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

Background: Reduced sequencing costs is facilitating comprehensive BioGenome projects and sequencing of specific species by individual research groups. While the choice between sequencing technologies is limited, the choice of assembly algorithm can be overwhelming. However, this choice will influence the accuracy of the assembly and downstream analysis and annotation. Here, we set out to demonstrate the extent and effect of assembly variation using three nematode species—*Caenorhabditis bovis*, *Haemonchus contortus*, *Heligmosomoides bakeri*—whose genomes differ in size and complexity. In addition to comparing the assembly accuracy and structure, we predicted gene models from each assembly and investigated the effect on large gene family membership.

Results: We found that assemblers using the Overlap-Layout-Consensus (OLC) algorithm were better choices for assembling more complex genomes while those that used the De-Bruijn Graph (DBG) algorithm performed better in smaller genomes. However, the DBG Redbean assembler performed comparably well in both scenarios, making it a versatile choice. Using standard assembly metrics, variations seemed relatively minor; however, we found noticeable intra-species/inter-assembly differences in the composition of four gene families—CYP, GST, UGT, and NHR—that are of interest in anti-nematode drug discovery and drug resistance research. Finally, through comparisons of highly conserved genes in the assemblies of *H. contortus*, we found probable errors in the highly curated reference assembly, demonstrating the need to support ongoing curation efforts.

Conclusions: While the choice of assembly algorithm can be daunting and the differences appear relatively minor, there is a need to consider differences in assemblies in terms of the assembly structure and subsequent annotation. We repeat the challenge, started by others, of relying on a small number of convenient assembly metrics. Finally, we demonstrate the value of exploring alternative assemblies, particularly in the study of content of protein-coding genes.

30 ## Introduction

31 Genome assembly software serves a critical, often underappreciated role in the life sciences. Its purpose is to
32 reconstruct, *de novo*, the complete sequence of an organism's DNA—its genome—from shorter DNA fragments
33 obtained through one or more sequencing technologies. A typical genome assembler is a pipeline of smaller
34 programs that perform specific tasks, such as error correction, read alignment and overlap detection, and
35 consensus building. The reduced costs and increase outputs of sequencing platforms have empowered the
36 launch of large BioGenome consortia which strive to sequence, assemble and annotation the genomes of as
37 many animal and plant species as possible (Darwin Tree of Life Project Consortium, 2022; Ebenezer et al.,
38 2022). There is also an increase of individual labs constructing the genomes of their favourite species. This
39 places the evolution and refinement of genome assemblers at the core of current and future breakthroughs in
40 health, agricultural, and environmental sciences.

41 Understanding accuracy of the output of the genome assembler—long strings of As, Gs, Cs, Ts and the
42 occasional N—is helpful; errors in the produced assembly will carry into and magnify through the annotation
43 and analysis, often without notice. There are many sources of error. One example is high and/or inconsistent
44 levels of heterozygosity in the organism’s genome leading to sequence reads being assembled as separate loci
45 rather than collapsed in the haploid assembly. A second example is long regions of repetitive DNA and
46 transposable elements scattered through the genome being incorrectly collapsed or rearranged, leading to
47 gaps. While some genome assemblers specialise in one particular problem, for example heterozygosity (Kajitani
48 et al., 2014), most assemblers are generalists and seek to identify and resolve as many potential errors as
49 possible. There are, naturally, different approaches to take. The result is that different assemblers—when given
50 the same sequence reads as input—will produce different assemblies.

51 Evaluations of genome assembler performance show that even within a phylum, there is no single optimum.
52 The Assemblathon2 competition used 102 metrics to compare assemblies from three chordate species
53 (Bradnam et al., 2013); Sun and colleagues showed that genome size variation within the molluscs required
54 different assemblers (Sun *et al.*, 2021); and we and others have shown the challenges and successes of
55 software to construct a haploid assembly from tetraploid protozoa and plants (Pollo et al., 2020; Jung et al.,
56 2020). These studies, and many others, provided the motivation to explore the optimal genome assembler—at
57 the current time—for nematodes; we use their genomes to understand host-parasite interactions.

58 Nematode species are parasites of humans, livestock, and crops; free-living contributors to environmental
59 health; and pivotal model organisms in understand human biology. Accordingly, nematodes have often led the
60 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 are available for at least 135 species (WormBase Parasite release 18) (Howe et al., 2017). These assemblies represent a broad range of different sequencing technologies and assembly protocols (Coghlan et al., 2019). This availability of genomic, transcriptomics, and proteomic information has led to comparative genomic studies on various scales, much of it focused on the evolution of gene families; their birth and death, expansions and contractions (Wasmuth et al., 2008; Coghlan et al., 2019; Baker et al., 2021). 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. For example, in the study of gene birth and death, if different genome annotation protocols are used to generate the gene models of the species being compared, there is an increase in the number of apparent lineage-specific genes (Weisman *et al.*, 2022). This motivated us to determine the effect of genome assembler heterogeneity on gene family comparisons.

We focused on 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) and Pacific Biosciences (PacBio).

C. bovis is associated with bovine parasitic otitis disease and is the only pathogenic member of its genus (Msolla et al., 1993). Its genome assembly is approximately 63Mb, making it the smallest *Caenorhabditis* and one of the smallest nematode genomes so far sequenced (Stevens et al., 2020). *H. 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 (Redman *et al.*, 2012; Gilleard, 2006; Kaminsky *et al.*, 2008; Taylor *et al.*, 2013; Bethony *et al.*, 2006; LeJambre *et al.*, 2008; Knox *et al.*, 2003). The *H. contortus* genome assembly has undergone multiple rounds of improvement; an initial short-read assembly of ~370Mb was later surpassed by a long-read assembly that was scaffolded through optical mapping and improved with expert manual curation (Laing et al., 2013; Doyle et al., 2018, 2020). Currently, the five autosomes and one X chromosome of the *H. contortus* genome is assembled into seven scaffolds, with a total length of ~283Mb. Finally, *H. bakeri* is an intestinal parasite of mice that serves as a model for studying host-parasite interactions and testing anthelmintic efficacy (Pollo *et al.*, 2023; Reynolds *et al.*, 2012; Maizels *et al.*, 2012). Please note, that the Chow assembly paper described the species sequenced as *Heligmosomoides polygyrus*, a long-standing miss-classification of the study species which has been recently corrected through genomic comparisons (Stevens et al., 2023). The two *Heligmosomoides* nematode species, *H. bakeri* and *H. polygyrus*, are separate species, with *H. bakeri* being the laboratory strain infecting domestic mice and *H. polygyrus* the wild strain infecting wood mice (Cable et al., 2006).

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

100

101 ## Methods

102 **Nematode species and sequence data**

103 DNA sequence reads for *C. bovis* and *H. contortus* were retrieved from the Sequence Read Archive (Leinonen,
104 Sugawara & Shumway, 2011), and their genomes downloaded from WormBase Parasite (release 17) (Howe et
105 al., 2017). The sequence reads and assembly for *H. bakeri* are new for this study. *C. bovis* DNA (BioProject:
106 PRJEB34497) was sequenced using the Oxford Nanopore MinION platform (Stevens et al., 2020). *H. contortus*
107 DNA (BioProject PRJEB506) was sequenced using the PacBio RS and Sequel platforms (Doyle et al., 2020). We
108 generated *H. bakeri* DNA was generated as part of de novo sequencing project. A full description will follow in a
109 future version of this preprint. Briefly: we extracted adult *H. bakeri* worms from experimentally infected mice;
110 separately females and males; pooled 20 worms of each sex. PacBio Hi-Fi sequencing was done at the
111 University of Delaware DNA Sequencing & Genotyping Center. For the assemblies in this study, we combined
112 sequence reads from female and male worms.

113 **Genome assembly and quality control.**

114 As a guide, we reproduced the protocol for assembling the *C. bovis* genome as outlined in (Stevens et al.,
115 2020). The same protocol was followed to generate draft assemblies for the *C. bovis* and *H. contortus* using six
116 long-read assemblers that support Oxford Nanopore and PacBio RS and Sequel reads. Similarly, the protocol
117 was used to generate draft assemblies for the *H. bakeri* using three long-read assemblers that support HiFi
118 data. The protocol is briefly outlined as follows:

119 **a) Generation of preliminary assemblies**

120 First, preliminary genome assemblies for *C. bovis* and *H. contortus* were generated using six long-read
121 assemblers: Redbean v.2.5 (also known as Wtdbg2) (Ruan & Li, 2020), Canu v.2.1.1 (Koren et al., 2017), Flye
122 v.2.8.2 (Lin et al., 2016; Kolmogorov et al., 2019), SMARTdenovo v.1.0.0 (Liu et al., 2021), along with the Pb-
123 assembly suite encompassing Falcon-kit v.1.8.1 and Falcon-unzip v.1.3.7 (Chin et al., 2013, 2016). For *H. bakeri*,
124 preliminary assemblies were generated using three HiFi-compatible long-read assemblers: Flye v.2.8.2 (Lin et
125 al., 2016; Kolmogorov et al., 2019), Hifiasm v.0.16.1-r375 (Cheng et al., 2021) and HiCanu v.2.1.1 (Nurk et al.,
126 2020). All the commands used to generate preliminary assemblies can be found in S1-Methods.

127 **b) Decontamination and polishing of preliminary assemblies**

128 For the *C. bovis* and *H. contortus* assemblies contaminated contigs were removed from the preliminary draft
129 assemblies using Blobtools v.1.1.1 (Laetsch & Blaxter, 2017). This process involved mapping of the long reads to
130 assemblies using minimap2 v.2.17-r941 (Li, 2018) and the taxonomic classification of the contigs based on the
131 Uniprot Proteome database (Pundir, Martin & O'Donovan, 2017) using DIAMOND v.2.0.6.144 (Buchfink, Xie &
132 Huson, 2015). The resulting *de novo* assemblies were polished to correct for sequencing errors using different
133 tools as follows: for both *C. bovis* and *H. contortus*, the long MinION and PacBio reads were mapped to each of
134 their respective assemblies using minimap2 and four rounds of polishing performed using Racon v.1.4.20
135 (Vaser et al., 2017). Following this, for *C. bovis*, one round of polishing was performed using Medaka v.1.3.2
136 (accessible at https://github.com/nanoporetech/medaka). For *H. contortus*, PacBio reads were mapped to
137 each of the *H. contortus* preliminary assemblies using pbmm2 v.1.4.0 (available at
138 https://github.com/PacificBiosciences/pbmm2) before performing one round of polishing using Arrow v2.3.3
139 (installed via the GenomicConsensus package v.2.3.3, available at
140 http://github.com/PacificBiosciences/pbbioconda). Lastly, to correct for any remaining errors, Illumina short-
141 reads from both species were mapped using BWA-MEM v.0.7.17-r1188 (Li, 2013) to each of the respective
142 assemblies and two rounds of Racon and Pilon v.1.24 (Walker et al., 2014) performed. All the commands
143 associated with these steps can be found in S1-Methods

144 **c) Removal of residual haplotype duplications**

145 Due to the high levels of duplications and heterozygosity in the *H. contortus* and *H. bakeri* genomes, an
146 additional step was taken to remove redundant haplotigs and contig overlaps from their assemblies. This step
147 was not included in the *C. bovis* protocol. To do this, we used the Purge_dups v.1.2.5 tool (Guan et al., 2020).
148 We initially applied the default cut-off threshold, followed by two manually set cut-offs derived from the
149 default cutoffs. This step is recommended by (Guan et al., 2020) because default cutoffs in the Purge_dups
150 pipeline may not be optimal for reducing low-coverage k-mers in highly heterozygous and/or repetitive diploid
151 organisms. However, this comes with a trade-off between achieving more accurate assemblies and maintaining
152 completeness. All the commands and parameters implemented in this step can be found in S1-Methods. We
153 considered the second manually set cut-offs for all the generated assemblies for the *H. contortus* and *H. bakeri*
154 (Table S4)

155 **d) Assembly evaluation**

156 All the generated assemblies for the *C. bovis*, *H. contortus* and *H. bakeri* were evaluated for contiguity and
157 completeness using various metrics, including the number of fragments, assembly size, N50, length of the
158 longest contig, and BUSCO v.4.1.4 (Simão et al., 2015). The command used for the BUSCO evaluation was:

159 (busco -i [PATH/TO/ASSEMBLY] --lineage nematoda_odb10 --out [UNIQUE_NAME] --mode genome --long --
160 augustus). Additionally, we assessed the alignment of sequence reads to each assembled genome using
161 Inspector v.1.0.2 (Chen et al., 2021) to identify potential large-scale (structural) and small-scale errors. For this
162 evaluation, we used the raw ONT reads for the *C. bovis* assemblies, the raw PacBio RS and Sequel reads for the
163 *H. contortus* assemblies, and the raw HiFi long reads for the *H. bakeri* assemblies. The command used to align
164 the reads to the assemblies was: (inspector.py --thread 24 --contig [PATH/TO/ASSEMBLY] --read
165 [PATH/TO/READS] --outpath [UNIQUE_NAME_FOR_OUTPUT_DIRECTORY] --datatype [nanopore for *C. bovis* or
166 clr for *H. contortus* or hifi for *H. bakeri*]. To generate the corrected assembly, we used the command:
167 (inspector-correct.py --thread 24 --inspector [UNIQUE_NAME_FOR_OUTPUT_DIRECTORY] --datatype [nano-
168 raw or pacbio-raw or pacbio-hifi] --outpath [DIRECTORY_FOR_CORRECTED_ASSEMBLY]).

169 **Genome synteny**

170 Genome-to-genome alignments of the final draft assemblies were performed using the Nucmer wrapper script
171 incorporated in MUMmer v.3.23 (Kurtz et al., 2004; Marçais et al., 2018), with the options: (nucmer --
172 maxmatch --prefix=[UNIQUE_NAME] [REFERENCE GENOME] [QUERY GENOME]). For *C. bovis*, we considered
173 the Redbean v2.5 assembly as the reference genome given its better contiguity while the Doyle assembly
174 served as the reference genome for *H. contortus*. For *H. bakeri*, the Hifiasm assembly was the reference
175 genome. Additionally, for the *C. bovis*, we used the dnadiff wrapper script found within Nucmer, with the
176 options (dnadiff -d [NUCMER_OUTPUT].delta), to report one-to-one alignment coordinates. The coordinates
177 were used to link scaffolds between assemblies for the visualization of synteny between *C. bovis* genomes
178 using the Circos tools v.0.69-8 (Krzywinski et al., 2009). Lastly, for all the generated assemblies, we used
179 NucDiff v.2.0.3 (Khelik et al., 2017) to quantify the structural and local genome differences between each
180 assembly and its respective reference genome. The options used for the NucDiff analyses were: (nucdiff --proc
181 24 --ref_name_full yes --query_name_full yes [REFERENCE GENOME] [QUERY GENOME] [PREFIX NAME]).

182 **BUSCO genes analyses**

183 The number of highly conserved orthologues correctly assembled across the generated assemblies were
184 determined using BUSCO v.4.1.4 (Simão et al., 2015) with the options: (busco -i [PATH/TO/ASSEMBLY] --lineage
185 nematoda_odb10 --out [UNIQUE_NAME] --mode genome --long --augustus). We then generated Sankey
186 diagrams using the SankeyMATIC builder (available at http://sankeymatic.com/build/) to visualize the
187 transitions of BUSCO genes—genes that changed categories—between assemblies i.e., those that were not S-S-
188 S, D-D-D, F-F-F or M-M-M. In the Sankey diagrams, each BUSCO category, Single (S), Duplicated (D),
189 Fragmented (F), or Missing (M), is represented by a distinct column and the connections or "flows" between

190 these columns indicate how elements move from one category to another in the next assembly. The thickness
191 of the connecting lines reflects the proportion of elements transitioning between categories. We highlight that
192 the Sankey diagrams illustrate pairwise comparisons of BUSCO genes flowing between two assemblies at a
193 time. Set-based BUSCO genes data was displayed using UpSet plots (Lex et al., 2014) with additional annotation
194 in Adobe Illustrator (available at https://www.adobe.com). For *H. contortus*, we created a heatmap with the
195 ggplot2 package in R (Wickham, 2016; Core, 2022) to visualize the pattern of BUSCO genes reported as missing
196 in the reference Doyle assembly but found complete (single or duplicated) in at least one of the other six
197 assemblies. Lastly, we used the BlastKOALA tool (Kanehisa, Sato & Morishima, 2016) to identify the functions of
198 BUSCO genes absent in both the *H. contortus* reference Doyle assembly and the HiFi-based *H. bakeri*
199 assemblies, Flye, Hicanu and Hifiasm.

200 **Genome annotation**

201 **a) Soft masking assemblies**

202 All the generated assemblies were soft masked prior to genome annotation, using RepeatModeler v2.0.3 (Flynn
203 et al., 2020) and RepeatMasker v4.1.2-p1 (Tarailo-Graovac & Chen, 2009). Briefly, we used the NCBI/RMBLAST
204 [2.11.0+] search engine to create a default repeat library for RepeatModeler from the Dfam database of
205 transposable elements and DNA repeats (Hubley et al., 2016). Next, we used the RepeatMasker util
206 (queryRepeatDatabase.pl script) to create a custom library specific to nematode species. Both repeat libraries
207 were combined as input for the RepeatMasker program.

208 **b) Gene prediction**

209 Protein coding genes (gene models) in the generated assemblies were predicted using Braker 3 (Stanke et al.,
210 2006, 2008; Hoff et al., 2016, 2019; Brůna et al., 2021) installed using the Singularity container (available at
211 https://hub.docker.com/r/teambraker/braker3). For *H. contortus* and *H. bakeri*, we used short-read RNA-seq
212 data from the Sequence Read Archive (Leinonen, Sugawara & Shumway, 2011) as external evidence for the
213 gene prediction (BioProjects: PRJEB506 (Laing et al., 2013) and PRJNA750155 (Pollo et al., 2023) respectively).
214 The RNA-seq data was first aligned using the STAR v.2.7.10a aligner (Dobin et al., 2013) and the generated
215 alignments, in bam-format, used to produce a training set for AUGUSTUS, using GeneMark-ET (Hoff *et al*.,
216 2016; Li *et al*., 2009; Barnett *et al*., 2011) and DIAMOND (Buchfink, Xie & Huson, 2015), to filter redundant
217 training gene structures. For *C. bovis*, we used the longest isoforms of the *C. elegans* gene models (BioProject:
218 PRJNA13758) retrieved from WormBase Parasite (release 17) (Howe et al., 2017), as the protein evidence for
219 gene prediction.

220 **Identification of gene family members**

221 Members of the Cytochrome P450 (CYP), Glutathione S-transferase (GST), UDP-glucuronosyltransferase (UGT)
222 and nuclear hormone receptors (NHR) gene families were identified by comparing each assembly's predicted
223 gene models against the full *C. elegans* gene models using BLASTP (Altschul et al., 1990; Camacho et al., 2009;
224 Boratyn et al., 2012). The top hit of each gene model was selected and the gene class annotation from
225 WormBase Parasite (release 17) (Howe et al., 2017) used to identify the gene family members. In each species
226 and for every gene family, we used OrthoFinder2 (Emms & Kelly, 2019), (with parameters set to -S blast,
227 otherwise default) to detect homologs and cluster them with the *C. elegans* proteins. We then computed the
228 percentage differences between the minimum and maximum counts of gene models and the members of the
229 CYP, GST, UGT and NHR gene families predicted in all the assemblies used in the study.

230

231

232 ## Results

233 We generated fifteen new assemblies—six for *C. bovis*, six for *H. contortus*, and three for *H. bakeri*—which we
234 used, along with the published genomes for each species, for intraspecies comparisons. For readability, we will
235 refer to the published genomes by their papers' first author—Stevens for *C. bovis*, Doyle for *H. contortus*, and
236 Chow for *H. bakeri*—and by the name of the genome assembly program for the others (Chow et al., 2019;
237 Stevens et al., 2020; Doyle et al., 2020). Each new assembly was subject to decontamination and polishing
238 protocols with an additional step of 'purging' duplicates in the highly heterozygous *H. contortus* and *H. bakeri*
239 assemblies. Decontamination—the removal of contigs from bacteria and other heterogenous sources—
240 removed approximately 20% of the nucleotides of any given intermediate assembly and, in *C. bovis*, up to 90%
241 of the scaffolds (Table S1, Table S2, Table S3). *C. bovis* samples were obtained directly from cattle ears in the
242 field (Stevens et al., 2020), where contamination levels were higher than in controlled lab settings. Also, the
243 ONT reads used to assemble the *C. bovis* genome have higher error rates, that increase sequencing errors
244 particularly in homopolymeric regions (Eid et al., 2009; Rhoads & Au, 2015; Kinkar et al., 2021). Polishing—
245 using the original sequence reads to identify and fix any small errors—led to notable improvements in the ONT
246 *C. bovis* assemblies; for example, the Redbean2.5 assembly was improved from 79% to 96% of the BUSCO
247 genes were found as complete (Table S1). However, polishing had a minor impact on the PacBio-based
248 assemblies, *H. contortus* (CSII) and *H. bakeri* (HiFi), presumably because the genome assemblers used had their
249 own in-built polishing steps. 'Purging' duplicates—phasing out allelic repeats in highly heterozygous genomes
250 based on read depth—led to notable reductions in the genome sizes and duplication of BUSCO genes; for
251 example, the BUSCO genes' duplication in the *H. contortus*' Canu assembly was reduced by 48%. In
252 heterozygous genomes, allelic contigs (haplotigs) are assembled into separate contigs, resulting in an increased
253 assembly size.

254 Below, we compare the assemblies of each species using general assembly statistics, synteny, and BUSCO gene
255 completeness. We then use four gene families to demonstrate how differences in assembly impact
256 downstream analysis. For the BUSCO analyses, we use the term *present* to define genes which are either single
257 or duplicated, and the term *absent* to define genes which are either fragmented or missing.

258 ### *C. bovis* assemblies

259 The Stevens *C. bovis* assembly was generated using ONT reads with the Redbean assembler v2.3 and polished
260 with Racon and Medaka. We generated assemblies using an updated Redbean (v2.5, hereafter Redbean2.5),
261 Flye, Canu, SMARTdenovo, Falcon, and Falcon-Unzip (Ruan & Li, 2020; Lin *et al.*, 2016; Kolmogorov *et al.*, 2018,
262 2019; Koren *et al.,* 2017, 2018; Liu *et al.*, 2021; Chin *et al*., 2013, 2016) . Following decontamination and

263 polishing (Table S1), the overall standard assembly metrics were relatively similar (Table 1) . Compared to the
264 Redbean v2.3 Stevens assembly, the Redbean2.5 assembly was near identical in size but on fewer scaffolds (21
265 vs 35) and the N50 was 400 kb larger. The Canu assembly size was 10% larger and was on almost four times as
266 many scaffolds. The other four assemblies were comparable to the Stevens assembly in terms of assembly size
267 and the number of scaffolds, but the N50s were between 25% and 50% smaller.

268 We used BUSCO to evaluate how well assembled the highly conserved genes were across seven assemblies
269 (including the Stevens-Redbean2.3 assembly). There are 3131 pan-nematode orthologues (v10) and six of the
270 assemblies had similar scores (Table 1). Of the 3131 BUSCO genes, 2972 (95%) were present (single or
271 duplicated) in all seven assemblies, with 2850 (91%) as singletons (Figure S1).

272

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

273 **Table 1**: **Assembly statistics for *Caenorhabditis bovis.*** We used BUSCO version 4.1.4 with Augustus. The
274 lineage option used is 'nematoda_odb10', and C is for complete genes, S for singletons, and D for duplicates (all
275 % from 3131 genes). We used Inspector version 1.0.2 to evaluate the alignment of *C. bovis* ONT long reads to
276 the generated scaffolds to identify both large-scale (structural) and small-scale errors.

277

278 Six genes were present in only one assembly. Ninety-one (3%) were absent in all seven assemblies. A notable
279 difference was the duplication rate in the Canu assembly, which was five times that of the Stevens assembly:
280 18 genes (0.6%) vs 97 genes (3.1%). Of the BUSCO orthologues, 97.1% (3040 genes) were complete (singleton
281 or duplicated) in at least one assembly and 94.9% (2972) were complete in all seven assemblies. Of these
282 complete genes, 15 were duplicated in all assemblies. Ninety-one genes were missing or fragmented in all
283 seven assemblies, which suggests they are undergoing pseudogenization or have been lost as a consequence of
284 a genome size reduction in *C. bovis* (Stevens et al., 2020). In total, the BUSCO classification differed for 178
285 genes, *e.g.* from Duplicated to Fragmented (Figure 1A). Expectedly, there was little disagreement between the
286 Stevens (Redbean2.3) and the new Redbean2.5 assemblies; nine genes complete and single in the Stevens
287 assembly were fragmented or missing in the Redbean2.5 assembly with six genes in the reverse situation
288 (Figure 1A). Disagreements between BUSCO classification was more striking between the other assemblies. For

example, SMARTdenovo and Falcon had near identical BUSCO scores, but 36 complete and singleton 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 but were found complete and single copy in Canu (Figure 1B). The full list of BUSCO genes found in each assembly can be found in Tables S5-to-S11.

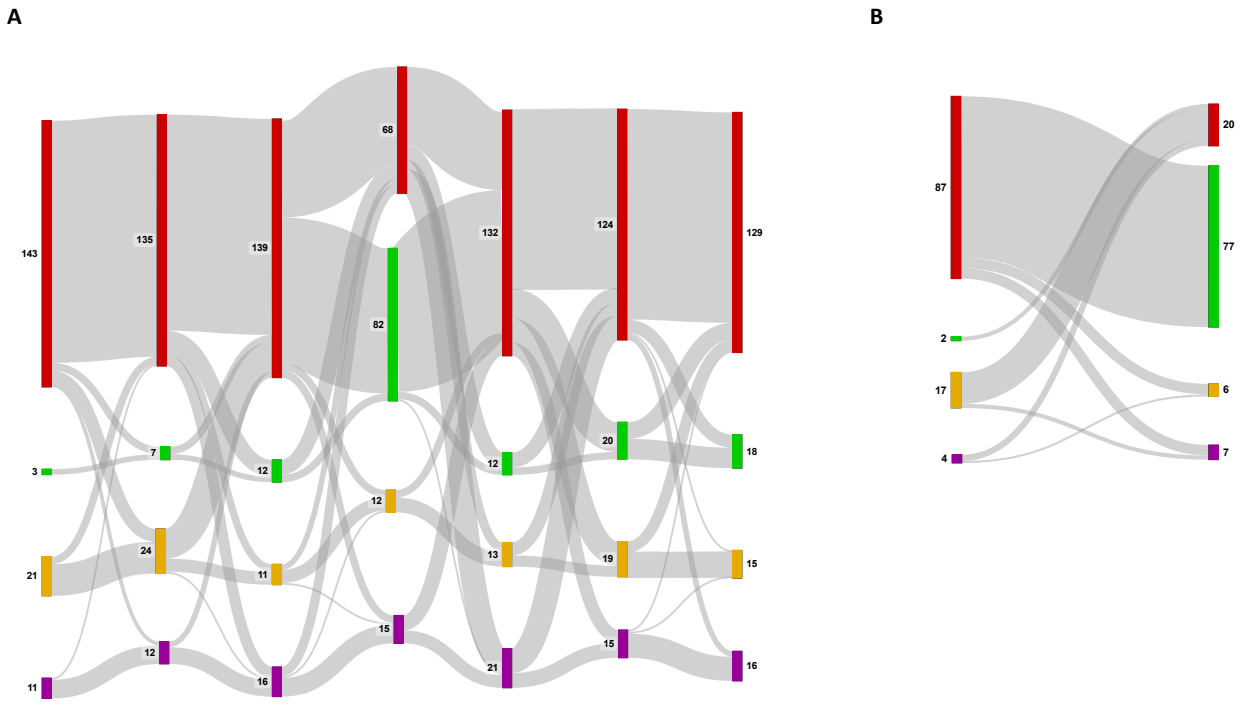

Figure 1: Transitions of BUSCO genes across multiple *C. bovis* assemblies.

(A) A Sankey diagram of the 178 genes that changed categories in at least one assembly. The order of the assemblies is consistent with Table 1. The data for this figure can be found in Data S2 and Data S3. (B) Comparison of the 110 BUSCO genes that were categorized differently in the Readbean2.5 and Canu assemblies. Data for this figure can be found in Data S4. The figures were generated using SankeyMATIC (https://sankeymatic.com/). Red – complete & single; Green – complete & duplicated; Yellow – Fragmented; Purple – Missing.

We used the original ONT reads to evaluate the reads alignment to each of the assemblies, identifying large structural and small errors with the Inspector program (Chen et al., 2021). Overall, all the assemblies had near identical QV scores; the Redbean2.5 assembly had the fewest large structural errors, and the Flye assembly had the fewest small-scale errors (Table 1 and Table S12). In considering the fundamental algorithm of each assembly program, the DBG assemblers—Redbean and Flye—typically performed best in the traditional

308 metrics—e.g., N50 and BUSCO (Table 1). The Inspector report was a more mixed, but it should be stressed that
309 the margins of difference were typically very small.

310 Next, we investigated the structural variation—insertions, deletions, and translocations—between assemblies.
311 Because we consider the Redbean2.5 assembly to give the best statistics, we used it as the reference and
312 aligned the other assemblies to it using MUMmer and the alignments summarized with NucDiff (Kurtz et al.,
313 2004; Khelik et al., 2017; Marçais et al., 2018). As expected, Stevens and Flye, which use the DBG algorithm,
314 had the fewest number of differences (Table 2 and Table S13). These were followed by the OLC-based
315 SMARTdenovo assembly, the related HGAP Falcon and Falcon-Unzip assemblies, with the OLC-based Canu
316 assembly having the most differences to the Redbean2.5 assembly. For all assemblies, at least 97% of the
317 variants were insertions or deletions, typically short regions (1-to-10bp). However, we did find longer
318 duplications and tandem duplications which could impact accurate curation of genes. To ascertain this, we
319 investigated whether any of the braker 3-annotated genes predicted in both Stevens and Redbean2.5
320 assemblies were on any of the duplications or tandem duplications. For the Stevens assembly, five genes were
321 found in these regions (four genes in tandem duplications and one gene on a duplication). For the Redbean2.5
322 assembly, 19 genes were found in these regions (17 genes in tandem duplications and two on duplications)
323 (Table S14).

324

	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

325 **Table 2 NucDiff results for *C. bovis* assemblies.** Here, the reference assembly is Redbean2.5.

326

327 We did find structural rearrangements, with translocations, relocations, reshufflings, and inversions (Table 2
328 and Figure 2). Again, we focus on assemblies by different versions of the same program—Redbean2.5 and the
329 Stevens's (Redbean2.3) assemblies—we found 194 translocations and the Redbean2.5 Scaffold 3 is split across
330 the Stevens Scaffolds 6, 7, and 11. These would likely have larger consequences in downstream analysis,
331 affecting gene finding and population-based association studies.

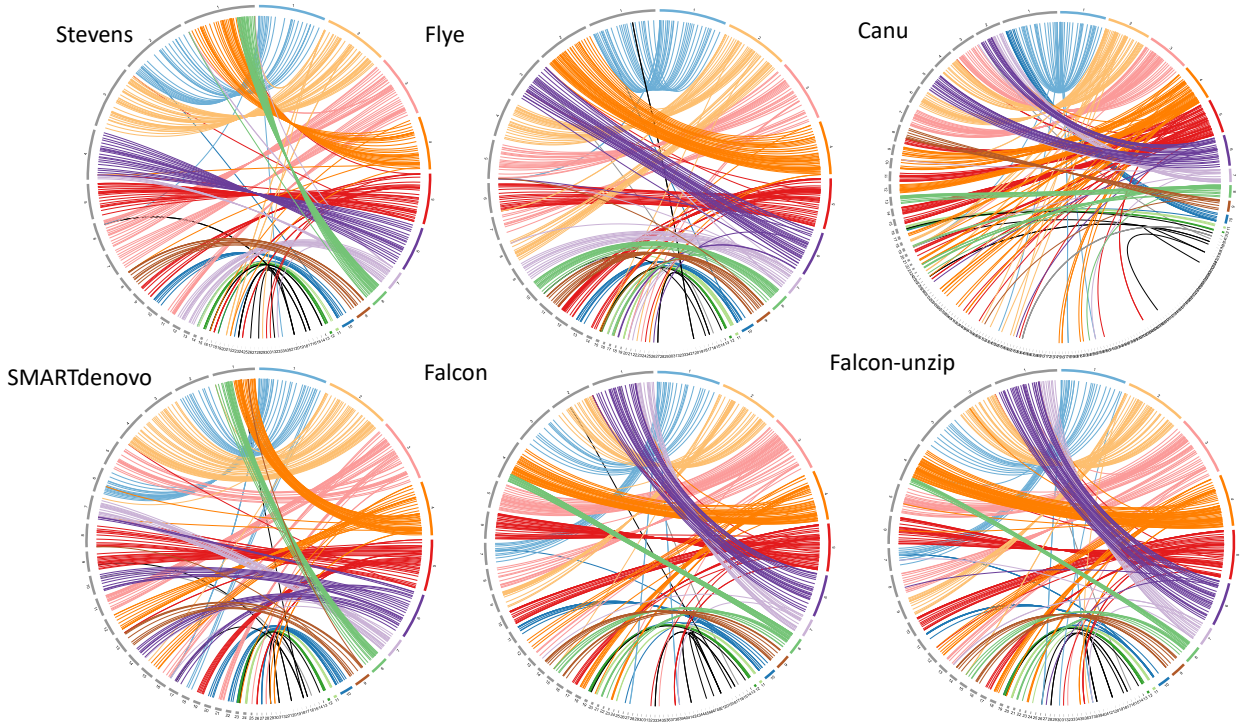

Figure 2: Circos plots showing synteny across six genome assemblies for the *C. bovis*. The colored chromosomes represent the reference genome (Redbean2.5) and the gray chromosomes represent different assemblies as query.

Finally, we note that two scaffolds in the Canu assembly did not align to the Redbean2.5 assembly. Through a BLASTX search against NCBI NR database (Camacho et al., 2009; Sayers et al., 2022), we were able to determine that these are most likely of bacterial origin and not removed by the decontamination protocols (data not shown).

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 (Laing et al., 2013; Doyle et al., 2018, 2020). Here, we use it as the gold-standard reference, termed the Doyle assembly, against which the other assemblies are compared, unless otherwise stated. We generated six assemblies using 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

350 Redbean2.5 and Falcon assemblies were the closest to the Doyle assembly in terms of assembly length, 13 Mb
351 longer and 16 Mb shorter, respectively.

352

	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

353 **Table 3: Assembly statistics for *H. contortus*.** We used BUSCO version 4.1.4 with Augustus. The lineage option
354 used is 'nematoda_odb10', and (C) is for complete genes, (S) for singletons, and (D) for duplicates (all % from
355 3131 genes). We used Inspector version 1.0.2 to evaluate the alignment of *H. contortus* PacBio RSII and Sequel
356 long reads to the generated scaffolds to identify both large-scale (structural) and small-scale errors.
357

358 For the Inspector reports, we used both the reads and the Doyle assembly for reference-free and reference-
359 guided metrics, respectively (Table 3 and Table S15). We were initially surprised to see that the Doyle assembly
360 had the lowest QV score. Here, it is important to note that this score is reference-free, and a function of read
361 alignments to the assembly. The Doyle assembly has undergone extensive curation, in which data, including
362 optical mapping, was used to identify mistakes in the preliminary long-read driven assembly. So, a higher level
363 of disagreement with the raw sequence reads is to be expected. For example, the Doyle assembly had the
364 highest number of haplotype switches (Table S15), likely a consequence of the manual improvement that
365 sought to maximize scaffold continuity. For the six assemblies that we generated, the Inspector metrics were
366 similar. The coverage of the assembly-to-Doyle alignments were relatively low (79%-to-84%). The effect of the
367 alignment algorithm is demonstrated with the 98% coverage Doyle-to-Doyle assembly. Therefore, the regions
368 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 Redbean, Falcon and Falcon-unzip 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 found in at least one assembly, with 2525 genes (80.6%) were complete (singleton or duplicated) in all seven assemblies and 94 genes (3.0%) missing or fragmented in all seven (Figure S2, Figure S3, Figure S4). It is striking that fewer genes were missing in the Canu and SMARTdenovo assemblies compared to the reference. Looking more broadly, of the 141 genes missing in the Doyle assembly, 56 (39.7%) were found complete in 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.

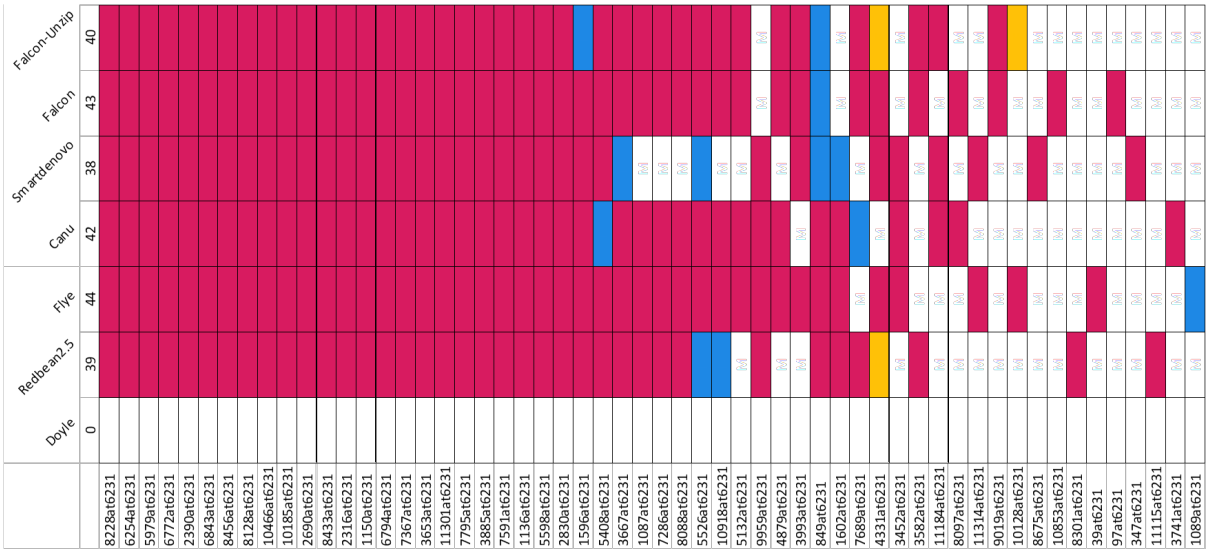

Figure 3: A heatmap showing BUSCO genes missing from the Doyle reference assembly but found present in at least one of the other six assemblies. Red – complete and single, Blue – complete and duplicated, Yellow – Fragmented.

We further examined the functions of the 141 genes absent in the Doyle assembly identifying several genes whose true absence could have significant implications on the *H. contortus* and other Clade V nematodes (Table S16). One example is the *let-756* gene, a lethal locus in the worm. This gene is involved in the fibroblast growth factor receptor signaling pathway essential in nematode larval development. The knock-down of *let-756* expression by RNA interference (RNAi) induces larval arrest in worms (Roubin et al., 1999). A second example is the *nck-1* gene, which enables kinase activity required for neuronal guidance (Mohamed & Chin-Sang, 2011). The knock-down of this gene expression results in worms with impairments in axon guidance,

392 neuronal cell positioning, and abnormalities in the excretory canal cell, gonad, and male mating (Mohamed &
393 Chin-Sang, 2011). A third example is the *gcn-2* gene which is a kinase involved in the nutrient-sensing
394 pathways, impacting several processes that include determining the adult lifespan. Low amino acid supply,
395 leads to loss of *gcn-2* and shortened life spans in worms, supporting its function in stress response under
396 nutrient deprivation (Rousakis et al., 2013).

397 Of the 35 BUSCO genes duplicated in the Doyle reference assembly, 29 were also duplicated in most of the new
398 assemblies and three were singletons in all six new assemblies (Figure S2). Comparing between the new
399 assemblies, there were large exchanges between singleton and duplicated categories, suggesting that assembly
400 algorithms successfully collapsed divergent haplotypes of different genes at different rates (Figure S5 and Data
401 S5). The presence or absence of BUSCOs between the new assemblies was striking. For example, 44 genes
402 considered missing or fragmented in the Doyle reference assembly ('red') were classified as complete and
403 singletons in the Redbean2.5 assembly (Figure S5). Another example is, 62 genes considered missing or
404 fragmented in the Redbean2.5 assembly ('green') were complete and singletons in Flye (Figure S5), with 66
405 genes reclassified in the other direction ('red') (Figure S6 and Data S6). Both assemblers use the DBG algorithm.
406 The full list of BUSCO genes for each assembly can be found in Tables S17-to-S23).

407 *H. bakeri* assemblies

408 At the start of this study, there were two published *H. bakeri* assemblies; one generated from short-reads only
409 (PRJEB1203) and one that used a low-coverage of PacBio reads to scaffold together a short-read assembly
410 (PRJEB15396) (Coghlan et al., 2019; Chow et al., 2019). We selected the second assembly as it was more
411 contiguous and had a higher BUSCO score (Chow et al., 2019). We acknowledge that as we finished our study, a
412 chromosome-scale assembly for *H. bakeri* was published (Stevens et al., 2023). While more contiguous than
413 the Chow assembly—essentially six chromosomes—we note that the Chow assembly had a higher BUSCO score
414 (Table 4).

415 As part of our own *H. bakeri* genome project, we generated HiFi PacBio reads and assembled them with three
416 different software—Flye, HiCanu, and Hifiasm. The assembly lengths differed less than 6%, with the Chow
417 genome being the longest (Table 4). As expected, the N50 and longest scaffold of the new HiFi-based
418 assemblies were considerably longer than the Chow assembly but themselves varied five-fold (from 944 kb to
419 4.9 Mb) for N50 and three-fold (5.9 Mb to 20 Mb) for the longest contigs.

420

421

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*. We used BUSCO version 4.1.4 with Augustus. The lineage option used is 'nematoda_odb10', and C is for complete genes, S for singletons, and D for duplicates (all % from 3131 genes). We used Inspector version 1.0.2 to evaluate the alignment of *H. bakeri* HiFi long reads to the generated scaffolds to identify both large-scale (structural) and small-scale errors. The chromosome-level Stevens *et al.* assembly was only published at the end of 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 (Table 4). As expected, the highly fragmented Chow assembly had a three-fold higher split-read rate (Table S24). Nonetheless, it had the fewest structural errors, suggesting well assembled contigs. The assembly also displayed a higher count of small-scale errors potentially attributed to divergence of sequence data as the HiFi data used for assembly was generated from a Canadian isolate of *H. bakeri*; the Chow assembly is from a British isolate (Table S24).

Marginally more BUSCO genes were found in the HiFi assemblies—Flye, HiCanu, Hifiasm—but at the cost of a duplication rate higher than for the Chow assembly. We found that 164 BUSCO genes were complete (single and duplicated) in at least one assembly; the majority of these were assembly specific (Figure 4A) with 19 found duplicated in all four assemblies and 10 found in the HiFi-based assemblies (Figure 4B). This tempts speculation that the ~29 genes are indeed duplicated. Similar to *C. bovis* and *H. contortus*, the *H. bakeri* assemblies contained different cohorts of BUSCO genes (Figure 5).

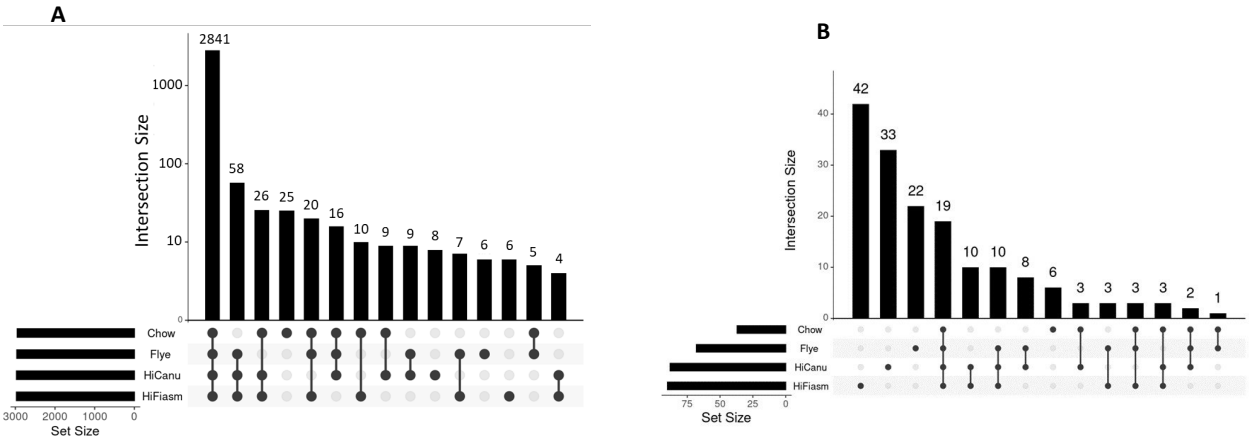

Figure 4: UpSet plots of (A) Single and Duplicated BUSCOs present in *H. bakeri* assemblies, and (B) Duplicated BUSCOs in the assemblies. The figures were generated with https://asntech.shinyapps.io/intervene/.

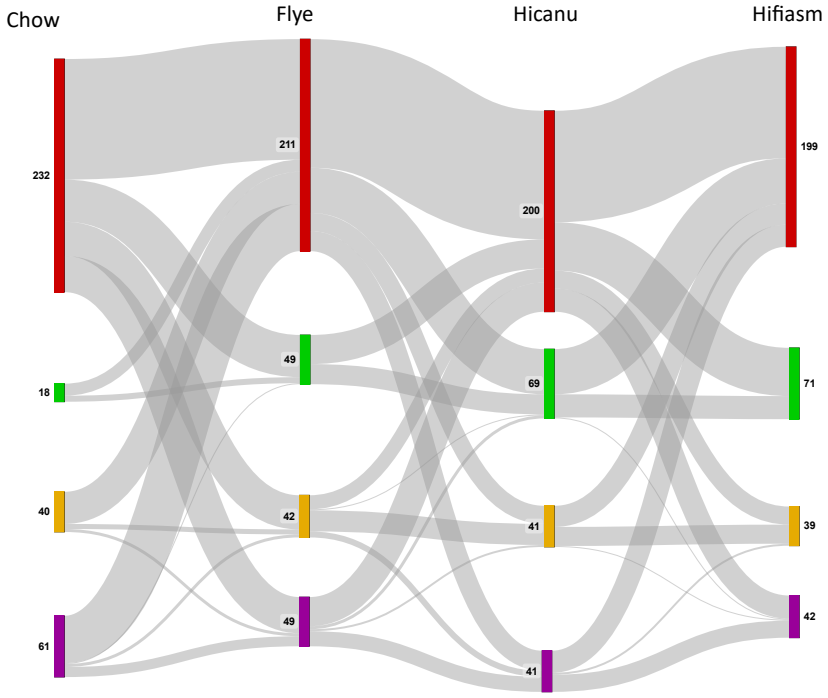

Figure 5: Transitions of BUSCO genes across multiple *H. bakeri* assemblies. A Sankey diagram of the 351 genes that changed categories in at least one *H. bakeri* assembly. The figures were generated using SankeyMATIC (https://sankeymatic.com/). Red – complete & single; Green – complete & duplicated; Yellow – Fragmented; Purple – Missing.

When we considered the union of Single and Duplicated categories as 'present' and the union of Fragmented and Missing as 'absent', we found that between 36 and 80 genes were present in at least one assembly and absent in another (Table 5).

		Absent: Fragmented/Missing			
		Chow *et al.*	Flye	HiCanu	Hifiasm
Present: Single/ Duplicated	Chow *et al.*	-	70	60	55
	Flye	80	-	38	36
	HiCanu	79	47	-	42
	Hifiasm	7 5	46	43	-

Table 5: BUSCO category transition data for *Heligmosomoides bakeri*. An alternative way to show the Sankey diagram data considering the union of Single and Duplicated categories as 'present' and the union of Fragmented and Missing as 'absent'.

A total of 120 genes were present in three assemblies and absent in one (Figure 4 **(A)**). We note that while 58 of these in-three-out-one genes were absent in the Chow assembly, 62 genes were absent 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 (Missing in Hicanu: Table S25, Missing in Hifiasm: Table S26, Missing in Flye: Table S27).

We found several genes whose absence, if real, could significantly impact a *H. bakeri* worm. One example is *mog-1*, absent only in the Flye assembly, encoding a mannosyl-oligosaccharide glucosidase which is involved in N-Glycan biosynthesis (Grinna & Robbins, 1979, 1980; Michael & Kornfeld, 1980; Elting, Chen & Lennarz, 1980; Kilker et al., 1981). The knock-down of *mogs-1* expression by RNA interference (RNAi) results in worms that exhibit slow growth and shortened lifespans (Katoh et al., 2013). A second example is *gpb-1,* the G protein beta subunit, which is also absent in the Flye assembly. RNAi knock-down of *gpb-1* expression is embryonic lethal (Kamath et al., 2000). A third example, this time absent only in the HiCanu assembly, is *fmo-4*, which encodes for flavin-containing monooxygenase and is necessary for the biosynthesis of taurine (Ziegler & Pettit, 1966; Jones & Ballou, 1986; Chiba et al., 1988, 1995; Cashman, 1995; Cashman & Zhang, 2006). RNAi knock-down of *fmo-4* expression causes defects in the nervous system development (axon pathfinding) and is lethal to *C. elegans* (Gujar, Stricker & Lundquist, 2017). The full list of BUSCO genes for each assembly can be found in Table S28, Table S29, Table S30, Table S31).

Impact of assembly variation on gene family analysis

We generated protein-coding gene models for each of the assemblies and identified members of four gene families: the cytochrome P450s (CYP), glutathione S-transferase (GST), UDP glucuronosyltransferases (UGT),

490 and nuclear hormone receptors (NHR). Members of these families have been implicated in anthelmintic drug
491 susceptibility and are among the largest in *C. elegans*: 79 CYPs, 74 GSTs, 67 UGTs and 277 NHRs. The full list of
492 the *C. elegans* gene family members can be found here: CYP450s: Table S32, GSTs: Table S33, UGTs: Table S34,
493 NHRS: Table S35.

494 In each species and for each gene family, we used sequence similarity to identify likely homologues and
495 clustered them, with the *C. elegans* proteins (Emms & Kelly, 2019). If the differences in genome assemblies did
496 not affect the genome annotation, we would expect that every cluster would have an identical number of
497 proteins from each assembly (excluding *C. elegans*). Any deviation from this would demonstrate that
498 differences in the assemblies led to different gene model predictions. We considered cluster membership
499 across all assemblies ('all') and just for those assemblies generated in this study ('new') (Table 6).

500

	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

501

502 **Table 6:** Showing percentage differences between the minimum and maximum number of the
503 predicted members of the CYP, GST, UGT and NHR gene families across assemblies. We considered
504 cluster membership across all assemblies ('all') and just for those assemblies generated in this study
505 ('new').
506

507 Overall, we found considerable variation within gene families, ranging from a difference of three in the *C. bovis*
508 GSTs to a difference of 12 genes in *C. bovis* UGTs (Table 6 and Table 7).

Species	Assembly	Number of protein-coding gene models	Number of CYP450s	Number of GSTs	Number of UGTs	Number of NHRs
C. elegans	**-**	**19,981**	**79**	**74**	**67**	**277**
C. bovis						
	Stevens	14,945	44	39	40	146
	Canu	15,010	50	41	51	160
	Falcon	14,143	58	39	45	144
	Flye	14,070	44	38	39	151
	Redbean2.5	14,001	44	38	41	148
	SMARTdenovo	13,863	44	39	43	145
	Falcon-unzip	13,988	50	38	45	149
H. contortus						
	Doyle	14,824	35	37	36	103
	Canu	18,808	36	39	37	136
	Falcon	16,259	46	40	37	103
	Flye	19,225	59	45	36	131
	Redbean2.5	17,220	47	46	36	121
	SMARTdenovo	17,589	53	49	32	114
	Falcon-unzip	15,890	54	41	31	103
H. bakeri						
	Chow	23,218	40	50	43	141
	Flye	20,801	35	48	37	132
	HiCanu	20,259	36	49	34	127
	Hifiasm	20,079	33	48	32	121

Table 7: Counts of protein-coding gene models and members of the CYP450, GST, UGT and NHR gene families in each assembly for the *C. bovis, H.contortus and H. bakeri*.

Species	Assembly	Number of CYPs	Number of GSTs	Number of UGTs	Number of NHRs
C. bovis	ρ w/ assembly size	0.51	0.85	0.95	0.77
	ρ w/ BUSCO duplication %	0.34	0.85	0.9	0.87
H. contortus	ρ w/ assembly size	0.07	0.39	0.26	0.9
	ρ w/ BUSCO duplication %	0.39	0.66	0.09	0.86
H. bakeri	ρ w/ assembly size	0.66	0.86	0.4	0.35
	ρ w/ BUSCO duplication %	-0.87	-0.69	-0.99	-0.97

Table 8: Showing correlations between gene family count differences between assemblies with assembly size and BUSCO duplication. The correlation coefficient (rho) is indicated by the symbol ρ. The correlations were calculated using the Correlation coefficient calculator - Pearson and Spearman's rank, with solution (statskingdom.com).

518 The difference in gene family counts between assemblies was strongly and positively correlated with the size of
519 the assembly and BUSCO duplications for UGTs in the *C. bovis* assemblies and between the size of the assembly
520 for NHRs in *H. contortus* assemblies (Table 8).

521 Across the 628 clusters that we generated, approximately 60% contained the same number of members for
522 each assembly. Put another way, ~40% of clusters contained a variable number of proteins for each assembly
523 (Table 6). Often, this variation was an additional member in just one assembly.

524 We provide three examples of variable cluster memberships from the CYPs. The first example is cluster bovis-
525 cyp-0000003, which has the same count for the Redbean2.5 and the *C. elegans* (one member each). When
526 compared to *C. elegans* the Falcon and Falcon-unzip assemblies showed expansions (three members each) and
527 the remaining *C. bovis* assemblies (two members each) (Table S36 and Table S37). The *C. elegans* protein is
528 CYP-33C9 whose function has been associated with fatty-acid desaturation hence reducing susceptibility of
529 xenobiotics by increasing desiccation tolerance (Erkut et al., 2013). A second example is cluster bovis-cyp-
530 0000013, which contains one protein from *C. elegans* (CYP-36A1), two members from the Canu assembly and
531 one member each for the remaining *C. bovis* assemblies (Table S36 and Table S37). CYP-36A1 is responsible for
532 the metabolism of tolbutamide (Harlow et al., 2018). A third example is cluster contortus-cyp-0000026 which
533 contains one protein in *C. elegans* (CYP-43A1), and the *H. contortus* falcon, flye and falcon-unzip assemblies but
534 is seemingly absent in the other assemblies (Table S38 and Table S39). The proposed orthologue of CYP-43A1 in
535 an earlier version of the *H. contortus* assembly was shown to be the most highly expressed CYP in the *H.*
536 *contortus* intestine (Laing et al., 2015). The full list of cluster information can be found in Table S36 and Table
537 S37 for *C. bovis* CYPs, Table S38 and Table S39 for *H. contortus* CYPs and Table S40 and Table S41 for *H. bakeri*
538 CYPs.

539

540 ## Discussion

541 The need for more contiguous and complete genomes have led to a myriad of novel computational approaches
542 with continuous advancements in sequencing technology rendering some assembly programs obsolete. We
543 acknowledge that newer long-read assembly programs are available, with others becoming available as we
544 carried out our study, we selected assemblers based on their popularity at the start of our study. The
545 performance of most of these programs has also been benchmarked in several studies using complex genomes
546 (Jung *et al.*, 2020; Sun *et al.*, 2021; Koo *et al.*, 2023; Low *et al.*, 2019; Pollo *et al.*, 2020). We may have reached a
547 point, at which the number of independent comparisons of assembly algorithms exceeds the number of
548 assembly algorithms. However, these comparisons remain useful, particularly considering the tremendous
549 diversity observed in genomes so far sequenced. Acknowledging the impracticality of testing every available or
550 emerging program and the researchers' likelihood to maintain their preferences amidst evolving tools, we
551 aimed to explore the effectiveness of these programs and the impact of the resulting assembly variations on
552 gene finding, focusing on nematodes, species which have a significant effect on animal (including human) and
553 crop health, and, led by *C. elegans*, the fundamental understanding of life's many processes.

554 Long-read assemblers implement two major classes of assembly algorithms: the Overlap-Layout-Consensus
555 (OLC) and the De-Bruijn Graph (DBG) (Flicek & Birney, 2009; Miller, Koren & Sutton, 2010; Schatz, Delcher &
556 Salzberg, 2010). Briefly, the OLC algorithm first identifies overlaps in reads, constructs a layout based on the
557 reads and overlap information and finally builds a consensus. On the other hand, the DBG algorithm works by
558 first trimming reads into shorter sequences of a given length (k-mer) and then using all the k-mers to form
559 repeat graphs and finally merging them to infer the genome sequence (Li et al., 2012). The algorithmic
560 differences and tailored refinements among assemblers, result in varying strategies to handle inherent
561 biological characteristics within genomes, such as repetitive regions and ploidy, as well as error read rates,
562 especially in regions with structural variations. For the smaller and less complex *C. bovis* genome, the DBG-
563 Redbean and Flye assemblers, produced the least fragmented assemblies, with larger N50 and highest BUSCO
564 scores. Further, these assemblers reported the highest mapping read rate to large contigs, alignment depth and
565 the least number of errors, with Redbean having the fewest structural errors. Despite the low heterozygosity in
566 *C. bovis*, the OLC- Canu assembler generated a 10% larger assembly with more duplicated BUSCO genes than
567 other assemblies. This highlights the potential difficulties in handling repeats and heterozygosity by Canu. For
568 the larger and more complex *H. contortus* genome, OLC- SMARTdenovo and Canu assemblers, reported larger
569 N50 sizes, higher BUSCO scores, higher mapping read rates, and lower read split rates. Although Canu
570 incorporates a base-correction step in its algorithm, it produced an assembly with the highest duplication rates.
571 The DBG- Redbean assembler had the lowest duplication rate and fewest structural errors, demonstrating its

572 effectiveness in handling genomes with repeats and allelic variations. SMARTdenovo, Canu, and Redbean
573 assemblers reported the highest percentage coverage to the Doyle reference assembly. We encourage
574 researchers to consider Redbean when selecting assemblers, even for genomes with greater complexities. For
575 *H. bakeri*, the Hifiasm assembler, despite producing the most contiguous assembly, displayed the highest
576 number of structural errors. This suggests potential challenges in handling highly heterozygous regions by the
577 assembly algorithm, despite its default haplotype-resolving capabilities.

578 Assembly variations, as expected, resulted in differentially classified BUSCO genes across assemblers. Given the
579 high conservation of BUSCO genes within a lineage, the shared presence or absence of these genes across
580 multiple assemblers may reflect inherent biological characteristics and evolutionary events, like duplications,
581 within a species. However, we highlight that while the count of duplicated BUSCO genes provides a metric for
582 duplication, the specific genes involved and the context in which they undergo duplication may vary among
583 assemblies. Similarly, BUSCO genes consistently absent in multiple assemblies may suggest gene loss
584 presumably due to pseudogenes. Pseudogenes are prevalent in eukaryotic genomes and typically, have short
585 life spans attributed to chance DNA replication or repair errors and accumulated mutational changes that
586 render them unrecognizable (Chu et al., 2021). Therefore, considering alternative assemblies may facilitate
587 more accurate conclusions regarding the presence or absence of true biological elements within genomes.
588 Alternative assemblies can also guide merging of multiple assemblies to increase contiguity and identify or
589 resolve mis-assemblies. For example, the larger translocations shared by multiple assemblies in *C. bovis* hint at
590 how supporting assemblies could be merged to give a more contiguous assembly. Joining Scaffolds 4 and 8 from
591 Redbean2.5 is supported by their alignments with four assemblies—Stevens, SmartDenovo, Falcon, and Falcon-
592 unzip—which share the translocation in the same region (Figure 2). Sequences from the five assemblies could
593 be aligned to each other and the overlapping regions integrated into a single assembly either by concatenating
594 sequences or by selecting the best representative for each region. However, merging Scaffolds 1 and 10 from
595 Redbean2.5 is only supported by their alignments with the Canu assembly (Figure 2) suggesting a possible mis-
596 assembly in these regions.

597 We identified the functions of some BUSCO genes absent in the *H. contortus* Doyle reference genome and the
598 HiFi-based *H. bakeri* assemblies. Given that these missing genes are part of the BUSCO set for the nematode
599 lineage, their presence across the nematodes is expected, especially considering the relatively close
600 relationship between *C. elegans*, *H. contortus* and *H. bakeri* (all being Clade V nematodes). Therefore, it is not
601 too bold a suggestion that the detrimental effects of loss-of-function in *C. elegans* would equally harm *H.*
602 *contortus* and *H. bakeri*. Faced with such striking loss in a key pathway of a favorite organism, a researcher may
603 formulate alternative and/or novel pathways. Instances of this phenomenon are evident in the literature, and

604 as peer-reviewers, we have encountered and experienced it in our research into nematode biology. This
605 phenomenon is not limited to nematodes and should be taken into consideration by researchers utilizing
606 genome assemblies from their favorite organisms.

607 Comparative analyses often involve species assembled and annotated using different methods across different
608 laboratories (Drukewitz & von Reumont, 2019; Coghlan et al., 2019). However, this practice has been shown to
609 inflate lineage-specific genes due to annotation heterogeneity, where orthologous DNA sequences are
610 annotated as genes in one species and as lineage-specific in another (Weisman *et al.*, 2022). To avoid
611 introducing such biases, we used a uniform annotation method, Braker 3. Unlike in the highly conserved BUSCO
612 genes, assembly variations were more pronounced in the less conserved genes and increased with genome
613 complexity. There was an 8% percentage difference in the predicted gene models for *C. bovis*, 14% for *H.*
614 *bakeri*, and 25% for *H. contortus* assemblies. We acknowledge that efforts to reduce heterozygosity in the *H.*
615 *contortus* and *H. bakeri* assemblies may have resulted in potential differential gene losses across assemblies.
616 Results also showed considerable variations in the member counts of the four gene families under study.
617 Although most variation, based on cluster membership, was an additional member in just one assembly, the
618 variation could be larger with important ramifications on biological interpretation. Using examples from the *C.*
619 *bovis* and *H. contortus* CYP family we demonstrated how variations in assembly can produce results that may
620 lead to inaccurate conclusions. The CYPs are arguably the family with the best functional annotation, however
621 studies into the other families are increasing and variation in the assemblies and annotation could have
622 important consequences. Particularly for gene families that influence drug susceptibility, accurate gene
623 prediction is key for drug design as it guides inferences crucial for determining toxicity and optimal drug
624 dosages.

625 Different assembly algorithms link contigs in different ways, potentially altering their order and orientation, and
626 consequently changing gene structures. The more common annotation errors in eukaryotic genomes are
627 attributed to the complexity of intron-exon structures and may result in either an overestimation or
628 underestimation of genes. Overestimations of genes potentially arise from either "split" or "cleaved" genes
629 (Denton et al., 2014). Split genes are a result of the splitting of haplotypes (allelic variations) in heterozygous
630 organisms into separate contigs which often appear as tandem duplications or with one copy independently on
631 a small scaffold (Denton et al., 2014). Cleaved genes on the other hand, are a result of the fragmentation of a
632 single gene onto multiple scaffolds. These two gene categories can cause an annotation program to predict a
633 single split gene as multiple separate genes, or alternatively, to fail in connecting distant exons altogether
634 (Denton et al., 2014). Underestimations of genes are often caused by insufficient representation of genes in
635 low-coverage regions, such as gaps, and the collapse of true paralogous gene copies into a single locus—a

636 phenomenon particularly affecting newly duplicated genes that have not yet undergone significant divergence
637 (Denton et al., 2014).

638 Eukaryotic organisms such as nematodes, therefore, require more contiguous assemblies to avoid splitting
639 genes across gene scaffolds. Thus, assembly methods that effectively resolve repeats and allelic variations are
640 critical. Similarly, it is important to examine the features of genes of interest and assess the functionalities
641 and/or parameters of the chosen assembly method. Genes that are short exons may go undetected or get
642 discarded by some assemblers. Canu, for instance, only selects the longest 40-fold of reads for correction
643 before assembly (Koren et al., 2017). This process could potentially lead to loss of genes.

644 In summary, algorithmic differences and genome characteristics of species under study impact the accuracy and
645 completeness of assembled genomes and the subsequent genome annotation process. Considering the varying
646 impact of each assembly method on genomic feature identification and the absence of a single gold standard
647 assembler, researchers should carefully select methods that best suit the genomic traits of their study species
648 and research objectives. For instance, contiguity and genome organization, particularly gene order, are crucial in
649 population genetics studies that seek to identify regions under selection. For this reason, assemblers that excel
650 in contiguity are better choices. Similarly, studies that investigate important gene families require assembly
651 methods that demonstrate better genome accuracy and completeness.

652 Conclusion

653 We encourage those charged with assembling genomes—whether nematode, nautilus, or natterjack—to be
654 aware of the important differences between seemingly very similar assemblies. We emphasize the importance
655 of exploring alternative assemblies generated by multiple assemblers to validate biological hypotheses. Finally,
656 these results demonstrate the need to support the continual curation of an organism's assembly and
657 annotation (Doyle, 2022).

658 Acknowledgements

659 We acknowledge the high-performance computing resources made available by the Faculty of Veterinary
660 Medicine and Research Computing at the University of Calgary.

661 Funding

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

664 ## References

665 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of*
666 *Molecular Biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

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

669 Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for
670 analyzing and managing BAM files. *Bioinformatics (Oxford, England)* 27:1691–1692. DOI:
671 10.1093/bioinformatics/btr174.

672 Bethony JM, Loukas A, Hotez PJ, Knox DP. 2006. Vaccines against blood-feeding nematodes of humans and
673 livestock. *Parasitology* 133 Suppl:S63-79. DOI: 10.1017/S0031182006001818.

674 Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. 2012. Domain enhanced lookup time
675 accelerated BLAST. *Biology Direct* 7:12. DOI: 10.1186/1745-6150-7-12.

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

687 Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome
688 annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and*
689 *Bioinformatics* 3:lqaa108. DOI: 10.1093/nargab/lqaa108.

690 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*
691 12:59–60. DOI: 10.1038/nmeth.3176.

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

694 Cable J, Harris PD, Lewis JW, Behnke JM. 2006. Molecular evidence that Heligmosomoides polygyrus from
695 laboratory mice and wood mice are separate species. *Parasitology* 133:111–122. DOI:
696 10.1017/S0031182006000047.

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

699 Cashman JR. 1995. Structural and catalytic properties of the mammalian flavin-containing monooxygenase.
700 *Chemical Research in Toxicology* 8:166–181. DOI: 10.1021/tx00044a001.

701 Cashman JR, Zhang J. 2006. Human flavin-containing monooxygenases. *Annual Review of Pharmacology and*
702 *Toxicology* 46:65–100. DOI: 10.1146/annurev.pharmtox.46.120604.141043.

703 Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. 2021. Accurate long-read de novo assembly evaluation with
704 Inspector. *Genome Biology* 22:312. DOI: 10.1186/s13059-021-02527-4.

705 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased
706 assembly graphs with hifiasm. *Nature Methods* 18:170–175. DOI: 10.1038/s41592-020-01056-5.

Chiba K, Kobayashi K, Itoh K, Itoh S, Chiba T, Ishizaki T, Kamataki T. 1995. N-oxygenation of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine by the rat liver flavin-containing monooxygenase expressed in yeast cells. *European Journal of Pharmacology* 293:97–100. DOI: 10.1016/0926-6917(95)90023-3.

Chiba K, Kubota E, Miyakawa T, Kato Y, Ishizaki T. 1988. Characterization of hepatic microsomal metabolism as an in vivo detoxication pathway of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine in mice. *The Journal of Pharmacology and Experimental Therapeutics* 246:1108–1115.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10:563–569. DOI: 10.1038/nmeth.2474.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13:1050–1054. DOI: 10.1038/nmeth.4035.

Chow FW-N, Koutsovoulos G, Ovando-Vázquez C, Neophytou K, Bermúdez-Barrientos JR, Laetsch DR, Robertson E, Kumar S, Claycomb JM, Blaxter M, Abreu-Goodger C, Buck AH. 2019. Secretion of an Argonaute protein by a parasitic nematode and the evolution of its siRNA guides. *Nucleic Acids Research* 47:3594–3606. DOI: 10.1093/nar/gkz142.

Chu X, Li S, Wang S, Luo D, Luo H. 2021. Gene loss through pseudogenization contributes to the ecological diversification of a generalist Roseobacter lineage. *The ISME Journal* 15:489–502. DOI: 10.1038/s41396-020-00790-0.

Coghlan A, Tyagi R, Cotton JA, Holroyd N, Rosa BA, Tsai IJ, Laetsch DR, Beech RN, Day TA, Hallsworth-Pepin K, Ke H-M, Kuo T-H, Lee TJ, Martin J, Maizels RM, Mutowo P, Ozersky P, Parkinson J, Reid AJ, Rawlings ND, Ribeiro DM, Swapna LS, Stanley E, Taylor DW, Wheeler NJ, Zamanian M, Zhang X, Allan F, Allen JE, Asano K, Babayan SA, Bah G, Beasley H, Bennett HM, Bisset SA, Castillo E, Cook J, Cooper PJ, Cruz-Bustos T, Cuéllar C, Devaney E, Doyle SR, Eberhard ML, Emery A, Eom KS, Gilleard JS, Gordon D, Harcus Y, Harsha B, Hawdon JM, Hill DE, Hodgkinson J, Horák P, Howe KL, Huckvale T, Kalbe M, Kaur G, Kikuchi T, Koutsovoulos G, Kumar S, Leach AR, Lomax J, Makepeace B, Matthews JB, Muro A, O'Boyle NM, Olson PD, Osuna A, Partono F, Pfarr K, Rinaldi G, Foronda P, Rollinson D, Samblas MG, Sato H, Schnyder M, Scholz T, Shafie M, Tanya VN, Toledo R, Tracey A, Urban JF, Wang L-C, Zarlenga D, Blaxter ML, Mitreva M, Berriman M, International Helminth Genomes Consortium. 2019. Comparative genomics of the major parasitic worms. *Nature Genetics* 51:163–174. DOI: 10.1038/s41588-018-0262-1.

Core R. 2022. TEAM, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *Online: https://www. r-project. org*.

Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences of the United States of America* 119:e2115642118. DOI: 10.1073/pnas.2115642118.

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. 2014. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLOS Computational Biology* 10:e1003998. DOI: 10.1371/journal.pcbi.1003998.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 29:15–21. DOI: 10.1093/bioinformatics/bts635.

Doyle SR. 2022. Improving helminth genome resources in the post-genomic era. *Trends in Parasitology* 38:831–840. DOI: 10.1016/j.pt.2022.06.002.

Doyle SR, Laing R, Bartley DJ, Britton C, Chaudhry U, Gilleard JS, Holroyd N, Mable BK, Maitland K, Morrison AA, Tait A, Tracey A, Berriman M, Devaney E, Cotton JA, Sargison ND. 2018. A Genome Resequencing-Based

753 Genetic Map Reveals the Recombination Landscape of an Outbred Parasitic Nematode in the Presence
754 of Polyploidy and Polyandry. *Genome Biology and Evolution* 10:396–409. DOI: 10.1093/gbe/evx269.

755 Doyle SR, Tracey A, Laing R, Holroyd N, Bartley D, Bazant W, Beasley H, Beech R, Britton C, Brooks K, Chaudhry
756 U, Maitland K, Martinelli A, Noonan JD, Paulini M, Quail MA, Redman E, Rodgers FH, Sallé G, Shabbir
757 MZ, Sankaranarayanan G, Wit J, Howe KL, Sargison N, Devaney E, Berriman M, Gilleard JS, Cotton JA.
758 2020. Genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus
759 contortus, a model gastrointestinal worm. *Communications Biology* 3:656. DOI: 10.1038/s42003-020-
760 01377-3.

761 Drukewitz SH, von Reumont BM. 2019. The Significance of Comparative Genomics in Modern Evolutionary
762 Venomics. *Frontiers in Ecology and Evolution* 7.

763 Ebenezer TE, Muigai AWT, Nouala S, Badaoui B, Blaxter M, Buddie AG, Jarvis ED, Korlach J, Kuja JO, Lewin HA,
764 Majewska R, Mapholi N, Maslamoney S, Mbo'o-Tchouawou M, Osuji JO, Seehausen O, Shorinola O,
765 Tiambo CK, Mulder N, Ziyomo C, Djikeng A. 2022. Africa: sequence 100,000 species to safeguard
766 biodiversity. *Nature* 603:388–392. DOI: 10.1038/d41586-022-00712-4.

767 Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K,
768 Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A,
769 Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma
770 C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J,
771 Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F,
772 Korlach J, Turner S. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*
773 323:133–138. DOI: 10.1126/science.1162986.

774 Elting JJ, Chen WW, Lennarz WJ. 1980. Characterization of a glucosidase involved in an initial step in the
775 processing of oligosaccharide chains. *The Journal of Biological Chemistry* 255:2325–2331.

776 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome*
777 *Biology* 20:238. DOI: 10.1186/s13059-019-1832-y.

778 Erkut C, Vasilj A, Boland S, Habermann B, Shevchenko A, Kurzchalia TV. 2013. Molecular Strategies of the
779 Caenorhabditis elegans Dauer Larva to Survive Extreme Desiccation. *PLOS ONE* 8:e82473. DOI:
780 10.1371/journal.pone.0082473.

781 Flicek P, Birney E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*
782 6:S6–S12. DOI: 10.1038/nmeth.1376.

783 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated
784 genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*
785 117:9451–9457. DOI: 10.1073/pnas.1921046117.

786 Gilleard JS. 2006. Understanding anthelmintic resistance: the need for genomics and genetics. *International*
787 *Journal for Parasitology* 36:1227–1239. DOI: 10.1016/j.ijpara.2006.06.010.

788 Grinna LS, Robbins PW. 1979. Glycoprotein biosynthesis. Rat liver microsomal glucosidases which process
789 oligosaccharides. *The Journal of Biological Chemistry* 254:8814–8818.

790 Grinna LS, Robbins PW. 1980. Substrate specificities of rat liver microsomal glucosidases which process
791 glycoproteins. *The Journal of Biological Chemistry* 255:2255–2258.

792 Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic
793 duplication in primary genome assemblies. *Bioinformatics (Oxford, England)* 36:2896–2898. DOI:
794 10.1093/bioinformatics/btaa025.

795 Gujar MR, Stricker AM, Lundquist EA. 2017. Flavin monooxygenases regulate Caenorhabditis elegans axon
796 guidance and growth cone protrusion with UNC-6/Netrin signaling and Rac GTPases. *PLoS genetics*
797 13:e1006998. DOI: 10.1371/journal.pgen.1006998.

798 Harlow PH, Perry SJ, Stevens AJ, Flemming AJ. 2018. Comparative metabolism of xenobiotic chemicals by
799 cytochrome P450s in the nematode Caenorhabditis elegans. *Scientific Reports* 8:13333. DOI:
800 10.1038/s41598-018-31215-w.

801 Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome
802 Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics (Oxford, England)* 32:767–769. DOI:
803 10.1093/bioinformatics/btv661.

804 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. *Methods in*
805 *Molecular Biology (Clifton, N.J.)* 1962:65–95. DOI: 10.1007/978-1-4939-9173-0_5.

806 Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. WormBase ParaSite − a comprehensive resource for
807 helminth genomics. *Molecular and Biochemical Parasitology* 215:2–10. DOI:
808 10.1016/j.molbiopara.2016.11.005.

809 Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of
810 repetitive DNA families. *Nucleic Acids Research* 44:D81–D89. DOI: 10.1093/nar/gkv1272.

811 Jones KC, Ballou DP. 1986. Reactions of the 4a-hydroperoxide of liver microsomal flavin-containing
812 monooxygenase with nucleophilic and electrophilic substrates. *The Journal of Biological Chemistry*
813 261:2553–2559.

814 Jung H, Jeon M-S, Hodgett M, Waterhouse P, Eyun S. 2020. Comparative Evaluation of Genome Assemblers
815 from Long-Read Sequencing for Plants and Crops. *Journal of Agricultural and Food Chemistry* 68:7670–
816 7677. DOI: 10.1021/acs.jafc.0c01647.

817 Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama
818 H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient de novo assembly of highly heterozygous
819 genomes from whole-genome shotgun short reads. *Genome Research*:gr.170720.113. DOI:
820 10.1101/gr.170720.113.

821 Kamath RS, Martinez-Campos M, Zipperlen P, Fraser AG, Ahringer J. 2000. Effectiveness of specific RNA-
822 mediated interference through ingested double-stranded RNA in Caenorhabditis elegans. *Genome*
823 *Biology* 2:research0002.1. DOI: 10.1186/gb-2000-2-1-research0002.

824 Kaminsky R, Ducray P, Jung M, Clover R, Rufener L, Bouvier J, Weber SS, Wenger A, Wieland-Berghausen S,
825 Goebel T, Gauvry N, Pautrat F, Skripsky T, Froelich O, Komoin-Oka C, Westlund B, Sluder A, Mäser P.
826 2008. A new class of anthelmintics effective against drug-resistant nematodes. *Nature* 452:176–180.
827 DOI: 10.1038/nature06722.

828 Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional
829 Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* 428:726–731.
830 DOI: 10.1016/j.jmb.2015.11.006.

831 Katoh T, Takase J, Tani Y, Amamoto R, Aoshima N, Tiemeyer M, Yamamoto K, Ashida H. 2013. Deficiency of α-
832 glucosidase I alters glycoprotein glycosylation and lifespan in Caenorhabditis elegans. *Glycobiology*
833 23:1142–1151. DOI: 10.1093/glycob/cwt051.

834 Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. 2017. NucDiff: in-depth characterization and
835 annotation of differences between two sets of DNA sequences. *BMC bioinformatics* 18:338. DOI:
836 10.1186/s12859-017-1748-z.

837 Kilker RD, Saunier B, Tkacz JS, Herscovics A. 1981. Partial purification from Saccharomyces cerevisiae of a
838 soluble glucosidase which removes the terminal glucose from the oligosaccharide Glc3Man9GlcNAc2.
839 *The Journal of Biological Chemistry* 256:5299–5603.

840 Kinkar L, Gasser RB, Webster BL, Rollinson D, Littlewood DTJ, Chang BCH, Stroehlein AJ, Korhonen PK, Young
841 ND. 2021. Nanopore Sequencing Resolves Elusive Long Tandem-Repeat Regions in Mitochondrial
842 Genomes. *International Journal of Molecular Sciences* 22:1811. DOI: 10.3390/ijms22041811.

843 Knox DP, Redmond DL, Newlands GF, Skuce PJ, Pettit D, Smith WD. 2003. The nature and prospects for gut
844 membrane proteins as vaccine candidates for Haemonchus contortus and other ruminant
845 trichostrongyloids. *International Journal for Parasitology* 33:1129–1137. DOI: 10.1016/S0020-
846 7519(03)00167-X.

847 Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, Odom D, Flicek P, Keane TM, Thybert D,
848 Paten B, Pham S. 2018. Chromosome assembly of large and complex genomes using multiple
849 references. *Genome Research* 28:1720–1732. DOI: 10.1101/gr.236273.118.

850 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs.
851 *Nature Biotechnology* 37:540–546. DOI: 10.1038/s41587-019-0072-8.

852 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
853 annotation of polyploidy woody plants, Hibiscus syriacus L. using PacBio and Nanopore platforms.
854 *Scientific Data* 10:713. DOI: 10.1038/s41597-023-02631-z.

855 Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy
856 AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*
857 36:1174–1182. DOI: 10.1038/nbt.4277.

858 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read
859 assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 27:722–736. DOI:
860 10.1101/gr.215087.116.

861 Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An
862 information aesthetic for comparative genomics. *Genome Research* 19:1639–1645. DOI:
863 10.1101/gr.092759.109.

864 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open
865 software for comparing large genomes. *Genome Biology*:9.

866 Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. DOI:
867 10.12688/f1000research.12232.1.

868 Laing R, Bartley DJ, Morrison AA, Rezansoff A, Martinelli A, Laing ST, Gilleard JS. 2015. The cytochrome P450
869 family in the parasitic nematode Haemonchus contortus. *International Journal for Parasitology* 45:243–
870 251. DOI: 10.1016/j.ijpara.2014.12.001.

871 Laing R, Kikuchi T, Martinelli A, Tsai IJ, Beech RN, Redman E, Holroyd N, Bartley DJ, Beasley H, Britton C, Curran
872 D, Devaney E, Gilabert A, Hunt M, Jackson F, Johnston SL, Kryukov I, Li K, Morrison AA, Reid AJ, Sargison
873 N, Saunders GI, Wasmuth JD, Wolstenholme A, Berriman M, Gilleard JS, Cotton JA. 2013. The genome
874 and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery.
875 *Genome biology* 14:R88. DOI: 10.1186/gb-2013-14-8-r88.

876 Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. *Nucleic Acids Research* 39:D19–D21.
877 DOI: 10.1093/nar/gkq1019.

878 LeJambre LF, Windon RG, Smith WD. 2008. Vaccination against Haemonchus contortus: Performance of native
879 parasite gut membrane glycoproteins in Merino lambs grazing contaminated pasture. *Veterinary*
880 *Parasitology* 153:302–312. DOI: 10.1016/j.vetpar.2008.01.032.

881 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE*
882 *transactions on visualization and computer graphics* 20:1983–1992. DOI: 10.1109/TVCG.2014.2346248.

883 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. DOI:
884 10.48550/arXiv.1303.3997.

885 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. DOI:
886 10.1093/bioinformatics/bty191.

887 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
888 major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in*
889 *Functional Genomics* 11:25–37. DOI: 10.1093/bfgp/elr035.

890 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence
891 Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI:
892 10.1093/bioinformatics/btp352.

893 Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. 2016. Assembly of long error-prone reads
894 using de Bruijn graphs. *Proceedings of the National Academy of Sciences* 113:E8396–E8405. DOI:
895 10.1073/pnas.1604560113.

896 Liu H, Wu S, Li A, Ruan J, Wu S, Li A, Ruan J. 2021. SMARTdenovo: a de novo assembler using long noisy reads.
897 *Gigabyte* 2021:1–9. DOI: 10.46471/gigabyte.15.

898 Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre
899 L, Hume DA, Collins A, Ajmone-Marsan P, Smith TPL, Williams JL. 2019. Chromosome-level assembly of
900 the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nature*
901 *Communications* 10:260. DOI: 10.1038/s41467-018-08260-0.

902 Maizels RM, Hewitson JP, Murray J, Harcus YM, Dayer B, Filbey KJ, Grainger JR, McSorley HJ, Reynolds LA, Smith
903 KA. 2012. Immune modulation and modulators in Heligmosomoides polygyrus infection. *Experimental*
904 *Parasitology* 132:76–89. DOI: 10.1016/j.exppara.2011.08.011.

905 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile
906 genome alignment system. *PLOS Computational Biology* 14:e1005944. DOI:
907 10.1371/journal.pcbi.1005944.

908 Michael JM, Kornfeld S. 1980. Partial purification and characterization of the glucosidases involved in the
909 processing of asparagine-linked oligosaccharides. *Archives of Biochemistry and Biophysics* 199:249–
910 258. DOI: 10.1016/0003-9861(80)90278-7.

911 Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics*
912 95:315–327. DOI: 10.1016/j.ygeno.2010.03.001.

913 Mohamed AM, Chin-Sang ID. 2011. The C. elegans nck-1 gene encodes two isoforms and is required for
914 neuronal guidance. *Developmental Biology* 354:55–66. DOI: 10.1016/j.ydbio.2011.03.019.

915 Msolla P, Semuguruka WD, Kasuku AA, Shoo MK. 1993. Clinical observations on bovine parasitic otitis in
916 Tanzania. *Tropical Animal Health and Production* 25:15–18. DOI: 10.1007/BF02236880.

917 Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020.
918 HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity
919 long reads. :2020.03.14.992248. DOI: 10.1101/2020.03.14.992248.

920 Pollo SMJ, Leon-Coria A, Liu H, Cruces-Gonzalez D, Finney CAM, Wasmuth JD. 2023. Transcriptional patterns of
921 sexual dimorphism and in host developmental programs in the model parasitic nematode
922 Heligmosomoides bakeri. *Parasites & Vectors* 16:171. DOI: 10.1186/s13071-023-05785-2.

923 Pollo SMJ, Reiling SJ, Wit J, Workentine ML, Guy RA, Batoff GW, Yee J, Dixon BR, Wasmuth JD. 2020.
924 Benchmarking hybrid assemblies of Giardia and prediction of widespread intra-isolate structural
925 variation. *Parasites & Vectors* 13:108. DOI: 10.1186/s13071-020-3968-8.

926 Pundir S, Martin MJ, O'Donovan C. 2017. UniProt Protein Knowledgebase. In: Wu CH, Arighi CN, Ross KE eds.
927 *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*. Methods in Molecular
928 Biology. New York, NY: Springer, 41–55. DOI: 10.1007/978-1-4939-6783-4_2.

929 Redman E, Sargison N, Whitelaw F, Jackson F, Morrison A, Bartley DJ, Gilleard JS. 2012. Introgression of
930 ivermectin resistance genes into a susceptible Haemonchus contortus strain by multiple backcrossing.
931 *PLoS pathogens* 8:e1002534. DOI: 10.1371/journal.ppat.1002534.

932 Reynolds LA, Filbey KJ, Maizels RM. 2012. Immunity to the model intestinal helminth parasite Heligmosomoides
933 polygyrus. *Seminars in Immunopathology* 34:829–846. DOI: 10.1007/s00281-012-0347-3.

934 Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*
935 13:278–289. DOI: 10.1016/j.gpb.2015.08.002.

936 Roubin R, Naert K, Popovici C, Vatcher G, Coulier F, Thierry-Mieg J, Pontarotti P, Birnbaum D, Baillie D, Thierry-
937 Mieg D. 1999. let-756, a C. elegans fgf essential for worm development. *Oncogene* 18:6741–6747. DOI:
938 10.1038/sj.onc.1203074.

939 Rousakis A, Vlassis A, Vlanti A, Patera S, Thireos G, Syntichaki P. 2013. The general control nonderepressible-2
940 kinase mediates stress response and longevity induced by target of rapamycin inactivation in
941 Caenorhabditis elegans. *Aging Cell* 12:742–751. DOI: 10.1111/acel.12101.

942 Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* 17:155–158. DOI:
943 10.1038/s41592-019-0669-3.

944 Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T,
945 Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse
946 T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. 2022. Database resources of the national
947 center for biotechnology information. *Nucleic Acids Research* 50:D20–D26. DOI:
948 10.1093/nar/gkab1112.

949 Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing.
950 *Genome Research* 20:1165–1173. DOI: 10.1101/gr.101360.109.

951 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome
952 assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*
953 31:3210–3212. DOI: 10.1093/bioinformatics/btv351.

954 Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to
955 improve de novo gene finding. *Bioinformatics* 24:637–644. DOI: 10.1093/bioinformatics/btn013.

956 Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized
957 hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62. DOI:
958 10.1186/1471-2105-7-62.

959 Stevens L, Martínez-Ugalde I, King E, Wagah M, Absolon D, Bancroft R, Gonzalez de la Rosa P, Hall JL, Kieninger
960 M, Kloch A, Pelan S, Robertson E, Pedersen AB, Abreu-Goodger C, Buck AH, Blaxter M. 2023. Ancient
961 diversity in host-parasite interaction genes in a model parasitic nematode. *Nature Communications*
962 14:7776. DOI: 10.1038/s41467-023-43556-w.

963 Stevens L, Rooke S, Falzon LC, Machuka EM, Momanyi K, Murungi MK, Njoroge SM, Odinga CO, Ogendo A,
964 Ogola J, Fèvre EM, Blaxter M. 2020. The Genome of Caenorhabditis bovis. *Current biology: CB* 30:1023-
965 1031.e4. DOI: 10.1016/j.cub.2020.01.074.

966 Sun J, Li R, Chen C, Sigwart JD, Kocot KM. 2021a. Benchmarking Oxford Nanopore read assemblers for high-
967 quality molluscan genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*
968 376:20200160. DOI: 10.1098/rstb.2020.0160.

969 Sun J, Li R, Chen C, Sigwart JD, Kocot KM. 2021b. Benchmarking Oxford Nanopore read assemblers for high-
970 quality molluscan genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*
971 376:20200160. DOI: 10.1098/rstb.2020.0160.

972 Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences.
973 *Current Protocols in Bioinformatics* Chapter 4:4.10.1-4.10.14. DOI: 10.1002/0471250953.bi0410s25.

974 Taylor CM, Martin J, Rao RU, Powell K, Abubucker S, Mitreva M. 2013. Using existing drugs as leads for broad
975 spectrum anthelmintics targeting protein kinases. *PLoS pathogens* 9:e1003149. DOI:
976 10.1371/journal.ppat.1003149.

977 Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long
978 uncorrected reads. *Genome Research* 27:737–746. DOI: 10.1101/gr.214270.116.

979 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK,
980 Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome
981 Assembly Improvement. *PLOS ONE* 9:e112963. DOI: 10.1371/journal.pone.0112963.

982 Wasmuth J, Schmid R, Hedley A, Blaxter M. 2008. On the Extent and Origins of Genic Novelty in the Phylum
983 Nematoda. *PLOS Neglected Tropical Diseases* 2:e258. DOI: 10.1371/journal.pntd.0000258.

984 Weisman CM, Murray AW, Eddy SR. 2022. Mixing genome annotation methods in a comparative analysis
985 inflates the apparent number of lineage-specific genes. *Current biology: CB* 32:2632-2639.e2. DOI:
986 10.1016/j.cub.2022.04.085.

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

989 Ziegler DM, Pettit FH. 1966. Microsomal oxidases. I. The isolation and dialkylarylamine oxygenase activity of
990 pork liver microsomes. *Biochemistry* 5:2932–2938. DOI: 10.1021/bi00873a024.

991