Generating reference-quality *de novo* genome assemblies from parasitic nematodes using the Oxford Nanopore Technologies MinION sequencing platform

Kaylee S. Herzog[1], Rachel Wu[1], John M. Hawdon[2], Peter Nejsum[3], and Joseph R. Fauver[1]*

[1]Department of Epidemiology, University of Nebraska Medical Center, Omaha, NE 68198

[2]Department of Microbiology, Immunology, and Tropical Medicine, The George Washington University, Washington, D.C. 20037

[3]Department of Clinical Medicine, Aarhus University, 8200 Aarhus, Denmark

*Author for Correspondence: Joseph R. Fauver, Department of Epidemiology, University of Nebraska Medical Center, Omaha, NE 68198, (402) 552-3397, (402) 559-4961, jfauver@unmc.edu

Abstract

Parasitic nematode infections represent a significant burden of disease in impoverished populations. High prevalence of infection coupled with widespread geographic distributions means populations of parasitic nematodes likely harbor high levels of genetic diversity that are not reflected in currently available genomic resources. Comparative genomic studies of parasitic nematodes have identified novel drug and vaccine targets and yielded unprecedented insights into helminth biology, highlighting the potential of high-quality reference genomes. Few genomic resources exist for parasitic nematodes, however, and most species are represented by only a single genome. This is partly due to challenges associated with generating genomic data for these parasites. Many species are small in size and yet have relatively large and complex genomes, making their genomes difficult and expensive to sequence and assemble. Short-read sequence data has primarily been used to generate genomic data for parasitic nematodes, but whole genome assemblies based on short-read data are highly discontiguous. Recent improvements in the accuracy of Oxford Nanopore Technologies (ONT) long-read data makes it an appealing approach for improving existing reference genomes and generating novel genome assemblies from uncharacterized species. This study targets three species of parasitic nematodes to compare the current gold standard hybrid *de novo* genome assembly approach to one that relies on long-read ONT MinION data only. Hybrid assemblies were found to be more contiguous, but nonetheless, MinION data alone were used successfully to generate highly contiguous, complete, and accurate assemblies for all species quickly and inexpensively.

Key Words

Parasitic nematodes, whole genome sequencing, genome assembly, Oxford Nanopore Technologies, MinION sequencing, BUSCO

Significant Statement

High quality helminth genome assemblies are invaluable tools for identifying novel drug and vaccine targets and monitoring the effectiveness of control campaigns, but generating regionally relevant reference-quality genomes from a broad diversity of populations and species can be challenging when relying on traditional approaches. This study demonstrates that Oxford Nanopore Technologies MinION long-read data alone can be used to quickly and inexpensively generate highly contiguous, complete, and accurate whole genome assemblies for parasitic nematodes. These findings highlight an accessible method for increasing the resources available for comparative helminth genomics and a path toward incorporating genomic epidemiology as a routine component of assessing control campaign effectiveness.

Data Availability Statement

Quality-controlled MinION and Illumina data for each species are deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number XXXX. Final assemblies for each species are publicly available on WormBase ParaSite.

Introduction

Parasitic nematodes inflict an enormous burden of human disease. It is estimated that 1.5 billion people alone are infected with soil-transmitted helminths, while approximately 120 million and 15 million people suffer from filariasis and onchocerciasis, respectively (Global Burden of Disease Study, 2020). In addition to the morbidity and mortality caused by nematodes themselves, infection with these parasites also exacerbates the disease burden of other existing conditions such as malaria, HIV, and tuberculosis, and can reduce immunological response to important vaccines (Ndyomugyenyi et al., 2008; Kizito et al., 2013; Biraro et al., 2014; Morawski et al., 2017; Nash et al., 2017). Accordingly, the World Health Organization (WHO) has established large-scale global control programs to reduce the incidence of nematode infections (Casulli, 2021; International Helminth Genomes Consortium, 2021). These programs are centered around broad distribution of anthelmintic drugs to all individuals in vulnerable communities regardless of their infection status. Because a limited number of inexpensive drugs can be used to treat multiple parasitic infections, these mass drug administration campaigns (i.e., MDAs) are touted as an inexpensive, broadly effective control measure for a variety of nematode parasites. As programmatic monitoring focuses almost exclusively on metrics of infected individuals themselves, however, the question of the impact of repeated exposure to anthelmintic drugs on nematode populations remains unanswered. Recent advances in next generation sequencing (NGS) technologies now allow for the monitoring of parasite population structure over the course of long-term treatment with anthelminthics, thus enabling the application of genomic epidemiological approaches to assess control program efficacy (Berger et al., 2021).

82 Genomic epidemiology requires high quality and regionally specific reference
83 genomes for the species of parasite(s) being treated and monitored (Formenti et al.,
84 2022). This is due to the fact that most genomic epidemiological approaches for
85 eukaryotes rely on population-level datasets in which individuals are genotyped based on
86 mapping reads for loci of interest to a conspecific reference-quality genome (Sims et al.,
87 2014; Hoban et al., 2016). The advent of NGS—and specifically, reductions in its cost
88 and improved ease of access—has recently made the generation of genomic data from
89 many populations of parasites feasible (Doyle 2022). While this has resulted in an
90 increased number of population-level studies for parasites that are based on genomic
91 data (e.g., Brabec et al., 2016; Sweet & Johnson, 2018; Berger et al., 2021; Fraija-
92 Fernández et al., 2021; Doyle et al., 2022; Virrueta Herrera et al., 2022), there has not
93 been an equivalent increase in the generation of reference genomes representative of
94 these populations.

95 The reference genomes that do exist for parasitic nematodes have proven to be
96 an invaluable resource for increasing our understanding of parasite biology, treatment,
97 and control. For example, the landmark comparative genomic study by the International
98 Helminth Genomes Consortium analyzed tens of complete genomes representing all
99 major helminth lineages, including nematodes. This expansive dataset allowed for
100 identification of multiple gene family expansions in various groups that are targets for
101 novel drug and vaccine development (International Helminth Genomes Consortium,
102 2019). Studies such as these highlight the importance of parasite genomes for not only
103 population-level genomic epidemiological approaches, but also as indispensable tools for
104 comparative biomedical research.

105 Historically, short-read sequences like those produced by Illumina NGS platforms
106 have been the data type of choice for assembling parasite genomes. Though highly
107 accurate, short-read data alone are unable to produce contiguous reference genomes for
108 most species (Bashir et al., 2012; English et al., 2012; Koren et al., 2012). The Oxford
109 Nanopore Technologies (ONT) MinION NGS platform offers an appealing complimentary
110 approach. Its benefits include reduced equipment costs and infrastructure requirements
111 as compared to traditional Illumina platforms, ease of use and portability, and, importantly,
112 the ability to generate the long-read sequence data needed to produce contiguous
113 reference-quality genomes. Recent updates to ONT chemistries put the error rates of
114 MinION sequence data on par with those of Illumina platforms, allowing for the potential
115 to generate parasite reference genomes using long-read data, only.
116 This study aims to compare whole genomes of parasitic nematodes assembled *de*
117 *novo* using the current gold standard hybrid approach of combining both short- and long-
118 read data during the assembly process versus using Q20+ ONT MinION long-read data,
119 only. We utilize three species for sequencing which collectively span the breadth of
120 nematode diversity: the filarial worm *Brugia malayi* (Spirurina), the whipworm *Trichuris*
121 *trichiura* (Trichinellida), and the dog hookworm *Ancylostoma caninum* (Rhabditina).
122 Ultimately, the goal of this study is to evaluate whether the improved quality of Q20+
123 MinION long reads make these data alone sufficient for generating reference-quality
124 assemblies, highlighting a quick, inexpensive, and accessible method for generating
125 parasite genomes.
126
127 **Results**

128 DNA extraction, library preparation, and sequencing

129 The total amount of gDNA extracted from each sample type and the mean
130 fragment lengths of extractions are provided in Supplementary Table 1.

131 *MinION library preparation and sequencing:* Figure 1 provides a summary of the
132 bioinformatic pipeline used to process MinION sequence data. The concentrations and
133 mean fragment sizes of MinION libraries are provided in Supplementary Table 1. The
134 amount of data generated from each flow cell and the amount of data successfully
135 basecalled by Guppy are also provided in Supplementary Table 1. Select NanoPlot
136 outputs summarizing each Guppy run are provided for each species separately in
137 Supplementary Figures 1–3. For *T. trichiura* and *A. caninum*, for which more than one
138 flow cell's worth of data were generated, NanoPlot plots are based on input from two (*T.*
139 *trichiura*) or three (*A. caninum*) sequencing summary files. Values for average depth of
140 coverage and proportion of basecalled MinION reads mapped to species-specific
141 reference genomes as estimated using Minimap2 and SAMtools were 124.85× (98.77%
142 of reads mapped) for *B. malayi*, 249.91× (99.91% of reads mapped) for *T. trichiura*, and
143 49.42× (99.85% of reads mapped) for *A. caninum.* For *B. malayi*, percent identity and
144 median multiplicity values output by Mash indicate the majority of MinION reads were
145 classified as *B. malayi* or its *Wolbachia* endosymbiont (see Supplementary Table 2). For
146 *T. trichiura* and *A. caninum*, Mash outputs were less straightforward. Neither species was
147 identified within the MinION read sets analyzed; rather, various lineages of virus, bacteria,
148 and fungi were assigned, largely with relatively low percent identity and median
149 multiplicity values (see Supplementary Table 2).

Figure 1. Bioinformatic pipeline used to generate MinION data (left) and hybrid (right) genome assemblies.

150 *Illumina library preparation and sequencing:* Figure 1 provides a summary of the
151 bioinformatic pipeline used to process Illumina sequence data. The amount of raw
152 sequencing data generated for each Illumina library, and the amount of data retained
153 following the first iteration of filtration with fastp, are provided in Supplementary Table 1.
154 Values for average depth of coverage and proportion of quality-controlled Illumina reads
155 mapped across species-specific reference genomes as estimated using BWA and
156 SAMtools were 82.95× (95.74% of reads mapped) for *B. malayi*, 48.58× (90.71% of reads
157 mapped) for *T. trichiura*, and 38.64× (90.31% of reads mapped) for *A. caninum*. As with
158 MinION data, Mash results indicated that Illumina read datasets for *B. malayi* were
159 overwhelmingly identified as *B. malayi* or its *Wolbachia* endosymbiont but indicated no
160 species-appropriate hits for *T. trichiura* or *A. caninum* (see Supplementary Table 2).
161
162 <u>Estimation of genome size, heterozygosity, and ploidy</u>
163 GenomeScope plots for each species are provided in Supplementary Figure 4.
164 Genome size (and heterozygosity) estimates from Jellyfish and GenomeScope were
165 85,917,606 bp (0.275%) for *B. malayi*, 68,538,097 bp (1.35%) for *T. trichiura,* and
166 329,957,709 (1.09%) for *A. caninum*. Genome size estimates for all three species were
167 lower than the current reference assembly lengths of ~88 Mb for *B. malayi* (Ghedin et al.,
168 2007), ~75–80 Mb for *T. trichiura* (Foth et al., 2014)
169 (https://github.com/stephenrdoyle/ancient_trichuris/tree/master/02_data; Doyle et al.,
170 2022), and ~465 Mb for *A. caninum* (International Helminth Genomes Consortium, 2019).
171 Model fit values estimated by GenomeScope ranged from ~93.4–94.5% (*B. malayi*),
172 ~91.6–93.3% (*T. trichiura*) and ~94.1–96.2% (*A. caninum*). For *T. trichiura*, based on log-

173 likelihood scores, nQuire preferred a triploid model over diploid and tetraploid models for
174 both Illumina and MinION read datasets (see Supplementary Table 3).
175
176 MinION data assembly and refinement
177 Histograms of Q scores for MinION reads versus Q scores for MinION reads
178 aligned to final MinION data assemblies are provided for each species in Supplementary
179 Figure 5. For all three species, a median Q score value of >Q16 was achieved in both
180 categories. *Brugia malayi* had the highest median Q score, approaching Q20 in both
181 categories, followed closely by *T. trichiura*. The BlobPlots and summary of by-contig
182 BLAST results produced by BlobTools for MinION data assemblies are provided in
183 Supplementary Figures 6–8 and Supplementary Table 4, respectively. For *B. malayi* and
184 *A. caninum*, four contigs totaling 49,030 bp and four contigs totaling 116,599 bp,
185 respectively, were confidently identified as potential contamination and were removed
186 from final assemblies (see Supplementary Table 4); however, contigs from the *T. trichiura*
187 and *A. caninum* assemblies returned largely off-target BLAST results (see Supplementary
188 Table 4; Supplementary Figures 7, 8).
189
190 Hybrid assembly and refinement
191 K-mer multiplicity plots output by KAT for purged hybrid assemblies are provided
192 for all three species in Supplementary Figure 9. A single peak in the plot for *B. malayi*
193 indicates low heterozygosity whereas two distinct peaks in the plots for *T. trichiura* and *A.*
194 *caninum* indicates higher levels of heterozygosity. For all three species, KAT plots
195 suggest that errors and—in the case of *T. trichiura* and *A. caninum*, heterozygosity—were

196 successfully purged from final hybrid assemblies. The BlobPlots and summary of by-
197 contig BLAST results produced by BlobTools for hybrid assemblies are provided in
198 Supplementary Figures 10–12 and Supplementary Table 4, respectively. Only a single
199 contig was confidently identified as contamination and eliminated from a hybrid assembly
200 (i.e., for *A. caninum*; see Supplementary Table 4). As with MinION data assemblies, *T.*
201 *trichiura* and *A. caninum* assemblies returned high proportions off-target BLAST results
202 (see Supplementary Table 4; Supplementary Figures 11, 12).
203
204 Assembly contiguity, accuracy, and completeness
205 Assembly length, contig number, N50 value, and percent GC and N content as
206 output by QUAST are provided for MinION data, hybrid, and reference assemblies in
207 Table 1. The number of gSNPs and gINdels identified by MuMmer4 between MinION data
208 and hybrid assemblies, and the nucleotide-level percent identity where these two
209 assemblies align, are also provided in Table 1. Graphs of contig number by genome size
210 are provided for each species in Figure 2. With the exception of the MinION data
211 assembly for *B. malayi*, all assemblies generated here were shorter in terms of total length
212 than the current reference assemblies available for each species (Table 1; Figure 2). In
213 comparison to the genomes available on WormBase ParaSite for *T. trichiura* (Foth et al.,
214 2014), and *A. caninum* (International Helminth Genomes Consortium, 2019), both MinION
215 data and hybrid assemblies were more contiguous (as measured by a smaller contig
216 number and a larger N50 value) than the existing reference assemblies, and hybrid
217 assemblies were in all cases more contiguous than MinION data assemblies (Table 1;
218 Figure 2). By species, all assemblies had similar levels of GC content within, and MinION

Table 1. Comparative quality metrics for the assemblies generated as part of this study and the reference assemblies available for each focal species.

		QUAST					MuMmer4		
		Length (bp)	No. contigs	N50 (bp)	GC content	N content	Nucleotide-level genome identity where aligned	No. gSNPS	No. gIndels
Brugia malayi	WormBase ParaSite reference	88,235,797	197	14,214,749	28.5%	0.30%			
	MinION data-only assembly	88,513,498	88	4,666,641	28.43%	0%	99.34%	150,681	9,434
	Hybrid assembly	84,528,932	49	3,943,761	28.57%	<0.001%	98.68%		
Trichuris trichiura	WormBase ParaSite reference	75,474,068	4,086	70,602	42.3%	0.30%			
	Doyle et al. (2022) reference	80,573,711	113	112,99,416	42.3%	0.31%			
	MinION data-only assembly	73,094,541	174	806,679	42.22%	0%	99.32%	35,115	36,140
	Hybrid assembly	72,705,678	63	27,73,302	41.15%	0%	99.60%		
Ancylostoma caninum	WormBase ParaSite Reference	465,750,606	25,339	256,700	42.8%	13.50%			
	MinION data-only assembly	358,159,610	1,578	638,901	43.36%	0%	97.33%	299,510	1,258,880
	Hybrid assembly	348,034,555	652	1,095,227	43.24%	<0.001%	96.16%		

219 data and hybrid assemblies were comprised of 0%, or essentially 0%, ambiguous
220 nucleotide calls, respectively (Table 1). Nucleotide-level alignment identity between
221 MinION data and hybrid assemblies as assessed by MuMmer4 ranged from ~96.2–99.6%
222 (Table 1).

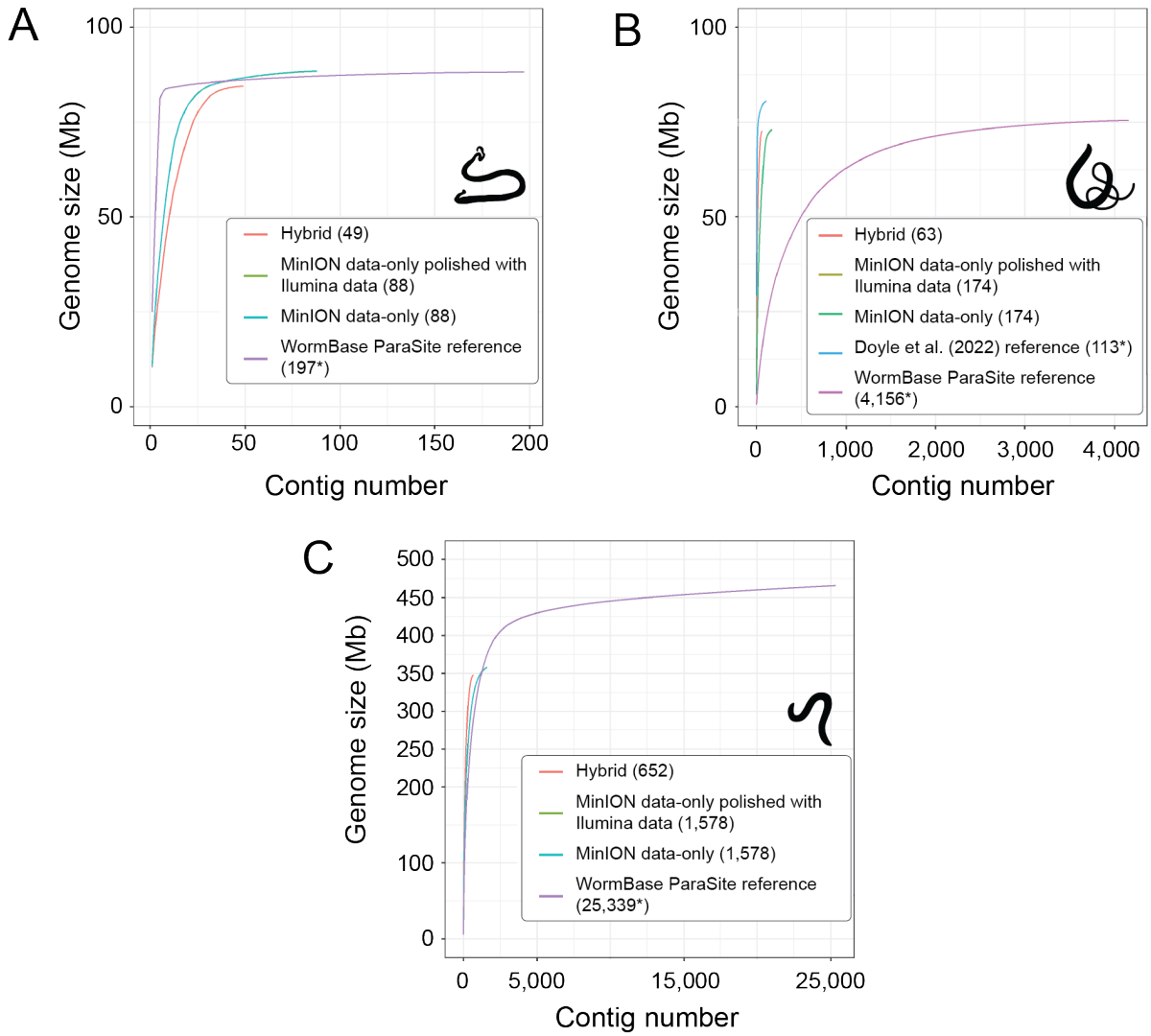

Figure 2. Comparison of contiguity of the assemblies generated as part of this study and existing reference assemblies for (A) *Brugia malayi*, (B) *Trichuris trichiura*, and (C) *Ancylostoma caninum*. Numbers in parenthesis within keys at lower right are the total number of contigs in each assembly; asterisks indicate a scaffold count rather than a contig count.

223

224 Scores from miniBUSCO for all assembly types are summarized for each species
225 in Table 2. For all three species, scores for each category are nearly identical among the
226 three assembly types generated herein (i.e., MinION data, MinION data polished with
227 Illumina data, and hybrid). For *A. caninum*, miniBUSCO scores indicate MinION and
228 hybrid assemblies are more complete as compared to the existing reference assembly.
229 All assemblies compared for *T. trichiura*, including both available references, had high
230 proportions of missing BUSCOs (i.e., ~40.8–42.4%; see Table 2). A summary of the
231 assessment of "tier 1" helminth genome status *sensu* the International Helminth Genomes
232 Consortium (International Helminth Genomes Consortium, 2019) for the assemblies
233 generated here, and assemblies of all nematodes that parasitize animal hosts available
234 from WormBase ParaSite, are presented in Figure 3. Both the MinION data and hybrid
235 assemblies for *B. malayi* and *A. caninum* are classified as "tier 1" genomes. For *T.*
236 *trichiura*, the contiguity of both the MinION data and hybrid assembly is sufficient to qualify
237 them each as "tier 1", but, as mentioned above, all assemblies for trichurids appear to be
238 missing a significant proportion of nematode BUSCOs, preventing any trichurid assembly
239 assessed from achieving "tier 1" status (see Figure 3).

240 In terms of assembly accuracy, one of the most salient questions is how different
241 MinION data assemblies are from assemblies that include both MinION and Illumina data.

Table 2. Scores from miniBUSCO for MinION data-only, MinION data-only polished with Illumina data, hybrid, and reference assemblies for each focal species. Scores are presented as number of BUSCOs recovered in each assembly followed in parentheses by proportion of the total number of nematode orthologs assessed by miniBUSCO (n=3,131).

		Single copy	Duplicated	Fragmented	Missing
Brugia malayi	WormBase ParaSite reference	3,100 (99.01%)	15 (0.48%)	8 (0.26%)	8 (0.26%)
	MinION data-only assembly	3,097 (98.91%)	17 (0.54%)	9 (0.29%)	8 (0.26%)
	MinION assembly + Illumina polishing	3,097 (98.91%)	17 (0.54%)	9 (0.29%)	8 (0.26%)
	Hybrid assembly	3,096 (98.88%)	17 (0.54%)	8 (0.26%)	10 (0.32%)
Trichuris trichiura	WormBase ParaSite reference	1,676 (53.53%)	15 (0.48%)	112 (3.58%)	1,328 (42.41%)
	Doyle et al. (2022) Reference	1,702 (54.36%)	46 (1.47%)	102 (3.26%)	1,281 (40.91%)
	MinION data-only assembly	1,717 (54.84%)	20 (0.64%)	99 (3.16%)	1,295 (41.36%)
	MinION assembly + Illumina polishing	1,716 (54.81%)	20 (0.64%)	100 (3.19%)	1,295 (41.36%)
	Hybrid assembly	1,725 (55.09%)	22 (0.70%)	104 (3.32%)	1,280 (40.88%)
Ancylostoma caninum	WormBase ParaSite reference	2,654 (84.77%)	239 (7.63%)	143 (4.57%)	95 (3.03%)
	MinION data-only assembly	2,865 (91.50%)	100 (3.19%)	66 (2.11%)	100 (3.19%)
	MinION assembly + Illumina polishing	2,867 (91.57%)	99 (3.16%)	65 (0.08%)	100 (3.19%)
	Hybrid assembly	2,883 (92.08%)	111 (3.55%)	57 (1.82%)	80 (2.56%)

To address this question, we utilized BUSCO to objectively generate comparative gene datasets. BUSCO outputs the DNA sequence for each single copy ortholog it identifies in an assembly. We compiled these sequences for MinION data assemblies, MinION data assemblies polished with Illumina data (i.e., “polished” assemblies) and hybrid assemblies, and for each species, we identified which sequences were shared between MinION data and polished, and MinION data and hybrid, assemblies. We then generated pairwise alignments for each shared gene and used comparative identity metrics to assess nucleotide-level differences. The goal of generating these comparisons

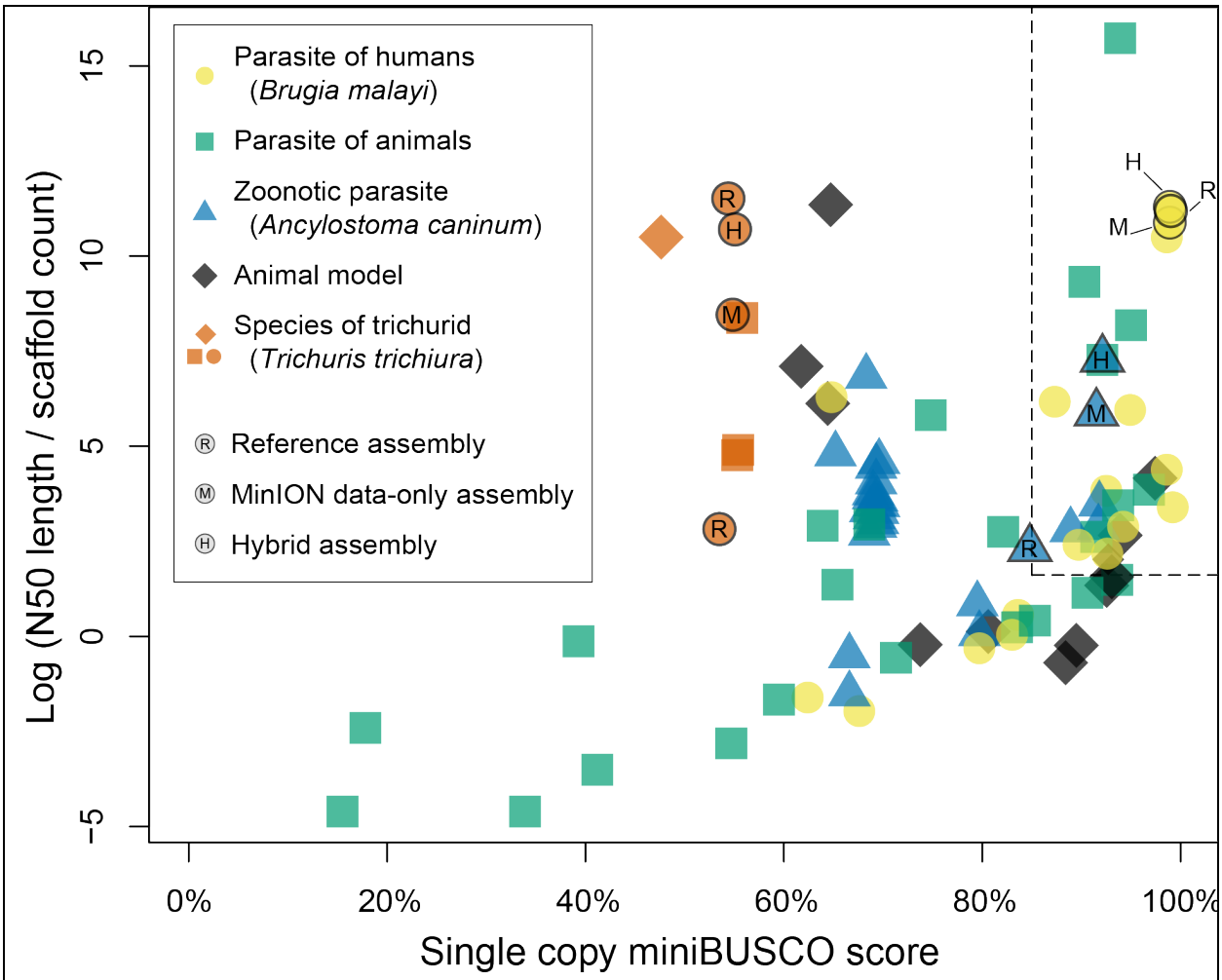

Figure 3. Plot of single copy miniBUSCO score versus the log value of N50 length divided by scaffold (or contig) count for the assemblies generated as part of this study, and for all genomes of nematodes that parasitize animals as adults available from WormBase ParaSite. Dotted lines indicate the cutoff for "tier 1" genome status *sensu* the International Helminth Genomes Consortium (i.e., >85% single copy BUSCO score and >1.6 log value for the contiguity metric). A key to symbolic designations is provided at top left.

was to assess whether the accuracy of assembled protein coding sequences is systematically affected by data type.

A summary of pairwise alignment designations and percent identities for genes shared between assembly types for each species is provided in Table 3. For all species, a greater proportion of identical genes were shared between MinION data and polished assemblies (2,893 for *B. malayi*, 1,485 for *T. trichiura*, and 1,801 for *A. caninum*) as

Table 3. Comparison of pairwise alignments for genes shared between assembly types generated as part of this study. Columns for each species represent values for the MinION data assembly compared to either the hybrid assembly (left) or the polished assembly (right). The total number of complete single copy BUSCOs shared between assemblies is out of the total number of nematode orthologs assessed by BUSCO (n=3,131).

	Brugia malayi		***Trichuris trichiura***		***Ancylostoma caninum***	
	Versus hybrid assembly	**Versus polished assembly**	**Versus hybrid assembly**	**Versus polished assembly**	**Versus hybrid assembly**	**Versus polished assembly**
Total no. complete single copy genes shared between assemblies	3,047	3,071	1,750	1,799	2,656	2,812
No. (percent) of identical alignments	1,959 (64.29%)	2,893 (94.20%)	622 (35.54%)	1,485 (82.55%)	1,207 (45.44%)	1,801 (64.05%)
No. (percent) of alignments that differ by SNPs only	411 (13.49%)	10 (0.36%)	420 (24.00%)	90 (5.00%)	316 (11.90%)	222 (7.89%)
No. (percent) of alignments that differ by SNPs and Indels	477 (15.65%)	93 (3.03%)	664 (37.94%)	177 (9.83%)	947 (35.66%)	600 (21.34%)
No. (percent) of alignments that differ by Indels only	200 (6.56%)	75 (2.44%)	44 (2.51%)	47 (2.61%)	186 (7.00%)	189 (6.72%)
Average percent identity | Average percent identity where aligned	**98.08% | 99.74%**	**99.59% | 99.95%**	**95.80% | 98.74%**	**99.02% | 99.75%**	**95.19% | 98.82%**	**97.45% | 99.52%**

257 compared to MinION data and hybrid assemblies (1,959 for *B. malayi*, 622 for *T. trichiura*,
258 and 1,207 for *A. caninum*) (see Table 3). Additionally, across species, average percent
259 identity, and average percent identity where aligned, were each greater for pairwise
260 alignments from MinION data versus polished assemblies as compared to hybrid
261 assemblies (see Table 3).

262 Dot plots of the lengths of genes shared between assembly types that differed by
263 length variants (i.e., Indels only, or SNPs and Indels), are provided for each species in
264 Figure 4A–F. (These alignments were of particular interest for determining whether one
265 assembly method systematically over- or under-represents gene length.) For all species,
266 a greater proportion of shared genes were found to differ by length variants in MinION
267 data versus hybrid assemblies as compared to polished assemblies. For each species
268 and comparison, no substantial difference was identified between the number of genes
269 identified as longer in MinION data versus polished or hybrid assemblies (Figure 4).

270 For all species, assembly comparisons, and pairwise alignment designations,
271 violin plots of percent identity and percent identity where aligned for alignments that
272 scored >90% in identity are provided in Figure 4G and 4F, respectively. For MinION data
273 versus hybrid assemblies, 88.65% and 98.36% of alignments had >90% identity and
274 >90% identity where aligned, respectively. For MinION data versus polished assemblies,
275 95.54% and 99.37% of alignments had >90% identity and >90% identity where aligned,
276 respectively. Across all species and comparisons, most pairwise alignments generated
277 for were found to be identical or nearly identical, with comparatively few alignments
278 scoring <97% in identity (see Figure 4G, F). Raw data characterizing each pairwise

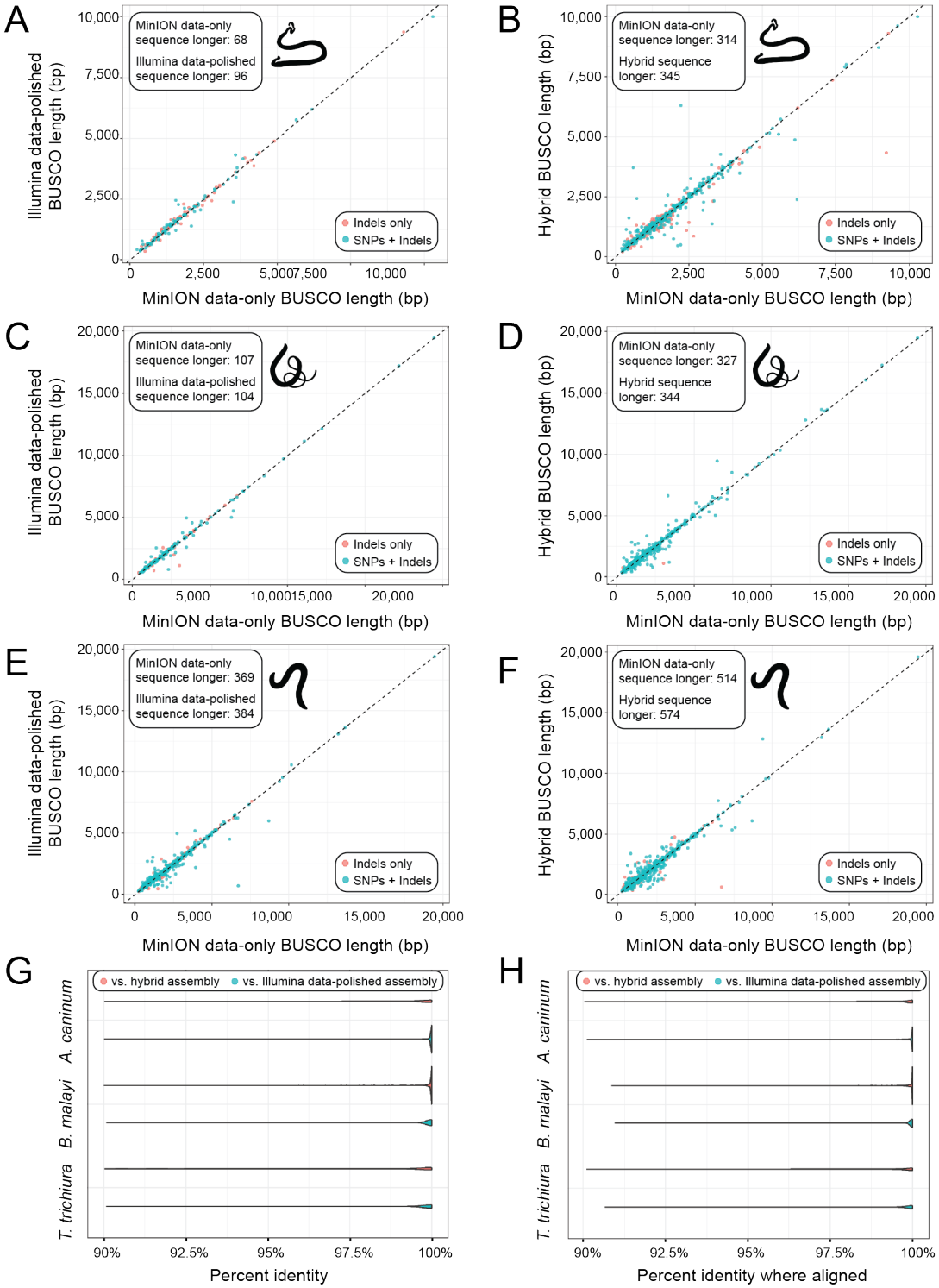

Figure 4. Comparisons of genes that differ by length variants shared between assembly types. Dot plots of lengths of shared genes for (A–B) *Brugia malayi*, (C–D) *Trichuris trichiura* and (E–F) *Ancylostoma caninum*. Dotted lines indicate equal lengths. Violin plots of (G) percent identity and (H) percent identity where aligned for all pairwise alignments generated for shared genes for each of the three focal species.

alignment generated for each assembly comparison are presented in Supplementary Table 5.

Discussion

Sample processing and sequencing

Genomic DNA extracted from a single individual is the ideal input for *de novo* whole genome assembly. This is in part because, for diploid organisms, assembly pipelines need only contend with two potential haplotypes, leading to more accurate assemblies with less haplotypic duplication. An alternative method would be to sequence multiple individuals from a highly inbred homozygous line. Many species cannot be maintained in a laboratory setting, however, making the establishment of inbred lines for these taxa challenging or impossible (Solares et al., 2021). It is also now well-recognized that inbreeding can significantly reduce genome size (Price, 1976; Fierst et al., 2015; Roessler et al., 2019), making these lines poor representatives of the species as whole. Furthermore, sampling from natural, rather than laboratory-maintained, populations is a requirement for most genomic studies (including genomic epidemiology), further highlighting the advantage of whole-genome sequencing from single individuals.

Considering these advantages, the DNA extraction and MinION library preparation protocols described here were optimized for low input gDNA extracted from individual parasitic nematodes. For *B. malayi* and *T. trichiura*, which both have relatively large adult stages (i.e., ~3–5 cm in length) and tractable genome sizes (i.e., <100 Mb), these protocols allowed for the generation of both short- and long-read data from a single adult worm. Adults of *A. caninum*, however, tend not to exceed ~1.5 cm in length, and yet have

302 the largest genome size of the three focal species in this study (i.e., ~400 Mb). Thus, for
303 *A. caninum*, even if we had had access to adult worms, we expect that using these
304 optimized protocols would still not have allowed for generation of genomic data from a
305 single adult. Moving forward, the approaches outlined here can be adapted for use with
306 additional species of parasitic nematodes.

307 There was an association between the average Q score of MinION read datasets
308 and the total amount of data generated by a flow cell: More efficient MinION sequencing
309 runs generated data with higher average Q scores (see Supplementary Table 1). For
310 example, MinION libraries for both *B. malayi* and *T. trichiura* were sequenced for ~63–72
311 hr each and generated ~10.8–11.8 Gb of data per flow cell post-basecalling with
312 comparably high Q scores. Conversely, libraries for *A. caninum*, which were sequenced
313 for ~68–80 hr each, only generated ~7.7–8.5 Gb of data per flow cell post-basecalling,
314 and these data had the lowest average Q score of the three species sequenced (see
315 Supplementary Table 1; Supplementary Figure 5). Differences in pore occupancy during
316 sequencing could potentially explain this correlation. According to MinKNOW run reports,
317 the flow cell used to sequence the *B. malayi* library had the greatest proportion of
318 available and actively sequencing pores throughout the sequencing run while the flow
319 cells that were used to sequence the three *A. caninum* libraries experiences the most
320 rapid decline in pore availability. Whatever the reason for this association, it is clear that
321 sequencing efficiency should be taken into consideration when attempting to achieve
322 Q20+ data using the latest ONT chemistries.

323

324 <u>Estimating contamination in read-level data</u>

325 In terms of detecting potential contamination in read-level data, Mash worked well
326 for *B. malayi*, but did not return species-appropriate hits for *T. trichiura* or *A. caninum* (see
327 Supplementary Table 2). Given that high proportions (~90.3–99.9%) of both MinION and
328 Illumina reads mapped to the reference genomes for all species, however, it is unlikely
329 that these results indicate significant read-level contamination for the later two species.
330 Of the three species, however, *B. malayi* is the least well-represented in the NCBI RefSeq
331 archive (the databased utilized by Mash) in terms of total number of records, so why Mash
332 performed most effectively for *B. malayi* is unclear. It is worth noting that read N50 lengths
333 for *B. malayi* greater than those for *T. trichiura* or *A. caninum* (Supplementary Table 1),
334 and so perhaps longer sequences allowed the NCBI RefSeq archive to be queried more
335 accurately by Mash.
336
337 <u>Estimation of genome size, heterozygosity, and ploidy</u>
338 For *B. malayi*, the MinION data assembly was slightly larger than the current
339 reference assembly, while the hybrid assembly was appreciably smaller (see Table 1).
340 For each of *T. trichiura* and *A. caninum*, both assemblies generated as part of this study
341 were considerably smaller than existing reference genomes (see Table 1). For *A.*
342 *caninum*, this different was most dramatic, on the order of >100 Mb. Assemblies
343 generated here are likely smaller than current references due to that fact that a purging
344 step to remove haplotypic duplication was included in the bioinformatic pipeline for both
345 assembly types (see Figure 1). This was not the case in the generation of existing
346 reference genomes (Ghedin et al., 2007; Foth et al., 2014; Consortium, 2019), and so
347 they likely include haplotypic duplication that artificially inflates genome size—particularly

348 for *T. trichiura* and *A. caninum*, which were sequenced from heterozygous sample types.
349 Three lines of evidence support that the genomes assembled here more accurately
350 represent the true genome size for each species: (1) The lengths of these assemblies
351 more closely match k-mer based genome size estimates than do those of the existing
352 reference genomes (see Table 1, Supplementary Figure 4); (2) miniBUSCO scores are
353 largely indistinguishable across all assembly types (including existing reference
354 assemblies), indicating that all are comparably complete despite ranging in size; in fact,
355 the Doyle et al. (2022) reference assembly for *T. trichiura* and the WormBase ParaSite
356 reference assembly for *A. caninum* had higher proportions of duplicated BUSCOs as
357 compared to the assemblies generated herein (see Table 2); and (3) for hybrid
358 assemblies, KAT plots suggest they are not "over-purged" and thus missing genomic
359 content (see Supplementary Figure 9). As advocated previously, purging haplotypic
360 duplication is an important step in genome curation regardless of assembly approach
361 (Howe et al., 2021; Rhie et al., 2021), and, unsurprisingly, becomes increasingly
362 important for highly heterozygous sample types such as that sequenced for *A. caninum*.
363 Results from GenomeScope (Supplementary Figure 4) and the k-mer multiplicity
364 plots output by KAT (Supplementary Figure 9) indicate low heterozygosity for *B. malayi*
365 and elevated heterozygosity for *T. trichiura* and *A. caninum*. These results are congruent
366 with the fact that the specimen of *B. malayi* sequenced came from a long-maintained
367 inbred laboratory strain while *T. trichiura* and *A. caninum* were sequenced from
368 specimens from infections maintained for shorter periods of time.
369 Preliminary data from nQuire suggest that *T. trichiura* may be triploid (see
370 Supplementary Table 3). This result was unexpected, given that the well-studied model

371 species for trichuriasis, *Trichuris muris*, is diploid (Špakulová et al., 1994). An early study
372 by Goswami (1978) on the karyotypes of trichurids reports that *T. trichiura* is also diploid.
373 Goswami does report supernumerary chromosomes (i.e., small abnormal chromosomes
374 with few or no banding landmarks; Spinner et al., 2013) in *Trichuris ovis* and *Trichuris*
375 *globulosa*, but makes no mention of supernumerary chromosomes in *T. trichiura*. Despite
376 the publication of Goswami's work, more recent literature perplexingly states that the
377 karyotype of *T. trichiura* remains unknown (Foth et al., 2014; Yousuf et al., 2022). The
378 estimation of ploidy for *T. trichiura* made here was based purely on a bioinformatic
379 approach and should therefore not be considered a definitive assessment. Moving
380 forward, it seems a worthwhile pursuit to confirm the chromosome number and copy
381 number for this medically important species using a modern karyotyping approach.
382
383 <u>Assembly refinement</u>
384 As with Mash, BlobTools was a more effective tool for assessing potential
385 contamination in final assemblies for *B. malayi* than for *T. trichiura* and *A. caninum*. For
386 example, for *B. malayi*, BlobTools identified four small contigs in the MinION data
387 assembly that represented sequence from the jird host, and contigs in both assembly
388 types that represented sequence from the *Wolbachia* endosymbiont, with all remaining
389 contigs identified as belonging to the phylum Nematoda (see Supplementary Tables 3
390 and 4). For *T. trichiura* and *A. caninum*, however, while many contigs could be identified
391 as belonging to the phylum Nematoda, a large proportion of contigs in both assembly
392 types returned off-target results from other phyla, or could not be assigned to a taxon at
393 all (see Supplementary Table 4). All three species are represented in the NIH BLAST

394 database on which BlobTools is based, but the assemblies generated for *B. malayi* were
395 more contiguous than those generated for *T. trichiura* and *A. caninum*. As noted above
396 for Mash, perhaps longer contig lengths allowed for the return of more on-target BLAST
397 results for *B. malayi*.
398
399 Assembly, contiguity, accuracy, and completeness
400 Hybrid assemblies were in all cases more contiguous than MinION data
401 assemblies (see Table 1, Figure 2). (As polishing with Illumina data improves accuracy
402 rather than contiguity, MinION data assemblies polished with Illumina data were no
403 included in contiguity comparisons.) This result was somewhat unexpected given that
404 including short-read data in whole genome assembly has traditionally been touted to
405 increase accuracy rather than contiguity (Bashir et al., 2012; English et al., 2012; Koren
406 et al., 2012). While this was indeed the case when utilizing hybrid approaches with older
407 ONT chemistries, the results of the comparisons made here demonstrate that the
408 improved read-level quality of Q20+ MinION data may lessen the value of short-read data
409 for improving assembly accuracy. For example, alignment identity metrics indicate a high
410 degree of identity (i.e., ~96–99%) between MinION data and hybrid assemblies for each
411 species (see MuMmer4 results in Table 1).
412 Though often reported (e.g., Fu et al., 2020; Wang et al., 2020; Yi et al., 2020;
413 Ludington & Sanders, 2021; Voelker et al., 2021; Okenmor et al., 2022), genome-wide
414 alignment identity metrics like those produced by MuMmer4 are a broad and imperfect
415 measure of assembly accuracy. Therefore, we endeavored to conduct a more in-depth
416 analysis of accuracy by characterizing pairwise alignments of BUSCO sequences shared

417 between assembly types. The long-read and hybrid assemblers utilized here are based
418 on different assembly algorithms that handle read-level data dissimilarly. Though both are
419 built on a Celera Assembler foundation, Canu's MinHash Alignment Process (MHAP) is
420 broadly an overlap layout consensus-based method (Koren et al., 2017) while MaSuRCA
421 relies on a "super-read" algorithm that combines aspects of both overlap layout
422 consensus and de Bruijn graph-based approaches (Zimin et al., 2013). For assessing the
423 ability of short-read data to improve long-read assembly accuracy, comparing MinION
424 data assemblies to MinION data assemblies polished with Illumina data is a more salient
425 comparison than comparing to hybrid assemblies. If polishing with Illumina data
426 substantially improved the accuracy of MinION data assemblies, we would expect the
427 majority of genes found in both assembly types to differ at the sequence level. This was
428 not the case, as for all three species, most genes shared between these two assembly
429 types were identical or nearly identical at the nucleotide level (see Table 3, Figure 4,
430 Supplementary Figures 13–15). The biggest differences were attributable to pairwise
431 alignments containing dramatic length variants, as evidenced by high average percent
432 identities where aligned (i.e., >99.5% across species; Table 3) and comparably lower
433 average percent identities (i.e., >97.4% across species; Table 3), which take into account
434 difference caused by Indels. Given that most alignments were identical or nearly identical,
435 the dramatic length variants observed between MinION data only and Illumina-polished
436 assemblies are more than likely the result of inconsistencies in the way BUSCO identifies
437 genes.

438 Assembly accuracy is a complex metric that is challenging to assess for large
439 eukaryotic genomes, and none of the methods utilized here to measure accuracy are

440 alone sufficient assessments. In combination, however, the results of these methods
441 suggest that when working with Q20+ MinION data for whole genome *de novo* assembly,
442 the added value of short-read data comes more measurably from improvements to
443 contiguity rather than accuracy. Increased contiguity of hybrid assemblies over long read-
444 only assemblies has been observed in bacteria whole genome sequencing (George et
445 al., 2017; Neal-McKinney et al., 2021), but to our knowledge has not been previously
446 reported for eukaryotes. It is worth noting that the long reads generated for each species
447 here did not approach the read lengths that MinION devices are capable of sequencing.
448 Across species, read length N50 values ranged from ~2.7–8.6 kb (Supplementary Table
449 1), which are relatively small values. This is likely in part due to the fact that generating
450 data from low-input samples, as was done here for single or larval worms, precludes
451 rigorous size selection. It is possible that the generation of ultra-long reads would have
452 lessened the gap in contiguity metrics between MinION data and hybrid assemblies.It
453 should also be noted that generating short-read data in addition to long-read data also
454 has the added benefit of allowing for the use of k-mer based approaches to estimate
455 genome size and heterozygosity and to assess the effectiveness of purging haplotypic
456 duplication.

457 Results from miniBUSCO indicate a high proportion of missingness (i.e., >40%)
458 for all trichurid assemblies (see Table 2; Figure 3). Why trichurids appear to be missing
459 nearly half of nematode BUSCOs is not well understood. Completeness of the existing
460 WormBase ParaSite reference assemblies for *T. trichiura* and *T. suis* were previously
461 assessed using the Core Eukaryotic Genes Mapping Approach (i.e., CEGMA) (Parra et
462 al., 2007), which indicated much higher rates of completeness for these assemblies (i.e.,

463 ~94–96%) (Foth et al., 2014; Jex et al., 2014). However, the creators of CEGMA have
464 since acknowledged BUSCO as a superior tool for assessing assembly completeness
465 (Bradnam, 2015). Its origins aside, this presumed genome reduction clearly presents a
466 challenge for assessing trichurid genome assemblies when using traditional
467 completeness metrics.
468
469 **Conclusions**
470 This study represents the first comprehensive comparison of Q20+ MinION data
471 versus hybrid assembly approaches for generating complete genomes from parasitic
472 nematodes, and from eukaryotes more broadly. Traditionally, hybrid approaches—
473 including both incorporating short- and long-read in assembly and polishing long-read
474 assemblies with short-read data—have been shown to improve assembly-level accuracy.
475 With the advent of improved read-level quality of Q20+ MinION data, however, it appears
476 that, at least in this study, where the generation of ultra-long reads was not possible, the
477 more evident benefit of a hybrid approach was improved contiguity rather than improved
478 assembly accuracy. Our pairwise alignment analysis of shared genes suggests that most
479 differences observed between assembly types result from differences in assembly
480 algorithms or BUSCO implementation rather than significant errors in read-level MinION
481 data that systematically affect final genome assemblies.
482 This study was of course met with limitations. For two of the three focal species,
483 gDNA was generated from adult worms. For most species, and particularly those that
484 infect humans, adult nematodes are difficult or impossible to obtain, somewhat limiting
485 the utility of this approach in practice. And although we highlight the potential utility of

486 Q20+ MinION data for population-level analyses, it was outside the scope of this study to
487 benchmark MinION data for population genomics in eukaryotes. Finally, the assemblies
488 generated here remain admittedly incomplete in terms of the modern reference genome
489 standards. They do not approach chromosome-scale, and they are not annotated, and
490 additional data types are required to complete them. Generating complete reference
491 assemblies was not the goal of this study, however; rather, we sought to evaluate the
492 feasibility and quality of whole genome assemblies generated from an accessible and
493 inexpensive sequence data type. For *B. malayi* and *T. trichiura*, existing reference
494 genomes were already exceptionally high-quality, "tier 1"-status assemblies *sensu* the
495 International Helminth Genomes Consortium Consortium (2019), and the MinION data
496 and hybrid assemblies generated in general matched them in terms of comparative
497 quality metrics. The comparison in which the added value of Q20+ MinION data is most
498 evident is that of *A. caninum*. Using gDNA from the same strain and sample type on which
499 the existing reference assembly was based, we were able to considerably improve the
500 genomic resources available for this species relatively easily and inexpensively. These
501 results highlight the generation of Q20+ MinION data as a simple and cost-effective way
502 to both shift existing reference genomes for parasitic nematodes into "tier 1" status, as
503 well to generate *de novo* whole genome assemblies.

504 In sum, we demonstrated that it is reasonably straightforward to generate high-
505 quality "tier 1" genomes for a variety of species of parasitic nematodes using Q20+
506 MinION data alone. Additionally, for species of sufficient size and tractable genome
507 length, the optimized DNA extraction and library preparation protocols outlined here allow
508 for the generation of complete genomes from individual specimens. More broadly, the

509 quality and accessibility of Q20+ MinION data have the potential to make generating real-
510 time genomic data a routine part of control program monitoring and surveillance,
511 ultimately accelerating progress toward disease elimination.
512

513 **Materials and Methods**

514 Specimen acquisition

515 The specimens sequenced were obtained from the same strains previously used
516 to generate existing reference genomes for each species. Adults of *B. malayi* (FR3 strain)
517 were acquired through the NIH/NIAID Filariasis Research Reagent Resource Center via
518 BEI Resources (Michalski et al., 2011). Worms were received frozen and stored at -80°C.
519 Prior to sequencing, worms were thawed at 4°C and subsequently preserved in DNA/RNA
520 Shield (Zymo Research, Irvine, CA, USA). Adults of *T. trichiura* were obtained from a
521 Danish patient after anthelmintic treatment for an infection acquired in Uganda. Worms
522 were received preserved in 70% ethanol and were stored at 4°C prior to sequencing.
523 Pooled third stage (L3) larvae of *A. caninum* (Baltimore strain) were obtained from
524 experimental infection in laboratory-maintained canines. Pooled L3s were received
525 frozen, stored at -80°C, and thawed at 4°C prior to sequencing.
526

527 DNA extraction

528 Total nucleic acid was extracted separately from each of a single adult female (*B.*
529 *malayi*), a single adult male (*T. trichuris*), and a pool of L3 larvae (*A. caninum*) using a
530 *Quick*-DNA™ HMW MagBead Kit (Zymo Research, Irvine, CA, USA) and modified Solid
531 Tissue extraction protocol. Prior to extraction, adult worms were placed in clean 1.5 mL

532 DNA LoBind® tubes (Eppendorf® North America, Enfield, CT, USA) using sterilized
533 forceps, while pooled L3s were centrifuged at 6,540 *g* for 3 min to allow for the removal
534 of excess supernatant. All samples were then homogenized mechanically using a
535 Monarch® Pestle Set single-use microtube pestle (New England Biolabs® Inc., Ipswich,
536 MA, USA). Homogenization was performed both prior to, and immediately following, the
537 addition of Proteinase K. Samples were then digested in a digital dry bath at 55°C for 24
538 hr with occasional flick mixing. The protocol for DNA purification was then followed, with
539 these modifications: After the addition of 33 µL of MagBinding Beads, samples were
540 placed on a benchtop rotating mixer for 40 min to 1.5 hr at RT; and after the addition of
541 52 µL DNA Elution Buffer, samples were incubated in a digital dry bath at 37°C for 2 hr
542 with occasional flick mixing, then incubated at RT overnight prior to collecting eluted DNA.
543 To avoid DNA fragmentation, samples were mixed by flicking rather than pipette mixing
544 wherever possible. The concentration of each extraction was measured using a Qubit4™
545 1X dsDNA High Sensitivity (HS) Assay (ThermoFisher Scientific, Waltham, MA, USA) and
546 fragment size distribution was determined using an Agilent 2200 TapeStation System and
547 associated protocol for genomic DNA ScreenTape analysis (Agilent, Santa Clara, CA,
548 USA). For *T. trichiura*, TapeStation analysis was not performed.

549 For *A. caninum*, initial attempts to sequence ONT libraries resulted in low
550 sequencing efficiency, poor data quality, and rapid pore loss. To ameliorate concerns of
551 protein contamination, extracted gDNA was purified via bead-based cleanup prior to final
552 ONT library preparation. This cleanup was performed as follows: Two aliquots of
553 extracted gDNA were each bound to a 0.5× volume of AMPure XP beads (Beckman
554 Coulter, Brea, CA, USA) on a benchtop rotator mixer for 1 hr at RT. Beads were then

555 washed twice with 80% EtOH, air-dried, and gDNA was eluted in 52 µL Zymo DNA Elution
556 Buffer for 2 hr at 37°C with occasional flick mixing, followed by overnight elution at RT.
557 This cleanup process was repeated a second time for both aliquots. Purified extractions
558 were then quantified using a Qubit4™ 1X dsDNA High HS Assay and the Agilent 2200
559 TapeStation System.
560
561 MinION library preparation
562 Aliquots of gDNA extracted for each species were used to prepare one or more
563 libraries for whole genome sequencing on an ONT MinION desktop sequencer (see
564 Supplementary Table 1). Libraries were prepared using a combination of the DNA Clean
565 & Concentrator MagBead (Zymo Research, Irvine, CA, USA) and ONT SQK-LSK114
566 Ligation Sequencing (Oxford Nanopore Technologies, Oxford, United Kingdom) Kits, and
567 a modified hybrid protocol. First, gDNA was incubated with ONT DNA repair and end
568 preparation reagents for 10 min at RT followed by 10 min at 65°C. End preparation
569 reactions were then bound to 20 µL Zymo MagBinding Beads in 4× volumes of Zymo
570 DNA MagBinding Buffer on a benchtop rotating mixer for 30 min to 2 hr. Beads were then
571 washed twice with Zymo DNA Wash Buffer, air-dried for 10 min, eluted in 51 µL Zymo
572 DNA Elution Buffer via manual agitation for 10 min, and quantified via Qubit4™ 1X dsDNA
573 HS Assay. Sequencing adaptors were then ligated via a 15 min incubation at RT, and
574 libraries were bound to a 0.4× volumes of AMPure XP beads on a benchtop rotator mixer
575 at RT for 1 hr. Beads were then washed twice with either ONT Long Fragment Buffer (*A.*
576 *caninum*) or a titrated wash mix of 1:3 ONT Short Fragment Buffer:ONT Long Fragment
577 Buffer (*B. malayi* and *T. trichiura*). After washing, beads were air-dried, then libraries were

578 eluted in 15–17 µL of ONT Elution Buffer for 2 hr at 37°C with occasional flick mixing
579 followed by overnight elution at RT. Final libraries were quantified using a Qubit4™ 1X
580 dsDNA High HS Assay and the Agilent 2200 TapeStation System. For *T. trichiura*,
581 TapeStation analysis was not performed. Additionally for *T. trichiura*, a portion of ONT
582 library reserved after sequencing on one flow cell was re-washed prior to sequencing on
583 a second flow cell (see Supplementary Table 1). This aliquot of library was bound to a
584 0.4× volume of AMPure XP beads on a benchtop rotator mixer for 1 hr, washed twice with
585 ONT Long Fragment Buffer, air-dried, then eluted in 17 µL of ONT Elution Buffer for 2 hr
586 at 37°C with occasional flick mixing followed by overnight elution at RT.
587
588 <u>MinION sequencing, basecalling, and read data quality control</u>
589 Portions of ONT libraries were sequenced for ~62–80 hr each on one (*B. malayi*),
590 two (*T. trichiura*), or three (*A. caninum*) ONT MinION R10.4.1 flow cells. The amount of
591 library loaded onto each flow cell and the total number of pores available at the start of
592 sequencing are provided in Supplementary Table 1. MinKNOW software (ONT) v. 22.10.7
593 (*B. malayi*), v. 22.10.10 (*T. trichiura*), or v. 22.12.7 (*A. caninum*) was used to run each
594 flow cell with pore scans every 1.5 hr. After sequencing, signal data (i.e., fast5 files) from
595 each flow cell were basecalled using Guppy v. 6.3.4 (ONT). Sequencing adaptors were
596 simultaneously removed by specifying the "--trim_adapters" flag. The results of each
597 Guppy run were summarized using NanoPlot v. 1.40.2 (De Coster & Rademakers, 2023).
598 Fastq files that passed basecalling were input to Mash v. 2.2.2 (Ondov et al., 2016) to
599 confirm there was no significant contamination in the read-level data prior to assembly.
600 Estimated genome sizes provided to Mash were based on the lengths of the existing

601 reference assemblies available for each species, and were thus set as 88 Mb for *B. malayi*
602 (Ghedin et al., 2007), 465 Mb for *A. caninum* (Consortium, 2019) and 75 Mb for *T. trichiura*
603 (https://github.com/stephenrdoyle/ancient_trichuris/tree/master/02_data) (Doyle et al.,
604 2022). Average depth of coverage of reads across the species-specific reference genome
605 and proportion of reads mapped were estimated using Minimap2 v. 2.16 (Li, 2018) and
606 SAMtools v. 1.9 (Li et al., 2009).
607
608 Illumina library preparation, sequencing, and read data quality control
609 Aliquots of gDNA extracted for each species were sent to the Yale Center for
610 Genome Analysis (YCGA) for Illumina whole genome library preparation and sequencing
611 (see Supplementary Table 1). Libraries were prepared using an xGenTM cfDNA & FFPE
612 DNA Library Prep Kit with unique dual indexing (Integrated DNA Technologies, Newark,
613 NJ, USA) and standard protocol. Libraries were subjected to 5–7 cycles of PCR to
614 increase concentration (see Supplementary Table 1) and multiplexed for 2×150 paired-
615 end sequencing on an Illumina NovaSeq 6000 targeting 50× depth of coverage genome-
616 wide for each species. Reads were demultiplexed by the YCGA and subsequently filtered
617 for remaining adaptor contamination, quality, and length using fastp (Chen et al., 2018).
618 Fastq files that passed fastp were input to Mash to confirm there was no significant
619 contamination in the read-level data prior to assembly. Estimated genome sizes provided
620 to Mash were based on the lengths of the existing reference assemblies available for
621 each species (see above). Average depth of coverage across the species-specific
622 reference genome and proportion of reads mapped were estimated using BWA v. 0.7.17
623 and SAMtools v. 1.9 (Li et al., 2009). A second round of fastp was run on the raw

624 demultiplexed Illumina data to remove sequencing adapters, only, by specifying
625 "--detect_adapter_for_pe", "--disable_length_filtering", and "--disable_quality_filtering".
626 These data were the used for hybrid genome assembly, as MaSuRCA utilizes built-in
627 error correction and cleaning.
628
629 Estimation of genome size, heterozygosity, and ploidy
630 Quality-controlled Ilumina data were used to estimate genome size and
631 heterozygosity for each species. First, Jellyfish v. 2.3.0. (Marçais & Kingsford, 2011) was
632 used to generate k-mer spectra, specifying a k-mer size of 21 and an initial hash size of
633 1,000,000,000. Histograms from Jellyfish were then provided to the GenomeScope web-
634 based graphical user interface (Vurture et al., 2017) to visualize k-mer spectra and
635 estimate heterozygosity. For *T. trichiura*, ploidy was estimated bioinformatically using
636 quality-controlled Ilumina data and basecalled MinION data in conjunction with nQuire
637 (Weiß et al., 2018).
638
639 MinION data genome assembly
640 Fastq files that passed basecalling via Guppy were concatenated and used as
641 input to Canu v. 2.1 (ONT) (Koren et al., 2017) to generate MinION data assemblies. For
642 *T. trichiura* and *A. caninum*, a 3 kb minimum read length requirement was specified, and
643 for *B. malayi*, a 5 kb minimum read length requirement was specified. Estimated genome
644 sizes provided to Canu were based on the lengths of the existing reference assemblies
645 available for each species (see above). Contigs in the resulting Canu assemblies that
646 were indicated as potential alternative alleles (i.e., with FASTA headers including

647 “suggestBubble=yes”) were removed using a combination of SeqKit v 0.10.1 (Shen et al.,
648 2016) and the filterbyname.sh script of BBMap v.38.84 (Bushnell, 2014). Resulting
649 “popped” assemblies were used as input to purge_dups v. 1.2.5 (Guan et al., 2020) to
650 remove false duplications. These “popped and purged” assemblies were then polished
651 with one round of Racon v. 1.5.0 (Vaser et al., 2017) followed by one round of Medaka v.
652 1.7.2 (https://github.com/nanoporetech/medaka) to generate “popped, purged, and
653 polished” assemblies.
654
655 Hybrid genome assembly
656 FASTQ files that passes basecalling by Guppy and Illumina data from which only
657 sequencing adaptors were removed by fastp were used as input to MaSuRCA v. 4.1.0
658 (Zimin et al., 2013) to generate hybrid assemblies. Default settings were used except for
659 setting “LHE_COVERAGE” to “35” and “cgwErrorRate=0.15”, and setting “JF_SIZE” to a
660 value ten times the genome size of the species-specific reference assembly. Assemblies
661 output by MaSuRCA were input to purge_dups v. 1.2.5 (Guan et al., 2020) to remove
662 false duplications.
663
664 Quality control and refinement of final assemblies
665 To ensure “purged” MaSuRCA hybrid assemblies were not over- or under-purged,
666 copy number spectrum plots were generated using the K-mer Analysis Toolkit (KAT)
667 v. 2.4.0 (Mapleson et al., 2017) specifying a k-mer size of 31. “Popped, purged, and
668 polished” Canu MinION data assemblies and “purged” MaSuRCA hybrid assemblies were
669 input to BlobTools v. 1.1.1 (Laetsch & Blaxter, 2017) to identify potential contaminants.

670 First, the reads used to generate each assembly were mapped to the assembly using
671 Minimap2 (MinION data) and/or BWA (Illumina data), and SAMtools was used to generate
672 sorted BAM files for each mapping. The “blastn” function in BLAST v. 2.14.0 (Altschul et
673 al., 1990) was then used to compare all contigs to reference sequences available in the
674 National Center for Biotechnology Information (NCBI)’s nucleotide databases, specifying
675 “-evalue 1e-25”, “-max_target_seqs 10” and “-max_hsps 1”. BLAST outputs and sorted
676 BAM files were then used as input to BlobTools. Contigs that could be confirmed as host,
677 human, or bacterial contamination were excluded from final assemblies. Quality scores
678 (i.e., Q scores) of basecalled MinION reads, and Q scores of reads when mapped to the
679 final “popped, purged, and polished” MinION data assemblies, were measured using
680 NanoPlot and plotted using the “geom_density” function in the ggplot2 package
681 (Wickham, 2016) in R v. 4.3.0 (Team, 2023) in RStudio v. 2023.03.1 (Team, 2020).
682
683 Mitogenome and *Wolbachia* genome assembly
684 For *B. malayi* and *A. caninum* contigs containing the mitogenomes—and, for *B.*
685 *malayi*, the genome of its *Wolbachia* endosymbiont—were extracted from “popped”
686 MinION and unpurged MaSuRCA assemblies. For *T. trichiura*, mitogenomes were
687 extracted from the original “unpopped” Canu assembly and the unpurged MaSuRCA
688 assembly. These organelle genomes contigs were then mapped to species-specific
689 reference organelle genomes using Geneious v. 2022.0.1 ("Geneious Prime v.
690 2022.0.1,") and manually reconfigured to match the linear orientation of the reference.
691 For *B. malayi* and *T. trichiura*, reference organelle genomes were extracted from the
692 reference genome assemblies for each species (see above). For *A. caninum*, for which a

693 mitochondrial contig is not labeled in the reference assembly, a reference mitogenome
694 was downloaded from GenBank (accession no. MN215971) (Xie et al., 2019). Contigs
695 containing complete organelle genomes were then manually added to the appropriate
696 final version of each assembly.
697
698 BUSCO comparison to assess assembly accuracy
699 Copies of final "popped, purged and polished" MinION data assemblies generated for
700 each species (i.e., assemblies from which potential contaminants identified by BlobTools
701 had been removed, and to which organelle genomes had been added) were separately
702 polished with Illumina data using Pilon v. 1.24 (Walker et al., 2014) and BWA. Three
703 iterations of Pilon were iteratively run specifying the "--diploid" flag. BUSCO v. 4.0.6
704 (Seppey et al., 2019) was then run for MinION data, polished, and hybrid assemblies for
705 each species specifying comparison to the Nematoda ortholog database
706 (nematoda_odb10). Two single copy ortholog datasets were generated per species from
707 the resulting BUSCO output: one containing sequences found in both the MinION data
708 and the polished assembly, and one containing sequences found in both the MinION data
709 and the hybrid assembly. For each dataset and species, the Genes shared between the
710 two assembly types were pairwise aligned using the "pairwiseAlignment" function in
711 Biostrings package v. 2.66.0 (Pages et al., 2013) in R v. 4.2.3 via RStudio v. 2023.06.1
712 specifying the alignment type as "global". Biostrings and custom R code were then used
713 to retain only pairwise alignments in which Genes from both assemblies started and
714 ended with eukaryotic start and stop codons, respectively. Biostrings was then used to
715 characterize each pairwise alignment in terms of percent identity, alignment length,

716 BUSCO lengths, and number of single nucleotide polymorphisms (i.e., SNPS) and
717 insertions/deletions (i.e., Indels). Summary figures for all comparisons were generated in
718 R v. 4.3.0 via RStudio v. 2023.03.1 using the ggplot2 package.
719
720 Assessing assembly contiguity, accuracy, and completeness
721 Contiguity and accuracy were respectively assessed by comparing the MinION
722 data assembly to hybrid and reference assemblies using QUAST v. 5.0.2 (Gurevich et
723 al., 2013), and to the hybrid assembly using the "nucmer" and "dnadiff" modules of
724 MuMmer4 v. 4.0.0beta2 (Marçais et al., 2018). To assess assembly completeness,
725 miniBUSCO v. 0.2 (Huang & Li, 2023) was run on all three final assemblies (i.e., MinION
726 data, MinION data polished with Illumina data, and hybrid) generated for each species,
727 specifying comparison to the Nematoda ortholog database (nematoda_odb10). For
728 comparison, miniBUSCO was also run on the aforementioned reference genomes for
729 each species, as well as for all genomes of nematodes that parasitize animal as adults
730 available from WormBase ParaSite (Howe et al., 2016; Howe et al., 2017). For each
731 assembly, the proportion of single copy BUSCOs present was plotted against the log
732 value of the assembly's N50 length divided by its scaffold or contig count using R v. 4.2.3
733 via RStudio v. 2023.06.1. Cutoff values of >85% (for proportion of single copy BUSCOs
734 present) and >1.6 (for the log value of the contiguity metric) were used to assess whether
735 each assembly could be classified as a "tier 1" genome following the standard for helminth
736 genomes established by the International Helminth Genomes Consortium (Consortium,
737 2019).
738

739 **Acknowledgements**

740 The authors thank the BEI Resources NIH/NIAID Filariasis Research Reagent
741 Resource Center for providing specimens of *Brugia malayi* for sequencing. Thanks are
742 also due to Dr. Michael Wiley, Dr. Shaun Cross, Kristen Bernhard, and Mahmood Al-
743 Haehm (University of Nebraska Medical Center) for providing access and support for use
744 of select laboratory equipment. The authors additionally thank Christopher Castaldi and
745 Irina Tikhonova (Yale Center for Genome Analysis, Yale School of Medicine) for
746 generation of, and correspondence regarding, Illumina sequence data. Bioinformatic
747 analysis was enabled by the University of Nebraska-Lincoln Holland Computing Center
748 computing cluster.

749

750 **Author Contributions**

751 KSH and JRF designed the study. JMH provided the specimens of *Ancylostoma*
752 *caninum* sequenced. PN provided the specimen of *Trichuris trichiura* sequenced. KSH
753 extracted genomic DNA and prepared and sequenced MinION libraries. KSH and JRF
754 determined and refined bioinformatic pipelines. KSH performed bioinformatic analyses.
755 RW aided KSH in the generation of comparative BUSCO data. KSH and JRF interpreted
756 data and generated figures and tables. KSH and JRF drafted the manuscript. All authors
757 contributed to revising and approving the final version of the manuscript.

Literature Cited

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. 1990. Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.

Bashir, A., Klammer, A. A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., & Peluso, P. 2012. A hybrid approach for the automated finishing of bacterial genomes. Nature Biotechnology, 30(7), 701–707. DOI: 10.1038/nbt.2288.

Berger, D. J., Crellen, T., Lamberton, P. H., Allan, F., Tracey, A., Noonan, J. D., Kabatereine, N. B., Tukahebwa, E. M., Adriko, M., & Holroyd, N. 2021. Whole-genome sequencing of *Schistosoma mansoni* reveals extensive diversity with limited selection despite mass drug administration. Nature Communications, 12(1), 4776. DOI: 10.1038/s41467-021-24958-0

Biraro, I. A., Egesa, M., Toulza, F., Levin, J., Cose, S., Joloba, M., Smith, S., Dockrell, H. M., Katamba, A., & Elliott, A. M. 2014. Impact of co-infections and BCG immunisation on immune responses among household contacts of tuberculosis patients in a Ugandan cohort. PLOS One, 9(11), e111517. DOI: 10.1371/journal.pone.0111517.

Brabec, J., Kuchta, R., Scholz, T., & Littlewood, D. T. J. 2016. Paralogues of nuclear ribosomal genes conceal phylogenetic signals within the invasive Asian fish tapeworm lineage: Evidence from next generation sequencing data. International Journal for Parasitology, 46(9), 555–562. DOI: 10.1016/j.ijpara.2016.03.009.

Bradnam, K. (2015). Goodbye CEGMA, hello BUSCO!

Bushnell, B. 2014. *BBMap: a fast, accurate, splice-aware aligner*. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States). Retrieved from.

780 Casulli, A. 2021. New global targets for NTDs in the WHO roadmap 2021–2030. PLOS Neglected
781 Tropical Diseases, 15(5), e0009373. DOI: 10.1371/journal.pntd.0009373.
782 Chen, S., Zhou, Y., Chen, Y., & Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
783 Bioinformatics, 34(17), i884–i890. DOI: 10.1093/bioinformatics/bty560.
784 Consortium, I. H. G. 2019. Comparative genomics of the major parasitic worms. Nature
785 Genetics, 51(1), 163–174. DOI: 10.1038/s41588-018-0262-1.
786 De Coster, W., & Rademakers, R. 2023. NanoPack2: population-scale evaluation of long-read
787 sequencing data. Bioinformatics, 39(5), btad311. DOI: 10.1093/bioinformatics/btad311.
788 Doyle, S. Improving helminth genome resources in the post-genomic era. Trends in
789 Parasitology, 38(10), 831–840. DOI: 10.1016/j.pt.2022.06.002.
790 Doyle, S. R., Søe, M. J., Nejsum, P., Betson, M., Cooper, P. J., Peng, L., Zhu, X.-Q., Sanchez, A.,
791 Matamoros, G., & Sandoval, G. A. F. 2022. Population genomics of ancient and modern
792 *Trichuris trichiura*. Nature Communications, 13(1), 3888. DOI: 10.1038/s41467-022-
793 31487-x.
794 English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., &
795 Worley, K. C. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-
796 read sequencing technology. PLOS One, 7(11), e47768. DOI:
797 10.1371/journal.pone.0047768.
798 Fierst, J. L., Willis, J. H., Thomas, C. G., Wang, W., Reynolds, R. M., Ahearne, T. E., Cutter, A. D.,
799 & Phillips, P. C. 2015. Reproductive mode and the evolution of genome size and
800 structure in *Caenorhabditis* nematodes. PLoS Genetics, 11(6), e1005323. DOI:
801 10.1371/journal.pgen.1005323.

802 Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C.,
803 Crottini, A., Godoy, J. A., & Höglund, J. 2022. The era of reference genomes in
804 conservation genomics. Trends in Ecology & Evolution, 37(3), 197–202. DOI:
805 10.1016/j.tree.2021.11.008.
806 Foth, B. J., Tsai, I. J., Reid, A. J., Bancroft, A. J., Nichol, S., Tracey, A., Holroyd, N., Cotton, J. A.,
807 Stanley, E. J., & Zarowiecki, M. 2014. Whipworm genome and dual-species
808 transcriptome analyses provide molecular insights into an intimate host-parasite
809 interaction. Nature Genetics, 46(7), 693–700. DOI: 10.1038/ng.3010.
810 Fraija-Fernández, N., Waeschenbach, A., Briscoe, A. G., Hocking, S., Kuchta, R., Nyman, T., &
811 Littlewood, D. T. J. 2021. Evolutionary transitions in broad tapeworms (Cestoda:
812 Diphyllobothriidea) revealed by mitogenome and nuclear ribosomal operon
813 phylogenetics. Molecular Phylogenetics and Evolution, 163(107262), 11 pp. DOI:
814 10.1016/j.ympev.2021.107262.
815 Fu, Z., Li, Y., Elling, A. A., & Snyder, W. E. 2020. A draft genome of a field-collected *Steinernema*
816 *feltiae* strain NW. Journal of Nematology, 52(1), 1–7. DOI: 10.21307/jofnem-2020-003.
817 Geneious Prime v. 2022.0.1. https://www.geneious.com.
818 George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., Mathers, A.,
819 Norris, R., Navickaite, I., & Eaton, C. 2017. Resolving plasmid structures in
820 Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and
821 MinION/Illumina hybrid data assembly approaches. Microbial Genomics, 3(8), e000118.
822 DOI: 10.1099/mgen.0.000118.

823 Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L.,
824 Guiliano, D. B., & Miranda-Saavedra, D. 2007. Draft genome of the filarial nematode
825 parasite *Brugia malayi*. Science, 317(5845), 1756–1760. DOI: 10.1126/science.1145406.
826 The Global Burden of Disease Study 2019. 2020. The Lancet, 396(10258), 1129–1306.
827 Goswami, U. 1978. Some cytogenetical aspects of genus *Trichuris* (Nematoda). Current Science,
828 47(11), 368–371.
829 Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. 2020. Identifying and
830 removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36(9),
831 2896–2898. DOI: 10.1093/bioinformatics/btaa025.
832 Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. 2013. QUAST: quality assessment tool for
833 genome assemblies. Bioinformatics, 29(8), 1072–1075. DOI:
834 10.1093/bioinformatics/btt086.
835 Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L.,
836 Reed, L. K., Storfer, A., & Whitlock, M. C. 2016. Finding the genomic basis of local
837 adaptation: pitfalls, practical solutions, and future directions. The American Naturalist,
838 188(4), 379–397. DOI: 10.1086/688018.
839 Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., Torrance, J., Tracey, A., &
840 Wood, J. 2021. Significantly improving the quality of genome assemblies through
841 curation. Gigascience, 10(1), giaa153. DOI: 10.1093/gigascience/giaa153.
842 Howe, K. L., Bolt, B. J., Cain, S., Chan, J., Chen, W. J., Davis, P., Done, J., Down, T., Gao, S., &
843 Grove, C. 2016. WormBase 2016: expanding to enable helminth genomic research.
844 Nucleic Acids Research, 44(D1), D774–D780. DOI: 10.1093/nar/gkv1217.

845 Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P., & Berriman, M. 2017. WormBase ParaSite− a
846 comprehensive resource for helminth genomics. Molecular and Biochemical
847 Parasitology, 215, 2–10. DOI: 10.1016/j.molbiopara.2016.11.005.
848 Huang, N., & Li, H. 2023. miniBUSCO: a faster and more accurate reimplementation of BUSCO.
849 bioRxiv, 2023.2006. 2003.543588. DOI: 10.1101/2023.06.03.543588.
850 Jex, A. R., Nejsum, P., Schwarz, E. M., Hu, L., Young, N. D., Hall, R. S., Korhonen, P. K., Liao, S.,
851 Thamsborg, S., & Xia, J. 2014. Genome and transcriptome of the porcine whipworm
852 *Trichuris suis*. Nature Genetics, 46(7), 701–706. DOI: 10.1038/ng.3012.
853 Kizito, D., Tweyongyere, R., Namatovu, A., Webb, E. L., Muhangi, L., Lule, S. A., Bukenya, H.,
854 Cose, S., & Elliott, A. M. 2013. Factors affecting the infant antibody response to measles
855 immunisation in Entebbe-Uganda. BMC Public Health, 13(619), 1–9. DOI: 10.1186/1471-
856 2458-13-619.
857 Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko,
858 D. A., McCombie, W. R., & Jarvis, E. D. 2012. Hybrid error correction and *de novo*
859 assembly of single-molecule sequencing reads. Nature Biotechnology, 30(7), 693–700.
860 DOI: 10.1038/nbt.2280.
861 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. 2017. Canu:
862 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat
863 separation. Genome Research, 27(5), 722–736. DOI: 10.1101/gr.215087.116.
864 Laetsch, D. R., & Blaxter, M. L. 2017. BlobTools: Interrogation of genome assemblies.
865 F1000Research, 6(1287). DOI: 10.12688/f1000research.12232.1.

866 Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18),
867 3094–3100. DOI: 10.1093/bioinformatics/bty191.
868 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,
869 Durbin, R., & Subgroup, G. P. D. P. 2009. The sequence alignment/map format and
870 SAMtools. Bioinformatics, 25(16), 2078-2079. DOI: 10.1093/bioinformatics/btp352.
871 Ludington, A. J., & Sanders, K. L. 2021. Demographic analyses of marine and terrestrial snakes
872 (Elapidae) using whole genome sequences. Molecular Ecology, 30(2), 545–554. DOI:
873 10.1111/mec.15726.
874 Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. 2017. KAT: a K-
875 mer analysis toolkit to quality control NGS datasets and genome assemblies.
876 Bioinformatics, 33(4), 574–576. DOI: 10.1093/bioinformatics/btw663.
877 Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. 2018.
878 MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology,
879 14(1), e1005944. DOI: 10.1371/journal.pcbi.1005944.
880 Marçais, G., & Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of
881 occurrences of k-mers. Bioinformatics, 27(6), 764–770. DOI:
882 10.1093/bioinformatics/btr011.
883 Michalski, M. L., Griffiths, K. G., Williams, S. A., Kaplan, R. M., & Moorhead, A. R. 2011. The NIH-
884 NIAID filariasis research reagent resource center. PLOS Neglected Tropical Diseases,
885 5(11), e1261. DOI: 10.1371/journal.pntd.0001261.
886 Morawski, B. M., Yunus, M., Kerukadho, E., Turyasingura, G., Barbra, L., Ojok, A. M., DiNardo, A.
887 R., Sowinski, S., Boulware, D. R., & Mejia, R. 2017. Hookworm infection is associated

888 with decreased CD4+ T cell counts in HIV-infected adult Ugandans. PLOS Neglected
889 Tropical Diseases, 11(5), e0005634. DOI: 10.1371/journal.pntd.0005634.
890 Nash, S., Mentzer, A. J., Lule, S. A., Kizito, D., Smits, G., van der Klis, F. R., & Elliott, A. M. 2017.
891 The impact of prenatal exposure to parasitic infections and to anthelminthic treatment
892 on antibody responses to routine immunisations given in infancy: Secondary analysis of
893 a randomised controlled trial. PLOS Neglected Tropical Diseases, 11(2), e0005213. DOI:
894 10.1371/journal.pntd.0005213.
895 Ndyomugyenyi, R., Kabatereine, N., Olsen, A., & Magnussen, P. 2008. Malaria and hookworm
896 infections in relation to haemoglobin and serum ferritin levels in pregnancy in Masindi
897 district, western Uganda. Transactions of the Royal Society of Tropical Medicine and
898 Hygiene, 102(2), 130–136. DOI: 10.1016/j.trstmh.2007.09.015.
899 Neal-McKinney, J. M., Liu, K. C., Lock, C. M., Wu, W.-H., & Hu, J. 2021. Comparison of MiSeq,
900 MinION, and hybrid genome sequencing for analysis of *Campylobacter jejuni*. Scientific
901 Reports, 11(1), 5676. DOI: 10.1038/s41598-021-84956-6.
902 Okenmor, G. A., Kutshik, R. J., & Longdet, I. Y. 2022. Draft genome sequence of *Lysinibacillus* sp.
903 strain BPa_S21, isolated from petroleum-contaminated soil in Delta State, Nigeria.
904 Microbiology Resource Announcements, 11(8), e00456-00422. DOI:
905 10.1128/mra.00456-22.
906 Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy,
907 A. M. 2016. Mash: fast genome and metagenome distance estimation using MinHash.
908 Genome Biology, 17(132), 1–14. DOI: 10.1186/s13059-016-0997-x.

909 Organization, W. H. 2021. Ending the neglect to attain the Sustainable Development Goals: A
910 road map for neglected tropical diseases 2021–2030 [Internet].
911 Pages, H., Aboyoun, P., Gentleman, R., DebRoy, S., Pages, M. H., DataImport, D., BSgenome, S.,
912 XStringSet-class, R., MaskedXString-class, R., & XStringSet-io, R. 2013. Package
913 'Biostrings'. Bioconductor, 18129.
914 Parra, G., Bradnam, K., & Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in
915 eukaryotic genomes. Bioinformatics, 23(9), 1061–1067. DOI:
916 10.1093/bioinformatics/btm071.
917 Price, H. J. 1976. Evolution of DNA content in higher plants. The Botanical Review, 42, 27–52.
918 DOI: 10.1007/BF02860861.
919 Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow,
920 W., Fungtammasan, A., & Kim, J. 2021. Towards complete and error-free genome
921 assemblies of all vertebrate species. Nature, 592(7856), 737–746. DOI: 10.1038/s41586-
922 021-03451-0.
923 Roessler, K., Muyle, A., Diez, C. M., Gaut, G. R., Bousios, A., Stitzer, M. C., Seymour, D. K.,
924 Doebley, J. F., Liu, Q., & Gaut, B. S. 2019. The genome-wide dynamics of purging during
925 selfing in maize. Nature Plants, 5(9), 980–990. DOI: 10.1038/s41477-019-0508-7.
926 Seppey, M., Manni, M., & Zdobnov, E. M. 2019. BUSCO: assessing genome assembly and
927 annotation completeness. Gene prediction: methods and protocols(33), 227–245. DOI:
928 10.1093/bioinformatics/btx153.
929 Shen, W., Le, S., Li, Y., & Hu, F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q
930 file manipulation. PLOS One, 11(10), e0163962. DOI: 10.1371/journal.pone.0163962.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics, 15(2), 121–132. DOI: 10.1038/nrg3642.

Solares, E. A., Tao, Y., Long, A. D., & Gaut, B. S. 2021. HapSolo: an optimization approach for removing secondary haplotigs during diploid genome assembly and scaffolding. BMC Bioinformatics, 22(1), 1–17. DOI: 10.1186/s12859-020-03939-y.

Špakulová, M., Kralova, I., & Cutillas, C. 1994. Studies on the karyotype and gametogenesis in *Trichuris muris*. Journal of Helminthology, 68(1), 67–72. DOI: 10.1017/S0022149X0001350X.

Spinner, N. B., Conlin, L. K., Mulchandani, S., & Emanuel, B. S. 2013. Deletions and other structural abnormalities of the autosomes. *In* Emery and Rimoin's Essential Medical Genetics. D. L. Rimoin, R. E. Pyeritz, & B. Korf (Eds.). Vol. 161. Oxford, UK: Elsevier.

Sweet, A. D., & Johnson, K. P. 2018. The role of parasite dispersal in shaping a host–parasite system at multiple evolutionary scales. Molecular Ecology, 27(24), 5104–5119. DOI: 10.1111/mec.14937.

Team, R. 2020. RStudio: Integrated Development for R. http://www.rstudio.com/. PBC, Boston, MA, USA.

Team, R. C. 2023. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research, 27(5), 737–746. DOI: 10.1101/gr.214270.116.

953 Virrueta Herrera, S., Johnson, K. P., Sweet, A. D., Ylinen, E., Kunnasranta, M., & Nyman, T. 2022.
954 High levels of inbreeding with spatial and host-associated structure in lice of an
955 endangered freshwater seal. Molecular Ecology, 31(18), 4593–4606. DOI:
956 10.1111/mec.16569.
957 Voelker, J., Shepherd, M., & Mauleon, R. 2021. A high-quality draft genome for *Melaleuca*
958 *alternifolia* (tea tree): a new platform for evolutionary genomics of myrtaceous terpene-
959 rich species. GigaByte(gigabyte28). DOI: 0.46471/gigabyte.28.
960 Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., &
961 Schatz, M. C. 2017. GenomeScope: fast reference-free genome profiling from short
962 reads. Bioinformatics, 33(14), 2202–2204. DOI: 10.1093/bioinformatics/btx153.
963 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng,
964 Q., Wortman, J., & Young, S. K. 2014. Pilon: an integrated tool for comprehensive
965 microbial variant detection and genome assembly improvement. PLOS One, 9(11),
966 e112963. DOI: 10.1371/journal.pone.0112963.
967 Wang, W., Das, A., Kainer, D., Schalamun, M., Morales-Suarez, A., Schwessinger, B., & Lanfear,
968 R. 2020. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for
969 comparing de novo assemblies. Gigascience, 9(1), giz160. DOI:
970 10.1093/gigascience/giz160.
971 Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., & Burbano, H. A. 2018. nQuire: a statistical
972 framework for ploidy estimation using next generation sequencing. BMC Bioinformatics,
973 19(122). DOI: 10.1186/s12859-018-2128-z.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis (978-3-319-24277-4, https://ggplot2.tidyverse.org ed.). Springer-Verlag New York. pp.

Xie, Y., Xu, Z., Zheng, Y., Li, Y., Liu, Y., Wang, L., Zhou, X., Zuo, Z., Gu, X., & Yang, G. 2019. The mitochondrial genome of the dog hookworm *Ancylostoma caninum* (Nematoda, Ancylostomatidae) from Southwest China. Mitochondrial DNA Part B, 4(2), 3002–3004. DOI: 10.1080/23802359.2019.1666048.

Yi, L., Dalai, M., Su, R., Lin, W., Erdenedalai, M., Luvsantseren, B., Chimedtseren, C., Wang, Z., & Hasi, S. 2020. Whole-genome sequencing of wild Siberian musk deer (*Moschus moschiferus*) provides insights into its genetic features. BMC Genomics, 21(108), 1–13. DOI: 10.1186/s12864-020-6495-2.

Yousuf, P., Parvaiz, S., Razzak, S., & Zehbi, N. (Eds.). (2022). *Trichuris: A critical review*: IntechOpen.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. 2013. The MaSuRCA genome assembler. Bioinformatics, 29(21), 2669–2677. DOI: 10.1093/bioinformatics/btt476.