PLOS NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

Long-read RNA sequencing of human and animal filarial parasites improves gene models and discovers operons

Nicolas J Wheeler, Paul M. Airs, Mostafa Zamanian*

Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

* mzamanian@wisc.edu

Citation: Wheeler NJ, Airs PM, Zamanian M (2020) Long-read RNA sequencing of human and animal filarial parasites improves gene models and discovers operons. PLoS Negl Trop Dis 14(11): e0008869. https://doi.org/10.1371/journal.pntd.0008869

Editor: Krystyna Cwiklinski, National University of Ireland Galway, IRELAND

Received: June 22, 2020

Accepted: October 9, 2020

Published: November 16, 2020

Data Availability Statement: All data analysis and visualization scripts are publicly available at https://github.com/zamanianlab/Filarid_IsoSeq-ms. The script for poly(A) analysis can be found at https://github.com/zamanianlab/polyAudit. Long-read sequencing data has been deposited into NIH BioProjects PRJNA548902 (B. malayi) and PRJNA640410 (D. immitis).

Funding: Funding for MZ is provided by an R01 grant from the National Institute of Allergy and Infectious Diseases (R01AI151171, NIH.gov). The

Abstract

Filarial parasitic nematodes (Filarioidea) cause substantial disease burden to humans and animals around the world. Recently there has been a coordinated global effort to generate, annotate, and curate genomic data from nematode species of medical and veterinary importance. This has resulted in two chromosome-level assemblies (*Brugia malayi* and *Onchocerca volvulus*) and 11 additional draft genomes from Filarioidea. These reference assemblies facilitate comparative genomics to explore basic helminth biology and prioritize new drug and vaccine targets. While the continual improvement of genome contiguity and completeness advances these goals, experimental functional annotation of genes is often hindered by poor gene models. Short-read RNA sequencing data and expressed sequence tags, in cooperation with *ab initio* prediction algorithms, are employed for gene prediction, but these can result in missing clade-specific genes, fragmented models, imperfect mapping of gene ends, and lack of isoform resolution. Long-read RNA sequencing can overcome these drawbacks and greatly improve gene model quality. Here, we present Iso-Seq data for *B. malayi* and *Dirofilaria immitis*, etiological agents of lymphatic filariasis and canine heartworm disease, respectively. These data cover approximately half of the known coding genomes and substantially improve gene models by extending untranslated regions, cataloging novel splice junctions from novel isoforms, and correcting mispredicted junctions. Furthermore, we validated computationally predicted operons, manually curated new operons, and merged fragmented gene models. We carried out analyses of poly(A) tails in both species, leading to the identification of non-canonical poly(A) signals. Finally, we prioritized and assessed known and putative anthelmintic targets, correcting or validating gene models for molecular cloning and target-based anthelmintic screening efforts. Overall, these data significantly improve the catalog of gene models for two important parasites, and they demonstrate how long-read RNA sequencing should be prioritized for ongoing improvement of parasitic nematode genome assemblies.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Filarial parasitic nematodes are vector-borne parasites that infect humans and animals. *Brugia malayi* and *Dirofilaria immitis* are transmitted by mosquitoes and cause human lymphatic filariasis and canine heartworm disease, respectively. Recent years have seen a dramatic increase in genomic and transcriptomic data sets and the concomitant increase in innovative strategies for drug target identification, validation, and screening. However, while the completeness of genome assemblies of filarial parasitic nematodes has seen steady improvements, the reliability of gene models has not kept pace, hindering cloning efforts. Long-read RNA sequencing technologies are uniquely able to improve gene models, but have not been widely used for the causative agents of neglected tropical diseases. Here, we report the improvement of gene models in both *B. malayi* and *D. immitis* by long-read RNA sequencing. We identified novel operons, deprecated false positive operons, identified dozens of novel genes, and described the parameters of polyadenylation. We also focused on putative anthelmintic targets, identifying novel isoforms and correcting gene models. These data substantially increase the trustworthiness of gene models in these two species and demonstrate how long-read sequencing approaches should be prioritized in the continued improvement of genome assemblies and their gene annotations.

Introduction

Infections caused by filarial parasitic nematodes inflict massive burdens upon humans and animals around the world. The mosquito-borne filarial nematode *Brugia malayi* is an etiological agent of the neglected tropical disease lymphatic filariasis (LF), which affects over 60 million people worldwide [1]. *Dirofilaria immitis*, a related mosquito-borne filarial nematode, is a causative agent of dog heartworm disease and a zoonotic concern [2]. These diseases are ultimately controlled by different strategies for preventive chemotherapy. In tropical regions endemic for LF, the World Health Organization has recommended mass drug administration (MDA) of two annual doses of albendazole (ABZ) or a combination of ABZ, ivermectin (IVM), and/or diethylcarbamazine citrate (DEC) [3]. For control of dog heartworm, the American Heartworm Society recommends monthly administration of macrocyclic lactones like IVM [4].

Preventive chemotherapies are active against parasite larvae; no adulticidal therapies are approved for the treatment of LF, and only the arsenic-containing drug melarsomine is available for clearing adult heartworm. Further, while the triple-drug combination is a new development for LF control, the absence of novel filaricidal chemical moieties combined with the massive drug pressure caused by preventive chemotherapy at a global scale continues to stoke fear of loss of drug efficacy for both the human and veterinary diseases [5–7]. Indeed, this is already an emerging problem for heartworm control [8,9].

LF and heartworm control and elimination strategies would be significantly helped by the development of new antifilarial drugs and therapies. New data and technologies have promised to speed the discovery and validation of novel therapeutic targets. The release of the *B. malayi* draft genome initiated a wave of target-based approaches that leverage genomic data, molecular biology, and rational target selection, which allow for the cloning, expression, and screening of putative targets in recombinant systems or exploring the repurposing of approved drugs with known targets [10–13]. Intrinsic to this process is the assumption that the desired transcript and isoform of a given target is known and amenable to straightforward cloning.

However, while the completeness of the *B. malayi* and *D. immitis* draft genomes is sufficient for target prioritization and selection via comparative genomics, the fragmentation and incompleteness of predicted gene models can delay facile cloning, expression, and characterization or screening (see [10], for instance). Despite the excellent chromosome-level assembly of the *B. malayi* genome [14], two of three cloned genes in a recent study [15] had incorrectly predicted splice junctions even though these genes share high similarity with their one-to-one *C. elegans* homologs. Gene prediction and curation in *Caenorhabditis* spp. is known to be a difficult task, and it is clear that is also the case in filarial nematodes [16].

In addition to incorrect gene models, both *B. malayi* and *D. immitis* have very few genes with predicted isoforms. *C. elegans* (WS277) has 1.97 predicted transcripts per gene, while *B. malayi* (WS277) has 1.44 and *D. immitis* (WBPS14) does not have a single predicted isoform. It is increasingly clear that anthelmintics can have selectivity for specific target isoforms. IVM exhibits strong isoform-specific action against *C. elegans* AVR-14 ion channel subunits expressed in *Xenopus* [17], and this has been validated in a range of parasitic nematodes, including *D. immitis* [18–20]. More recently, emodepside was shown to cause sex-specific effects in adult *B. malayi* due to alternative splicing of SLO-1, one of the drug's targets [21]. Thus, identifying transcript isoforms should be a high priority for the continual curation of the genomes of parasitic nematodes.

A number of techniques are available to generate datasets that aid supervised gene prediction algorithms. Cap analysis gene expression (CAGE) has been used to characterize the landscape of transcriptional start sites (TSS) in a given species or tissue [22]. However, many nematode pre-mRNAs lose their original 5' end through *trans*-splicing to splice-leader (SL) sequences (~70% in *B. malayi* [23]), obscuring the true TSS and making the CAGE technique unsuitable. In *C. elegans*, a modified technique has been used that leveraged the ability to rear worms at lower temperatures to slow the processing of *trans*-spliced transcripts [24], but most parasitic nematodes are not amenable to these modifications in animal husbandry. Additionally, the integration of ever-increasing short-read RNA sequencing data provides support for exon usage, but it is difficult to generate abundant high-quality mRNA across many parasitic nematode life stages. This difficulty has resulted in databases with a strong bias in reads mapping to the 3' end of transcripts, leading to errors on the 5' ends. Even where RNA quality and yield are not limiting, alternative splicing and isoform prediction are not straightforward with exclusively short-read datasets.

Long-read RNA sequencing can overcome many of the limitations inherent to short-read sequencing and can substantially improve gene predictions and annotations in genome assemblies (Fig 1). Only two long-read RNA sequencing datasets exist for parasitic nematodes, the zoonotic hookworm *Ancylostoma ceylanicum* and the barber's pole worm *Haemonchus contortus* [25,26]. Here, we report long-read RNA sequencing of the filarial parasites *B. malayi* and *D. immitis*. We separately sequenced both sexes of these species, allowing for the analysis of sex-specific isoform usage. We show a substantial increase in the lengths of predicted untranslated regions, identify operons, correct gene models, and catalog the parameters of poly(A) tail usage. This resource will significantly improve the confidence in gene predictions in these organisms and enhance the utility of the reference genomes.

Methods

Parasite source

B. malayi (strain FR3) adult males and females were sent to the University of Wisconsin-Madison from the Filariasis Reagent Resource Center (FR3)[27], washed, and incubated at 37˚C in fresh complete media (RPMI 1640, 10% FBS, gentamicin, and penicillin/streptomycin)

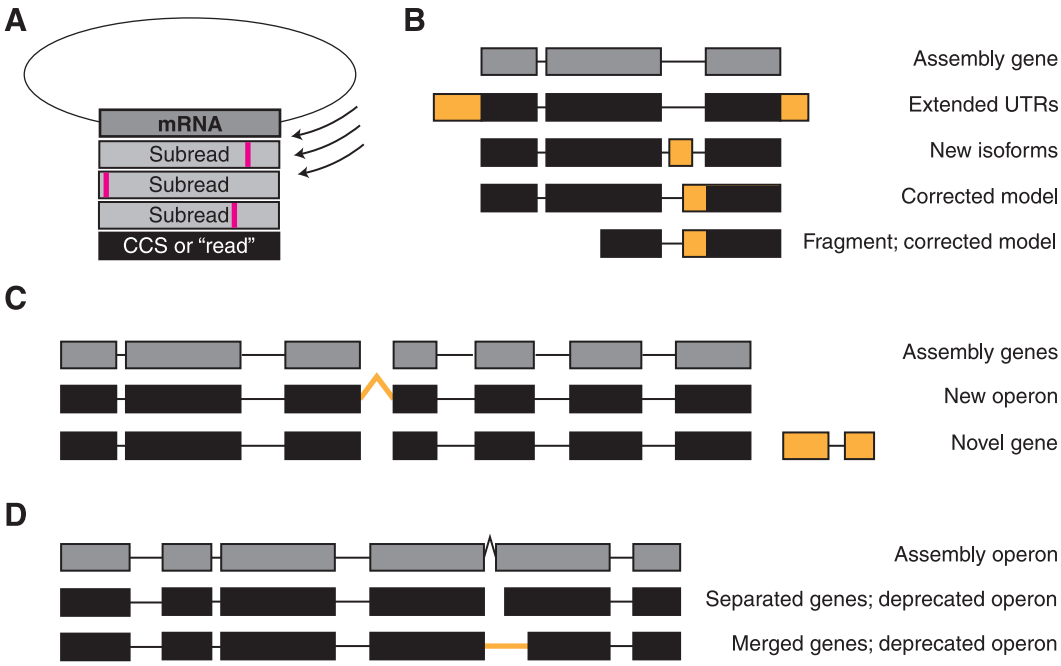

Fig 1. Improvements in gene prediction and annotation by Iso-Seq. (A) Iso-Seq libraries are prepared by isolating and ligating mRNA into a circular backbone. The polymerase can circle the template multiple times, creating multiple "subreads" per a given template. Subreads are often error-prone, but errors are corrected by alignment and consensus calling, resulting in a circular consensus sequence (CCS, termed "read" here). (B) Reads aligning to individual genes can validate the reference model and extend UTRs, identify new isoforms by discovering novel splice sites and exons, or correct the assembly model. Notably, models can be corrected even by reads arising from fragmented mRNA with 5' degradation. (C) Reads aligning to multiple genes can indicate a new operon, and reads mapping to intergenic sequences can identify novel genes. (D) Reads aligning to an assembly operon may direct the deprecation of the operon by separating the genes or by merging the genes into a single model.

https://doi.org/10.1371/journal.pntd.0008869.g001

overnight prior to RNA extraction. *D. immitis* (isolate ZoeLA) adult males and females were received at the University of Wisconsin-Madison from Zoetis, washed, and incubated at 39˚C in fresh complete media (RPMI 1640, 20 mM HEPES, 1% D-glucose, 1% HCO_3, 10% FBS, and penicillin/streptomycin). After overnight incubation, worms were individually dipped in PBS, transferred to an empty 50 mL conical tube, and flash-frozen in liquid N_2. Frozen worms were sent to Smith College for RNA extraction.

RNA extraction

15 *B. malayi* worms from each sex were combined and transferred to 750 μL TRIzol LS with 250 μL nuclease-free water. Samples were incubated at room temperature for 1 hour prior to homogenization in a TissueLyser II for two cycles of 3 minutes at 30 Hz separated by a 2 minute incubation on ice. RNA was extracted according to the manufacturer's recommendation, DNase treated, and re-purified with phenol:chloroform. RNA pellets were resuspended in nuclease-free water and assessed for quality and concentration with a Bioanalyzer 2100 (Agilent, Santa Clara, CA) RNA Pico chip. RNA was extracted from a single *D. immitis* worm from each sex using the PureLink RNA Mini kit (Thermo Fisher Scientific, Waltham, MA). RNA was DNase treated and subsequently dialyzed and assessed for quality and concentration with a Bioanalyzer 2100 (Agilent) RNA Nano chip. Extraction using these protocols provided RNA with RNA integrity (RIN) values between 6 and 8.

SMRTbell library preparation, quality control, and sequencing

cDNA was synthesized from parasite RNA using the SMARTer PCR cDNA synthesis kit (Takara, Mountain View, CA), which incorporates a poly(A) selection step. 1 μg of RNA input was used and replicate cDNA reactions were performed where excess RNA was available. PCR cycles were optimized, and 12 cycles were used for all 4 samples. PCR reactions were pooled at 1X and 0.5X for each library, cleaned with 1X and 0.5X AMPure XP beads, and quantified on a 2100 Bioanalyzer (Agilent) with an HS DNA chip. Equimolar fractions of 1X and 0.5X cDNA were pooled with a target of >2 μg of each fraction, and libraries were created with the SMARTbell Template Library kit (Pacific Biosciences, Menlo Park, CA). The final library quality and concentration were assessed using a Qubit DNA HS kit (Thermo Fisher Scientific) and a 2100 Bioanalyzer (Agilent) HS DNA chip. A single SMRT cell was used for each library and sequenced on the Sequel system (PacBio) with diffusion loading and a 20 hour movie.

Sequencing quality control and transcriptome analysis

Circular consensus sequences (hereafter, "reads") were generated from subreads with IsoSeq v3.1.0 using the genome assemblies from WormBase ParaSite (WBPS14) [14,28,29]. Transcripts were filtered with SQANTI3, which reference-corrects all transcripts and filters them based on hallmarks of artifacts and SVM classifiers [30]. Sequencing summary statistics were generated with SAMtools [31], and read depth was calculated with BEDtools [32,33].

Identification of novel genes

We filtered reads to maintain those categorized as antisense (aligning to the opposite strand of a locus with a predicted gene) and intergenic (aligning to a genomic locus without a predicted gene). These reads were used as queries for blastx [34] searches against the respective predicted proteomes [14,29]. Reads with blast hits with e-values > 1e-3 (i.e., no significant hits to a predicted protein) were retained and used as queries for blastx searches against the predicted proteomes of *Brugia pahangi* [35,36], *Caenorhabditis elegans* [37], *Ascaris suum* [38], *Onchocerca volvulus* [39], *Strongyloides ratti* [40], and *Toxocara canis* [41]. Reads with blast hits with e-values < 1e-3 (i.e., significant hits to predicted proteins from other nematodes) were retained and finally used as queries in a blastn search against the *B. malayi* genomic scaffolds to determine the number of unique loci mapped.

Analysis of poly(A) tails

We developed a single script to analyze poly(A) tails in untrimmed reads (https://github.com/zamanianlab/polyAudit). Analyses include (1) measurement of poly(A) tail length, (2) discovery and tally of k-mers in the region upstream of the 3' cleavage site, (3) identification of the most likely poly(A) signal (PAS), (4) location of the PAS, and (5) generation of a position-specific score matrix (PSSM) to calculate nucleotide frequencies at every position in the region flanking the 3' cleavage site. The tool is open-source and utilizes open-source Python libraries.

Poly(A) tails are measured by counting A nucleotides until reaching a non-A dinucleotide. To identify k-mers enriched in 3' UTRs, k-mers from the translational stop site to the 3' cleavage site were tallied and compared to k-mer frequencies across all non-coding intergenic sequences from the relevant genome; this comparison resulted in a list of k-mers enriched in 3' UTRs. The most likely PAS was identified by iteratively searching the upstream region for the k-mers from this enriched list, sorted by frequency; the search was terminated when a

k-mer was found and the location of the PAS relative to the cleavage site was recorded. Finally, to analyze nucleotide frequencies in the region immediately upstream of the cleavage site for a given PAS, the poly(A) tail was trimmed from each sequence, reads were pooled by PAS, and the 50 bp upstream of the cleavage site were aligned. The PSSM function of BioPython [42] was implemented to calculate nucleotide frequencies at each position of the multiple sequence alignment.

Operon identification

The GTF file for *B. malayi* was downloaded from WormBase ParaSite (WBPS14) [28], converted to BED format with the gtf2bed script from BEDOPS v2.4.37 [43], and "gene" features were retained. Reads were aligned to reference genomes with minimap2 [44] and BAM files were converted to BED format with bamtobed from BEDTOOLS with the -split flag [32,33]. BEDTOOLS "intersect" with the -wa and -wb flags were used to retain exonic regions to which reads had been mapped. BED files were imported into R for further analysis.

Any read that mapped to more than one gene was kept as a potential polycistronic read. The distance between cistrons was calculated by taking the difference between the reference 3' end of the 5' coding sequence (CDS) and the reference 5' end of the 3' CDS; thus, incorrect reference annotations of the gene ends inevitably caused some errors in the distance calculation. We kept all putative operons that had a maximum intercistronic distance of $<$5000 nt, as the vast majority of assembly-annotated operons had intergenic distances below this threshold. We manually inspected each putative operon and made true/false decisions using a variety of evidence, including 1) the number of ORFs in reads spanning the genes (calculated using the ExPASy translate server [45]), 2) the sequence similarity of ORFs within reads to nematode orthologs (calculated using blastx on WormBase [46] or WormBase ParaSite [28]), the 3) differences in lengths between these orthologs, 4) and public short-read RNA-seq data viewed on Jbrowse at WormBase [46,47]. GO enrichment of curated operons was performed using the R package topGO [48].

Analysis of operon expression

Short-read RNA sequencing data from across the *B. malayi* life cycle [49] were acquired from NCBI SRA; reads were remapped to Bmal-4.0 and counts were generated with HISAT2 [50] and StringTie2 [51]. The RNA-seq pipeline was implemented using Nextflow [52] and is publicly available (https://github.com/zamanianlab/BmalayiRNAseq-nf).

Only the first two cistrons of each operon were kept, and pseudo-operons were created by randomly selecting neighboring genes with intergenic distances $<$5000 nt. Expression abundances for each gene pair—putative operons and pseudo-operons—during each life cycle stage (26 RNA-seq data sets, S1 Table [53]) were fit using a linear model, and R^2 values were calculated.

Drug target analysis

We generated a curated list of genes that are known anthelmintic targets and receptors belonging to druggable protein families in parasitic nematodes, including selected ligand-gated ion channels (LGICs), voltage-gated ion channels, CNG channels, TRP channels, GPCRs, and beta-tubulins. Reads overlapping target loci were extracted with BEDTools [33] and manually assessed with IGV [54] and R.

Genome curation

Genome curation was performed in collaboration with WormBase and WormBase ParaSite staff. An Apollo [55] instance was created and hosted by WormBase ParaSite. Iso-Seq data along with public RNA-seq datasets were used as evidence for manual curation of operons, novel genes, annotated genes, and the merging of spurious operons into single gene models.

Results

Summary of sequencing results

We extracted high-quality RNA from *B. malayi* and *D. immitis* adult worms and subjected it to long-read sequencing with PacBio Iso-Seq, which incorporated a poly(A) selection step. A single *D. immitis* worm generated sufficient quantities of RNA, while multiple *B. malayi* worms were pooled to achieve adequate RNA input. Sequencing of the four libraries and data preparation with IsoSeq3 resulted in a combined 160 Mb of high-quality isoforms from almost 134,000 reads. More detailed sequencing statistics can be found in S2 Table.

We assessed the quality of reads with SQANTI [30], which reference-corrects, filters, and categorizes putative isoforms based on a number of quality control metrics, and we removed reads that had characteristics of reverse transcriptase template switching. After filtration, 60,230 and 71,267 reads remained from *B. malayi* and *D. immitis*, respectively (Fig 2A). We identified reads mapping to 1,847/1,873 known genes in females, 1,651/1,433 in males, and 2,430/2,423 in both sexes (Fig 2B), covering 54.2% and 44.6% of the *B. malayi* and *D. immitis* coding genomes, respectively. For *B. malayi*, we compared these long-read data to short-read RNA-seq data from adult males and females [53] and found long-reads mapping to 67.5% of genes that had a mean transcripts per million (TPM) value of greater than 10 in adults. Of genes that had mapped long-reads, there was a significant correlation between the number of reads and the TPM values (Pearson's $\rho = 0.33$, $p < 2.2e^{-16}$), suggesting that deeper sequencing could potentially capture lowly expressed transcripts. From these genes, we identified reads mapping to 6,197/10,522 unique transcripts in females, 6,646/5,542 in males, and 2,958/2,833 in both sexes (Fig 2C). These tallies of unique transcripts are partially inflated by the inclusion of truncated transcripts caused by RNA degradation, as we detected a 3' bias in libraries from both species (Fig 2D).

Improvement and correction of gene models by Iso-Seq

We used SQANTI categories to identify gene models that were improved, corrected, and/or validated by our dataset (Fig 2E). Full-splice matches (FSMs) indicate models that were validated, and many of these reads extended UTRs. 57/77% extended the 5' UTRs in *B. malayi* and *D. immitis*, respectively (Fig 3A), while 68/89% extended the 3' UTR (Fig 3B). Reads categorized as incomplete-splice matches (ISMs) were primarily 3' fragments, but these were still able to extend gene models at the 3' end (59/75% in *B. malayi* and *D. immitis*, respectively) (Fig 3B). In total, 1,632/3,713 transcripts in *B. malayi* and 1,481/3,502 in *D. immitis* had 5'/3' UTRs extended.

In addition to extended UTRs, we also added to the catalog of splice junctions used by these two filarial species. We identified 38,973/33,904 known and 4,527/4,651 novel canonical (G[UC]-AG) splice junctions and a total of 33/7 known and 818/947 novel non-canonical splice junctions in *B. malayi* and *D. immitis*, respectively (the top 10 splice junctions from *B. malayi* are shown in Fig 4A).

Finally, we identified reads that correspond to gene models that require concatenation (i.e., represent individual genes but are fragmented into multiple models). We identified reads that

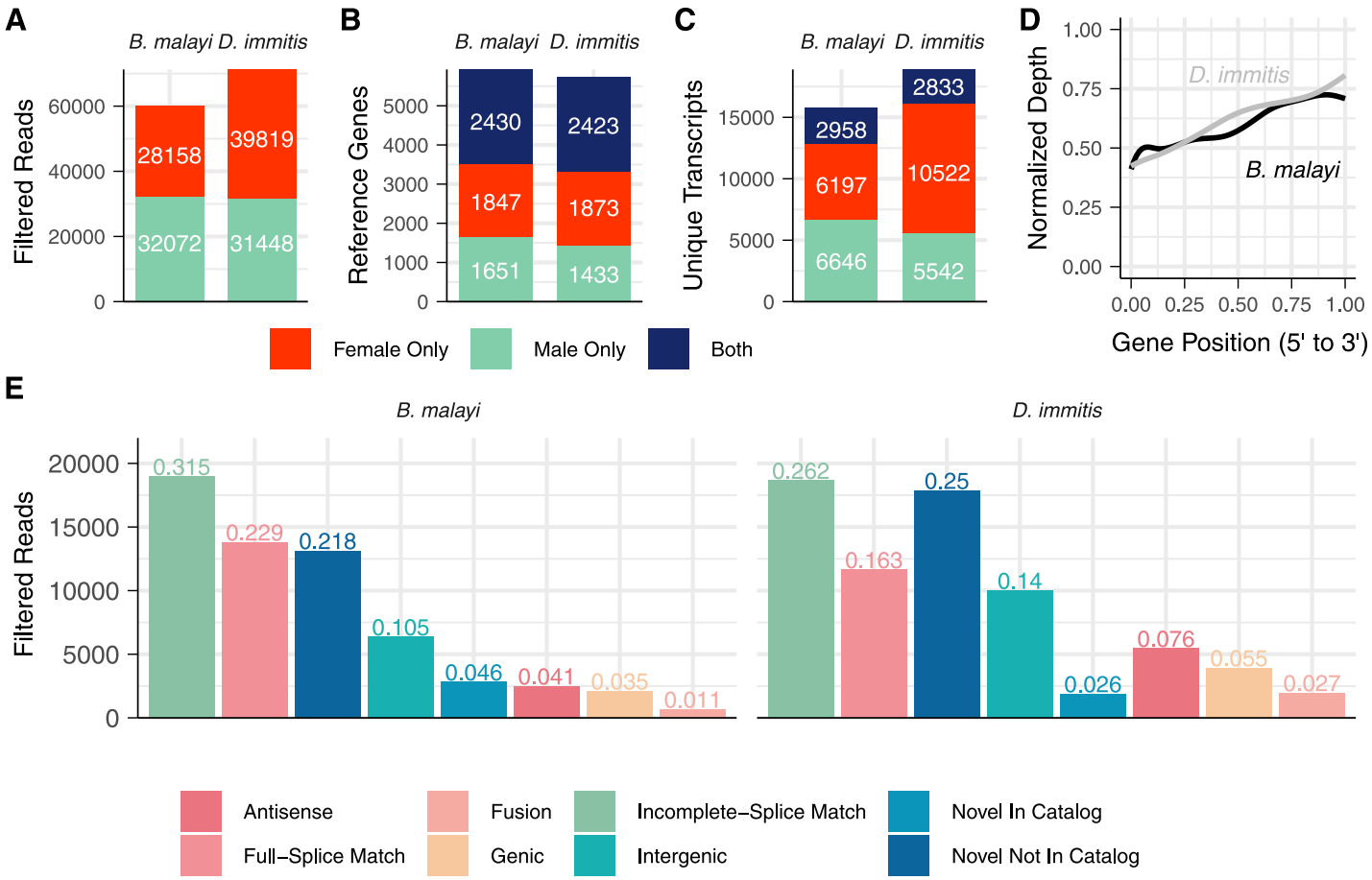

Fig 2. Summary of Iso-Seq data from adult male and female *B. malayi* and *D. immitis*. (A) The number of filtered reads generated from each sex and species. (B) The number of reference genes with at least one mapped read from only females, only males, or both sexes. (C) The number of unique transcripts captured from only females, only males, or both sexes. (D) Libraries arising from both species had a strong 3' bias due to poly(A) selection and mRNA degradation. Depth was normalized to the maximum and minimum read depth for each species. (E) Distribution of SQANTI structural category assignments. Numbers above bars indicate the percentage of reads in each respective category. Antisense: maps to the opposite strand of a predicted gene. Full-splice match: confirms the annotated gene model. Fusion: maps to two consecutive genes. Genic: maps to a single exon within a predicted gene. Incomplete-splice match: partially confirms the annotated gene model. Intergenic: maps to an intergenic region. Novel in catalog: utilizes known splice sites in novel combinations. Novel not in catalog: uses novel splice sites.

https://doi.org/10.1371/journal.pntd.0008869.g002

aligned to >1 gene, extracted sequences from the aligned genes, performed a search of the predicted proteins against the *C. elegans* predicted proteome, and looked for reads that had a single hit spanning the majority of the read (see S1 Fig for an example). These data were used to distinguish incorrectly fragmented gene models and putative operons, discussed in subsequent sections.

Discovery of novel isoforms and genes encoding proteins and potential long non-coding RNAs

In addition to identifying ways in which currently annotated gene models were improved, we used these data to search for entirely novel genes and isoforms. Reads categorized as NIC or NNC align to either new combinations of splice junctions or novel unannotated junctions, making these categories rich in potential new isoforms. In known genes with high-quality models, splice junctions are equally abundant across the middle portion of transcripts (as shown by the FSM data in Fig 4B). Interestingly, junctions covered by NNC and NIC reads

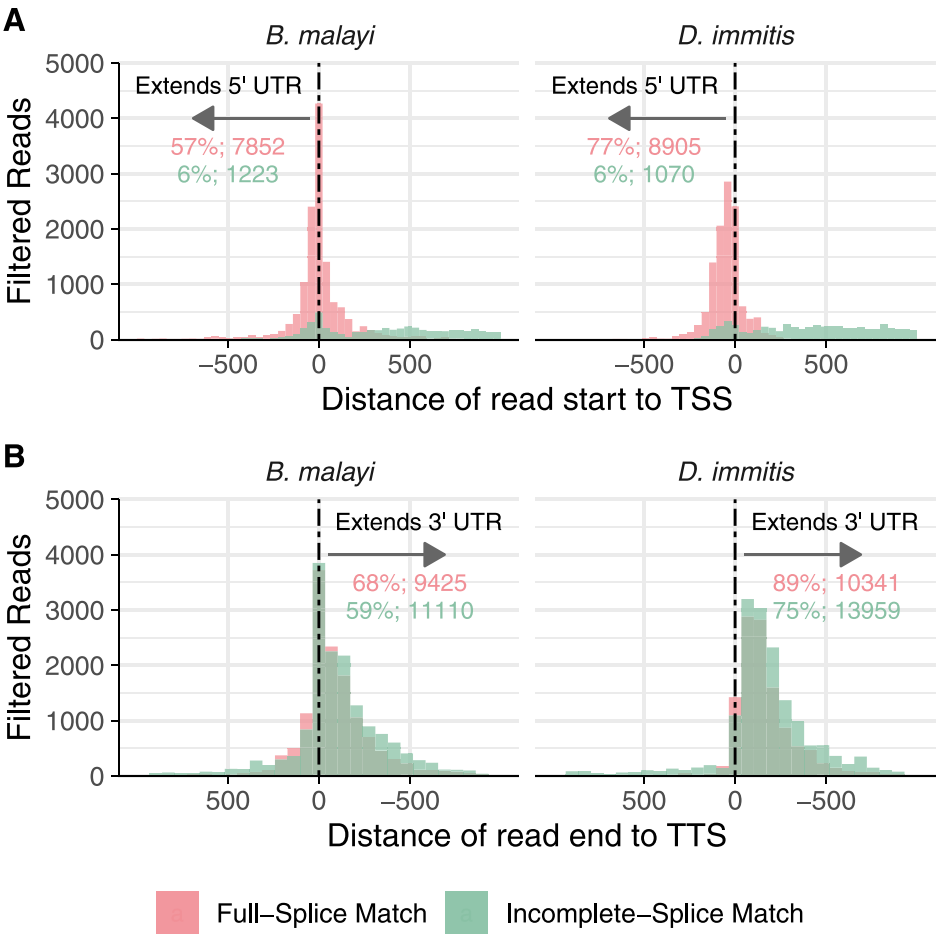

Fig 3. Iso-Seq extends UTRs of known transcripts. Filtered reads categorized as full or incomplete-splice matches (ISM) were examined for whether they extended the 5' and 3' untranslated regions. (A) A high portion of full-splice match (FSM) reads extend the annotated 5' UTRs of known transcripts in both *B. malayi* (left) and *D. immitis* (right), while ISM reads rarely cover the 5' end. (B) Both FSM and ISM reads extend the 3' UTRs of known transcripts. Negative values indicate reads that cover the transcriptional start or termination sites. Percentages and values printed on the plot represent the percentage and number of total reads in that category that extend the given UTR. TSS: transcriptional start site. TTS: transcriptional termination site.

https://doi.org/10.1371/journal.pntd.0008869.g003

show a similar distribution to those covered by FSMs, while junctions covered by ISM reads are more abundant within the first 40 nt, causing a statistically significant 1.5% shift in mean position ($p < 10^{-10}$, Tukey's honestly significant difference). Reads classified as ISMs tend to be 3' fragments, explaining why splice junctions had a greater propensity to map to these first 40 nt (Fig 4C). In contrast, the similar splice junction distributions among FSMs, NNCs, and NICs suggest that many NNC and NIC reads are full-length transcripts that represent unannotated isoforms.

We next looked for reads that may represent novel polyadenylated genes. We found a high number of reads categorized as antisense or intergenic, particularly in *D. immitis* (Fig 2E). These categories represent reads that map to regions of the genome that do not have predicted genes (intergenic) or have predicted genes on the opposite strand (antisense). We found 1670/1878 multi-exonic loci of these potentially novel genes in *B. malayi* and *D. immitis*, respectively. We reasoned that these reads could represent genes for either novel proteins or non-coding RNAs (ncRNAs).

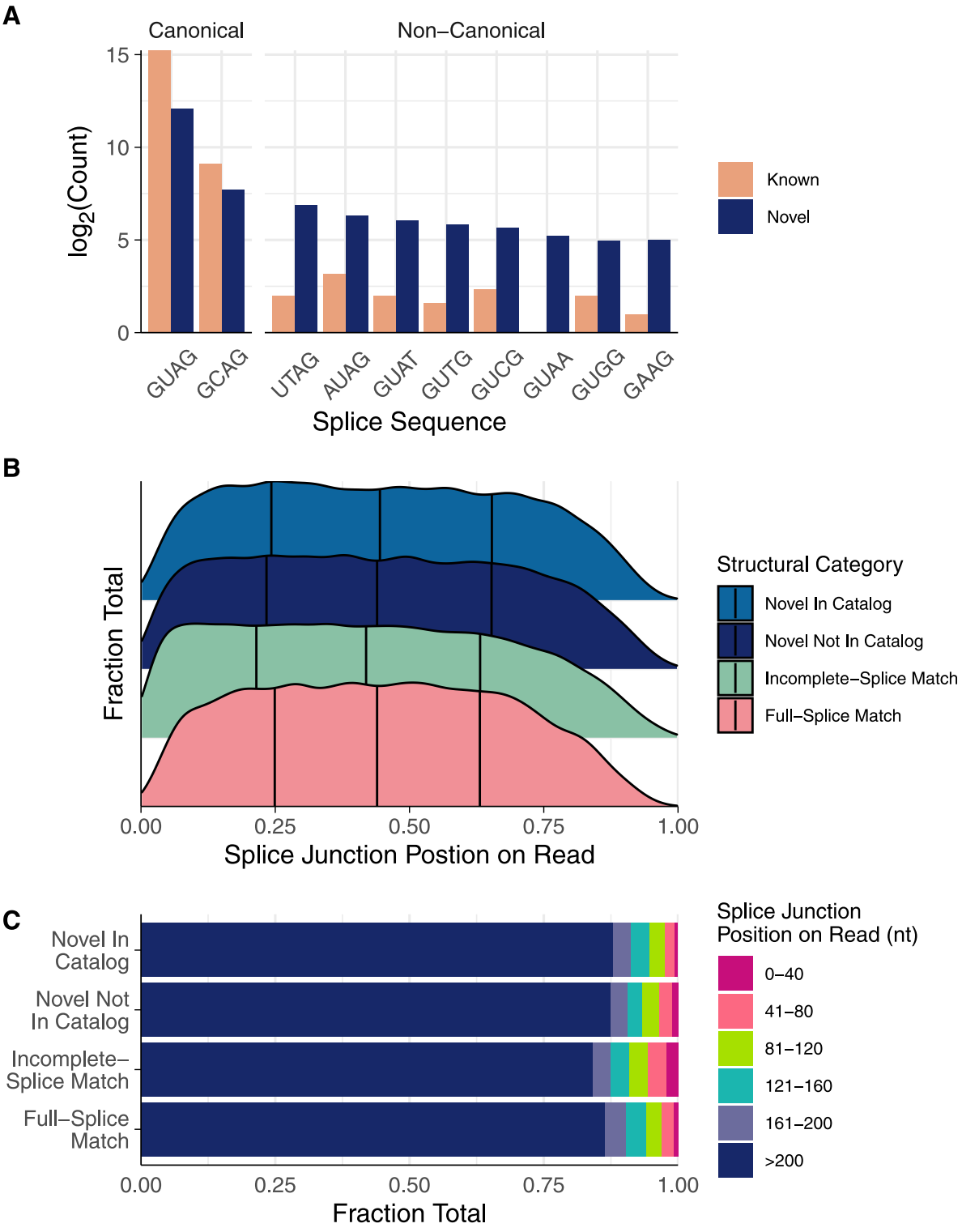

Fig 4. Iso-Seq adds to the catalog of splice junctions and identifies unannotated isoforms. (A) The 10 most abundant splice junctions in *B. malayi*. Most splice junctions mapped by reads were either the canonical G[UC]-AG, but a large portion of novel splice junctions identified were non-canonical, indicating the utility of Iso-Seq for identifying non-canonical junctions. (B and C) Transcripts typically contain splice junctions equally within the middle 50% of the transcript, while junctions in the first 10% and last 25% are rarer. Incomplete-splice match (ISM) reads contain a higher proportion of splice junctions at the 5' end of the read when compared to full-splice matches, novel in catalog (NIC) reads, and novel not in catalog (NNC) reads. This suggests that ISMs represent degraded mRNA, while NIC and NNC reads likely represent true novel isoforms or transcripts with corrected splice sites.

https://doi.org/10.1371/journal.pntd.0008869.g004

We focused on potentially novel protein-coding genes by filtering for reads that had a multi-exonic open reading frame (ORF), as mono-exonic reads are more likely to be the result of oligo(dT) priming of poly(A) stretches within contaminating genomic DNA. We found 209/501 of these reads mapping to 50/108 loci in *B. malayi* and *D. immitis*, respectively (S3 and S4 Tables), and these reads are not similar to already predicted proteins but are similar to predicted proteins from other nematodes. We manually curated these potential novel genes in both species and found that 6/56 of these putatively genic loci were artifacts, 6/12 contained genes requiring extension by the addition of new exons, 2/4 contained genes with novel isoforms, 8/0 contained pseudogenes, and 17/29 contained novel protein coding genes in *B. malayi* and *D. immitis*, respectively (S3 and S4 Tables). The majority of these were nematode and filarid-specific hypothetical genes, but we also identified genes such as *Dim-serp-1.1*, *Dim-tomm-7*, *Dim-rpl-39*, *and Dim-ser-3*, among others.

The remaining antisense and intergenic reads could represent poly-adenylated ncRNAs. Only 39 ncRNAs are annotated in the *B. malayi* genome, and none have been identified in *D. immitis*. For *B. malayi*, all of these are mono-exonic, even though some ncRNAs such as long non-coding (lncRNAs) can be multi-exonic and use the same splicing machinery as pre-mRNAs. We found 67 reads mapping to 4 of these annotated ncRNA genes (WBGene00221733, WBGene00255379, WBGene00230800, and WBGene00220274). Among our potential novel protein-coding genes, we found 5/2 coded for ncRNAs (S3 and S4 Tables). We attempted to determine whether the remaining reads could represent lncRNAs, but lncRNAs are typically expressed in lower abundance than mRNAs, and confidently predicting them would require the integration of a large amount of short-read RNA-seq datasets combined with *de novo* gene prediction.

Identification of new operons

A mixture of experimental and computational approaches have been used to annotate operons in the *B. malayi* genome, relying heavily on evidence from *C. elegans* [56,57]. Iso-Seq captures transcripts at a variety of processing steps, enabling the sequencing of some unspliced polycistronic transcripts. For operon prediction, we focused on the *B. malayi* data, which was less likely to include reads that mapped to single genes composed of multiple fragmented gene models. We identified 1,775 reads that aligned to 443 loci containing multiple unique, consecutive genes.

Manual inspection of these putative operons revealed that some were indicative of fragmented gene models or were false positives resulting from misalignment, reads aligning to the opposite strand of neighboring genes, or genomic DNA contamination. We manually curated all putative operons that had an intergenic distance of less than 5000, 338 loci in total.

As expected, we found that some predicted operons, both in our dataset and in the assembly annotations, were actually fragmented models that belonged to a single gene. These fragments tended to include one long fragment that was highly similar, but shorter, than its nearest nematode ortholog, and a second shorter fragment with an ORF of 30–60 amino acids. Reads spanning these fragments had a single ORF that was significantly similar to a single nematode ortholog. We found 52 of these fragmented loci that require concatenation, and 19 of these were annotated operons in the Bmal-4.0 assembly that we deprecated. On the other hand, true operons had ORFs that were of similar size and typically >100 amino acids, and the reads spanning them often had unspliced intercistronic introns that separated two large ORFs with similarity to multiple nematode orthologs. We identified 34 novel operons and validated 102 assembly-annotated operons. These 136 operons contain 354 genes for an average of 2.60 genes per operon (Fig 5A, S5 Table).

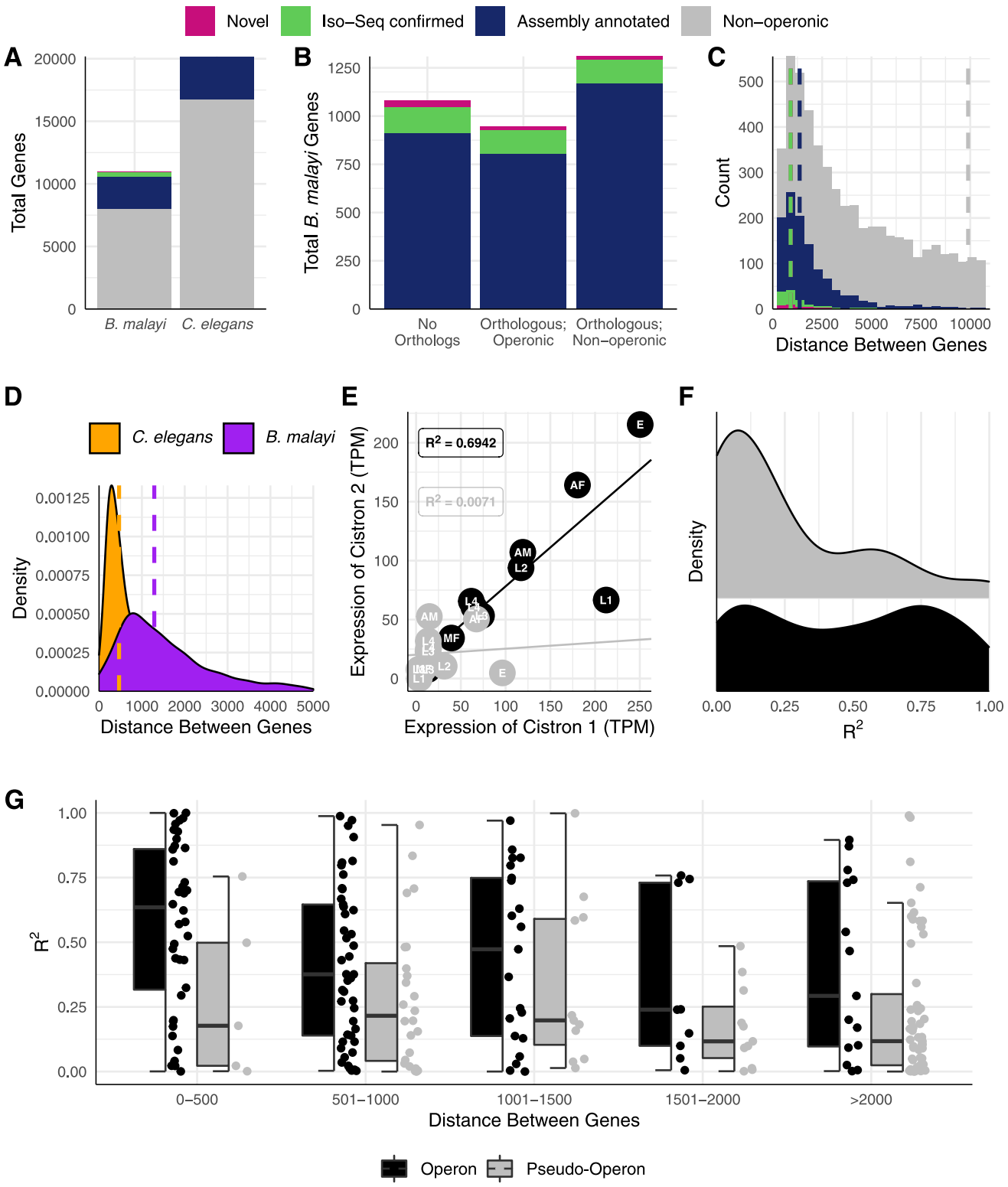

Fig 5. Identification and validation of over a hundred operons in *B. malayi*. (A) Iso-Seq identified new operons and validated assembly-annotated operons in *B. malayi*, which has a similar proportion of its genome arranged in operons as *C. elegans*. (B) Most *B. malayi* genes in operons have orthologs in *C. elegans* that are not in operons (39%), while 32% do not have orthologs, and 28% have orthologs that are also in operons. (C) Intercistronic distances between genes in operons are much shorter than intergenic distances between non-operonic genes (note that green and magenta median lines overlap). Operons identified via Iso-Seq (a median of 915.0 nucleotides for new operons, 906.5 for validated operons) on average had shorter intercistronic distances than those already annotated in the Bmal-4.0 assembly (1362.0). (D) *B. malayi* operons on average have larger (median = 1289 contrasted to 467 for *C.*

elegans) and a broader range (interquartile range of 1422.5 contrasted to 700.5 for *C. elegans*) intercistronic distances than those in *C. elegans*. (E) An example of linear fits to staged RNA-seq data from genes in a pseudo-operon (non-operonic neighboring genes) and an operon. (F) Linear fits better explain the relationship between expression of neighboring genes in operons than neighboring genes in pseudo-operons, and (G) this contrast does not stratify by intergenic distance.

https://doi.org/10.1371/journal.pntd.0008869.g005

We pulled *C. elegans* orthologs from WormBase ParaSite [28] and found that most of the newly identified operonic genes do not have a *C. elegans* ortholog (Fig 5B). This is likely explained by the previous utilization of *C. elegans* operons to inform *B. malayi* operon annotation, highlighting the power of unbiased sequencing approaches to identifying operons. In *C. elegans*, many operonic genes function in RNA processing and protein synthesis in ribosomes, as well as in mitochondria [58]. Likewise, we found that genes in *B. malayi* operons are enriched in GO terms signifying function in protein synthesis (e.g., mRNA and tRNA processing) (S2 Fig), reflecting the orthology between the *C. elegans* and *B. malayi* datasets.

The distances between consecutive cistrons in *B. malayi* operons are substantially shorter than monocistronic genes (Fig 5C, 1865.64 nucleotides for cistrons and 16968.05 for non-operonic genes), as in other nematodes, but the distribution of distances is much broader than that of *C. elegans* (Fig 5D, interquartile range of 1422.5 contrasted to 700.5 for *C. elegans*). This may be explained by the lack of the SL2 splice-leader sequence in *B. malayi*. While SL2 is the predominant SL used for trans-splicing of downstream cistrons in *C. elegans*, SL1-spliced downstream cistrons are often associated with longer intergenic distances [59].

Operons allow for the co-regulation of neighboring genes, which often results in transcript abundances that are correlated between paired cistrons [60]. We found that the expression abundances of the first pair of neighboring cistrons in *B. malayi* operons show a stronger linear relationship than pseudo-operons created by neighboring monocistronic genes (Fig 5E–5G). We used publicly available short-read RNA-seq data from staged parasites and fit a linear model to these values using the abundance of the first cistron as a predictor for the abundance of the second cistron (example shown in Fig 5E). In general, a linear model did not fit well to consecutive monocistronic genes (median $R^2 = 0.158$), but performed better for operonic genes (median $R^2 = 0.484$) (Fig 5F), and this contrast is not altered by the intergenic distance (Fig 5G). Interestingly, the R^2 values of the linear fits of operonic genes appear to be bimodal, which could indicate post-transcriptional regulation that may cause a deviation from the linear relationship.

Analysis of poly(A) tails in filarial nematodes identifies alternative poly(A) addition sites

PacBio chemistry is adept at sequencing long homopolymers, allowing for the analysis of poly (A) tails and poly(A) addition sites (PAS). Though there are alternative focused methods for determining poly(A) tail lengths, derivatives of Iso-Seq have been shown to correlate well in *C. elegans* [61]. We found the median poly(A) tail length for reads from *B. malayi* was 33 nt (range of 1–541) and 35 (range of 1–536) for *D. immitis* (Fig 6A), shorter than the reported *C. elegans* median of around ~50 nt (depending on life stage and technique) and much shorter than the typical lengths for human organoids, iPSCs, and cell lines [61–63]. Nearly 40% of *B. malayi* reads contained the canonical AAUAAA PAS motif, with AUAAAN as the second most abundant (approximately 15% in total) (Fig 6B). About 8% of reads did not have an identifiable PAS but instead included an A-rich region in the 50 nt upstream of the poly(A) addition site (Fig 6B–6D). Interestingly, we also identified an upstream sequence element (USE) that is a known recognition site for some RNA binding proteins (UUUUGU) in approximately 4% of reads that did not have other PASs [64,65].

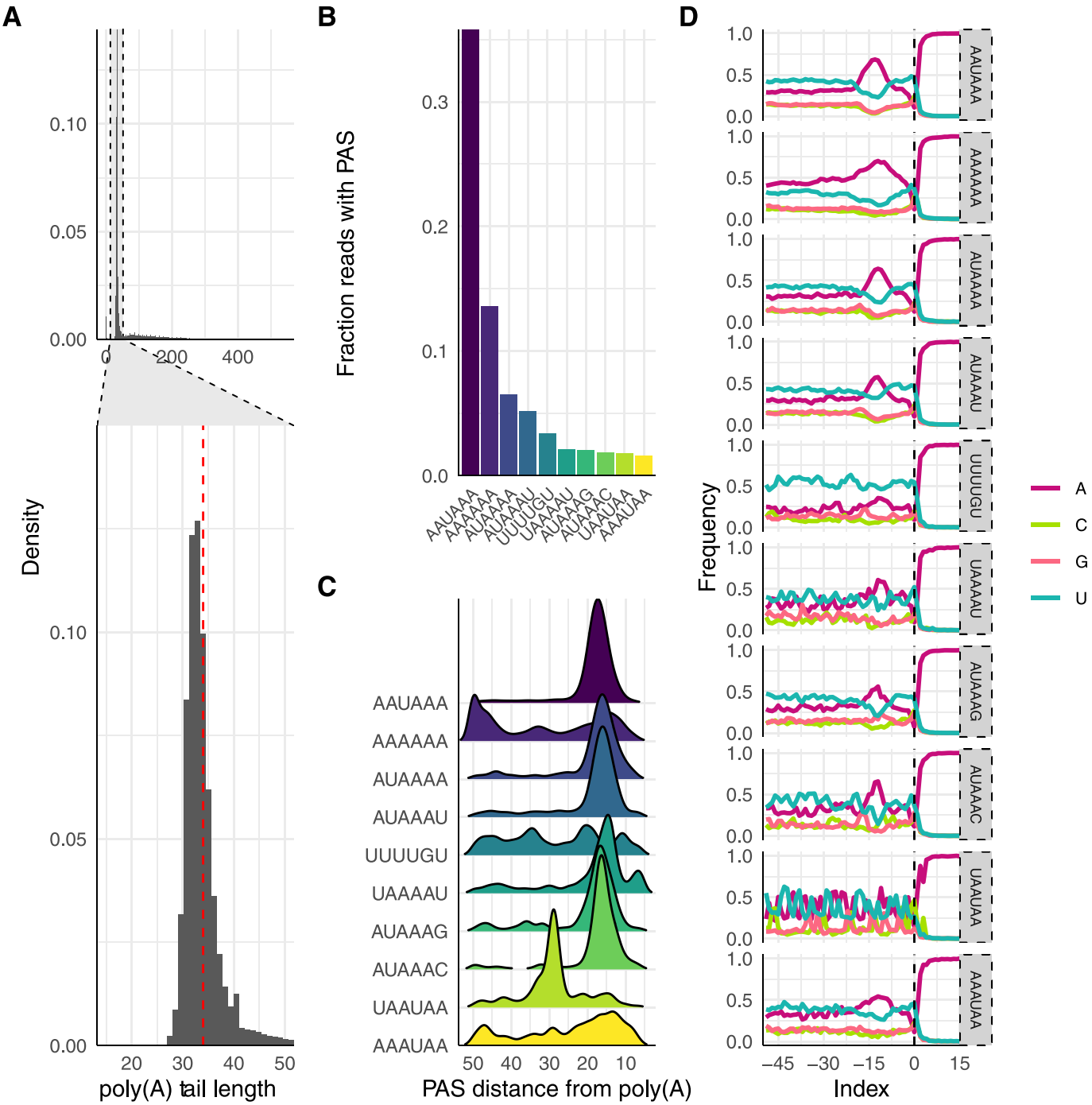

Fig 6. Analysis of poly(A) tails in *B. malayi*. (A) Poly(A) tails on reads from *B. malayi* ranged from 0–541 and had a median length of 33. (B) Greater than 35% of reads contained the canonical poly(A) signal (PAS) AAUAA, while A-rich regions were the second most abundant PAS. (C) All PASs occur 10–20 nucleotides upstream of the poly(A) cleavage site, other than AAAAA, which has two A-rich sites at 10–20 and 45–50 nucleotides upstream, and UAAUAA, which may not be a PAS but is a motif for RNA binding proteins [64,65]. (D) All PASs also include an A-rich region at 15 nucleotides upstream of the cleavage site.

https://doi.org/10.1371/journal.pntd.0008869.g006

Improvement of gene models for known and putative anthelmintic targets

We curated a list of 205 genes of interest in *B. malayi* and *D. immitis*, consisting of known anthelmintic targets and receptors belonging to druggable protein families in parasitic

nematodes (S6 Table). We examined the genomic loci of these targets for reads that mapped to these locations and cross-referenced these reads with SQANTI classifications. In *B. malayi*, 28 of these targets had mapped reads, most with multiple reads, and 11 of these had FSMs that confirmed the reference model and included 5' and 3' UTRs. For example, *Bma-slo-1* (Bm6719) had FSMs from both sexes (Fig 7A) and *Bma-gar-3* (Bm13584) had the b isoform validated (Fig 7C). *Bma-slo-1* has been shown to have sex-specific splice patterns that likely contribute to the differential response to emodepside [21], and our Iso-Seq data detected the usage of both the a and f isoforms in females and only the f isoform in males, consistent with this interpretation (Fig 7A). We also captured in females an additional isoform that contains both the exons that show alternative usage in the a and f isoforms; this isoform would not have been picked up by PCR and restriction digest test used to previously define *Bma-slo-1* isoform usage.

Other targets had models corrected. While only a single full-length *Bma-osm-9* read was captured, multiple reads from degraded transcripts validated the previously corrected splice recognition site in *Bma-osm-9* (Fig 7B) [15]. This is one of many examples encountered during our curatorial efforts where the evidence weighed by gene prediction algorithms made errors based on preferences for canonical splice junctions, while Iso-Seq reads clearly supported a nearby non-canonical junction. Finally, a read mapping to *Bm17517* (an *srt* family chemoreceptor [15]), showed evidence for a corrected or alternative TSS and start codon (Fig 7D). The reference model for *Bm17517* contains a 9 nt first exon that has limited evidence from short-read and EST data sources, while Iso-Seq reads suggests that this exon is spurious and instead the ORF of the second exon extends further upstream. These data demonstrate the ways in which Iso-Seq data can be used to correct the models for genes that are under active research and cloning efforts. The remaining 24 *B. malayi* targets with mapped reads (including acetylcholine and glutamate-gated ion channels, TRP channels, CNG channels, and GPCRs) can be found in S2–S6 Figs.

In addition to anthelmintic targets, we used our Iso-Seq data to manually curate the two longest scaffolds of the *D. immitis* genome assembly. Scaffolds 1 and 2 contain 163 and 133 genes, respectively, and we curated 58 and 46 of these using both short-read and long-read data as evidence, as well as 44 additional genes scattered throughout the assembly. Nearly all of these curations included the addition of UTRs, while many also included the addition of new isoforms, adding exons and adjusting the translational start and stop sites, and correcting splice sites. This curation will be ongoing, and these data will be of particular use as the *D. immitis* assembly becomes more complete and contiguous.

Discussion

In recent years there has been a concerted push to increase the availability of genomic resources for parasitic worms of medical, agricultural, and veterinary importance, highlighted by the recent release of 81 draft genomes from both free-living and parasitic flatworms and nematodes [36]. These resources promise to considerably bolster understanding of the evolution of parasitism and parasitic biological processes through the use of comparative genomics.

Although many important analyses can be performed with draft genomes of suboptimal contiguity, experimentation upon individual genes requires gene model predictions in which one can be highly confident. *Ab initio* gene prediction algorithms can be optimized to provide high BUSCO or CEGMA scores but often miss the genes in which parasitologists are most interested: those that are not conserved in hosts and are likely to be involved in parasitism [66,67].

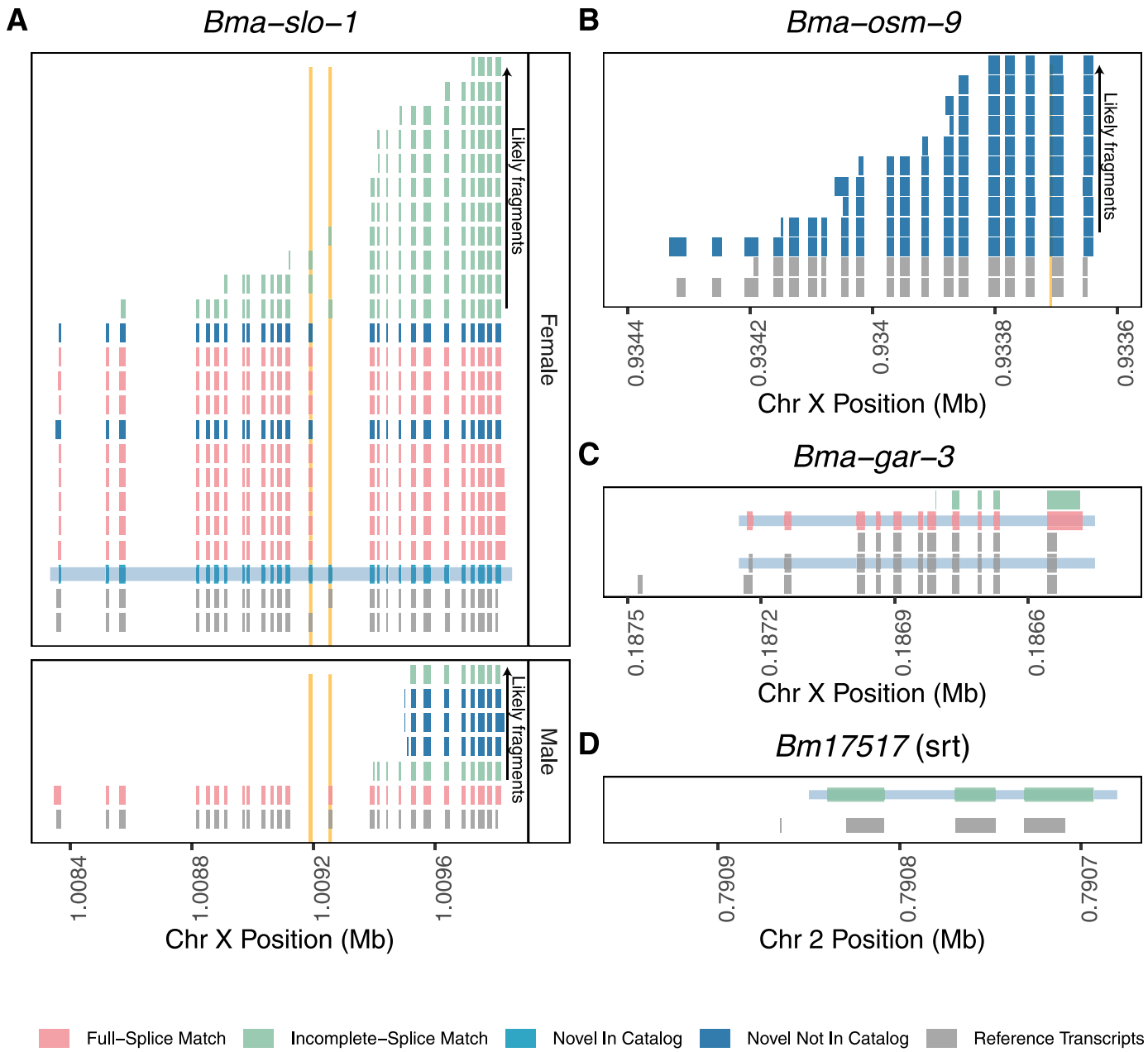

Fig 7. Correction of models for genes currently studied as potential anthelmintic targets. All transcripts/reads are shown in 5' to 3' orientation; the x-axis has been flipped for transcripts on the (-) strand. (A) SLO-1, a target for emodepside, has multiple alternative isoforms that play a role in emodepside response [21]. Iso-Seq captured a previously unidentified isoform in *B. malayi* females (highlighted in light blue). This read was categorized as "Novel In Catalog" because it contains a novel arrangement of known exons (highlighted in orange) and splice sites. Females also express *slo-1* isoforms with one or the other exon, which is seen in all the remaining reads. Males only express the isoform with the second highlighted exon. (B) All reads mapping to *osm-9* showed a mispredicted splice junction (highlighted in orange) in the penultimate exon. (A & B) Both *slo-1* and *osm-9* had multiple reads that arose from fragmented transcripts. These transcripts were captured by poly(A) selection but did not contain the full-length transcript. (C) Iso-Seq confirms the b isoform of *gar-3* (highlighted in light blue), but could not rule out expression of the other isoforms. (D) Iso-Seq shows the incorrect prediction of the first exon and transcriptional start site of the chemoreceptor *Bm17517* (small gray exon at the 5' end) and extends the coding sequence of exon 1.

https://doi.org/10.1371/journal.pntd.0008869.g007

Thus, the next substantial push for curators of helminth genomes should be to improve the gene model prediction and annotation pipeline. Short-read RNA-seq data has been instrumental in this aim and will continue to play an important role, but the technology is limited in its ability to differentiate between isoforms or resolve full-length transcripts. The recent development of long-read sequencing technologies can overcome these obstacles.

We report here the first long-read RNA-seq data for filarial nematodes, an important group of parasites that infect humans and animals. While the genome assembly of *B. malayi* is chromosome-scale, there remains serious errors in gene models, even for genes that are highly conserved among nematodes and have been cloned in other species [15]. *D. immitis*, on the other hand, has a far less contiguous genome assembly and likely contains even more mispredicted gene models than *B. malayi*. The completeness of the *B. malayi* and *D. immitis* transcriptomes stands to be significantly improved once these data are integrated into gene and transcript prediction pipelines. The transcriptome assembly of *D. immitis*, in particular, will be substantially improved by these data. Toward this end, we have used long-read evidence to curate the >150 genes on the two largest scaffolds of the *D. immitis* draft genome and 31 novel genes scattered throughout the genome. Until now, there was not a single alternative isoform predicted in this species, predicted UTRs were sparse, and very few non-canonical splice sites were annotated. Likewise, we used these data to curate novel genes (S3 Table) and operons (S5 Table) in *B. malayi*. Filtered reads for *B. malayi* have been included in WormBase WS277 [46] as genome browser tracks, which will be integrated into gene prediction pipelines in future releases, and the *D. immitis* data has been integrated as a genome browser tracks on WormBase ParaSite [28].

Our data had a high 3' bias due to RNA degradation and poly(A) selection and covered approximately 50% of the coding transcriptomes of *B. malayi* and *D. immitis*. Recent improvements in the Iso-Seq library preparation workflow will increase transcriptome coverage and isoform discovery, and allow for the sequencing of RNA from RNA-poor larvae and individual worms or tissues. A reduction in minimum RNA input will also enable techniques that enrich for full-length RNA molecules that have both the m7G cap and the poly(A) tail [68,69]. These approaches promise to further improve upon the enhancements we report here and should be used in the cases of other nematode species.

Supporting information

S1 Fig. Example of a fragmented gene model (*Dim-tax-4*) that requires concatenation.
(PDF)

S2 Fig. GO enrichment analysis of operonic genes from *B. malayi*.
(PDF)

S3 Fig. Reads mapping to the two *ben-1* orthologs in *B. malayi*.
(PDF)

S4 Fig. Reads mapping to glutamate and acetylcholine-gated ion channels in *B. malayi*.
(PDF)

S5 Fig. Reads mapping to miscellaneous ion channels in *B. malayi*.
(PDF)

S6 Fig. Reads mapping to GPCRs in *B. malayi*.
(PDF)

S1 Table. NCBI SRA information for public RNA-seq datasets utilized.
(CSV)

S2 Table. Sequencing statistics and metadata.
(CSV)

S3 Table. Curated genomic loci in *B. malayi* predicted to contain novel genes.
(CSV)

S4 Table. Curated genomic loci in *D. immitis* predicted to contain novel genes.
(CSV)

S5 Table. List of validated and novel operons and their genes in *B. malayi*.
(CSV)

S6 Table. List of 205 genes of interest in *B. malayi* and *D. immitis*, consisting of known anthelmintic targets and receptors belonging to druggable protein families.
(CSV)

Acknowledgments

Some parasite materials were provided by the NIH/NIAID Filariasis Research Reagent Resource Center (www.filariasiscenter.org). *D. immitis* RNA extractions were carried out by Jessica Grant and Steve Williams at Smith College. RNA sequencing was carried out at the University of Wisconsin-Madison Biotechnology Center. *B. malayi* data was uploaded to WormBase and *B. malayi* gene and operon curation was performed by Michael Paulini. Faye Rogers at WormBase ParaSite assisted in setting up the Apollo instance for *D. immitis* gene curation. The authors would like to thank members of the Zamanian laboratory for assistance in *D. immitis* gene curation and critical comments on the manuscript.

Author Contributions

Conceptualization: Mostafa Zamanian.

Data curation: Nicolas J Wheeler, Mostafa Zamanian.

Formal analysis: Nicolas J Wheeler, Mostafa Zamanian.

Funding acquisition: Mostafa Zamanian.

Investigation: Nicolas J Wheeler, Mostafa Zamanian.

Methodology: Nicolas J Wheeler, Paul M. Airs, Mostafa Zamanian.

Project administration: Nicolas J Wheeler, Mostafa Zamanian.

Resources: Nicolas J Wheeler, Paul M. Airs, Mostafa Zamanian.

Software: Nicolas J Wheeler, Mostafa Zamanian.

Supervision: Mostafa Zamanian.

Validation: Nicolas J Wheeler, Mostafa Zamanian.

Visualization: Nicolas J Wheeler, Mostafa Zamanian.

Writing – original draft: Nicolas J Wheeler, Mostafa Zamanian.

Writing – review & editing: Nicolas J Wheeler, Paul M. Airs, Mostafa Zamanian.

References

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392:1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7 PMID: 30496104
2. Simón F, Siles-Lucas M, Morchón R, González-Miguel J, Mellado I, Carretón E, et al. Human and animal dirofilariasis: the emergence of a zoonotic mosaic. Clin Microbiol Rev. 2012; 25:507–544. https://doi.org/10.1128/CMR.00012-12 PMID: 22763636
3. Lymphatic filariasis. In: Lymphatic filariasis. 6 Oct 2019 [cited 3 Feb 2020]. https://www.who.int/news-room/fact-sheets/detail/lymphatic-filariasis.
4. American Heartworm Society. Current Canine Guidelines for the Prevention, Diagnosis, and Management of Heartworm (Dirofilaria immitis) Infection in Dogs. 2018. https://d3ft8sckhnqim2.cloudfront.net/images/pdf/2018_AHS_Canine_Guidelines_rev_7-25-19.pdf?1564157216.
5. McCarthy J. Is anthelmintic resistance a threat to the program to eliminate lymphatic filariasis? Am J Trop Med Hyg. 2005; 73:232–233. PMID: 16103580
6. Eberhard ML, Lammie PJ, Dickinson CM, Roberts JM. Evidence of nonsusceptibility to diethylcarbamazine in Wuchereria bancrofti. J Infect Dis. 1991; 163:1157–1160. https://doi.org/10.1093/infdis/163.5.1157 PMID: 2019765
7. Schwab AE, Boakye DA, Kyelem D, Prichard RK. Detection of benzimidazole resistance-associated mutations in the filarial nematode Wuchereria bancrofti and evidence for selection by albendazole and ivermectin combination treatment. Am J Trop Med Hyg. 2005; 73:234–238. PMID: 16103581
8. Pulaski CN, Malone JB, Bourguinat C, Prichard R, Geary T, Ward D, et al. Establishment of macrocyclic lactone resistant Dirofilaria immitis isolates in experimentally infected laboratory dogs. Parasit Vectors. 2014; 7:494. https://doi.org/10.1186/s13071-014-0494-6 PMID: 25376278
9. Ballesteros C, Pulaski CN, Bourguinat C, Keller K, Prichard RK, Geary TG. Clinical validation of molecular markers of macrocyclic lactone resistance in Dirofilaria immitis. Int J Parasitol Drugs Drug Resist. 2018; 8:596–606. https://doi.org/10.1016/j.ijpddr.2018.06.006 PMID: 30031685
10. Bilsland E, Bean DM, Devaney E, Oliver SG. Yeast-Based High-Throughput Screens to Identify Novel Compounds Active against Brugia malayi. PLoS Negl Trop Dis. 2016; 10:e0004401. https://doi.org/10.1371/journal.pntd.0004401 PMID: 26812604
11. Bennuru S, O'Connell EM, Drame PM, Nutman TB. Mining Filarial Genomes for Diagnostic and Therapeutic Targets. Trends Parasitol. 2018; 34:80–90. https://doi.org/10.1016/j.pt.2017.09.003 PMID: 29031509
12. Geary TG, Thompson DP, Klein RD. Mechanism-based screening: discovery of the next generation of anthelmintics depends upon more basic research. Int J Parasitol. 1999; 29:105–12; discussion 113–4. https://doi.org/10.1016/s0020-7519(98)00170-2 PMID: 10048823
13. Woods DJ, Knauer CS. Discovery of veterinary antiparasitic agents in the 21st century: a view from industry. Int J Parasitol. 2010; 40:1177–1181. https://doi.org/10.1016/j.ijpara.2010.04.005 PMID: 20430031
14. Tracey A, Foster JM, Paulini M, Grote A, Mattick J, Tsai Y-C, et al. Nearly Complete Genome Sequence of Brugia malayi Strain FR3. Microbiol Resour Announc. 2020; 9. https://doi.org/10.1128/MRA.00154-20 PMID: 32527783
15. Wheeler NJ, Heimark ZW, Airs PM, Mann A, Bartholomay LC, Zamanian M. Genetic and functional diversification of chemosensory pathway receptors in mosquito-borne filarial nematodes. PLoS Biol. 2020; 18:e3000723. https://doi.org/10.1371/journal.pbio.3000723 PMID: 32511224
16. Williams GW, Davis PA, Rogers AS, Bieri T, Ozersky P, Spieth J. Methods and strategies for gene structure curation in WormBase. Database. 2011; 2011:baq039. https://doi.org/10.1093/database/baq039 PMID: 21543339
17. Dent JA, Smith MM, Vassilatis DK, Avery L. The genetics of ivermectin resistance in Caenorhabditis elegans. Proc Natl Acad Sci U S A. 2000; 97:2674–2679. https://doi.org/10.1073/pnas.97.6.2674 PMID: 10716995
18. McCavera S, Rogers AT, Yates DM, Woods DJ, Wolstenholme AJ. An ivermectin-sensitive glutamate-gated chloride channel from the parasitic nematode Haemonchus contortus. Mol Pharmacol. 2009; 75:1347–1355. https://doi.org/10.1124/mol.108.053363 PMID: 19336526
19. Yates DM, Wolstenholme AJ. An ivermectin-sensitive glutamate-gated chloride channel subunit from Dirofilaria immitis. Int J Parasitol. 2004; 34:1075–1081. https://doi.org/10.1016/j.ijpara.2004.04.010 PMID: 15313134
20. Cheeseman CL, Delany NS, Woods DJ, Wolstenholme AJ. High-affinity ivermectin binding to recombinant subunits of the Haemonchus contortus glutamate-gated chloride channel. Mol Biochem Parasitol. 2001; 114:161–168. https://doi.org/10.1016/s0166-6851(01)00258-4 PMID: 11378196

21. Kashyap SS, Verma S, Voronin D, Lustigman S, Kulke D, Robertson AP, et al. Emodepside has sex-dependent immobilizing effects on adult Brugia malayi due to a differentially spliced binding pocket in the RCK1 region of the SLO-1 K channel. PLoS Pathog. 2019; 15:e1008041. https://doi.org/10.1371/journal.ppat.1008041 PMID: 31553770

22. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003; 100:15776–15781. https://doi.org/10.1073/pnas.2136655100 PMID: 14663149

23. Foster JM, Grote A, Mattick J, Tracey A, Tsai Y-C, Chung M, et al. Sex chromosome evolution in parasitic nematodes of humans. Nat Commun. 2020; 11:1964. https://doi.org/10.1038/s41467-020-15654-6 PMID: 32327641

24. Saito TL, Hashimoto S-I, Gu SG, Morton JJ, Stadler M, Blumenthal T, et al. The transcription start site landscape of C. elegans. Genome Res. 2013; 23:1348–1361. https://doi.org/10.1101/gr.151571.112 PMID: 23636945

25. Magrini V, Gao X, Rosa BA, McGrath S, Zhang X, Hallsworth-Pepin K, et al. Improving eukaryotic genome annotation using single molecule mRNA sequencing. BMC Genomics. 2018; 19:172. https://doi.org/10.1186/s12864-018-4555-7 PMID: 29495964

26. Doyle SR, Tracey A, Laing R, Holroyd N, Bartley D, Bazant W, et al. Extensive genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus contortus, a model gastrointestinal worm. bioRxiv. 2020. p. 2020.02.18.945246. https://doi.org/10.1101/2020.02.18.945246

27. Michalski ML, Griffiths KG, Williams SA, Kaplan RM, Moorhead AR. The NIH-NIAID Filariasis Research Reagent Resource Center. PLoS Negl Trop Dis. 2011; 5:e1261. https://doi.org/10.1371/journal.pntd.0001261 PMID: 22140585

28. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite—a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 2017; 215:2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005 PMID: 27899279

29. Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, et al. The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. FASEB J. 2012; 26:4650–4661. https://doi.org/10.1096/fj.12-205096 PMID: 22889830

30. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 2018. https://doi.org/10.1101/gr.222976.117 PMID: 29440222

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

32. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014; 47:11.12.1–34. https://doi.org/10.1002/0471250953.bi1112s47 PMID: 25199790

33. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

34. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

35. Mattick J, Libro S, Sparklin BC, Chung M, Bromley RE, Nadendla S, et al. Nearly Complete Genome Sequence of Brugia pahangi FR3. Microbiol Resour Announc. 2020; 9. https://doi.org/10.1128/MRA.00479-20 PMID: 32616635

36. International Helminth Genomes Consortium. Comparative genomics of the major parasitic worms. Nat Genet. 2019; 51:163–174. https://doi.org/10.1038/s41588-018-0262-1 PMID: 30397333

37. elegans Sequencing Consortium C. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998; 282:2012–2018. https://doi.org/10.1126/science.282.5396.2012 PMID: 9851916

38. Wang J, Gao S, Mostovoy Y, Kang Y, Zagoskin M, Sun Y, et al. Comparative genome analysis of programmed DNA elimination in nematodes. Genome Res. 2017; 27:2001–2014. https://doi.org/10.1101/gr.225730.117 PMID: 29118011

39. Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, et al. The genome of Onchocerca volvulus, agent of river blindness. Nat Microbiol. 2016; 2:16216. https://doi.org/10.1038/nmicrobiol.2016.216 PMID: 27869790

40. Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, et al. The genomic basis of parasitism in the Strongyloides clade of nematodes. Nat Genet. 2016; 48:299–307. https://doi.org/10.1038/ng.3495 PMID: 26829753

41. Zhu X-Q, Korhonen PK, Cai H, Young ND, Nejsum P, von Samson-Himmelstjerna G, et al. Genetic blueprint of the zoonotic pathogen Toxocara canis. Nature Communications. 2015. https://doi.org/10.1038/ncomms7145 PMID: 25649139

42. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

43. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012; 28:1919–1920. https://doi.org/10.1093/bioinformatics/bts277 PMID: 22576172

44. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018; 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191 PMID: 29750242

45. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003; 31:3784–3788. https://doi.org/10.1093/nar/gkg563 PMID: 12824418

46. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 2010; 38:D463–7. https://doi.org/10.1093/nar/gkp952 PMID: 19910365

47. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016; 17:66. https://doi.org/10.1186/s13059-016-0924-1 PMID: 27072794

48. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. 2016.

49. Chung M, Teigen L, Liu H, Libro S, Shetty A, Kumar N, et al. Targeted enrichment outperforms other enrichment techniques and enables more multi-species RNA-Seq analyses. Sci Rep. 2018; 8:13377. https://doi.org/10.1038/s41598-018-31420-7 PMID: 30190541

50. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019; 37:907–915. https://doi.org/10.1038/s41587-019-0201-4 PMID: 31375807

51. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019; 20:278. https://doi.org/10.1186/s13059-019-1910-1 PMID: 31842956

52. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017; 35:316–319. https://doi.org/10.1038/nbt.3820 PMID: 28398311

53. Chung M, Teigen L, Libro S, Bromley RE, Kumar N, Sadzewicz L, et al. Multispecies Transcriptomics Data Set of Brugia malayi, Its Wolbachia Endosymbiont wBm, and Aedes aegypti across the B. malayi Life Cycle. Microbiol Resour Announc. 2018; 7. https://doi.org/10.1128/MRA.01306-18 PMID: 30533772

54. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14:178–192. https://doi.org/10.1093/bib/bbs017 PMID: 22517427

55. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol. 2019; 15:e1006790. https://doi.org/10.1371/journal.pcbi.1006790 PMID: 30726205

56. Liu C, Oliveira A, Chauhan C, Ghedin E, Unnasch TR. Functional analysis of putative operons in Brugia malayi. Int J Parasitol. 2010; 40:63–71. https://doi.org/10.1016/j.ijpara.2009.07.001 PMID: 19631652

57. Guiliano DB, Blaxter ML. Operon conservation and the evolution of trans-splicing in the phylum Nematoda. PLoS Genet. 2006; 2:e198. https://doi.org/10.1371/journal.pgen.0020198 PMID: 17121468

58. Blumenthal T, Gleason KS. Caenorhabditis elegans operons: form and function. Nat Rev Genet. 2003; 4:112–120. https://doi.org/10.1038/nrg995 PMID: 12560808

59. Graber JH, Salisbury J, Hutchins LN, Blumenthal T. C. elegans sequences that control trans-splicing and operon pre-mRNA processing. RNA. 2007; 13:1409–1426. https://doi.org/10.1261/rna.596707 PMID: 17630324

60. Lercher MJ, Blumenthal T, Hurst LD. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. 2003; 13:238–243. https://doi.org/10.1101/gr.553803 PMID: 12566401

61. Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. Nat Methods. 2019; 16:879–886. https://doi.org/10.1038/s41592-019-0503-y PMID: 31384046

62. Lima SA, Chipman LB, Nicholson AL, Chen Y-H, Yee BA, Yeo GW, et al. Short poly(A) tails are a conserved feature of highly expressed genes. Nat Struct Mol Biol. 2017; 24:1057–1063. https://doi.org/10.1038/nsmb.3499 PMID: 29106412

63. Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. The full-length transcriptome of C. elegans using direct RNA sequencing. Genome Res. 2020; 30:299–312. https://doi.org/10.1101/gr.251314.119 PMID: 32024661

64. Danckwardt S, Kaufmann I, Gentzel M, Foerstner KU, Gantzert A-S, Gehring NH, et al. Splicing factors stimulate polyadenylation via USEs at non-canonical 3' end formation signals. EMBO J. 2007; 26:2658–2669. https://doi.org/10.1038/sj.emboj.7601699 PMID: 17464285

65. Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. 2020; 30:214–226. https://doi.org/10.1101/gr.247494.118 PMID: 31992613

66. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

67. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007; 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071 PMID: 17332020

68. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017; 18:323. https://doi.org/10.1186/s12864-017-3691-9 PMID: 28438136

69. Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. PLoS One. 2016; 11:e0157779. https://doi.org/10.1371/journal.pone.0157779 PMID: 27327613