Published in final edited form as:

Int J Parasitol. 2018 July ; 48(8): 585–590. doi:10.1016/j.ijpara.2018.02.001.

Refined ab initio gene predictions of *Heterorhabditis bacteriophora* using RNA-seq*

Jonathan Vadnal^a, Olivia G. Granger^a, Ramesh Ratnappan^a, Ioannis Eleftherianos^{b,*}, Damien M. O'Halloran^{b,c,*}, and John M. Hawdon^{a,*}

^aDepartment of Microbiology Immunology and Tropical Medicine, George Washington University Medical Center, Washington, DC 20037, USA

^bDepartment of Biological Sciences, George Washington University, Science and Engineering Hall, suite 6000, 800 22nd Street NW, Washington DC 20052, USA

^cInstitute for Neuroscience, George Washington University, 636A Ross Hall, 2300 I Street NW, Washington DC 20052, USA

Abstract

Interest has recently grown in developing the entomopathogenic nematode *Heterorhabditis bacteriophora* as a model to genetically dissect the process of parasitic infection. Despite the availability of a full genome assembly, there is substantial variation in gene model accuracy. Here, a methodology is presented for leveraging RNA-seq evidence to generate improved annotations using ab initio gene prediction software. After alignment of reads and subsequent generation of a RNA-seq supported annotation, the new gene prediction models were verified on a selection of genes by comparison with sequenced 5' and 3' rapid amplification of cDNA ends (RACE) products. By utilizing a whole transcriptome for genome annotation, the current reference annotation was enriched, demonstrating the importance of coupling transcriptional data with genome assemblies.

Graphical abstract

*Note: Nucleotide sequence data reported in this paper are available in NCBI GenBank under accession numbers **MG518372** (*Hba-cpf-1*), **MG788341** (*Hba-dpy-13*), and **MG518373** (*Hba-pcp-1*).

*Co-corresponding authors. Ioannis Eleftherianos (ioannise@gwu.edu); Damien M. O'Halloran (damienoh@gwu.edu); John M. Hawdon (jhawdon@gwu.edu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Nematode; *Heterorhabditis bacteriophora*; Gene modeling; RNA-seq; Gene annotation

Parasitic nematode infection places a considerable burden on both human health and agricultural production by creating a variety of unfavorable conditions ranging from lethargy and fatigue to lethality. Despite advances in the study of innate immunity, the current understanding of the host response to parasitic nematode infections is limited (Erb, 2009; Cooper and Eleftherianos, 2016). The obligate requirement for a host makes studies in vertebrate nematode parasites very difficult and costly. Considerable advances in understanding nematode biology have been made using the model organism *Caenorhabditis elegans*, but its free living lifestyle limits its use as a model for parasitic nematode infections. For this reason, there is growing interest in developing entomopathogenic nematodes as model organisms to study the molecular basis of parasitism. One of these nematodes is the insect-specific parasite *Heterorhabditis bacteriophora*. Similar to the vertebrate hookworms *Ancylostoma caninum* and *Necator americanus*, *H. bacteriophora* is a member of the *Eurhabditis* clade. Carrying a mutualistic bacteria, *Photorhabdus luminescens*, *H. bacteriophora* infective juveniles (IJs) infect insect hosts, where they reproduce (Ciche et al., 2006). The free-living, developmentally arrested IJ stage is analogous to the infective stages of hookworms and the dauer stage of *C. elegans* (Hotez et al., 1993; Crook, 2014). Once established in a host, the IJs resume their development and progress through their complete life cycle, using the host and *P. luminescens* as a food source. After two to three generations of reproduction, the resources of the insect host's cadaver are exhausted and juveniles begin to arrest as IJs. These IJs leave the cadaver and disperse into the surrounding environment searching for a new host.

Unlike most other parasitic nematodes, including hookworms, *H. bacteriophora* and its symbiont *P. luminescens* are amenable to manipulations and maintenance in vitro (Somvanshi et al., 2012). Additionally, insect hosts are easy to obtain and manipulate for in vivo culturing of *H. bacteriophora* (Han and Ehlers, 2000; Castillo et al., 2012). The genome of *H. bacteriophora* has been published, and advanced molecular tools such as gene silencing by RNA interference (RNAi) have been developed, making *H. bacteriophora* a potentially excellent candidate as a model of nematode parasitism (Bai et al., 2013; Ratnappan et al., 2016). Despite these recent advances, information on its genes is sparse, with gene models and annotation largely based on orthologous proteins. Necessary techniques such as cloning, promoter identification, gene reporter constructs and protein expression are difficult without a more robust annotation. In order to dissect the molecular mechanisms of infection in *H. bacteriophora*, a more precise annotation is necessary to clearly identify uninformative genes versus novel parasitic candidate genes. In an attempt to improve the gene annotation for *H. bacteriophora*, we leveraged a relatively small RNA-seq dataset with ab initio gene prediction software to produce a more accurate and precise series of gene predictions.

The *H. bacteriophora* life cycle was maintained as previously described (Vadnal et al., 2017). Total RNA was extracted from *H. bacteriophora* IJs incubated in *Manduca sexta* hemolymph

plasma or vehicle and used for RNA-Seq on the Illumina HiSeq4000 platform with 150 bp, paired-end sequencing, as previously described (BioProject ID PRJNA418589, Vadnal et al., 2017). Reads were trimmed to remove Illumina adapter sequences, leading and trailing bases and low quality reads (Phred-64 score < 15) using Trimmomatic v0.33 (Bolger et al., 2014). Resulting reads 25 bp or longer were retained and sequence corrected using Rcorrector v1.0.2 (Song and Florea, 2015). Corrected reads were then aligned to the published *H. bacteriophora* reference genome and annotations (**PRJNA13977, WBPS9**) using a two-pass methodology with STAR v2.5.1b (Bai et al., 2013; Dobin et al., 2013; Howe et al., 2016). The first alignment pass aligned reads to the genome without preference to junctions identified by the published annotations. A second alignment procedure was then performed using the spliced junctions identified during the first pass alignment. The resulting SAM files for each sample ($n = 6$) were then converted and merged into a single BAM file for use as extrinsic evidence for ab initio gene prediction.

In order to reduce prediction errors, a repeat library was created to mask transposable elements and repeats within the *H. bacteriophora* genome. The methodology used was based on the methodology published by Campbell et al. (2014). A brief summary of the methodology with any deviations from the published methodology are outlined below. Miniature inverted repeat transposable element (MITE) sequences were captured using MITE-Hunter (November 2011 release) using default settings and the reference *H. bacteriophora* genome as input (Han and Wessler, 2010). Both relatively new and old long terminal repeats (LTRs) were discovered (LTR length between 100 and 6000 bp) and screened using GenomeTools v1.5.10 LTRharvest and LTRdigest tools (Gremme et al., 2013). Discovered LTRs were further screened for false positives using CRL Scripts v1.0 and a series of BLAST searches (BLASTX, BLASTN) against transposase databases (Camacho et al., 2009; Campbell et al., 2014). The CRL scripts are a set of Perl scripts designed to filter for eukaryotic tRNA sequences, LTR boundaries, nested insertions, and to finally create a set of exemplary LTRs. The discovered MITEs and LTRs were then used to mask the *H. bacteriophora* genome using RepeatMasker v4.0.7 with the slow search option on (<http://www.repeatmasker.org/>). Repetitive sequences were then collected using RepeatModeler v1.0.11 and the MITE/LTR masked genome (<http://www.repeatmasker.org/RepeatModeler/>). The MITE, LTR and RepeatModeler repeats were screened for gene fragments using BLASTX and a protein library generated from NCBI RefSeq proteins (www.ncbi.nlm.nih.gov/refseq/) identified as belonging to phylum Nematoda and the UniProt Swiss-Prot Invertebrate protein database (<http://www.uniprot.org/downloads>) (O'Leary et al., 2016; The UniProt Consortium, 2017). Only proteins which were supported by expressed sequence tag (EST) sequences identified as Nematoda deposited on NCBI dbEST (www.ncbi.nlm.nih.gov/dbEST/) (TBLASTN e-value < 0.00001) but not identified as being transposases (BLASTP e-value > 0.00001) were included in the library. The resulting repeat libraries were combined and used to soft-mask the *H. bacteriophora* genome using RepeatMasker with slow search.

In order to further increase evidence for gene predictions, a set of protein hints was created using BLAST search results and phylum Nematoda proteins gathered from NCBI RefSeq and the UniProt Swiss-Prot Invertebrate protein database (O'Leary et al., 2016; The UniProt

Consortium, 2017). Possible protein hints were first screened for support by EST sequences by performing a TBLASTN search against Nematoda sequences gathered from the NCBI dbEST (e-value < 0.00001). The filtered proteins were then searched against the *H. bacteriophora* genome using TBLASTN. Due to the possible cross-genus/species alignment of proteins, a stringent filter was used. Only protein hits with an e-value < 1E-50, percent query identity > 70%, and a percent query coverage > 15.625% were used as hints for BRAKER (<http://bioinf.uni-greifswald.de/bioinf/braker/>). The percent query coverage cut-off value was based on the average number of exons making up a *C. elegans* gene (~6.4 exons per gene) (Spieth et al., 2005). After screening, 3373 proteins in total were identified as suitable for use as protein hints.

The previously described merged BAM file, soft-masked genome, and filtered nematode protein library were passed to BRAKER v2.0.1 as inputs for new gene predictions (Hoff et al., 2016). BRAKER was run with the options for conditional random field training (--crf), soft-masking (--softmasking 1), and GFF3 output (--gff3) turned on. Spaln v2.3.0a was used as the aligner for protein hints (--prg=spaln) against the genome (Iwata and Gotoh, 2012). The final GFF3 output of BRAKER was compared with the reference annotations using gffcompare v0.10.1 (Trapnell et al., 2010). Differences between the final output and reference annotations of a selection of candidate infection genes were visualized with Integrative Genomics Viewer v2.3.90 (Getz et al., 2011; Thorvaldsdóttir et al., 2013). Alignments of reference, prediction and rapid amplification of cDNA ends (RACE) sequences were generated with ggbio (Yin et al., 2012) and Inkscape (www.inkscape.org).

To perform RACE on genes of interest, gene-specific primers were designed to target regions containing conserved protein domains identified using the NCBI CD-Search (Marchler-Bauer and Bryant, 2004). Primers were designed to target conserved regions using the Primer3Plus Cloning task with lengths between 23 – 28 nucleotides and predicted melting temperatures > 70°C (Untergasser et al., 2007). Primers were checked for specificity by standard PCR using *H. bacteriophora* IJ cDNA as a template and running the products on a 1.0% agarose gel to confirm band size, followed by sequencing. The Clontech (USA) SMARTer RACE 5'/3' Kit, using the manufacturer's guidelines for total RNA samples, was used for 5' and 3' RACE. Briefly, total RNA (the same samples used for RNA-seq) was converted to either 5'-RACE-Ready or 3'-RACE-Ready cDNA. The 5' and 3' ends were amplified using touchdown PCR with either the 3' or 5' gene-specific primer, respectively, and a universal primer mix. The gene-specific primers identified as above were modified to contain 5' overlap sequences given by the manufacturer for subsequent cloning into a vector. The amplified RACE products were separated on a 1.2% agarose gel, excised, purified and subsequently ligated into a sequencing vector. The resulting vector was transformed into Clontech Stellar Competent Cells and screened for vector and insert via ampicillin (100 µg/mL) selection on lysogeny broth (LB) agar plates. Positive colonies were picked, screened for the insert via PCR with gene-specific primers, and grown overnight in LB supplemented with ampicillin (100 µg/mL). Plasmids were harvested using the Promega (USA) Wizard Plus SV Miniprep DNA Purification System as per the manufacturer's protocol. Purified plasmid was used as template for PCR with M13(–40) and M13 Reverse primers. PCR products were then sent to Eurofins Genomics (USA) for sequencing with M13(–21) Forward and M13 Reverse primers. The 5' and 3' end sequences were analyzed,

assembled, and joined into consensus sequences using the Staden Package v2.0.0b11-2016 (Bonfield et al., 1995). Consensus RACE products were aligned to genomic DNA and visualized using Splign (Kapustin et al., 2008). To determine if the RNA-seq supported annotations were only predicting splice variants for a single life stage (IJ), products were further validated by amplifying and sequencing full-length genes from L4 stage cDNA. Additionally, *Hba-dpy-13* was validated in adult stages. The RNA-seq supported annotation was compared with the reference annotation using the gffcompare program (Table 1). A total of 15,597 mRNAs were contained in the RNA-seq supported annotation compared with 20,964 mRNAs in the reference annotation. Of the 15,597 mRNAs, 15,090 (96.7%) were identified as being multi-exon transcripts (versus 19,236 for the reference). The numbers of novel and missing exons and introns are also reported in Table 1.

In order to validate the newly predicted annotation, a BUSCO (Benchmarking Universal Single-Copy Orthologs; <http://busco.ezlab.org/>) assessment of proteins predicted from the RNA-seq annotation was performed using the nematode lineage set (Simao et al., 2015). Out of a database of 982 BUSCO groups searched, 909 (92.6%) orthologs were identified as being complete based on the RNA-seq annotation. Of these, 909, 709 (72.2%) were identified as being single-copy BUSCOs and 200 (20.4%) were identified as being duplicated. Of the remaining 73 groups, 52 were identified as being fragmented and 21 were missing. In comparison, BUSCO assessment of the proteins predicted by the reference annotation contains 308 (31.4%) complete BUSCOs with 674 (68.6%) fragmented and missing, suggesting an improvement with the RNA-seq annotation prediction of near-universal, single-copy orthologs. To verify the accuracy of the RNA-seq supported annotation and the reference annotation on a per gene level, 5' and 3' RACE was performed on three genes, *Hba_15308*, *Hba_06590* and *Hba_13000* (Fig. 1). *Hba_15308* and *Hba_13000* were chosen due to being identified as candidate genes involved in parasitic nematode infections (Vadnal et al., 2017). *Hba_06590* was chosen due to earlier work within our laboratory examining a dumpy phenotype produced by microinjection of double-stranded (ds)RNA in *H. bacteriophora* (Ratnappan et al., 2016).

The reference annotation identified *Hba_15308* as being composed of seven exons and six introns, while the RNA-seq supported annotation predicted three additional exons (E1, E7 and E9 in Fig. 1A), for a total of 10 exons and nine introns. Additionally, several of the newly predicted exons differed in length from the reference annotation (Fig. 1A). Exons E4 and E8 were truncated by approximately 24 nucleotides (eight amino acids) and 67 nucleotides (22 amino acids), respectively, while exon E5 was determined to have a longer 3' end of approximately 111 nucleotides (37 amino acids), compared with the reference annotation. While the addition of exon E1 appears to have no functional consequence, the novel exons E7 and E9 contain conserved protein motifs for the peptidase C1 superfamily (cl23744). The increased coding region of exon E5 also contains a conserved motif for the inhibitor I29 superfamily (cl22448). These novel and revised exons suggest key functional segments of protein were missed upon the initial annotation. Sequencing and alignment of the RACE products of *Hba_15308* matched the number and length of exons predicted by the RNA-seq supported annotation, and identified a 27 bp 5'-untranslated region (UTR) and 129 bp 3'-UTR, with a polyA signal sequence identified within the last 20 bp of the 3'-UTR. Due to its orthology with the *C. elegans* cathepsin W/cathepsin F protein R07E3.1

(WormBase ID WP:CE39608; www.wormbase.org), the sequence was submitted to GenBank with the gene name *Hba-cpf-1* (GenBank accession number **MG518372**).

Based on the reference annotation, *Hba_06590* was identified as a nematode collagen candidate composed of 10 exons spanning an approximately 6000 bp region (Fig. 1B). The RNA-seq annotation suggests only the first three exons of the reference *Hba_06590* make up the entirety of the gene over a 1400 bp span (Fig. 1B). Sequencing of RACE PCR products confirmed this prediction. While the first exon of the RACE products, E1, is conserved across both annotations, E2 and E3 differ from the reference annotation. The 5' end of exon E2 is extended 484 bp to a total of 603 bp (201 amino acids) compared with the 119 bp of the reference. Exon E3 is extended in the 3' direction from 137 bp to 172 bp (35 bp difference) within the RACE products and RNA-seq annotation. Additionally, RACE sequencing identified a 10 bp 5' UTR and 58 bp 3' UTR. Conserved protein domains for nematode cuticle collagen N-terminal domain (smart01088) and collagen triple helix repeat (pfam01391) are found in exons E1 and E3, respectively. No conserved protein domains were identified in exon E2. The protein product of this RACE sequence shares orthology with the *C. elegans* DPY-13 protein (WormBase ID WP:CE04462) and was submitted to GenBank with the gene name *Hba-dpy-13* (GenBank accession number **MG788341**).

For *Hba_13000*, the reference annotation predicted a gene composed of three exons and two introns (Fig. 1C). In contrast, the RNA-seq supported annotation predicted a composition of 14 exons with some variations in exon size spanning not only the reference genes *Hba_13000* and *Hba_13002*. Alignment of the products of the 5' and 3' RACE PCR confirmed *Hba_13000* is actually a single large gene composed of *Hba_13000* and *Hba_13002*, separated by an approximately 1700 bp intron. The RACE sequence indicates that the full-length *Hba_13002-13000* is composed of a total of 14 exons (equivalent to those predicted by the RNA-seq annotation), whereas the combined *Hba_13002-13000* hybrid from the reference annotation yields a 10 exon gene. The newly identified exons in the hybrid gene are E7, E12, E13 and E14 (Fig. 1C). Similar to the novel exons and coding regions predicted in *Hba-cpf-1*, the novel exons of the hybrid gene contain conserved domains for alpha/beta hydrolase superfamily (cl21494), the superfamily for peptidase S28, again suggesting important functional components of the gene product were missed in the reference annotation. A 4 bp 5'-UTR and 116 bp 3'-UTR were also predicted from RACE products. RACE sequence data also showed a small truncation in exon E6 of approximately 58 nucleotides (19 amino acids), compared with the reference, and a truncation in E8 of 18 nucleotides (six amino acids) compared with the reference annotation. A small increase (48 nucleotides, 16 amino acids) in length of exon E9 (Fig. 1C) was also demonstrated. Due to its orthology with the *C. elegans* predicted carboxypeptidase protein F56F10.1 (WormBase ID WP:CE34034), the RACE generated sequence was submitted to GenBank with the gene name *Hba-pcp-1* (GenBank accession number **MG518373**).

Here we report using RNA-seq alignments, a soft-masked reference genome, a set of protein hints, and ab initio gene prediction software to produce a revised annotation for *H. bacteriophora* genes. A number of novel exons and introns were identified by the evidence-supported annotation, as well as a reduction in the fragmentation of transcripts/genes and a sizeable increase in BUSCO protein coverage.

Exon and intron predictions were verified by performing RACE PCR on three genes. While BRAKER has been previously demonstrated as a way to improve genome annotations on model organisms, to our knowledge this is the first time BRAKER has been used to improve a genome annotation for a less well described organism using a relatively small RNA-seq data set from a single developmental stage (Hoff et al., 2016). While the BRAKER pipeline can produce an improved annotation, care must be taken when setting options and providing inputs. Initial BRAKER predictions using unmasked genome sequences predicted *Hba-pcp-1* as two separate genes (similar to the reference annotation) with similar exon structure to the RACE products. However, inclusion of a soft-masked genome sequence removed this errant prediction. While *Hba-pcp-1* is transcribed on the reverse strand, the 1700 bp intronic region separating E8 and E9 contains a possibly errantly predicted (within the reference annotation) TRD-1 repeat/gene on the forward strand. Masking of this inter-exonic region likely allows the prediction software to accurately connect exons E8 and E9 as being part of the same coding sequence.

The refined annotation for *H. bacteriophora* presented here demonstrates the utility of using relatively small amounts of empirical evidence to further enrich and define gene models. Additionally, automated modeling software with relatively low computational overhead can provide improvements in accuracy and precision of predictions. While further experimental data is needed to fully confirm the new gene models throughout all life stages of *H. bacteriophora*, the current prediction provides an immediate improvement to move forward with dissection of the functions of genes. While this RNA-seq supported annotation offers improved gene predictions, it should still only be considered as a single step in an iterative process of improvement and a starting point for fully studying other genes of interest. As this resource and others are utilized to study *H. bacteriophora*, more gene structure evidence will be generated and the annotation can be further improved. In order to facilitate further research of *H. bacteriophora*, the refined annotation GFF/GTF, transcript sequences, and protein sequences, as well as any updates, are freely available at the following GitHub repository: <https://github.com/dohalloran/Heterorhabditis-bacteriophora-Re-annotation>.

Acknowledgments

The project was supported by grant 5R21AI109517 from the United States National Institute of Allergy and Infectious Diseases (NIAID). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the National Institutes of Health (USA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Bai X, Adams BJ, Ciche TA, Clifton S, Gaugler R, Kim K, Spieth J, Sternberg PW, Wilson RK, Grewal PS. A lover and a fighter: The genome sequence of an entomopathogenic nematode *Heterorhabditis bacteriophora*. PLOS One. 2013; 8:e69618. [PubMed: 23874975]
- Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics. 2014; 30:2114. [PubMed: 24695404]
- Bonfield JK, Smith Kf, Staden R. A new DNA sequence assembly program. Nucleic Acids Res. 1995; 23:4992–4999. [PubMed: 8559656]
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: Architecture and applications. BMC Bioinform. 2009; 10:421.

- Campbell MS, Law MeiYee, Holt Carson, Stein Joshua C, Moghe Gaurav D, Hufnagel David E, Lei Jikai, Achawanantakun Rujira, Jiao Dian, Lawrence Carolyn J, Ware Doreen, Shiu Shin-Han, Childs Kevin L, Sun Yanni, Jiang Ning, Yandell Mark. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* 2014; 164:513–524. [PubMed: 24306534]
- Castillo JC, Shokal U, Eleftherianos I. A novel method for infecting *Drosophila* adult flies with insect pathogenic nematodes. *Virulence.* 2012; 3:339–347. [PubMed: 22546901]
- Ciche TA, Darby C, Ehlers R, Forst S, Goodrich-Blair H. Dangerous liaisons: The symbiosis of entomopathogenic nematodes and bacteria. *Biol Control.* 2006; 38:22–46.
- Cooper D, Eleftherianos I. Parasitic nematode immunomodulatory strategies: Recent advances and perspectives. *Pathogens.* 2016; 5:58.
- Crook M. The dauer hypothesis and the evolution of parasitism: 20 years on and still going strong. *Int J Parasitol.* 2014; 44:1–8. [PubMed: 24095839]
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013; 29:15–21. [PubMed: 23104886]
- Erb KJ. Can helminths or helminth-derived products be used in humans to prevent or treat allergic diseases? *Trends Immunol.* 2009; 30:75–82. [PubMed: 19138565]
- Getz G, Lander ES, Guttman M, Thorvaldsdóttir H, Winckler W, Mesirov JP, Robinson JT. Integrative genomics viewer. *Nat Biotechnol.* 2011; 29:24–26. [PubMed: 21221095]
- Gremme G, Steinbiss S, Kurtz S. GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform.* 2013; 10:645–656. [PubMed: 24091398]
- Han R, Ehlers R. Pathogenicity, development, and reproduction of *Heterorhabditis bacteriophora* and *Steinernema carpocapsae* under axenic in vivo conditions. *J Invertebr Pathol.* 2000; 75:55–58. [PubMed: 10631058]
- Han Y, Wessler SR. MITE-hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 2010; 38:e199. [PubMed: 20880995]
- Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics.* 2016; 32:767. [PubMed: 26559507]
- Hotez P, Hawdon J, Schad GA. Hookworm larval infectivity, arrest and amphiparatensis: The *Caenorhabditis elegans daf-c* paradigm. *Parasitol Today.* 1993; 9:23–26. [PubMed: 15463660]
- Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite – a comprehensive resource for helminth genomics. *Mol Biochem Parasitol.* 2016; 215:2–10. [PubMed: 27899279]
- Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of spaln that incorporates additional species-specific features. *Nucleic Acids Res.* 2012; 40:e161. [PubMed: 22848105]
- Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: Algorithms for computing spliced alignments with identification of paralogs. *Biol Direct.* 2008; 3:20. [PubMed: 18495041]
- Marchler-Bauer A, Bryant SH. CD-search: Protein domain annotations on the fly. *Nucleic Acids Res.* 2004; 32:W331.
- O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O’Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016; 44:D745.
- Ratnappan R, Vadnal J, Keaney M, Eleftherianos I, O’Halloran D, Hawdon JM. RNAi-mediated gene knockdown by microinjection in the model entomopathogenic nematode *Heterorhabditis bacteriophora*. *Parasit Vectors.* 2016; 9:160. [PubMed: 26993791]

- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* (Oxford, England). 2015; 31:3210–3212.
- Somvanshi VS, Sloup RE, Crawford JM, Martin AR, Heidt AJ, Kim K, Clardy J, Ciche TA. A single promoter inversion switches *Photorhabdus* between pathogenic and mutualistic states. *Science*. 2012; 337:88. [PubMed: 22767929]
- Song L, Florea L. Rcorrector: Efficient and accurate error correction for illumina RNA-seq reads. *GigaScience*. 2015; 4:48. [PubMed: 26500767]
- Spieth, J., Lawson, D., Davis, P. WormBook: The online review of *C. elegans* biology. Wormbook; Pasadena, CA, USA: 2005. Overview of gene structure in *C. elegans*.
- The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res*. 2017; 45:D169.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*. 2013; 14:178–192. [PubMed: 22517427]
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010; 28:511–515. [PubMed: 20436464]
- Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res*. 2007; 35:W74.
- Vadnal J, Ratnappan R, Keaney M, Kenney E, Eleftherianos I, O'Halloran D, Hawdon JM. Identification of candidate infection genes from the model entomopathogenic nematode *Heterorhabditis bacteriophora*. *BMC Genomics*. 2017; 18:8. [PubMed: 28049427]
- Vishal S, Yin T, Cook D, Lawrence M. Ggbio: An R package for extending the grammar of graphics for genomic data. *Genome Biol*. 2012; 13:R77. [PubMed: 22937822]

Highlights

- Using RNA-seq alignments and ab initio gene prediction software we report a re-annotation for *Heterorhabditis bacteriophora*
- Novel exons and introns identified by our evidence supported re-annotation
- We demonstrated that a relatively small RNA-sequencing data set can be leveraged to improve genome annotation

Fig. 1.

Alignment of reference annotation, RNA-seq supported annotation, and sequenced rapid amplification of cDNA ends (RACE) products for *Hba-cpf-1* (A), *Hba-dpy-13* (B), and *Hba-pcp-1* (C). RACE (5' and 3') was performed on the *Heterorhabditis bacteriophora* reference genes *Hba_15308*, *Hba_06590* and *Hba_13000*, and the products were sequenced. The reference annotation is shown in white, the RNA-seq supported annotation in blue (grey), and the sequenced RACE product in purple (dark grey). Thick colored rectangles represent coding sequence while thin colored blocks represent untranslated regions (UTRs) predicted from sequenced RACE products. Thin black lines represent intronic regions with coding direction shown by arrows. Cross-hatched exons have identical nucleotide sequences among the three gene models. For C, the reference prediction of the gene model originally identified as *Hba_13002* is highlighted with a light blue (light grey) background while the predicted *Hba_13000* is not highlighted. Horizontal axes show nucleotide position on the titled scaffold.

Table 1

Summary of annotations. Summary of mRNAs, exons and introns predicted by the RNA-seq supported annotation, relative to the reference annotation for *Heterorhabditis bacteriophora*. Comparison between annotations was made using the gffcompare program. Removed exons and introns are those not present in the RNA-seq supported annotation. Novel exons and introns are those not contained within the reference annotation.

	mRNA	Exons	Introns
Reference annotation	20,964	123,238	102,274
RNA-seq annotation	15,597	152,922	140,497
Changes between annotations:			
			Percent (%)
Exons	Removed	36,002	29.2
	Novel	63,604	41.6
Introns	Removed	17,970	17.6
	Novel	36,775	26.5