Genome assembly variation and its implications for gene discovery in nematode species

Grace M. Mariene[1,2], James D. Wasmuth[1,2,*]

1. Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada

2. Host-Parasite Interactions Research Training Network, University of Calgary, Calgary, Alberta, Canada

* Corresponding author: jwasmuth@ucalgary.ca

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 popular 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 critical gene families like *cyp*, *gst*, *ugt*, and *nhr*. Despite broadly similar performance using various assembly metrics, comparisons of assemblies with a single species revealed underlying structural rearrangements and inconsistencies in gene content. This emphasizes the imperative for continuous refinement of genomic resources. Our findings advocate for a cautious and informed approach to genome assembly and annotation to ensure reliable and insightful genomic interpretations.

Introduction

Genome assembly software serves a critical, often underappreciated role in the life sciences. Its purpose is to reconstruct, *de novo*, 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 decrease cost of sequencing has empowered the launch of large BioGenome consortia which aim to sequence, assemble and annotation the genomes of as many animal and plant species as possible (Mieszkowska et al., 2022; Ebenezer 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 the 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 leading 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 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 single optimum: 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 (Pollo et al., 2020; Jung et al., 2020). There 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 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* (*C. elegans* Sequencing Consortium, 1998); at least 177 genome assemblies—generated from a broad range of assemblers—are available for at least 135 species

52 (WormBase Parasite release 18) (Howe et al., 2017). This availability of genomic information has led to
53 comparative genomic studies on various scales, much of it focused on the evolution of gene families (Wasmuth
54 et al., 2008; Coghlan et al., 2019; Baker et al., 2021). For this present study, in addition to understanding how
55 different genome assemblers perform with different nematode species, we wanted to look deeper at the
56 downstream consequences of the technical variation. It has been shown that the number of apparent lineage-
57 specific genes is inflated if datasets are generated by different gene finding algorithms (Weisman, Murray &
58 Eddy, 2022), motivated us to examine the effect of genome assembler heterogeneity on gene family
59 comparisons.

60 We examined assemblies from three nematode species—*Caenorhabditis bovis*, *Haemonchus contortus*, and
61 *Heligmosomoides bakeri*—which reflect different genomic features—size and heterozygosity—and sequencing
62 platforms—Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio).

63 *C. bovis* is associated with bovine parasitic otitis disease and is the only pathogenic member of its genus (Kreis,
64 1964). Its genome assembly is approximately 63Mb, making it the smallest *Caenorhabditis* and one of the
65 smallest nematode genomes so far sequenced (Stevens et al., 2020). *H. contortus* parasitizes small ruminants
66 (sheep and goats) and the emergence of drug resistant strains motivated the study of its genome to
67 understand mechanisms of resistance and develop new drugs and vaccines (Laing et al., 2013). The *H.*
68 *contortus* genome assembly has undergone multiple rounds of improvement—a long-read assembly was
69 scaffolded with optical mapping and manually curated—to give seven scaffolds, with a total length of ~283Mb
70 (Doyle et al., 2020). Finally, *H. bakeri* is an intestinal parasite of mice that serves as a model for studying host-
71 parasite interactions and testing anthelmintic efficacy (Hu et al., 2013; Colomb et al., 2024). Note that earlier
72 descriptions of the *H. bakeri* genome, were described as *Heligmosomoides polygyrus*, a long-standing miss-
73 classification of the study species which has been recently corrected through genomic comparisons (Chow et
74 al., 2019; Stevens et al., 2023).

75 We generated genome assemblies for each species using several popular genome assemblers and annotated
76 these assemblies with a single, widely used pipeline. We found that most programs returned assemblies with
77 encouraging and similar assembly metrics—size, number of scaffolds, gene completeness—but intra-species
78 comparisons revealed structural rearrangements in the scaffolds and diverse complements of conserved
79 orthologues. We explored the effect of assembly differences on annotation more deeply by carrying out a
80 phylogenomic comparison of large gene families—the cytochrome P450s (CYP), glutathione S-transferases
81 (GST), UDP glucuronosyltransferases (UGT), and nuclear hormone receptors (NHR)—and found that despite
82 having the same sequence reads as input, the assemblies' annotations could vary by as much as 51%.

83

84 # Methods

85 ## Nematode species and sequence data

86 DNA sequence reads for *C. bovis* and *H. contortus* were retrieved from the Sequence Read Archive, and their
87 genomes downloaded from WormBase Parasite (release 17) (Leinonen, Sugawara & Shumway, 2011; Howe et
88 al., 2017). The sequence reads and assembly for *H. bakeri* are new for this study. *C. bovis* DNA (BioProject:
89 PRJEB34497) was sequenced using the Oxford Nanopore MinION platform (Stevens et al., 2020). *H. contortus*
90 DNA (BioProject PRJEB506) was sequenced using the PacBio RS and Sequel platforms (Doyle et al., 2020). We
91 generated *H. bakeri* DNA as part of a *de novo* sequencing project. Briefly: we extracted adult *H. bakeri* worms
92 from experimentally infected mice; separated females and males; pooled 20 worms of each sex. PacBio Hi-Fi
93 sequencing was done at the University of Delaware DNA Sequencing & Genotyping Center. For the assemblies
94 in this study, we combined sequence reads from female and male worms. The sequence reads are available
95 from the NCBI SRA: SRX23938287 and SRX23938287.

96 ## Genome assembly and quality control.

97 ### *Generation of preliminary assemblies*

98 For *C. bovis* and *H. contortus* we used six long-read assemblers: Redbean v2.5 (Ruan & Li, 2020), Canu v2.1.1
99 (Koren et al., 2017), Flye v2.8.2 (Lin et al., 2016; Kolmogorov et al., 2019), SMARTdenovo v1.0.0 (Liu et al.,
100 2021), Falcon-kit v1.8.1 and Falcon-unzip v1.3.7 (Chin et al., 2016). For *H. bakeri*, we used three HiFi-compatible
101 assemblers: Flye v2.8.2 (Lin et al., 2016; Kolmogorov et al., 2019), Hifiasm v0.16.1-r375 (Cheng et al., 2021) and
102 HiCanu v2.1.1 (Nurk et al., 2020). All the commands and parameters can be found in File S1.

103 ### *Decontamination and polishing of preliminary assemblies*

104 For the *C. bovis* and *H. contortus*, we used Blobtools v1.1.1 (Laetsch & Blaxter, 2017). Long reads were aligned
105 to assembly contigs using minimap2 v2.17-r941 (Li, 2018) and the taxonomic classification determined by
106 searching contigs against the Uniprot Proteome database using DIAMOND v2.0.6.144 (Buchfink, Xie & Huson,
107 2015; Pundir, Martin & O'Donovan, 2017). For polishing—identification and correction of small errors—the *C.*
108 *bovis* and *H. contortus* assemblies underwent multiple rounds of polishing using tools appropriate for the
109 sequencing platforms used. Briefly:

110 1. Four rounds with Racon with long reads for both *C. bovis* and *H. contortus* assemblies (Vaser et al.,
111 2017),
112 2. One round with Medaka v1.3.2 (Oxford Nanopore Technologies., 2018) with long reads for *C. bovis*
113 assemblies (https://github.com/nanoporetech/medaka),
114 3. One round with Arrow v2.3.3 with long reads for *H. contortus* assemblies (installed via the
115 GenomicConsensus package v2.3.3, http://github.com/PacificBiosciences/pbbioconda),

4. Two rounds of Racon and Pilon v1.14 Assemblies with Illumina reads for both *C. bovis* and *H. contortus* assemblies (Walker et al., 2014; Vaser et al., 2017).

All the commands and parameters can be found in File S1.

Removal of residual haplotype duplications

This was done for all *H. contortus* and *H. bakeri* assemblies generated as part of this study. 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 (Table S10). All the commands and parameters can be found in File S1.

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, length of the longest contig, and BUSCO v4.1.4 (Simão et al., 2015). All the commands and parameters can be found in File S1. Briefly, we used BUSCO v4.1.4 to evaluate completeness of single copy orthologues (Manni et al., 2021) and Inspector v1.0.2 to identify both large- and small-scale errors (Chen et al., 2021).

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 of *C. bovis*; Doyle for *H. contortus*; Hifiasm for *H. contortus*. For *C. bovis*, we used the dnadiff script (part of Nucmer) to report one-to-one alignment coordinates and viewed the alignments with Cirocs 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). All the commands and parameters can be found in File S1.

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; R Core Team, 2022). 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/).

Functional annotation of BUSCO genes

We used BlastKOALA webserver (Kanehisa, Sato & Morishima, 2016).

146 ## Genome annotation

147 ### *Soft masking assemblies*

148 We used RepeatModeler v2.0.3 and RepeatMasker v4.1.2-p1 (Tarailo-Graovac & Chen, 2009; Flynn et al., 2020).
149 For RepeatModeler, we used NCBI/RMBLAST v2.11.0+ to create repeat library from the Dfam database of
150 transposable elements and DNA repeats (Storer et al., 2021). Next, we used the RepeatMasker util
151 (queryRepeatDatabase.pl script) to create a custom library specific to nematode species. Both repeat libraries
152 were combined as input for the RepeatMasker program.

153 ### *Gene prediction*

154 Protein coding genes (gene models) in the generated assemblies were predicted using Braker 3 (Gabriel et al.,
155 2023) installed using the Singularity container (available at https://hub.docker.com/r/teambraker/braker3).
156 For *H. contortus* and *H. bakeri*, we used short-read RNA-seq data from the Sequence Read Archive as external
157 evidence for the gene prediction (BioProjects: PRJEB506 (Laing et al., 2013) and PRJNA750155 (Pollo et al.,
158 2023) respectively). RNA-seq data was aligned using the STAR v2.7.10a aligner (Dobin et al., 2013) and the
159 generated alignments, in bam-format, used to produce a training set for AUGUSTUS v3.5.0, using GeneMark-ET
160 v4.71 (Hoff et al., 2016). We filtered redundant training gene structures with DIAMOND v0.9.24 (Buchfink, Xie
161 & Huson, 2015). For *C. bovis*—for which no RNA-Seq is available—we used the longest isoforms from *C. elegans*
162 gene models (BioProject: PRJNA13758) retrieved from WormBase Parasite (release 17) (Howe et al., 2017), and
163 the default protein sequences retrieved from UniProtKB/Swiss-Prot database (Boutet et al., 2007) as the
164 protein evidence for gene prediction. Braker 3 used the ProtHint pipeline to produce hints using GeneMark-EP
165 v4.71, DIAMOND v0.9.24 and Spaln v2.3.3d tools (Lomsadze et al., 2005; Iwata & Gotoh, 2012; Gotoh, Morita &
166 Nelson, 2014; Buchfink, Xie & Huson, 2015; Brůna, Lomsadze & Borodovsky, 2020). All the commands and
167 parameters can be found in File S1.

168 ### *Identification of gene family members*

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

174 For the TGM protein family, the protein sequences were taken from the publication (Smyth et al., 2018). The
175 latest ARI and BARI protein sequences from provided by Dr. Henry McSorley (University of Dundee) (Osbourn
176 et al., 2017). These protein sequences were mapped to the genome assemblies using the splice-aware aligners
177 exonerate and miniprot (Slater & Birney, 2005; Li, 2023). For both, we a scoring threshold which only accepted
178 an alignment which scored within a specified percentage of the top score for a given query. For miniprot we

179 used '--outs 90', and for exonerate we used '--percent 90' and '--percent 70'. We found that the results for
180 'exonerate --percent 70' were broadly similar to 'miniprot --outs 90' (Table S69).

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 intraspecies comparisons. All the assemblies are available in File S2. 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; Stevens et al., 2020; Doyle 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 (Tables S7-S9). 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 were found as complete (Table S7). However, polishing had a minor impact on the PacBio-based assemblies, *H. contortus* (CSII) and *H. bakeri* (HiFi), presumably because the genome assemblers used had their own in-built polishing steps. 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% (Table S8).

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.

C. 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; Lin et al., 2016; Koren et al., 2017; Kolmogorov et al., 2019; Ruan & Li, 2020; Liu et al., 2021). Following decontamination and polishing (Table S7), the overall standard assembly metrics were relatively similar (Table 1) . Compared to the Stevens assembly, the Redbean2.5 assembly was near 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.

211 Through BUSCO analysis, we found a similar percentage of genes were complete across the seven assemblies
212 (Table 1 and Tables S11-S17). A notable difference was the duplication rate in the Canu assembly, which was
213 five times that of the Stevens assembly: 18 genes vs 97 genes.

214

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	C:96.0 S:95.3 D:0.7	C:96.4 S:95.5 D:0.7	C:96.3 S:93.2 D:3.1	C:96.1 S:93.2 D:0.9	C:96.1 S:95.0 D:1.1	C:96.2 S:95.1 D:1.1
Inspector QV	25.1	25.1	25.0	25.3	25.1	25.0	25.2
Structural errors	12	8	9	10	16	21	9

215

216 **Table 1**: **Assembly statistics for *Caenorhabditis bovis.*** Complete 'C' gene the sum of single copy 'S' and
217 duplicate 'D' (all % from 3131 genes).

218

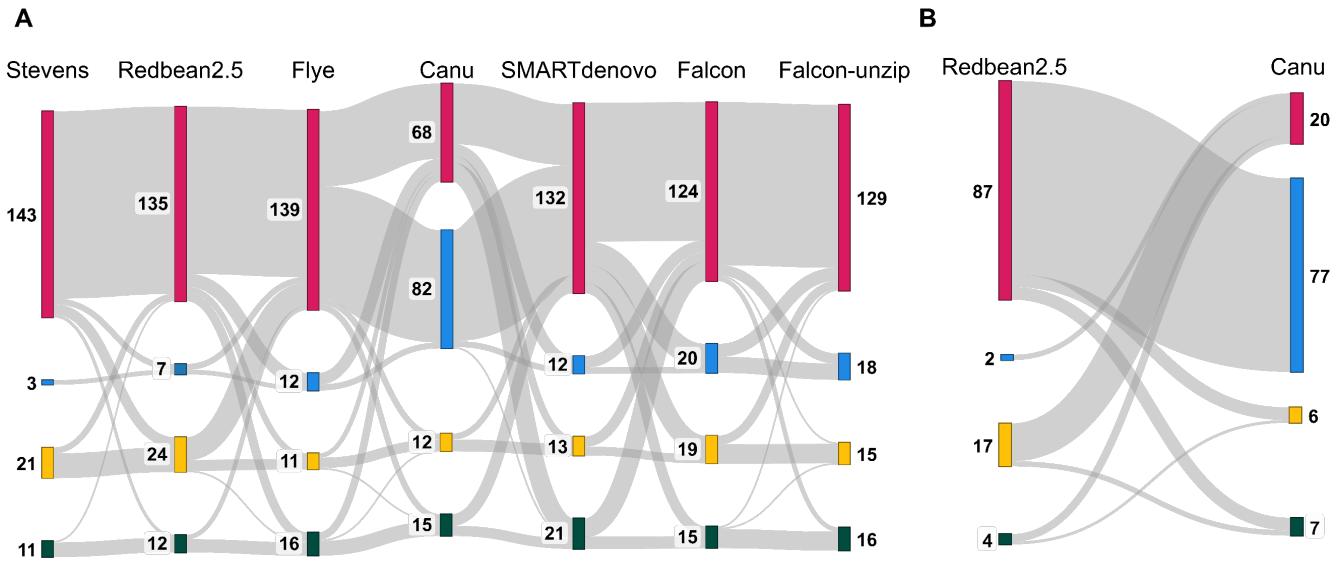
219 Of the 3131 BUSCO orthologues, 3040 genes (97%) were complete (single copy or duplicated) in at least one
220 assembly and 94.9% (2972) were complete in all seven assemblies (Figure S1). Of these complete genes, 15
221 were duplicated in all assemblies. Ninety-one genes were missing or fragmented in all seven assemblies, which
222 suggests they are undergoing pseudogenization or have been lost as a consequence of a genome size reduction
223 in *C. bovis* (Stevens et al., 2020). Ninety-one (3%) were absent (fragmented or missing) in all seven assemblies.
224 Six genes were complete in only one assembly. In total, the BUSCO classification differed for 178 genes, *e.g.*
225 from Duplicated to Fragmented (Figure 1A). Expectedly, there was strong agreement between the Stevens
226 (Redbean2.3) and the new Redbean2.5 assemblies; nine genes complete and single in the Stevens assembly
227 were fragmented or missing in the Redbean2.5 assembly with six genes in the reverse situation (Figure 1A).
228 Disagreements on BUSCO classification was more striking between the other assemblies. For example,
229 SMARTdenovo and Falcon had near identical BUSCO scores, but 36 single copy genes in the SMARTdenovo
230 assembly were classified as duplicated (16), fragmented (13) or missing (7) in the Falcon assembly. In a direct
231 comparison of Redbean2.5 and Canu assemblies, 18 genes considered fragmented or missing in Redbean2.5
232 but were found complete and single copy in Canu (Figure 1B).

233

234

235

236



237

238 **Figure 1: Transitions of BUSCO genes across multiple *C. bovis* assemblies.** The Sankey diagrams show how
239 BUSCO genes switch categories: red – complete and single copy genes; blue – complete and duplicated; yellow
240 – fragmented; green – missing. (A) The 178 genes that changed BUSCO categories in at least one assembly. The
241 order of the assemblies is arbitrary and follows Table 1. The data for this panel is in Table S1 and Table S2. (B)
242 Comparison of the 110 genes which differed between Redbean2.5 and Canu assemblies. The data for this panel
243 is in Table S3.

244

245 We used the original ONT reads to evaluate the reads alignment to each of the assemblies, identifying large
246 structural and small errors (Chen et al., 2021). All the assemblies had near identical QV scores, the Redbean2.5
247 assembly had the fewest large structural errors, and the Flye assembly had the fewest small-scale errors (Table
248 1 and Table S18).

249 Next, we investigated the structural variation—insertions, deletions, and translocations—between assemblies.
250 We needed a reference assembly, and considered Redbean2.5 assembly to have the best accuracy metrics,
251 albeit slightly. As expected, the Stevens and Flye assemblies—both use the DBG algorithm—had the fewest
252 number of differences (Table 2 and Table S19). These were followed by the OLC-based SMARTdenovo
253 assembly, the related HGAP Falcon and Falcon-Unzip assemblies, with the OLC-based Canu assembly having the
254 most differences to the Redbean2.5 assembly. For all assemblies, at least 97% of the variants were insertions
255 or deletions of typically short regions (1-to-10bp). However, we did find longer duplications and tandem
256 duplications which could impact accurate curation of genes. Using Braker3 to annotate the assemblies with
257 gene models, we found five genes from the Stevens assembly—four genes in tandem duplications and one

258 gene on a duplication—and 19 genes from the Redbean2.5—17 genes in tandem duplications and two on
259 duplications—in these regions (Table S20).

	Stevens	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
Translocation	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

260 **Table 2: NucDiff results for *C. bovis* assemblies.** Each assembly was aligned to the reference assembly,
261 Redbean2.5.

262 Larger structural variants—translocations, relocations, reshufflings, and inversions—were present even
263 between Redbean assemblies; we found that the RedBean2.5 Scaffold 3 was split across the Stevens Scaffolds
264 6, 7, and 11 (Figure 2). However, for five of the assemblies, all scaffolds could be aligned to the reference. The
265 exception was the Canu assembly; two scaffolds did not align to the Redbean2.5 assembly. Through a BLASTX
266 search against NCBI NR database (Camacho et al., 2009), we found that these were most likely of bacterial
267 origin and were not removed by the decontamination protocols.

268

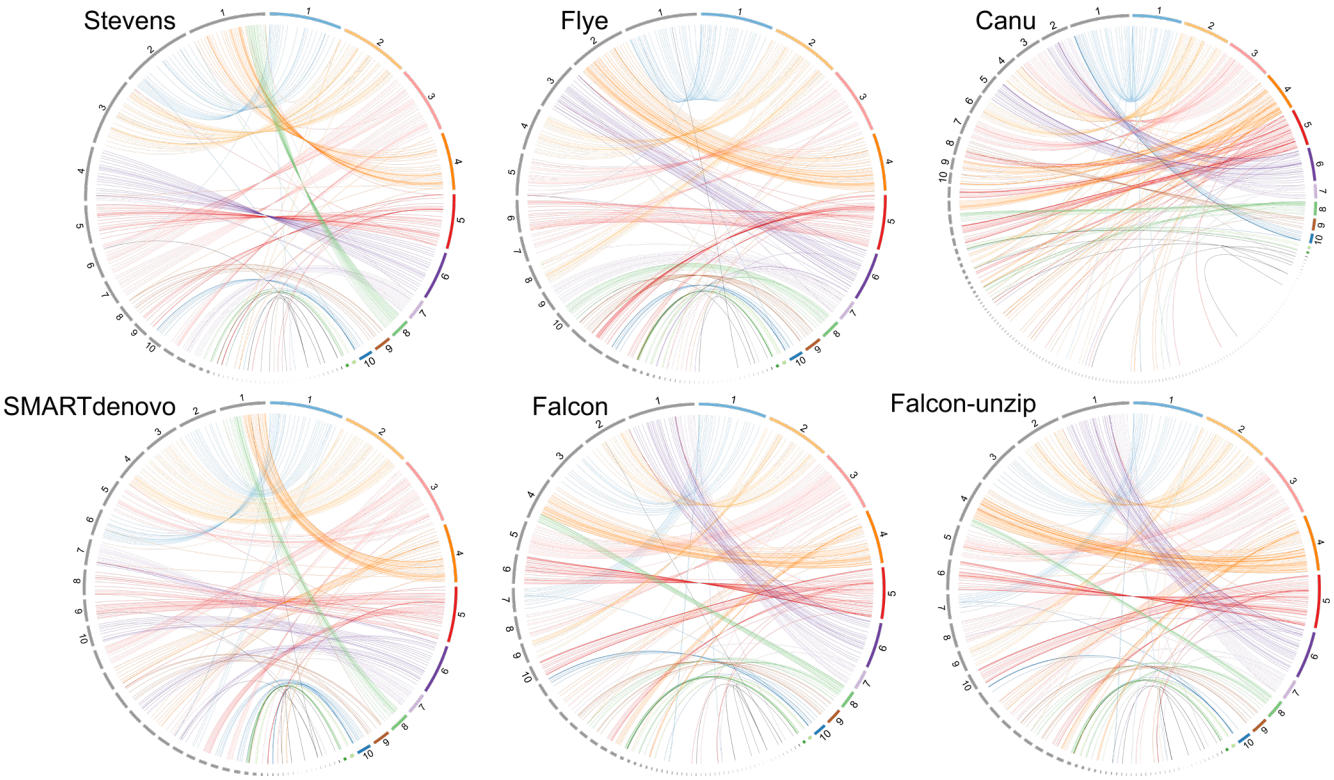


Figure 2: Synteny across six genome assemblies for the *C. bovis*. In each Circos plot, the reference assembly (Redbean2.5) is represented by the coloured scaffolds and the labeled assembly is represented by the gray scaffolds. The lines linking syntenic regions are coloured according to the reference scaffold. The scaffolds are ordered by length (longest to shortest) and for clarity we only label the first ten scaffolds.

H. 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 Doyle assembly. These six assemblies, once decontaminated and polished, varied considerably in their length, fragmentation, and completeness (Table 3). There was a 23% difference between the longest (Flye 331 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.

	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	C:94.8 S:93.7 D:1.1	C:93.9 S:86.7 D:7.2	C:94.1 S:85.0 D:9.1	C:95.2 S:86.6 D:8.6	C:94.7 S:86.5 D:8.2	C:88.2 S:84.2 D:4.0	C:87.7 S:84.2 D:3.5
Inspector Reference-free (mapping reads-to-assembly):							
Inspector QV	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

Table 3: Assembly statistics for *H. contortus*. Complete 'C' gene the sum of single copy 'S' and duplicate 'D' (all % from 3131 genes).

For the Inspector reports, we used both the reads and the Doyle assembly for reference-free and reference-guided metrics, respectively (Table 3 and 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 (Table S21). For the six assemblies that we generated, the Inspector metrics were similar. The coverage of the assembly-to-Doyle alignments were relatively low (79%-to-84%). The effect of the alignment algorithm is demonstrated with the 98% coverage Doyle-to-Doyle assembly. Therefore, the regions not covered predominantly are due to differences between each of the six assemblies and the Doyle reference. 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.

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 (Figures 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 at least one of the new assemblies, 27 in all six assemblies (Figure 3). It is likely that these represent errors in the current reference assembly.

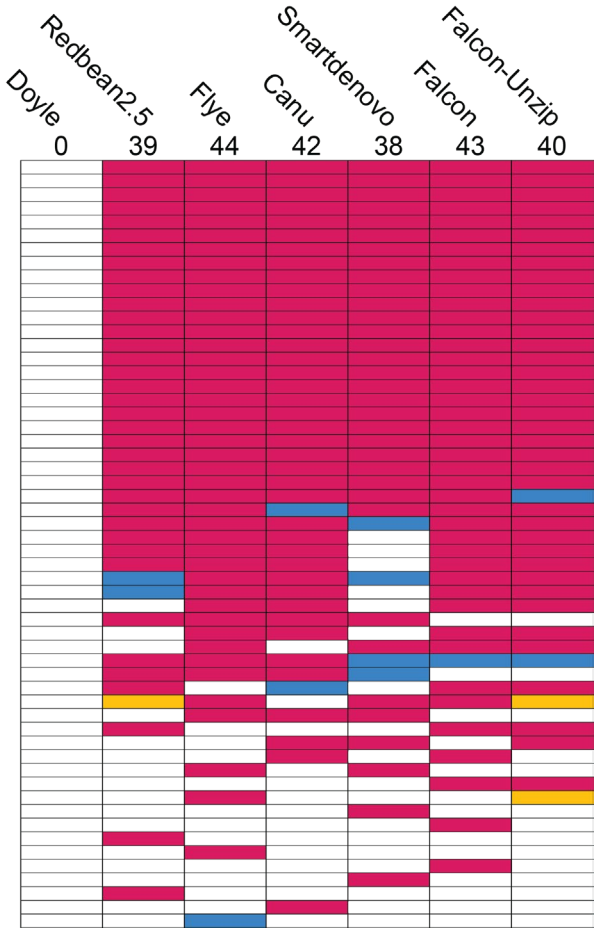


Figure 3: BUSCO genes missing from the *H. contortus* Doyle assembly but complete in at least one other assembly. Red – complete and single copy genes; blue – complete and duplicated; yellow – fragmented; white – missing. The data for this figure is in Table S6.

We used BlastKOALA to annotate these 141 genes and found at least three noteworthy examples (Table S22) (Kanehisa, Sato & Morishima, 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 in axon guidance, neuronal cell positioning, and abnormalities in the excretory canal cell, gonad, and male mating (Mohamed & 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 severally 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 (Figure 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 S5 and Table S4). 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 S5). Another example is, 62 genes considered missing or fragmented in the Redbean2.5 assembly single copy in Flye (Figure S5), with 66 genes reclassified in the other direction (Figure S6 and Table S5). Both assemblers use the DBG algorithm. The full list of BUSCO genes for each assembly can be found in Tables S23-S29).

H. 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 (Coghlan et al., 2019; Chow et al., 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 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 4).

Assembly	Chow	Flye	HiCanu	Hifiasm	Stevens *et al.* (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	C:94.3 S:93.1 D:1.2	C:94.6 S:92.5 D:2.1	C:95.0 S:92.2 D:2.8	C:94.9 S:92.0 D:2.9	C:93.8 S:92.6 D:1.2
Inspector QV	22.1	36.0	35.5	32.5	-
Structural errors	34	116	74	268	-

Table 4: Assembly statistics for *H. bakeri*. Complete 'C' gene the sum of single copy 'S' and duplicate 'D' (all % from 3131 genes). The chromosome-level Stevens *et al.* assembly was only published as we completed our study and was not included in our comparisons.

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 4 and S30).

Similar to *C. bovis* and *H. contortus*, the *H. bakeri* assemblies, despite extremely similar BUSCO scores, contained different cohorts of BUSCO genes (Figure 4 and Table S31). A total of 120 genes were complete in three assemblies and absent (fragmented or missing) in one (Figure 5A. We note that while 58 of these in-three-out-one 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 (Tables S32-S38).

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 embryoic lethal (Rual et al., 2004; Hirani et al., 2016).

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 in the HiFi-based assemblies (Figure 5B). This tempts speculation that the ~29 genes are indeed duplicated.

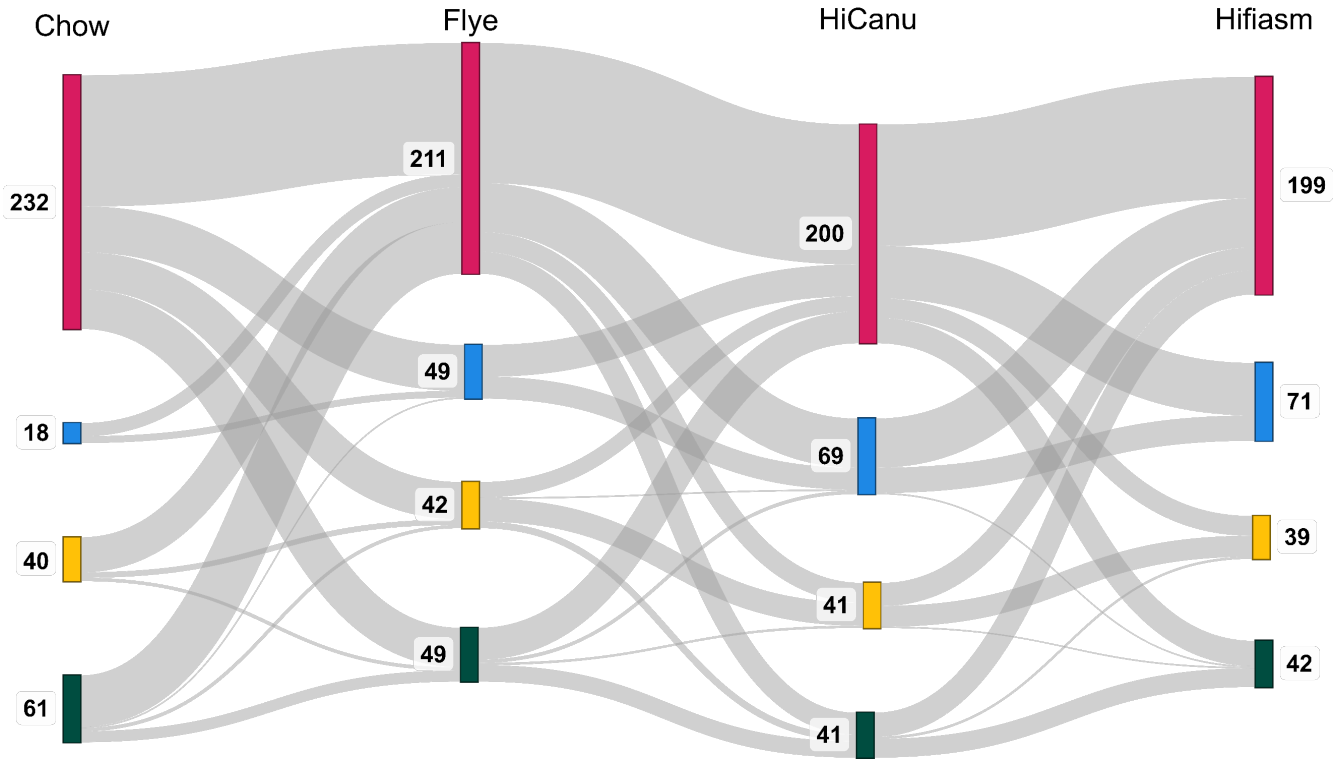


Figure 4: Transitions of BUSCO genes across multiple *H. bakeri* assemblies. The Sankey diagram shows how 351 genes changed categories in at least one *H. bakeri* assembly: red – complete & single copy genes; blue – complete & duplicated; yellow – fragmented; green – missing. The order of the assemblies is arbitrary and follows Table 4. The data for this figure is in Table S31.

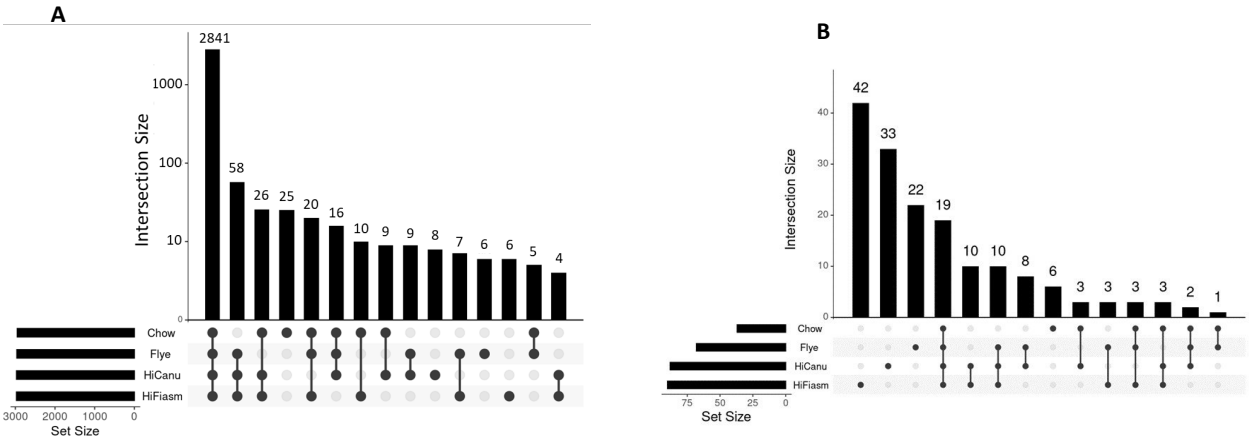


Figure 5: UpSet plots of (A) Complete BUSCOs present in *H. bakeri* assemblies, and (B) Duplicated BUSCOs in the assemblies. The figures were generated with https://asntech.shinyapps.io/intervene/.

391 ## Impact of assembly variation on gene family analysis

392 We generated protein-coding gene models for all assemblies using Braker3 (Table 5 and File S2) (Gabriel et al.,
393 2023). For consistency, we used Braker3-generated gene models for the published genomes (see Discussion).

Species	Assembly	Number of protein-coding genes	BUSCO % n=3131
C. elegans	WS285	19,981	C:99.7 [S:99.3, D:0.4], F:0.1, M:0.2
C. 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
H. 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
H. 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

394

395 **Table 5:** Protein-coding genes predicted for the different assemblies. For the published genomes, we report the
396 both the published protein-coding gene models and our reannotation. For all other assemblies, the protein-
397 coding gene models were generated with Braker3. BUSCO scores (C=complete, S=single copy, D=duplicated,
398 F=fragmented, M=missing) are percentages where *n*=3131.

399 For all protein sets, we identified members of four gene families—cytochrome P450s (CYP), glutathione S-
400 transferase (GST), UDP glucuronosyltransferases (UGT), and nuclear hormone receptors (NHR)—which have
401 been implicated in anthelmintic studies and are among the largest in *C. elegans* (Tables S39-S43). For each
402 gene family, within each species, we clustered the proteins with those from C. elegans (Emms & Kelly, 2019).
403 Here, our assumption is that if the differences in genome assemblies did not affect the genome annotation, we
404 would expect that every cluster would have an identical number of proteins from each assembly and that these
405 would form perfect one-to-one orthologous groups to the exclusion of *C. elegans*. Any deviation from this

406 would demonstrate that differences in the assemblies led to different gene model predictions. We considered
407 cluster membership across all assemblies ('all') and just for those assemblies generated in this study ('new')
408 (Table 6).

409

	CYP	**GST**	**UGT**	**NHR**
C. 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	all: 26 new: 26	all: 32 new: 32	all: 9 new: 10	all: 120 new: 122
H. 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	all: 14 new: 14	all: 12 new: 13	all: 8 new: 9	all: 56 new: 57
H. 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	all: 17 new: 20	all: 19 new: 24	all: 9 new: 11	all: 86 new: 100

410

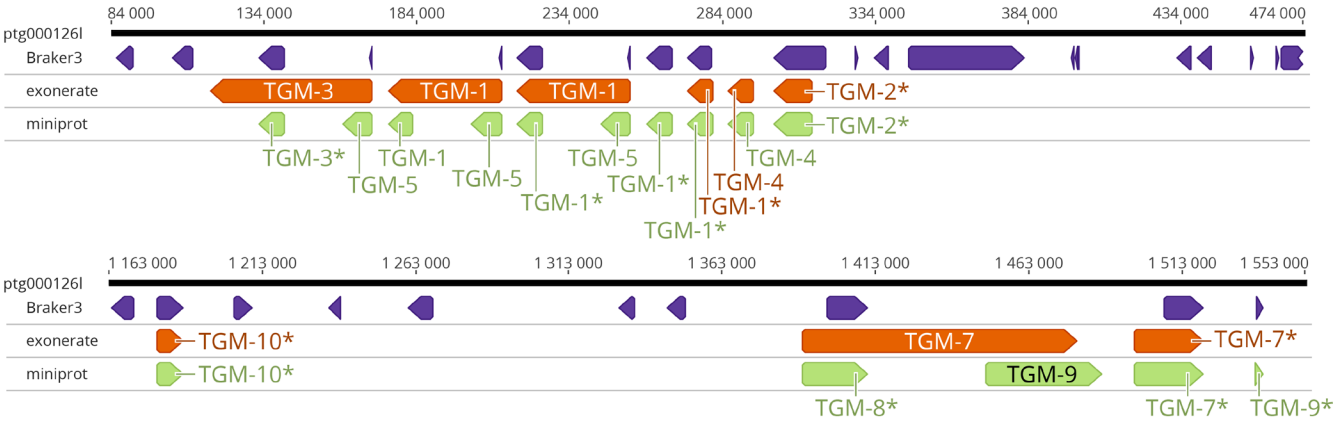
411 **Table 6:** Showing percentage differences between the minimum and maximum number of the predicted
412 members of the CYP, GST, UGT and NHR gene families across assemblies. We considered cluster membership
413 across all assemblies ('all') and also restricted to those assemblies generated in this study ('new').
414

415 Overall, we found considerable variation within gene family membership (Table 6 and Tables S45-S68). We
416 looked at the correlation between the size of the gene families and the size of the assembly or the BUSCO
417 duplication rate. The results were inconsistent. We saw a strong and positive correlated (ρ>0.7) for: GSTs,
418 UGTs, and NHRs in *C. bovis* against both assembly size and duplication rate; and NHRs in *H. contortus* against
419 both assembly size and duplication rate. However, we saw a strong and negative correlation (ρ<-0.7) for CYPs,
420 GTS, NHRs in *H. bakeri* against the duplication rate (Table S44). The other comparisons could not be considered
421 noteworthy ($|\rho|$<0.7). Across the 628 clusters that we generated, approximately 60% contained the same
422 number of members for each assembly. Put another way, ~40% of clusters contained a variable number of
423 proteins and often this variation was an additional member in just one assembly (Tables 6 and S45-S68).

424 We provide two examples of variable cluster memberships from the CYPs. The first is cluster bovis-cyp-
425 0000003 (Table S45 and Table S46), in which: *C. elegans* and the Redbean2.5 assembly have one member each;
426 the Stevens, Canu and Flye assemblies have two members each; the Falcon and Falcon-unzip assemblies have 3

427 members each. The *C. elegans* protein is CYP-33C9 whose function has been associated with fatty-acid
428 desaturation hence reducing susceptibility of xenobiotics by increasing desiccation tolerance (Erkut et al.,
429 2013). The second example is cluster contortus-cyp-0000026, in which: *C. elegans*, in which: *C. elegans* and the
430 Flye, Falcon, and Falcon-unzip assemblies have one member each; the Doyle, Canu, Redbean2.5, and
431 SMARTdenovo assemblies have no members (Table S47 and Table S48). The *C. elegans* protein is CYP-43A1 and
432 is highly expressed in the intestine and serotonergic neuron ADF (Packer et al., 2019). Interestingly, an
433 orthologue to C. elegans CYP-43A1 was found in an earlier version of the *H. contortus* assembly and shown to
434 be the most highly expressed CYP in the *H. contortus* intestine (Laing et al., 2015).

435 In *H. bakeri*, three protein families—TGF-b mimics (TGM), ARI, and BARI—have been shown to modulate the
436 host's immune system (Osbourn et al., 2017; Johnston et al., 2017; Smyth et al., 2018). The available protein
437 sequences for these families aligned poorly to the *H. bakeri* Braker3 gene models (data not shown). We
438 determine the likely locations of these proteins on the genome assemblies using two splice-aware aligners,
439 exonerate and miniprot (Slater & Birney, 2005; Li, 2023). Similar to the other protein families, we found
440 variation in gene counts across the three HiFi long-read assemblies: nine-to-16 TGMs and 13-to-23 ARIs/BARIs
441 (Table 7 and Table S69). We checked whether the protein-to-genome alignments overlapped with Braker3
442 gene models and found that while most alignments (TGM 81-100%; ARI/BARI 68-100%) overlapped with a gene
443 model, far fewer alignments (TGM 40-91%; ARI/BARI 27-67%) overlapped gene models across by greater than
444 50% of their length (Table 7). Focusing on scaffold ptg000126l from the Hifiasm assembly, we saw a full set of
445 examples, from complete congruence between Braker3, exonerate, and miniprot, to evidence from just one
446 source (Figure 6).



447

448 **Figure 6: Location of TGM protein-coding genes on *H. bakeri* scaffold ptg000126l.** TGM genes were found to
449 cluster in two locations on this scaffold. The purple track is the Braker3 gene model predictions, the orange
450 track is the Exonerate protein-to-genome alignments, the green track is the miniprot protein-to-genome
451 alignments. For both exonerate and miniprot alignments, an * signifies ≥50% reciprocal overlap with a Braker3
452 gene model. The black bar is the scaffold, with nucleotide coordinates given above the bar.

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**

453 **Table 7: Immunomodulators annotated in *H. bakeri* assemblies.**

454 ## Discussion

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

464 Long-read assemblers predominantly use two core algorithmic frameworks: Overlap-Layout-Consensus (OLC)
465 and De-Bruijn Graph (DBG) (reviewed in (Li et al., 2012)). Each framework is decorated with a myriad of
466 developer adaptations to address specific genome features or challenges, such as repetitive sequences, varying
467 levels of ploidy, and differential sequencing error profiles. The process of implementing and optimising these
468 adaptations is not trivial; they significantly affect the assemblers' effectiveness, especially in regions exhibiting
469 intra-population structural complexities or high heterogeneity.

470 Arguably, the least complex genome in this study was *C. bovis* and DBG-based assemblers—Redbean2.5 and
471 Flye—performed best, yielding less fragmented assemblies and higher BUSCO scores. They also did better with
472 the more granular measures of assembly performance: read mapping rates, alignment depths, and structural
473 errors (large and small). Determining the best assemblies for *H. contortus* was less obvious. Here, OLC-based
474 assemblers—SMARTdenovo and Canu—achieved larger N50s and marginally better BUSCO scores. The
475 Redbean2.5 (DBG) assembly had the lowest BUSCO duplication rate and fewest structural errors. For *H. bakeri*,
476 the Hifiasm assembler, produced the most contiguous assembly but according to the Inspector report displayed
477 the highest number of structural errors. This suggests potential challenges in handling highly heterozygous
478 regions by the assembly algorithm, despite its default haplotype-resolving capabilities.

479 From our reading of the literature, the BUSCO score may be the most widely used metric to determine
480 assembly accuracy, and we too relied heavily on it in the presented analysis. Often, we found that the scores
481 were largely similar between assemblers, suggesting that the gene complements of each genome would be
482 similar. However, diving deeper into the classification of the 3131 nematode single copy orthologous
483 demonstrated a surprising situation, where a gene may be found as a single copy in half the assemblies and
484 missing in another half. We saw differences between Redbean versions 2.3 and 2.5 which should motivate

485 researchers to carefully consider updated annotations; that is to ask, what gets lost when you improve your
486 annotation? We found at least 27 genes which were potentially incorrectly absent in the published *H. contortus*
487 assembly. This is one of the most carefully curated parasite genome assemblies and demonstrated the need to
488 fund ongoing curation for important species (Doyle et al., 2020; Doyle, 2022).

489 The reliance on BUSCO scores is that small variations in score can be considered trivial. For example, 1% change
490 in nematodes is 31 genes. Does this matter? It is important to remember that BUSCO genes are the most
491 conserved genes in an organism. They typically perform critical functions and their absence from an assembly, if
492 true *in vivo*, would point to a significant change in an organism's biology. However, If the absence is due to a
493 technical problem correctly assembling the reads, one might assume that assembly of regions containing less
494 well conserved single copy genes would be more error-prone, particularly for multi-copy gene families. This is
495 what we find with the *cyp*, *gst*, *ugt*, and *nhr* gene family. Comparative analyses often involve species assembled
496 and annotated using different methods across different laboratories (Coghlan et al., 2019). This annotation
497 heterogeneity inflates the number of lineages-specific genes (Weisman, Murray & Eddy, 2022). An alternative is
498 to reannotate the assemblies with a uniform approach, as we did here. A limitation of this approach, and our
499 study, is that we then ignore expert curation. For example, our reannotation of the Doyle *H. contortus* assembly
500 resulted in approximately 4,800 genes then the published annotation, and a lower BUSCO score. In contrast,
501 our reannotation of the published *C. bovis* and *H. bakeri* assemblies resulted in more genes, and importantly,
502 improved BUSCO scores. Together, this demonstrates the requirement to validate the absence of a gene, which
503 may actually be present in the assembly but score just the wrong side of a gene finder's score cut-off (Gilabert
504 et al., 2016; Stroehlein, Young & Gasser, 2018; Baker et al., 2021).

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

513 Determining the usefulness of an assembly will remain a personal choice; one that should be guided by a range
514 of metrics. It is noteworthy that in a comparison of assemblies generated for several invertebrates, the
515 assemblies with the highest BUSCO score were sometimes the most fragmented (Sutton et al., 2021).

516 ## Conclusion

517 We encourage those charged with assembling genomes—be it *Heligmosomoides*, *Haemaphysalis*, or
518 *Henneguya*—to be aware of the important differences between seemingly very similar assemblies. For users of
519 genome assemblies and annotation, we emphasize the importance of exploring alternative assemblies
520 generated by multiple assemblers to validate biological hypotheses. Finally, we note that much of the above
521 discussion can be summed up neatly by the first recommendation from Assemblethon2: "Don't trust the results
522 of a single assembly (Bradnam et al., 2013)."

523

524 ## Acknowledgements

525 We acknowledge the high-performance computing resources made available by the Faculty of Veterinary
526 Medicine and Research Computing at the University of Calgary. We thank Drs. Stephen Doyle, John Gilleard,
527 Graham Plastow, John Soghigian, and Frank van der Meer for discussions around this topic.

528 ## Competing Interests

529 The authors declare there are no competing interests.

530 ## Author contribution

531 GMM: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Visualization;
532 Writing – original draft; Writing – review & editing

533 JDW: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Supervision;
534 Validation; Visualization; Writing – original draft; Writing – review & editing

535 ## Funding

536 This work was supported by grants to JDW from Results Driven Agricultural Research (RDAR, Alberta) and the
537 Natural Sciences and Engineering Research Council of Canada (NSERC).

538 ## Data availability

539 Sequence reads for the *Heligmosomoides bakeri* genome are available on the sequence read archive (SRA):
540 SRX23938287 and SRX23938287. File S2 contains the genome assemblies and Braker3 gene models generated
541 in this study and is available on Datadryad (doi.org/10.5061/dryad.p2ngf1vzh)

542 ## References

543 Baker EA, Gilbert SPR, Shimeld SM, Woollard A. 2021. Extensive non-redundancy in a recently duplicated
544 developmental gene family. *BMC Ecology and Evolution* 21:33. DOI: 10.1186/s12862-020-01735-z.

545 Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. 2007. UniProtKB/Swiss-Prot. *Methods in Molecular*
546 *Biology (Clifton, N.J.)* 406:89–112. DOI: 10.1007/978-1-59745-535-0_4.

547 Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R,
548 Chitsaz H, Chou W-C, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca
549 NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis É, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB,
550 Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO,
551 Knight JR, Koren S, Lam T-W, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, MacCallum I,
552 MacManes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM,
553 Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin
554 S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H,
555 Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu S-M, Yuan J, Zhang G, Zhang H,
556 Zhou S, Korf IF. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three
557 vertebrate species. *GigaScience* 2:10. DOI: 10.1186/2047-217X-2-10.

558 Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the
559 space of genes and proteins. *NAR genomics and bioinformatics* 2:lqaa026. DOI:
560 10.1093/nargab/lqaa026.

561 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*
562 12:59–60. DOI: 10.1038/nmeth.3176.

563 *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for
564 investigating biology. *Science (New York, N.Y.)* 282:2012–2018. DOI: 10.1126/science.282.5396.2012.

565 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture
566 and applications. *BMC Bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.

567 Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. 2021. Accurate long-read de novo assembly evaluation with
568 Inspector. *Genome Biology* 22:312. DOI: 10.1186/s13059-021-02527-4.
569 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased
570 assembly graphs with hifiasm. *Nature Methods* 18:170–175. DOI: 10.1038/s41592-020-01056-5.
571 Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas
572 R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016.
573 Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*
574 13:1050–1054. DOI: 10.1038/nmeth.4035.
575 Chow FW-N, Koutsovoulos G, Ovando-Vázquez C, Neophytou K, Bermúdez-Barrientos JR, Laetsch DR,
576 Robertson E, Kumar S, Claycomb JM, Blaxter M, Abreu-Goodger C, Buck AH. 2019. Secretion of an
577 Argonaute protein by a parasitic nematode and the evolution of its siRNA guides. *Nucleic Acids*
578 *Research* 47:3594–3606. DOI: 10.1093/nar/gkz142.
579 Coghlan A, Tyagi R, Cotton JA, Holroyd N, Rosa BA, Tsai IJ, Laetsch DR, Beech RN, Day TA, Hallsworth-Pepin K,
580 Ke H-M, Kuo T-H, Lee TJ, Martin J, Maizels RM, Mutowo P, Ozersky P, Parkinson J, Reid AJ, Rawlings ND,
581 Ribeiro DM, Swapna LS, Stanley E, Taylor DW, Wheeler NJ, Zamanian M, Zhang X, Allan F, Allen JE,
582 Asano K, Babayan SA, Bah G, Beasley H, Bennett HM, Bisset SA, Castillo E, Cook J, Cooper PJ, Cruz-
583 Bustos T, Cuéllar C, Devaney E, Doyle SR, Eberhard ML, Emery A, Eom KS, Gilleard JS, Gordon D, Harcus
584 Y, Harsha B, Hawdon JM, Hill DE, Hodgkinson J, Horák P, Howe KL, Huckvale T, Kalbe M, Kaur G, Kikuchi
585 T, Koutsovoulos G, Kumar S, Leach AR, Lomax J, Makepeace B, Matthews JB, Muro A, O'Boyle NM,
586 Olson PD, Osuna A, Partono F, Pfarr K, Rinaldi G, Foronda P, Rollinson D, Samblas MG, Sato H, Schnyder
587 M, Scholz T, Shafie M, Tanya VN, Toledo R, Tracey A, Urban JF, Wang L-C, Zarlenga D, Blaxter ML,
588 Mitreva M, Berriman M, International Helminth Genomes Consortium. 2019. Comparative genomics of
589 the major parasitic worms. *Nature Genetics* 51:163–174. DOI: 10.1038/s41588-018-0262-1.

590 Colomb F, Ogunkanbi A, Jamwal A, Dong B, Maizels RM, Finney CAM, Wasmuth JD, Higgins MK, McSorley HJ.
591 2024. IL-33-binding HpARI family homologues with divergent effects in suppressing or enhancing type 2
592 immune responses. *Infection and Immunity* 0:e00395-23. DOI: 10.1128/iai.00395-23.
593 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR:
594 ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 29:15–21. DOI:
595 10.1093/bioinformatics/bts635.
596 Doyle SR. 2022. Improving helminth genome resources in the post-genomic era. *Trends in Parasitology* 38:831–
597 840. DOI: 10.1016/j.pt.2022.06.002.
598 Doyle SR, Tracey A, Laing R, Holroyd N, Bartley D, Bazant W, Beasley H, Beech R, Britton C, Brooks K, Chaudhry
599 U, Maitland K, Martinelli A, Noonan JD, Paulini M, Quail MA, Redman E, Rodgers FH, Sallé G, Shabbir
600 MZ, Sankaranarayanan G, Wit J, Howe KL, Sargison N, Devaney E, Berriman M, Gilleard JS, Cotton JA.
601 2020. Genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus
602 contortus, a model gastrointestinal worm. *Communications Biology* 3:656. DOI: 10.1038/s42003-020-
603 01377-3.
604 Ebenezer TE, Muigai AWT, Nouala S, Badaoui B, Blaxter M, Buddie AG, Jarvis ED, Korlach J, Kuja JO, Lewin HA,
605 Majewska R, Mapholi N, Maslamoney S, Mbo'o-Tchouawou M, Osuji JO, Seehausen O, Shorinola O,
606 Tiambo CK, Mulder N, Ziyomo C, Djikeng A. 2022. Africa: sequence 100,000 species to safeguard
607 biodiversity. *Nature* 603:388–392. DOI: 10.1038/d41586-022-00712-4.
608 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome*
609 *Biology* 20:238. DOI: 10.1186/s13059-019-1832-y.
610 Erkut C, Vasilj A, Boland S, Habermann B, Shevchenko A, Kurzchalia TV. 2013. Molecular Strategies of the
611 Caenorhabditis elegans Dauer Larva to Survive Extreme Desiccation. *PLOS ONE* 8:e82473. DOI:
612 10.1371/journal.pone.0082473.

613 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated
614 genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*
615 117:9451–9457. DOI: 10.1073/pnas.1921046117.
616 Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. 2023. BRAKER3: Fully automated
617 genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA.
618 DOI: 10.1101/2023.06.10.544449.
619 Geng S, Li Q, Zhou X, Zheng J, Liu H, Zeng J, Yang R, Fu H, Hao F, Feng Q, Qi B. 2022. Gut commensal E. coli outer
620 membrane proteins activate the host food digestive system through neural-immune communication.
621 *Cell Host & Microbe* 30:1401-1416.e8. DOI: 10.1016/j.chom.2022.08.004.
622 Gilabert A, Curran DM, Harvey SC, Wasmuth JD. 2016. Expanding the view on the evolution of the nematode
623 dauer signalling pathways: refinement through gene gain and pathway co-option. *BMC Genomics*
624 17:476. DOI: 10.1186/s12864-016-2770-7.
625 Gotoh O, Morita M, Nelson DR. 2014. Assessment and refinement of eukaryotic gene structure prediction with
626 gene-structure-aware multiple protein sequence alignment. *BMC bioinformatics* 15:189. DOI:
627 10.1186/1471-2105-15-189.
628 Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic
629 duplication in primary genome assemblies. *Bioinformatics (Oxford, England)* 36:2896–2898. DOI:
630 10.1093/bioinformatics/btaa025.
631 Hirani N, Westenberg M, Seed PT, Petalcorin MIR, Dolphin CT. 2016. C. elegans flavin-containing
632 monooxygenase-4 is essential for osmoregulation in hypotonic stress. *Biology Open* 5:537–549. DOI:
633 10.1242/bio.017400.
634 Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome
635 Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics (Oxford, England)* 32:767–769. DOI:
636 10.1093/bioinformatics/btv661.

637 Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. WormBase ParaSite − a comprehensive resource for
638 helminth genomics. *Molecular and Biochemical Parasitology* 215:2–10. DOI:
639 10.1016/j.molbiopara.2016.11.005.
640 Hu Y, Ellis BL, Yiu YY, Miller MM, Urban JF, Shi LZ, Aroian RV. 2013. An Extensive Comparison of the Effect of
641 Anthelmintic Classes on Diverse Nematodes. *PLOS ONE* 8:e70702. DOI: 10.1371/journal.pone.0070702.
642 Iwata H, Gotoh O. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of
643 Spaln that incorporates additional species-specific features. *Nucleic Acids Research* 40:e161. DOI:
644 10.1093/nar/gks708.
645 Johnston CJC, Smyth DJ, Kodali RB, White MPJ, Harcus Y, Filbey KJ, Hewitson JP, Hinck CS, Ivens A, Kemter AM,
646 Kildemoes AO, Le Bihan T, Soares DC, Anderton SM, Brenn T, Wigmore SJ, Woodcock HV, Chambers RC,
647 Hinck AP, McSorley HJ, Maizels RM. 2017. A structurally distinct TGF-β mimic from an intestinal
648 helminth parasite potently induces regulatory T cells. *Nature Communications* 8:1741. DOI:
649 10.1038/s41467-017-01886-6.
650 Jung H, Jeon M-S, Hodgett M, Waterhouse P, Eyun S. 2020. Comparative Evaluation of Genome Assemblers
651 from Long-Read Sequencing for Plants and Crops. *Journal of Agricultural and Food Chemistry* 68:7670–
652 7677. DOI: 10.1021/acs.jafc.0c01647.
653 Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama
654 H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient de novo assembly of highly heterozygous
655 genomes from whole-genome shotgun short reads. *Genome Research*:gr.170720.113. DOI:
656 10.1101/gr.170720.113.
657 Kamath RS, Martinez-Campos M, Zipperlen P, Fraser AG, Ahringer J. 2000. Effectiveness of specific RNA-
658 mediated interference through ingested double-stranded RNA in Caenorhabditis elegans. *Genome*
659 *Biology* 2:research0002.1. DOI: 10.1186/gb-2000-2-1-research0002.

660 Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional
661 Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* 428:726–731.
662 DOI: 10.1016/j.jmb.2015.11.006.
663 Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. 2017. NucDiff: in-depth characterization and
664 annotation of differences between two sets of DNA sequences. *BMC bioinformatics* 18:338. DOI:
665 10.1186/s12859-017-1748-z.
666 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs.
667 *Nature Biotechnology* 37:540–546. DOI: 10.1038/s41587-019-0072-8.
668 Koo H, Lee G-W, Ko S-R, Go S, Kwon S-Y, Kim Y-M, Shin A-Y. 2023. Two long read-based genome assembly and
669 annotation of polyploidy woody plants, Hibiscus syriacus L. using PacBio and Nanopore platforms.
670 *Scientific Data* 10:713. DOI: 10.1038/s41597-023-02631-z.
671 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read
672 assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 27:722–736. DOI:
673 10.1101/gr.215087.116.
674 Kreis HA. 1964. Ein neuer Nematode aus dem äusseren Gehörgang von Zeburindern in Ostafrika, Rhabditis
675 bovis n.sp. (Rhabditidoidea, Rhabditidae). *Schweizer Archiv fur Tierheilkunde* 106:371–378.
676 Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An
677 information aesthetic for comparative genomics. *Genome Research* 19:1639–1645. DOI:
678 10.1101/gr.092759.109.
679 Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. DOI:
680 10.12688/f1000research.12232.1.
681 Laing R, Bartley DJ, Morrison AA, Rezansoff A, Martinelli A, Laing ST, Gilleard JS. 2015. The cytochrome P450
682 family in the parasitic nematode Haemonchus contortus. *International Journal for Parasitology* 45:243–
683 251. DOI: 10.1016/j.ijpara.2014.12.001.

684 Laing R, Kikuchi T, Martinelli A, Tsai IJ, Beech RN, Redman E, Holroyd N, Bartley DJ, Beasley H, Britton C, Curran
685 D, Devaney E, Gilabert A, Hunt M, Jackson F, Johnston SL, Kryukov I, Li K, Morrison AA, Reid AJ, Sargison
686 N, Saunders GI, Wasmuth JD, Wolstenholme A, Berriman M, Gilleard JS, Cotton JA. 2013. The genome
687 and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery.
688 *Genome Biology* 14:R88. DOI: 10.1186/gb-2013-14-8-r88.
689 Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. *Nucleic Acids Research* 39:D19–D21.
690 DOI: 10.1093/nar/gkq1019.
691 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE*
692 *transactions on visualization and computer graphics* 20:1983–1992. DOI: 10.1109/TVCG.2014.2346248.
693 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. DOI:
694 10.1093/bioinformatics/bty191.
695 Li H. 2023. Protein-to-genome alignment with miniprot. *Bioinformatics* 39:btad014. DOI:
696 10.1093/bioinformatics/btad014.
697 Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan W. 2012. Comparison of the two
698 major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in*
699 *Functional Genomics* 11:25–37. DOI: 10.1093/bfgp/elr035.
700 Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. 2016. Assembly of long error-prone reads
701 using de Bruijn graphs. *Proceedings of the National Academy of Sciences* 113:E8396–E8405. DOI:
702 10.1073/pnas.1604560113.
703 Liu H, Wu S, Li A, Ruan J, Wu S, Li A, Ruan J. 2021. SMARTdenovo: a de novo assembler using long noisy reads.
704 *Gigabyte* 2021:1–9. DOI: 10.46471/gigabyte.15.
705 Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic
706 genomes by self-training algorithm. *Nucleic Acids Research* 33:6494–6506. DOI: 10.1093/nar/gki937.
707 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined
708 Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic,

709 Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* 38:4647–4654. DOI:
710 10.1093/molbev/msab199.
711 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile
712 genome alignment system. *PLOS Computational Biology* 14:e1005944. DOI:
713 10.1371/journal.pcbi.1005944.
714 Mieszkowska N, Wilson WH, Jenkins HL, Mrowicki R, Cunliffe M, Adkins P, Lonsdale P, Harley J, Bishop JDD,
715 Warring S, Blaxter ML. 2022. Sequence locally, think globally: The Darwin Tree of Life Project.
716 *Proceedings of the National Academy of Sciences* 119. DOI: 10.1073/pnas.2115642118.
717 Mohamed AM, Chin-Sang ID. 2011. The C. elegans nck-1 gene encodes two isoforms and is required for
718 neuronal guidance. *Developmental Biology* 354:55–66. DOI: 10.1016/j.ydbio.2011.03.019.
719 Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020.
720 HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity
721 long reads. :2020.03.14.992248. DOI: 10.1101/2020.03.14.992248.
722 Osbourn M, Soares DC, Vacca F, Cohen ES, Scott IC, Gregory WF, Smyth DJ, Toivakka M, Kemter AM, le Bihan T,
723 Wear M, Hoving D, Filbey KJ, Hewitson JP, Henderson H, Gonzàlez-Cìscar A, Errington C, Vermeren S,
724 Astier AL, Wallace WA, Schwarze J, Ivens AC, Maizels RM, McSorley HJ. 2017. HpARI Protein Secreted
725 by a Helminth Parasite Suppresses Interleukin-33. *Immunity* 47:739-751.e5. DOI:
726 10.1016/j.immuni.2017.09.015.
727 Oxford Nanopore Technologies. 2018. Medaka: Sequence correction provided by ONT Research.
728 Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim J,
729 Waterston RH, Murray JI. 2019. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at
730 single-cell resolution. *Science* 365:eaax1971. DOI: 10.1126/science.aax1971.
731 Pollo SMJ, Leon-Coria A, Liu H, Cruces-Gonzalez D, Finney CAM, Wasmuth JD. 2023. Transcriptional patterns of
732 sexual dimorphism and in host developmental programs in the model parasitic nematode
733 Heligmosomoides bakeri. *Parasites & Vectors* 16:171. DOI: 10.1186/s13071-023-05785-2.

734 Pollo SMJ, Reiling SJ, Wit J, Workentine ML, Guy RA, Batoff GW, Yee J, Dixon BR, Wasmuth JD. 2020.
735 Benchmarking hybrid assemblies of Giardia and prediction of widespread intra-isolate structural
736 variation. *Parasites & Vectors* 13:108. DOI: 10.1186/s13071-020-3968-8.
737 Pundir S, Martin MJ, O'Donovan C. 2017. UniProt Protein Knowledgebase. In: Wu CH, Arighi CN, Ross KE eds.
738 *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*. Methods in Molecular
739 Biology. New York, NY: Springer, 41–55. DOI: 10.1007/978-1-4939-6783-4_2.
740 R Core Team. 2022. R: A Language and Environment for Statistical Computing.
741 Roubin R, Naert K, Popovici C, Vatcher G, Coulier F, Thierry-Mieg J, Pontarotti P, Birnbaum D, Baillie D, Thierry-
742 Mieg D. 1999. let-756, a C. elegans fgf essential for worm development. *Oncogene* 18:6741–6747. DOI:
743 10.1038/sj.onc.1203074.
744 Rousakis A, Vlassis A, Vlanti A, Patera S, Thireos G, Syntichaki P. 2013. The general control nonderepressible-2
745 kinase mediates stress response and longevity induced by target of rapamycin inactivation in
746 Caenorhabditis elegans. *Aging Cell* 12:742–751. DOI: 10.1111/acel.12101.
747 Rual J-F, Ceron J, Koreth J, Hao T, Nicot A-S, Hirozane-Kishikawa T, Vandenhaute J, Orkin SH, Hill DE, Heuvel S
748 van den, Vidal M. 2004. Toward Improving Caenorhabditis elegans Phenome Mapping With an
749 ORFeome-Based RNAi Library. *Genome Research* 14:2162–2168. DOI: 10.1101/gr.2505604.
750 Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* 17:155–158. DOI:
751 10.1038/s41592-019-0669-3.
752 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome
753 assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*
754 31:3210–3212. DOI: 10.1093/bioinformatics/btv351.
755 Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC*
756 *Bioinformatics* 6:31. DOI: 10.1186/1471-2105-6-31.
757 Smyth DJ, Harcus Y, White MPJ, Gregory WF, Nahler J, Stephens I, Toke-Bjolgerud E, Hewitson JP, Ivens A,
758 McSorley HJ, Maizels RM. 2018. TGF-β mimic proteins form an extended gene family in the murine

759 parasite *Heligmosomoides polygyrus*. *International Journal for Parasitology* 48:379–385. DOI:
760 10.1016/j.ijpara.2017.12.004.
761 Stevens L, Martínez-Ugalde I, King E, Wagah M, Absolon D, Bancroft R, Gonzalez de la Rosa P, Hall JL, Kieninger
762 M, Kloch A, Pelan S, Robertson E, Pedersen AB, Abreu-Goodger C, Buck AH, Blaxter M. 2023. Ancient
763 diversity in host-parasite interaction genes in a model parasitic nematode. *Nature Communications*
764 14:7776. DOI: 10.1038/s41467-023-43556-w.
765 Stevens L, Rooke S, Falzon LC, Machuka EM, Momanyi K, Murungi MK, Njoroge SM, Odinga CO, Ogendo A,
766 Ogola J, Fèvre EM, Blaxter M. 2020. The Genome of Caenorhabditis bovis. *Current biology: CB* 30:1023-
767 1031.e4. DOI: 10.1016/j.cub.2020.01.074.
768 Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element
769 families, sequence models, and genome annotations. *Mobile DNA* 12:2. DOI: 10.1186/s13100-020-
770 00230-y.
771 Stroehlein AJ, Young ND, Gasser RB. 2018. Improved strategy for the curation and classification of kinases, with
772 broad applicability to other eukaryotic protein groups. *Scientific Reports* 8:6808. DOI: 10.1038/s41598-
773 018-25020-8.
774 Sun J, Li R, Chen C, Sigwart JD, Kocot KM. 2021. Benchmarking Oxford Nanopore read assemblers for high-
775 quality molluscan genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*
776 376:20200160. DOI: 10.1098/rstb.2020.0160.
777 Sutton JM, Millwood JD, McCormack AC, Fierst JL. 2021. Optimizing experimental design for genome
778 sequencing and assembly with Oxford Nanopore Technologies. *GigaByte* 2021. DOI:
779 10.46471/gigabyte.27.
780 Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences.
781 *Current Protocols in Bioinformatics* Chapter 4:4.10.1-4.10.14. DOI: 10.1002/0471250953.bi0410s25.
782 Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long
783 uncorrected reads. *Genome Research* 27:737–746. DOI: 10.1101/gr.214270.116.

784 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK,
785 Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome
786 Assembly Improvement. *PLOS ONE* 9:e112963. DOI: 10.1371/journal.pone.0112963.
787 Wasmuth J, Schmid R, Hedley A, Blaxter M. 2008. On the Extent and Origins of Genic Novelty in the Phylum
788 Nematoda. *PLOS Neglected Tropical Diseases* 2:e258. DOI: 10.1371/journal.pntd.0000258.
789 Weisman CM, Murray AW, Eddy SR. 2022. Mixing genome annotation methods in a comparative analysis
790 inflates the apparent number of lineage-specific genes. *Current biology: CB* 32:2632-2639.e2. DOI:
791 10.1016/j.cub.2022.04.085.
792 Wickham H. 2016. Data Analysis. In: Wickham H ed. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Cham:
793 Springer International Publishing, 189–201. DOI: 10.1007/978-3-319-24277-4_9.
794