Improved structural annotation of protein-coding genes in the *Meloidogyne hapla* genome using RNA-Seq

Yuelong Guo¹, David McK Bird^{1,2}, and Dahlia M Nielsen^{1,*}

¹Bioinformatics Research Center; NC State University; Raleigh NC USA; ²Department of Plant Pathology; NC State University; Raleigh NC USA

Keywords: *Meloidogyne hapla*, annotation, NGS, RNA sequencing, trans-splicing

As high-throughput cDNA sequencing (RNA-Seq) is increasingly applied to hypothesis-driven biological studies, the prediction of protein coding genes based on these data are usurping strictly in silico approaches. Compared with computationally derived gene predictions, structural annotation is more accurate when based on biological evidence, particularly RNA-Seq data. Here, we refine the current genome annotation for the *Meloidogyne hapla* genome utilizing RNA-Seq data. Published structural annotation defines 14420 protein-coding genes in the *M. hapla* genome. Of these, 25% (3751) were found to exhibit some incongruence with RNA-Seq data. Manual annotation enabled these discrepancies to be resolved. Our analysis revealed 544 new gene models that were missing from the prior annotation. Additionally, 1457 transcribed regions were newly identified on the ends of as-yet-unjoined contigs. We also searched for trans-spliced leaders, and based on RNA-Seq data, identified genes that appear to be trans-spliced. Four 22-bp trans-spliced leaders were identified using our pipeline, including the known trans-spliced leader, which is the *M. hapla* ortholog of SL1. In silico predictions of trans-splicing were validated by comparison with earlier results derived from an independent cDNA library constructed to capture trans-spliced transcripts. The new annotation, which we term HapPep5, is publicly available at www.hapla.org.

Background

The first metazoan genome sequenced was that of the free-living nematode worm *Caenorhabditis elegans*.¹ Arguably, this is the only animal genome for which every base is known with high certainty. Because of their importance as parasites of humans, animals, and plants, draft sequences of a growing number of nematode genomes have been obtained.²⁻⁵ Included in this list are those of two root-knot nematodes (RKN), *Meloidogyne hapla*² and *Meloidogyne incognita*.³ *Meloidogyne* is a large genus of plant parasites that are distributed worldwide and cause substantial loss to agricultural production. Despite being members of the same genus, *M. incognita* and *M. hapla* exhibit substantial differences in their parasitic ability including host range and resistance breaking. Comparison of these diverse species is predicated on robust annotation of the reference genome.

The *M. hapla* genome is diploid, consistent with sexual reproduction. The sequence spans 56Mb on 16 chromosomes,^{6,7} and accounts for > 98% of the large number of available ESTs.² Empowering the sequence has been the development of a robust linkage map,⁷ which permits the mapping of Mendelian and quantitative traits, and because the map and sequence are tightly anchored, the isolation of candidate genes. Collectively, these

resources establish *M. hapla* as a model organism to study the mechanisms of plant parasitism. Further, *M. hapla* is an emerging model for studying cross-kingdom horizontal gene transfer.^{8,9}

By contrast, *M. incognita* has both a complex karyotype that exhibits variable aneuploidy, and a complex genome organization that complicates assembly. First published as spanning 86 Mb,³ it now appears that the final assembly may approach 140 Mb.¹⁰ It is widely assumed that the complexity of the *M. incognita* genome is related to the obligate asexual lifestyle of this species. However, a study of polymorphisms in *Meloidogyne* mtDNA concluded that interspecific crosses occasionally occurred although at a frequency too low to repeat in a laboratory setting.^{11,12} Using more comprehensive methodology, a compelling case for *M. incognita* being derived from an interspecific cross has been made;¹⁰ modern *M. incognita* appears to be an allo-hexaploid, currently undergoing chromosomal decay.^{10,13}

The current annotation of the *M. hapla* genome independently used the algorithms FgenesH and Glimmer, trained with EST data, to predict gene models in the genome assembly. Concordance of prediction was deemed evidence of a reliable gene model. To improve the reliability of gene predictions, the TimeLogic GeneDetective algorithm (Active Motif Inc.) was employed to yield the HapPep3 release. The algorithm

*Correspondence to: Dahlia M Nielsen; Email: dahlia@statgen.ncsu.edu

Submitted: 01/29/2014; Revised: 04/26/2014; Accepted: 05/07/2014; Published Online: 05/16/2014
<http://dx.doi.org/10.4161/worm.29158>

Table 1. Summary statistics for modified gene models

Number of gene models	HapPep 3 annotation	HapPep5 annotation
confirmed	2556	2556
partially confirmed	1321	1321
confirmed single exon	188	188
merged in HapPep5	1670	839
split in HapPep5	236	504
modified in HapPep5	1845	1845
novel in HapPep5	0	544
unchanged	6604	6604
total genes	14420	14401

reconstructs splice junctions in order to accurately align EST/ cDNAs, proteins, or uses Hidden Markov Models to genomic DNA to produce a graphical gene model.¹⁴ As the community has begun to use the *M. hapla* genome as a tool to study biology, it has become clear that some of the gene predictions remain either ambiguous or are incorrect. Redressing those deficiencies is the focus of this work.

As RNA-Seq is increasingly applied to biological studies, prediction of protein coding genes utilizing biological evidence, ideally full-length expressed mRNA sequences, has been made possible. Several studies improving genome annotation using RNA-Seq data have been published, e.g., references 15–18. Beyond simply predicting primary gene structure, annotation using RNA-Seq data has the potential to find non-coding transcripts, as well as alternative splicing and editing events.^{19,20} Important for nematode genomes, we also deduced trans-splicing events based on RNA-Seq data.

In many nematode species (perhaps all), trans-splicing is a mechanism in which a short (typically 22 nucleotide) leader is spliced to a primary transcript, thus creating the 5'-most exon of that gene. In *C. elegans*, this sequence, termed a “spliced leader” (SL), is transcribed as a capped transcript by RNA Polymerase III from a locus that is part of the 5S cluster. The 5' end of this sequence, destined to be the SL, is 22 residues, ending with a splice donor. This donor transcript resembles a half exon. Transcripts to be trans-spliced have a splice acceptor sequence without a more 5' splice donor motif. The standard cellular splicing machinery is able to splice the donor to the target mRNA acceptor, i.e., in trans.²¹

The role of trans-splicing is not fully explored. However, the discovery that many nematode genes reside in closely spaced gene clusters similar to bacterial operons points to one need. Such genes are transcribed together as a single poly-cistronic message, which must be disrupted to permit translational initiation, usually from the most 5' ATG (depending on the Kozak motif). In *C. elegans* this conversion from a polycistronic transcript to monocistronic mRNA able to be translated is realized by “internal” trans-splicing with SL2.²² Other SL variants have been discovered^{23,24} and it has been estimated that about 70% of *C. elegans* mRNAs are trans-spliced to one of two 22 nucleotide spliced

leaders.²⁵ A spliced leader similar to SL1 in *C. elegans* has been identified in genus *Meloidogyne*,^{3,26} and it is estimated at least 10% of *M. hapla* mRNAs are trans-spliced.^{27–30} However, to date, other SL, including those similar to SL2, had yet to be identified in *Meloidogyne*.

The main aim of this study was to improve the annotation of protein-coding genes in the *M. hapla* genome by incorporating Illumina RNA-Seq reads. To do this, RNA-Seq reads were assembled into transcripts, from which open reading frames were predicted. Annotation discrepancies between this approach and the current HapPep3 release were further arbitrated based on BLAST analysis, with HapPep3 models deemed to be incorrect only if additional evidence from BLAST showed that the newly predicted genes are more similar to a known protein. In addition, we applied a method for extracting novel trans-spliced leader sequences from high-throughput RNA-Seq data. For each identified spliced leader, genes that showed evidence of being trans-spliced based on the RNA-Seq data were recorded.

Results and Discussion

Annotation based on RNA-Seq evidence

Updates to version 3 annotation

We obtained Illumina RNA-Seq reads from 98 independent samples of RKN-infected *Medicago truncatula* plants. RNA-Seq reads were submitted to and stored in the NCBI SRA database (BioProject accession: PRJNA229407). Approximately 1 billion reads from the first 32 samples were used in this study, which provides sufficient coverage for gene structural annotation. We were able to map 149 million (15.2%) of these reads unambiguously to the *M. hapla* genome. The majority of the other mapped reads aligned to the host plant genome. Based on a reference-guided assembly of the *M. hapla* RNA-Seq reads, 16896 transcripts, including products of alternative splicing, were identified using Cufflinks.³¹ Open reading frames (ORF) beginning with ATG and ending with a stop codon were found in 16678 (98.7%) of transcripts. The current annotation (HapPep3) of the *M. hapla* genome predicts a total of 14420 protein-coding genes.¹⁴ As summarized in Table 1, comparison of the new annotation with the HapPep3 models fully confirmed 2556, and partially confirmed 1321 HapPep3 genes. Additionally, new annotation confirmed the expression of 188 single exon genes, although due to the presence of 5' and 3' untranslated regions in mRNA, the start and stop codon may not be precisely evaluated solely based on RNA-seq data. Split or merge gene model comparison as defined by the different number of gene at one locus were observed in 5343 HapPep3 genes. Three thousand three hundred and twenty-eight genes were merged into 1728 gene models at corresponding loci, while other 1136 genes were split into 2422 genes in the new annotation. Although the merge/split models were supported by RNA-Sequencing evidence, they may still represent incorrect predictions. For example, incomplete pre-mRNA-splicing of operons would lead to short reads spanning two genes in one operon, which would in turn cause the false prediction of a merged gene model. Therefore, the merge/split models were

Figure 1. RNA-Seq data reveal annotation discrepancies in HapPep3 gene models. Examples of updates to the HapPep3 annotation are shown in this figure, visualized by IGV.^{38,39} (A) Three HapPep3 genes were merged into one gene model; (B) one HapPep3 gene was split into two genes; (C) RNA-Seq predictions revise intron-exon boundaries for a few exons as predicted in HapPep3 annotation; (D) one or more missing exons were added to the HapPep3 annotation. Panels (E and F) show two additional examples of (A), with arrows indicating the strandedness of the corresponding gene. In these examples, the merged gene spans a long intron, with another gene located on the other strand within the intronic region.

further evaluated by BLAST as described in the methods section. Eight hundred and thirty-nine merged models (Fig. 1A, E, and F) and 236 splits (504 gene models, Fig. 1B) were deemed more plausible than the corresponding HapPep3 gene models. Moreover, 1845 new gene models that were in one-to-one correspondence to HapPep3 annotation were updated (Fig. 1C and D). We also found 544 gene models that were missing from the HapPep3 annotation (Table S1).

M. hapla samples used in this study were progeny of a cross between inbred strains VW9 (the reference strain) and LM. Because of this, the data used to improve gene structure annotation represents a mixture from the LM and VW9 strains. As these two strains are the same species, we do not expect large chromosomal differences between them, and any smaller-scale differences (e.g., SNPs or small indels) do not affect structural

annotation predictions. The annotation derived from this mixture represents a useful general reference resource for the species.

RNA-Seq evidence recovers expressed regions near ends of contigs

Although it is estimated that the *M. hapla* genome assembly spans more than 99% of the genome, the current version spans 3452 contigs (1523 super contigs) anchored to the map that have yet to be joined.² Genes that are close to the ends of contigs are often missed from the annotation using either protein prediction methods or methods based on hidden Markov models, as these methods rely heavily on information about the entire gene structure, including coordinates for the start and stop codon and intron-exon splicing boundaries. These transcribed regions, however, can be identified using RNA-Seq data. A total of 1457 expressed regions that lie on both ends of 3452 contigs were identified in the new annotation. Although it can be far from

Figure 2. Analysis flowchart. The protocol is divided into two parts: (i) structural annotation and (ii) trans-spliced leader identification.

Table 2. List of *M. hapla* Trans-splice leaders

<i>C. elegans</i> SL1	GGTTTAATTACCCAAGTTTGAG	refs. 21
<i>M. incognita</i> SL1	GGTTTAATTACCCAAGTTTAAG	refs. 26
<i>M. hapla</i> SL1	GGTTTAATTACCCAAGTTTAAG	refs. 28–30
<i>M. hapla</i> SL1.2	GGTTTAATTACCAATTGAG	This paper
<i>M. hapla</i> SL1.3	GGTTAATTACCCAAGTAAAAG	This paper
<i>M. hapla</i> SL1.4	TTAATACCCAAGCATTTGAGG	This paper

trivial to accurately deduce the start and stop codon from primary sequence (or sometimes even the true open reading frame), the identification of these expressed regions aids studies using this annotation, including looking for genes being differentially expressed. The updated annotation is publicly available through the project site.³² The current (January 2014) public version of *M. hapla* gene predictions, called “HapPep,” is version 3, although we use a partially revised draft (HapPep4) in house. Consequently, we named this public release HapPep5.

Mortazavi, et al. has shown scaffolding yet-to-be-joined contigs with cDNA,³³ because the ends of contigs often had apparent open reading frames, we attempted to assemble the 3452 contigs of *M. hapla* genome using RNA-Seq data by detecting mRNA reads that mapped to the ends of two contigs. We reasoned that if part of an expressed transcript is located at the end of one contig, with the remainder being located at the beginning of the contig next to it, we would anticipate finding RNA-Seq reads spanning the unclosed gap. However, only 282 pairs were identified using this method, 199 of which are supported with only one or two RNA-Seq reads. Thus, only a limited amount of information was obtained from RNA-Seq data on the closure of contigs and super-contigs.

Detection of trans-spliced leaders

We employed a two step approach in detecting trans-spliced leaders in *M. hapla*. First, plausible spliced leaders (pSL) were identified using a protocol to detect and filter RNA-Seq reads

with ends that map to two different genomic locations (Fig. 2; also see Methods). To verify these pSL, we extracted all RNA-Seq reads (from the entire set of reads) that contained these pSL sequences or their reverse complement sequences. In order to evaluate whether these pSL appeared to be trans-spliced to several genomic loci we used BLAST to align the reads that contained these pSL to the *M. hapla* genome. It was observed that for six out of 10 identified pSL, more than 90% of the reads containing these pSL appeared to show evidence of cis-splicing instead of trans-splicing (the non-pSL portion of these reads mapped to genomic locations adjacent to that of the pSL). Therefore, these pSL were considered to be false positives. For the other four pSL (listed in Table 2), the non-pSL portion of their RNA-Seq reads mapped to distal genomic locations, as expected for trans-splicing. It was also observed that sequences concatenated to these pSL mapped to multiple genes (i.e., multiple genes were trans-spliced by the same pSL). These four pSL were therefore retained as trans-spliced leaders. Two of the identified SL differ by one base pair from SL1 in *C. elegans*, and one of these concurs with the known SL1 in *M. hapla*. The other two putative SL are novel SL1 variants. It was observed that transcripts trans-spliced by SL1 may also be trans-spliced by the identified SL1 variants, although trans-splicing by SL1 variants are less abundant. Moreover, instances have been found that SL1 trans-splice to genes on the same strand that are less than 500 bp apart, i.e., likely resolving polycistronic genes within an operon. Given that no SL2-like splice leaders were identified, *M. hapla* could be one of the species that uses SL1 for all trans-splicing, including trans-splicing to downstream operonic genes, as discussed in Guiliano, et al.³⁴

Genes that were determined, based on the RNA-Seq evidence, to encode transcripts that were subsequently trans-spliced were compared with sequences from an earlier SL-cDNA library that was derived independently.²⁸⁻³⁰ This cDNA library was constructed by PCR amplification of total RNA using a primer based on the known SL and an oligo(dT) primer, so that only

mRNA sequences with the known spliced leader were amplified. Sequencing was then performed from both ends using the same primers. Upon aligning to the nematode genome, a total of 1193 genes were determined to be trans-spliced based on the cDNA library evidence. In contrast, 585 genes were identified as being trans-spliced based on the combined RNA-Seq data of 32 independent samples of RKN-infected plant root tissue. Although nearly half (266 out of 585) of the genes identified through the RNA-Seq data coincided with ones identified through the cDNA library, we had anticipated a higher degree of match. One possibility stems from a feature of our pipeline in which the identification of trans-splicing based on RNA-Seq data requires the presence of a full-length SL (~22 bp). While the fragmented cDNA short reads are around 78 bp, the probability of retaining a full-length SL is low during fragmentation before sequencing. Thus trans-spliced products, although present in the library, are likely not recognizable, especially for mRNAs with low steady-state levels. Because of this, lowly expressed genes are likely to be missed in the identification. In reexamining our data, we did observe that the number of RNA-Seq instances of each trans-spliced gene was low (with an average of 2.185 times coverage). Further, we performed the same analysis with 18 additional independent samples (50 out of total 98 available), to test whether more trans-spliced genes may be detected with more samples and coverage. With these data, a total of 1270 genes were identified as being trans-spliced and 494 out of 1270 (38.9%) genes were the same as the ones obtained from the SL cDNA library. From these results, we concluded that increasing RNA-Seq sample size results in increasing the number of genes identified as being trans-spliced, and that the list of genes identified as being trans-spliced is not complete. While it is likely this list also contain false-positive results, we expect that many newly identified trans-spliced genes are novel true positives, and thus, that RNA-Seq data can provide insight into trans-splicing events.

Future directions

RNA-Seq data has been proven to be valuable for the identification of gene structures, including alternative splicing. In regions with low expression however, often little or no information can be obtained. In this study, with approximately 149 million reads (about 200 fold coverage on average combining all samples), we failed to identify 3373 genes (out of 16896, 19.96%) predicted in the HapPep3 annotation. Since all biological samples used in this study were in the same developmental stage, nematode genes that are not expressed 3 wk after inoculation will not be identified in our analysis. We expect that further improvements can be made in both *M. hapla* genome structural annotation and alternative-splicing identification by sequencing nematode transcriptomes sampled from different developmental stages.

Many nematode genes reside in closely spaced gene clusters similar to bacterial operons. Such genes are transcribed together as single poly-cistronic messages, which must be disrupted to permit translational initiation. One role of trans-splicing is that it creates mono-cistronic mRNAs that are translated separately. Although trans-spliced genes were successfully identified in this study, the low number of unclosed gaps in the existing *M. hapla* genome assembly creates difficulties for making further

inferences about operons (i.e., total number of operons, average number of genes per operons) in *M. hapla*. As genes in the same operon are in close proximity, trans-spliced genes that are physically near to each other are likely processed from one operon. Upon closing the 3452 contigs of the current *M. hapla* genome assembly, it will be worth repeating the trans-spliced analysis and extending the methods to allow identification of operons.

Conclusions

Genome annotation is a process of continuous refinement. Although our updated structural annotation approach utilizing RNA-Seq data are not perfect, it has improved the HapPep3 annotation. In the new assembly, 3751 protein-coding genes were updated and 544 genes were newly predicted. Moreover, in the process of identifying trans-spliced leaders, four SLs, including one known previously, were successfully identified. Further experimental evidence will be needed to confirm the other three SL identified in this study and to acquire a complete list of trans-spliced genes.

Materials and Methods

Experimental design

Medicago truncatula plants were grown in sand in separate containers in a growth chamber for 16 h at 22 °C under 200–300 $\mu\text{mol/s}\cdot\text{m}^{-2}$ of light (day) and for 6 h at 18 °C in the dark (night); humidity 60%. Plants were watered daily and fertilized three times a week with half-strength Hoaglands solution. Plants were inoculated with 2 000 freshly hatched *M. hapla* J2 per plant. Three weeks after inoculation, roots were gently removed from pots, and sand removed by washing. Small portions of root (several mm) containing RKN-induced galls were resected by hand and immediately placed on dry ice. To reduce random variation (and facilitate identification of real signals), we pooled gall samples from three plants grown separately and infected with the same nematode sample. This pool was considered to be one replicate. Statistical power was obtained by performing 32 biological replicates. Total RNA was extracted using QIAGEN's RNeasy Mini Kit (Qiagen) and examined on the Agilent Bioanalyzer 2100 (Agilent) for quality and concentration using a RNA Pico Chip. Libraries were prepared using Illumina's TruSeq RNA library prep kit v1 (Illumina). After quantification of the library, the fragments were annealed to a flow cell and clonally amplified using a cBot (Illumina). The amplified sequencing libraries were sequenced using Illumina's Genome Analyzer IIx sequencing platform as single reads for 68 cycles. Each sample was run on a separate lane. Single read runs of 76 nt were generated.

Reconstructing transcripts

A flowchart of the analysis protocol is shown in Figure 2. Reference-guided assembly was performed using Bowtie³⁵ and the spliced aligner TopHat³⁶ with minimum intron length set to 10 bp, and minimum anchor length set to 5 bp. The appropriate plant and nematode genome sequence files were merged and the concatenated file served as the reference genome sequence. Only reads that mapped unambiguously to the *M. hapla* genome were used for subsequent analysis. Mapped reads from 32 independent

Figure 3. A schematic representation of the identification of trans-spliced leaders. RNA-Seq reads that remained unmapped following reference guided assembly were aligned to the *M. hapla* genome using BLAST. Trans-spliced genes were identified by locating reads for which one end aligned to one position in the genome, while the other end aligned to another genome position. Examples of such reads and their alignments are shown here. In each case, the DNA sequence of the reads (in fasta format) is displayed in purple, followed by the BLAST results. (A) An example of a BLAST alignment indicating a putative trans-spliced leader and trans-spliced gene at distant genome locations (i.e., Contig 2540 and 2912) is shown. Here, as expected, the position of the putative trans-spliced leader is immediately adjacent to the position of the gene. Due to the nature of the BLAST algorithm however, an identified alignment may overextend or under-extend the true alignment boundaries, creating the appearance of a gap or overlap between the portion of the read belonging to the trans-spliced leader and the gene. Shaded boxes represent nucleotides that may have been incorrectly aligned by BLAST algorithm. To account for this in our analysis, we allowed up to a 5nt overlap or gap between the putative trans-spliced leader and the gene. (B) An example in which the BLAST algorithm overextends the true alignment, with the 4nt overlapping nucleotides indicated by black rectangles. (C) An example in which the BLAST algorithm under-extends the true alignment, with the 2nt gap indicated by a black rectangle. (D) To reduce false positives, RNA-Seq reads with BLAST results showing more than a 5nt overlap or gap were discarded in the initial screening of trans-spliced leaders, although it is possible that these RNA-Seq reads did originate from trans-spliced genes.

lanes were merged for deeper coverage. Cufflinks³¹ was then used to assemble the results of TopHat into transcript assemblies, with minimal intron length set to be 10 bp due to the compact nature of the *M. hapla* genome. *getorf* in EMBOSS³⁷ was utilized to find ORFs in the predicted transcripts. New evidence of gene structure was compared with HapPep3 annotations if inconsistent prediction occurred at the same genome locus.

In case split or merge gene structures were observed, i.e., HapPep3 annotation and new evidence differs in number of genes at a same locus, BLAST analysis similar to Li, et al. was used to determine whether the split or the merge gene model is more plausible.¹⁶ If two or more of the split gene model mapped to same top hits, then the merged gene model is preferred. While two or more of the split gene model mapped to different top hits, the split gene model is preferred. However, if only one of the split gene has BLAST hits, further comparison was made to identify the most plausible model. If the BLAST hit of merged model was longer than that of the split model, with extension to the other split genes, the merged model was determined more plausible. Otherwise, the HapPep3 gene model was remained unchanged.

In cases where the new evidence was incongruence with the HapPep3 annotation, but no split or merge gene structure was observed, the HapPep3 annotation was determined to be incorrect only if additional evidence from BLAST showed that the newly predicted genes displayed higher sequence similarity to a known protein. Genes assembled by RNA-Seq evidence and the ones annotated in HapPep3 were both BLASTed to the UniProt database of all nematodes. If BLAST returned no significant hits for either predicted protein, the current HapPep3 annotation was left unchanged. Newly identified genes not included in the HapPep3 annotation were also BLASTed against this protein database. Only predicted transcripts with significant hits were kept and recorded as protein-coding genes. Expressed regions near the ends of not-yet-closed contigs were recorded, but not definitively designated as being protein-coding genes.

Detection of trans-splice leaders

To detect trans-splicing events, a data filter followed by a two-step identification and verification procedure was applied to the RNA-Seq data. At the outset, reads that remained unmapped after performing the reference-guided transcript assembly were BLASTed against both the plant and nematode genome, and the subset of reads that mapped to the plant genome were discarded. All remaining reads (those that were not included in the reference-guided transcript assembly and that did not align to the plant genome) were used for the following trans-splicing event identification and verification steps.

In the identification step, the goal was to identify plausible spliced leaders (pSL). This was accomplished by searching for RNA-Seq reads for which one end of the read aligned (via BLAST) to one genomic location, while the other end of the read aligned to another, distal location (Fig. 3). Only reads for which both ends aligned (1) exclusively to nematode genome and (2) to two distinct genome locations were retained. To filter out reads that might have mapped by random chance, only reads that displayed 100 percent identity were considered. Several additional stringent filters were then applied to further reduce random mapping: (1) the match length of one of the ends of each read, hypothetically the spliced-leader sequence, had to be 22 ± 5 nt (Fig. 3C–E), as all of the currently known spliced-leaders in nematodes are around 22 nt; and (2) the overlap between the two halves that mapped to different locations had to be less than 5 nt (Fig. 3E).

The second step of the trans-spliced detection procedure was verification. Each pSL sequence identified in the first step was used as a probe against the entire set of RNA-Seq reads (every RNA-Seq read containing this pSL sequence was identified via BLAST). Each of these reads was then aligned via BLAST to the nematode genome, and the location of the non-pSL portion of these reads was recorded. These recorded locations were then used to determine if the non-pSL component of the reads (the sequences that were presumably derived from a transcribed gene) tended to be nearby the location of the pSL (i.e., *cis*) or if they tended to be distal (i.e., *trans*). If $\geq 90\%$ of these reads aligned to a region adjacent to the location of the pSL, the pSL was considered to be a false positive and was discarded.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was supported by a grant from the National Science Foundation (IOS-1025840).

Authors' Contributions

Nielsen D designed and coordinated the project, and with Bird DM obtained the funding. Guo Y was responsible for the annotation analysis and the drafting of the manuscript. Nielsen D and Bird DM helped to draft the manuscript. All authors read and approved the final manuscript.

Supplemental Material

Supplemental material may be found here: www.landesbioscience.com/journals/worm/article/29158/

References

1. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998; 282:2012-8; PMID:9851916; <http://dx.doi.org/10.1126/science.282.5396.2012>
2. Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, Cromer J, Diener S, Gajan J, Graham S, et al. Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *Proc Natl Acad Sci U S A* 2008; 105:14802-7; PMID:18809916; <http://dx.doi.org/10.1073/pnas.0805946105>
3. Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EGJ, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, et al. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol* 2008; 26:909-15; PMID:18660804; <http://dx.doi.org/10.1038/nbt.1482>
4. Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S, et al. The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets. *FASEB J* 2012; 26:4650-61; PMID:22889830; <http://dx.doi.org/10.1096/fj.12-205096>
5. Bai X, Adams BJ, Ciche TA, Clifton S, Gaugler R, Kim KS, 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; PMID:23874975; <http://dx.doi.org/10.1371/journal.pone.0069618>
6. Liu QL, Williamson VM. Host-specific pathogenicity and genome differences between inbred strains of *Meloidogyne hapla*. *J Nematol* 2006; 38:158-64; PMID:19259441

7. Thomas VP, Fudali SL, Schaff JE, Liu Q, Scholl EH, Opperman CH, Bird DM, Williamson VM. A sequence-anchored linkage map of the plant-parasitic nematode *Meloidogyne hapla* reveals exceptionally high genome-wide recombination. *G3 (Bethesda)* 2012; 2:815-24; PMID:22870404; <http://dx.doi.org/10.1534/g3.112.002261>
8. Scholl EH, Thorne JL, McCarter JP, Bird DM. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. *Genome Biol* 2003; 4:R39; PMID:12801413; <http://dx.doi.org/10.1186/gb-2003-4-6-r39>
9. Mayer WE, Schuster LN, Bartelme G, Dieterich C, Sommer RJ. Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. *BMC Evol Biol* 2011; 11:13; PMID:21232122; <http://dx.doi.org/10.1186/1471-2148-11-13>
10. Lunt DH, Kumar S, Koutsovoulos G, Blaxter ML. The complex hybrid origins of the root knot nematodes revealed through comparative genomics. *PeerJ* 2013; 2:e356; <http://dx.doi.org/10.7717/peerj.356>
11. Peloquin JJ, Bird DM, Platzner EG. Morphologically homogeneous isolates of *M. hapla* have two distinct mitochondrial genome sizes. *J Nematol* 1989; 21:579
12. Peloquin JJ, Bird DM, Kaloshian I, Matthews WC. Isolates of *meloidogyne-hapla* with distinct mitochondrial genomes. *J Nematol* 1993; 25:239-43; PMID:19279763
13. Lunt DH. Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins. *BMC Evol Biol* 2008; 8:194; PMID:18606000; <http://dx.doi.org/10.1186/1471-2148-8-194>
14. Mbeunkui F, Scholl EH, Opperman CH, Goshe MB, Bird DM. Proteomic and bioinformatic analysis of the root-knot nematode *Meloidogyne hapla*: the basis for plant parasitism. *J Proteome Res* 2010; 9:5370-81; PMID:20804128; <http://dx.doi.org/10.1021/pr1006069>
15. Tisserant E, Da Silva C, Kohler A, Morin E, Wincker P, Martin F. Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. *New Phytol* 2011; 189:883-91; PMID:21223284; <http://dx.doi.org/10.1111/j.1469-8137.2010.03597.x>
16. Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* 2011; 12:540; PMID:22047402; <http://dx.doi.org/10.1186/1471-2164-12-540>
17. Uyar B, Chu JSC, Vergara IA, Chua SY, Jones MR, Wong T, Baillie DL, Chen N. RNA-seq analysis of the *C. briggsae* transcriptome. *Genome Res* 2012; 22:1567-80; PMID:22772596; <http://dx.doi.org/10.1101/gr.134601.111>
18. Laing R, Hunt M, Protasio AV, Saunders G, Mungall K, Laing S, Jackson F, Quail M, Beech R, Berriman M, et al. Annotation of two large contiguous regions from the *Haemonchus contortus* genome using RNA-seq and comparative analysis with *Caenorhabditis elegans*. *PLoS One* 2011; 6:e23216; PMID:21858033; <http://dx.doi.org/10.1371/journal.pone.0023216>
19. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 2011; 471:473-9; PMID:21179090; <http://dx.doi.org/10.1038/nature09715>
20. Derrien T, Vaysse A, André C, Hitte C. Annotation of the domestic dog genome sequence: finding the missing genes. *Mamm Genome* 2012; 23:124-31; PMID:22076420; <http://dx.doi.org/10.1007/s00335-011-9372-0>
21. Krause M, Hirsh D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* 1987; 49:753-61; PMID:3581169; [http://dx.doi.org/10.1016/0092-8674\(87\)90613-1](http://dx.doi.org/10.1016/0092-8674(87)90613-1)
22. Huang XY, Hirsh D. A second trans-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 1989; 86:8640-4; PMID:2813415; <http://dx.doi.org/10.1073/pnas.86.22.8640>
23. Ross LH, Freedman JH, Rubin CS. Structure and expression of novel spliced leader RNA genes in *Caenorhabditis elegans*. *J Biol Chem* 1995; 270:22066-75; PMID:7665629; <http://dx.doi.org/10.1074/jbc.270.37.22066>
24. Allen MA, Hillier LW, Waterston RH, Blumenthal T. A global analysis of *C. elegans* trans-splicing. *Genome Res* 2011; 21:255-64; PMID:21177958; <http://dx.doi.org/10.1101/gr.113811.110>
25. Blumenthal T. Trans-splicing and operons. *WormBook* 2005; 1-9; PMID:18050426
26. Ray C, Abbott AG, Hussey RS. Trans-splicing of a *Meloidogyne incognita* mRNA encoding a putative esophageal gland protein. *Mol Biochem Parasitol* 1994; 68:93-101; PMID:7891751; [http://dx.doi.org/10.1016/0166-6851\(94\)00153-7](http://dx.doi.org/10.1016/0166-6851(94)00153-7)
27. Bird DM, Williamson VM, Abad P, McCarter J, Danchin EGJ, Castagnone-Sereno P, Opperman CH. The genomes of root-knot nematodes. *Annu Rev Phytopathol* 2009; 47:333-51; PMID:19400640; <http://dx.doi.org/10.1146/annurev-phyto-080508-081839>
28. McCarter JP, Bird DM, Clifton SW, Waterston RH. Nematode gene sequences, december 2000 update. *J Nematol* 2000; 32:331-3; PMID:19270985
29. McCarter JP, Clifton SW, Bird DM, Waterston RH. Nematode gene sequences, update for june 2002. *J Nematol* 2002; 34:71-4; PMID:19265911
30. McCarter JP, Mitreva M, Clifton SW, Bird DM, Waterston RH. Nematode gene sequences: update for december 2003. *J Nematol* 2003; 35:465-9; PMID:19262781
31. 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; 28:511-5; PMID:20436464; <http://dx.doi.org/10.1038/nbt.1621>
32. Opperman CH, Bird DM. Plant nematode genomics group. 2012
33. Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW. Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. *Genome Res* 2010; 20:1740-7; PMID:20980554; <http://dx.doi.org/10.1101/gr.111021.110>
34. Guiliano DB, Blaxter ML. Operon conservation and the evolution of trans-splicing in the phylum Nematoda. *PLoS Genet* 2006; 2:e198; PMID:17121468; <http://dx.doi.org/10.1371/journal.pgen.0020198>
35. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; 10:R25; PMID:19261174; <http://dx.doi.org/10.1186/gb-2009-10-3-r25>
36. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; 25:1105-11; PMID:19289445; <http://dx.doi.org/10.1093/bioinformatics/btp120>
37. Rice P, Longden I, Bleasby A. EMBOSS: The european molecular biology open software suite. *Trends Genet* 2000; 16:276-7; PMID:10827456; [http://dx.doi.org/10.1016/S0168-9525\(00\)02024-2](http://dx.doi.org/10.1016/S0168-9525(00)02024-2)
38. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013; 14:178-92; PMID:22517427; <http://dx.doi.org/10.1093/bib/bbs017>
39. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol* 2011; 29:24-6; PMID:21221095; <http://dx.doi.org/10.1038/nbt.1754>