2024

Assessing De Novo Parasite Genomes Assembled Using Only Oxford Nanopore Technologies MinION Data

Kaylee Herzog

Rachel Wu

John M. Hawdon

Peter Nejsum

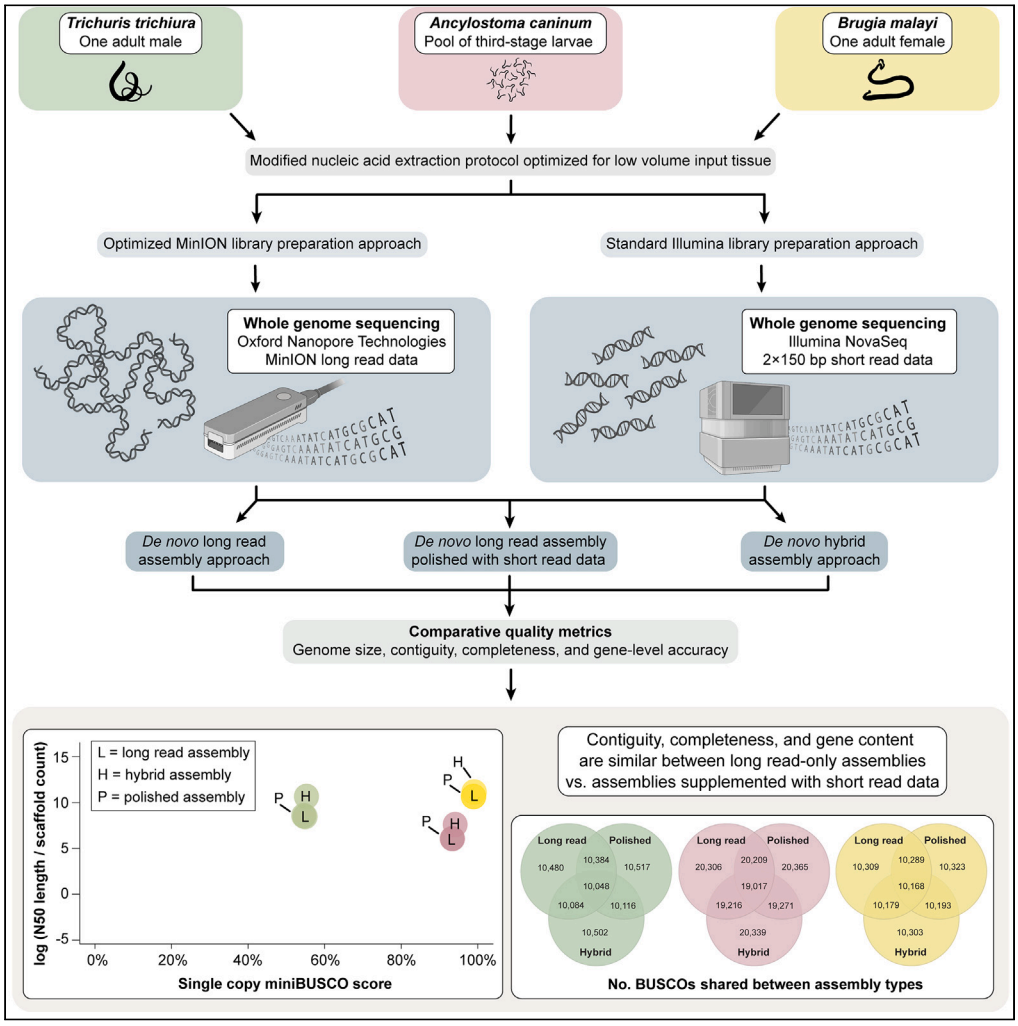
Joseph R. Fauver

iScience

CellPress
OPEN ACCESS

Article

Assessing *de novo* parasite genomes assembled using only Oxford Nanopore Technologies MinION data



Kaylee S. Herzog, Rachel Wu, John M. Hawdon, Peter Nejsum, Joseph R. Fauver

jfauver@unmc.edu

Highlights

Reference-quality helminth genomes are assembled using only MinION data

Polishing MinION genomes with Illumina data moderately improves gene-level accuracy

Modified protocols allow for whole genome sequencing and assembly from single helminths

Herzog et al., iScience 27, 110614
September 20, 2024
https://doi.org/10.1016/j.isci.2024.110614

iScience

CellPress
OPEN ACCESS

Article

Assessing *de novo* parasite genomes assembled using only Oxford Nanopore Technologies MinION data

Kaylee S. Herzog,[1] Rachel Wu,[1] John M. Hawdon,[2] Peter Nejsum,[3] and Joseph R. Fauver[1,4,*]

SUMMARY

In this study, we assessed the quality of *de novo* genome assemblies for three species of parasitic nematodes (*Brugia malayi*, *Trichuris trichiura*, and *Ancylostoma caninum*) generated using only Oxford Nanopore Technologies MinION data. Assemblies were compared to current reference genomes and against additional assemblies that were supplemented with short-read Illumina data through polishing or hybrid assembly approaches. For each species, assemblies generated using only MinION data had similar or superior measures of contiguity, completeness, and gene content. In terms of gene composition, depending on the species, between 88.9 and 97.6% of complete coding sequences predicted in MinION data only assemblies were identical to those predicted in assemblies polished with Illumina data. Polishing MinION data only assemblies with Illumina data therefore improved gene-level accuracy to a degree. Furthermore, modified DNA extraction and library preparation protocols produced sufficient genomic DNA from *B. malayi* and *T. trichiura* to generate *de novo* assemblies from individual specimens.

INTRODUCTION

Parasitic nematodes represent an enormous burden of disease. It is estimated that 1.5 billion people are infected with soil-transmitted helminths, while approximately 120 million and 15 million people suffer from filariasis and onchocerciasis, respectively.[1] Helminths are responsible for nearly 12 million disability-adjusted life years lost annually.[1] In addition to the morbidity caused by nematodes themselves, infection also exacerbates the disease burden of other existing conditions such as malaria, HIV, and tuberculosis, and can reduce immunological response to important vaccines.[2–6] Reference genomes for parasitic nematodes have proven to be an invaluable resource for increasing our understanding of helminth biology, treatment, and control. For example, the landmark comparative genomic study by the International Helminth Genomes Consortium[7] analyzed dozens of complete genomes representing all major helminth lineages. This extensive dataset allowed for the identification of multiple gene family expansions in various groups that are targets for novel drug and vaccine development. Additional more focused comparative genomic work has revealed differences between helminth and host metabolic pathways and extracellular vesicle protein content that represent potential druggable targets, evidence for positive selection in gene families that are uniquely expanded in flatworms and implicated in the biology of endoparasitism, conserved immunomodulatory proteins that are potential vaccine targets for soil-transmitted helminths, and patterns of coevolution between parasitic roundworms and their mammalian hosts.[8–14] These advances further highlight helminth reference genomes as indispensable tools for comparative biological and biomedical research.

Reductions in the cost and improved ease of access of next-generation sequencing (NGS) have made the assembly of whole genomes from helminths more feasible.[15] As a result, reference genomes have now been generated for the majority of species of medical and veterinary importance.[16,17] Most of these species, however, are still represented by only a single reference genome. These solitary references are often generated from laboratory models that have been maintained for decades and therefore cannot represent the biological diversity observed in natural helminth populations.[18,19] Generating multiple reference genomes per species from contemporary and geographically disparate populations would begin to capture the genomic diversity of helminths, and allow for characterization of differences in, for example, genome organization, gene copy number and structural variants, and adaptation to local selective pressures. When generating a genome from a natural population, using genomic DNA from one specimen is ideal to avoid complications in the assembly process. Therefore, streamlined laboratory and computational workflows that allow high-quality assemblies to be generated from an individual specimen using a single data source would be invaluable. The Oxford Nanopore Technologies (ONT) MinION is a single molecule sequencing platform capable of generating long reads ideal for *de novo* genome assembly. Lower read-level accuracy of ONT data have previously required long-read assemblies to be error-corrected with more accurate short-read data. Recent updates to ONT chemistries have decreased the error rate of

[1]Department of Epidemiology, University of Nebraska Medical Center, Omaha, NE 68198, USA
[2]Department of Microbiology, Immunology, and Tropical Medicine, The George Washington University, Washington, DC 20037, USA
[3]Department of Clinical Medicine, Aarhus University, 8200 Aarhus, Denmark
[4]Lead contact
*Correspondence: jfauver@unmc.edu
https://doi.org/10.1016/j.isci.2024.110614

MinION reads, allowing for the potential to generate helminth reference genomes using only long-read data. In fact, Lee et al. (2023) were able to generate complete genomes from individual free-living nematodes using ONT and transcriptome data.[20] However, similar approaches have not been explored in parasitic nematodes.

This study aims to assess the contiguity, completeness, gene content, and gene composition of whole genome *de novo* assemblies of parasitic nematodes using MinION data only compared to assemblies supplemented with accurate short-read Illumina data through polishing and/or hybrid assembly approaches. Three species that collectively span the breadth of parasitic nematode diversity were utilized for sequencing: the filarial worm *Brugia malayi* (Spirurina), the whipworm *Trichuris trichiura* (Trichinellida), and the dog hookworm *Ancylostoma caninum* (Rhabditina). Assemblies generated using only ONT MinION data were compared to reference genomes for each species, as well as to assemblies supplemented with short-read Illumina data through polishing or hybrid assembly approaches that were generated as a part of this study. This work highlights a straightforward approach for generating high-quality *de novo* genome assemblies from parasitic nematodes using only data generated from the ONT MinION.

RESULTS

Data generation

Total gDNA was successfully extracted from a single adult worm for each of *B. malayi* and *T. trichiura*, and from a pool of L3 larvae for *A. caninum*. For each species, both MinION and Illumina sequencing libraries were prepared from the same source of gDNA. The bioinformatic pipeline used to generate each assembly type is outlined in Figure 1. Total amounts sequence data generated for each species prior to basecalling and/or quality control were as follows: ~15.9 Gb (MinION) and ~7.9 Gb (Illumina) for *B. malayi*; ~26.8 Gb (MinION) and ~4.5 Gb (Illumina) for *T. trichiura*; and ~33.1 Gb (MinION) and ~20.8 Gb (Illumina) for *A. caninum* (Table S1). For all three species, a median quality score value of >Q16 was achieved for MinION reads, and for MinION reads aligned to final MinION data only assemblies (Figure S1). Values for average depth of coverage of quality-controlled reads mapped to species-specific reference assemblies were as follows: 124.85× (MinION) and 82.95× (Illumina) for *B. malayi*; 249.91× (MinION) and 48.58× (Illumina) for for *T. trichiura*; and 49.42× (MinION) 38.64× (Illumina) for *A. caninum*.

Selection of assembly approaches

A total for four MinION-only assembly approaches and three hybrid assembly approaches were compared to determine which assembly algorithms would produce the highest quality genome assemblies from our datasets. For the MinION data only assembly of the *B. malayi* genome, Canu outperformed WTDBG2, Shasta, and Flye for contiguity and completeness metrics. For hybrid genome assembly of the *B. malayi* genome, MaSuRCA outperformed WENGAN and HASLR for contiguity and completeness metrics (Table S2). All additional comparative analyses for *B. malayi*, *T. trichiura*, and *A. caninum* were conducted using the MinION data only assembly generated from Canu and the hybrid assembly from MaSuRCA, respectively.

Assembly size, heterozygosity, and contiguity

All assemblies generated in this study were shorter in terms of total length compared to current reference genomes, with the exception of the MinION data only assembly for *B. malayi* (Table 1). Genome sizes estimated by GenomeScope were similarly shorter than those of existing references (Table S1; Figure S2). The best-fit GenomeScope model was that of the *B. malayi* dataset, which estimated low heterozygosity. GenomeScope models for *T. trichiura* and *A. caninum* each show two peaks in the frequency spectra, consistent with estimates of elevated levels of heterozygosity. Of the assemblies generated in this study, hybrid assemblies were consistently the most contiguous as represented by higher N50 values (Table 1). The most improvement in terms of contiguity was observed in *A. caninum*, where each assembly approach resulted in a shorter genome and more contiguous assembly compared to the current reference genome. Within a species, all assemblies had similar levels of GC content.

Assembly completeness

For all species and assembly types, complete mitochondrial genomes were concurrently generated. For *B. malayi*, the complete genome for the *Wolbachia* endosymbiont was also assembled in a single circular genome for both the MinION data only assembly and the hybrid genome assembly. BUSCO scores were nearly identical among the three assembly types generated for all three species and matched or exceeded completeness scores of current reference assemblies (Tables 2 and S3). For the Nematoda reference ortholog database, specifically, proportions of single copy orthologs identified in each assembly generated ranged from 98.79 to 98.88% (*B. malayi*), 56.47–56.72% (*T. trichiura*), and 91.50–92.40% (*A. caninum*). For *A. caninum*, BUSCO scores indicate that the assemblies generated here are more complete than the existing International Helminth Genomes Consortium7 reference assembly. All assemblies compared for *T. trichiura*, including both available references, had high proportions of missing BUSCOs when analyzed with the Nematoda reference ortholog database (i.e., ~39%; see Table 2). According to the assessment of contiguity versus completeness, the assemblies generated here for *B. malayi* and *A. caninum* can be classified as "tier 1" genomes sensu the International Helminth Genomes Consortium7 (i.e., >85% single copy BUSCO score and >1.6 log value for the defined contiguity metric; Figure 2). For *T. trichiura*, the contiguity of assemblies generated here is sufficient to qualify them each as "tier 1", but, as mentioned previously, high proportions of BUSCO missingness prevented any trichurid assembly assessed from achieving "tier 1" status (see Figure 2). For *B. malayi* and *A. caninum*, four contigs totaling 49,030 bp and four contigs totaling 116,599 bp, respectively,

Figure 1. Bioinformatic pipelines used to generate *de novo* whole genome assemblies

The MinION data only pipeline is provided at left and the hybrid pipeline is provided at right.

were confidently identified by BlobTools as potential contamination and were removed from final assemblies (see Table S4). Blobplots for all species are presented in Figures S3–S5 (Minion data only assemblies) and Figures S6–S8 (hybrid assemblies).

Gene content and composition

For organelle genomes, within a species, the mitochondrial and *Wolbachia* genomes generated were nearly identical (i.e., >99.9% pairwise nucleotide identity) across assembly types (Table S4). Genome-wide nucleotide-level pairwise identity varied across species and assembly type comparisons, ranging from 99.04 to 99.74% for *A. caninum* to 99.22–99.86% for *T. trichiura* to 99.57–99.89% for B. malayi (Table S5). For gene datasets produced by GeMoMa, the number of predicted genes was roughly equal across assembly types for each species (Figure 3). Additionally, the majority of these genes were shared between assembly types. Within a species, genes predicted across assembly types were similar or identical in mean gene length, mean length of introns, exons, and coding sequences, and mean number of exons per coding sequence (Table 3). Pairwise nucleotide comparisons of genes shared between MinION only data assemblies and MinION assemblies polished with Illumina data showed 88–98% of these genes to be identical at the nucleotide level (Figure 4A). The majority of differences in gene composition were the result of single SNPs or single indels, with few shared genes demonstrating greater than ten mismatches (Figure 4B).

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

	Length (bp)	No. contigs	N50 (bp)	GC content	N content
Brugia malayi					
Ghedin et al. (2007) reference assembly[a]	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%
MinION assembly polished with Illumina data	88,428,546	88	4,667,161	28.45%	0%
Hybrid assembly	84,528,932	49	3,943,761	28.57%	<0.001%
Trichuris trichiura					
Foth et al. (2014) reference assembly[b]	75,474,068	4,086	70,602	42.3%	0.30%
Doyle et al. (2022) reference assembly	80,573,711	113	11,299,416	42.3%	0.31%
MinION data only assembly	73,094,541	174	806,679	42.22%	0%
MinION assembly polished with Illumina data	72,959,001	174	805,117	42.23%	0%
Hybrid assembly	72,705,678	63	2,773,302	41.15%	0%
Ancylostoma caninum					
International Helminth Genomes Consortium (2019) reference assembly[c]	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%
MinION assembly polished with Illumina data	357,986,713	1,578	637,973	43.38%	0%
Hybrid assembly	348,034,555	652	1,095,227	43.24%	<0.001%

Abbreviations: bp = base pairs.
[a]Reference assembly included PacBio, optical mapping, and Sanger sequencing data.
[b]Reference assembly generated using Illumina data.
[c]Reference assembly generated using 454 sequencing.

DISCUSSION

Sample processing and sequencing

The DNA extraction and MinION library preparation protocols described here were optimized for low input gDNA extracted from individual parasitic nematodes, allowing us to retain the majority of input gDNA through library preparation. Genomic DNA from a single individual is the ideal input for *de novo* whole genome assembly for diploid organisms. This allows assembly pipelines to contend with only 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.[21] Additionally, sampling from natural, rather than laboratory-maintained, populations is preferred for many genomic studies, further highlighting the advantage of whole-genome sequencing from single individuals.[22] For *B. malayi* and *T. trichiura*, which both have relatively large adult stages (~3–5 cm in length) and tractable genome sizes (<100 Mb), optimized 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 have the largest genome size of the three focal species in this study (i.e., ~400 Mb). Thus, for *A. caninum*, even if adult worms had been accessible, these protocols would likely still not have allowed for generation of sufficient genomic data from a single adult. Despite using a pooled sample, however, copy number spectrum plots suggest that incorporation of a purging step produces results similar to those for single individuals (see Figure S9). For nematodes where gDNA from single individuals does not meet the requisite quantity for sequencing on the ONT MinION, additional strategies such as whole genome amplification can be pursued.

There was an association between the average Q score of MinION read datasets and the total amount of data generated by a flow cell: more efficient MinION sequencing runs generated data with higher average Q scores (see Table S1; Figure S1). For example, MinION libraries for both B. malayi and *T. trichiura* generated ~10.8–11.8 Gb of data per flow cell post-basecalling with comparably high Q scores. Conversely, libraries for *A. caninum* generated ~7.7–8.5 Gb of data per flow cell post-basecalling, and these data had the lowest average Q score of the three species sequenced (see Table S1; Figure S1). We observed that flow cells that were used to sequence the *A. caninum* libraries experienced the most rapid decline in pore availability. This observation highlights the variability in data generation and quality depending on the sample type when using the latest ONT chemistries.

Estimation of genome size and heterozygosity

There were appreciable differences in overall size of assemblies produced in this study compared to reference genomes (see Table 1). These differences were notable for both *T. trichiura* and *A. caninum*, but most dramatic for *A. caninum*, where assemblies produced here

Table 2. Scores from compleasm for the assemblies generated as part of this study and the reference assemblies available for each species

	Single copy	Duplicated	Fragmented	Missing
Brugia malayi				
Ghedin et al. (2007) reference assembly	3,097 (98.91%)	16 (0.51%)	10 (0.32%)	8 (0.26%)
MinION data only assembly	3,096 (98.88%)	17 (0.54%)	10 (0.32%)	8 (0.26%)
MinION assembly polished with Illumina data	3,096 (98.88%)	17 (0.54%)	10 (0.32%)	8 (0.26%)
Hybrid assembly	3,093 (98.79%)	17 (0.54%)	11 (0.35%)	10 (0.32%)
Trichuris trichiura				
Foth et al. (2014) reference assembly	1,746 (55.76%)	55 (1.76%)	99 (3.16%)	1,231 (39.32%)
Doyle et al. (2022) reference assembly	1,746 (55.76%)	55 (1.76%)	99 (3.16%)	1,231 (39.32%)
MinION data only assembly	1,768 (56.47%)	26 (0.83%)	93 (2.97%)	1,244 (39.73%)
MinION assembly polished with Illumina data	1,768 (56.47%)	26 (0.83%)	94 (3.00%)	1,243 (39.70%)
Hybrid assembly	1,776 (56.72%)	27 (0.86%)	97 (3.10%)	1,231 (39.32%)
Ancylostoma caninum				
International Helminth Genomes Consortium (2019) reference assembly	2,657 (84.86%)	242 (7.73%)	146 (4.66%)	86 (2.75%)
MinION data only assembly	2,868 (91.60%)	103 (3.29%)	67 (2.14%)	93 (2.97%)
MinION assembly polished with Illumina data	2,865 (91.50%)	105 (3.35%)	69 (2.20%)	92 (2.94%)
Hybrid assembly	2,893 (92.40%)	110 (3.51%)	54 (1.72%)	74 (2.36%)

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).

were >100 Mb smaller than the existing reference. This is likely due to the fact that a purging step to remove haplotypic duplication was included in the bioinformatic pipeline for all assembly types generated in this study (see Figure 1). This was not the case for the reference genome for *A. caninum*7, and thus it likely includes haplotypic duplication that artificially inflates genome size. Three lines of evidence support that the genomes assembled here more accurately represent the true haplotypic genome size for both *T. trichiura* and *A. caninum*. First, the lengths of these assemblies more closely match k-mer based genome size estimates than do those of the existing reference genomes (see Table 1). Second, BUSCO scores were largely indistinguishable across all assembly types (including existing reference assemblies), indicating that all are comparably complete despite ranging in size. In fact, the Doyle et al. reference assembly for *T. trichiura*[23] and the International Helminth Genomes Consortium reference assembly for *A. caninum*[7] had higher proportions of duplicated BUSCOs as compared to the assemblies generated herein (see Tables 2 and S3). Finally, for hybrid assemblies, copy number spectrum plots suggest they were not "over-purged" and thus missing genomic content (see Figure S9). As previously advocated by other authors, purging haplotypic duplication is an important step in genome curation regardless of assembly approach.[24,25] Unsurprisingly, purging becomes increasingly important for highly heterozygous sample types, as observed here for *A. caninum*.

Results from GenomeScope (see Table S1; Figure S2) and copy number spectrum plots (see Figure S9) indicate low heterozygosity for *B. malayi* and elevated heterozygosity for *T. trichiura* and *A. caninum*. These results are congruent with the fact that the single individual of *B. malayi* sequenced came from a long-maintained inbred laboratory strain while the individual of *T. trichiura* sequenced came from a naturally infected human host, and sequence data for *A. caninum* were generated from a pool of individuals. When sequencing adult females that are potentially gravid, as was the case for *B. malayi*, analyses like GenomeScope and KAT that utilize k-mer spectra can be altered by the presence of paternal haplotypes in eggs. Given that the female of *B. malayi* sequenced came from an inbred strain, however, maternal and paternal haplotypes are likely to be identical, or highly similar, ameliorating this concern. It is worth noting that the k-mer based approaches used to estimate genome size and heterozygosity and to assess the effectiveness of purging rely on short-read data as input, and therefore cannot be used to evaluate long-read only assemblies at this time.

Wolbachia genome assembly from *Brugia malayi*

Complete circular *Wolbachia* genomes 1.08 Mb in length were assembled alongside the *B. malayi* genome for both assembly approaches. *Wolbachia* genomes generated in this study were highly similar to one another (>99.9%; 19 nucleotide mismatches) at the nucleotide level upon comparison via multisequence alignment. This was expected, given they were generated from the same individual sample. This high level of nucleotide conservation extends to the publicly available *Wolbachia* genomes generated from the FR3 strain of *B. malayi* (GenBank: CP034333.1, AE017321.1; Table S4).[26,27] The majority of nucleotide mismatches identified between the *Wolbachia* genomes generated from the MinION data only and hybrid assembly occur in homopolymer regions with >5 nucleotide repeats, and are most commonly represented by a single base insertion in these genomes as compared to the references. The hybrid genome contained more nucleotide mismatches

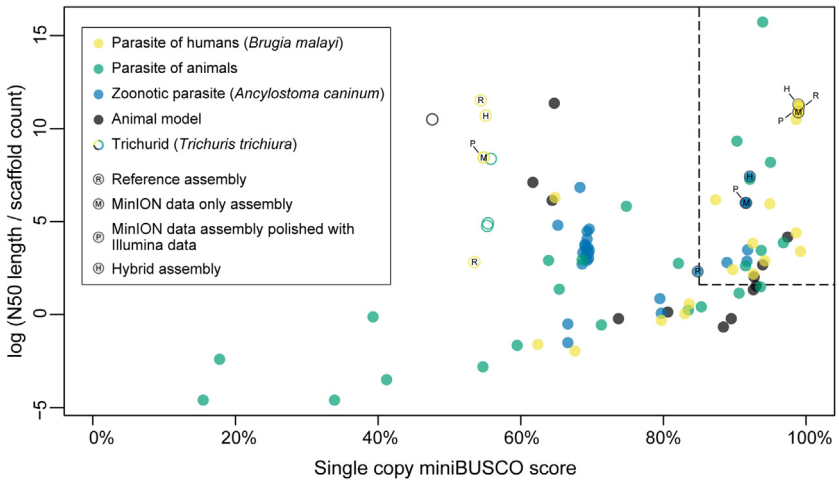


Figure 2. Plot of single copy BUSCO score versus the log value of N50 length divided by contig (or scaffold) count for the assemblies generated as part of this study, and for all genomes of nematodes that parasitize animals as adults available from WormBase ParaSite

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. The Nematoda ortholog database (n = 3,131 BUSCOs) was used for comparison.

compared to the reference genomes (185 and 185 mismatches, respectively) than the MinION data only assembly (23 and 18 mismatches, respectively). These discrepancies are due to two sections of the genome that contain a large number of multiple base pair indels and large (>5bp) runs of SNPs, which is likely the result of a misassembly event. Overall, the high level of nucleotide identity between all Wolbachia genomes assessed in this study is expected given that they are generated from the same laboratory strain of *B. malayi* that has been in isolation since 1970.[28]

Assembly contiguity

We assessed a number of different MinION data only and hybrid assembly approaches to assess which program gave us the highest quality genomes for our datasets generated in the study in order to compare genomes generated with MinION data only compared to genomes supplemented with accurate short read data. The reference genome assemblies for both *B. malayi* and *T. trichiura* are highly contiguous and nearing chromosome scale. For *B. malayi* and *T. trichiura*, the N50 values for the assemblies generated in this study were substantially lower compared to those of the highly contiguous reference genome. This is due in part to the additional data sources used to generate these references, which for *B. malayi* included optical mapping[29] and for *T. trichiura* included unspecified updates that resulted in improvements to contiguity.[23] For *A. caninum*, however, the contiguity of the assemblies generated here represents marked improvements as compared to the reference genome. Hybrid assemblies were in all cases more contiguous than MinION data assemblies (see Table 1). This result was somewhat unexpected given that including short-read data in whole genome assembly has traditionally been touted to increase accuracy rather than contiguity.[30–32] Increased contiguity of hybrid assemblies over long read-only assemblies has been observed in bacteria whole genome sequencing[33,34] but does not appear to have been reported previously for eukaryotes. It is worth noting that the long reads generated for each species here did not approach the read lengths that MinION devices are capable of sequencing. Across species, read length N50 values ranged from ~2.7 to 8.6 kb (see Table S1), which are relatively small. This may be due to the optimized MinION library preparation approach used here, which relied on washing final libraries with a titration of short and long fragment buffer to retain some short fragments of gDNA in addition to long fragments (see STAR Methods). Rigorous size selection which could have improved read lengths was therefore sacrificed to enable sequencing from low quantities of input gDNA. Had it been possible, generation of ultra-long reads would have likely lessened the gap in contiguity metrics between MinION data and hybrid assemblies. Additionally, a single assembly algorithm was chosen to generate long-read and hybrid assemblies (i.e., Canu and MaSuRCA, respectively). Given the same read-level data, different assembly algorithms may have produced assemblies with different contiguity metrics.

Assembly completeness

In general, completeness values as measured by BUSCO scores from compleasm are indistinguishable between assembly types for *B. malayi* and *T. trichiura*, regardless of which reference ortholog database is used for comparison. Assemblies generated here for *A. caninum* showed an improvement in the number of single copy genes identified and a reduction in the number of duplicated and fragmented genes identified as compared to the reference genome across comparative databases. Results from compleasm indicate a high proportion of missingness for all trichurid assemblies when compared to the Nematoda reference ortholog database (see Table 2; Figure 2). Underrepresentation in the BUSCO "nematoda_odb10" dataset has been noted to result in biases and underestimation in ortholog detection for some species of

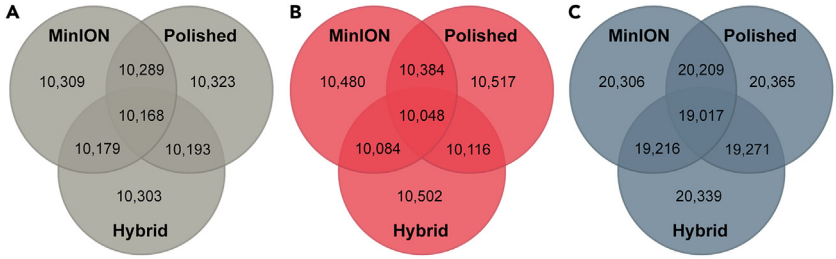


Figure 3. Venn diagrams of the number of genes predicted by GeMoMa that were shared among the assembly types generated
(A) *Brugia malayi*, (B) *Trichuris trichiura*, and (C) *Ancylostoma caninum*.

nematodes.[35] When comparing to higher-level Metazoa and Eukaryota databases, elevated proportions of missing BUSCOs were still found in the trichurid assemblies (~25% for Metazoa and ~7–8% for Eukaryota; see Table S3), albeit not high as those found for the Nematoda database. This presents a challenge when using BUSCO scores as a completeness metric to assess trichurid genome assemblies.

Gene content and composition

In terms of gene content, within a species, similar numbers of genes were predicted by GeMoMA for each of the three assembly types generated (see Figure 3). Though not all predicted genes were identified in all three assemblies, large proportions were found to be shared among them (i.e., ~93–98%, depending on the species). In particular, MinION only assemblies and MinION assemblies polished with Illumina data were highly similar in terms of gene content, with ~98.7–99.7% of predicted genes shared between them within a species. Furthermore, predicted genes sets had, by species, identical or nearly identical mean genes lengths, mean intron, exon, and coding sequence lengths, and mean numbers of exons per coding sequence (see Table 3), providing further evidence for a high degree of similarity in gene content across assembly types.

In terms of gene composition, a major goal of this study was to assess the potential of Illumina data to correct errors that may be present in the final assembly as a result of less accurate MinION data. We took multiple approaches to evaluate this, including calculating pairwise nucleotide-level identity between different assembly approaches at both the genome and gene level. Genome level nucleotide identities ranged from 99.04% to 99.89% (Table S5). Gene datasets predicted by GeMoMa for MinION data only assemblies and for MinION assemblies polished with Illumina data were pairwise aligned and assessed for mismatches. Comparisons between MinION data only assemblies and hybrid assemblies are not tenable because long-read assemblers like Canu and hybrid assemblers like MaSuRCA utilize assembly algorithms that handle read-level data dissimilarly.[36,37] Given that gene predictions for each assembly were based on existing reference genomes and their corresponding annotations (see STAR Methods), plus these reference genomes were generated from different biological samples, comparison to reference genomes was also not salient.

Table 3. Comparative summary statistics output by AGAT for genes predicted by GeMoMa in assembly types for each species

	Mean gene length	Mean exon length	Mean intron length	Mean CDS length	Mean no. exons per CDS
Brugia malay					
MinION data assembly	4,299	148	346	1,394	9.4
MinION assembly polished with Illumina data	4,291	148	346	1,394	9.4
Hybrid assembly	4,267	148	344	1,389	9.4
Trichuris trichiura					
MinION data assembly	2,916	212	325	1,280	6.0
MinION assembly polished with Illumina data	2,909	212	325	1,279	6.0
Hybrid assembly	2,918	212	324	1,283	6.0
Ancylostoma caninum					
MinION data assembly	2,663	125	332	822	6.5
MinION assembly polished with Illumina data	2,655	125	330	821	6.5
Hybrid assembly	2,681	125	332	825	6.6

Mean lengths are presented as number of base pairs. Abbreviations: CDS = coding sequence.

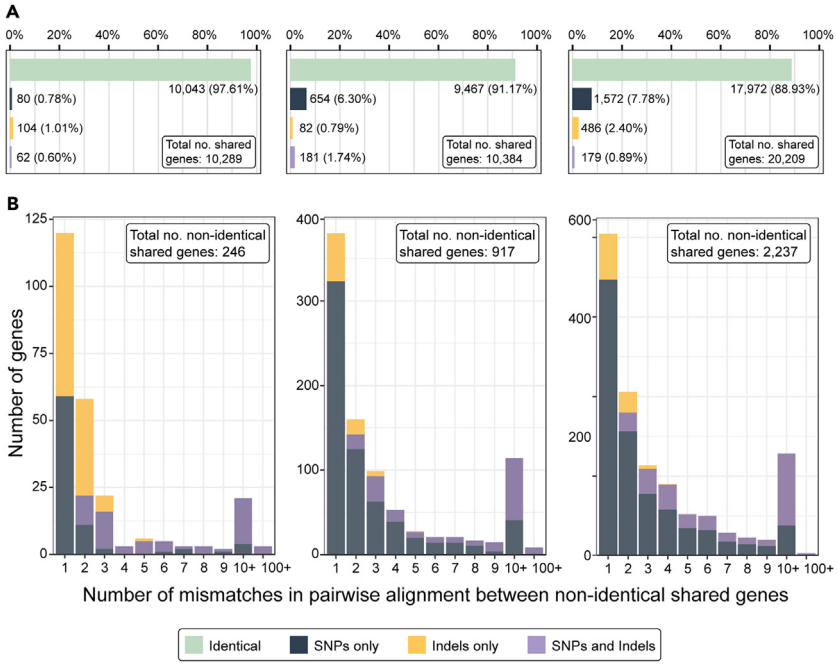


Figure 4. Summary of differences in gene composition identified in predicted genes shared between MinION data assemblies and MinION assemblies polished with Illumina data

(A) Proportions of shared genes identified at the nucleotide level as identical, differing by SNPs only, differing by indels only, or differing by both SNPS and indels for *Brugia malayi* (left), *Trichuris trichiura* (center), and *Ancylostoma caninum* (right).

(B) Number of differences identified between non-identical shared genes for *Brugia malayi* (left), *Trichuris trichiura* (center), and *Ancylostoma caninum* (right). A key to bar plot colors is provided at bottom center.

The most direct comparison to determine whether short reads improve gene accuracy in final assemblies are the MinION data only assembly versus that same assembly polished with Illumina short reads. If Illumina polishing substantially improves accuracy, homology-based gene prediction would be expected to result in a low proportion of genes shared between these two assembly types, and or/the majority of shared genes to differ at the nucleotide level. This was not the case as all three species showed high proportions of predicted genes shared between assembly types, and the majority of the genes shared between the two assembly types were identical (see Figures 3 and 4A). For those that were not identical, most differed by only a single SNP or indel. These single differences are presumably corrections made during Illumina polishing (Figure 4B). Despite this, a substantial number of gene comparisons still contained 10–100+ mismatches (Figure 4B). It is unlikely these differences result from polishing errors in MinION data only assemblies; rather, they are more likely the result of homology-based gene prediction models identifying incongruent genes between assembly types in a small number of cases. In summary, depending on the species, correcting with short reads did not change nucleotide calls in coding sequences for ~88–98% of the genes compared.

Conclusions

For this study, *de novo* whole genome assemblies were generated for three species of parasitic nematodes (*Brugia malayi*, *Trichuris trichiura*, and *Ancylostoma caninum*) using only MinION long-read data, using MinION data polished with Illumina short reads, and using a combined hybrid approach. For B. malayi and *T. trichiura*, optimized gDNA extraction and library preparation protocols allowed for the generation of complete genomes from individual adult worms. For all species sequenced, MinION data only assemblies had similar, or superior, measures of contiguity and completeness as compared to existing reference genomes. The most substantial improvements in quality metrics were observed for *A. caninum*, which was the only one of the three focal species for which the existing reference genome is not a near-chromosome scale assembly. Among the three assembly types generated, predicted gene content was nearly identical with a species, and the vast majority of predicted genes shared between MinION data only assemblies and MinION assemblies polished with short reads were identical at the nucleotide level. For some genes, however, polishing did result in the correction of SNPs or indels, indicating that the inclusion of short reads results in more accurate final genome assemblies. Although additional data types beyond MinION long reads are needed to produce reference-quality, chromosome-scale assemblies, the results of this study demonstrate that MinION data alone can be used to generate highly contiguous and complete *de novo* whole genomes from parasitic helminths. Importantly, these approaches are accessible and can use individual worms as input, allowing for the generation of more genome assemblies that will ultimately increase our understanding of genomic diversity and facilitate population-level genomic analyses for these important parasites.

Limitations of the study

This study has multiple limitations. For two of the three focal species, gDNA was generated from adult worms. For most species, and particularly those that infect humans, adult nematodes are difficult or impossible to obtain, somewhat limiting the utility of this approach in practice. The assemblies generated here remain incomplete in terms of modern reference genome standards. They are not chromosome-scale, they are not annotated, structural variation has not been explored, and additional data types (e.g., RNA-seq) are required to complete them. For *B. malayi* and *T. trichiura*, existing reference genomes were already exceptionally high-quality "tier 1"-status assemblies sensu the International Helminth Genomes Consortium,[7] and the assemblies generated here using a single data source approached their quality in contiguity and matched their quality in completeness. The comparison in which the added value of long-read MinION data are most evident is that of *A. caninum*. Using gDNA from the same strain and sample type on which the existing reference assembly was based, the genomic resources available for this species were improved considerably.

Comparisons of MinION data only versus hybrid assemblies generated from the same sample were limited to genomes assembled using a single approach. There are dozens of comparable tools available for each step of the genome assembly pipeline (e.g., *de novo* assemblers, methods for purging duplication, polishing algorithms, pipelines for identifying contamination, etc.) and comparing all combinations of these many tools was outside of the scope of this study. Benchmarking studies typically rely on prokaryotic genomes with "truth datasets" in which base calls at all positions are known with confidence. In contrast, the nematodes sequenced here are non-model eukaryotes with comparably large and complex genomes for which there are no truth datasets available for comparison. Benchmarking assembly pipelines was therefore not feasible, nor indeed was it the goal of this study. Rather, we sought to evaluate the feasibility and quality of whole genome assemblies generated from an accessible NGS data type. Ultimately, we constructed a straightforward pipeline of well-validated tools (Figure 1) that provided the best results for our data, establishing a roadmap that others seeking to generate high-quality helminth genomes using similar data types can follow moving forward.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Specimen acquisition
  - DNA extraction
  - MinION library preparation
  - MinION sequencing and basecalling
  - Illumina library preparation and sequencing
  - Estimation of genome size and heterozygosity
  - MinION data genome assembly
  - Hybrid genome assembly
  - Refinement of final assemblies
  - Mitogenome and Wolbachia genome assembly
  - Assembly contiguity and completeness
  - Gene content and composition

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110614.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTIONS

Conceptualization: K.S.H. and J.R.F.; methodology: K.S.H. and J.R.F.; investigation; K.S.H., R.W., J.M.H., P.N., and J.R.F.; writing: K.S.H. and J.R.F.; review and editing: K.S.H., J.M.H., P.N., and J.R.F.; resources: J.M.H., P.N., and J.R.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 13, 2024
Revised: June 9, 2024
Accepted: July 26, 2024
Published: July 30, 2024

REFERENCES

1. Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., et al. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet *396*, 1204–1222.
2. Biraro, I.A., Egesa, M., Toulza, F., Levin, J., Cose, S., Joloba, M., Smith, S., Dockrell, H.M., Katamba, A., and 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*, e111517. https://doi.org/10.1371/journal.pone.0111517.
3. Kizito, D., Tweyongyere, R., Namatovu, A., Webb, E.L., Muhangi, L., Lule, S.A., Bukenya, H., Cose, S., and Elliott, A.M. (2013). Factors affecting the infant antibody response to measles immunisation in Entebbe-Uganda. BMC Publ. Health *13*, 619.
4. Morawski, B.M., Yunus, M., Kerukadho, E., Turyasingura, G., Barbra, L., Ojok, A.M., DiNardo, A.R., Sowinski, S., Boulware, D.R., and Mejia, R. (2017). Hookworm infection is associated with decreased CD4+ T cell counts in HIV-infected adult Ugandans. PLoS Negl. Trop. Dis. *11*, e0005634.
5. Nash, S., Mentzer, A.J., Lule, S.A., Kizito, D., Smits, G., van der Klis, F.R.M., and Elliott, A.M. (2017). The impact of prenatal exposure to parasitic infections and to anthelminthic treatment on antibody responses to routine immunisations given in infancy: Secondary analysis of a randomised controlled trial. PLoS Negl. Trop. Dis. *11*, e0005213.
6. Ndyomugyenyi, R., Kabatereine, N., Olsen, A., and Magnussen, P. (2008). Malaria and hookworm infections in relation to haemoglobin and serum ferritin levels in pregnancy in Masindi district, western Uganda. Trans. R. Soc. Trop. Med. Hyg. *102*, 130–136.
7. International Helminth Genomes Consortium (2019). Comparative genomics of the major parasitic worms. Nat. Genet. *51*, 163–174.
8. Bennett, A.P.S., de la Torre-Escudero, E., and Robinson, M.W. (2020). Helminth genome analysis reveals conservation of extracellular vesicle biogenesis pathways but divergence of RNA loading machinery between phyla. Int. J. Parasitol. *50*, 655–661.
9. Collington, E., Lobb, B., Mazen, N.A., Doxey, A.C., and Glerum, D.M. (2023). Phylogenomic Analysis of 155 Helminth Species Reveals Widespread Absence of Oxygen Metabolic Capacity. Genome Biol. Evol. *15*, evad135. https://doi.org/10.1093/gbe/evad135.
10. Hu, Y., Yu, L., Fan, H., Huang, G., Wu, Q., Nie, Y., Liu, S., Yan, L., and Wei, F. (2021). Genomic Signatures of Coevolution between Nonmodel Mammals and Parasitic Roundworms. Mol. Biol. Evol. *38*, 531–544.
11. Luo, G., Gong, R., Li, P., Li, Q., and Wei, X. (2022). Comparative genomic analysis of Echinococcus multilocularis with other tapeworms. Biologia *77*, 2743–2750.
12. Montaño, K.J., Cuéllar, C., and Sotillo, J. (2021). Rodent Models for the Study of Soil-Transmitted Helminths: A Proteomics Approach. Front. Cell. Infect. Microbiol. *11*, 639573.
13. Rosa, B.A., Choi, Y.-J., McNulty, S.N., Jung, H., Martin, J., Agatsuma, T., Sugiyama, H., Le, T.H., Doanh, P.N., Maleewong, W., et al. (2020). Comparative genomics and transcriptomics of 4 Paragonimus species provide insights into lung fluke parasitism and pathogenesis. GigaScience *9*, giaa073. https://doi.org/10.1093/gigascience/giaa073.
14. Wang, J. (2021). Genomics of the Parasitic Nematode Ascaris and Its Relatives. Genes *12*, 493. https://doi.org/10.3390/genes12040493.
15. Doyle, S.R. (2022). Improving helminth genome resources in the post-genomic era. Trends Parasitol. *38*, 831–840.
16. Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P., and Berriman, M. (2017). WormBase ParaSite-a comprehensive resource for helminth genomics. Mol. Biochem. Parasitol. *215*, 2–10.
17. Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., et al. (2016). WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. *44*, D774–D780.
18. Valiente-Mullor, C., Beamud, B., Ansari, I., Francés-Cuesta, C., García-González, N., Mejía, L., Ruiz-Hueso, P., and González-Candelas, F. (2021). One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput. Biol. *17*, e1008678.
19. Yang, X., Lee, W.-P., Ye, K., and Lee, C. (2019). One reference genome is not enough. Genome Biol. *20*, 104.
20. Lee, Y.-C., Ke, H.-M., Liu, Y.-C., Lee, H.-H., Wang, M.-C., Tseng, Y.-C., Kikuchi, T., and Tsai, I.J. (2023). Single-worm long-read sequencing reveals genome diversity in free-living nematodes. Nucleic Acids Res. *51*, 8035–8047.
21. Solares, E.A., Tao, Y., Long, A.D., and Gaut, B.S. (2021). HapSolo: an optimization approach for removing secondary haplotigs during diploid genome assembly and scaffolding. BMC Bioinf. *22*, 9.
22. Adams, M., McBroome, J., Maurer, N., Pepper-Tunick, E., Saremi, N.F., Green, R.E., Vollmers, C., and Corbett-Detig, R.B. (2020). One fly-one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster. Nucleic Acids Res. *48*, e75.
23. Doyle, S.R., Søe, M.J., Nejsum, P., Betson, M., Cooper, P.J., Peng, L., Zhu, X.-Q., Sanchez, A., Matamoros, G., Sandoval, G.A.F., et al. (2022). Population genomics of ancient and modern Trichuris trichiura. Nat. Commun. *13*, 3888.
24. Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., Torrance, J., Tracey, A., and Wood, J. (2021). Significantly improving the quality of genome assemblies through curation. GigaScience *10*, giaa153. https://doi.org/10.1093/gigascience/giaa153.
25. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746.
26. Lefoulon, E., Vaisman, N., Frydman, H.M., Sun, L., Voland, L., Foster, J.M., and Slatko, B.E. (2019). Large Enriched Fragment Targeted Sequencing (LEFT-SEQ) Applied to Capture of Wolbachia Genomes. Sci. Rep. *9*, 5939.
27. Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., Bhattacharyya, A., Kapatral, V., Kumar, S., Posfai, J., et al. (2005). The Wolbachia genome of Brugia malayi: endosymbiont evolution within a human pathogenic nematode. PLoS Biol. *3*, e121.
28. Michalski, M.L., Griffiths, K.G., Williams, S.A., Kaplan, R.M., and Moorhead, A.R. (2011). The NIH-NIAID Filariasis Research Reagent Resource Center. PLoS Negl. Trop. Dis. *5*, e1261.
29. Tracey, A., Foster, J.M., Paulini, M., Grote, A., Mattick, J., Tsai, Y.-C., Chung, M., Cotton, J.A., Clark, T.A., Geber, A., et al. (2020). Nearly Complete Genome Sequence of Brugia malayi Strain FR3. Microbiol. Resour. Announc. *9*, e00154-20. https://doi.org/10.1128/MRA.00154-20.
30. Bashir, A., Klammer, A., Robins, W.P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., et al. (2012). A hybrid approach for the automated finishing of bacterial genomes. Nat. Biotechnol. *30*, 701–707.

31. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One *7*, e47768.
32. Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., and Phillippy, A.M. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat. Biotechnol. *30*, 693–700.
33. Neal-McKinney, J.M., Liu, K.C., Lock, C.M., Wu, W.-H., and Hu, J. (2021). Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of Campylobacter jejuni. Sci. Rep. *11*, 5676.
34. George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A.E., Mathers, A., Norris, R., Navickaite, I., Eaton, C., et al. (2017). Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. Microb. Genom. *3*, e000118.
35. Rödelsperger, C. (2021). The community-curated Pristionchus pacificus genome facilitates automated gene annotation improvement in related nematodes. BMC Genom. *22*, 216.
36. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.
37. Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics *29*, 2669–2677.
38. Foth, B.J., Tsai, I.J., Reid, A.J., Bancroft, A.J., Nichol, S., Tracey, A., Holroyd, N., Cotton, J.A., Stanley, E.J., Zarowiecki, M., et al. (2014). Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. Nat. Genet. *46*, 693–700. https://doi.org/10.1038/ng.3010.
39. Xie, Y., Xu, Z., Zheng, Y., Li, Y., Liu, Y., Wang, L., Zhou, X., Zuo, Z., Gu, X., and Yang, G. (2019). The mitochondrial genome of the dog hookworm *Ancylostoma caninum* (Nematoda, Ancylostomatidae) from Southwest China. Mitochondrial DNA. B Resour. *4*, 3002–3004.
40. De Coster, W. NanoPlot: Plotting Scripts for Long Read Sequencing Data (Github).
41. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.
42. Li, H. (2021). New strategies to improve minimap2 alignment accuracy. Bioinformatics *37*, 4572–4574.
43. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
44. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.
45. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
46. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.
47. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158.
48. Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H.E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., et al. (2020). Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. Nat. Biotechnol. *38*, 1044–1053.
49. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546.
50. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962.
51. Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner (Lawrence Berkeley National Lab.(LBNL)).
52. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896–2898.
53. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. *27*, 737–746.
54. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.
55. Di Genova, A., Buena-Atienza, E., Ossowski, S., and Sagot, M.-F. (2021). Efficient hybrid *de novo* assembly of human genomes with WENGAN. Nat. Biotechnol. *39*, 422–430.
56. Haghshenas, E., Asghari, H., Stoye, J., Chauve, C., and Hach, F. (2020). HASLR: Fast Hybrid Assembly of Long Reads. iScience *23*, 101389.
57. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B.J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics *33*, 574–576.
58. Challis, R., Richards, E., Rajan, J., Cochrane, G., and Blaxter, M. (2020). BlobToolKit - Interactive Quality Assessment of Genome Assemblies. G3 *10*, 1361–1374.
59. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
60. Wickham, H. ggplot2 (Springer New York).
61. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
62. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.
63. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.
64. Huang, N., and Li, H. (2023). compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics *39*, btad595. https://doi.org/10.1093/bioinformatics/btad595.
65. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinf. *19*, 189.
66. Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. *44*, e89.
67. Dainat, J.. AGAT: Another Gff Analysis Toolkit to Handle Annotations in Any GTF/GFF Format Version v0.7.0. https://doi.org/10.5281/zenodo.3552717.
68. Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2020). Biostrings: Efficient Manipulation of Biological Strings. https://rdrr.io/bioc/Biostrings/.
69. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Brugia malayi adults	NIH/NIAID Filariasis Research Reagent Resource Center via BEI Resources (Michalski et al.[28])	FR3 strain
Trichuris trichiura adults	Peter Nejsum, Department of Clinical Medicine, Aarhus University	N/A
Ancylostoma caninum third-stage larvae	John M. Hawdon, Department of Microbiology, Immunology, and Tropical Medicine, The George Washington University	Baltimore strain
Chemicals, peptides, and recombinant proteins		
DNA/RNA Shield	Zymo Research, Irvine, CA, USA	Cat#R1200-25
Critical commercial assays		
Quick-DNA HMW MagBead Kit	Zymo Research, Irvine, CA, USA	Cat#D6060
Monarch® Pestle Set single-use microtube pestle	New England Biolabs® Inc., Ipswich, MA, USA	Cat#T3000L
Qubit™ 4 1X dsDNA High Sensitivity (HS) Assay	ThermoFisher Scientific, Waltham, MA, USA	Cat#Q33231
Agilent 2200 TapeStation System Genomic DNA ScreenTape and Reagents	Agilent, Santa Clara, CA, USA	Cat#5067-5365; 5067-5366
AMPure XP	Beckman Coulter, Brea, CA, USA	Cat#A63880
DNA Clean & Concentrator Magbee Kit	Zymo Research, Irvine, CA, USA	Cat#D4012
Ligation Sequencing Kit	Oxford Nanopore Technologies, Oxford, United Kingdom	Cat#SQK-LSK114
NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing	New England Biolabs® Inc., Ipswich, MA, USA	Cat#E7180S
xGen™ cfDNA & FFPE DNA Library Prep Kit	Integrated DNA Technologies, Newark, NJ, USA	Cat#10010207
Deposited data		
Quality-controlled MinION and Illumina read data for *Brugia malayi*	National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)	BioProject accession ID PRJNA1074771; BioSample accession no. SAMN39888962
Quality-controlled MinION and Illumina read data for *Trichuris trichiura*	National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)	BioProject accession ID PRJNA1074771; BioSample accession no. SAMN39888963
Quality-controlled MinION and Illumina read data for *Ancylostoma caninum*	National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)	BioProject accession ID PRJNA1074771; BioSample accession no. SAMN39888964
Final assemblies generated for *Brugia malayi*	National Center for Biotechnology Information (NCBI) Genome database	Accession nos. GCA_037903335.1; GCA_037903345.1; GCA_037903365.1
Final assemblies generated for *Trichuris trichiura*	National Center for Biotechnology Information (NCBI) Genome database	Accession nos. GCA_037903355.1; GCA_037903375.1;GCA_037903465.1
Final assemblies generated for *Ancylostoma caninum*	National Center for Biotechnology Information (NCBI) Genome database	Accession nos. GCA_037903475.1; GCA_037903715.1; GCA_037903745.1
Reference genome for *Brugia malayi*	Tracy et al.[29]	BioProject accession ID PRJNA10729
Reference genome for *Trichuris trichiura*	Foth et al.[38]	BioProject accession ID PRJEB535
Reference genome for *Trichuris trichiura*	Doyle et al.[23]	https://github.com/stephenrdoyle/ancient_trichuris/tree/master/02_data

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Reference genome for *Ancylostoma caninum*	International Helminth Genomes Consortium[7]	BioProject accession ID PRJNA72585
Mitochondrial reference genome for *Ancylostoma caninum*	Xie et al.[39]	GenBank accession ID MN215971
Reference genomes for nematodes that parasitize animals	WormBase ParaSite[16,17]	https://parasite.wormbase.org/species.html#Nematoda
Software and algorithms		
MinKNOW v. 22.10.7; 22.10.10; 22.12.7	Oxford Nanopore Technologies	https://community.nanoporetech.com/docs/prepare/library_prep_protocols/experiment-companion-minknow/v/mke_1013_v1_revdc_11apr2016/installing-minknow-on-linu
Guppy v. 6.3.4	Oxford Nanopore Technologies	https://community.nanoporetech.com/downloads?from=support
NanoPlot v. 1.40.2	De Coster[40]	https://github.com/wdecoster/NanoPlot
Mash v. 2.2.2	Ondov et al.[41]	https://mash.readthedocs.io/en/latest/#
Minimap2 v. 2.16	Li[42]	https://github.com/lh3/minimap2
SAMtools v. 1.9	Danecek et al.[43]	https://github.com/samtools/samtools
fastp v. 0.23.2	Chen et al.[44]	https://github.com/OpenGene/fastp
BWA v. 0.7.17	Li and Durbin[45]	https://github.com/lh3/bwa
Jellyfish v. 2.3.0	Marçais and Kingsford[46]	https://github.com/gmarcais/Jellyfish
GenomeScope web-based graphical user interface	Ranallo-Benavidez et al.[44]	https://github.com/schatzlab/genomescope
Canu v. 2.1	Koren et al.[36]	https://github.com/marbl/canu
wtdbg2 v. 0.0	Ruan and Li[47]	https://github.com/ruanjue/wtdbg2
Shasta 0.11.1	Shafin et al.[48]	https://github.com/chanzuckerberg/shasta
Flye v. 2.9.1-b1780	Kolmogorov et al.[49]	https://github.com/mikolmogorov/Flye
SeqKit v 0.10.1	Shen et al.[50]	https://bioinf.shenwei.me/seqkit/
BBMap v.38.84	Bushnell[51]	https://github.com/BioInfoTools/BBMap
purge_dups v. 1.2.5	Guan et al.[52]	https://github.com/dfguan/purge_dups
Racon v. 1.5.0	Vaser et al.[53]	https://github.com/isovic/racon
Medaka v. 1.7.2	Oxford Nanopore Technologies	https://github.com/nanoporetech/medaka
Pilon v. 1.24	Walker et al.[54]	https://github.com/broadinstitute/pilon
MaSuRCA v. 4.1.0	Zimin et al.[37]	https://github.com/alekseyzimin/masurca
WENGAN v. 0.2	Di Genova et al.[55]	https://github.com/adigenova/wengan
HASLR v. 0.8a1	Haghshenas et al.[56]	https://github.com/vpc-ccg/haslr
K-mer Analysis Toolkit (KAT) v. 2.4.0	Mapleson et al.[57]	https://github.com/TGAC/KAT
BlobTools v. 1.1.1	Challis et al.[58]	https://github.com/DRL/blobtools
BLAST v. 2.14.0	Altschul et al.[59]	https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html#downloadblastdata
ggplot2 R package	Wickham[60]	https://ggplot2-book.org/
R v. 4.3.0; 4.2.3	Core R Team[61]	https://cran.rstudio.com/
RStudio v. 2023.03.1; 2023.06.1	N/A	https://posit.co/download/rstudio-desktop/
Geneious Prime v. 2022.0.1	Kearse et al.[62]	https://www.geneious.com/updates
QUAST v. 5.0.2	Gurevich et al.[63]	https://github.com/ablab/quast
compleasm v. 0.2.6	Huang and Li[64]	https://github.com/huangnengCSU/compleasm

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GeMoMA v. 1.9	Keilwagen et al.[65,66]	https://www.jstacs.de/index.php/GeMoMa
AGAT v. 0.9.1	Dainat et al.[67]	https://github.com/NBISweden/AGAT
Biostrings R package v. 2.66.0	Pagès et al.[68]	https://bioconductor.org/packages/release/bioc/html/Biostrings.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Joseph R. Fauver (jfauver@unmc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Quality-controlled MinION and Illumina data for each species are deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under the BioProject accession ID PRJNA1074771 and under BioSample accession nos. SAMN39888962 (*Brugia malayi*), SAMN39888963 (*Trichuris trichiura*) and SAMN39888964 (*Ancylostoma caninum*), and are publicly available as of the date of publication. Final assemblies for each species are deposited in the NCBI Genome database under the accession numbers GCA_037903335.1, GCA_037903345.1, and GCA_037903365.1 (*Brugia malayi*), GCA_037903355.1, GCA_037903375.1, and GCA_037903465.1 (*Trichuris trichiura*), and GCA_037903475.1, GCA_037903715.1, and GCA_037903745.1 (*Ancylostoma caninum*), and are publicly available as of the date of publication.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Specimen acquisition

The specimens sequenced were obtained from the same strains previously used to generate existing reference genomes for each species. Adults of *B. malayi* (FR3 strain) were acquired through the NIH/NIAID Filariasis Research Reagent Resource Center via BEI Resources.[28] Worms were received frozen and stored at -80°C. Prior to sequencing, worms were thawed at 4°C and subsequently preserved in DNA/RNA Shield (Zymo Research, Irvine, CA, USA). Adults of *T. trichiura* were obtained from a Danish patient after anthelmintic treatment for an infection acquired in Uganda. Worms were received preserved in 70% ethanol and were stored at 4°C prior to sequencing. Pooled third-stage (L3) larvae of *A. caninum* (Baltimore strain) were obtained from experimental infection in laboratory-maintained canines. Pooled L3s were received frozen, stored at -80°C, and thawed at 4°C prior to sequencing.

DNA extraction

Total nucleic acid was extracted separately from each of a single adult female (*B. malayi*), a single adult male (*T. trichiura*), and a pool of ~10,000 L3 larvae (*A. caninum*) using a Quick-DNA HMW MagBead Kit (Zymo Research, Irvine, CA, USA) and modified Solid Tissue extraction protocol (Table S1). Prior to extraction, adult worms were placed in clean 1.5 mL DNA LoBind® tubes (Eppendorf® North America, Enfield, CT, USA) using sterilized forceps, while pooled L3s were centrifuged at 6,540 g for 3 min to allow for the removal of excess supernatant. All samples were then homogenized mechanically using a Monarch® Pestle Set single-use microtube pestle (New England Biolabs® Inc., Ipswich, MA, USA). Homogenization was performed both prior to, and immediately following, the addition of Proteinase K. Samples were then digested in a digital dry bath at 55°C for 24 hr with occasional flick mixing. The protocol for DNA purification was then followed, with these modifications: After the addition of 33 μL of MagBinding Beads, samples were placed on a benchtop rotating mixer for 40 min to 1.5 hr at RT; and after the addition of 52 μL DNA Elution Buffer, samples were incubated in a digital dry bath at 37°C for 2 hr with occasional flick mixing, then incubated at RT overnight prior to collecting eluted DNA. To avoid DNA fragmentation, samples were mixed by flicking rather than pipette mixing wherever possible. The concentration of each extraction was measured using a Qubit™ 4 1X dsDNA High Sensitivity (HS) Assay (ThermoFisher Scientific, Waltham, MA, USA) and fragment size distribution was determined using an Agilent 2200 TapeStation System and associated protocol for genomic DNA ScreenTape analysis (Agilent, Santa Clara, CA, USA). For *T. trichiura*, TapeStation analysis was not performed.

For *A. caninum*, initial attempts to sequence ONT libraries resulted in low sequencing efficiency, poor data quality, and rapid pore loss. To ameliorate concerns of protein contamination, extracted genomic DNA (gDNA) was purified via bead-based cleanup prior to final ONT library

preparation. This cleanup was performed as follows: Two aliquots of extracted gDNA were each bound to a 0.5× volume of AMPure XP beads (Beckman Coulter, Brea, CA, USA) on a benchtop rotator mixer for 1 hr at RT. Beads were then washed twice with 80% EtOH, air-dried, and gDNA was eluted in 52 μL Zymo DNA Elution Buffer for 2 hr at 37°C with occasional flick mixing, followed by overnight elution at RT. This cleanup process was repeated a second time for both aliquots. Purified extractions were then quantified using a Qubit™ 4 1X dsDNA HS Assay and the Agilent 2200 TapeStation System.

MinION library preparation

Aliquots of gDNA extracted for each species were used to prepare one or more libraries for whole genome sequencing on an ONT MinION desktop sequencer (see Table S1). Libraries were prepared using a combination of the DNA Clean & Concentrator MagBead Kit (Zymo Research, Irvine, CA, USA), the ONT SQK-LSK114 Ligation Sequencing Kit (Oxford Nanopore Technologies, Oxford, United Kingdom), the NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing (New England Biolabs® Inc., Ipswich, MA, USA), and a modified hybrid protocol. First, gDNA was incubated with ONT DNA repair and end preparation reagents for 10 min at RT followed by 10 min at 65°C. End preparation reactions were then bound to 20 μL Zymo MagBinding Beads in 4× volumes of Zymo DNA MagBinding Buffer on a benchtop rotating mixer for 30 min to 2 hr. Beads were then washed twice with Zymo DNA Wash Buffer, air-dried for 10 min, eluted in 51 μL Zymo DNA Elution Buffer via manual agitation for 10 min, and quantified via Qubit™ 1X dsDNA HS Assay. Sequencing adaptors were then ligated via a 15 min incubation at RT, and libraries were bound to a 0.4× volumes of AMPure XP beads on a benchtop rotator mixer at RT for 1 hr. Beads were then washed twice with either ONT Long Fragment Buffer (*A. caninum*) or a titrated wash mix of 1:3 ONT Short Fragment Buffer:ONT Long Fragment Buffer (*B. malayi* and *T. trichiura*). After washing, beads were air-dried, then libraries were eluted in 15–17 μL of ONT Elution Buffer for 2 hr at 37°C with occasional flick mixing followed by overnight elution at RT. Final libraries were quantified using a Qubit™ 4 1X dsDNA HS Assay and the Agilent 2200 TapeStation System. For *T. trichiura*, TapeStation analysis was not performed. Additionally, for *T. trichiura*, a portion of ONT library reserved after sequencing on one flow cell was re-washed prior to sequencing on a second flow cell (see Table S1). This aliquot of library was bound to a 0.4× volume of AMPure XP beads on a benchtop rotator mixer for 1 hr, washed twice with ONT Long Fragment Buffer, air-dried, then eluted in 17 μL of ONT Elution Buffer for 2 hr at 37°C with occasional flick mixing followed by overnight elution at RT.

MinION sequencing and basecalling

Portions of ONT libraries were sequenced for ~62–80 hr each on one (*B. malayi*), two (*T. trichiura*), or three (*A. caninum*) ONT MinION R10.4.1 flow cells. The amount of library loaded onto each flow cell and the total number of pores available at the start of sequencing are provided in Table S1. MinKNOW software (ONT) v. 22.10.7 (*B. malayi*), v. 22.10.10 (*T. trichiura*), or v. 22.12.7 (*A. caninum*) was used to run each flow cell with pore scans every 1.5 hr. After sequencing, signal data (i.e., fast5 files) from each flow cell were basecalled using Guppy v. 6.3.4 (ONT). The bioinformatic pipeline used to generate MinION data assemblies is outlined in Figure 1. Sequencing adaptors were simultaneously removed by specifying the "–trim_adapters" flag. The results of each Guppy run were summarized using NanoPlot v. 1.40.2.[40] Fastq files that passed basecalling were input to Mash v. 2.2.2[41] to confirm there was no significant contamination in the read-level data prior to assembly. Estimated genome sizes provided to Mash were based on the lengths of the existing reference assemblies available for each species, and were thus set as 88 Mb for *B. malayi*,[29] 465 Mb for *A. caninum*[7] and 75 Mb for *T. trichiura*[23] (https://github.com/stephenrdoyle/ancient_trichuris/tree/master/02_data). Average depth of coverage of reads across the species-specific reference genome and proportion of reads mapped were estimated using Minimap2 v. 2.16[42] and SAMtools v. 1.9.[43]

Illumina library preparation and sequencing

Aliquots of gDNA extracted for each species were sent to the Yale Center for Genome Analysis (YCGA) for Illumina whole genome library preparation and sequencing (see Table S1). Libraries were prepared using an xGen™ cfDNA & FFPE DNA Library Prep Kit with unique dual indexing (Integrated DNA Technologies, Newark, NJ, USA) and standard protocol. Libraries were subjected to 5–7 cycles of PCR to increase concentration (see Table S1) and multiplexed for 2×150 paired-end sequencing on an Illumina NovaSeq 6000 targeting 50× depth of coverage genome-wide for each species. Reads were demultiplexed by the YCGA and subsequently filtered for remaining adaptor contamination, quality, and length using fastp v. 0.23.2.[44] The bioinformatic pipeline used to process Illumina data and generate hybrid assemblies is outlined in Figure 1. Fastq files that passed fastp were input to Mash to confirm there was no significant contamination in the read-level data prior to assembly. Estimated genome sizes provided to Mash were based on the lengths of the existing reference assemblies available for each species (see above). Average depth of coverage across the species-specific reference genome and proportion of reads mapped were estimated using BWA v. 0.7.17[45] and SAMtools v. 1.9. A second round of fastp was run on the raw demultiplexed Illumina data to remove sequencing adapters, only, by specifying "–detect_adapter_for_pe", "–disable_length_filtering", and "–disable_quality_filtering". These data were the used for hybrid genome assembly, as MaSuRCA utilizes built-in error correction and cleaning.

Estimation of genome size and heterozygosity

Quality-controlled Ilumina data were used to estimate genome size and heterozygosity for each species. First, Jellyfish v. 2.3.0[46] was used to generate k-mer spectra, specifying a k-mer size of 21 and an initial hash size of 1,000,000,000. Histograms from Jellyfish were then provided to the GenomeScope web-based graphical user interface[69] to visualize k-mer spectra and estimate heterozygosity.

MinION data genome assembly

FASTQ files that passed basecalling by Guppy were concatenated and used as input for whole genome assembly. A total of four long-read only assembly algorithms were tested using the MinION data for *B. malayi* to determine which approach provided the most contiguous and complete genome assembly, including Canu v. 2.1,[36] wtdbg2 v. 0.0,[47] Shasta v. 0.11.1,[48] and Flye v. 2.9.1-b1780.[49] Assembly metrics included genome size, N50, contig count, GC content, and BUSCO scores. Following comparisons, we determined that Canu provided the highest quality genome assemblies from our data (see Table S2) and it was used to generate MinION data assemblies for all three species sequenced. For *T. trichiura* and *A. caninum*, a 3 kb minimum read length requirement was specified to Canu, and for *B. malayi*, a 5 kb minimum read length requirement was specified. Estimated genome sizes provided to Canu were based on the lengths of the existing reference assemblies available for each species (see above). Contigs in the resulting Canu assemblies that were indicated as potential alternative alleles (i.e., with FASTA headers including "suggestBubble=yes") were removed using a combination of SeqKit v 0.10.1[50] and the filterbyname.sh script of BBMap v.38.84.[51] Resulting "popped" assemblies were used as input to purge_dups v. 1.2.5[52] to remove false duplications. These "popped and purged" assemblies were then polished with one round of Racon v. 1.5.0[53] followed by one round of Medaka v. 1.7.2 (https://github.com/nanoporetech/medaka) to generate "popped, purged, and polished" assemblies. Copies of these final assemblies were also separately polished with Illumina data using Pilon v. 1.24[54] and BWA. Three iterations of Pilon were iteratively run specifying the "–diploid" flag.

Hybrid genome assembly

A total of three hybrid genome assembly algorithms were tested using sequence data for *B. malayi* to determine which approach provided the most contiguous and complete assembly, including MaSuRCA v. 4.1.0,[37] WENGAN v. 0.2,[55] and HASLR v. 0.8a1.[56] Assembly metrics included genome size, N50, contig count, GC content, and BUSCO scores. Following comparisons, we determined that MaSuRCA provided the highest quality hybrid genome assemblies from our data (see Table S2) and it was used to generate hybrid assemblies for all three species sequenced. FASTQ files that passed basecalling by Guppy and Illumina data from which only sequencing adaptors were removed by fastp were used as input to MaSuRCA. Default settings were used except for setting "LHE_COVERAGE" to "35" and "cgwErrorRate=0.15", and setting "JF_SIZE" to a value ten times the genome size of the species-specific reference assembly. Assemblies output by MaSuRCA were input to purge_dups v. 1.2.544 to remove false duplications.

Refinement of final assemblies

To ensure "purged" MaSuRCA hybrid assemblies were not over- or under-purged, copy number spectrum plots were generated using the K-mer Analysis Toolkit (KAT) v. 2.4.0[57] specifying a k-mer size of 31. "Popped, purged, and polished" Canu MinION data assemblies and "purged" MaSuRCA hybrid assemblies were input to BlobTools v. 1.1.1[58] to identify potential contaminants. First, the reads used to generate each assembly were mapped to the assembly using Minimap2 (MinION data) and/or BWA (Illumina data), and SAMtools was used to generate sorted BAM files for each mapping. The "blastn" function in BLAST v. 2.14.0[59] was then used to compare all contigs to reference sequences available in the National Center for Biotechnology Information (NCBI)'s nucleotide databases, specifying "-evalue 1e-25", "-max_target_seqs 10" and "-max_hsps 1". BLAST outputs and sorted BAM files were then used as input to BlobTools. Contigs that could be confirmed as host, human, or bacterial contamination were excluded from final assemblies. Quality scores (i.e., Q scores) of basecalled MinION reads, and Q scores of reads when mapped to the final "popped, purged, and polished" MinION data assemblies, were measured using NanoPlot and plotted using the "geom_density" function in the ggplot2 package[60] in R v. 4.3.0[61] in RStudio v. 2023.03.1.

Mitogenome and Wolbachia genome assembly

For *B. malayi* and *A. caninum*, contigs containing the mitogenomes—and, for *B. malayi*, the genome of its *Wolbachia* endosymbiont—were extracted from "popped" MinION and unpurged MaSuRCA assemblies. For *T. trichiura*, mitogenomes were extracted from the original "unpopped" Canu assembly and the unpurged MaSuRCA assembly. These organelle genome contigs were then mapped to species-specific reference organelle genomes using Geneious Prime v. 2022.0.1[62] and manually reconfigured to match the linear orientation of the reference. For *B. malayi* and *T. trichiura*, reference organelle genomes were extracted from the reference genome assemblies for each species (see above). For *A. caninum*, for which a mitochondrial contig is not labeled in the reference assembly, a reference mitogenome was downloaded from GenBank (accession no. MN215971).[39] Contigs containing complete organelle genomes were then manually added to the appropriate final version of each assembly.

Assembly contiguity and completeness

Contiguity and completeness were assessed for each assembly using QUAST v. 5.0.2[63] and compleasm v. 0.2.6,[64] respectively. For compleasm, the Nematoda, Metazoa, and Eukaryota ortholog databases (nematoda_odb10, metazoa_odb10, and eukaryota_odb10, respectively) were chosen for comparison. Compleasm was also run on the aforementioned reference genomes for each species, as well as for all genomes of nematodes that parasitize animals as adults available from WormBase ParaSite.[16,17] For each assembly, the proportion of single copy BUSCOs present was plotted against the log value of the assembly's N50 length divided by its scaffold or contig count using R v. 4.2.3 via RStudio v. 2023.06.1. Cutoff values of >85% (for proportion of single copy BUSCOs present) and >1.6 (for the log value of the contiguity metric) were used to assess whether each assembly could be classified as a "tier 1" genome following the standard for helminth genomes established by the International Helminth Genomes Consortium.[7]

Gene content and composition

Homology-based gene prediction was performed for each assembly generated here using GeMoMA v. 1.9.[65,66] Predictions were made using default settings and the species-specific reference genome and annotation as input. Resulting GFF files were summarized using AGAT v. 0.9.1.[67] For each species, the predicted genes shared between the MinION data assembly and the MinION assembly polished with Illumina data were pairwise aligned using the "pairwiseAlignment" function in Biostrings package v. 2.66.0[68] in R v. 4.2.3 via RStudio v. 2023.06.1 specifying the alignment type as "global". Biostrings was then used to characterize each pairwise alignment in terms of percent identity, alignment length, gene lengths, and number of single nucleotide polymorphisms (i.e., SNPS) and insertions/deletions (i.e., indels). Summary figures for all comparisons were generated in R v. 4.3.0 via RStudio v. 2023.03.1 using the ggplot2 package.