Supplementary Table S24). The presence or absence of BUSCOs between the new assemblies was striking. For example, 44 genes considered missing or fragmented in the Doyle reference assembly ('red') were classified as complete and single copies in the Redbean2.5 assembly (Figure S 2.1 and Supplementary Table S24). Another example is the 62 genes considered missing or fragmented in the Redbean2.5 assembly but found complete and single copy in Flye (Figure S 2.1), with 66 genes reclassified in the other direction (Figure S 2.2 and Supplementary Table S25). Both assemblers use the DBG algorithm. The full list of BUSCO genes for each assembly can be found in Supplementary Tables S26-S32.

2.4.3 *Heligmosomoides bakeri* assemblies

At the start of our study, there were two published *H. bakeri* assemblies; one generated from short-reads only and one that used a low-coverage of PacBio reads to scaffold together a short-read assembly (Chow *et al.*, 2019; International Helminth Genomes Consortium, 2019). We selected the Chow assembly as it was more contiguous and had a higher BUSCO score. As we finished our study, a chromosome-scale assembly for *H. bakeri* was published (Stevens *et al.*, 2023). While more contiguous than the Chow assembly, we note that the Chow assembly had a higher BUSCO score (Table 2.4). As part of our own *H. bakeri* genome project, we generated

HiFi PacBio reads and assembled them with three different software: Flye, HiCanu, and Hifiasm. The assembly lengths differed less than 6%, but there was considerable variation of N50 and longest contig (Table 2.4). The three newly generated assemblies exhibited comparable performance in the reference-free Inspector analysis. However, the Hifiasm assembly, despite its higher contiguity, exhibited more structural errors (Tables 2.4 and Supplementary Table S33).

Table 2.4. Assembly statistics for *Heligmosomoides bakeri*.

Assembly Stats	**Chow**	**Flye**	**HiCanu**	**Hifiasm**	**Stevens *et al.* 2023 (Chromosome-scale)**
# Scaffolds	23647	2380	1115	284	321
Assembly size (Mb)	696.95	658.60	685.42	678.45	649.16
N50 (Mb)	0.18	0.94	1.9	4.94	110.8
Longest contig (Mb)	1.25	5.94	11.70	20.45	114.7
BUSCO % n=3131 lineage=nematode	C:94.3 [S:93.1, D:1.2], F:1.8, M:3.9	C:94.6 [S:92.5, D:2.1], F:1.8, M:3.6	C:95.0 [S:92.2, D:2.8], F:1.9, M:3.1	C:94.9 [S:92.0, D:2.9], F:1.7, M:3.4	C:93.8 [S:92.6, D:1.2], F:0.7, M:5.5
Inspector QV (Quality Value)	22.1	36.0	35.5	32.5	-
Structural errors	34	116	74	268	-

BUSCO genes' categories: Complete 'C', is the sum of single copy 'S' and duplicated 'D' genes; 'F' denotes Fragmented genes, and 'M' denotes Missing genes (all % from 3131 genes). N50 indicates the length of the longest contigs that cover 50% of the genome. The chromosome-level Stevens *et al.* 2023 assembly was only published as we completed our study and was not included in our comparisons.

Similar to *C. bovis* and *H. contortus*, the *H. bakeri* assemblies, despite having extremely similar BUSCO scores, contained different cohorts of BUSCO genes (Figure 2.4 and Supplementary Table S34). A total of 120 genes were complete in three assemblies and absent (fragmented or missing) in one (Figure 2.5A). We note that while 58 of these genes were absent in the Chow assembly, 62 genes were fragmented or missing in one of the HiFi-based assemblies. We explored the functions of these 62 genes to determine their potential impact on hypotheses regarding *H. bakeri* biology (Supplementary Tables S35-S41). We found several genes whose absence, if real, could significantly affect a *H. bakeri* worm. One example is *mogs-1*, absent only in the Flye assembly, which encodes a mannosyl-oligosaccharide glucosidase which is involved in N-glycan biosynthesis and plays a pivotal role in digestion of bacteria (Geng *et al.*, 2022). A second example is *gpb-1,* also absent in the Flye assembly, which is a G-protein beta subunit

and whose RNAi knock-down is embryonic lethal (Kamath *et al.*, 2000). A third example, absent only in the HiCanu assembly, is *fmo-4*, which encodes for flavin-containing monooxygenase crucial for osmoregulation and RNAi knock-down is also embryonic lethal (Hirani *et al.*, 2016; Rual *et al.*, 2004).

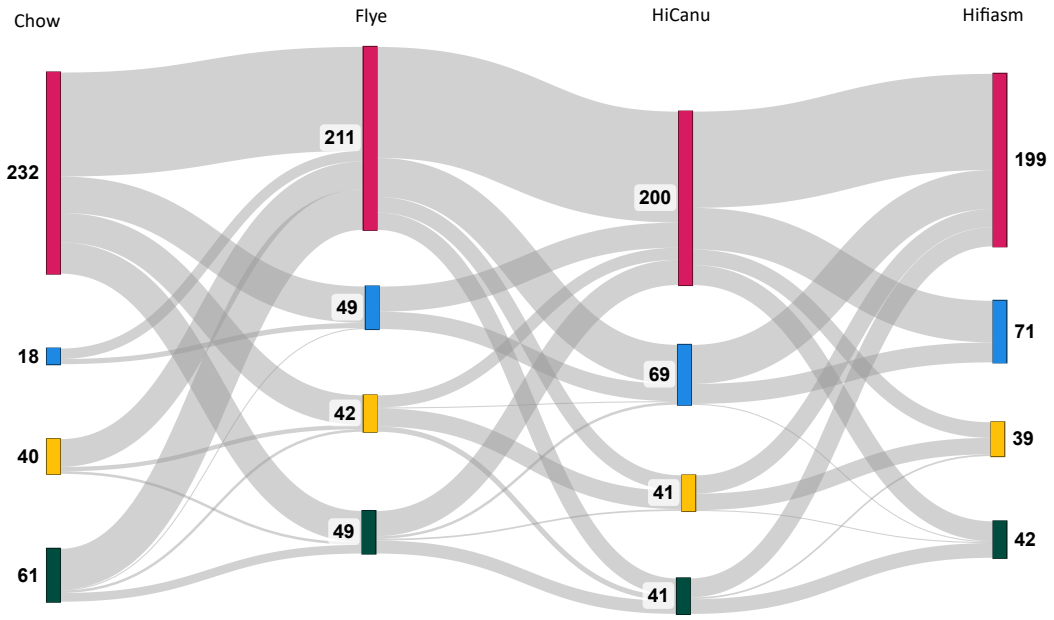

Figure 2.4. Transitions of BUSCO genes across multiple *Heligmosomoides bakeri* assemblies. The Sankey diagram shows how 351 genes changed categories in at least one *H. bakeri* assembly: red – complete & single copy genes; blue – complete & duplicated; yellow – fragmented; green – missing. The order of the assemblies is arbitrary and follows Table 2.4. The data for this figure is in Supplementary Table S34.

While, marginally more BUSCO genes were found in the HiFi assemblies compared to the Chow assembly, it came at the cost of a higher duplication. We found that 165 BUSCO genes were duplicated in at least one assembly; the majority of these were assembly specific with 19 found duplicated in all four assemblies and a further 10 found duplicated in the HiFi-based assemblies (Figure 2.5B). This tempts speculation that the ~29 genes are indeed duplicated. Moreover, these observations illustrate that assembly-specific differences affect the

completeness of the most highly conserved genes and, consequently, the overall recovery of other genes.

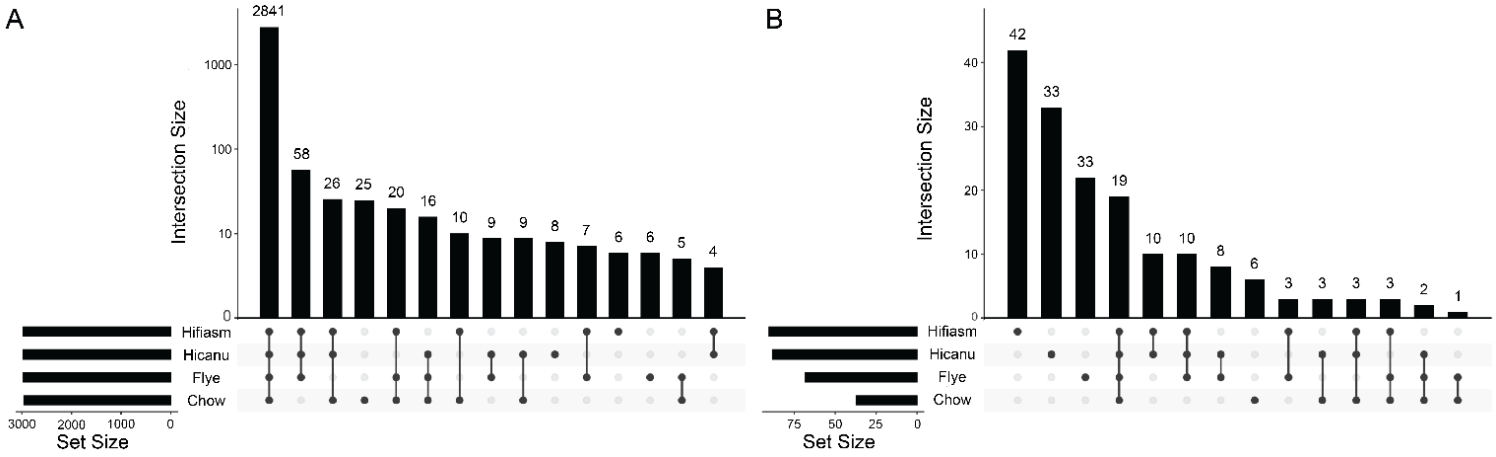

Figure 2.5. UpSet plots of (A) Complete BUSCOs present in *Heligmosomoides bakeri* assemblies, and (B) Duplicated BUSCOs in the assemblies.

2.4.4 Impact of assembly variation on gene family analysis

We generated protein-coding gene models for all assemblies using Braker3 (Table 2.5 and the Data availability section) (Gabriel *et al.*, 2024). For consistency, we used Braker3-generated gene models for the published genomes (see Discussion).

Table 2.5. Protein-coding genes predicted for the different assemblies.

Species	Assembly	Number of protein-coding genes	BUSCO % n=3131
Caenorhabditis elegans	WS285	19,981	C:99.7 [S:99.3, D:0.4], F:0.1, M:0.2
Caenorhabditis bovis	Stevens – published	13,128	C:93.2 [S:92.4, D:0.8], F:1.0, M:5.8
	Stevens – Braker3	14,945	C:97.3 [S:96.5, D:0.8], F:0.7, M:2.0
	Canu	15,010	C:97.3 [S:94.3, D:3.0], F:0.5, M:2.2
	Falcon	14,143	C:96.9 [S:95.8, D:1.1], F:0.5, M:2.6
	Flye	14,070	C:97.1 [S:96.2, D:0.9], F:0.6, M:2.3
	Redbean2.5	14,001	C:97.2 [S:96.4, D:0.8], F:0.5, M:2.3
	SMARTdenovo	13,863	C:97.2 [S:96.2, D:1.0], F:0.5, M:2.3
	Falcon-unzip	13,988	C:97.1 [S:95.8, D:1.3], F:0.5, M:2.4
Haemonchus contortus	Doyle – published	19,621	C:96.2 [S:94.6, D:1.6], F:0.5, M:3.3
	Doyle – Braker3	14,824	C:92.7 [S:91.1, D:1.6], F:0.8, M:6.5
	Canu	18,808	C:93.0 [S:83.1, D:9.9], F:1.1, M:5.9
	Falcon	16,259	C:86.0 [S:81.7, D:4.3], F:2.4, M:11.6
	Flye	19,225	C:91.6 [S:81.6, D:10.0], F:1.4, M:7.0
	Redbean2.5	17,220	C:92.7 [S:84.7, D:8.0], F:1.1, M:6.2
	SMARTdenovo	17,589	C:93.7 [S:85.1, D:8.6], F:1.2, M:5.1
	Falcon-unzip	15,890	C:85.8 [S:81.9, D:3.9], F:1.7, M:12.5
Heligmosomoides bakeri	Chow – published	23,471	C:90.5 [S:88.8, D:1.7], F:2.9, M:6.6
	Chow – Braker3	23,218	C:92.1 [S:90.5, D:1.6], F:2.5, M:5.4
	Flye	20,801	C:95.1 [S:92.7, D:2.4], F:1.3, M:3.6
	HiCanu	20,259	C:94.9 [S:92.0, D:2.9], F:1.2, M:3.9
	Hifiasm	20,079	C:95.1 [S:92.0, D:3.1], F:1.0, M:3.9

For the published genomes, we report both the published protein-coding gene models and our reannotation. For all other assemblies, the protein-coding gene models were generated with Braker3. BUSCO scores (C=complete, S=single copy, D=duplicated, F=fragmented, M=missing) are percentages where *n*=3131.

For all protein sets, we identified members of four gene families—cytochrome P450s (CYP), glutathione S-transferases (GST), UDP glucuronosyltransferases (UGT), and nuclear hormone receptors (NHR)—which have been implicated in anthelmintic studies and are among the largest in *C. elegans* (Supplementary Tables S42-S46). For each gene family, within each species, we clustered the proteins with those from *C. elegans* (Emms and Kelly, 2019). Here, our assumption is that if the differences in genome assemblies did not affect the genome annotation, we would expect that every cluster would have an identical number of proteins from

each assembly and that these would form perfect one-to-one orthologous groups to the exclusion of *C. elegans*. Any deviation from this would demonstrate that differences in the assemblies led to different gene model predictions.

Table 2.6. Percentage differences between the minimum and maximum number of the predicted members of the CYP, GST, UGT and NHR gene families across assemblies.

	CYP	**GST**	**UGT**	**NHR**
Caenorhabditis bovis:				
Number found	min: 44 max: 58 diff: 27%	min: 38 max: 41 diff: 8%	min: 39 max: 51 diff: 27 %	min: 144 max: 160 diff: 11%
Number of clusters	42	35	16	144
Identical size[a]	all: 26 new: 26	all: 32 new: 32	all: 9 new: 10	all: 120 new: 122
Haemonchus contortus:				
Number found	min: 35 max: 59 diff: 51%	min: 37 max: 49 diff: 28%	min: 31 max: 37 diff: 18%	min:103 max: 136 diff: 28 %
Number of clusters	35	29	14	107
Identical size[a]	all: 14 new: 14	all: 12 new: 13	all: 8 new: 9	all: 56 new: 57
Heligmosomoides bakeri:				
Number found	min: 33 max: 40 diff: 19%	min: 48 max: 50 diff: 4%	min: 32 max: 43 diff: 29%	min: 121 max: 141 diff: 15%
Number of clusters	29	31	18	128
Identical size[a]	all: 17 new: 20	all: 19 new: 24	all: 9 new: 11	all: 86 new: 100

[a] Each assembly contributed the same number of genes to the cluster. 'All' describes every assembly used in the study, including the published reference assemblies. 'New' describes those that we generated ourselves, so excludes the published reference assemblies.

Overall, we found considerable variation within gene family membership (Table 2.6 and Supplementary Tables S48-S71). We looked at the correlation between the size of the gene families and the size of the assembly or the BUSCO duplication rate. The results were inconsistent. We saw a strong and positive correlation (r>0.7) for: GSTs, UGTs, and NHRs in *C. bovis* against both assembly size and duplication rate; and NHRs in *H. contortus* against both assembly size and duplication rate. However, we saw a strong and negative correlation (r<-0.7) for CYPs, GSTs, NHRs in *H. bakeri* against the duplication rate (Supplementary Table S47). The other comparisons could not be considered noteworthy ($|r|$<0.7). Across the 628 clusters that we generated, approximately 60% contained the same number of members for each assembly. Put another way, ~40% of clusters contained a variable number of proteins and often this variation

was an additional member in just one assembly (Tables 2.6 and Supplementary Tables S48-S71).

We provide two examples of variable cluster memberships from the CYPs. The first is cluster bovis-cyp-0000003 (Supplementary Tables S48 and S49), in which: *C. elegans* and the Redbean2.5 assembly have one member each; the Stevens, Canu, SMARTdenovo and Flye assemblies have two members each; the Falcon and Falcon-unzip assemblies have three members each. The *C. elegans* protein is CYP-33C9 whose function has been associated with fatty-acid desaturation hence reducing susceptibility of xenobiotics by increasing desiccation tolerance (Erkut *et al.*, 2013). The second example is cluster contortus-cyp-0000026, in which: *C. elegans* and the Flye, Falcon, and Falcon-unzip assemblies have one member each; the Doyle, Canu, Redbean2.5, and SMARTdenovo assemblies have no members (Supplementary Tables S50 and S51). The *C. elegans* protein is CYP-43A1 and is highly expressed in the intestine and serotonergic neuron ADF (Packer *et al.*, 2019). Interestingly, an orthologue to *C. elegans* CYP-43A1 was found in an earlier version of the *H. contortus* assembly and shown to be the most highly expressed CYP in the *H. contortus* intestine (Laing *et al.*, 2015).

In *H. bakeri*, three protein families—TGF-b mimics (TGM), ARI, and BARI—have been shown to modulate the host's immune system (Johnston *et al.*, 2017; Osbourn *et al.*, 2017; Smyth *et al.*, 2018). The available protein sequences for these families aligned poorly to the *H. bakeri* Braker3 gene models (data not shown). We determined the likely locations of these proteins on the genome assemblies using two splice-aware aligners, exonerate and miniprot (Slater and Birney, 2005; Li, 2023). Similar to the other protein families, we found variation in gene counts across the three HiFi long-read assemblies: nine-to-16 TGMs and 13-to-23 ARIs/BARIs (Table 2.7 and Supplementary Table S4).

Table 2.7. Immunomodulators annotated in *Heligmosomoides bakeri* assemblies.

Assembly	Chow		Flye		HiCanu		Hifiasm	
Algorithm	**Miniprot**	**Exonerate**	**Miniprot**	**Exonerate**	**Miniprot**	**Exonerate**	**Miniprot**	**Exonerate**
TGM-1	0	0	1	1	2	1	4	2
TGM-2	1	1	2	1	1	1	1	1
TGM-3	0	0	0	0	1	1	1	1
TGM-4	1	1	1	1	1	1	1	1
TGM-5	1	1	1	1	1	1	3	1
TGM-6	1	1	1	1	1	1	1	1
TGM-7	1	1	1	1	1	2	1	2
TGM-8	1	1	2	1	1	0	2	0
TGM-9	2	2	1	1	1	1	1	0
TGM-10	1	1	1	1	1	1	1	1
Total TGMs	**9**	**9**	**11**	**9**	**11**	**10**	**16**	**10**
Overlap gene model (any)	**9**	**9**	**11**	**9**	**11**	**10**	**13**	**9**
Overlap gene model (≥50%)	**7**	**6**	**10**	**6**	**10**	**8**	**9**	**4**
ARI1	1	1	3	2	7	3	8	4
ARI2	1	1	1	2	1	2	1	2
ARI3	4	3	3	3	3	3	3	3
BARI	2	1	2	2	3	3	3	3
BARI_Hom1	2	4	3	5	7	4	7	4
BARI_Hom2	1	1	1	1	1	1	1	1
Total ARI/BARI	**11**	**11**	**13**	**15**	**22**	**16**	**23**	**17**
Overlap gene model (any)	**10**	**10**	**13**	**14**	**15**	**14**	**17**	**14**
Overlap gene model (≥50%)	**3**	**3**	**6**	**10**	**10**	**8**	**8**	**8**

TGM, transforming growth factor beta mimics; ARI, alarmin release inhibitor; BARI, 'binds alarmin receptor and inhibits' proteins.

We checked whether the protein-to-genome alignments overlapped with Braker3 gene models and found that while most alignments (TGM 81-100%; ARI/BARI 68-100%) overlapped with a gene model, far fewer alignments (TGM 40-91%; ARI/BARI 27-67%) overlapped gene models by greater than 50% of their length (Table 2.7). Focusing on scaffold ptg000126l from the Hifiasm assembly, we saw a full set of examples, from complete congruence between Braker3, exonerate, and miniprot, to evidence from just one source (Figure 2.6).

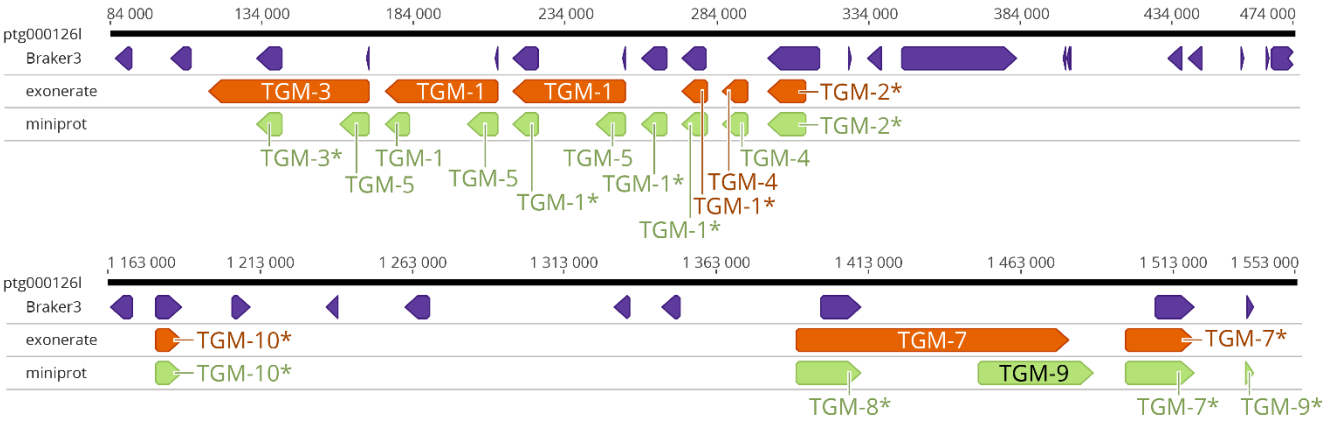

Figure 2.6. Location of TGM protein-coding genes on *Heligmosomoides bakeri* scaffold ptg000126l. TGM genes were found to cluster in two locations on this scaffold. The purple track is the Braker3 gene model predictions, the orange track is the exonerate protein-to-genome alignments, and the green track is the miniprot protein-to-genome alignments. For both exonerate and miniprot alignments, * signifies ≥50% reciprocal overlap with a Braker3 gene model. The black bar is the scaffold, with nucleotide coordinates given above the bar.

2.5 Discussion

Improvements in sequencing read length and accuracy is fuelling the demand for more contiguous genomes and has led to the development of many novel computational strategies. The technologies are rapidly evolving, and we have not attempted to be comprehensive; rather, we selected genome assemblers based on their use at the start of our study. Benchmarking in other organisms has shown that there remains value in comparative analyses, especially given the remarkable genetic diversity encountered across the genomes sequenced to date (Jung *et al.*, 2020; Koo *et al.*, 2023; Pollo *et al.*, 2020; Sun *et al.*, 2021). Our objectives here have been to determine the relative efficacy of different assembly programs, particularly their effect on gene identification within nematodes species. These species cover a wide range of genomic characteristics, which should make our findings of interest to those sequencing and assembling their own favourite, non-nematode, organisms.

Long-read assemblers predominantly use two core algorithmic frameworks: Overlap-Layout-Consensus (OLC) and De Bruijn Graph (DBG) (reviewed in (Li *et al.*, 2012)). Each framework is adapted with a myriad of developer adaptations to address specific genome features or challenges, such as repetitive sequences, varying levels of ploidy, and differential sequencing

error profiles. The process of implementing and optimising these adaptations is not trivial; they significantly affect the assemblers' effectiveness, especially in regions exhibiting intra-population structural complexities or high heterogeneity.

In this study, we have used genomic sequence data from different long-read sequence platforms to generate multiple draft assemblies for three nematode species with varying genome sizes, repeat contents, and levels of heterozygosity. The *C. bovis* genome—sequenced on the ONT MinION platform—has the smallest assembly of the three (~63Mb) and exhibits low heterozygosity and repeat content (Stevens *et al.*, 2020). The assemblies for *H. contortus*—sequenced with PacBio RSII—and *H. bakeri*—sequenced with PacBio HiFi—were larger (~280Mb and ~680Mb, respectively) and had higher heterozygosity and repeat content (Chow *et al.*, 2019; Doyle *et al.*, 2020; Stevens *et al.*, 2023). For *C. bovis*, the DBG-based assemblers—Redbean2.5 and Flye—performed best, yielding less fragmented assemblies and higher BUSCO scores. They also did better with the more granular measures of assembly performance: read mapping rates, alignment depths, and structural errors (large and small). Determining the best assemblies for *H. contortus* was less obvious. Here, OLC-based assemblers—SMARTdenovo and Canu—achieved larger N50s and marginally better BUSCO scores. The Redbean2.5 (DBG) assembly had the lowest BUSCO duplication rate and fewest structural errors. For *H. bakeri*, the Hifiasm assembler, produced the most contiguous assembly but according to the Inspector report displayed the highest number of structural errors. This suggests potential challenges in handling highly heterozygous regions by the assembly algorithm, despite its default haplotype-resolving capabilities.

From our reading of the literature, the BUSCO score may be the most widely used metric to determine assembly accuracy (Conte *et al.*, 2017; Cornet *et al.*, 2024; Park *et al.*, 2023; Timilsena *et al.*, 2022), and we also relied heavily on it in the presented analysis. Often, we found that the scores were largely similar between assemblers, suggesting that the gene complements of each genome would be similar. However, diving deeper into the classification of the 3131 nematode single copy orthologous genes demonstrated a surprising situation, where a gene may be found as a single copy in half the assemblies and missing in another half. We saw differences between Redbean versions 2.3 and 2.5 which should motivate researchers to carefully consider updated annotations; that is, to ask, what gets lost when you improve your annotation? We found at least 27 genes which were potentially incorrectly absent in the published *H. contortus* assembly. This is one of the most carefully curated parasite genome

assemblies and demonstrated the need to fund ongoing curation for important species (Doyle, 2022; Doyle *et al.*, 2020).

Small variations in BUSCO scores can be considered trivial. For example, 1% change in nematodes is 31 genes. Does this matter? It is important to remember that BUSCO genes are the most conserved genes in an organism. They typically perform critical functions and their absence from an assembly, if true *in vivo*, would point to a significant change in an organism's biology. However, if the absence is due to a technical problem in correctly assembling the reads, one might assume that assembly of regions containing less well conserved single copy genes would be more error-prone, particularly for multi-copy gene families. This is what we find with the CYP, GST, UGT, and NHR gene families. Comparative analyses often involve species assembled and annotated using different methods across different laboratories (International Helminth Genomes Consortium, 2019). This annotation heterogeneity inflates the number of lineage-specific genes (Weisman *et al.*, 2022). An alternative is to reannotate the assemblies with a uniform approach, as we did here. A limitation of this approach in our study, is that we then ignore expert curation. For example, our reannotation of the Doyle *H. contortus* assembly resulted in approximately 4,800 less genes than the published annotation, and a lower BUSCO score. In contrast, our reannotation of the published *C. bovis* and *H. bakeri* assemblies resulted in more genes, and importantly, improved BUSCO scores. Together, this demonstrates the requirement to validate the absence of a gene, which may actually be present in the assembly but score just the wrong side of a gene finder's score cut-off (Baker *et al.*, 2021; Gilabert *et al.*, 2016; Stroehlein *et al.*, 2018).

The TGM proteins were discovered through a proteomic survey of *H. bakeri* secreted products and later mapped to the Chow assembly, where only two of the ten proteins could be confidently assigned to gene models and another two proteins mapped directly to the genome (Johnston *et al.*, 2017; Smyth *et al.*, 2018). Similarly, the alarmin release inhibitors (ARI and BARI) evaded direct mapping to the published Chow assembly gene models. We find that even with improved assemblies, these proteins could not be reliably reconstructed with arguably the most popular gene finding software. However, targeted searches did provide strong evidence for genes for all proteins, some with multiple gene copies. As the variation across the assemblies shows, a more careful curation is required.

Our study demonstrates that the accuracy and completeness of assembled genomes, and consequently gene predictions, are influenced by both the choice of algorithm and the genomic

characteristics of the species being studied. Assemblers based on the Overlap-Layout-Consensus (OLC) approach were found to be more suitable for assembling complex genomes, while De Bruijn Graph (DBG)-based assemblers performed best with simpler genomes. However, the DBG-based Redbean assembler performed comparably well across both small and large nematode genomes, suggesting it is a versatile option to consider when selecting an assembly method.

Ultimately, determining the usefulness of an assembly will remain a personal choice; one that must be guided by a range of metrics and the specific research goals. For example, in population genetics studies aiming to identify regions under selection, contiguity and genome organization, such as gene order, are critical; therefore, assemblers that excel in producing highly contiguous assemblies are preferable. Conversely, studies focused on investigating specific gene families require assembly methods that provide superior genome accuracy and completeness. It is noteworthy that in a comparison of assemblies generated for several invertebrates, the assemblies with the highest BUSCO score were sometimes the most fragmented (Sutton *et al.*, 2021).

Therefore, we encourage those charged with assembling genomes—be it *Heligmosomoides*, *Haemaphysalis*, or *Henneguya*—to be aware of the important differences between seemingly very similar assemblies. For users of genome assemblies and annotation, we emphasize the importance of exploring alternative assemblies generated by multiple assemblers to validate biological hypotheses. Finally, we note that much of the above discussion can be summed up neatly by the first recommendation from Assemblathon2: Don’t trust the results of a single assembly (Bradnam *et al.*, 2013).

2.6 Acknowledgements

We thank Drs. Aralia Leon Coria and Constance Finney (University of Calgary) and the team at the University of Delaware DNA Sequencing & Genotyping Center for their hard work to generate sequence data for *Heligmosomoides bakeri*. We acknowledge the high-performance computing resources made available by the Faculty of Veterinary Medicine and Research Computing at the University of Calgary. We thank Drs. Stephen Doyle, John Gilleard, Graham Plastow, John Soghigian, and Frank van der Meer for robust discussions on this work. This work was supported by Results Driven Agricultural Research (2022T035R) and the Natural Sciences and Engineering Research Council of Canada (04589-2020).

2.7 Supplementary Material

Supplementary Material can be found in the attached Excel and Word files entitled “Chapter 2 Supplementary Tables.xlsx” and “Chapter 2 Supplementary Figures.docx”. Supplementary Figures are referred to in text as Figure S 2.X (and as Supplementary Fig. SX in the separate Word document), while Supplementary Tables are referred to as Table SXX in the separate Excel file.

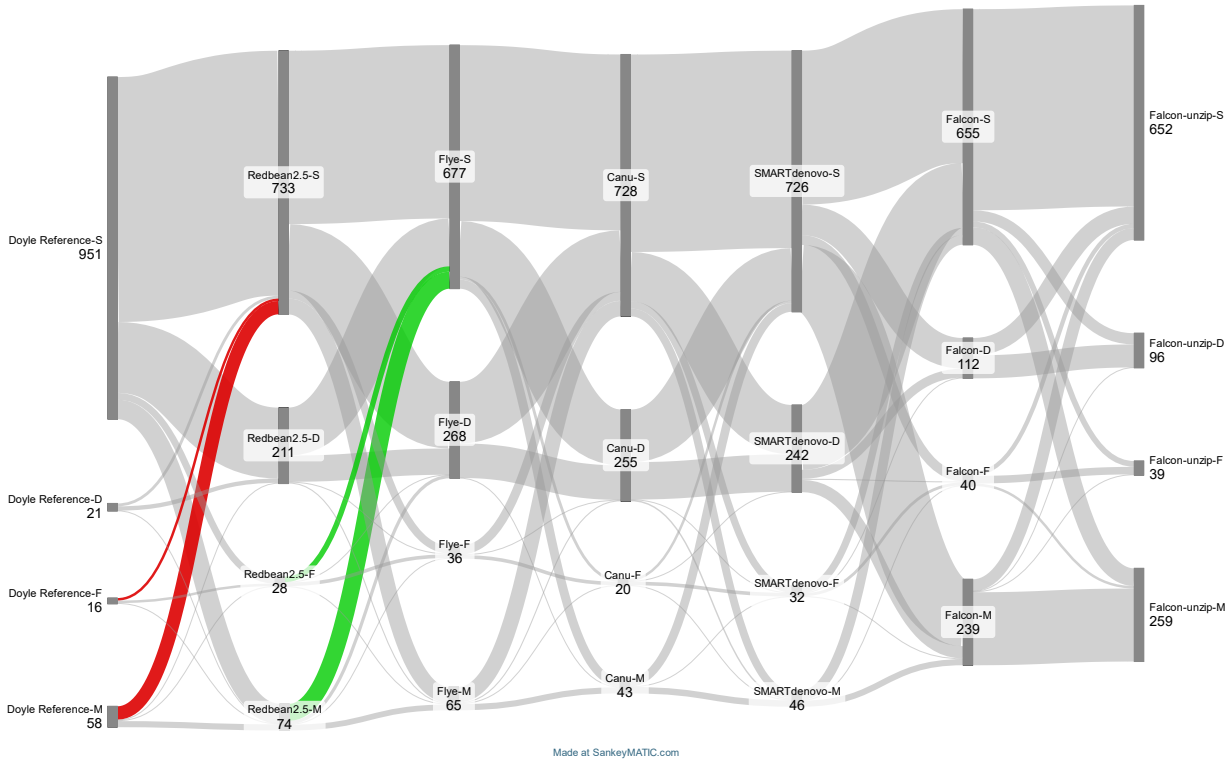

Figure S 2.1. Transitions of BUSCO genes across multiple *Haemonchus contortus* assemblies. A Sankey diagram of the 1046 genes that changed categories in at least one *H. contortus* assembly. In red are 37 BUSCO genes missing from the Doyle assembly but found complete in the Redbean2.5 assembly, and seven BUSCO genes found fragmented in the Doyle assembly but complete in the Redbean2.5 assembly. In green are 47 BUSCO genes fragmented in the Doyle assembly but found complete in the Redbean2.5 assembly, and 15 BUSCO genes fragmented in the Redbean2.5 assembly but found complete in the Flye assembly. The data for this panel is in Supplementary Table S24.

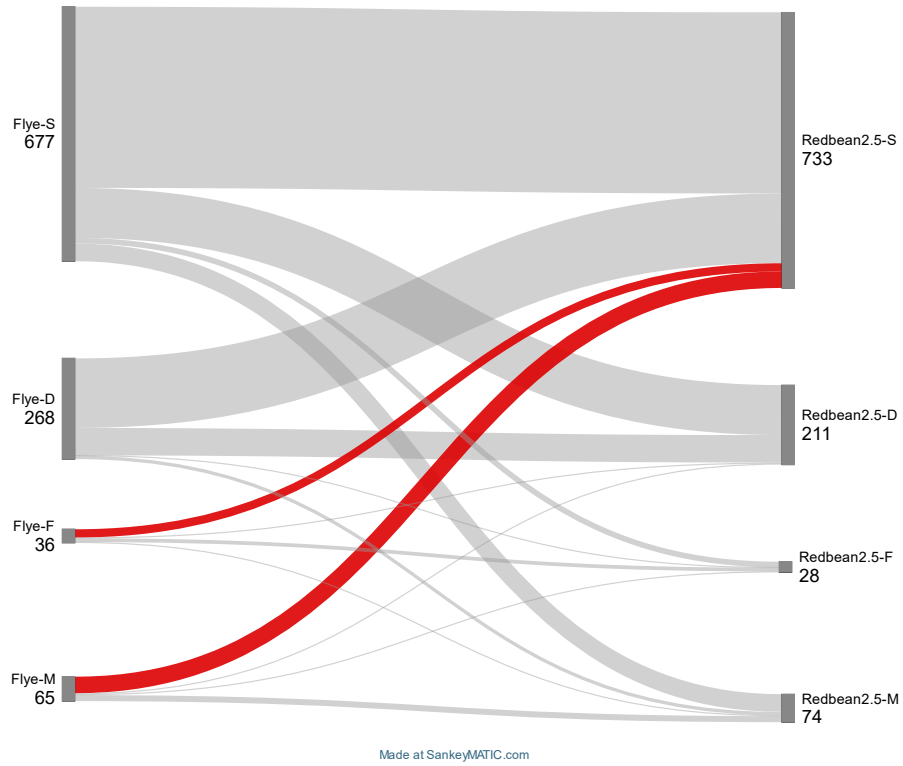

Figure S 2.2. Transitions of BUSCO genes between Flye and Redbean2.5 assemblies *in Haemonchus contortus*. A Sankey diagram of the 1046 genes that changed categories between Flye and Redbean2.5 assemblies in *H. contortus*. In red are 44 BUSCO genes missing in the Flye assembly but found complete in the Redbean2.5 assembly, and 22 BUSCO genes fragmented in the Flye assembly but found complete in the Redbean2.5 assembly. The data for this panel is in Supplementary Table S25.

Chapter 3

Exploring technical variation in the scaffolding of the *Heligmosomoides* genome

3.1 Abstract

Genome scaffolding tools facilitate joining of assembly contigs into scaffolds. Here, I investigate technical variation across Hi-C-scaffolded *Heligmosomoides* assemblies by analyzing assembly statistics and assessing potential biases introduced by two scaffolding tools. Results showed that the choice of a scaffolding tool was not a confounding bias; the technical bias introduced by the scaffolding software was small. This suggests that employing different scaffolding software does not introduce large-scale biases that could impede comparative genomic studies of species scaffolded by different methods.

3.2 Introduction

The typical ambition of a genome sequencing project is to generate a chromosome-level assembly. The continued advances in sequencing technologies are making this more achievable by providing longer and more accurate sequence reads. These reads must be clustered into contigs during assembly, and then correctly oriented and ordered. This process, known as scaffolding, connects assembly contigs separated by gaps or repetitive sequences into larger segments that could represent chromosome fragments or entire chromosomes.

There are different approaches used to scaffold genome assemblies. The earliest method involved genome mapping methods, developed in response to limitations in the conventional shotgun DNA sequencing approach. In this method, long DNA fragments were sheared into smaller fragments and sequence overlaps were used to computationally reconstruct the original DNA sequence (Staden, 1979; Anderson, 1981). While this approach was effective for the smaller prokaryotic genomes, larger eukaryotic genomes posed challenges due to the increased number of sequence fragments and the difficulty of handling repetitive regions during assembly (Blattner *et al.*, 1997; Tettelin *et al.*, 2001; Staden, 1979; Anderson, 1981). Genome maps helped overcome these challenges by outlining the positions and distances of unique genetic or physical features in a genome (National Research Council, 1988).

Two genome mapping methods—genetic and physical mapping—became essential for the earliest assemblies of animal genomes, *Caenorhabditis elegans* and *Drosophila melanogaster* (The *C. elegans* Sequencing Consortium, 1998; Adams *et al.*, 2000). Initially, genetic markers were genes for visible traits, but were superseded by non-gene DNA markers: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and

single nucleotide polymorphisms (SNPs) (reviewed in (Brown, 2002)). RFLPs are DNA sequence variations between individuals at sites recognized by restriction enzymes resulting in length polymorphisms (Lander and Botstein, 1989). SSLPs are repetitive sequences of varying base lengths. They include minisatellites (with repeat units up to 25 bp) and microsatellites (typically di-or tetranucleotide repeats) (Marwal and Gaur, 2020; Rosenberg *et al.*, 2002; Jorde, 2002; Brown, 2002). Lastly, SNPs are DNA variations that differ at a single base and are the most common type of genetic variation (Sherry *et al.*, 1999).

The presence and unique positions of molecular markers on a chromosome or contig (loci) provide information about gene organization and variants in genomes. Genetic maps are created by association of these markers through genetic linkage: markers that are in close proximity on the same chromosome tend to be coinherited (Elston, 1995; Pulst, 1999; Brown, 2002). Thus, inheritance patterns of these alleles' loci are used to determine their genomic order and positions along chromosomes (Heras *et al.*, 2016; Brown, 2002). To improve genome assemblies, genetic maps (which are identified through population mapping studies and stored in databases) have been used to assign contigs or scaffolds to linkage groups (chromosomes), by aligning the primer sequences of genetic markers to an assembly (Figure 3.1). This is followed by gap identification, scaffold ordering and orientation based on the marker locations, a process referred to as anchoring (Mascher and Stein, 2014). Genetic markers can also be aligned to distinctively numbered chromosome bands creating cytogenetic maps, which can also be mapped to DNA sequences, thus guiding scaffolding (Figure 3.1). Although genetic maps improved genome assemblies, their construction was time-consuming, expensive, and impractical for many species, especially for species with long generation times or complicated life cycles (*e.g.* parasites). Additionally, recombination events, on which genetic mapping relies, are not evenly distributed across chromosomes, leading to discrepancies between distances on genetic and physical maps (Sturtevant, 1913; Jensen-Seaman *et al.*, 2004; Backström *et al.*, 2010; Lichten and Goldman, 1995; Brown, 2002).

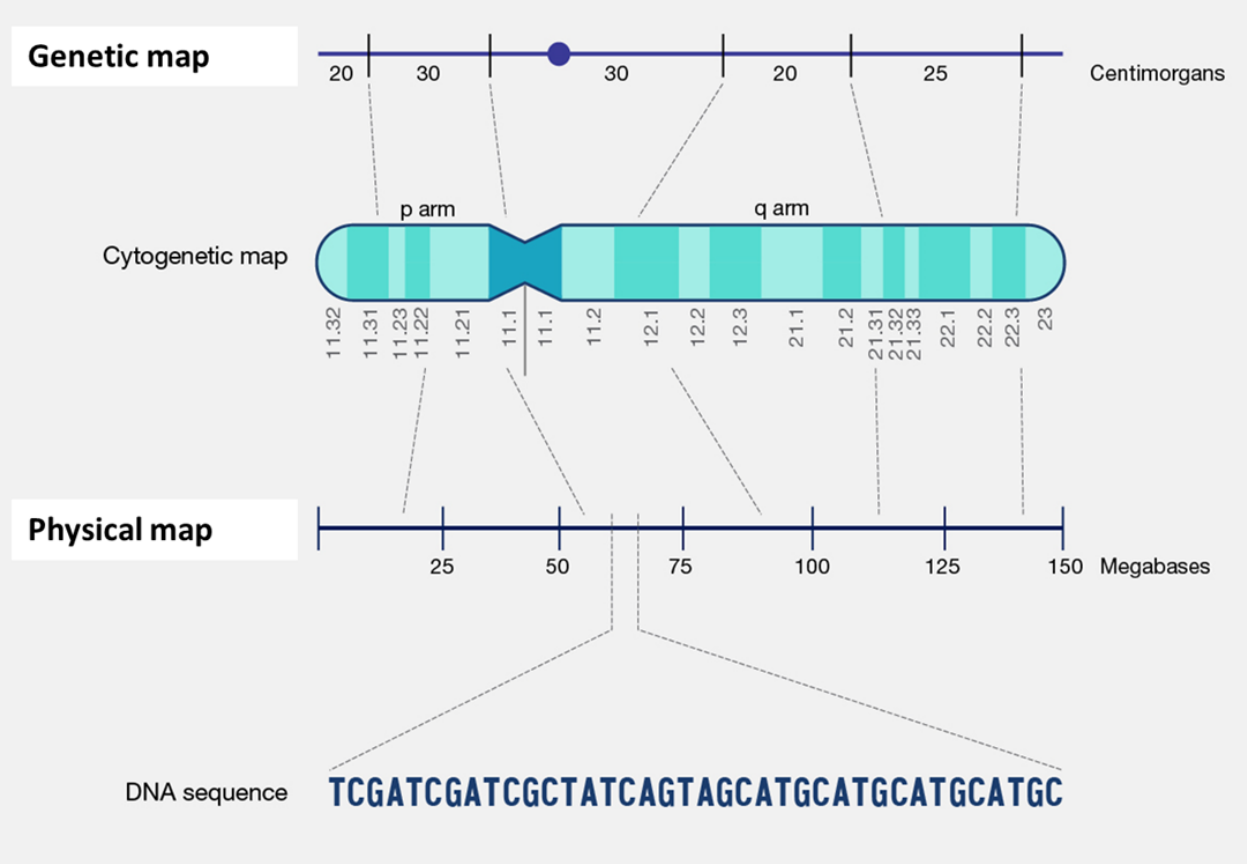

Figure 3.1. Genome mapping. The figure is downloaded from https://www.genome.gov/genetics-glossary/Mapping; (Mapping, 2023). The distances between genes in genetic maps are measured in centimorgan while physical map distances are measures in base pairs.

To address these limitations, physical mapping techniques were developed. While genetic mapping identifies gene positions indirectly by analyzing genetic markers, physical mapping directly analyzes gene locations and their characteristics on chromosomes through techniques such as restriction mapping, fluorescence in situ hybridization (FISH), and sequence tagged site (STS) mapping. These techniques use established maps based on physical distances between markers, often indicating how much DNA (measured in base pairs) separates genes (Figure 3.1). Restriction mapping is a technique used to identify positions of restriction enzyme recognition sites in a DNA sequence (Peck *et al.*, 1997). One common approach used to increase the genetic marker density on a genome map is to construct a restriction map by sequencing the entire genome and computationally identifying restriction sites (identified

experimentally and stored in a database) for all known restriction enzymes using specific restriction mapper programs (Bishop, 1984; Raschke, 1993). STS mapping, on the other hand, involves alignment of known short DNA sequences (100-500 bp) to genomes. These sequences are found in a unique location within a chromosome or genome and have distinctive and conserved flanking regions making them distinguishable from other DNA sequences thus facilitating their identification in genomes (Olson *et al.*, 1989). Sources of STS include SSLPs (like microsatellites), random genomic sequences and expressed sequence tags (ESTs) (Brown, 2002). ESTs are complementary DNA sequences (cDNA) produced by sequencing of cDNA clones that can be used to identify expression of genes at a particular time (ESTs Factsheet, 2002; Marra *et al.*, 1998). The FISH physical mapping method uses fluorescently labeled probes to determine the locations of known DNA markers along chromosomes (Espinosa and Le Beau, 2000; Heiskanen *et al.*, 1996).

One more recent physical mapping technique is mate paired-end sequencing. DNA fragments of a known size range, often 5-30kb, are circularized and then sheared into smaller pieces. Sequencing is performed on both ends of these smaller fragments, providing information about the relative positions and orientations of DNA sequences that are separated by a known distance within the genome (Figure 3.2).

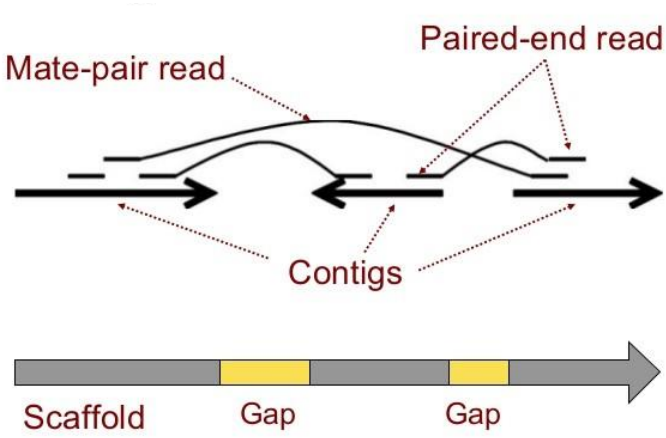

Figure 3.2. Scaffolding using mate-paired-end reads. This figure is downloaded from http://bch709.plantgenomicslab.org/fig/mate.png; CC BY 4.0 license.

Many nematode genome assemblies, including *Haemonchus contortus, Ascaris suum, Strongyloides ratti* and *Strongyloides stercoralis* benefitted from mate paired-end driven

scaffolding (Doyle *et al.*, 2018; Laing *et al.*, 2013; Wang *et al.*, 2012, 2017; Hunt *et al.*, 2016). In a similar way, RNA-seq data can be used to link exons from the same transcript which were initially assembled on the different contigs as demonstrated in the RNA-mediated scaffolding of the *Caenorhabditis* species PS1010 (Mortazavi *et al.*, 2010) and the *Heligmosomoides bakeri* assembly (Chow *et al.*, 2019).

Optical mapping is another physical mapping technique that constructs physical maps from single molecules of DNA (Zhou *et al.*, 2007). Briefly, long DNA molecules are randomly nicked and then repaired by a polymerase which introduces fluorescently labeled nucleotides. A short string of these nucleotides serves as markers along the DNA molecules and can be mapped to the assembled contigs. Scaffolding *de novo* assemblies through optical mapping gained popularity and was used for the current *Haemonchus contortus* reference assembly, but its high cost has meant its popularity in nematode research was short-lived (Doyle *et al.*, 2018, 2020; Jiao *et al.*, 2017; Yuan *et al.*, 2020). The FISH technique, described earlier, is similar to optical mapping with the difference being that in FISH, the marker is a DNA sequence visualized by hybridization with a fluorescent probe while in optical mapping, the marker is a restriction site visualized as a gap in a DNA sequence (Brown, 2002).

With the advent of long-read sequencing, new scaffolding approaches that use long reads as proxies for physical maps were developed. Two long-read assembly boosting methods—read threading and scaffolding/gap-filling programs—were introduced (Koren and Phillippy, 2015). Read threading was used to resolve short-read assemblies using long reads. Here, a De Bruijn assembly graph from sequence reads was first generated, then long reads were mapped on the graph to establish congruency and generate linking information which was subsequently used to order and orient scaffolds. Examples of long-read tools that used this approach include Allpaths-LG, SPAdes and Cerulean (Ribeiro *et al.*, 2012; Boetzer and Pirovano, 2014; Deshpande *et al.*, 2013). Exclusive scaffolding/gap-filling software were also used to improve genome assemblies by mapping gap-spanning long reads to an existing initial assembly to close gaps, order and orient contigs. Examples of such programs include ‘A Hybrid Assembler’ (AHA), SSPACE-LongRead and PBJelly (Bashir *et al.*, 2012; Boetzer and Pirovano, 2014; English *et al.*, 2012). PBJelly was used to scaffold and gap fill the earlier preliminary short-read Illumina *Heligmosomoides bakeri* contig assembly, using low coverage PacBio long reads (Chow *et al.*, 2019).

To further improve genome assembly, long-range genome scaffolding techniques, such as high-throughput chromosome conformation capture (Hi-C) have become increasingly popular and is

now a standard for large BioGenome consortia (Lieberman-Aiden *et al.*, 2009; Lewin *et al.*, 2018; Rhie *et al.*, 2021). Hi-C maps the three-dimensional organization of chromatin in the genome (Figure 3.3).

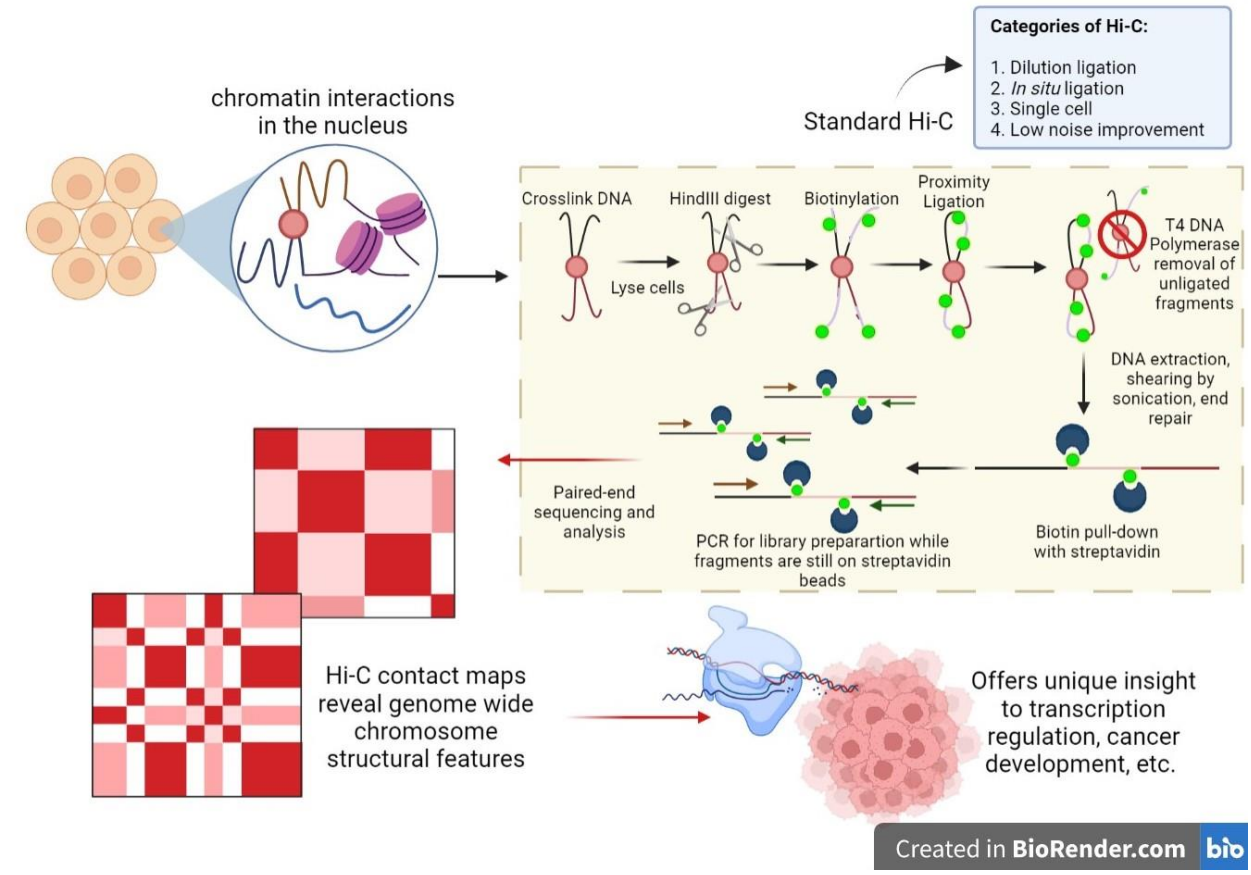

Figure 3.3. Hi-C. An overview of the Hi-C workflow and its applications in research. Figure by Prakrutiuday - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=115662977 and downloaded from https://upload.wikimedia.org/wikipedia/commons/thumb/a/a4/HiCschematic.png/1000px-HiCschematic.png

Chromatin is a DNA and protein complex which tightly packs the genome into the nucleus, and the key assumption in Hi-C sequencing is that DNA packed into individual chromatin molecules are typically proximally close in the genome. DNA physically adjacent, in three-dimensional space, is cross-linked with formaldehyde, and then the genome is cut into fragments using restriction endonucleases. A proximity ligation step preferentially attaches biotin to the cross-linked DNA, allowing the non-crosslinked DNA to be subsequently removed. This step allows

the capture of long-range interactions between different genomic regions. The remaining ligated chimeric DNA molecules are then fragmented and sequenced, generating Hi-C paired end data which are mapped back to the contig-based assembly, generating a map of the 3D contacts in the nucleus. Bioinformatic analysis of this map identifies regions between contigs with statistically significant elevated frequencies of interactions. These are termed topologically associated domains (TADs). Contigs that share TADs can then be linked into longer chromosome-scale scaffolds (Figure 3.3).

Unlike optical mapping which requires special instruments and laborious sample preparation steps, Hi-C libraries can be generated using standard short-read sequencing machines. Additionally, Hi-C employs direct laboratory protocols with several companies distributing their commercial Hi-C kits. The application of Hi-C data for scaffolding has been demonstrated in various invertebrates, e.g. *Lepidurus packardi*, an endangered Californian crustacean, and *Lytechinus pictus*, a sea urchin that is amenable to transgenic manipulation (Warner *et al.*, 2021; Kieran Blair *et al.*, 2022).

In our efforts to generate a chromosome-level assembly for the parasitic *Heligmosomoides bakeri*, we have used Hi-C approaches to improve the preliminary assembly that was presented in Chapter 2. The sequence reads for our *H. bakeri* were generated with PacBio HiFi technology. For scaffolding, we used a commercial service from Dovetail Genomics, who performed the Hi-C sequencing with their 'Omni-C' protocols, and their proprietary 'HiRise' software (Putnam *et al.*, 2016). As of December 2023, these Dovetail Genomics' protocols have been applied to scaffold genome assemblies of different species e.g., a sunburst anemone, a desert turtle, a snake, a desert plant, a tick, a camel and a plant parasitic nematode, among others (Cornwell *et al.*, 2022; Todd *et al.*, 2022; Grismer *et al.*, 2022; Anghel *et al.*, 2022; Nuss *et al.*, 2018; Elbers *et al.*, 2019; Masonbrink *et al.*, 2021).

At the same time, the genomes of UK isolates *H. bakeri* and *Heligmosomoides polygyrus* were sequenced as part of the Darwin Tree of Life (ToL) project, the UK's BioGenome initative (Stevens *et al.*, 2023; Darwin Tree of Life Project Consortium, 2022). For these, Hi-C capture and sequencing used kits from Arima Genomics and the 'YaHS' software (Zhou *et al.*, 2023). As *H. bakeri* is a foundation of research into host-parasite interactions and anthelmintic drug discovery in many laboratories, it was important to determine which assembly was likely more accurate. Further, I and others have shown that comparative genomic studies, between different species, can be negatively affected by technical assembly variation (Chapter 2). I wanted to see to what extent Hi-C scaffolding increased or decreased such variation.

It is worth noting that *Heligmosomoides* species identification in past research has been contradicting with numerous immunological and related studies reporting the use of a lab-attenuated strain of *Heligmosomoides polygyrus*. However, it has been determined that the nematodes used in these studies were *H. bakeri*—which infects the domestic mouse, *Mus musculus*—and not *H. polygyrus*, which infects the wood mouse *Apodemus sylvaticus* (Behnke, Menge, *et al.*, 2009; Behnke, Eira, *et al.*, 2009). Genetic analyses have shown sufficient divergence between these two species of *Heligmosomoides* to warrant their classification as distinct species (Cable *et al.*, 2006). This misclassification has been rectified in recent updates (Stevens *et al.*, 2023). Henceforth, the accurate species names are used in this thesis.

I compared the two Hi-C scaffolded *H. bakeri* assemblies, which I refer to as LS-Hbak and GM-Hbak, the earlier PBJelly-scaffolded *H. bakeri* assembly (FC-Hbak), and the *H. polygyrus* Hi-C scaffolded assembly (LS-Hpol). It is important to note that I am not directly comparing the performance of Dovetail and Arima kits. These assemblies were generated using different DNA isolation methods, sequencing technologies, at different times, in different locations, by different people. However, I could compare the performance of the HiRise and YaHS scaffolding tools using the unscaffolded GM-Hbak assembly with the same Omni-C-generated reads.

I found that the GM-Hbak-HiRise assembly was a slight improvement over GM-Hbak-YaHS; it was less fragmented and more complete in terms of highly conserved genes. Similarly, the GM-Hbak-HiRise assembly was a slight improvement over the LS-Hbak-YaHS assembly. In summary, my findings suggest that the choice of a Hi-C scaffolding tool does not introduce substantial technical biases and thereby allows comparative studies between Hi-C-scaffolded assemblies. Overall, across the assemblies, GM-Hbak-HiRise reported the best statistics for contiguity and completeness, making it the *H. bakeri* assembly of choice for future research in our laboratory.

3.3 Methods

3.3.1 Hi-C sequencing and scaffolding of our *Heligmosomoides bakeri* assembly (GM-Hbak)

The draft contig-level GM-Hbak assembly is described in Chapter 2, and referred to here, as GM-Hbak-Unscaff. I selected the assembly generated with Hifiasm (v0.16.1-r375) (Cheng *et al.*, 2021). Approximately 200 frozen *H. bakeri* adults, mixed male and female, were shipped to

Dovetail Genomics for Hi-C sequencing using their Omni-C protocols. The Omni-C library was prepared using the Dovetail® Omni-C® Kit according to the manufacturer's protocol. Briefly, the chromatin was fixed with disuccinimidyl glutarate (DSG) and formaldehyde in the nucleus. The cross-linked chromatin was then digested in situ with DNase I. Following digestion, the cells were lysed with SDS to extract the chromatin fragments, and the chromatin fragments were bound to Chromatin Capture Beads. Next, the chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, the crosslinks were reversed, the associated proteins were degraded, and the DNA was purified then converted into a sequencing library using Illumina-compatible adaptors. Biotin-containing fragments were isolated using streptavidin beads prior to PCR amplification. The library was sequenced to generate 7.6 million 2 x 150 bp read pairs.

The Dovetail bioinformatics analysts used their proprietary HiRise software to generate the TAD maps and scaffold the contigs into a final assembly: GM-Hbak-HiRise. To generate the alternative scaffolded assembly, GM-Hbak-YaHS, I mapped the Omni-C reads to the Hifiasm assembly using the Arima Hi-C mapping pipeline (available at https://usermanual.wiki/Document/ArimaMappingUserGuideA160156v02.1083656696). Briefly, the Hifiasm assembly was first indexed using BWA v0.7.17-r1188 (parameters set to -a bwtsw) and the Omni-C paired-end reads aligned to the assembly using BWA-MEM (Li, 2013). The paired-end reads were initially mapped independently as single-ends and were later paired in a subsequent step. The generated alignment BAM file was then marked for duplicates using the Picard v3.0.0 (available at http://broadinstitute.github.io/picard) before being used as the input for YaHS (v1.2a.2 with default settings) (Zhou *et al.*, 2023).

3.3.2 Other *Heligmosomoides* genome assemblies

The HiFi sequence reads and YaHS-scaffolded assemblies from the Darwin ToL project were downloaded from the European Nucleotide Archive (ENA): LS-Hbak-YaHS, BioProject PRJEB57615, assembly file GCA_947359475.1_nxHelBake1.1_genomic.fna, and LS-Hpol-YaHS, BioProject PRJEB57641, assembly file GCA_947396885.1_ngHelPoly1.1_genomic.fna (Stevens *et al.*, 2023). PacBio HiFi reads for the LS-Hbak-YaHS assembly were derived from a single male *H. bakeri* worm while those for the LS-Hpol-YaHS assembly were obtained from a single female *H. polygyrus* worm. Both assemblies were generated following the same protocol (Stevens *et al.*, 2023). Briefly, preliminary assemblies were generated using Hifiasm, followed by the removal of redundant haplotigs using the Purge_dups pipeline (Guan *et al.*, 2020) and contaminated contigs using the Blobtoolkit (Challis *et al.*, 2020). The final assembly was

scaffolded with YaHS and improved through manual curation using the Paired REad TEXTure Viewer (PretextView).

The earlier *H. bakeri* assembly, BioProject PRJEB15396 (Chow *et al.*, 2019), (referred to here, as FC-Hbak-PBJelly), was downloaded from WormBase Parasite (WBPS18; WS285) (Howe *et al.*, 2017) for comparison.

3.3.3 Assembly metrics

I evaluated the assemblies for contiguity using the assembly-stats software (available at https://github.com/sanger-pathogens/assembly-stats). To determine the number of highly conserved orthologues correctly assembled, I used BUSCO v5.4.6 (Simão *et al.*, 2015), with options set at: '--lineage nematoda_odb10' (n=3131 proteins), '--mode genome', '--long' and '--augustus'. I opted for the AUGUSTUS software over the then-default MetaEuk as BUSCO returned a higher proportion of proteins with AUGUSTUS. I used Inspector v.1.0.2 (Chen *et al.*, 2021) to evaluate the alignment of HiFi reads associated with the GM-Hbak assemblies (GM-Hbak-HiRise and GM-Hbak-YaHS) for identification of potential large- and small-scale errors in the assemblies. All assembly metrics are reported in Table 3.1.

3.3.4 Genome synteny

I used Nucmer (from MUMmer v.3.23) to generate assembly-to-assembly alignments for the GM-Hbak assemblies (GM-Hbak-HiRise and GM-Hbak-YaHS) (Marçais *et al.*, 2018). I considered the GM-Hbak-HiRise assembly as the reference genome given its better contiguity. Similarly, I used the dnadiff script (part of Nucmer) to report one-to-one alignment coordinates between the GM-Hbak assemblies. The options for Nucmer and dnadiff scripts were set at: '--mincluster=500' and '--maxmatch=100'.

3.3.5 Data visualization

The one-to-one coordinates from the dnadiff script were used to link scaffolds between the GM-Hbak-HiRise and GM-Hbak-YaHS assemblies for genome synteny visualization. The generated alignments were viewed with Circos v.0.69-8 (Krzywinski *et al.*, 2009). For BUSCO content analysis, I generated Sankey diagrams using the SankeyMATIC web-based tool (available at http://sankeymatic.com/build/) to visualize the transitions of BUSCO genes—genes that changed categories between assemblies. Set-based BUSCO genes data was displayed using UpSet plots (Lex *et al.*, 2014).

3.4 Results and Discussion

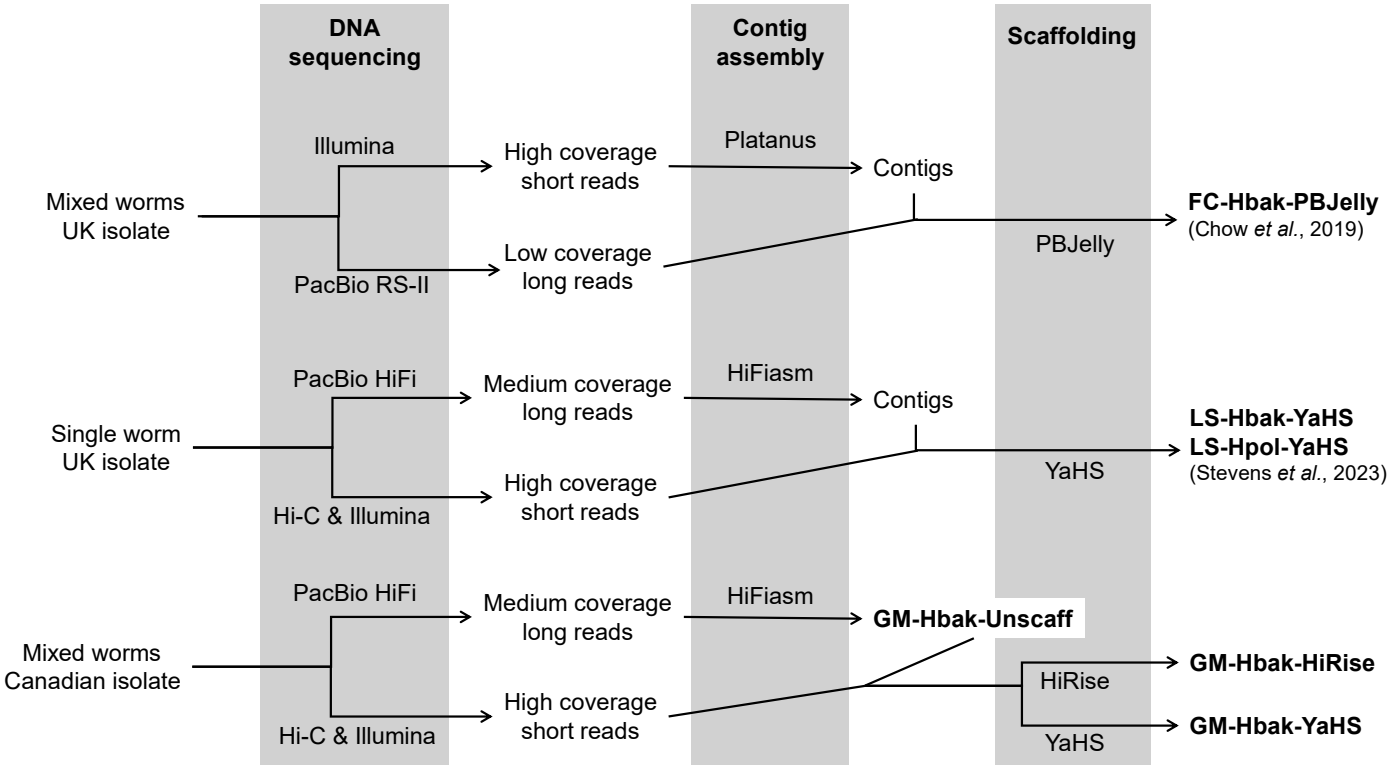

Figure 3.4. *Heligmosomoides* genomes. I compared six genome assemblies—five from *H. bakeri* and one from *H. polygyrus*—that made use of different sequencing technologies and scaffolding approaches.

3.4.1 Scaffolded assemblies were more contiguous but not necessarily more complete

Across all assemblies, it was no surprise that the FC-HBak-PBJelly assembly was the most fragmented—highest number of pieces—and had the highest number of intra-scaffold gaps (Table 3.1). The explanation for this could lie in several places. One, the assembly relies on the correct assembly of short Illumina reads into contigs, which for a large, highly heterozygous genome is not trivial. It is unlikely that poorly assembled contigs, particularly those from repetitive regions, would be mapped to the long PacBio reads and would therefore not be

scaffolded together. Two, the PacBio read coverage itself was relatively low (~12-fold coverage) (Chow *et al.*, 2019), limiting their usefulness in scaffolding.

Table 3.1. Genome assembly statistics for the *Heligmosomoides* assemblies.

Assembly Stats	**FC-Hbak-PBJelly**	**GM-Hbak-Unscaff**	**GM-Hbak-HiRise**	**GM-Hbak-YaHS**	**LS-Hbak-YaHS**	**LS-Hpol-YaHS**
# Fragments (n)	23,647	284	25	46	321	1,278
Length (Mbp)	697.0	678.4	678.4	678.5	649.2	656.6
Average (mean, Mbp)	0.03	2.4	27.1	14.7	1.5	0.5
Largest (Mbp)	1.2	20.5	119.9	119.0	114.8	111.6
N50 (Mbp)	0.2	4.9	116.6	115.1	110.8	107.4
N90 (Mbp)	0.03 n=4429	1.2 n=144	100.3 n=6	98.8 n=6	92.0 n=6	97.4 n=6
N_count	4,503,213	161	26,561	52,961	679,414	691,617
Number of Gaps	39,371	7	271	271	3,398	3,460
BUSCO (%, n=3131, mode=genome)	C:94.3 [S:93.1,D:1.2], F:1.8, M:3.9	C:94.9 [S:92.0,D:2.9], F:1.7, M:3.4	C:94.7 [S:93.1,D:1.6], F:1.4, M:3.9	C:94.5 [S:92.5,D:2.0], F:1.6, M:3.9	C:93.8 [S:92.6,D:1.2], F:0.7, M:5.5	C:95.1 [S:92.7,D:2.4], F:0.9, M:4.0

BUSCO genes' categories: Complete 'C', is the sum of single copy 'S' and duplicated 'D' genes; 'F' denotes Fragmented genes, and 'M' denotes Missing genes (all % from 3131 genes). N50 is the number of bases in (n) fragments that make up 50% of the assembly, N90 is the number of bases in (n) fragments that make up 90% of the assembly, N_count is the number of undetermined nucleotide bases (Ns) across the assembly and gaps are any repeated number of Ns of any length.

The Hi-C-scaffolded assemblies were all essentially at a chromosome-level (≥90% of the assembly being on six scaffolds) with a small number of unplaced contigs (Table 3.1). The GM-Hbak-HiRise and GM-Hbak-YaHS assemblies were less fragmented than the LS-Hbak-YaHS assembly. The DNA library for the latter had been isolated and generated from a single worm in an effort to reduce the effect of heterozygosity in the assembly process (Stevens *et al.*, 2023). Typically, high heterozygosity leads to more fragmented assemblies as multiple divergent alleles cannot be easily collapsed and may be misclassified as independent loci in the assembly. However, a whole genome amplification step was required to obtain enough DNA to sequence

from a single worm (Stevens *et al.*, 2023), and this PCR-based process may lead to uneven amplification and confound the assembler. The LS-Hpol-YaHS assembly was the most fragmented among the Hi-C scaffolded assemblies. This could be attributed to the increased heterozygosity expected in the wild strain *H. polygyrus* than in the laboratory strain *H. bakeri*. Adult *H. polygyrus* worms used for sequencing were removed directly from the gastrointestinal tracts and small intestines of trapped wild *Apodemus sylvaticus* wood mice while for the *H. bakeri,* infective larvae were passaged into laboratory mice (*Mus musculus*) and adult worms harvested from the small intestines (Stevens *et al.*, 2023).

The completeness of the assemblies was measured with BUSCO genes; genes which have orthologs across a taxonomic lineage and are expected to be found once and only once in the assembly. Here, I measured the completeness of the 3131 genes considered to be orthologous across the Nematoda lineage. In considering just the proportion of BUSCO genes identified, I found similar scores across all the assemblies. The most complete assembly was GM-Hbak-Unscaff (C: 94.9%) and I note that the scaffolding process, whether by HiRise or YaHS led to the loss of eight and 11 genes respectively (Table 3.1). It was striking that the FC-Hbak-PBJelly assembly, despite being the most fragmented by a considerable margin, contained a higher proportion of BUSCO genes than LS-Hbak-YaHS: 94.3% vs 93.8%. While this represents 16 genes, it demonstrates the value of the earlier assembly and that the short reads from coding regions of the genome were probably well assembled.

3.4.2 The HiRise scaffolder slightly outperformed YaHS

I compared the GM-Hbak-HiRise and GM-Hbak-YaHS assemblies. I mapped the PacBio HiFi long reads to each assembly and the Inspector software identified potential errors. There was little to distinguish the two assemblies, with both assemblies having the same total number of potential structural errors, though individual categories were different (Table 3.2).

Table 3.2. Assembly evaluations of the two differently scaffolded GM-Hbak assemblies.

Read to Contig alignment		
Assembly	**GM-Hbak-HiRise**	**GM-Hbak-YaHS**
Mapping rate /%	99.82	99.82
Split-read rate /%	9.04	9.05
Alignment depth	25.6	25.6
Mapping rate in large contigs /%	99.05	98.58
Split-read rate in large contigs /%	9.08	9.08
Errors		
Total structural errors	319	319
Expansion	223	216
Collapse	68	67
Haplotype switch	26	32
Inversion	2	4
Small-scale assembly error /per Mbp	109	110
Total small-scale assembly error	74,089	74,818
Base substitution	62,196	62,928
Small-scale expansion	6,254	6,234
Small-scale collapse	5,639	5,656
QV (Quality Value)	31.7	31.7

The GM-Hbak-YaHS assembly had more haplotype switches, inversions, and small-scale errors, which may explain the slightly worse BUSCO score compared with GM-Hbak-HiRise (Table 3.1). It was striking that a union of the two assemblies contained 3000 of the 3131 BUSCO genes (for a completeness of 95.8%) (Figure 3.5C). This is higher than the completeness of the GM-Hbak-Unscaff assembly, which indicates that not only are BUSCO genes lost in the scaffolding process, but they are also restored, likely by bringing together contigs over which genes were previously split.

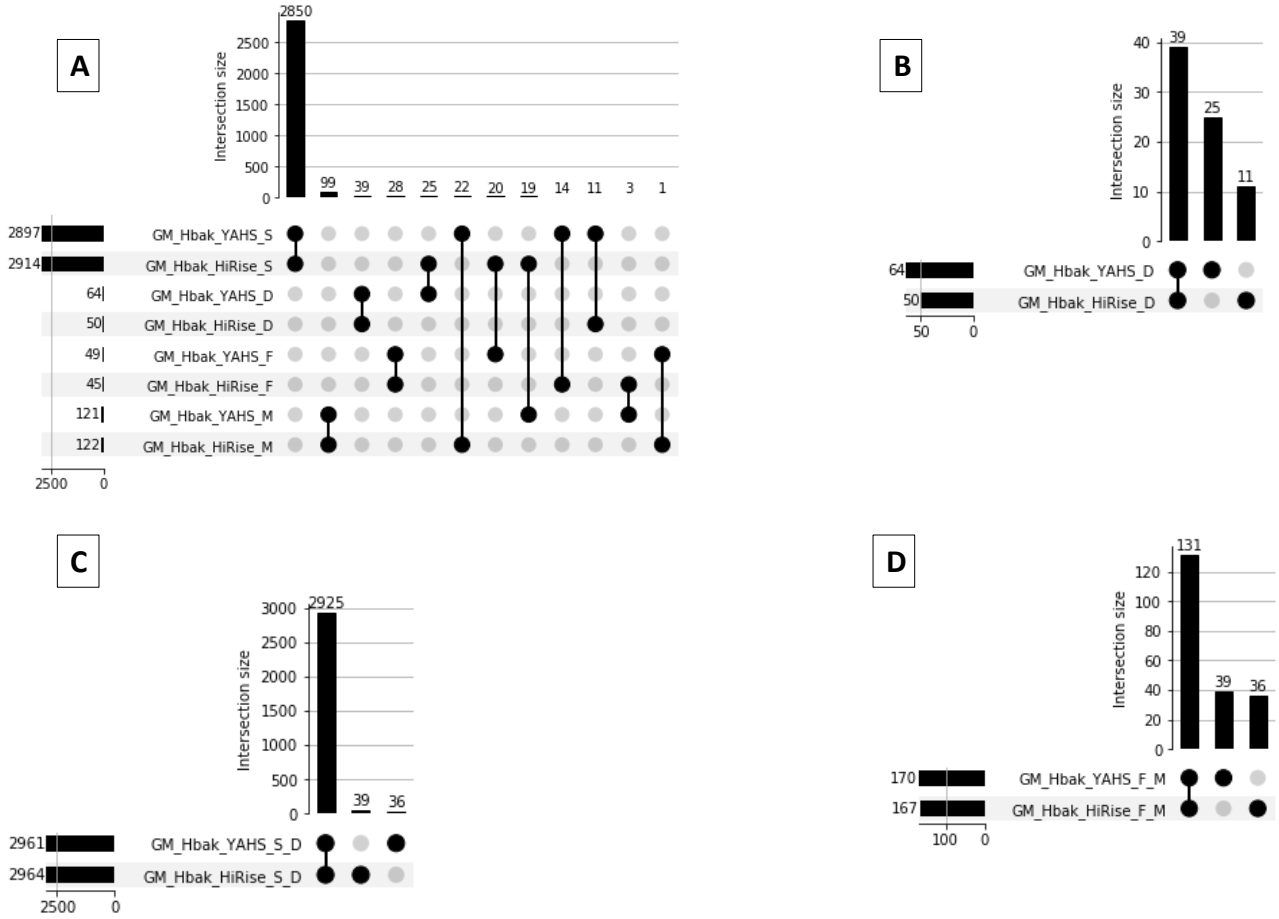

Figure 3.5. UpSet plots of the two differently scaffolded GM-Hbak assemblies (GM-Hbak-HiRise and GM-Hbak-YaHS). A) Upset plot of all BUSCO gene categories: complete and single copy (S), complete and duplicated (D), fragmented (F), missing (M),
B) Upset plot of only complete and duplicated BUSCO genes (D),
C) Upset plot of complete, single copy and duplicated BUSCO genes (S_D),
D) Upset plot of fragmented and missing BUSCO genes (F_M).

3.4.3 Choice of scaffolding software does not increase bias to comparative genomic analysis

Inferring genomic and genic differences between species can be adversely affected by the datasets being generated by different software (Florea *et al.*, 2011; Weisman *et al.*, 2022; Jung *et al.*, 2020; Pollo *et al.*, 2020; Sun *et al.*, 2021; Bradnam *et al.*, 2013). I wanted to know to what extent this may extend to scaffolding software, YaHS and HiRise (Zhou *et al.*, 2023; Putnam *et*

al., 2016). To do this, I carried out analysis using the nematode BUSCO genes to compare orthologs between GM-Hbak-HiRise, GM-Hbak-YaHS, and LS-Hbak-YaHS. If the choice of the scaffolding software was a major influence, then there would be more shared orthologs between GM-Hbak-YaHS and LS-Hbak-YaHS to the exclusion of GM-Hbak-HiRise. However, this is not what was observed (Figure 3.6).

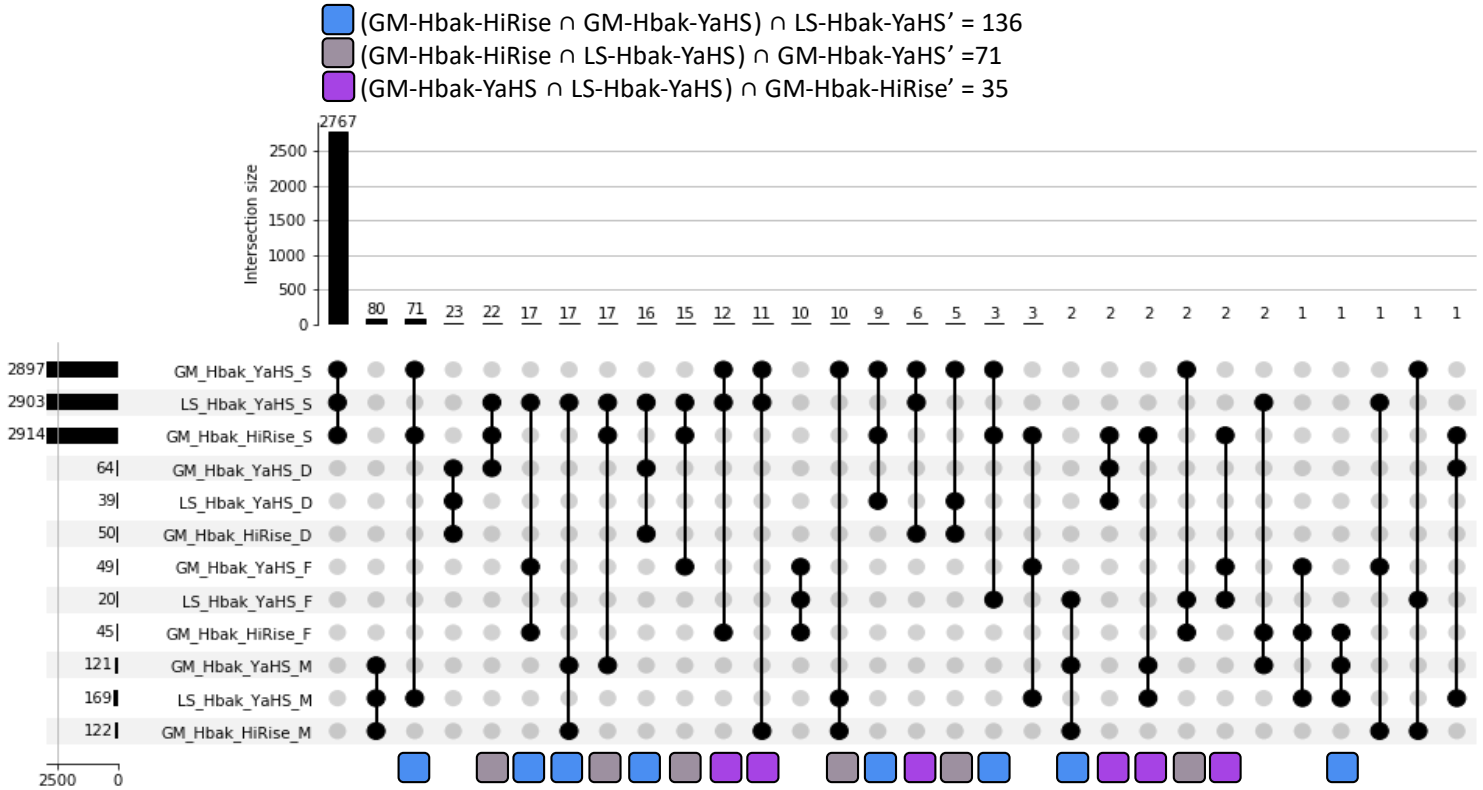

Figure 3.6. An UpSet plot of the three *Heligmosomoides bakeri* scaffolded assemblies, GM-Hbak-HiRise, GM-Hbak-YaHS and LS-Hbak-YaHS. The blue, brown and purple boxes are BUSCO categories shared between pairs of *H. bakeri* HiRise vs YaHS scaffolded assemblies to the exclusion of the third. (GM-Hbak-HiRise ∩ GM-Hbak-YaHS) ∩ LS-Hbak-YaHS' denotes shared BUSCO categories between GM-Hbak-HiRise and GM-Hbak-YaHS to the exclusion of LS-Hbak-YaHS, e.g., S-S-D.

In fact, there were twice as many orthologs between GM-Hbak-HiRise and LS-Hbak-YaHS to the exclusion of GM-Hbak-YaHS (71 vs 35). As HiRise is a proprietary software run by the Dovetail Genomics analysts, details of its algorithm are not known. However, for scaffolding Hi-C data, there are limited variations on the theme, and I suspect the primary difference with YaHS would be arbitrary cut-offs that determine whether two contigs are joined or not. This is supported by the assembly-to-assembly alignment (Figure 3.7); most differences are short contigs that incorporated into the six chromosomes of the HiRise assembly (coloured scaffolds) but remain unplaced in the YaHS assembly (gray scaffolds).

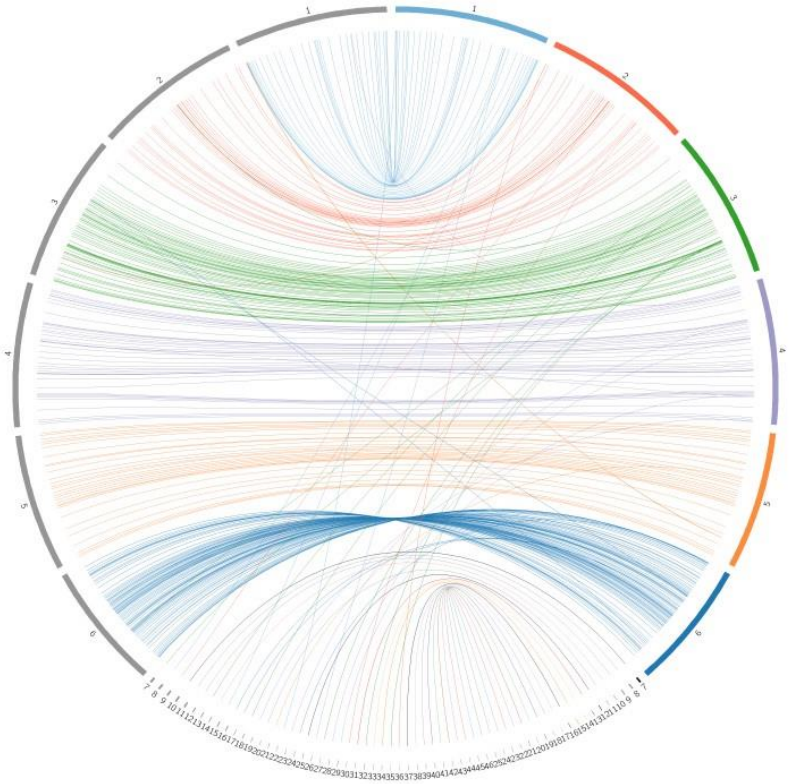

Figure 3.7. A Circos plot showing synteny between two differently scaffolded GM-Hbak assemblies. The colored scaffolds represent the reference genome (GM-Hbak-HiRise), and the gray chromosomes represent the GM-Hbak-YaHS genome as the query.

3.4.4 The GM-Hbak-HiRise assembly is the better choice for future *Heligmosomoides bakeri* analysis but is still missing genes

Having accepted the GM-Hbak-HiRise assembly as the best assembly for our Canadian isolate, I compared it with LS-Hbak-YaHS and LS-Hpol-YaHS. In terms of BUSCO completeness, my GM-Hbak-HiRise assembly had 50 fewer missing BUSCO genes than LS-Hbak-YaHS (Table 3.1). In Chapter 2, I have shown that small changes in the highly conserved BUSCO gene set can scale to considerably higher variation for less well conserved genes, including those of important large gene families. In the three-assembly comparison, 2715 BUSCO genes were found to be complete and single copy in all three, and 72 genes missing in all three (Figure 3.8).

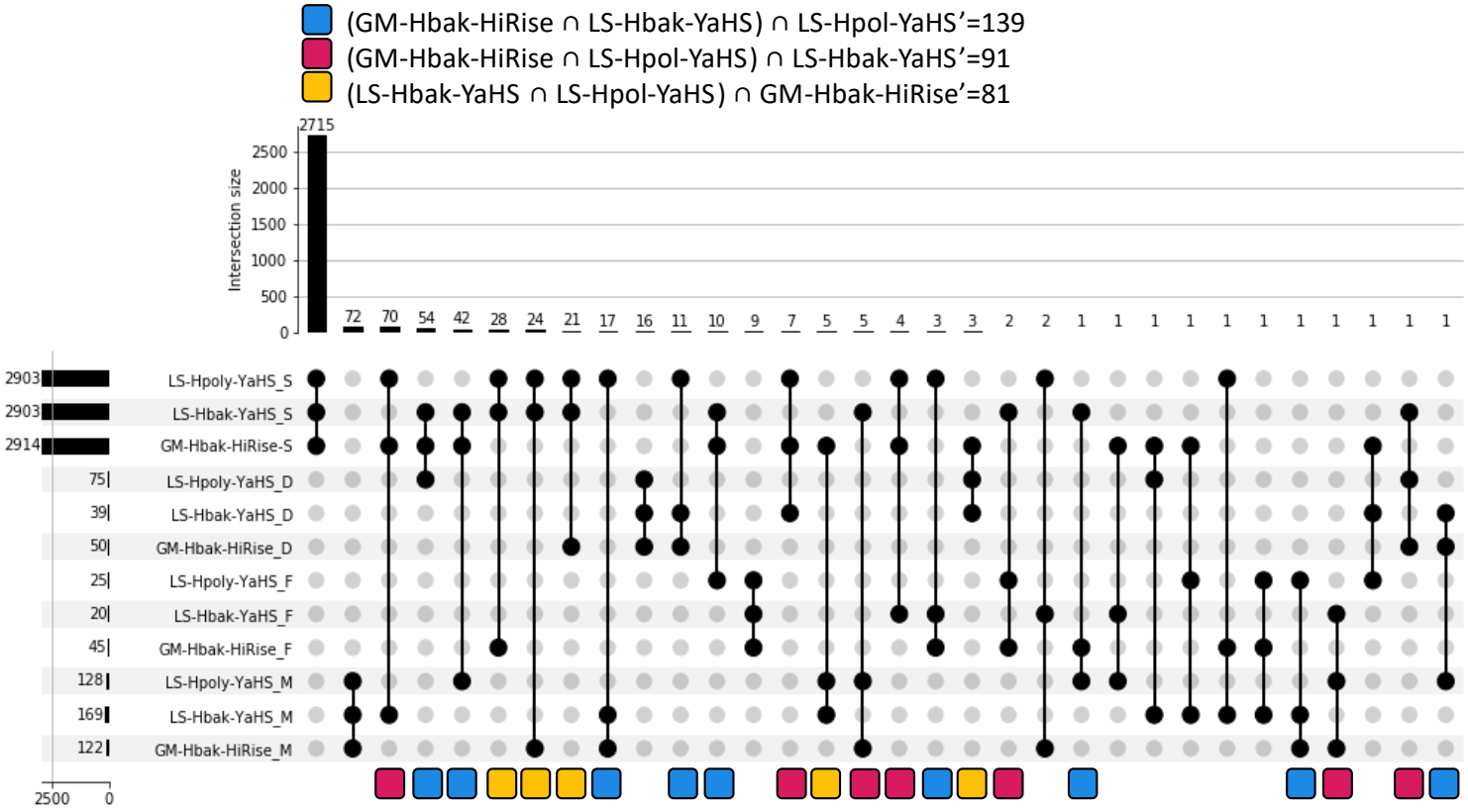

Figure 3.8. An UpSet plot of the three *Heligmosomoides* scaffolded assemblies, GM-Hbak-HiRise, LS-Hbak-YaHS and LS-Hpol-YaHS. The blue, red and orange boxes are BUSCO categories shared between pairs of *Heligmosomoides* assemblies to the exclusion of the third. (GM-Hbak-HiRise ∩ LS-Hbak-YaHS) ∩ LS-Hpol-YaHS' denotes shared BUSCO categories between GM-Hbak-HiRise and LS-Hbak-YaHS to the exclusion of (complement) LS-Hpol-YaHS, e.g., S-S-F.

I saw no bias between the two single-worm YaHS scaffolded assemblies, which adds further confidence to future analyses. When I take the union of the single and duplicated BUSCO genes in the two *H. bakeri* assemblies (GM-Hbak-HiRise and LS-Hbak-YaHS), I find 3024 of the 3131 BUSCO genes (for a completeness of 96.6%), implying that the LS-Hbak-YaHS assembly has BUSCO genes that could be missing in the GM-Hbak-HiRise. This could be the result of biological variation at the population level (Canadian vs UK lab isolates). However, when coupled with the comparison of GM-Hbak-HiRise and GM-Hbak-YaHS assemblies, I propose that many of the genes are present in the *in vivo* Canadian *H. bakeri* genome.

3.4.5 Conclusion

The technical variation observed between the HiRise- and the YaHS-scaffolded GM-Hbak assemblies was minimal. This suggests that employing different scaffolding software does not introduce large-scale biases that could impede comparative genomic studies of species scaffolded by different methods. Moreover, the GM-Hbak-HiRise assembly was slightly more contiguous and complete in terms of highly conserved genes than the LS-Hbak-YaHS assembly, making it the *H. bakeri* assembly of choice for future research in our laboratory.

Chapter 4

Exploring *Heligmosomoides bakeri* as a model for Clade V parasitic nematodes

4.1 Abstract

Evolution is based on the concept of homology of genes, describing the relationship between genes related by common ancestry. In this Chapter, I test the utility of *Heligmosomoides bakeri* as a model for Clade V parasitic nematodes, using a genomics approach of orthology and phylogenomic comparisons focusing on gene families involved in xenobiotic metabolism. Results show that *H. bakeri* shares more genes with the worms of interest—*Haemonchus contortus*, a livestock parasite with confirmed anthelmintic resistance, and *Necator americanus*, a human parasite with emerging resistance concerns—a finding that is congruent with the species tree. Similarly, phylogenetic trees of the CYP, GST and UGT families confirm that *H. bakeri* is more closely related to the parasites of interest than *Caenorhabditis elegans*, except for the GST-3 gene, where the *C. elegans* GST-3 was equally similar.

4.2 Introduction

Most scientific discovery has been made possible using model organisms. Extensive studies of these carefully selected non-human organisms have led to key biological insights of other species through shared ancestries and conserved genetic similarities over evolutionary times (Fields and Johnston, 2005; Allmon and Ross, 2018).

Historically, model organisms are selected based on characteristics that make them attractive genetic tools to study. These characteristics include, rapid life cycles, short generation times, low cost of maintenance, ease of experimental manipulation in a laboratory, genetic simplicity and similarity to other organisms (Leonelli and Ankeny, 2013). Despite these advantages, the use of model organisms has several limitations including ethical concerns, an inability for models to translate accurately and directly to other species or replicate complex biological systems and natural environments or an over-reliance on a limited range of model species which can bias results, limiting the broader applicability of conclusions (Hunter, 2008; Haas, 2011; Levy and Currie, 2015).

Among the earliest model organisms widely studied are the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the flowering plant *Arabidopsis thaliana,* and the nematode *Caenorhabditis elegans,* all of which were among the first organisms to have their genomes sequenced (Goffeau *et al.*, 1996; Blattner *et al.*, 1997;

Adams *et al.*, 2000; The *C. elegans* Sequencing Consortium, 1998; The Arabidopsis Genome Initiative, 2000).

Caenorhabditis elegans, has significantly advanced our understanding of human and animal biology. For example, it has enabled the generation of the complete cell lineage from fertilization to adulthood, as well as the discovery of genes involved in the developmental apoptosis pathway (Sulston and Horvitz, 1977; Kimble and Hirsh, 1979; Sulston *et al.*, 1983; Ellis and Horvitz, 1986). Furthermore, it has been instrumental in the study of human health and disease; approximately 60-80% of human genes have orthologs in *C. elegans*, including ~40% of genes linked to human diseases (Kaletta and Hengartner, 2006; Culetto and Sattelle, 2000). Given its multicellular nature, ease of study and maintenance, transparency for cellular differentiation studies, and well-established cellular lineages, *C. elegans* is the simplest multicellular model organism and was the first to have its genome sequenced (The *C. elegans* Sequencing Consortium, 1998). Despite being a free-living nematode, *C. elegans* has been an invaluable model for studying parasitic nematodes including anthelmintic resistance mechanisms, drug targets and parasite gene function and regulation (Gilleard, 2004; Geary and Thompson, 2001). Parasitic nematodes are difficult to study in controlled laboratory environments due to their reliance on a host to complete their life cycles and their significant genetic diversity, which complicates the study of key genes (Blaxter, 1998; Wasmuth *et al.*, 2008). Nevertheless, a deeper understanding of these organisms is required.

Three approaches—RNA interference (RNAi), heterologous expression and CRISPR/Cas9—have significantly advanced complementary research on parasitic nematodes by extrapolating insights gained from studies on *C. elegans*. RNAi, first discovered in *C. elegans*, selectively silences specific genes by inhibiting their expression to determine their functions. This method is especially useful for parasitic species that are refractory to genetic transformation or challenging to culture and manipulate at a genetic level (Sugimoto, 2004). Examining *C. elegans* genes for knockdown phenotypes in parasitic nematodes, for instance, has led to the discovery of important genes required for the larval and neuronal development of nematodes (Roubin *et al.*, 1999; Mohamed and Chin-Sang, 2011). Moreover, *C. elegans* has been shown to be responsive to anthelmintics and RNAi has provided insight into the mechanism of action of these drugs (Kaminsky *et al.*, 2008; Holden-Dye and Walker, 2007). Heterologous expression, on the other hand, involves expressing a gene in a host that lacks it to determine the gene's function (Tamás and Shewry, 2006). Used as a successful heterologous expression system, *C. elegans* has enhanced its value as a model by testing the function of specific genes or gene

mutants from parasitic nematodes. For instance, drug sensitivity can be restored in *C. elegans* mutants through heterologous expression of drug targets from *H. contortus* (Kwa *et al.*, 1995; Glendinning *et al.*, 2011). However, the function of many genes may not be conserved using this approach requiring the need to study genes directly in parasitic species through RNAi (Gilleard, 2013). Lastly, CRISPR/Cas9 (Clustered regularly interspaced palindromic repeats) is a gene editing technology consisting of two key components—a programmable, sequence-specific CRISPR guide RNA (crRNA) that matches a target gene and Cas9 (CRISPR-associated protein 9), an endonuclease that cleaves DNA, generating double-stranded breaks at target sites—which enable manipulation of specific genomic elements, to determine target gene functions (Redman *et al.*, 2016; Bak *et al.*, 2018; Zhang *et al.*, 2014). Targeted, heritable genetic modifications using CRISPR/Ca9 have been successfully achieved in *C. elegans*, enabling generation of loss-of-function mutants. Phenotypic analysis and characterization of *C. elegans* mutants have enabled the identification of key genes essential for nervous system function and collagen formation in nematodes; these genes have been successfully used in the CRISPR/Cas9 genome editing approach in *C. elegans* (von Mende *et al.*, 1988; Maduro and Pilgrim, 1995; Friedland *et al.*, 2013). Another example is a CRISPR/Cas9-mediated deletion that removed seven of the eight drug-metabolizing UDP-glucuronosyltransferase (UGT) genes within the UGT-9 cluster in *C. elegans* enabling investigations into albendazole drug sensitivity in *H. contortus* (Sharma *et al*., 2024).

The exponentially accumulating wealth of nematode genome resources has enabled comparative studies to understand phenotypes and test the utility of parasitic nematodes as models. *Haemonchus contortus*, a blood-feeding parasitic nematode found mainly in sheep, has been a particularly valuable model for drug resistance studies (Prichard, 2001; Yates *et al.*, 2003; Gilleard, 2006; Wolstenholme and Kaplan, 2012; Saunders *et al.*, 2013). However, like other parasitic nematodes, it lacks a laboratory animal host. In this chapter, I have focused on *Heligmosomoides bakeri*, a parasitic nematode with a laboratory animal host, the mouse. *H. bakeri* is a naturally occurring roundworm found in the mouse intestine and serves as a model for human helminth infection due to its ability to establish chronic infections in various mouse strains (Behnke, Menge, *et al.*, 2009). Additionally, it is used as a model to study host-parasite interactions and host immune responses (Maizels *et al.*, 2012; Reynolds *et al.*, 2012; Pollo *et al.*, 2023; Finney *et al.*, 2007). *H. bakeri* is found in the Rhabditina suborder (Clade V) of the nematode phylogeny and is closely related to *C. elegans*, and gastrointestinal parasitic nematodes with significant importance to animal and human health. Examples include the dog hookworm (*Ancylostoma caninum*), the human hookworm (*Necator americanus*), and

Trichostrongylids such as *Haemonchus contortus*, *Ostertagia ostertagi*, and *Cooperia oncophora* (Figure 1.5). Anthelmintic resistance has been confirmed in the three Trichostrongylids and there is concern that mass drug administration (MDA) could lead to resistance emerging in human hookworms. Notably, resistance has already emerged in the dog hookworm, *Ancylostoma caninum* (Venkatesan *et al.*, 2023).

H. bakeri has a direct life cycle (i.e., no intermediate host is required) and an infective L3 stage similar to that of other parasitic nematodes within the same clade (Figure 4.1). Unlike *H. contortus*, which requires sacrificing a sheep host for adult worm collection, a costly process, *H. bakeri* adult worms are relatively easy and cost effective to obtain from mice intestines.

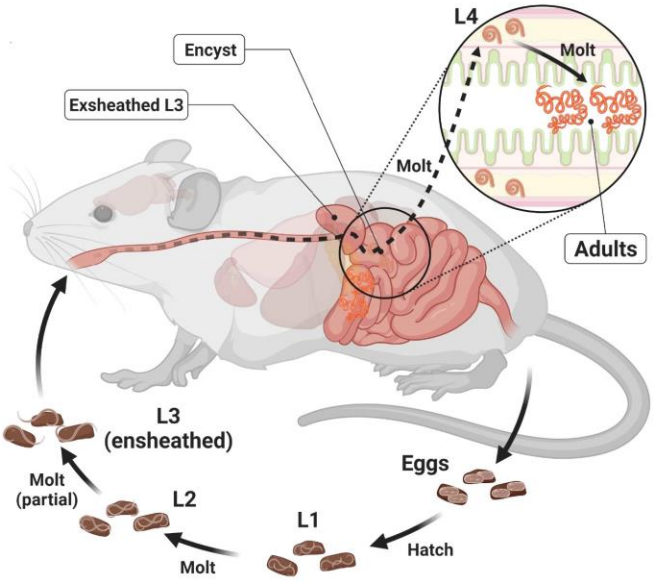

Figure 4.1. The *Heligmosomoides bakeri* life cycle. This figure is from (Pollo *et al*., 2023).

Immune responses to infections have been shown to differ significantly between wild and laboratory mice (Graham, 2021). Similarly, vaccine efficacy against the livestock parasite *Teladorsagia circumcincta* in lambing ewes, has been shown to differ significantly between ewes in pen trials contaminated with *T. circumcincta* and those in the natural field environments exposed to multiple common species of parasitic nematodes of sheep including *T. circumcincta*, *Chabertia ovina, Oesophagostomum venulosum* and *Haemonchus contortus* (Nisbet *et al.*, 2013, 2024). Although I could not identify a similar study on anthelmintic efficacy, it is reasonable to infer that the same factors likely apply to anthelmintic drug responses, with wild strains of parasitic nematodes exhibiting different responses compared to laboratory strains.

Indeed, *in silico* molecular docking studies have highlighted differences in drug binding between wild-type and mutated *Ascaris* beta-tubulin genes (Jones *et al.* 2022a).

Natural variation is useful for identifying the genetic causes of anthelmintic resistance. Differential drug susceptibilities and dose-responses have been observed in genetically diverse (wild-type) *C. elegans* strains making it a useful model for screening potential anthelmintics (Wit *et al.*, 2021; Shaver *et al.*, 2023). An earlier study showed that *C. elegans* maintained on ampicillin-treated *Escherichia coli* demonstrated sensitivity to benzimidazole anthelmintics, enhancing its utility for high-throughput in vitro pre-screens of anthelmintics, as well as for investigating drug modes of action and mechanisms of nematode resistance to anthelmintics (Simpkin and Coles, 1981). Another study by Burns *et al.*, (2015), conducted high-throughput liquid-based chemical assay screening for anthelmintic lead compounds that disrupted the *C. elegans* life cycle, ultimately killing the worm. Successful anthelmintic candidates were subsequently re-screened against two parasitic nematodes, *Cooperia onchophora and H. contortus*, revealing that targeted drug molecules that were effective against *C. elegans* (i.e., lethal to *C. elegans*) were more than 15 times more likely to demonstrate similar activity against parasitic species compared to randomly selected compounds (Burns *et al.*, 2015). Similar to *C. elegans*, *H. bakeri*, a naturally occurring parasite, also demonstrates genetic variability across different strains, which are propagated as it is shared live between experimental laboratories; it has also been used in drug screening studies (Stevens *et al.*, 2023; Hu *et al.*, 2010).

The *H. bakeri* genome is large (~650 Mb) and is characterized by high heterozygosity, hyper-divergent haplotypes in vaccine-targeted loci, and trans-specific polymorphisms in homologues of vaccine antigens used against both human hookworms (like *N. americanus*) and the livestock parasite *H. contortus* (Stevens *et al.*, 2023). While this complexity may limit its effectiveness in vaccine trials against the human and livestock parasites, the recently proposed larval blood-feeding phenotype makes it a potentially useful model to study other aspects of hookworm infections (Szabo *et al.*, 2024). Other studies on *H. bakeri* have examined its extracellular vesicles, revealing the presence of microRNAs that may exert immunomodulatory effects in the host (White *et al.*, 2020; Buck *et al.*, 2014). Additionally, three protein families in *H. bakeri*—TGF-b mimics (TGM), ARI, and BARI—have been shown to modulate the host's immune system (Johnston *et al.*, 2017; Osbourn *et al.*, 2017; Smyth *et al.*, 2018). While these studies primarily focus on the host's response to the worms, they also aim to investigate the worms' behavior and mechanisms of parasitism. Given the conserved core biology and expected shared mechanisms with other Clade V nematodes, *H. bakeri* may be useful for studying the

diversity and conservation of genes, such as those involved in key metabolic pathways, including drug metabolism.

Several gene families are involved in drug metabolism in nematodes and play important roles in modulating drug sensitivity. Examples of these families include: the Cytochrome P450s (CYP), Glutathione S-transferases (GST), and UDP-glucuronosyltransferases (UGT). While the function of most genes in these families remain unknown, the expression of these multigene family members is induced by drugs, highlighting their potential role in drug resistance (Kawalek *et al.*, 1984; Menzel *et al.*, 2001, 2005; Laing *et al.*, 2010; Matoušková *et al.*, 2018).

The *C. elegans* genome encodes approximately 79 CYP genes, 74 GST genes, and 67 UGT genes (WormBase Parasite (WBPS18; WS285) (Howe *et al.*, 2017). These numbers vary significantly across other nematode species likely reflecting true biological genic differences and/or discrepancies arising from the assembly and gene-prediction approaches used. Using Braker3 (Chapter 2), I annotated 52 CYP, 37 GST, and 35 UGT genes in the *H. contortus* (Doyle) genome, and 35 CYP, 46 GST, and 43 UGT genes in the *H. bakeri* (Hifiasm) genome. Drug-metabolizing gene families are expanded in the genomes of free-living nematodes, likely due to greater environmental exposure to complex metabolic substrates and xenobiotics compared to obligate endoparasites (Nelson, 1998; Dieterich *et al.*, 2008; Zarowiecki and Berriman, 2015; Markov *et al.*, 2015). In this chapter, I focused on a few examples of genes from the CYP, GST, and UGT families, with known literature on their role in xenobiotic detoxification, as discussed below.

For the CYPs, CYP-35D1 has been shown to be involved in the detoxification of the anthelmintic thiobenzamide (TBZ) in *C. elegans* (Jones *et al.*, 2015). TBZ directly binds to the nuclear hormone receptor NHR-176, which is essential for the TBZ-induced expression of the cytochrome P450 enzyme encoded by CYP-35D1 (Jones *et al.*, 2015). A second example is CYP-13A11, whose expression has been shown to induce reduced susceptibility to the anthelmintic avermectin in *H. contortus,* an effect not attributed to genetic polymorphisms (Jakobs *et al.*, 2022). CYP-43A1 is another example. In an earlier assembly of the *H. contortus* genome, the proposed orthologue of this gene (HCOI02017000) was the most highly expressed CYP (Laing *et al.*, 2013). However, this CYP was found missing in the current reference assembly (results are shown in Chapter 2) (Doyle *et al.*, 2020).

For the GSTs, examples include: GST-25, GST-30 and GST-3, whose expression has been shown to be induced by the soil fumigant and nematicide, dazomet in *C. elegans* (Jones *et al.*,

2013). Results suggested that dazomet may serve as a substrate for Phase II metabolism without requiring prior Phase I metabolism, which explained its induction of Phase II enzymes but not Phase I enzymes (Jones *et al.*, 2013). For the UGT gene family, the expression of UGT-13 and UGT-8 have been shown to be induced by thiabendazole and albendazole in *C. elegans* in a mechanism involving induction by thiabendazole and catalysis of a glycosylated metabolite identified during albendazole metabolism (Jones *et al.*, 2013). UGT-9, a member of the "ugt-9 cluster" in *C. elegans* (indicated in Figure 4.2), was one the UGTs prioritized in a study by Sharma *et al.* (2024). RNAi knockdown of UGT-9 expression in *C. elegans* showed strongest effects of increased susceptibility to the anthelmintic albendazole in *C. elegans* (Sharma *et al.*, 2024).

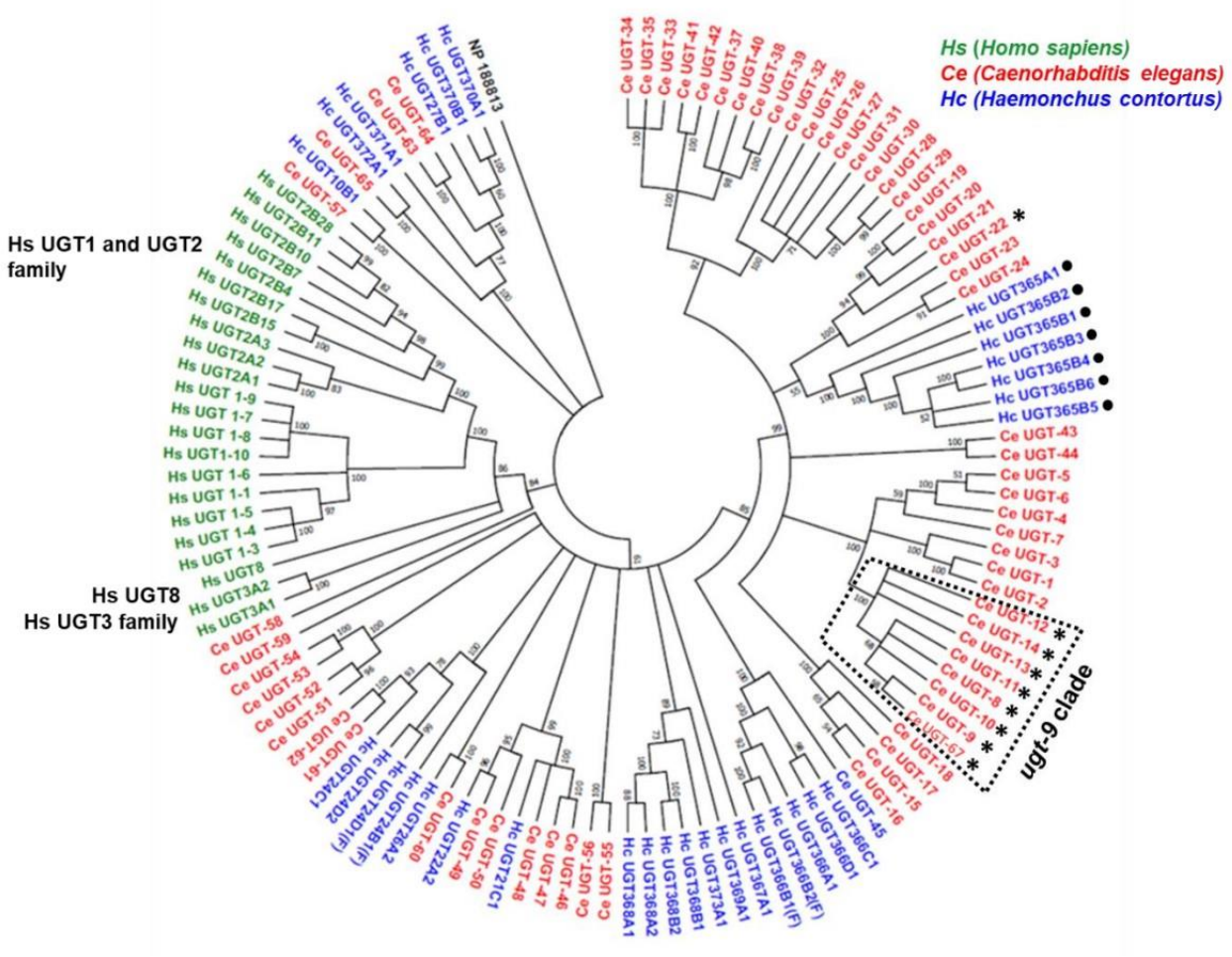

Figure 4.2. Phylogenetic analysis of *Caenorhabditis elegans*, *Haemonchus contortus* and Homo sapiens UGTs. This figure is from (Sharma *et al.,* 2024). The ugt-9 clade is indicated.

As a contribution to these studies, I wanted to find out the utility of *H. bakeri* as a genetic model for understanding gene evolution and resistance mechanisms in livestock and human parasites. Specifically, I used a genomics approach of orthology and phylogenomic comparisons focussing on the specific members of each gene family, previously outlined, i.e., Cytochrome P450s (CYP): (CYP-35D1, CYP-13A11, CYP-43A1), Glutathione S-transferases (GST): (GST-25, GST-30, GST-3) and UDP-glucuronosyltransferases (UGT): (UGT-13, UGT-8, UGT-9). The aim was to test the effectiveness of *H. bakeri* as a model for three Clade V nematodes found in the gastrointestinal tract of their hosts: *Nippostrongylus brasiliensis* (rat hookworm), *Necator americanus* (human hookworm) and *Haemonchus contortus* (sheep roundworm). I chose to investigate *N. americanus* over other human *Ancylostoma* hookworms due to its global prevalence which make it the most important human hookworm.

The utility of any model organism largely depends on the evolutionary conservation between model and target organisms, the phenotype of interest and specific study objectives. Here, I examined one-to-one orthologues, shared orthogroups, and nine genes (three from each of the CYP, GST, and UGT families) providing a preliminary comparison between the *H. bakeri* and *C. elegans* models, rather than a comprehensive analysis. Results from the phylogenomic comparisons, based on shared homologous genes across the study species, aligned with the established nematode phylogeny (Figure 1.5). However, it was insufficient to validate the usefulness of any model organism. Comparisons of gene family trees showed partial congruency with the species tree, except for the GST-3 gene, where *C. elegans'* CYP-GST-3, was equally similar to *H. bakeri* CYP-GST-3. In all other cases, *H. bakeri* was more closely related to other genes from the parasites of interest, confirming more evolutionary closeness and therefore potentially conserved drug metabolizing pathways. However, being a much easier Clade V study system to use, the utility of *C. elegans* as a model for nematodes remains unmatched. Therefore, *H. bakeri* should be used in parallel with *C. elegans* to study genes in Clade V nematodes.

4.3 Methods

4.3.1 Selection of genome assemblies and completeness evaluation

I downloaded high-quality genome assemblies (determined by the BUSCO completeness score) for four nematodes—*Nippostrongylus brasiliensis,* (BioProject PRJNA994163), *Necator americanus,* (BioProject PRJNA1007425), *Haemonchus contortus*, (BioProject PRJEB506), and

Caenorhabditis elegans, (BioProject PRJNA13758) from WormBase Parasite (WBPS19; WS291) (Howe *et al.*, 2017). As part of the analysis, I also included our Dovetail-scaffolded *Heligmosomoides bakeri* (BioProject PRJNA1087270), referred to as GM-Hbak-HiRise in Chapter 3, which I accepted as the assembly of choice for future research in our laboratory based on its better assembly statistics.

The sequence reads for the scaffolded assemblies of the rat and human hookworms, (*N. brasiliensis* and *N. americanus*), were generated using the long-read Oxford Nanopore Technologies (ONT) platform and assembled into genome sizes estimated at 257.3 and 234.4 Mb respectively (Howe *et al.*, 2017). The *H. contortus* chromosome-scale assembly was an improvement using PacBio RS II/Sequel long reads and optical mapping resulting into a genome size of 283.4 Mb (Doyle *et al.*, 2020). The *C. elegans* assembly, was generated from a combination of physical and genetic maps and manual gap filling approaches and assembled into a genome size of ~100.3 Mb (Howe *et al.*, 2017; The *C. elegans* Sequencing Consortium, 1998; Hillier *et al.*, 2005). Lastly, our scaffolded *H. bakeri* assembly was generated from PacBio HiFi sequence reads resulting into an assembly of approximately 678.4 Mb.

Assembly completeness was evaluated by determining the proportion of the nematode single-copy orthologues in the assemblies. To do this, I used BUSCO v5.4.6 (Simão *et al.*, 2015) with options set at: '--lineage nematoda_odb10' (n=3131 proteins), '--mode genome', '--long' and '--augustus'.

4.3.2 Assembly reannotation

For uniformity, I reannotated the assemblies of the *N. brasiliensis*, *N. americanus* and *C. elegans*, using Braker v3.0.0 (Gabriel *et al.*, 2024). The program was installed using the Singularity container (available at https://hub.docker.com/r/teambraker/braker3). I then compared all the species' annotations to the original annotations found in WormBase Parasite (WBPS19; WS291) (Howe *et al.*, 2017). For the *H. contortus* and our scaffolded *H. bakeri* assemblies, I used the previously predicted braker3-gene models outlined in Chapter 2. Prior to the reannotation of the three worms, (*N. brasiliensis*, *N. americanus* and *C. elegans*), I performed the following two steps:

4.3.2.1 Assembly soft masking

I used RepeatModeler v2.0.3 (Flynn *et al.*, 2020) with the NCBI/RMBLAST v2.11.0+ repetitive sequence search engine, to create a repeat library from the Dfam database of transposable elements and DNA repeats (Storer *et al.*, 2021). This was followed by the creation of a

customized repeat library specific to nematode species using the RepeatMasker util (queryRepeatDatabase.pl script) from RepeatMasker v4.1.2-p1 (Tarailo-Graovac and Chen, 2009). The two repeat libraries were concatenated and used as input for the RepeatMasker program which quantified and masked repeat elements in the assemblies.

4.3.2.2 External evidence, annotation and completeness evaluation

For *N. brasiliensis* and *N. americanus*, I used RNA-seq data from the Sequence Read Archive as external evidence for the gene prediction (BioProjects: PRJNA994163 and PRJNA1007425 respectively). The STAR v2.7.10a aligner (Dobin *et al.*, 2013) was used to map the RNA-seq data, producing bam-formatted alignment sets used to train AUGUSTUS v3.5.0, using GeneMark-ET v4.71 (Hoff *et al.*, 2016). I used DIAMOND v0.9.24 (Buchfink *et al.*, 2015) to filter redundant gene structures predicted by AUGUSTUS. For *C. elegans*, I used the longest isoforms of its predicted gene set, (BioProject: PRJNA13758), retrieved from WormBase Parasite (WBPS19; WS291) and the default protein sequences retrieved from UniProtKB/Swiss-Prot database as the protein evidence for gene prediction (Boutet *et al.*, 2007; Howe *et al.*, 2017). Here, Braker3 used the ProtHint pipeline to produce hints using GeneMark-EP v4.71, DIAMOND v0.9.24 and Spaln v2.3.3d tools (Brůna *et al.*, 2020; Buchfink *et al.*, 2015; Gotoh *et al.*, 2014; Iwata and Gotoh, 2012; Lomsadze *et al.*, 2005).

To assess the completeness of the predicted gene sets, I ran BUSCO v5.4.6 (Simão *et al.*, 2015) on the longest isoforms, with options set at: '--lineage nematoda_odb10' (n=3131 proteins), '--mode proteins', '--long' and '--augustus'.

4.3.3 Determining orthology

To cluster the orthologous braker3-annotated genes in the five worms, I used OrthoFinder2 with the, '-S blast' option; all other parameters were set at default (Emms and Kelly, 2019). I focused my results on orthogroups that reported one-to-one relationships between species pairs and those that were shared between species.

Additionally, I identified how genes were shared at specific cluster levels between *C. elegans*/*H. bakeri* to the exclusion of everything else: a) Genes shared between *H. bakeri*/*C. elegans* and *N. brasiliensis* (rat hookworm) but not others, b) Genes shared between *H. bakeri*/*C. elegans* and *N. americanus* (human hookworm) but not others, c) Genes shared between *H. bakeri*/*C. elegans* and *H. contortus* (livestock parasite) but not others, d) Genes shared between *H. bakeri*/*C. elegans* and rat and human hookworms but not others and e) Genes shared across all four species but not *H. bakeri*/*C. elegans*.

4.3.4 Identification of gene family members and phylogenetic analysis

I focused on specific members of three gene families involved in the detoxification of xenobiotics: Cytochrome P450s (CYP)—CYP-35D1, CYP-13A11, CYP-43A1; Glutathione S-transferases (GST)—GST-25, GST-30, GST-3; and UDP-glucuronosyltransferases (UGT)—UGT-13, UGT-8, UGT-9. To identify these genes in other genomes, I compared known gene family members from the *C. elegans* genome against the full protein sets of the four species using BLASTP v2.9.0+ (Camacho *et al.*, 2009). To distinguish the gene family members from each species, I used the following abbreviations: Nb for *Nippostrongylus brasiliensis*, Na for *Necator americanus*, Hc for *Haemonchus contortus*, Hb for *Heligmosomoides bakeri*, and Ce for *Caenorhabditis elegans*. The abbreviations were followed by the gene family member name, e.g., Hb_ CYP-35D1 for *H. bakeri*-CYP-35D1.

The identified protein sequences were then aligned with MUSCLE and evolutionary analysis conducted in MEGA version 11 (Tamura *et al.*, 2021). For each gene family, the evolutionary histories were inferred by using the Maximum Likelihood method and Le_Gascuel_2008 model (Le and Gascuel, 2008). The bootstrap consensus trees were inferred from 100 replicates (Felsenstein, 1985) and taken to represent the evolutionary histories of the species analyzed. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model and then selecting the topology with superior log likelihood value. The gene trees were drawn to scale; branches with at least 50% bootstrap replicates were shown, while those below 50% were collapsed.

4.4 Results

4.4.1 Assembly and annotation completeness evaluation

I selected the genome assemblies of the species included in this study from WormBase Parasite (WBPS19; WS291) based on their BUSCO scores and included only those with high scores. I then evaluated the assemblies and annotations for completeness using BUSCOv5.4.6 (Simão *et al.*, 2015). BUSCO identified the proportion of the 3131 single-copy nematode orthologues in each assembly and annotation (Table 4.1 and Table 4.2 respectively). The completeness of all the assemblies was >93% for complete and single copy BUSCO genes with

no more than 1.6% for complete and duplicated BUSCO genes. The *H. contortus* assembly had the highest proportion of missing BUSCOs (4.5%) followed by the *H. bakeri* assembly (3.9%).

Table 4.1. Genome assembly statistics for the Clade V species used in this study.

Assembly	**Project Number**	**Assembly Size (Mbp)**	**BUSCO (%, n=3131, mode=genome)**
GM-*Hbakeri*-HiRise	PRJNA1087270	678.4	C:94.7 [S:93.1, D:1.6], F:1.4, M:3.9
***N. brasiliensis* (WormBase19)**	PRJNA994163	257.3	C:96.8 [S:95.5, D:1.3], F:0.7, M:2.5
***H. contortus* (WormBase19)**	PRJEB506	283.4	C:94.8 [S:93.7%, D:1.1], F:0.7, M:4.5
***N. americanus* (WormBase19)**	PRJNA1007425	234.4	C:97.3 [S:96.4, D:0.9], F:0.8, M:1.9
***C. elegans* (WormBase19)**	PRJNA13758	100.3	C:99.4 [S:98.9, D:0.5], F:0.1, M:0.5

BUSCO genes' categories: Complete 'C', is the sum of single copy 'S' and duplicated 'D' genes; 'F' denotes Fragmented genes, and 'M' denotes Missing genes (all % from 3131 genes). Four assemblies—*N. brasiliensis, H. contortus, N. americanus* and *C. elegans*—were retrieved from WormBase Parasite (WBPS19; WS291). The GM-Hbakeri-HiRise assembly was our scaffolded *H. bakeri* assembly from Chapter 3.

For the annotations of the four species—*N. brasiliensis*, *N. americanus*, *H. contortus* and *C. elegans*—I compared the original annotations from WormBase Parasite (WBPS19; WS291) with my reannotations from Braker3 (Gabriel *et al.*, 2024; Howe *et al.*, 2017). It was not surprising that there were differences in the results between gene sets called by different annotation protocols, likely due to technical variation and an inflation in lineage-specific genes (Weisman *et al.*, 2022). Although here I am not comparing annotations/gene finding protocols, it was striking that the percentage differences between the number of gene models from WormBase Parasite and Braker3 ranged between 8% for *N. brasiliensis* and 63% for *N. americanus* (Table 4.2).

BUSCO scores were similar between the WormBase and Braker3 annotations, even for *N. brasiliensis*. However, for *H. contortus*, it is noteworthy that there was a 28% difference in

predicted gene models between the two annotations: a 4% difference in complete and single copy BUSCO genes, and an approximately 65% difference in missing BUSCOs. It is likely that either the Braker3 gene prediction tool failed to call all genes in the assembly, or errors were introduced in the reference *H. contortus* assembly through over-collapsing true repeats during the extensive manual curation that aimed at reducing heterozygous contigs.

Table 4.2. Annotation statistics for the Clade V species used in this study.

Assembly	**Number of gene models**	**Percentage difference between WormBase19 and Braker3 gene models**	**BUSCO (%, n=3131, mode=protein)**
GM-*Hbakeri*-HiRise	19,201	-	C:95.2 [S:92.2, D:3.0], F:1.0, M:3.8
N. brasiliensis			
WormBase19	16,461	8%	C:94.6 [S:92.9, D:1.7], F:0.7, M:4.7
Braker3	15,214		C:94.7 [S:93.2, D:1.5], F:1.0, M:4.3
H. contortus			
WormBase19	19,623	28%	C:96.2 [S:94.6, D:1.6], F:0.5, M:3.3
Braker3	14,824		C:92.7 [S:91.1, D:1.6], F:0.8, M:6.5
N. americanus			
WormBase19	26,617	63%	C:94.3 [S:93.2, D:1.1], F:1.2, M:4.5
Braker3	13,800		C:95.4 [S:94.3, D:1.1], F:0.8, M:3.8
C. elegans			
WormBase19	19,984	9%	C:99.7 [S:99.3, D:0.4], F:0.1, M:0.2
Braker3	21,945		C:99.2 [S:98.7, D:0.5], F:0.6, M:0.2

BUSCO genes' categories: Complete 'C', is the sum of single copy 'S' and duplicated 'D' genes; 'F' denotes Fragmented genes, and 'M' denotes Missing genes (all % from 3131 genes). I included both the original gene predictions from WormBase Parasite (WBPS19; WS291) and those from Braker3.

4.4.2 Orthology analysis

The next step was to determine evolutionary relationships by identifying orthologous genes present in the species of interest. To achieve this, I used OrthoFinder2 (Emms and Kelly, 2019), using all the protein sets predicted using Braker3. This program clusters orthologous genes into different orthogroups—groups of genes that have descended from a single gene in the last common ancestor of a group of species (Emms and Kelly, 2019).

All genes within orthogroups trace their origin to a single ancestral gene and are assumed to have initially shared the same sequence and function. However, over the course of evolution, processes such as gene loss and gene expansion (primarily through duplication) occur frequently, leading to various types of relationships among genes, including one-to-one, many-to-one, and many-to-many associations (Theißen, 2002; Emms and Kelly, 2019).

I considered only orthogroups that reported one-to-one relationships between species pairs and those that were shared between species. I recognize that focusing exclusively on one-to-one relationships limits the analysis to a small subset of gene relationships. However, given that this is a preliminary analysis, including all relationship types would generate an overwhelming volume of data, potentially expanding the study significantly. A higher number of shared one-to-one orthologs between species suggests a closer evolutionary relationship. *H. bakeri* shared the highest number of one-to-one orthologs with *N. brasiliensis*, followed by *H. contortus*, *N. americanus* and lastly, *C. elegans* (Table 4.3).

Table 4.3. Number of one-to-one orthologues between each pair of species.

	C. elegans	***H. contortus***	***H. bakeri***	***N. americanus***	***N. brasiliensis***
C. elegans	0	7256	7239	7279	7285
H. contortus	7256	0	8996	8724	8837
H. bakeri	7239	8996	0	8735	9108
N. americanus	7279	8724	8735	0	8717
N. brasiliensis	7285	8837	9108	8717	0

The same pattern was observed in the orthogroups shared between species (Table 4.4). These patterns align with the established nematode phylogeny lending confidence to the results (Figure 1.5).

Table 4.4. Orthogroups Species Overlaps: Orthogroups shared between species.

	C. elegans	***H. contortus***	***H. bakeri***	***N. americanus***	***N. brasiliensis***
C. elegans	10163	8502	8778	8510	8631
H. contortus	8502	10492	9810	9263	9484
H. bakeri	8778	9810	11313	9602	9920
N. americanus	8510	9263	9602	10272	9328
N. brasiliensis	8631	9484	9920	9328	10638

I then investigated specific sets of orthologous genes of interest to determine how genes were shared across different orthogroups at specific levels to the exclusion of everything else (Table 4.5).

A total of 7757 genes were shared in all five nematodes. *H. bakeri, H. contortus,* and *C. elegans* shared approximately 40%, 52% and 35% of their genes with all other nematodes respectively, while the rat and human hookworms—*N. brasiliensis* and *N. americanus*—shared 51% and 56%, respectively (Table 4.2 and Table 4.5). Considered as sister species, a higher number of genes shared between *H. bakeri* and *N. brasiliensis* was expected (Table 4.5a). Given this close relationship, it would have been surprising if *H. bakeri* shared genes with *N. americanus* that *N. americanus* and *N. brasiliensis* did not share (Table 4.5b). Similarly, there would be more genes shared between *H. bakeri* and *N. brasiliensis* than to *H. contortus*. Therefore, excluding *N. brasiliensis* likely excluded all the shared genes (Table 4.5c). Congruent with the species tree, it is expected that *H. bakeri* would share a higher number of orthologous genes with the rat and human hookworms than *C. elegans* (Table 4.5d). It is also likely that *H. bakeri* underwent lineage-specific loss of 226 genes and *C. elegans* lost 1177 genes shared with all other four parasitic nematodes (Table 4.5e). These could be genes that have either been gained during the evolutionary path leading to the Strongylida (gastrointestinal nematodes) or lost during the evolutionary path leading to *C. elegans*. Including a broader range of species in the analysis would provide a more robust comparative study, yielding more conclusive insights into the evolutionary paths of these genes.

Table 4.5. Table of shared genes at specific cluster levels between *Caenorhabditis elegans/Heligmosomoides bakeri* to the exclusion of everything else.

	C. elegans	*H. contortus*	*H. bakeri*	*N. americanus*	*N. brasiliensis*	Total
Genes shared in all species	✓	✓	✓	✓	✓	7757
a) C. elegans	✓	-	-	-	✓	667
H. bakeri	-	-	✓	-	✓	966
b) C. elegans	✓	-	-	✓	-	0
H. bakeri	-	-	✓	✓	-	0
c) C. elegans	✓	✓	-	-	-	0
H. bakeri	-	✓	✓	-	-	0
d) C. elegans	✓	-	-	✓	✓	132
H. bakeri	-	-	✓	✓	✓	273
e) C. elegans	-	✓	✓	✓	✓	1177
H. bakeri	✓	✓	-	✓	✓	226

a) Genes shared between *H. bakeri/C. elegans* and *N. brasiliensis* (rat hookworm) but not others

b) Genes shared between *H. bakeri/C. elegans* and *N. americanus* (human hookworm) but not others

c) Genes shared between *H. bakeri/C. elegans* and *H. contortus* (livestock parasite) but not others

d) Genes shared between *H. bakeri/C. elegans* and rat and human hookworms but not others

e) Genes shared across all four species but not *H. bakeri/C. elegans*

4.4.3 Phylogenetic analysis of gene family members

Next, I investigated whether specific members of the CYP, GST and UGT gene families implicated in anthelmintic resistance, could be better understood using *H. bakeri* than *C. elegans*. Based on the presence or absence of these genes across the study species, a pattern of expansion and loss within the gene families was observed, with 27 of the expected 45 genes found present (Table 4.6).

Table 4.6. Table of specific members of the CYP, GST and UGT gene families present or absent in the five Clade V species used in this study.

Gene family members	*H. bakeri*	*H. contortus*	*N. americanus*	*N. brasiliensis*	*C. elegans*	Total
CYP-13A11	✓	-	-	✓	✓	3
CYP-35D1	✓	✓	✓	-	✓	4
CYP-43A1	-	-	✓	-	✓	2
GST-25	-	-	-	-	✓	1
GST-30	✓	✓	✓	✓	✓	5
GST-3	✓	✓	✓	✓	✓	5
UGT-13	-	✓	✓	-	✓	3
UGT-8	-	✓	-	-	✓	2
UGT-9	-	✓	-	-	✓	2
Total	**4**	6	5	3	**9**	27

In total, *H. bakeri* had five fewer genes than *C. elegans*, suggesting potential gene loss in *H. bakeri* and/or expansion in *C. elegans* (Table 4.6). Similarly, the inferred phylogenies of these genes, constructed with MEGA 11 (Tamura *et al.*, 2021), displayed this characteristic pattern of large gene families (International Helminth Genomes Consortium, 2019).

4.4.3.1 Glutathione S-transferases (GSTs)

GST-30 and GST-3 were present in all the study species while GST-25 was exclusive to *C. elegans,* suggesting an expansion of this gene in *C. elegans* (Table 4.6). Unlike single-copy gene trees which typically mirror species trees, multi-copy genes, such as genes within these large families, often show branching patterns that conflict with species trees. These patterns are indicated by either a distinct separation or a complex mixing of similar genes within a species or across species. The latter pattern was observed in the GST gene tree (Figure 4.3A). For example, Ce_GST-30 did not cluster with other GST-30 genes from other species. Similarly, Hc_GST-3 clustered with Hc_GST-30, suggesting that the genes are the result of a recent gene duplication, rather than separate orthologs of the *C. elegans* genes. Ce_GST-25 appeared to be more similar to Ce_GST-3 than to GST-3 genes from other species, suggesting a possible

duplication of Ce_GST-3 or a very close similarity of Ce_GST-25 to Ce_GST-3. GST genes, Ce_GST-25 and Ce_GST-3 from the free-living *C. elegans*, clustered with Hb_GST-3 and Nb_GST-3 (which are sister parasitic species) than Na_GST-3. Notably, Nb_GST-30 appeared to be the least similar gene to everything else.

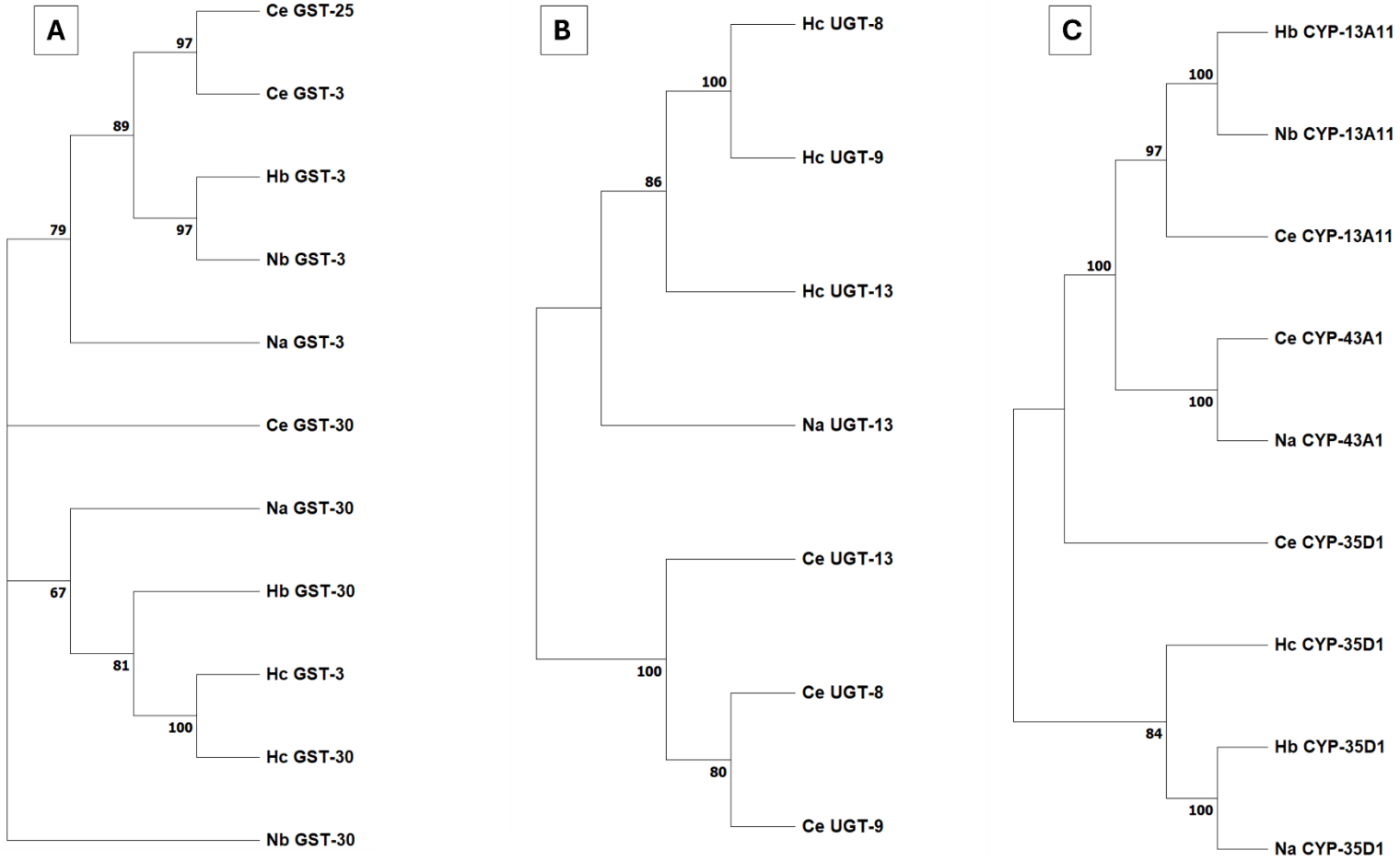

Figure 4.3. Phylogenetic trees of gene families. A) GSTs phylogenetic tree, B) UGTs phylogenetic tree and, C) CYPs phylogenetic tree. For each gene family, the evolutionary histories were inferred by using the Maximum Likelihood method and Le_Gascuel_2008 model (Le and Gascuel, 2008). The bootstrap consensus trees were inferred from 100 replicates (Felsenstein, 1985) and taken to represent the evolutionary histories of the species analyzed. The gene trees were drawn to scale; branches with at least 50% bootstrap replicates were shown, while those below 50% were collapsed.

The aim of this study was to investigate whether *H. bakeri* was a useful model to understand evolution of gene families in livestock and human parasites (*H. contortus* and *N. americanus*). For the GST-3 gene, *N. americanus'* (Na_GST-3), was similar to both *H. bakeri* and *C. elegans*. Here, the choice of a model organism would be *C. elegans*—it is free living and a much easier to use study system than *H. bakeri*. For GST-30, both *H. contortus* and *N. americanus* GST-30s clustered more closely to *H. bakeri* GST-30 than to *C. elegans*'. In this case, *H. bakeri* would be a more useful model to use than *C. elegans*.

4.4.3.2 UDP glucuronosyltransferases (UGTs)

All UGT genes of prior interest were present in *H. contortus* and *C. elegans* but absent in *N. brasiliensis* and *H. bakeri*, which are sister species. As these two species were sequenced and assembled by independent groups, the loss of these genes is likely to be biological than technical. *N. americanus* had only one UGT gene present: UGT-13 (Table 4.6). Based on these results alone, *C. elegans* would be the suitable model species. However, the gene tree shows that the UGT genes present were much more similar within a species than between species (Figure 4.3B). Na_UGT-13 was more similar to *H. contortus* than to *C. elegans,* consistent with the species tree. No additional information could be inferred from the gene tree suggesting that in this case, a model organism would not be very useful for extrapolating the evolution of this gene family.

4.4.3.3 Cytochrome P450s (CYPs)

N. americanus had two CYPs of prior interest (CYP-35D1 and CYP-43A1), *H. bakeri* had two CYPs (CYP-13A11 and CYP-35D1), and *H. contortus* had one CYP (CYP-35D1) (Table 4.6). CYP-43A1 was present only in *N. americanus* and *C. elegans*. From Chapter 2, this gene was absent from the current *H. contortus* reference assembly (Doyle *et al.*, 2020) although its proposed orthologue was shown to be most highly expressed in an earlier *H. contortus* assembly version (Laing *et al.*, 2013). This gene was also absent in *H. bakeri*, making *C. elegans* the model to consider for the study of this gene. CYP-13A11 was present in *H. bakeri*, *N. brasiliensis* and *C. elegans*, while CYP-35D1 was present in *H. bakeri*, *H. contortus, N. americanus* and *C. elegans.*

The CYPs phylogenetic tree, (Figure 4.3C), was largely congruent with the species tree except for Hb_CYP-35D1, which was more similar to CYP-35D1 gene in *N. americanus* than in *H. contortus*. Here, CYP-35D1 gene from *N. americanus* can be best extrapolated by *H. bakeri*. Additionally, with the exception of Ce_CYP-35D1 which did not cluster with other CYP-35D1

genes, other identical genes found in different species were clustered together: Hb_CYP-13A11 and Nb_ CYP-13A11, Ce_ 43A1 and Na_CYP-43A1, and Hb_CYP-35D1 and Na_CYP-35D1. In these instances, *H. bakeri* would be advantageous to use than *C. elegans*. Furthermore, since the nematode phylogeny places *H. bakeri* closer to the livestock and human parasites, the argument would be that it is a better model to use than the free-living *C. elegans*. However, *C. elegans* is a much easier study system than *H. bakeri* and has well-established resources and tools.

4.5 Discussion

The conserved genetic code in protein-coding sequences has advanced studies in gene evolution through the identification of homologous gene sequences—sequences derived from a common ancestral gene sequence – across species (Gabaldón, 2008).

Homologous genes are categorized into two main types: orthologues and paralogues, with xenologues—homologous genes that originate from horizontal gene transfer—also identified (Theißen, 2002; Fitch, 1970; Gabaldón and Koonin, 2013). Several proposed subtypes of paralogs also exist (Sonnhammer and Koonin, 2002). Orthologues are homologous genes that descend from a single common ancestor through speciation while paralogues are homologous genes that evolved through duplication events (Jensen, 2001; Gabaldón and Koonin, 2013). Genes in different species are considered orthologs if they originated from a single gene in the last common ancestor and diverged following speciation events leading to the formation of the distinct species (Anisimova, 2012). Orthologues can exhibit one-to-one, one-to-many, or many-to-many relationships, depending on the gene duplication events that occurred after their divergence (Theißen, 2002; Emms and Kelly, 2019).

Since phylogenetic reconstruction of species aims at representing the paths of past speciation events, orthologous gene relationships, rather than paralogous, are considered (Baldauf, 2003). Additionally, orthologues are presumed to retain the function of the ancestral gene and are conserved across species while paralogues, over time, either diverge functionally or result in the loss of all but one copy (with a few exceptions) (Theißen, 2002; Studer and Robinson-Rechavi, 2009). While functional similarity may reflect orthology, it can also arise independently through convergent evolution (Theißen, 2002). Moreover, orthologs of highly conserved genes are more likely to retain similar functions, a case not always true for genes that are members of a large multigene family; functional divergence between orthologs is expected (Cui *et al.*, 2021; Deng *et*

al., 2024). Hence, orthology and functional similarity should be assessed independently through phylogenetic and comparative functional analyses (Theißen, 2002).

There are several approaches for inferring orthology from sequence data. Traditionally, orthology was inferred using heuristic analysis of pairwise sequence comparisons, primarily based on similarity score thresholds using programs such as BLAST and DIAMOND (Buchfink *et al.*, 2015; Camacho *et al.*, 2009; Emms and Kelly, 2019). An example is the best bi-directional best hits (BBH), also known as best reciprocal hits (BRH), which detects pairs of sequences from different species that are, reciprocally, the best hit of each other in a sequence search (Huynen and Bork, 1998). The BRH methods reports one-to-one orthology relationships (Gabaldón, 2008). However, given the variable evolution rates between genes, false-positive and negative errors occur, making the approach less effective at representing one-to-many or many-to-many relationships, often missing true orthologs (Gabaldón, 2008; Emms and Kelly, 2019).

The BRH approach has since then been extended to incorporate clusters of orthologous groups (COGs) (Tatusov *et al.*, 1997). As orthology and paralogy relationships extends through subsequent speciation and duplication events, groups of orthologs, rather than individual pairs, more accurately represent the ancestral relationships in genes across species (Gabaldón, 2008). Examples of methods that implement this approach include Inparanoid, OrthoMCL and OrthoFinder; each method uses different approaches to analyze sequence similarity scores and identify paralogs, orthologs and paralogs or orthogroups (O'Brien *et al.*, 2005; Li *et al.*, 2003; Emms and Kelly, 2019). I used OrthoFinder in my analysis to cluster orthologous genes. Briefly, OrthoFinder identifies orthologous and paralogous genes by performing sequence similarity searches on proteomes of different species with DIAMOND or BLAST (Buchfink *et al.*, 2015; Camacho *et al.*, 2009), then uses a Markov Cluster Algorithm to assign proteins to orthogroups and infer gene trees of the orthogroups using methods like DendroBLAST (Kelly and Maini, 2013). The output from OrthoFinder includes genes in orthologous groups; genes duplicated after speciation are also reported in the same group.

Including model organisms in comparing orthology and evolutionary relationships across species can enhance the understanding of biological insights in species of interest. In this study, I assessed whether *H. bakeri* could serve as a useful model to study gene evolution in Clade V nematodes. *C. elegans*, a well-established free-living model organism, was included in comparisons to the parasitic nematodes. I compared both species for their effectiveness in studying specific members of the drug metabolism gene families. I acknowledge the limitation of

the study in lacking a suitable outgroup for the comparative analysis, as *C. elegans*, also a Clade V nematode, was one of the species targeted in the comparisons.

Among the key features that constitute a model organism are, ease of maintenance and manipulation, genetic simplicity, established experimental protocols, well-understood biological mechanisms, short life cycles and shared genetic similarity across species (Leonelli and Ankeny, 2013). Genetic similarity includes shared genes and can be used to infer orthology. However, interpreting that one organism is a more useful model than another solely based on the number of shared genes without assessing the similarity and exact functions of these genes, can be an oversimplification. Similar pathways may function effectively despite missing genes in different species. Thus, relying on the number of shared genes alone is not sufficient. The orthology results from this study aligned with the nematode phylogeny, confirming the accuracy of the annotation and orthology analysis.

A better consideration to investigate genetic similarities across species is the phenotype of interest. For example, the sheep parasite, *H. contortus*, and the human hookworm, *N. americanus*, are pathogenic, and understanding the genetic basis of their blood-feeding phenotypes is crucial. Since studying them directly is challenging, there is need for a model that replicates the blood-feeding behaviour. Studies have shown evidence of blood-feeding behavior in *N. brasiliensis* and *H. bakeri* (Bouchery *et al.*, 2018; Szabo *et al.*, 2024), making them potentially useful models to study human hookworms and pathogenic livestock nematodes. Similarly, investigating intestinal functions and gene families involved in drug metabolism in parasitic worms can help understand drug resistance, particularly since the intestine is hypothesized as the primary site for detoxification (Laing *et al.*, 2015; An and Blackwell, 2003). However, a useful model must metabolize drugs similarly to the species of interest.

The CYP, GST, and UGT gene families are involved in drug metabolism in *C. elegans* and other parasitic worms (Sharma *et al.*, 2024; Stasiuk *et al.*, 2019; Laing *et al.*, 2010). However, members of such large gene families can have multiple similar looking genes with unique roles that metabolize distinct or overlapping substrates resulting in different protein products (Sharma *et al.*, 2024). As a result, even small differences in these genes that are shared between organisms could produce real differences in a metabolic pathway. A study done by Hu *et al.*, (2013) showed that all the tested worms—*Ancylostoma ceylanicum* (human hookworm), *Trichuris muris* (mouse whipworm), *Ascaris suum* (pig and human roundworm), *H. bakeri* and *C. elegans*—responded differently to different doses of major anthelmintic drug classes, benzimidazoles, nicotinic acetylcholine receptor agonists, macrocyclic lactones and

nitazoxanide. These results suggest that even with closely similar drug metabolizing genes, species may not demonstrate similar effects, demonstrating a need to consider each species independently. One likely cause for this observation is genetic variation which can alter proteins and their expression in response to drugs enabling the organism to survive different dosages of treatment, potentially reflecting evolutionary adaptation to a toxic environment where drug resistance promotes 'survival of the fittest' (James *et al.*, 2009). Alternatively, gene interaction, also known as epistasis, can suppress the phenotypic effects of other genes, highlighting that most phenotypes result from complex gene interactions (Phillips, 2008; de Visser *et al.*, 2011).

Another likely cause for differential drug responses in nematodes has been investigated through *in silico* molecular docking studies. This computational method which simulates molecular interactions to predict protein-drug interactions and assess ligand binding to protein active sites, has been used to identify key interactions of benzimidazole resistance-associated amino acid mutations in several nematodes including *Haemonchus contortus*, *Trichinella spiralis*, *Ascaris lumbricoides*, *Ascaris suum* and filarial nematodes (Robinson *et al.*, 2004; Aguayo-Ortiz *et al.*, 2013; Halder *et al.*, 2019; Jones *et al.*, 2022a; 2002b). *In silico* docking studies have also demonstrated that variation in the protein sequence can induce small changes to the protein structure, potentially affecting drug binding. For example, Jones *et al.* (2022a) showed that in the *Ascaris lumbricoides* and *Ascaris suum* genomes, benzimidazole resistance-associated mutations at amino acids E198A and F200Y altered binding of benzimidazoles, while the F167Y mutation had no direct effect, but instead acted by binding to E198 and destabilizing the drug within the binding site. Furthermore, the number of beta-tubulin gene isotypes varies across the genomes of different helminth groups. For example, ascarids share a similar set of seven beta-tubulin isotypes whereas strongyles share four (Saunders *et al.*, 2013; Roose *et al.*, 2021). A study by Jones *et al.*, (2022b), demonstrated that differences in resistance and susceptibility patterns to benzimidazole drugs among ascarid, strongyle, and whipworm parasite groups were potentially associated to variations in beta-tubulin gene abundance and expression.

Most results from my phylogenomic analyses placed *H. bakeri* closer to the livestock and human parasites, suggesting similar pathways. However, results from Hu *et al.*, (2013) and the observed hyper-divergent haplotypes and polymorphisms in *H. bakeri* at homologues of the ASP and H11 antigens used in *N. americanus* and *H. contortus* vaccines (Stevens *et al.*, 2023), challenge *H. bakeri's* reliability as a model for drug-related research. Moreover, while an established model like *C. elegans* may not offer detailed insights into specific pathways, especially in species with different life cycles and host interactions, it is not sufficient to

definitively assert that *H. bakeri* is a more useful model than *C. elegans* solely based on *H. bakeri* being a parasite and *C. elegans* not being one. The argument that *C. elegans* is not a parasite overlooks the fact that not all parasites are the same. A parasite can be any organism (including helminths, protozoa and ectoparasites) that lives on or in a host organism and gets its food from or at the expense of its host (Centers for Disease Control and Prevention, 2024).

4.6 Conclusion

In summary, *H. bakeri* shared more genes with the worms of interest (*H. contortus* and *N. americanus*) than *C. elegans*. However, a higher number of shared homologous genes is not sufficient to validate the utility of a potential model organism as these genes may have distinct functions. For example, they may metabolize different substrates and influence different phenotypic traits, such as drug responses. Additionally, shared pathways can still remain functional despite species-specific gene differences, such as the absence of certain genes. Phylogenetic analysis of gene family members showed loose congruency with the species tree, confirming that *H. bakeri* is more closely related to the parasites of interest than *C. elegans*, except for the GST-3 gene, where Ce_GST-3 was equally similar. It is highly likely that *H. bakeri* shares more pathway similarities with the parasites making it more advantageous to use over *C. elegans*. However, *H. bakeri is* not as easy a study system to use as *C. elegans* is; *C. elegans* has a small genome size (100Mb) and the use of a single cryopreserved *C. elegans* stock prevents genetic diversity which may result in low levels of heterozygosity. In *H. bakeri*, genetic variability is observed in different strains as a result of it being shared live between experimental laboratories (Stevens *et al.*, 2023). In this case, *C. elegans* is a preferred model. This was a preliminary study. A more comprehensive phylogenetic analysis incorporating all the members of the CYP, GST and UGT genes families across the nematode species under study would provide more conclusive insights into patterns of gene family evolution related to parasitism. Additionally, it would identify more definitive instances to ascertain which species, *C. elegans* or *H. bakeri*, serves as a better model for understanding resistance mechanisms in Clade V parasitic nematodes. Notwithstanding this, *C. elegans* is well-studied and has well established protocols. Therefore, parasitic nematode models, such as *H. bakeri,* should be used in parallel with *C. elegans* to study genes.

Chapter 5

General Discussion

The goal of my studies was to generate a high quality, chromosome-level assembly for the parasitic nematode *Heligmosomoides bakeri*. This assembly, and the associated annotation are needed by a broad spectrum of researchers, including, those searching for new anthelmintic drugs, those striving to understand how parasitism evolves, and those investigating how the mammalian immune system suppresses infection.

I propose that I have succeeded in this goal, and I am excited to see how this resource will be used by colleagues. In pursuing this goal, I have also investigated the complex nuances of genome assembly. Those findings were critical to helping generate the required quality of the *H. bakeri* assembly. Arguably more importantly, it offers a framework for others, from BioGenome projects to individual researchers to pause and consider how they wish to move forward as they generate genome assemblies that will uncover insights into their species of interest.

5.1 Impact of algorithm choice on assembly and annotation

The development of long-read DNA sequencing technologies has led to a predictable surge in the number of available assembly programs and the accuracy of genome assemblies for eukaryotic species has certainly improved (Logsdon *et al.*, 2020; Laing *et al.*, 2013; Marks *et al.*, 2021; Doyle *et al.*, 2020). However, selecting the most suitable program for specific species remains challenging; the inherent algorithmic differences between assembly programs affect how biological features like genome size, ploidy, and repeat regions are interpreted. These differences introduce biases that can lead to variable assembly accuracy and, in some cases, compromise the preservation of crucial genetic data, which may have downstream impacts on research findings and conclusions (Jung *et al.*, 2020; Pollo *et al.*, 2020; Sun *et al.*, 2021).

In Chapter 2, I presented an evaluation of eight long-read assembly programs applied to three nematode species—*Caenorhabditis bovis*, *Haemonchus contortus*, and *Heligmosomoides bakeri*—which differ in terms of size, heterozygosity, and repeat content. The DNA sequence reads were generated by three different technologies, respectively, Oxford Nanopore's (ONT) MinION, Pacific BioSciences' (PacBio) RS II, and PacBio's HiFi. Through an array of assembly accuracy metrics and comparisons, I showed that different algorithms, even with identical input data, can produce considerably different assemblies. Generally, assemblers using Overlap-Layout-Consensus (OLC) algorithms were most effective for the larger, more heterozygous and repetitive genomes of *H. contortus* and *H. bakeri*, while those using De Bruijn Graph (DBG) algorithms excelled in assembling the 'simpler' genome of *C. bovis*. Nevertheless, the DBG-based Redbean assembler exhibited versatility, performing comparably well across all

categories. The technical variation in the assemblies affected gene prediction and, subsequently, composition of gene families, which has important implications for downstream analysis that spans biological research.

5.2 Sources of assembly variations

Nematodes are multi-cellular animals and, like many of their fellow metazoans, their genomes present significant assembly challenges: large intergenic and intronic regions; large multigene families; and repetitive DNA, which includes short tandem repeats (STRs) and transposable elements (TEs) (The *C. elegans* Sequencing Consortium, 1998; Chen and Stein, 2006). STRs are often found within coding sequences, while TEs are interspersed throughout non-coding regions, and these patterns complicate assembly (Cutter *et al.*, 2009). Alternative splicing, which is common in nematodes, further complicates assembly, as alternative splicing sites frequently contain transcribed STRs that bring exons into proximity, resulting in multiple gene isoforms (Lian and Garner, 2005). Furthermore, as diploid organisms, nematodes carry two chromosome sets with potentially distinct alleles at each locus, introducing heterozygosity, which can result in spurious duplications.

Long reads can help overcome many complexities of the *in vivo* genome. For example, they can span the repetitive regions. However, the sequencing platforms themselves introduce error. Both PacBio (RS and CLR methods) and ONT platforms have historically struggled with homopolymer stretches of DNA, particularly those longer than five nucleotide bases, and ONT devices have also struggled with the accurate reading of STRs and heavily methylated sites (Audano *et al.*, 2019; Eid *et al.*, 2009; Weirather *et al.*, 2017; Rhoads and Au, 2015; Giesselmann *et al.*, 2019; Mitsuhashi *et al.*, 2019). These errors manifest as short insertions/deletions (indels). When these occur in gene coding regions, they introduce frameshifts that can negatively affect gene-finding methods and, if found, the genes may be affected through a premature stop codon or error in the intron/exon boundaries, resulting in altered or truncated gene structures (Sutton *et al.*, 2021; Watson and Warr, 2019; Sacristán-Horcajada *et al.*, 2021; Vallender, 2017). Some of these errors, particularly those that are distributed randomly in the sequence reads, can be resolved during the assembly process. If sequenced at sufficient depth, the overlapping of sequence reads will reveal errors which can be fixed in the consensus contigs. A process of polishing, such as using highly accurate short reads, can also find and fix errors (Walker *et al.*, 2014). However, many of the errors are not randomly distributed, such as those on homopolymers, and correcting them requires more

sophisticated approaches that are an area of active development (https://github.com/nanoporetech/medaka; http://github.com/PacificBiosciences/pbbioconda).

Despite the notable improvements in sequencing technology and assembly programs, repetitive DNA remains a major confounding factor. These regions often create bifurcated paths in assembly graphs, which assemblers interpret as expansions or collapses, depending on the repeat type (Chen *et al.*, 2021). Tandem repeats, for example, are prone to collapsing, while interspersed repeats often lead to expansions. These problems are compounded by haplotypic differences at a locus between sister chromosomes within an individual organism, or between individuals if multiple samples were needed to obtain sufficient starting DNA material. When assemblers fail to discriminate between haplotypes, heterozygous breakpoints are reported as structural errors, often leading to collapsed or expanded regions. In some cases, highly divergent alleles may be misassembled into separate contigs, artificially inflating genome sizes and gene dosages, while in other cases, homologous haplotypes maybe collapsed, generating false gene duplications (Korlach *et al.*, 2017; Kelley and Salzberg, 2010; Roach *et al.*, 2018; Guan *et al.*, 2020)

5.3 BUSCO genes as an assembly accuracy indicator

The identification of BUSCO (Benchmarking Universal Single-Copy Orthologs) genes serves as a reliable metric for genome assembly accuracy and completeness (Simão *et al.*, 2015). Given that BUSCO genes within a lineage are highly conserved, shared classifications—such as genes reported as ‘duplicated’ or ‘missing’ across multiple assemblies for a given species—may reflect true biological characteristics. However, poor assembly of haplotypes and repetitive regions will lead to an increase in the number of genes classified as ‘duplicated’, ‘fragmented’, or ‘missing’ (Manni *et al.*, 2021). Typically, BUSCO reports the proportion of genes recovered – ideally as ‘single copy’—with the assembly containing the highest proportion considered the best. While such a strategy is understandable, it ignores useful information. I have shown that many of the assemblers generated assemblies with extremely similar BUSCO scores. However, there were notable differences in the actual repertoire of BUSCO genes recovered; some genes classified as ‘missing’ in the best scoring assembly were classified as ‘single copy’ in a worse scoring assembly. For example, 18 BUSCO genes classified as ‘fragmented’ or ‘missing’ in the best scoring *C. bovis* Redbean2.5 assembly were classified ‘single copy’ in the Canu assembly. Indeed, I found that in the published and highly curated *H. contortus* reference assembly (referred to as Doyle), 27 BUSCO genes were absent but were reported present in all the other newly generated assemblies for *H. contortus.*

5.4 Annotating the genome with gene families

It is important to bear in mind that BUSCO genes are genes that are conserved across a broad taxonomic range; here, I used the pan-nematode gene set. Their inclusion in the set indicates that these genes have not undergone recent duplications and are therefore among the easiest elements of the genome to assemble. However, many genes are members of large gene families; the seven transmembrane (7TM) receptor family in *C. elegans* contains approximately 1,500 genes, representing about 7% of all protein-coding genes, though numbers vary considerably in other nematode species (Troemel, 1999; Langeland *et al.*, 2020). Motivated by a long-standing interest of the Wasmuth lab in detoxification pathways, I looked at how assembly variation affected the discovery of genes from four families: CYP, GST, UGT, and NHR. I found considerable variation in the size of each family. Unfortunately, there was no consistent trend; a specific assembly method did not always give smaller families across the board. The same was true for recently discovered immunomodulators in *H. bakeri*. These findings should be a cause of concern for all researchers that make use of genome annotations. Further, the problem is not limited to nematodes; discrepancies in the accuracy of the assembly and its annotation can lead to inaccurate interpretation of evolutionary and functional capabilities of gene families.

5.5 The importance of genome resource accuracy: sources and consequences of errors

Genomics and bioinformatics have become essential components of most research in the biological, medical, and veterinary sciences. To facilitate the use of these data, web-based genome resources are essential. With bioinformatics still not a core competency of most biological-related undergraduate and graduate programs and the ever-increasing demand to publish, it is understandable that the data provided on these resources is taken as correct and there is little appreciation for errors in sequencing, assembly, and annotation that can significantly affect the reliability of research findings.

All non-model organisms, and many model organisms, depend on computational annotations for gene models and predicted molecular function. These annotations are typically based on sequence similarity with sequences deposited in repositories like GenBank and UniProt (Sayers *et al.*, 2020; The UniProt Consortium, 2023). Sequence similarity is determined through an algorithm, such as BLAST or DIAMOND, and is a proxy for homology and sometimes, orthology (Camacho *et al.*, 2009; Buchfink *et al.*, 2015). In this way the gene sequence is assigned a

functional annotation using a system like the Gene Ontology (GO) (Higgins *et al.*, 2022; Blum *et al.*, 2021).

The reliance on automated methods, with little or no human curation, will create errors that propagate through the system. An example from parasitic nematodes involves the search for orthologs to the *C. elegans* beta-tubulin gene, *ben-1*. The beta-tubulins are the target of the benzimidazoles, a widely used class of anthelmintics. Mutations in *ben-1* confer resistance to benzimidazole in *C. elegans* and the study of beta-tubulins in parasitic nematodes has been intense (Driscoll *et al.*, 1989; Redman *et al.*, 2015; Morrison *et al.*, 2022; Dolinská *et al.*, 2023). As there are multiple copies of the beta-tubulins in each species, the search has been on for each parasite's one-to-one ortholog with *ben-1* (Geldhof *et al.*, 2006). Sequence similarity has been used to annotate several genes as the *ben-1* ortholog and these exist in WormBase Parasite (Howe *et al.*, 2017). This is despite clear phylogenetic evidence that the repertoire of beta-tubulins in nematodes is phylogenetically complex and there is no one-to-one orthologue between *C. elegans* and any parasitic nematodes (Saunders *et al.*, 2013; Roose *et al.*, 2021).

Incremental assembly improvement without comprehensive re-annotation introduces further errors. Advances in sequencing, assembly algorithms, and RNA-seq data increase the need for re-annotation, yet many projects simply track gene releases, making only minimal updates. Comparative analyses of draft and improved assemblies in chickens, chimpanzees, and cattle, for example, showed discrepancies of up to 40% in gene annotations (Florea *et al.*, 2011; Denton *et al.*, 2014). In addition, up to 15% of the SNP database (dbSNP) records could not be recovered when the cattle genome assembly was improved (Florea *et al.*, 2011).

When carrying out any comparison between genome annotation datasets, it is important to consider heterogeneity in the gene prediction methods. In nematodes, WormBase Parasite, stores annotations generated by different labs, which used different annotation methods (Howe *et al.*, 2017). This heterogeneity can inflate species-specific gene counts, leading to misinterpretations (Weisman *et al.*, 2022).

Here, I have listed a few of the major challenges facing those that generate, curate, maintain, and use genome resources. The reliance on public genome resources is immense but their sustainability remains at the whims of funding agencies. This was demonstrated recently with the withdrawal of funding of VEuPathDB by the National Institute of Allergy and Infectious Diseases (Fernandez-Prada *et al.*, 2024). WormBase Parasite is also without funding and will no longer be updated.

5.6 Improving genome resource generation: practical recommendations

The process of constructing a high-quality genome assembly for any animal species is non-trivial and likely demands a tailored approach. Each step in the process—from the choice of the sequencing platform to post-assembly validation—can introduce new errors or amplify existing errors. My work, and that of others, demonstrates that currently there is no single assembler or annotation software that performs best across all species; the approaches can vary in their performance between relatively closely related species. Therefore, researchers would benefit by adapting their bioinformatics strategy according to specific research objectives and characteristics of their organisms of study (Bradnam *et al.*, 2013). Here, I provide four practical recommendations.

1. **Compare multiple assemblies from different software and parameter settings**

I have shown that even in the 'best' assembly, there are mistakes that are revealed by alternative assemblies. Therefore, finding the union or consensus between assemblies can improve the accuracy of a genome assembly. This approach was used to generate the final improved rhesus macaque genome by combining multiple assemblies from three assembly methods: Atlas whole-genome shotgun, parallel contig assembly program (PCAP), and the Celera Assembler (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.*, 2007). By aligning alternative *C. bovis* assemblies, I identified splits in the superior Redbean2.5 assembly that were better resolved in the Redbean2.3 assembly.

All assembly software has a wealth of parameters that control the software's ability to align reads, build the alignment graphs, resolve multifurcations, and generate consensus. These are rarely changed from default. A better understanding of the software and taking time to explore the 'parameter space' may lead to assemblies that better resolve repetitive regions and heterozygosity.

2. **Use multiple measures of assembly accuracy**

The quality of an assembly is often measured by contiguity and completeness. Each metric highlights different aspects of assembly accuracy: N50, for example, provides insights into contiguity but may overlook shorter, potentially significant contigs. Completeness metrics like BUSCO, which evaluates expected gene content, and Merqury, a k-mer-based tool, offer complementary perspectives on assembly reliability (Simão *et al.*, 2015; Rhie *et al.*, 2020). If

RNA-Seq data is available, this should be mapped to the alternative assemblies, as the one with the highest mapping rate presumably has more genes more accurately assembled (Liao *et al.*, 2014; Conesa *et al.*, 2016).

As with genome assembly software, it is important to consider the parameter choices of the quality assurance software. For example, in the pursuit of shorter run times, the default gene-finding algorithm in BUSCO has changed from AUGUSTUS to metaEuk, and most recently, to miniprot (Stanke *et al.*, 2006; Levy Karin *et al.*, 2020; Li, 2023). In my work on nematodes, I found that using metaEuk (which was the default at the time of my study), frequently identified 10% fewer BUSCO genes than AUGUSTUS (data not shown). However, for other species (non-nematode), it performed better (Wasmuth pers. comms.).

3. **Annotating the genome requires equal or greater caution**

There are considerably fewer annotation (gene-finding) strategies than assembly programs. The most popular, based on citations and anecdotal interest in the community is Braker, a suite of pipelines that uses a heterogeneity of evidence—including RNA-seq and protein sequences from other species—to train its models. While the release of new versions is likely to improve annotation accuracy is most species, colleagues in the lab and I have noticed that this is not always the case (Muthye pers. comms.). An alternative is Funannotate, originally designed to annotate fungal genome assemblies (Palmer and Stajich, 2023). I found that it was fast and accurate with *C. bovis* but dramatically underperformed on the larger *H. bakeri* and *H. contortus* assemblies (data not shown). While I did not have time to investigate this further, as the BioGenome projects increase the pace of submitting genome assemblies, added attention will be focused on finding the genes contained therein.

4. **Validate absence of genes from assemblies**

When genes of interest are known, it is crucial to validate their absence in assemblies, as they may exist in the *in vivo* genome but be missing from the *in silico* genome due to assembly errors. Moreover, some post-processing programs, like Purge_dups, reduce haplotypic duplications but at the cost of assembly completeness; genes get lost as haplotypic duplications reduce (Guan *et al.*, 2020). Validation of genes can be achieved computationally, such as by performing BLAST searches of known gene sequences against the *in silico* genome, or experimentally, through PCR, where primers are designed to amplify the target gene, followed by Sanger sequencing or cloning of the PCR product.

5.7 Conclusion

This thesis describes my investigations into the performance of widely used long-read assembly programs, with a focus on identifying bioinformatics protocols that produce high-quality genome resources for highly heterozygous parasitic nematodes. Accurate genome assemblies are essential for understanding anthelmintic resistance and host-parasite interactions, providing a foundation for drug development and functional genomics studies. Additionally, the work described creates awareness and cautions researchers, particularly those interested in comparative genomics and genic analysis, of the critical impacts of “small” variations in the preliminary draft genomes and the need for informed and more careful assembly generation approaches and selection.

References

Adams,M.D. *et al.* (2000) The Genome Sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Adams,P.E. *et al.* (2020) Genome Evolution: On the Road to Parasitism. *Current Biology*, **30**, R272–R274.

Aguayo-Ortiz,R. *et al.* (2013) Towards the identification of the binding site of benzimidazoles to β-tubulin of *Trichinella spiralis*: Insights from computational and experimental data. *Journal of Molecular Graphics and Modelling*, **41**, 12–19.

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**, R18.

Allendorf (2014) Heterozygosity. *Evolutionary Biology*. *Oxford Bibliographies Online*, doi: https://doi.org/10.1093/obo/9780199941728-0039

Allmon,W.D. and Ross,R.M. (2018) Evolutionary remnants as widely accessible evidence for evolution: the structure of the argument for application to evolution education. *Evolution: Education and Outreach*, **11**, 1.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

An,J.H. and Blackwell,T.K. (2003) SKN-1 links *C. elegans* mesendodermal specification to a conserved oxidative stress response. *Genes Dev*, **17**, 1882–1893.

Anderson,S. (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res*, **9**, 3015–3027.

Anghel,I.G. *et al.* (2022) Reference genome of the color polymorphic desert annual plant sandblossoms, *Linanthus parryae*. *Journal of Heredity*, **113**, 712–721.

Anisimova,M. (2012) Evolutionary Genomics: Statistical and Computational Methods | SpringerLink. Humana Press, New York, NY.

Ardui,S. *et al.* (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*, **46**, 2159–2168.

Audano,P.A. *et al.* (2019) Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, **176**, 663-675.e19.

Backström,N. *et al.* (2010) A High-Density Scan of the Z Chromosome in *Ficedula* Flycatchers Reveals Candidate Loci for Diversifying Selection. *Evolution*, **64**, 3461–3475.

Bak,R.O. *et al.* (2018) Gene Editing on Center Stage. *Trends Genet*, **34**, 600–611.

Baker,E.A. *et al.* (2021) Extensive non-redundancy in a recently duplicated developmental gene family. *BMC Ecology and Evolution*, **21**, 33.

Baldauf,S.L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends Genet*, **19**, 345–351.

Bao,W. *et al.* (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

Bashir,A. *et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol*, **30**, 701–707.

Batut,B. *et al.* (2018) Community-Driven Data Analysis Training for Biology. *cels*, **6**, 752-758.e1.

Behnke,J.M., Menge,D.M., *et al.* (2009) *Heligmosomoides bakeri*: a model for exploring the biology and genetics of resistance to chronic gastrointestinal nematode infections. *Parasitology*, **136**, 1565–1580.

Behnke,J.M., Eira,C., *et al.* (2009) Helminth species richness in wild wood mice, *Apodemus sylvaticus*, is enhanced by the presence of the intestinal nematode *Heligmosomoides polygyrus*. *Parasitology*, **136**, 793–804.

Bishop,M.J. (1984) Software club: Software for molecular biology. II. Restriction mapping and DNA sequencing programs. *BioEssays*, **1**, 75–77.

Blattner,F.R. *et al.* (1997) The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

Blaxter,M. (1998) *Caenorhabditis elegans* Is a Nematode. *Science*, **282**, 2041–2046.

Blaxter,M.L. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.

Blouin,M.S. *et al.* (1995) Host Movement and the Genetic Structure of Populations of Parasitic Nematodes. *Genetics*, **141**, 1007–1014.

Blum,M. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, **49**, D344–D354.

Boetzer,M. and Pirovano,W. (2014) SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, **15**, 211.

Bouchery,T. *et al.* (2018) A novel blood-feeding detoxification pathway in *Nippostrongylus brasiliensis* L3 reveals a potential checkpoint for arresting hookworm development. *PLOS Pathogens*, **14**, e1006931. doi: 10.1371/journal.ppat.1006931.

Boutet,E. *et al.* (2007) UniProtKB/Swiss-Prot. *Methods Mol Biol*, **406**, 89–112.

Bradnam,K.R. *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, **2**, 2047-217X-2–10.

Brasil,B.S.A.F. *et al.* (2012) Genetic diversity patterns of *Haemonchus placei* and *Haemonchus contortus* populations isolated from domestic ruminants in Brazil. *Int J Parasitol*, **42**, 469–479.

Brown,T.A. (2002) Mapping Genomes. In, *Genomes. 2nd edition*. Wiley-Liss. Genomes - NCBI Bookshelf.

de Bruijn,N.G. (1946) A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, **49**, 758–764.

Brůna,T. *et al.* (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, **3**, lqaa108.

Brůna,T. *et al.* (2020) GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform*, **2**, lqaa026.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, **12**, 59–60.

Buck,A.H. *et al.* (2014) Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat Commun*, **5**, 5488.

Burns,A.R. *et al.* (2015) *Caenorhabditis elegans* is a useful model for anthelmintic discovery. *Nat Commun*, **6**, 7485.

Cable,J. *et al.* (2006) Molecular evidence that *Heligmosomoides polygyrus* from laboratory mice and wood mice are separate species. *Parasitology*, **133**, 111–122.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Casasa,S. *et al.* (2021) Polyphenism of a Novel Trait Integrated Rapidly Evolving Genes into Ancestrally Plastic Networks. *Mol Biol Evol*, **38**, 331–343.

CBC Radio (2022) Scientists sequence complete, gap-free human genome for the first time | CBC Radio *Quirks and Quarks. Q&A*

Centers for Disease Control and Prevention (2024) About Parasites. *Parasites.* https://www.cdc.gov/parasites/about/index.html

Challis,R. *et al.* (2020) BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)*, **10**, 1361–1374.

Chen,N. and Stein,L.D. (2006) Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*. *Genome Res.*, **16**, 606–617.

Chen,Y. *et al.* (2021) Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology*, **22**, 312.

Chen,Y.C. *et al.* (2013) Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One*, **8**, doi: 10.1371/journal.pone.0062856.

Cheng,H. *et al.* (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, **18**, 170–175.

Chin,C.S. *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, **10**, 563–569.

Chin,C.S. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*, **13**, 1050–1054.

Chow,F.W.N. *et al.* (2019) Secretion of an Argonaute protein by a parasitic nematode and the evolution of its siRNA guides. *Nucleic Acids Res*, **47**, 3594–3606.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Coghlan,A. *et al.* (2023) A drug repurposing screen for whipworms informed by comparative genomics. *PLOS Neglected Tropical Diseases*, **17**, e0011205.

Colomb,F. *et al.* (2024) IL-33-binding HpARI family homologues with divergent effects in suppressing or enhancing type 2 immune responses. *Infection and Immunity*, **0**, e00395-23.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biology*, **17**, 13.

Conte,M.A. *et al.* (2017) A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics*, **18**, 341.

Cook,D.E. *et al.* (2017) CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Research*, **45**, D650–D657.

Cornet,L. *et al.* (2024) Evaluation of Genomic Contamination Detection Tools and Influence of Horizontal Gene Transfer on Their Efficiency through Contamination Simulations at Various Taxonomic Ranks. *Applied Microbiology*, **4**, 124–132.

Cornwell,B.H. *et al.* (2022) Reference genome assembly of the sunburst anemone, *Anthopleura sola*. *Journal of Heredity*, **113**, 699–705.

Cotton,J.A. *et al.* (2016) The genome of *Onchocerca volvulus*, agent of river blindness. *Nat Microbiol*, **2**, 1–12.

Cui,C. *et al.* (2021) Defining the functional divergence of orthologous genes between human and mouse in the context of miRNA regulation. *Briefings in Bioinformatics*, **22**, bbab253.

Culetto,E. and Sattelle,D.B. (2000) A role for *Caenorhabditis elegans* in understanding the function and interactions of human disease genes. *Human Molecular Genetics*, **9**, 869–877.

Cutter,A.D. *et al.* (2009) Evolution of the *Caenorhabditis elegans* Genome. *Molecular Biology and Evolution*, **26**, 1199–1234.

Daborn,P.J. *et al.* (2002) A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. *Science*, **297**, 2253–2256.

Darwin Tree of Life Project Consortium (2022) Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl Acad Sci U S A*, **119**, e2115642118.

David,J.P. *et al.* (2013) Role of cytochrome P450s in insecticide resistance: impact on the control of mosquito-borne diseases and use of insecticides on Earth. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, 20120429.

Deng,D. *et al.* (2024) Functional Divergence in Orthologous Transcription Factors: Insights from AtCBF2/3/1 and OsDREB1C. *Molecular Biology and Evolution*, **41**, msae089.

Denton,J.F. *et al.* (2014) Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLOS Computational Biology*, **10**, e1003998.

Deshpande,V. *et al.* (2013) Cerulean: A Hybrid Assembly Using High Throughput Short and Long Reads. In, Darling,A. and Stoye,J. (eds), *Algorithms in Bioinformatics*. Springer, Berlin, Heidelberg, pp. 349–363. doi: 10.48550/arXiv.1307.7933.

Devlin,H. (2022) First complete gap-free human genome sequence published | UpwardPost.*The Guardian*.

Dieterich,C. *et al.* (2008) The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet*, **40**, 1193–1198.

van Dijk,E.L. *et al.* (2018) The Third Revolution in Sequencing Technology. *Trends Genet*, **34**, 666–681.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Dohm,J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, **36**, e105.

Dolinská,M.U. *et al.* (2023) Variation in allele frequencies in benzimidazole resistant and susceptible isolates of *Haemonchus contortus* during patent infection in lambs. *Sci Rep*, **13**, 1296.

Dominguez Del Angel,V. *et al.* (2018) Ten steps to get started in Genome Assembly and Annotation. *F1000Res*, **7**, ELIXIR-148.

Doyle,S.R. *et al.* (2018) A Genome Resequencing-Based Genetic Map Reveals the Recombination Landscape of an Outbred Parasitic Nematode in the Presence of Polyploidy and Polyandry. *Genome Biol Evol*, **10**, 396–409.

Doyle,S.R. *et al.* (2020) Genomic and transcriptomic variation defines the chromosome-scale assembly of *Haemonchus contortus*, a model gastrointestinal worm. *Commun Biol*, **3**, 656.

Doyle,S.R. (2022) Improving helminth genome resources in the post-genomic era. *Trends Parasitol*, **38**, 831–840.

Doyle,S.R. and Cotton,J.A. (2019) Genome-wide Approaches to Investigate Anthelmintic Resistance. *Trends Parasitol*, **35**, 289–301.

Driscoll,M. *et al.* (1989) Genetic and molecular analysis of a *Caenorhabditis elegans* beta-tubulin that conveys benzimidazole sensitivity. *J Cell Biol*, **109**, 2993–3003.

Ebenezer,T.E. *et al.* (2022) Africa: sequence 100,000 species to safeguard biodiversity. *Nature*, **603**, 388–392.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Computational Biology*, **7**, e1002195.

Eid,J. *et al.* (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, **323**, 133–138.

Eilbeck,K. *et al.* (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, **10**, 67.

Elbers,J.P. *et al.* (2019) Improving Illumina assemblies with Hi-C and long reads: An example with the North African dromedary. *Mol Ecol Resour*, **19**, 1015–1026.

Ellis,H.M. and Horvitz,H.R. (1986) Genetic control of programmed cell death in the nematode *C. elegans*. *Cell*, **44**, 817–829.

Elston,R.C. (1995) Linkage and association to genetic markers. *Exp Clin Immunogenet*, **12**, 129–140.

Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 238.

English,A.C. *et al.* (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLOS ONE*, **7**, e47768.

Erkut,C. *et al.* (2013) Molecular Strategies of the *Caenorhabditis elegans* Dauer Larva to Survive Extreme Desiccation. *PLOS ONE*, **8**, e82473.

Espinosa,R. and Le Beau,M.M. (2000) Gene Mapping by FISH. In, Rapley,R. (ed), *The Nucleic Acid Protocols Handbook*, Springer Protocols Handbooks. Humana Press, Totowa, NJ, pp. 991–1010. doi: 10.1385/1-59259-038-1:991.

ESTs Factsheet (2002). https://web.archive.org/web/20020507082436/http://www.ncbi.nlm.nih.gov/About/primer/est.html.

Evans,R.M. and Mangelsdorf,D.J. (2014) Nuclear Receptors, RXR & the Big Bang. *Cell*, **157**, 255.

Felsenstein,J. (1985) CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution*, **39**, 783–791.

Fernandez-Prada,C. *et al.* (2024) Critical loss: the effects of VEuPathDB defunding on global health. *The Lancet Microbe*, **0**. doi: 10.1016/j.lanmic.2024.100980.

Fields,S. and Johnston,M. (2005) Cell biology. Whither model organism research? *Science*, **307**, 1885–1886.

Fierst,J.L. and Murdock,D.A. (2017) Decontaminating eukaryotic genome assemblies with machine learning. *BMC Bioinformatics*, **18**, 533.

Finney,C.A.M. *et al.* (2007) Expansion and activation of CD4+CD25+ regulatory T cells in *Heligmosomoides polygyrus* infection. *European Journal of Immunology*, **37**, 1874.

Fitch,W.M. (1970) Distinguishing Homologous from Analogous Proteins. *Systematic Biology*, **19**, 99–113.

Florea,L. *et al.* (2011) Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two *Bos taurus* Assemblies. *PLOS ONE*, **6**, e21400.

Flynn,J.M. *et al.* (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, **117**, 9451–9457.

Friedland,A.E. *et al.* (2013) Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat Methods*, **10**, 741–743.

Fukasawa,Y. *et al.* (2020) LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3: Genes|Genomes|Genetics*, **10**, 1193.

Gabaldón,T. (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biology*, **9**, 235.

Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics*, **14**, 360.

Gabriel,L. *et al.* (2024) BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*, **34**, 769.

Geary,T.G. and Thompson,D.P. (2001) *Caenorhabditis elegans*: how good a model for veterinary parasites? *Veterinary Parasitology*, **101**, 371–386.

Geldhof,P. *et al.* (2006) Testing the efficacy of RNA interference in *Haemonchus contortus*. *International Journal for Parasitology*, **36**, 801–810.

Geng,S. *et al.* (2022) Gut commensal *E. coli* outer membrane proteins activate the host food digestive system through neural-immune communication. *Cell Host & Microbe*, **30**, 1401-1416.e8.

Giesselmann,P. *et al.* (2019) Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol*, **37**, 1478–1481.

Gilabert,A. *et al.* (2016) Expanding the view on the evolution of the nematode dauer signalling pathways: refinement through gene gain and pathway co-option. *BMC Genomics*, **17**, 476.

Gilleard,J.S. (2013) *Haemonchus contortus* as a paradigm and model to study anthelmintic drug resistance. *Parasitology*, **140**, 1506–1522.

Gilleard,J.S. (2004) The use of *Caenorhabditis elegans* in parasitic nematode research. *Parasitology*, **128**, S49–S70.

Gilleard,J.S. (2006) Understanding anthelmintic resistance: the need for genomics and genetics. *Int J Parasitol*, **36**, 1227–1239.

Gilleard,J.S. and Beech,R.N. (2007) Population genetics of anthelmintic resistance in parasitic nematodes. *Parasitology*, **134**, 1133–1147.

Glendinning,S.K. *et al.* (2011) Glutamate-Gated Chloride Channels of *Haemonchus contortus* Restore Drug Sensitivity to Ivermectin Resistant *Caenorhabditis elegans*. *PLOS ONE*, **6**, e22390.

Goffeau,A. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Gotoh,O. *et al.* (2014) Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics*, **15**, 189.

Goussarov,G. *et al.* (2022) Introduction to the principles and methods underlying the recovery of metagenome-assembled genomes from metagenomic data. *MicrobiologyOpen*, **11**, e1298.

Graham,A.L. (2021) Naturalizing mouse models for immunology. *Nat Immunol*, **22**, 111–117.

Gray,J.C. and Cutter,A.D. (2014) Mainstreaming *Caenorhabditis elegans* in experimental evolution. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20133055. doi: 10.1098/rspb.2013.3055

Greener,M. (2021) What has *Caenorhabditis elegans* ever done for us? *Prescriber*, **32**, 29–32.

Grisi,L. *et al.* (2014) Reassessment of the potential economic impact of cattle parasites in Brazil. *Rev Bras Parasitol Vet*, **23**, 150–156.

Grismer,J.L. *et al.* (2022) Reference genome of the rubber boa, *Charina bottae* (Serpentes: Boidae). *Journal of Heredity*, **113**, 641–648.

Guan,D. *et al.* (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**, 2896–2898.

Haas,D. (2011) On the virtues and dangers of models. *FEMS Microbiology Reviews*, **35**, 1–2.

Halder,S.T. *et al.* (2019) Molecular Docking Studies of Filarial β-Tubulin Protein Models with Antifilarial Phytochemicals. *Biomedical and Biotechnology Research Journal (BBRJ)*, **3**, 162.

Heiskanen,M. *et al.* (1996) Visual mapping by high resolution FISH. *Trends Genet*, **12**, 379–382.

Heras,J. *et al.* (2016) A survey of tools for analysing DNA fingerprints. *Briefings in Bioinformatics*, **17**, 903–911.

Higgins,D.P. *et al.* (2022) Defining characteristics and conservation of poorly annotated genes in *Caenorhabditis elegans* using WormCat 2.0. *Genetics*, **221**, iyac085.

Hillier,L.W. *et al.* (2005) Genomics in *C. elegans*: So many genes, such a little worm. *Genome Res.*, **15**, 1651–1660.

Hiltemann,S. *et al.* (2023) Galaxy Training: A powerful framework for teaching! *PLOS Computational Biology*, **19**, e1010752.

Hirani,N. *et al.* (2016) *C. elegans* flavin-containing monooxygenase-4 is essential for osmoregulation in hypotonic stress. *Biology Open*, **5**, 537–549.

Hoff,K.J. *et al.* (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.

Holden-Dye,L. and Walker,R.J. (2007) Anthelmintic drugs. *WormBook*, 1–13. doi: 10.1895/wormbook.1.143.2

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

Howe,K.L. *et al.* (2017) WormBase ParaSite − a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology*, **215**, 2–10.

Hu,Y. *et al.* (2013) An Extensive Comparison of the Effect of Anthelmintic Classes on Diverse Nematodes. *PLOS ONE*, **8**, e70702.

Hu,Y. *et al.* (2010) *Bacillus thuringiensis* Cry5B Protein Is Highly Efficacious as a Single-Dose Therapy against an Intestinal Roundworm Infection in Mice. *PLOS Neglected Tropical Diseases*, **4**, e614.

Hubley,R. *et al.* (2016) The Dfam database of repetitive DNA families. *Nucleic Acids Res*, **44**, D81–D89.

Huelsmann,M. *et al.* (2019) Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. *Science Advances*, **5**, eaaw6671.

Hunt,V.L. *et al.* (2016) The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nat Genet*, **48**, 299–307.

Hunter,P. (2008) The paradox of model organisms. The use of model organisms in research will continue despite their shortcomings. *EMBO Rep*, **9**, 717–720.

Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proceedings of the National Academy of Sciences*, **95**, 5849–5856.

Idury,R.M. and Waterman,M.S. (1995) A new algorithm for DNA sequence assembly. *J Comput Biol*, **2**, 291–306.

International Helminth Genomes Consortium (2019) Comparative genomics of the major parasitic worms. *Nat Genet*, **51**, 163–174.

Iwata,H. and Gotoh,O. (2012) Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res*, **40**, e161.

Jain,M. *et al.* (2022) Advances in nanopore direct RNA sequencing. *Nat Methods*, **19**, 1160–1164.

Jain,M. *et al.* (2017) MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res*, **6**, 760.

Jakobs,N. *et al.* (2022) Transgenic Expression of *Haemonchus contortus* Cytochrome P450 Hco-cyp-13A11 Decreases Susceptibility to Particular but Not All Macrocyclic Lactones in the Model Organism *Caenorhabditis elegans*. *Int J Mol Sci*, **23**, 9155.

James,C.E. *et al.* (2009) Drug resistance mechanisms in helminths: is it survival of the fittest? *Trends in Parasitology*, **25**, 328–335.

Jauhal,A.A. and Newcomb,R.D. (2021) Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. *Mol Ecol Resour*, **21**, 1416–1421.

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*, **20**, 866–876.

Jensen,R.A. (2001) Orthologs and paralogs - we need to get it right. *Genome Biology*, **2**, interactions1002.1. doi: 10.1186/gb-2001-2-8-interactions1002.

Jensen-Seaman,M.I. *et al.* (2004) Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Res.*, **14**, 528–538.

Jiao,W.B. *et al.* (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res*, **27**, 778–786.

Johnston,C.J.C. *et al.* (2017) A structurally distinct TGF-β mimic from an intestinal helminth parasite potently induces regulatory T cells. *Nat Commun*, **8**, 1741.

Jones,B.P. *et al.* (2022a) Identification of key interactions of benzimidazole resistance-associated amino acid mutations in *Ascaris* β-tubulins by molecular docking simulations. *Sci Rep*, **12**, 13725.

Jones,B.P. *et al.* (2022b) *In Silico* Docking of Nematode β-Tubulins With Benzimidazoles Points to Gene Expression and Orthologue Variation as Factors in Anthelmintic Resistance. *Front. Trop. Dis*, **3**. 898814. doi: 10.3389/fitd.2022.898814.

Jones,L.M. *et al.* (2013) Adaptive and Specialised Transcriptional Responses to Xenobiotic Stress in *Caenorhabditis elegans* Are Regulated by Nuclear Hormone Receptors. *PLoS One*, **8**, e69956.

Jones,L.M. *et al.* (2015) NHR-176 regulates cyp-35d1 to control hydroxylation-dependent metabolism of thiabendazole in *Caenorhabditis elegans*. *Biochem J*, **466**, 37–44.

Jorde,L. (2002) Mapping Human History: Discovering the Past through Our Genes. *Am J Hum Genet*, **71**, 1484–1485.

Jung,H. *et al.* (2020) Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops. *J. Agric. Food Chem.*, **68**, 7670–7677.

Jung,H. *et al.* (2019) Tools and Strategies for Long-Read Sequencing and *De Novo* Assembly of Plant Genomes. *Trends in Plant Science*, **24**, 700–724.

Kajitani,R. *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 184-1395.

Kaletta,T. and Hengartner,M.O. (2006) Finding function in novel targets: *C. elegans* as a model organism. *Nat Rev Drug Discov*, **5**, 387–399.

Kamath,R.S. *et al.* (2000) Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*. *Genome Biology*, **2**, research0002.1.

Kaminsky,R. *et al.* (2008) A new class of anthelmintics effective against drug-resistant nematodes. *Nature*, **452**, 176–180.

Kanehisa,M. *et al.* (2016) BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol*, **428**, 726–731.

Kapitonov,V.V. and Jurka,J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet*, **9**, 411–412.

Kaplan,R.M. (2004) Drug resistance in nematodes of veterinary importance: a status report. *Trends Parasitol*, **20**, 477–481.

Kawalek,J.C. *et al.* (1984) Glutathione-S-transferase, a possible drug-metabolizing enzyme, in *Haemonchus contortus*: comparative activity of a cambendazole-resistant and a susceptible strain. *Int J Parasitol*, **14**, 173–175.

Kelley,D.R. and Salzberg,S.L. (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biology*, **11**, R28.

Kelly,S. and Maini,P.K. (2013) DendroBLAST: Approximate Phylogenetic Trees in the Absence of Multiple Sequence Alignments. *PLOS ONE*, **8**, e58537.

Kenyon,C. (1988) The Nematode *Caenorhabditis elegans*. *Science*, **240**, 1448–1453.

Khelik,K. *et al.* (2017) NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. *BMC Bioinformatics*, **18**, 338.

Kieran Blair,S.R. *et al.* (2022) The reference genome of the Vernal Pool Tadpole Shrimp, *Lepidurus packardi*. *Journal of Heredity*, **113**, 706–711.

Kimble,J. and Hirsh,D. (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Developmental Biology*, **70**, 396–417.

Kolmogorov,M. *et al.* (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, **37**, 540–546.

Koo,H. *et al.* (2023) Two long read-based genome assembly and annotation of polyploidy woody plants, *Hibiscus syriacus* using PacBio and Nanopore platforms. *Sci Data*, **10**, 713.

Koren,S. *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

Koren,S. *et al.* (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol*, **36**, 1174–1182.

Koren,S. and Phillippy,A.M. (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, **23**, 110–120.

Korlach,J. *et al.* (2017) De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, **6**, gix085.

Kreis,H.A. (1964) Ein neuer Nematode aus dem äusseren Gehörgang von Zeburindern in Ostafrika, Rhabditis *bovis* n.sp. (Rhabditidoidea, Rhabditidae). *Schweizer Archiv fur Tierheilkunde*, **106**, 371–378.

Krzywinski,M. *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Kuznetsov,D. *et al.* (2022) OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, **51**, D445.

Kwa,M.S.G. *et al.* (1995) β-Tubulin Genes from the Parasitic Nematode *Haemonchus contortus* Modulate Drug Resistance in *Caenorhabditis elegans*. *Journal of Molecular Biology*, **246**, 500–510.

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

Laing,R. *et al.* (2016) Chapter Thirteen - *Haemonchus contortus*: Genome Structure, Organization and Comparative Genomics. In, Gasser,R.B. and Samson-Himmelstjerna,G.V. (eds), *Advances in Parasitology*, Academic Press, **93**, 569–598.

Laing,R. *et al.* (2015) The cytochrome P450 family in the parasitic nematode *Haemonchus contortus*. *Int J Parasitol*, **45**, 243–251.

Laing,R. *et al.* (2013) The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery. *Genome Biol*, **14**, R88.

Laing,S.T. *et al.* (2010) Characterization of the xenobiotic response of *Caenorhabditis elegans* to the anthelmintic drug albendazole and the identification of novel drug glucoside metabolites. *Biochemical Journal*, **432**, 505–516.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lander,E.S. and Botstein,D. (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

Langeland,A. *et al.* (2020) NemChR-DB: a database of parasitic nematode chemosensory G-Protein Coupled Receptors. *International journal for parasitology*, **51**, 333.

Le,S.Q. and Gascuel,O. (2008) An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, **25**, 1307–1320.

Lee,D.F. *et al.* (2016) Mapping DNA polymerase errors by single-molecule sequencing. *Nucleic Acids Res*, **44**, e118.

Leinonen,R. *et al.* (2011) The Sequence Read Archive. *Nucleic Acids Res*, **39**, D19–D21.

Leonelli,S. and Ankeny,R.A. (2013) What makes a model organism? *Endeavour*, **37**, 209–212.

Leung,M.C.K. *et al.* (2008) *Caenorhabditis elegans*: An Emerging Model in Biomedical and Environmental Toxicology. *Toxicological Sciences*, **106**, 5.

Levy,A. and Currie,A. (2015) Model Organisms are Not (Theoretical) Models. *The British Journal for the Philosophy of Science*, **66**, 327–348.

Levy Karin,E. *et al.* (2020) MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, **8**, 48.

Lewin,H.A. *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A*, **115**, 4325–4333.

Lewis,S.E. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol*, **3**, RESEARCH0082.

Lex,A. *et al.* (2014) UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph*, **20**, 1983–1992.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv,* **1303**. doi: 10.48550/arXiv.1303.3997.

Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li,H. (2023) Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**, 2178–2189.

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**, 25–37.

Lian,Y. and Garner,H.R. (2005) Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. *Bioinformatics*, **21**, 1358–1364.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Lichten,M. and Goldman,A.S.H. (1995) Meiotic Recombination Hotspots. *Annual Review of Genetics*, **29**, 423–444.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289–293.

Lin,Y. *et al.* (2016) Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences*, **113**, E8396–E8405.

Liu,H. *et al.* (2021) SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte*, **2021**, 1–9.

Liu-Wei,W. *et al.* (2024) Sequencing accuracy and systematic errors of nanopore direct RNA sequencing. *BMC Genomics*, **25**, 528.

Logsdon,G.A. *et al.* (2020) Long-read human genome sequencing and its applications. *Nat Rev Genet*, **21**, 597–614.

Lomsadze,A. *et al.* (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, **33**, 6494–6506.

Ma,X. *et al.* (2021) Improved chromosome-level genome assembly and annotation of the seagrass, *Zostera marina* (eelgrass). *F1000Res*, **10**, 289.

Maduro,M. and Pilgrim,D. (1995) Identification and cloning of *unc-119*, a gene expressed in the *Caenorhabditis elegans* nervous system. *Genetics*, **141**, 977–988.

Maizels,R.M. *et al.* (2012) Immune modulation and modulators in *Heligmosomoides polygyrus* infection. *Experimental Parasitology*, **132**, 76–89.

Manni,M. *et al.* (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, **38**, 4647–4654.

Mapping,Genome. gov. (2023) Mapping. *Retrieved 3 May 2023.* https://www.genome.gov/genetics-glossary/Mapping.

Marçais,G. *et al.* (2018) MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, **14**, e1005944.

Markov,G.V. *et al.* (2015) The same or not the same: lineage-specific gene expansions and homology relationships in multigene families in nematodes. *J Mol Evol*, **80**, 18–36.

Marks,R.A. *et al.* (2021) Representation and participation across 20 years of plant genome sequencing. *Nat. Plants*, **7**, 1571–1578.

Marra,M.A. *et al.* (1998) Expressed sequence tags--ESTablishing bridges between genomes. *Trends Genet*, **14**, 4–7.

Marwal,A. and Gaur,R.K. (2020) Chapter 18 - Molecular markers: tool for genetic analysis. In, Verma,A.S. and Singh,A. (eds), *Animal Biotechnology (Second Edition)*. Academic Press, Boston, pp. 353–372.

Mascher,M. and Stein,N. (2014) Genetic anchoring of whole-genome shotgun assemblies. *Frontiers in Genetics*, **5,** 208. doi: https://doi.org/10.3389/fgene.2014.00208.

Masonbrink,R.E. *et al.* (2021) A chromosomal assembly of the soybean cyst nematode genome. *Molecular Ecology Resources*, **21**, 2407–2422.

Matoušková,P. *et al.* (2018) UDP-glycosyltransferase family in *Haemonchus contortus*: Phylogenetic analysis, constitutive expression, sex-differences and resistance-related differences. *Int J Parasitol Drugs Drug Resist*, **8**, 420–429.

McCarthy,T.C. and Sinal,C.J. (2005) Biotransformation. In, Wexler,P. (ed), *Encyclopedia of Toxicology (Second Edition)*. Elsevier, New York, pp. 299–312. doi: https://doi.org/10.1016/B0-12-369400-0/00137-X.

McCartney,A.M. *et al.* (2021) An exploration of assembly strategies and quality metrics on the accuracy of the rewarewa (*Knightia excelsa*) genome. *Mol Ecol Resour*, **21**, 2125–2144.

von Mende,N. *et al.* (1988) *dpy-13*: A nematode collagen gene that affects body shape. *Cell*, **55**, 567–576.

Ménez,C. *et al.* (2019) The transcription factor NHR-8: A new target to increase ivermectin efficacy in nematodes. *PLOS Pathogens*, **15**, e1007598.

Menzel,R. *et al.* (2001) A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. *Arch Biochem Biophys*, **395**, 158–168.

Menzel,R. *et al.* (2005) CYP35: xenobiotically induced gene expression in the nematode *Caenorhabditis elegans*. *Arch Biochem Biophys*, **438**, 93–102.

Mieszkowska,N. *et al.* (2022) Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, **119**. e2115642118.

Miga,K.H. *et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, **585**, 79–84.

Mitsuhashi,S. *et al.* (2019) Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol*, **20**, 58.

Mitsuhashi,S. and Matsumoto,N. (2020) Long-read sequencing for rare human genetic diseases. *J Hum Genet*, **65**, 11–19.

Mohamed,A.M. and Chin-Sang,I.D. (2011) The *C. elegans* nck-1 gene encodes two isoforms and is required for neuronal guidance. *Dev Biol*, **354**, 55–66.

Morrison,A.A. *et al.* (2022) Phenotypic and genotypic analysis of benzimidazole resistance in reciprocal genetic crosses of *Haemonchus contortus*. *International Journal for Parasitology: Drugs and Drug Resistance*, **18**, 1–11.

Mortazavi,A. *et al.* (2010) Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. *Genome Res*, **20**, 1740–1747.

Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol*, **396**, 59–70.

National Research Council (1988) Mapping and Sequencing the Human Genome. The National Academies Press, Washington, DC. doi: 10.17226/1097.

Nelson,D.R. (1998) Metazoan cytochrome P450 evolution1. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology*, **121**, 15–22.

Nevers,Y. *et al.* (2024) Quality assessment of gene repertoire annotations with OMArk. *Nat Biotechnol*, 1–10. doi: 10.1038/s41587-024-02147-w.

Nisbet,A.J. *et al.* (2013) Successful immunization against a parasitic nematode by vaccination with recombinant proteins. *Vaccine*, **31**, 4017–4023.

Nisbet,A.J. *et al.* (2024) Field testing of recombinant subunit vaccines against *Teladorsagia circumcincta* in lambing ewes demonstrates a lack of efficacy in the face of a multi-species parasite challenge. *Front. Parasitol.*, **3**. 360029. doi: 10.3389/fpara.2024.1360029.

Nurk,S. *et al.* (2020) HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*, **30**, 1291–1305.

Nuss,A.B. *et al.* (2018) Chicago and Dovetail Hi-C proximity ligation yield chromosome length scaffolds of Ixodes scapularis genome. *bioRxiv* 2018. doi: 10.1101/392126.

O'Brien,K.P. *et al.* (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, **33**, D476–D480.

Olson,M. *et al.* (1989) A common language for physical mapping of the human genome. *Science*, **245**, 1434–1435.

Osbourn,M. *et al.* (2017) HpARI Protein Secreted by a Helminth Parasite Suppresses Interleukin-33. *Immunity*, **47**, 739-751.e5.

Oyola,S.O. *et al.* (2012) Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics*, **13**, 1.

Packer,J.S. *et al.* (2019) A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science*, **365**, eaax1971.

Palfalvi,G. *et al.* (2020) Genomes of the Venus Flytrap and Close Relatives Unveil the Roots of Plant Carnivory. *Curr Biol*, **30**, 2312-2320.e5.

Palmer,J.M. and Stajich,J.E. (2023) Funannotate (Version 1.8.16) [Computer software]. https://github.com/nextgenusfs/funannotate.

Panfilio,K.A. *et al.* (2019) Molecular evolutionary trends and feeding ecology diversification in the *Hemiptera*, anchored by the milkweed bug genome. *Genome Biology*, **20**, 64.

Park,S. *et al.* (2023) Benchmark study for evaluating the quality of reference genomes and gene annotations in 114 species. *Front. Vet. Sci.*, **10**.

Parra,G. *et al.* (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Peck,K. *et al.* (1997) Restriction Mapping of Genes by Capillary Electrophoresis with Laser-Induced Fluorescence Detection. *Anal. Chem.*, **69**, 1380–1384.

Phillippy,A.M. (2017) New advances in sequence assembly. *Genome Res*, **27**, xi–xiii.

Phillips,P.C. (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, **9**, 855–867.

Pinard,R. *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.

Pollo,S.M.J. *et al.* (2020) Benchmarking hybrid assemblies of *Giardia* and prediction of widespread intra-isolate structural variation. *Parasites & Vectors*, **13**, 108.

Pollo,S.M.J. *et al.* (2023) Transcriptional patterns of sexual dimorphism and in host developmental programs in the model parasitic nematode *Heligmosomoides bakeri*. *Parasit Vectors*, **16**, 171.

Prichard,R. (2001) Genetic variability following selection of *Haemonchus contortus* with anthelmintics. *Trends Parasitol*, **17**, 445–453.

Pulst,S.M. (1999) Genetic linkage analysis. *Arch Neurol*, **56**, 667–672.

Pundir,S. *et al.* (2017) UniProt Protein Knowledgebase. In, Wu,C.H. *et al.* (eds), *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, Methods in Molecular Biology. Springer, New York, NY, pp. 41–55.

Putnam,N.H. *et al.* (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.*, **26**, 342–350.

Raschke,E. (1993) Comprehensive restriction enzyme lists to update any DNA sequence computer program. *Genet Anal Tech Appl*, **10**, 49–60.

Redman,E. *et al.* (2015) The emergence of resistance to the benzimidazole anthlemintics in parasitic nematodes of livestock is characterised by multiple independent hard and soft selective sweeps. *PLoS Negl Trop Dis*, **9**, e0003494.

Redman,M. *et al.* (2016) What is CRISPR/Cas9? *Arch Dis Child Educ Pract Ed*, **101**, 213–215.

Reynolds,L.A. *et al.* (2012) Immunity to the model intestinal helminth parasite *Heligmosomoides polygyrus*. *Semin Immunopathol*, **34**, 829–846.

Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.

Rhie,A. *et al.* (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, **21**, 245.

Rhie,A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.

Rhoads,A. and Au,K.F. (2015) PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, **13**, 278–289.

Ribeiro,F.J. *et al.* (2012) Finished bacterial genomes from shotgun sequence data. *Genome Res*, **22**, 2270–2277.

Roach,M.J. *et al.* (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, **19**, 460.

Robinson,M.W. *et al.* (2004) A possible model of benzimidazole binding to β-tubulin disclosed by invoking an inter-domain movement. *Journal of Molecular Graphics and Modelling*, **23**, 275–284.

Rodríguez-Vivas,R.I. *et al.* (2017) Potential economic impact assessment for cattle parasites in Mexico. Review. *Revista mexicana de ciencias pecuarias*, **8**, 61–74.

Roose,S. *et al.* (2021) Characterization of the β-tubulin gene family in *Ascaris lumbricoides* and *Ascaris suum* and its implication for the molecular detection of benzimidazole resistance. *PLOS Neglected Tropical Diseases*, **15**, e0009777.

Rosenberg,N.A. *et al.* (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Ross,M.G. *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biology*, **14**, R51.

Roubin,R. *et al.* (1999) let-756, a *C. elegans* fgf essential for worm development. *Oncogene*, **18**, 6741–6747.

Rousakis,A. *et al.* (2013) The general control nonderepressible-2 kinase mediates stress response and longevity induced by target of rapamycin inactivation in *Caenorhabditis elegans*. *Aging Cell*, **12**, 742–751.

Rual,J.F. *et al.* (2004) Toward Improving *Caenorhabditis elegans* Phenome Mapping With an ORFeome-Based RNAi Library. *Genome Res.*, **14**, 2162–2168.

Ruan,J. and Li,H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nat Methods*, **17**, 155–158.

Rutherford,K. *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Sacristán-Horcajada,E. *et al.* (2021) ARAMIS: From systematic errors of NGS long reads to accurate assemblies. *Brief Bioinform*, **22**, bbab170.

Sangster,N.C. and Dobson,R.J. (2002) Anthelmintic resistance. In, *The Biology of Nematodes*. Taylor & Francis, **22**, 531–567.

Saunders,G.I. *et al.* (2013) Characterization and comparative analysis of the complete *Haemonchus contortus* β-tubulin gene family and implications for benzimidazole resistance in strongylid nematodes. *Int J Parasitol*, **43**, 465–475.

Sayers,E.W. *et al.* (2020) GenBank. *Nucleic Acids Research*, **48**, D84–D86.

Sharma,N. *et al.* (2024) Multiple UDP glycosyltransferases modulate benzimidazole drug sensitivity in the nematode *Caenorhabditis elegans* in an additive manner. *International Journal for Parasitology*, **54**, 535-549. doi: 10.1016/j.ijpara.2024.05.003.

Shaver,A.O. *et al.* (2023) Variation in anthelmintic responses are driven by genetic differences among diverse *C. elegans* wild strains. *PLoS Pathog*, **19**, e1011285.

Sherry,S.T. *et al.* (1999) dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.*, **9**, 677–679.

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

Simon,A. (2010) FastQC: Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.

Simpkin,K.G. and Coles,G.C. (1981) The use of *Caenorhabditis elegans* for anthelmintic screening. *Journal of Chemical Technology and Biotechnology*, **31**, 66–69.

Slater,G.S.C. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Smyth,D.J. *et al.* (2018) TGF-β mimic proteins form an extended gene family in the murine parasite *Heligmosomoides polygyrus*. *International Journal for Parasitology*, **48**, 379.

Smythe,A.B. *et al.* (2019) Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC Evolutionary Biology*, **19**, 121.

Sonnhammer,E.L.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, **18**, 619–620.

Staden,R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*, **6**, 2601–2610.

Stanke,M. *et al.* (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Stasiuk,S.J. *et al.* (2019) Similarities and differences in the biotransformation and transcriptomic responses of *Caenorhabditis elegans* and *Haemonchus contortus* to five different benzimidazole drugs. *International Journal for Parasitology: Drugs and Drug Resistance*, **11**, 13–29.

Stein,L.D. *et al.* (2003) The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics. *PLOS Biology*, **1**, e45.

Stevens,L. *et al.* (2023) Ancient diversity in host-parasite interaction genes in a model parasitic nematode. *Nat Commun*, **14**, 7776.

Stevens,L. *et al.* (2020) The Genome of *Caenorhabditis bovis*. *Current Biology*, **30**, 1023-1031.e4.

Storer,J. *et al.* (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, **12**, 2.

Straiton,J. *et al.* (2019) From Sanger sequencing to genome databases and beyond. *Biotechniques*, **66**, 60–63.

Stroehlein,A.J. *et al.* (2018) Improved strategy for the curation and classification of kinases, with broad applicability to other eukaryotic protein groups. *Sci Rep*, **8**, 6808.

Stromberg,B.E. and Gasbarre,L.C. (2006) Gastrointestinal nematode control programs with an emphasis on cattle. *Vet Clin North Am Food Anim Pract*, **22**, 543–565.

Studer,R.A. and Robinson-Rechavi,M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet*, **25**, 210–216.

Sturtevant,A.H. (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, **14**, 43–59.

Sugimoto,A. (2004) High-throughput RNAi in *Caenorhabditis elegans*: Genome-wide screens and functional genomics. *DIFFERENTIATION*, **72**, 81–91.

Sulston,J.E. *et al.* (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*, **100**, 64–119.

Sulston,J.E. and Horvitz,H.R. (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev Biol*, **56**, 110–156.

Sun,J. *et al.* (2021) Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **376**, 20200160.

Susa,S.T. *et al.* (2024) Drug Metabolism. In, *StatPearls*. StatPearls Publishing, Treasure Island (FL). http://www.ncbi.nlm.nih.gov/books/NBK442023/.

Sutton,J.M. *et al.* (2021) Optimizing experimental design for genome sequencing and assembly with Oxford Nanopore Technologies. *GigaByte*, **2021**, gigabyte27. doi: 10.46471/gigabyte.27.

Szabo,E.K. *et al.* (2024) Evidence of opportunistic blood feeding in the parasitic nematode *Heligmosomoides bakeri*. *bioRxiv* 2024. doi: 10.1101/2024.10.09.617485.

Tamás,L. and Shewry,P.R. (2006) Heterologous expression and protein engineering of wheat gluten proteins. *Journal of Cereal Science*, **43**, 259–274.

Tamura,K. *et al.* (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, **38**, 3022–3027.

Tarailo-Graovac,M. and Chen,N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*, **Chapter 4**, 4.10.1-4.10.14.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Taubert,S. *et al.* (2010) Nuclear Hormone Receptors in Nematodes: Evolution and Function. *Molecular and cellular endocrinology*, **334**, 49.

Tettelin,H. *et al.* (2001) Complete Genome Sequence of a Virulent Isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

The Bovine HapMap Consortium (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science (New York, N.Y.)*, **324**, 528.

The *C. elegans* Sequencing Consortium (1998) Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science*, **282**, 2012–2018.

The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, **51**, D523–D531.

Theißen,G. (2002) Orthology: Secret life of genes. *Nature*, **415**, 741–741.

Thrash,A. *et al.* (2020) Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics*, **21**, 249.

Timilsena,P.R. *et al.* (2022) Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. *Front. Plant Sci.*, **13**.

Todd,B.D. *et al.* (2022) Reference Genome of the Northwestern Pond Turtle, *Actinemys marmorata*. *Journal of Heredity*, **113**, 624–631.

Troell,K. *et al.* (2006) Global patterns reveal strong population structure in *Haemonchus contortus*, a nematode parasite of domesticated ruminants. *Int J Parasitol*, **36**, 1305–1316.

Troemel,E.R. (1999) Chemosensory signaling in *C. elegans*. *Bioessays*, **21**, 1011–1020.

Ungar,R.A. *et al.* (2024) Impact of genome build on RNA-seq interpretation and diagnostics. *The American Journal of Human Genetics*, **111**, 1282–1300.

Vallender,E.J. (2017) Molecular Evolution and Phenotypic Change. In, Kaas,J.H. (ed), *Evolution of Nervous Systems (Second Edition)*. Academic Press, Oxford, **4**,101–119. doi: https://doi.org/10.1016/B978-0-12-804042-3.00108-1.

Vaser,R. *et al.* (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, **27**, 737–746.

Venkatesan,A. *et al.* (2023) Molecular evidence of widespread benzimidazole drug resistance in *Ancylostoma caninum* from domestic dogs throughout the USA and discovery of a novel β-tubulin benzimidazole resistance mutation. *PLOS Pathogens*, **19**, e1011146.

de Visser,J.A.G.M. *et al.* (2011) The causes of epistasis. *Proc Biol Sci*, **278**, 3617–3624.

Wajid,B. and Serpedin,E. (2012) Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers. *Genomics, Proteomics & Bioinformatics*, **10**, 58.

Walker,B.J. *et al.* (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, **9**, e112963.

Wang,J. *et al.* (2017) Comparative genome analysis of programmed DNA elimination in nematodes. *Genome Res*, **27**, 2001–2014.

Wang,J. *et al.* (2012) Silencing of germline-expressed genes by DNA elimination in somatic cells. *Dev Cell*, **23**, 1072–1080.

Wang,P. and Wang,F. (2023) A proposed metric set for evaluation of genome assembly quality. *Trends in Genetics*, **39**, 175–186.

Wang,Z. *et al.* (2017) Nuclear receptors: emerging drug targets for parasitic diseases. *The Journal of Clinical Investigation*, **127**, 1165.

Warner,J.F. *et al.* (2021) Chromosomal-Level Genome Assembly of the Painted Sea Urchin *Lytechinus pictus*: A Genetically Enabled Model System for Cell Biology and Embryonic Development. *Genome Biology and Evolution*, **13**, evab061.

Wasmuth,J. *et al.* (2008) On the Extent and Origins of Genic Novelty in the Phylum Nematoda. *PLOS Neglected Tropical Diseases*, **2**, e258.

Watson,M. and Warr,A. (2019) Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol*, **37**, 124–126.

Weirather,J.L. *et al.* (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res*, **6**, 100.

Weisman,C.M. *et al.* (2022) Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr Biol*, **32**, 2632-2639.e2.

White,R. *et al.* (2020) Extracellular vesicles from *Heligmosomoides bakeri* and *Trichuris muris* contain distinct microRNA families and small RNAs that could underpin different functions in the host. *Int J Parasitol*, **50**, 719–729.

Wick,R.R. *et al.* (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*, **20**, 129.

Wickham,H. (2016) Data Analysis. In, Wickham,H. (ed), *ggplot2: Elegant Graphics for Data Analysis*, Use R! Springer International Publishing, Cham, pp. 189–201. doi: 10.1007/978-3-319-24277-4_9.

Wilson,R.K. (1999) How the worm was won. The *C. elegans* genome sequencing project. *Trends Genet*, **15**, 51–58.

Wintersinger,J.A. *et al.* (2018) One species, two genomes: A critical assessment of inter-isolate variation and identification of assembly incongruence in *Haemonchus contortus*. *bioRxiv* 2018. doi: 10.1101/384008.

Wit,J. *et al.* (2021) Natural variation in *Caenorhabditis elegans* responses to the anthelmintic emodepside. *Int J Parasitol Drugs Drug Resist*, **16**, 1–8.

Wolstenholme,A.J. and Kaplan,R.M. (2012) Resistance to macrocyclic lactones. *Curr Pharm Biotechnol*, **13**, 873–887.

World Health Organization (2022) Soil-transmitted helminth infections. https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections.

Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*, **13**, 329–342.

Yates,D.M. *et al.* (2003) The avermectin receptors of *Haemonchus contortus* and *Caenorhabditis elegans*. *International Journal for Parasitology*, **33**, 1183–1193.

Yuan,Y. *et al.* (2020) Advances in optical mapping for genomic research. *Comput Struct Biotechnol J*, **18**, 2051–2062.

Zarowiecki,M. and Berriman,M. (2015) What helminth genomes have taught us about parasite evolution. *Parasitology*, **142**, S85.

Zhang,F. *et al.* (2014) CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum Mol Genet*, **23**, R40-46.

Zhou,C. *et al.* (2023) YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*, **39**, btac808. doi: https://doi.org/10.1093/bioinformatics/btac808.

Zhou,S. *et al.* (2007) Chapter 9 A Single Molecule System for Whole Genome Analysis. In, Mitchelson,K.R. (ed), *Perspectives in Bioanalysis*, New High Throughput Technologies for DNA Sequencing and Genomics. Elsevier, **2**, 265–300. doi: 10.1016/S1871-0069(06)02009-X.