

Quantitative analysis of the coding capacity of *C. elegans* using RNA-Seq data

by
Matthew J Douglas

B.Sc., University of Victoria, 2014

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

© Matthew J Douglas 2018
SIMON FRASER UNIVERSITY
Fall 2018

Copyright in this work rests with the author. Please ensure that any reproduction
or re-use is done in accordance with the relevant national copyright legislation.

Approval

Name: Matthew Douglas

Degree: Master of Science (Molecular Biology and Biochemistry)

Title: **Quantitative analysis of the coding capacity of *C. elegans* using RNA-Seq data**

Examining Committee:

Chair: **Peter Unrau**
Professor

Jack Chen
Senior Supervisor
Professor

David Baillie
Supervisor
Professor Emeritus

Ryan Morin
Supervisor
Associate Professor

Fiona Brinkman
Internal Examiner
Professor

Date Defended/Approved: November 30, 2018

Abstract

Annotating the genome of the nematode *Caenorhabditis elegans* has been an ongoing challenge for the last twenty years. Studies have leveraged high-throughput RNA-sequencing (RNA-Seq) to uncover evidence for thousands of novel splicing events, indicating that the current annotations are far from complete. Yet, there is some uncertainty whether the many rare events represent functional transcripts, or simply biological noise. We developed a method that leverages the wealth of publicly available RNA-Seq data to perform a quantitative evaluation of the completeness of the current *C. elegans* genome annotation. We identified 134,949 and 204,812 novel high-quality introns and exons, respectively. We find that many introns and exons are rarely expressed overall, but strongly expressed at specific developmental stages suggesting a functional role. We assembled a high-quality set of 72,274 protein-coding transcripts to show that only a fraction of the coding transcriptome of *C. elegans* is represented in the current genome annotation.

Keywords: coding capacity; alternative splicing, *Caenorhabditis elegans*; RNA-Seq; transcriptome; bioinformatics

Acknowledgements

Completing this work would not have been possible without many people. I am extremely grateful for all the guidance and support of my senior supervisor, Dr. Jack Chen. Your patience, dedication, and genuine enthusiasm has shown me what a true mentor should be. I am also grateful for Dr. Jiarui Li, who's unending willingness to help taught me so much and truly made me feel welcome in the lab.

I am grateful to my committee members, Dr. David Baillie and Dr. Ryan Morin, for their feedback and guidance through my graduate degree. I am also grateful to Dr. Fiona Brinkman and Dr. Peter Unrau, both for serving on my examining committee and the opportunities they provided via course work and TA-ships.

I would like to thank past and present members of the Chen lab: Marija Jovanovic, Shinta Thio, Kate Gibson, Shirley Yin, Zhaozhao Qin, and Justin White. I'm thankful for all your friendship and encouragement. Thank you to Vincent Ji and Frank Lin, who were volunteers partially under my supervision, for their excellent meta-analysis of publicly available RNA-Seq data and help selecting the libraries used in this thesis.

Thank you to my parents who despite claiming not to understand what I do, have never faltered in their faith in me and pride in who I've become. Finally, thank you Sam. You've made me a better person and I love you (and Bea) so much.

Table of Contents

Approval.....	ii
Abstract.....	iii
Acknowledgements	iv
Table of Contents.....	v
List of Tables.....	vii
List of Figures.....	viii
List of Acronyms.....	xii
Glossary.....	xiii
Chapter 1. Introduction	1
1.1. Organism complexity and alternative splicing	1
1.2. Coding Capacity	3
1.3. <i>Caenorhabditis elegans</i> as a model organism	6
1.4. Thesis aims and organization	8
Chapter 2. Building a high-quality intron database	10
2.1. Introduction.....	10
2.2. Data set selection and quality filtration.....	10
2.2.1. Data set selection	10
2.2.2. Pre-alignment processing of raw reads.....	13
2.2.3. Tandem-duplication filtering.....	16
2.2.4. Selecting a splice-aware alignment program	17
2.2.5. Selecting intron length thresholds.....	20
2.2.6. Selecting a minimum support threshold for introns	21
2.3. Results	22
2.3.1. Validation of WormBase curated introns	22
2.3.2. Identification of novel introns	25
2.3.3. Globally rare vs. locally rare introns.....	25
2.3.4. Identification of embryo stage-specific introns.....	27
2.4. Discussion	28
Chapter 3. Building a high-quality exon database	31
3.1. Introduction.....	31
3.2. Algorithm	31
3.3. Results	35
3.3.1. Evaluating the accuracy of exon reconstruction	35
3.3.2. Validation of the WormBase transcripts	36
3.3.3. Identification of novel exons.....	37
3.3.4. Globally rare vs. locally rare exons	38
3.4. Discussion	39
Chapter 4. Assembling transcripts and evaluating coding capacity of <i>C. elegans</i>	

4.1. Introduction.....	41
4.2. Constructing a fully-supported set of transcripts	41
4.2.1. Assembling transcripts from RNA-Seq data.....	41
4.2.2. Selecting transcripts that are fully-supported by our intron and exon databases.....	42
4.2.3. Selecting full-length supported transcripts	42
4.2.4. Assessing supported transcripts for coding potential	43
4.2.5. Evaluating the accuracy of supported transcripts.....	44
4.3. Evaluating the Coding Capacity of <i>C. elegans</i>	45
4.4. Discussion	48
Chapter 5. Conclusion	52
Chapter 6. Future Directions	54
References.....	55
Appendix A. Supplemental Materials and Methods	62
6.1. <i>C. elegans</i> reference annotation.....	62
6.2. Iso-Seq.....	62
6.3. Programs used.....	63
6.4. ExonTrap availability	63
Appendix B. Supplemental Tables	64

List of Tables

Table 1. Modifications to WormBase protein-coding gene models based on our intron database	25
Table 2. Modifications to WormBase protein-coding gene models based on our exon database	38
Table 3. List of external programs used in this thesis.	63
Supplemental Table 1. RNA-Seq libraries selected from SRA.....	64
Supplemental Table 2. Introns listed as "confirmed" in WormBase that were not detected.	83

List of Figures

Figure 1. Illustration of how seven different modes of alternative splicing produce distinct transcripts from one gene.....	2
Figure 2. Tissue-specific alternative splicing of ERBB4. (A) Simplified schematic of the exon structure of ERBB4. (B) Illustration showing how alternative splicing of exons 15b and 16 of ERBB4 produces transcripts JM-a through -d. (C) Relative expression of ERBB4 mRNA in normal and cancer tissues (adapted from Veikkolainen <i>et al.</i> , 2011).	2
Figure 3. Alternative splicing of exons 18 and 18b of FOXP1 differ between stages of embryo development. Inclusion of exon 18 promotes cell differentiation, whereas exon 18b inclusion promotes maintenance of pluripotency.	3
Figure 4. GBrowse screenshot of a subset of the Dscam1 transcript models in <i>D. melanogaster</i> taken from FlyBase, version FB2018_04 (Gramates <i>et al.</i> , 2017).....	4
Figure 5. Workflow diagram for constructing the intron database.	10
Figure 6. Distribution of the average read length per RNA-Seq library available for <i>C. elegans</i> from NCBI's SRA database, as of February 2018.	11
Figure 7. Distribution of <i>C. elegans</i> developmental stages included in the 802 libraries selected. Stage information is as listed in SRA, except for capitalization and typos (e.g. "daeur" to "dauer")......	12
Figure 8. A total of 84 RNA-Seq libraries were generated from synchronized <i>C. elegans</i> embryos sampled at 30-minute intervals across embryo development as part of an experiment by (Boeck <i>et al.</i> , 2016).	13
Figure 9. IGV (Robinson <i>et al.</i> , 2011) screenshot showing alignments split across multiple rRNA genes, resulting in false introns. Over 93% of reads in dataset SRR1746748 (left) mapped across <i>C. elegans</i> rRNA genes (alignment depth shown in blue); 65% of these mapped to multiple loci. 690 introns were identified from these reads (green). Filtering reads using BBduk prior to alignment minimizes the number of false introns reported from these genes (right).	13
Figure 10. Example paired-end dataset SRR1536037, showing how the distribution of average quality scores per read has two peaks: one at the minimal Phred score 2, and one at Phred score 30. Charts generated by FASTQC.	14
Figure 11. Stringent filtering criteria introduces a significant number of false negatives. (A) WormBase WS250 gene model for <i>atm-1</i> and the introns identified from RNA-Seq data showing that more stringent filtering reduces read support for valid introns, completely removing them in some cases. (B) Introns identified at each quality filtering threshold. The first number in the X-axis labels indicate the value for the LEADING and TRAILING parameters; the second number is the value for the SLIDINGWINDOW parameter (1.5X the value of LEADING) as used by Trimmomatic. (C) WormBase introns not detected at each quality filtering threshold.....	15
Figure 12. Illustration of how a read may be aligned to multiple, identical loci in the genome sequence, and how filtering can uniquely place these reads (top). Example showing false introns spanning between genes <i>fbf-1</i> , <i>fbf-2</i> ,	

and <i>ptc-2</i> , due to their sequence similarity (bottom). Filtering significantly reduces the number of false introns in many cases.....	17
Figure 13. Illustration of how a split alignment is created so that a single RNA-Seq can be correctly aligned back to the gene.	18
Figure 14. Twenty program-specific introns were randomly sampled for each program and manually called as true-positive or false-positive. (Left) False-positive rates for each program based on the randomly sampled program-specific introns. (Right) Example of an intron identified by TopHat2 that was called as a false positive based on the following criteria: (1) The intron is located on the opposite strand to the gene model. (2) Only 2 reads support this intron. (3) At least half of the supporting reads are aligned (by TopHat) with multiple mismatches. (4) The other programs align the same reads to a different splice junction, without mismatches.	19
Figure 15. Distribution of intron lengths identified using Iso-Seq (top) and in WormBase (bottom). Y-axis is log10 scale. The shaded region indicates 30-5000 bp window containing >99% of Iso-Seq and WormBase introns, respectively.	21
Figure 16. Effect of different minimum score thresholds on intron identification. (A) Example of the effect of setting various minimum scores on introns in <i>asic-1</i> . Introns in red are removed as filtering increases to a threshold of 10 reads. (B) Number of novel introns with the indicated maximum read support in any library. “% Included at threshold” (red line) denotes the number of introns retained when the minimum score is set to that value. A minimum score of 5 support was selected for use (light blue bar). (C) Introns present in WormBase that had 10 or less read support in any library.	22
Figure 17. Introns identified from RNA-Seq was used to validate WormBase introns. (A) All introns in the gene models for <i>gcy-17</i> are represented in our intron database. (B) Individual introns in WormBase are either supported by our intron database or only found in WormBase. (C) Transcripts in protein-coding genes were validated at the level of introns. Transcripts were designated “complete” if all introns were represented in our database, “partial” if only some introns were represented, or “none” if no introns were represented. (D) Indicated WormBase intron (red arrow) at the 5’ end of the gene model for <i>C50E3.9</i> is not represented in our database. (E) Indicated intron in WormBase gene <i>Y57A10A.3</i> (red arrow) is not represented in our database. Iso-Seq data supports an alternative transcript that excludes this intron. (F) Breakdown of WormBase introns not represented in our database. Some introns fell below our minimum score threshold (“filtered score”), or outside our intron length thresholds (“filtered length”). Other introns are labelled according to their position within the transcript: “terminal” if they are the first or last intron, otherwise “internal,” or “single-intron” if the transcript only contains one.....	24
Figure 18. Identification of novel introns using RNA-Seq. (A) Breakdown of introns in our database that are either unique to our RNA-Seq analysis or represented in WormBase. (B) Pictured is the WormBase gene models for <i>aex-1</i> and our intron database. Novel introns are highlighted in orange	25
Figure 19. Relative usage of a novel intron and adjacent novel exon (orange) in <i>cki-2</i> (Cyclin-dependent Kinase Inhibitor) shown at 30, 240, and 510 minutes	

into embryo development. Height of the intron represent the relative ratio of read support compared to other introns in the gene in that library. Orange line (bottom) is the average ratio of read support between biological replicates (open circles) at that time point (shaded or non-shaded bars).	26
Figure 20. Number of WormBase introns and novel introns that are non-rare in the global context (ratio calculated for all libraries pooled together), and non-rare in the local context (ratio calculated for each library, individually).....	27
Figure 21. Set of novel introns that show strong expression in one stage of embryo development were found in a gene set enriched for transmembrane components. (A) 135 novel introns show very specific, strong expression at a certain time point during development (average usage ratio ≥ 0.7 in one stage and < 0.1 in the other two stages). (B) Over-representation analysis of the gene set containing the novel embryo-stage specific introns is enriched for GO terms and upregulated pathways involved in membrane components and transmembrane helices. Lists of genes containing the introns showing strong expression in (C) early, (D) mid, or (E) late embryo.....	28
Figure 22. (A) Schematic representation of the exon reconstruction algorithm. (B) Examples of internal, 5' terminal, and 3' terminal exons with the relevant exon and translation block highlighted in orange.	32
Figure 23. Approaches taken to resolving 5' terminal exon boundaries. (A) When multiple putative start codons (methionines) are in frame, the one most distal from the splice donor is selected as the start of the 5' terminal exon. (B) Alternative start sites overlapping a longer transcript are partially missed because internal exons are identified first and are mutually exclusive with terminal exons. Cases like this are considered a partial match if the 3' end of the exon matches a 3' end of one of our exons. (C) Number of WormBase 5' terminal exons that have an exact match, a partial match, or no match in our exon database.	34
Figure 24. (A) Example illustrating how translation blocks (relevant blocks in purple) and introns (blue) are used to define the boundaries of coding exons (green). The exons defined in this example match those of the gene model for <i>ric-19</i> exactly. (B) Breakdown of WormBase coding exons that are represented in our database. Terminal exons that differs only at the terminal end are deemed a partial match. WormBase coding exons not represented in our database are "no match.".....	36
Figure 25. Percentage of WormBase protein-coding transcripts that are completely, partially, or not supported by our intron and exon databases. Example gene models, from top to bottom, are <i>srd-29</i> , <i>irld-61</i> , and <i>srh-141</i>	37
Figure 26. Identification of novel exons using RNA-Seq. (A) Breakdown of exons in our database that are either unique to our RNA-Seq analysis or represented in WormBase. (B) Novel exons (orange) identified for <i>ceh-93</i> from novel introns (blue). The highest scored novel exon is supported by Iso-Seq data.....	38
Figure 27. Relative usage of WormBase exons and novel exons in the global context (ratio calculated for all libraries pooled together), and the local context (ratio calculated for each library, individually).	39

Figure 28. Percentage of transcripts per library, from 802 RNA-Seq libraries, that are fully supported by our intron and exon databases. Box boundaries denote first and third quartile, horizontal line denotes the median, whiskers denote minima and maxima. Note: single-exon transcripts are not supported by our databases.	42
Figure 29. (A) Illustration of how assembled transcripts with the same introns, or a subset of introns of a longer series are merged. (B) Number of transcripts after merging that are fully supported by our intron and exon databases.	43
Figure 30. Identifying candidate coding regions in assembled transcripts. (A) Candidate coding regions identified within supported transcripts mapping to <i>sptf-2</i> . (B) Total number of supported transcripts with, or without, a candidate coding region.	44
Figure 31. Comparison of supported RNA-Seq transcripts to WormBase transcripts and Iso-Seq reads. Comparison is based on the series of introns in each transcript: (Left) The WormBase transcript model for <i>cic-1</i> , the supported RNA-Seq transcripts, and the Iso-Seq transcripts overlapping the gene. (Right) Percentage of supported RNA-Seq transcripts that match or extend a WormBase or Iso-Seq transcript.	45
Figure 32. Number and types of alternative splicing in WormBase gene models and our supported transcript set (RNA-Seq).	46
Figure 33. Number of genes with one (-) or multiple (+) transcripts in WormBase compared to the number of transcripts identified from RNA-Seq. Only transcripts that are fully supported by our intron and exon databases are counted. "Coding only" refers to supported transcripts with a predicted coding sequence.	47
Figure 34. A potentially novel protein-coding gene with multiple transcripts identified between genomic coordinates 13,292,958 and 13,296,957 on chromosome III (- strand). Shown (from top to bottom) are the nearby WormBase gene models, our intron database, our exon database, and supported transcripts with candidate coding regions highlighted in orange.	48

List of Acronyms

EST	Expressed Sequence Tag
OST	Open-reading-frame Sequence Tag
NCBI	The National Center for Biotechnology
NMD	Nonsense-Mediated Decay
RACE	Rapid Amplification of cDNA Ends
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
SRA	Sequence Read Archive
TEC-RED	<i>Trans</i> -spliced Exon Coupled RNA End Detection
UTR	Untranslated Region

Glossary

ASTALAVISTA	Program for identifying alternative splicing events from gene annotations.
BBDuk	Program for performing quality filtration of raw RNA-Seq reads. Used to remove rRNA contamination.
Cufflinks	Program for assembling transcripts from RNA-Seq reads mapped to a genome.
DAVID	A web-based tool for performing gene set enrichment analysis.
Ensembl	Genome browser and source of genome annotations started by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute.
FlyBase	A consortium of scientists curating data on the genetics, genomics and biology of <i>Drosophila</i> species.
GBrowse	Program for storing and visualizing genome annotations in a web browser.
GffCompare	Program for processing and comparing one or more sets of transcript data in GFF-format.
GMAP	Program for aligning RNA-Seq reads to a reference genome.
HISAT2	Program for aligning RNA-Seq reads to a reference genome.
Iso-Seq	Single-molecule, real-time sequencing method developed by PacBio that uses long reads to sequence full-length transcripts
modENCODE	Project with goal of building a comprehensive encyclopedia of genomic functions for model organisms <i>C. elegans</i> and <i>D. melanogaster</i> .
Python	Programming language.
RefSeq	NCBI Reference Sequence Database. Curated databases for genomic, transcriptomic, and proteomic data.
RNA-Seq	High-throughput transcript sequencing using next-generation sequencing technology
STAR	Program for aligning RNA-Seq reads to a reference genome.
Stringtie	Program for assembling transcripts from RNA-Seq reads mapped to a genome.
TransDecoder	Program for identifying candidate coding sequences in transcripts by identifying ORFs and scoring them based on sequence composition.
Trimmomatic	Program for performing quality filtration of raw RNA-Seq reads

TopHat2	Program for aligning RNA-Seq reads to a reference genome.
<i>Trans</i> -ABYSS	Program for assembling transcripts from RNA-Seq data <i>ab initio</i> .
WormBase	A consortium of scientists curating data on the genetics, genomics and biology of <i>C. elegans</i> and related nematodes.

Chapter 1. Introduction

1.1. Organism complexity and alternative splicing

One of the most surprising findings coming out of the sequencing, and subsequent annotation, of the human genome and the genomes of other organisms was a lack of obvious correlation between the number of protein-coding genes and the perceived complexity of the organism. While the human genome has just over 19,000 protein-coding genes (Ezkurdia *et al.*, 2014), the genomes of seemingly less complex animals such as the mouse *Mus musculus*, the zebrafish *Danio rerio*, and the nematode *Caenorhabditis elegans* have 25,592, 22,628 (Ensembl version 93), and 20,359 (WormBase release WS250) protein-coding genes, respectively.

One possible mechanism underlying organismal complexity is alternative splicing of protein-coding genes (Chen *et al.*, 2014). Alternative splicing is a mechanism common in eukaryotic genomes where different exons (or parts of exons) of a protein-coding transcript are included (or excluded) in the mature mRNA (Figure 1), allowing one gene to encode multiple distinct proteins (Wang *et al.*, 2015). Genome-wide surveys of the human transcriptome estimate that 90-95% of human multi-exon protein-coding genes undergo some form of alternative splicing, in total encoding up to 100,000 distinct transcripts (Pan *et al.*, 2008; Wang *et al.*, 2008). Alternative splicing is therefore a key method for increasing proteomic diversity, which in turn may be the key to increasing organismal complexity.

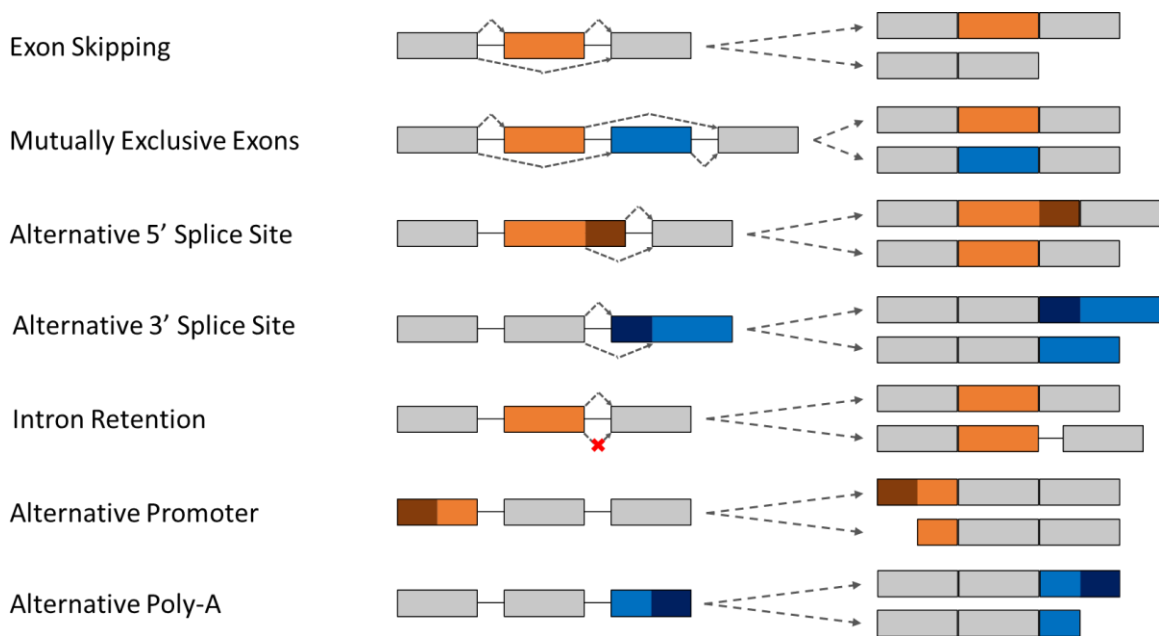


Figure 1. Illustration of how seven different modes of alternative splicing produce distinct transcripts from one gene.

Different proteins encoded by the same gene may play different roles in different biological contexts. As an example, ERBB4 (a.k.a. HER4) in humans encodes a receptor tyrosine kinase required for the proper development of several organs including the heart and central nervous system. It is also an oncogene associated with breast and ovarian cancers (Gilmour *et al.*, 2001). Under normal conditions, alternative splicing of ERBB4 occurs in a highly tissue-specific manner. ERBB4 contains two exons of interest: exon 15b and exon 16. Exon 16 is included in essentially all ERBB4 transcripts expressed in the kidneys (JM-a), whereas transcripts expressed in skeletal muscle (JM-b) generally lack exon 16, including exon 15b instead (Figure 2). Dysregulation of this coordinated splicing produces abnormal transcripts (JM-c and JM-d) which have only been detected in cancerous tissue (Veikkolainen *et al.*, 2011).

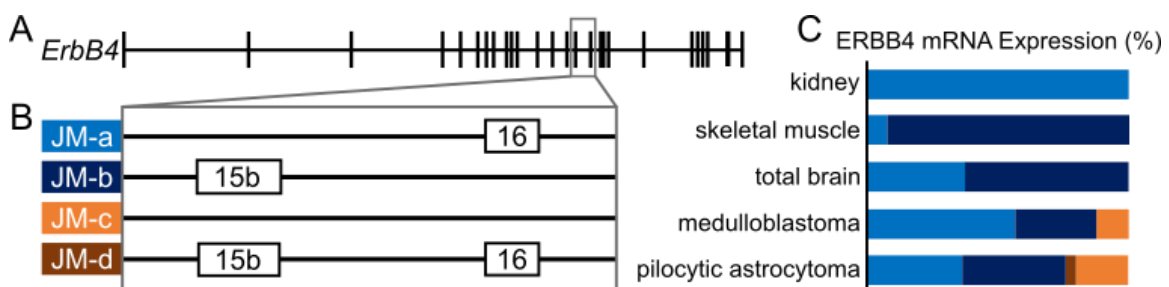


Figure 2. Tissue-specific alternative splicing of ERBB4. (A) Simplified schematic of the exon structure of ERBB4. (B) Illustration showing how alternative splicing of exons 15b and 16 of ERBB4 produces transcripts JM-a through -d. (C) Relative expression of ERBB4 mRNA in normal and cancer tissues (adapted from Veikkolainen *et al.*, 2011).

Normally, alternative splicing is a tightly regulated process. Changing the relative usage of different alternative splicing events, with the proper timing, is essential for development (Kalsotra *et al.*, 2008; Mantina *et al.*, 2009; Shaham and Horvitz, 1996). Changes in the patterns of alternative splicing are frequent during embryo development where tight control of cell differentiation is essential (Revil *et al.*, 2010). For example, mutually exclusive splicing of exons 18 and 18b of the transcription factor, FOXP1, has a profound effect on cell fate. Early in embryo development exon 18 is skipped while the downstream exon 18b is retained (Figure 3). The resulting protein promotes the expression of genes involved in differentiation. As development progresses, exon 18 is preferentially included and exon 18b is skipped. The alternative protein induces a suite of genes that promote maintenance of pluripotency (Gabut *et al.*, 2011). Changes in alternative splicing can have far reaching effects on development, independent of an overall change in gene expression (Dillman *et al.*, 2013).

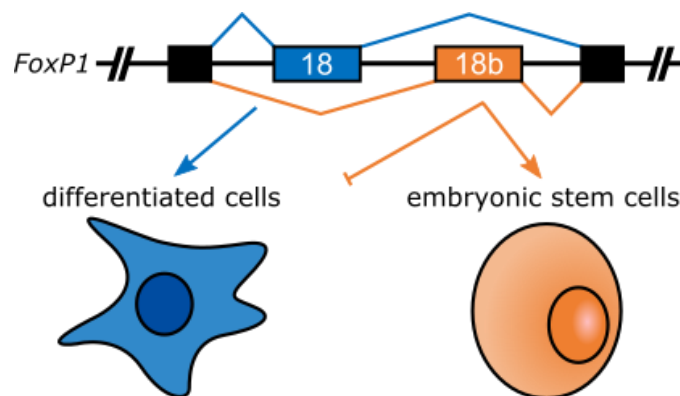


Figure 3. Alternative splicing of exons 18 and 18b of FOXP1 differ between stages of embryo development. Inclusion of exon 18 promotes cell differentiation, whereas exon 18b inclusion promotes maintenance of pluripotency.

1.2. Coding Capacity

It is of primary importance to define the complete set of protein-coding transcripts encoded by a genome, as it serves as a starting point for understanding organismal complexity. A given multi-exon gene in a eukaryote has the capacity to code for multiple transcripts through alternative splicing. The “coding capacity” of a gene is the number of distinct protein-coding transcripts (*i.e.*, all different combinations of exons) expressed across all normal conditions (excluding aberrant splicing causing a disease state). The coding capacity of a gene may be as small as a single transcript if the gene does not undergo any alternatively splicing, or as large as tens-of-thousands of transcripts –

Dscam in *Drosophila melanogaster* is an extreme example, encoding over 38,000 distinct transcripts (Figure 4) which are all expressed, though only a handful of may be found in any given cell (Neves *et al.*, 2004; Schmucker *et al.*, 2000).

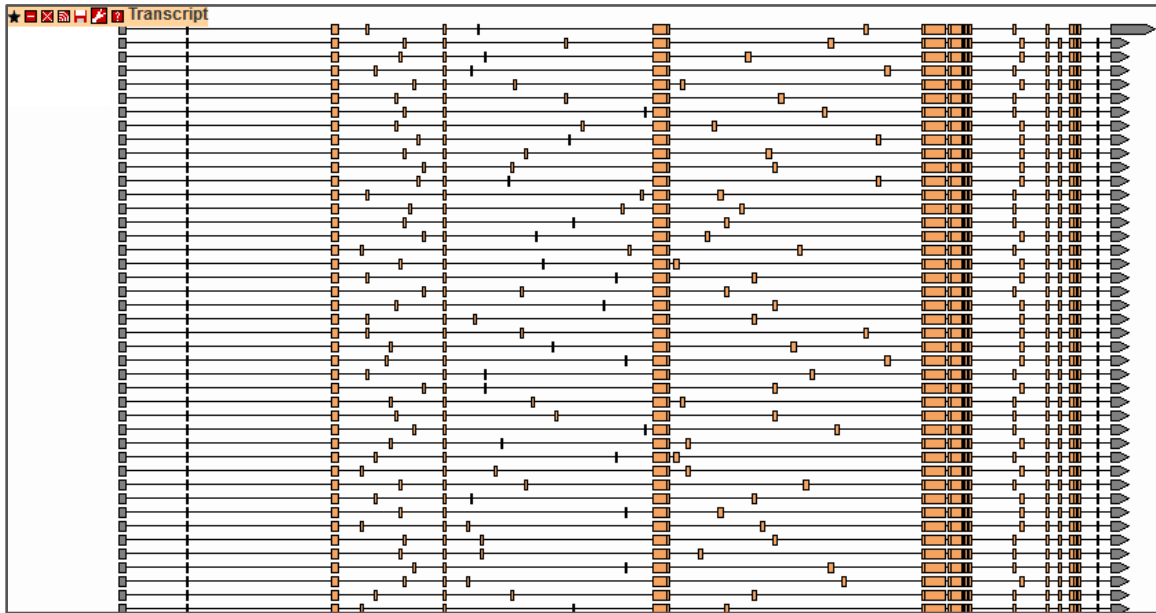


Figure 4. GBrowse screenshot of a subset of the *Dscam1* transcript models in *D. melanogaster* taken from FlyBase, version FB2018_04 (Gramates *et al.*, 2017).

At the whole-genome scale, coding capacity refers to the complete set of coding transcripts encoded by the entire genome (Abdel-Ghany *et al.*, 2016). The current genome annotations count 31,574 coding transcripts in *C. elegans* (WormBase release WS250), 57,388 in mice (GENCODE version M18), and 82,335 in humans (GENCODE version 28). So, while organismal complexity may not correlate with the number of coding genes, it may correlate with coding capacity. However, to make any meaningful inferences based on coding capacity we must first evaluate whether the coding capacity of the subject organisms has been fully defined – *i.e.* all protein-coding transcripts have been identified.

An enormous amount of effort has gone into identifying all protein-coding transcripts encoded in the genomes of several model organisms and humans. At the time when the first genome sequences were becoming available, the amount of sequence evidence for individual genes was limited. As such, a number of computational methods were developed to predict genes *ab initio* (directly from the genome sequence). For prokaryotic genomes, identifying the longest open reading frame (ORF) is typically an adequate method of protein-coding gene prediction as the mRNA is generally

translated without significant modification. In contrast, eukaryotic mRNA is often heavily modified, namely through the removal of introns. Introns present in protein-coding genes can contain stop codons which would otherwise interrupt the ORF, were they not spliced out prior to translation. As a result, gene prediction in eukaryotes faces distinct challenges associated with precisely defining the intron-exon structure of genes. One class of gene prediction software predicts genes *ab initio* based on known gene features including: transcriptional and translational signals, splicing motifs, feature length distribution, and sequence composition changes between coding and non-coding DNA. The most successful of these programs, including GENSCAN (which was used to predict genes as part of the Human Genome Project), structure the search for these features as a generalized hidden Markov model (Burge and Karlin, 1997). While this offers a statistically robust approach to gene identification – indeed, GENSCAN predictions continue to be a fundamental component of new releases of the human genome (Aken *et al.*, 2016) – the accuracy of such predictions is limited by our current understanding of gene structure. The second class of gene prediction software bases predictions on similarity to other genomes, proteins, or sequenced DNA fragments such as expressed-sequence tags (ESTs). The assumption underlying this approach is that functional regions of the genome (*i.e.* exons) are more conserved than non-functional regions (introns or intergenic regions) (Wang *et al.*, 2004). While this can be a valuable approach for identifying both gene structure and function, its usefulness is limited in two regards: First, the number of similarities identified is a function of the number of sequences available to compare to, which in the early days of gene prediction were comparatively limited (relative to today). Second, the degree of sequence conservation of genes between species may vary widely depending on the gene, meaning less conserved genes can be difficult to identify.

Comparative genomics-based evaluation of the initial set of predicted human gene models found that only a minority (<15%) could be experimentally validated (Flicek, 2007; Guigó *et al.*, 2006), highlighting a need for direct sequence evidence to accurately define gene models. Most of the current genome annotations, such as those from GENCODE (Harrow *et al.*, 2012) and RefSeq (Pruitt *et al.*, 2014), are primarily derived from *ab initio* gene prediction, followed by experimental validation using a combination of EST, cDNA, and protein alignments (Thibaud-Nissen *et al.*, 2013). This combined approach has produced high-quality genome annotations invaluable for biological

studies.

However, the classic sequencing techniques used to construct the current gene models are generally low-throughput, which has made it difficult to achieve a level of sequencing depth sufficient to reliably detect low-expression transcripts (Wang *et al.*, 2000). The advent of high-throughput transcriptome sequencing (RNA-Seq) overcomes this limitation and has been incredibly valuable for identifying novel transcripts. With RNA-Seq, investigators can achieve unparalleled coverage of the transcriptome. The technique's popularity means that there is now a wealth of sequence data available for most model organisms, which has been leveraged to identify many potentially novel splicing events (Nellore *et al.*, 2016; Tress *et al.*, 2007). For example, a novel transcript of the huntingtin gene that includes a human-specific exon had been missed by conventional sequencing techniques, which has implications in designing treatments for Huntington's disease (Ruzo *et al.*, 2015).

RNA-Seq has become a powerful tool for evaluating coding capacity – One that potentially allows us to answer the questions: Do we know the full coding capacity of any organism, and if not, how much remains to be identified?

1.3. *Caenorhabditis elegans* as a model organism

Caenorhabditis elegans is a multicellular, free-living nematode roughly one millimetre long. Proposed as a model organism by Sydney Brenner in 1974, it was the first multi-cellular genome to be sequenced (The *C. elegans* Sequencing Consortium, 1998). It remains a popular choice for genetic studies to this day due to the ease of reproduction (self-fertilizing hermaphrodite), rapid generation time (3-4 days), and relatively simple, 100 Mbp genome.

As is common for eukaryotes, nearly all protein-coding genes in *C. elegans* contain one or more introns. Many of these genes undergo alternative splicing, which increases the coding capacity of the worm. Over twenty years of computational and experimental efforts have gone into identifying all the protein-coding transcripts produced in the *C. elegans*: The initial gene models were predicted *ab initio* using Genefinder (Green, P., unpublished data). These models were refined based on supporting experimental evidence from large-scale sequencing projects using conventional

sequencing techniques including ESTs (Kohara, 1996, unpublished; Shin *et al.*, 2008), OSTs (Reboul *et al.*, 2003), SAGE (Ruzanov *et al.*, 2007), and RACE (Salehi-Ashtiani *et al.*, 2009). All this data has been used to curate a high-quality set of transcript models constructed by WormBase (Chen *et al.*, 2005; Lee *et al.*, 2018; Stein *et al.*, 2001). These efforts indicate that alternative splicing is less prevalent in *C. elegans* relative to other “more complex” vertebrates. While nearly all human multi-exon protein-coding genes show some evidence of alternative splicing, only 28% of *C. elegans* protein-coding genes have multiple annotated isoforms (WormBase WS250).

Beyond alternative splicing, which is a *cis*-splicing reaction, *C. elegans* also employs the mechanism of *trans*-splicing. *Trans*-splicing in *C. elegans* involves the cleavage of the pre-mRNA at a 3' splice site very close to the 5' end of the gene, catalyzed by the spliceosome in largely the same manner as would occur for an intron (Hannon *et al.*, 1991). This is followed by a 22-nucleotide “spliced leader” (SL), preceding a 5' splice site on an entirely separate small nuclear ribonucleoprotein particle, being cleaved and added to the pre-mRNA. This splicing event occurs between two different transcripts, hence why it is referred to as a *trans*-splicing reaction. While *trans*-splicing has been detected in mammals (Frenkel-Morgenstern *et al.*, 2013; Herai and Yamagishi, 2010), only a small fraction of transcripts were affected. In contrast, an estimated 70-84% of *C. elegans* protein-coding genes undergo *trans*-splicing (Riddle *et al.*, 1997a; Tourasse *et al.*, 2017). Just over half of the *C. elegans* protein-coding genes use one form of splice leader, SL1, which is added to the 5' end of the pre-mRNA a short distance upstream of the start codon. The second form of splice leader, SL2, is exclusively used on the 15% of *C. elegans* genes that are expressed as part of an operon. Operons are clusters of genes (often part of the same biological pathway) under the control of a single promoter which, when transcribed, produce polycistronic RNA. For a long time operons were thought to be a distinctly prokaryotic feature, prior to being identified in *C. elegans* (Spieth *et al.*, 1993). In many cases the first gene in the *C. elegans* operon uses SL1, but all subsequent genes in the operon (1-7 additional genes) are *trans*-spliced with SL2.

The near ubiquity of splice-leader *trans*-splicing in *C. elegans* mRNA complicates transcript annotation. Because the 5' end of most protein-coding transcripts is removed prior to sequencing and replaced with a sequence (*i.e.* the SL) that is not readily mapped back to the genome with the rest of the transcript, means that precisely defining the 5'

UTR of transcripts can be a challenging task. This is only compounded by the fact that most RNA-Seq libraries (*i.e.* those prepared using poly-A selection) already have a distinct bias towards the 3' end of transcripts (Li *et al.*, 2014). A sequencing protocol, TEC-RED, was developed specifically to annotate the 5' end of *C. elegans* transcripts by capturing transcript by the SL sequence. This technique has been successful at validating a subset of predicted *C. elegans* transcript 5' ends and identifying novel alternative transcripts (Hwang *et al.*, 2004). However sequencing depth has been limited, making it difficult to validate rare transcripts. Currently, 37% (11,603/31,574) of annotated protein-coding transcripts lack an annotated 5' UTR (WormBase release WS250).

More recently, RNA-Seq has been used to define and revise *C. elegans* transcript models. Large-scale RNA-Seq projects, including modENCODE (Gerstein *et al.*, 2010), have identified thousands of novel introns (Hillier *et al.*, 2009; Boeck *et al.*, 2016; Tourasse *et al.*, 2017). While some of the novel introns identified this way have been incorporated into gene models, others were neglected because they were observed to be relatively rare compared to those that make up the extant gene models; interpreted instead as spurious splicing events (Tourasse *et al.*, 2017). Conclusions from these types of large-scale RNA-seq meta-analyses were generally drawn from pooled sets of RNA-Seq data, which, we predict, mask stage- or tissue-specific expression patterns and underestimate the significance of seemingly rare transcripts. In this thesis project, I hypothesize that the coding capacity of the *C. elegans* genome has not been fully defined. Many new transcripts remain to be defined, especially those that may be rare in the overall context of *C. elegans*, but abundant in specific developmental stages, tissues, or cells. These transcript may play important functional roles in specific circumstances (Mullen *et al.*, 1999).

1.4. Thesis aims and organization

In this thesis, we will use *C. elegans* as a model to develop a method for evaluating completeness of coding capacity at the genome-scale, and for reconstructing missing transcripts using RNA-Seq data. Our first hypothesis is that after two decades of research, using a large array of technologies, the coding capacity of *C. elegans* (as represented in the WormBase gene models) is still far from complete. Our second hypothesis is that most introns identified by RNA-Seq are rare in the context of the entire

C. elegans life cycle yet may be highly expressed under specific circumstances. We tested this hypothesis with a particular focus on *C. elegans* embryo development, as stage-specific transcripts may have important roles in development.

Our research on the coding capacity of *C. elegans* consists of four aims. The first aim was to develop a pipeline to build a high-quality database of *C. elegans* introns using a set of empirically-derived filtering criteria. We applied this pipeline to hundreds of publicly available *C. elegans* RNA-Seq libraries. The resulting intron database was used to identify novel introns, and to investigate the changes in the relative usage of alternative splicing events across *C. elegans* embryo development.

The second aim was to build a high-quality database of *C. elegans* coding exons. Here we developed an algorithm that uses a set of introns (from the first aim) and a genome sequence to precisely define the boundaries of coding exons across the genome. We used our exon database to identify novel exons and investigate the relative usage of exons across all libraries.

The third aim was to evaluate the completeness of the WormBase gene models, using both the intron and exons databases generated as part of the first and second aims.

The fourth aim was to assemble a high-quality set of transcripts, guided by our intron and exon databases. These transcripts were used to evaluate the coding capacity of *C. elegans*.

Chapter 2. Building a high-quality intron database

2.1. Introduction

In this chapter we describe the construction of a high-quality database of introns in *C. elegans* using publicly available RNA-Seq libraries. Introns serve as one metric which we used to evaluate the completeness of the current protein-coding gene models. To minimize the effect of technical and biological noise on our interpretation of the results, we employed a set of empirically derived quality control steps for data set selection, read processing, program selection, alignment filtering, and intron identification (Figure 5). To demonstrate the validity of our approach, we use our intron database to validate introns in the WormBase gene models. Our database contains introns not represented in the WormBase models. A subset of these show specific expression patterns related to embryo development, which we explore.

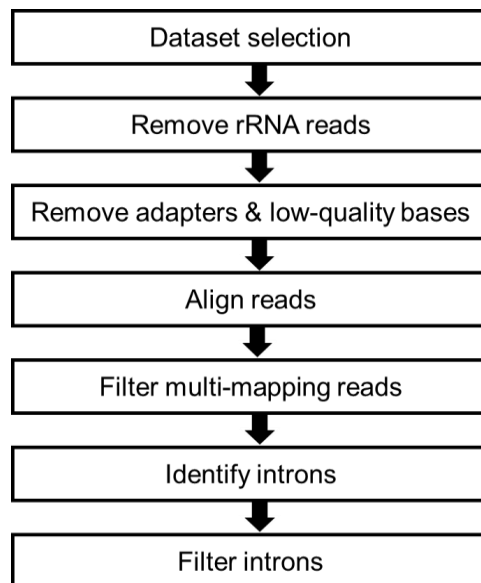


Figure 5. Workflow diagram for constructing the intron database.

2.2. Data set selection and quality filtration

2.2.1. Data set selection

Read length can have a significant impact on alternative splicing analyses. Longer, paired-end reads (>50 bp) are favoured for investigating alternative splicing

because their increased breadth of transcript coverage allows splicing events to be detected more reliably (Chhangawala *et al.*, 2015). Read length has been found to be negatively-correlated with variability in measured intron expression level between samples, with longer reads offering a more consistent measure of intron expression.

The RNA-Seq libraries used in this thesis were selected from the pool of publicly available libraries in SRA. A list of libraries was obtained by querying SRA with the term: "*Caenorhabditis elegans*"[Organism] AND "biomol rna"[Properties]. Our goal was to use reads with the greatest length possible, while still compiling enough data to achieve generally deep transcriptome coverage. As a compromise between read length and data volume, we chose to use paired-end libraries with an average read length of 140 bp or longer. 802 RNA-Seq libraries (1075 runs) met these requirements, totaling 53.9 billion reads – 5.6 Tbp of sequenced cDNA (Figure 6). Reads in FASTQ format were downloaded using “fastq-dump” from the “SRA Toolkit 2.8.2” (<http://ncbi.github.io/sra-tools/fastq-dump.html>).

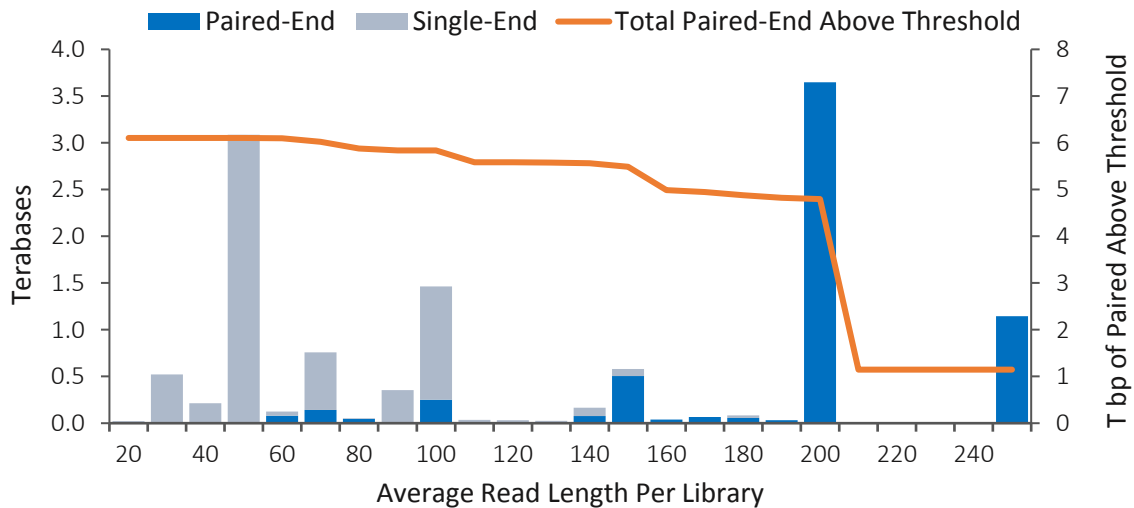


Figure 6. Distribution of the average read length per RNA-Seq library available for *C. elegans* from NCBI's SRA database, as of February 2018.

Out of the selected libraries, 349 cover all standard stages of the *C. elegans* life cycle (*i.e.* embryo, L1 to L4, young adult, adult, and dauer) in addition to 271 which were obtained at more specific time points during development (pertinent to their respective studies). 55 libraries were obtained from mixed populations. 127 had no stage metadata available in SRA (Figure 7).

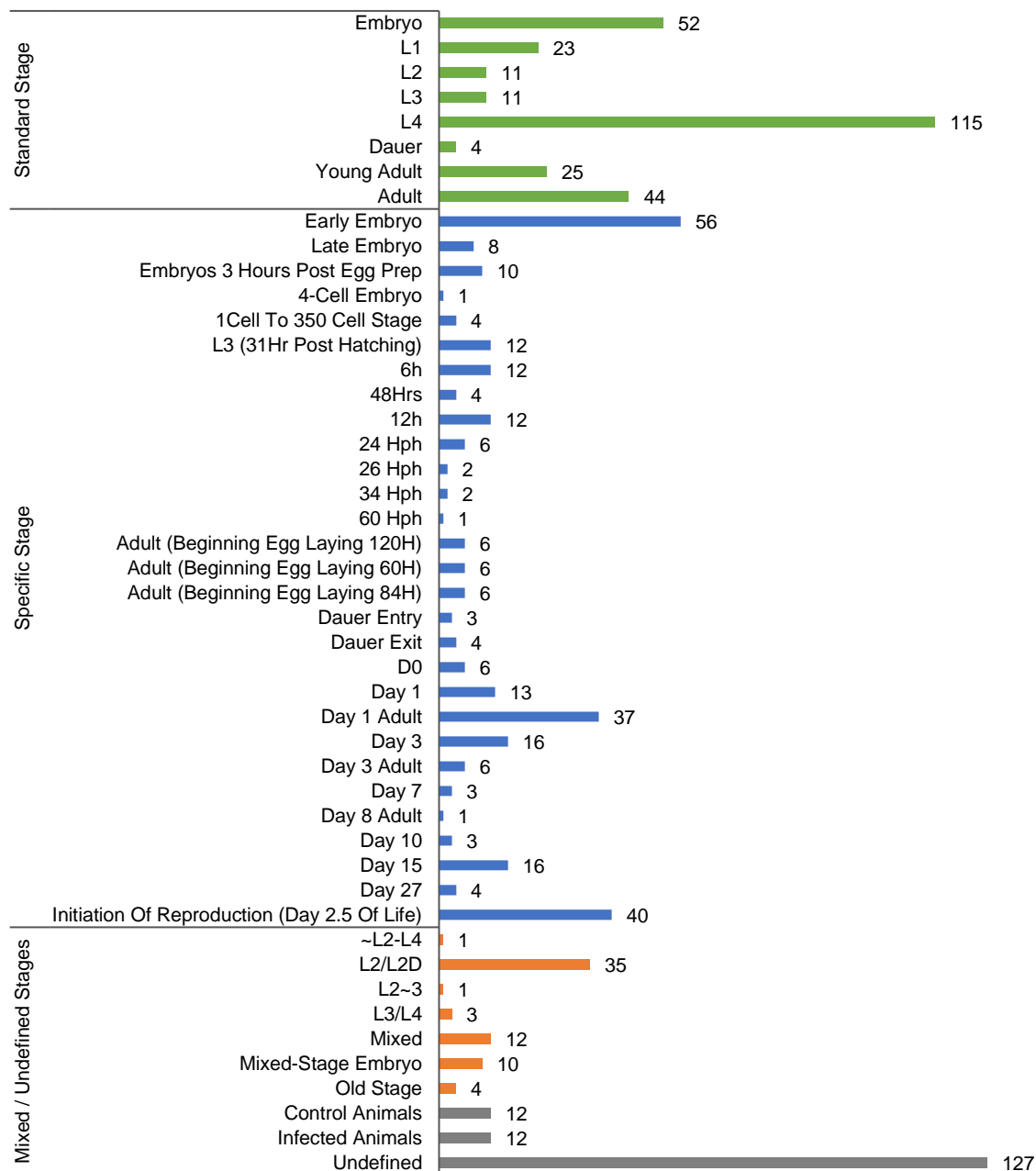


Figure 7. Distribution of *C. elegans* developmental stages included in the 802 libraries selected. Stage information is as listed in SRA, except for capitalization and typos (e.g. “daeur” to “dauer”).

84 of the selected libraries were generated by Boeck *et al.* (2016) where synchronized embryos were sampled at 30 minute intervals, as part of an experiment to assay gene expression across the *C. elegans* life cycle (Figure 8). These libraries – hereafter referred to as the “embryo time-series libraries” – facilitated our investigation of embryo stage-specific changes in intron expression.

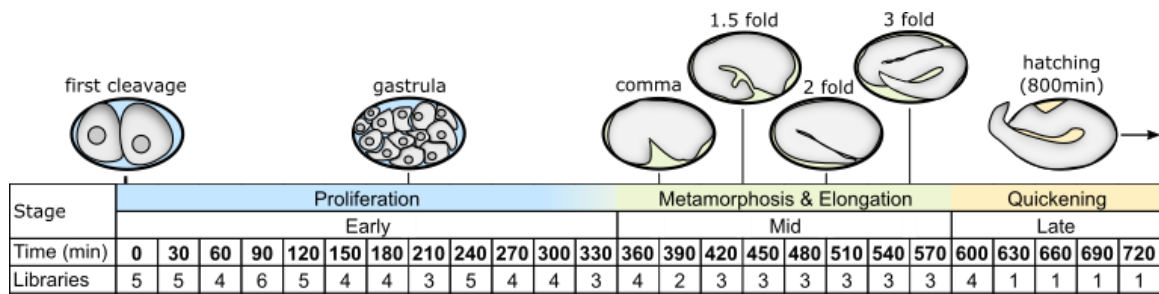


Figure 8. A total of 84 RNA-Seq libraries were generated from synchronized *C. elegans* embryos sampled at 30-minute intervals across embryo development as part of an experiment by (Boeck *et al.*, 2016).

2.2.2. Pre-alignment processing of raw reads

We observed that some RNA-Seq libraries have a high proportion of reads (in some cases >95%) mapping to ribosomal RNA (rRNA) genes. Most of these reads mapped across multiple rRNA genes and resulted in hundreds of overlapping introns (Figure 9). The fact that these introns: A) overlap multiple genes, B) overlap genes not known to contain introns, and C) often originate from multi-mapping reads, indicate they are false positives. Reads matching rRNA sequences were removed using “BBDuk 37.36” (<http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/BBDuk-guide>) supplied with the genomic sequences of all the genes labelled “rRNA” in WormBase. In total, 8.06 (Quinlan and Hall, 2010) billion out of the 53.9 billion individual reads (15.0%) were removed using BBDuk.

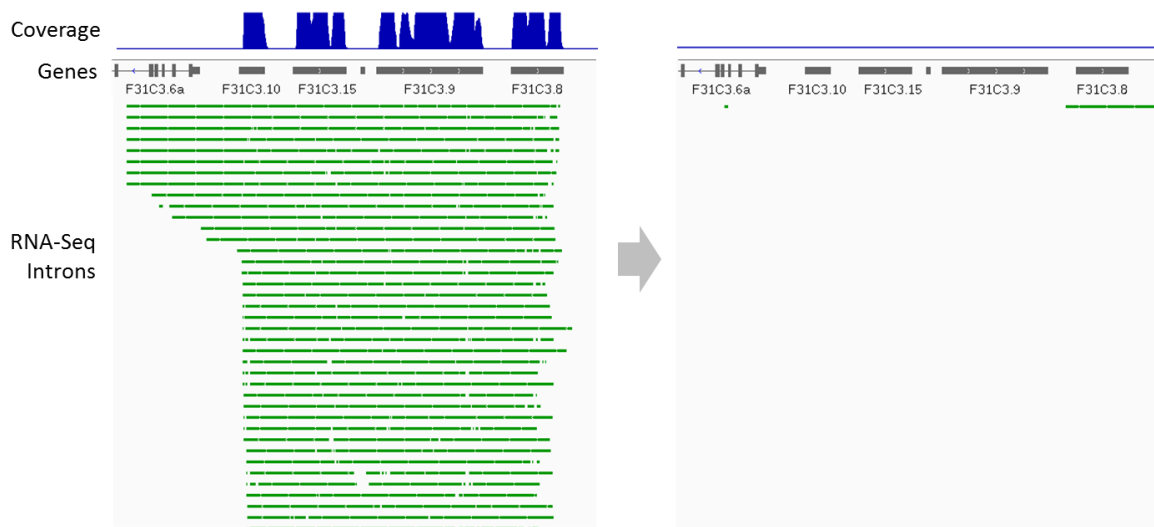


Figure 9. IGV (Robinson *et al.*, 2011) screenshot showing alignments split across multiple rRNA genes, resulting in false introns. Over 93% of reads in dataset SRR1746748 (left) mapped across *C. elegans* rRNA genes (alignment depth shown in blue); 65% of these mapped to multiple loci. 690 introns were identified from these reads (green). Filtering reads using BBDuk prior to alignment minimizes the number of false introns reported from these genes (right).

Read trimming, where low-quality bases and adapter sequences are removed, is a common pre-alignment quality control step in RNA-Seq analyses. However, the specific criteria used to perform the trimming varies between studies.

Average read quality for the 802 libraries was manually assessed using “FASTQC 0.11.5” (<http://bioinformatics.babraham.ac.uk/projects/fastqc>). We commonly observed populations of reads in individual libraries with overwhelmingly low quality-scores (Figure 10). These reads are almost certainly technical artifacts introduced during sequencing and should be removed.

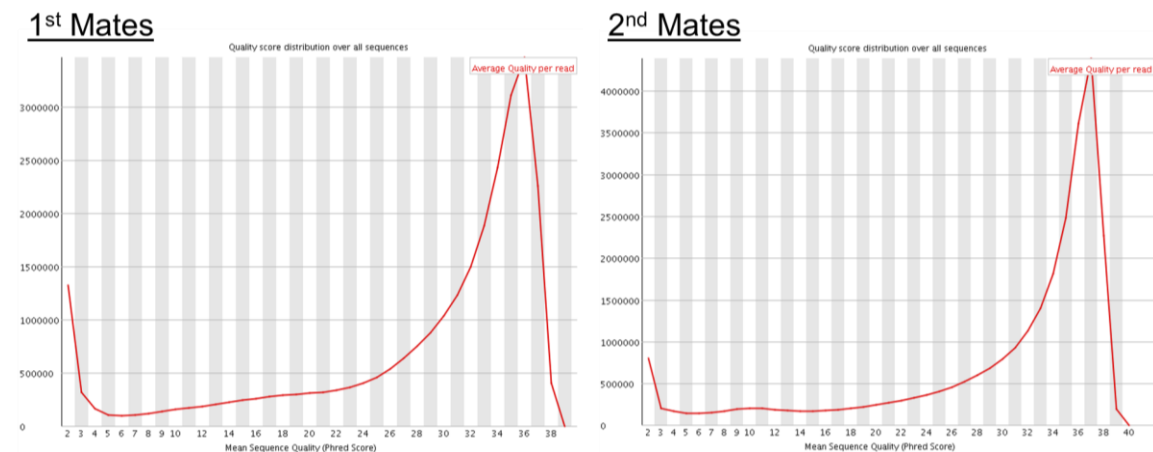


Figure 10. Example paired-end dataset SRR1536037, showing how the distribution of average quality scores per read has two peaks: one at the minimal Phred score 2, and one at Phred score 30. Charts generated by FASTQC.

We investigated if applying more stringent quality trimming criteria produced a higher-quality set of introns. One option would have been to set a threshold such that only the highest-quality reads were accepted (e.g. Phred score 30 or higher). While this increases our confidence that the base calls within each read have been called correctly it may needlessly exclude perfectly valid intron-supporting reads, leading to an underestimation of intron expression levels. As a test, dataset SRR1536037 was filtered using “Trimmomatic” 0.36 (<http://usadellab.org/cms/?page=trimmomatic>) at a set of thresholds. Adapter trimming was enabled with the “ILLUMINACLIP” parameter with either the NexteraPE, TruSeq2, or TruSeq3 adapter sequence in FASTA format (bundled with Trimmomatic), where appropriate. A suite of minimum quality thresholds was tested using the parameters: “LEADING:X TRAILING:X SLIDINGWINDOW:Y:30 MINLEN:50.” Where X was the minimum base quality score to accept, and Y is equal to X multiplied by 1.5. Only read-pairs where both reads survived filtering were kept. We

counted the total number of introns detected, as well as the number of WormBase introns that were not detected, at each filtration level. We did this with the assumption that most, if not all, WormBase introns have been correctly identified. Therefore, a useful filtration level will allow us to identify as many WormBase introns as possible while still removing low quality reads.

We observed that increasing the filtering threshold underrepresents intron expression levels; in many cases, eliminating intron support altogether (Figure 11). To maintain our ability to accurately report intron expression, we decided to set a low filtering threshold. We find that there was little change in results between filtering threshold three and fifteen, so we selected a value in the middle – seven. Our results are consistent with an evaluation by (Williams *et al.*, 2016b) that showed stringent filtering removes an excessive amount of valid reads, which has a significant negative impact on subsequent transcript assembly and gene expression estimates.

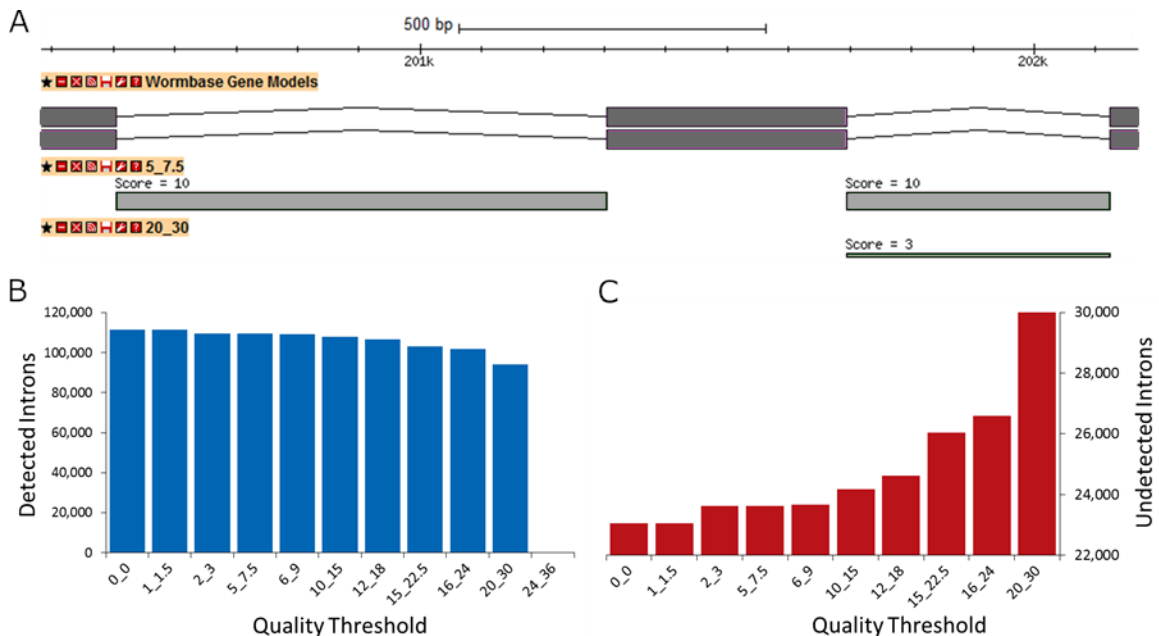


Figure 11. Stringent filtering criteria introduces a significant number of false negatives. (A) WormBase WS250 gene model for *atm-1* and the introns identified from RNA-Seq data showing that more stringent filtering reduces read support for valid introns, completely removing them in some cases. (B) Introns identified at each quality filtering threshold. The first number in the X-axis labels indicate the value for the LEADING and TRAILING parameters; the second number is the value for the SLIDINGWINDOW parameter (1.5X the value of LEADING) as used by Trimmomatic. (C) WormBase introns not detected at each quality filtering threshold.

Adapter removal and quality trimming resulted in 1.96 billion reads being discarded. In total, pre-alignment filtering (including those filtered out by BBDuk) removed 6.00 billion reads – 21.5% of the initial dataset obtained from SRA.

2.2.3. Tandem-duplication filtering

Tandem-duplications are characterized as two or more (nearly) identical loci that are close in the genome sequence. These can be a significant source of false positives during intron identification. Often RNA-Seq reads will map across introns at both loci equally well and count as support for both, artificially inflating read support. Reads can also be erroneously split between loci; one read fragment maps to one locus and the other maps to another resulting in a false intron that join two genes. It is not uncommon in RNA-Seq studies to simply discard multi-mapping reads. However, we feel this would needlessly discard ~3.3% of reads in our dataset and underrepresent the level of splicing in the affected genes. Several statistically robust algorithms exist to assign reads to a particular locus (Hashimoto *et al.*, 2009; Kahles *et al.*, 2016; Zhang *et al.*, 2013). However, these treat all alignments as correct and we needed to selectively discard misaligned intron-supporting reads. To avoid these issues, we filtered the alignments using two criteria (Figure 12):

1. When a read has multiple alignments, we select the shortest – this favours intron(s) localized within a gene over ones spanning between genes.
2. When a read has multiple alignments that are equally short, we select one at random – this ensures a read is only counted once and does not artificially inflate expression levels for transcripts.

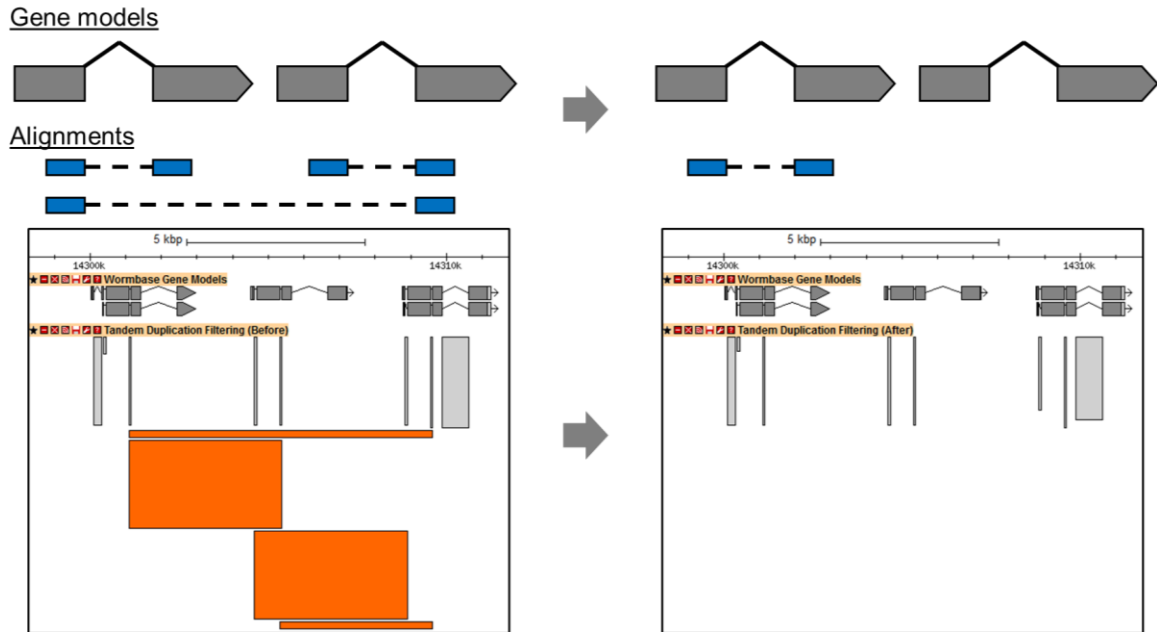


Figure 12. Illustration of how a read may be aligned to multiple, identical loci in the genome sequence, and how filtering can uniquely place these reads (top). Example showing false introns spanning between genes *fbf-1*, *fbf-2*, and *ptc-2*, due to their sequence similarity (bottom). Filtering significantly reduces the number of false introns in many cases.

Selecting alignments at random tends to produce a stochastic distribution of reads between the multiple loci. While this approach does not perfectly reflect the complex expression patterns of each loci, it does adequately address the heavy bias false positive introns would have on our analyses.

Tandem-duplication filtering removed 2.1 billion alignments (5.7% of alignments total). A total of 6,498 (6%) introns were eliminated (*i.e.* all reads supporting these introns were filtered out). The final number of aligned reads, after all filtering steps, is ~35 billion alignments, with just over 9 billion supporting an intron.

2.2.4. Selecting a splice-aware alignment program

Identification of novel splicing events necessitates the use of a splice-aware alignment program. Reads originating from intron-less cDNA that cross the boundaries of a splice junction must be split in order to be successfully mapped back to the intron-containing genome (Figure 13).

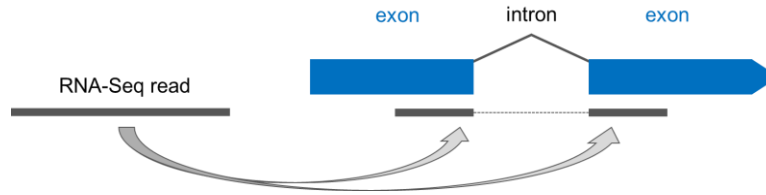


Figure 13. Illustration of how a split alignment is created so that a single RNA-Seq can be correctly aligned back to the gene.

Often an intron may split an RNA-Seq read close to the end of the read, such that a very small fragment (10 bp or less) of the original read needs to be mapped separately from the rest of the read (*i.e.* to the other end of the intron). These kinds of alignments are particularly challenging when the intron is not known. To address the challenges associated with split-alignments, a generation of so-called “splice-aware” alignment programs has been developed. We note that it is technically possible to identify novel introns with other alignment programs, such as BWA (Li and Durbin, 2009), where the novel intron is reported as an indel and it is up to the investigator to distinguish it from true insertions or deletions. However, even this requires careful tuning of run parameters to achieve any degree of accuracy, furthering the idea that a splice-aware aligner is necessary for accurate intron detection.

TopHat (Trapnell *et al.*, 2009) was the first splice-aware alignment program, but since its inception many more have been developed – each with their own strengths and weaknesses. The accuracy, and hence the utility, of the intron database we aimed to construct depended heavily on the performance of the splice-aware aligner we used. Therefore, our goal was to choose a program that identified novel introns with the least false positives. Here we compared three popular programs: “STAR v2.6.0” (Dobin *et al.*, 2013), “TopHat2 v2.1.0” (Kim *et al.*, 2013), and “HISAT2 v2.1.0” (Kim *et al.*, 2015) the successor to TopHat. To save on computational resources, we subsampled our initial dataset of 802 libraries. We randomly selected 100 RNA-Seq runs (from 57 libraries, noted in Supplemental Table 1) and aligned each separately with TopHat2, STAR, and HISAT2. Pre-processing of the libraries and tandem-duplication filtering of the alignments was carried out as described in sections 2.2.1, 2.2.2, and 2.2.3 of this thesis. Introns reported by each program were pooled into program-specific sets. Introns with only one supporting read in any given library, or introns with an ambiguous strand were discarded.

Each program identified a subset of introns specific to that program. These are

cases where one program has either erroneously identified an intron (a false positive) or has correctly identified an intron that the other two programs failed to detect. Our aim was to select the program with the lowest false-positive rate among the program-specific introns. However, the issue was that we have no “ground truth” about which introns are correct – just because a reported intron does not match a known intron, it does not necessarily mean that the intron is false. Based on our observations from the manual inspection of randomly sampled program-specific introns, we identified four criteria that we believe correlate with a false positive intron:

- A. Intron is not located within 100 bp of a known gene
- B. Intron has less than five supporting reads total
- C. Supporting reads have poor alignment scores (e.g. due to multiple mismatches)
- D. Other program(s) align the supporting reads to a different position with better alignment scores

If two or more of these criteria were met for a given intron, the intron was ruled as a false positive. To determine false-positive rates for each program, twenty introns were randomly sampled from each pool of program-specific introns and manually evaluated using the aforementioned criteria (Figure 14).

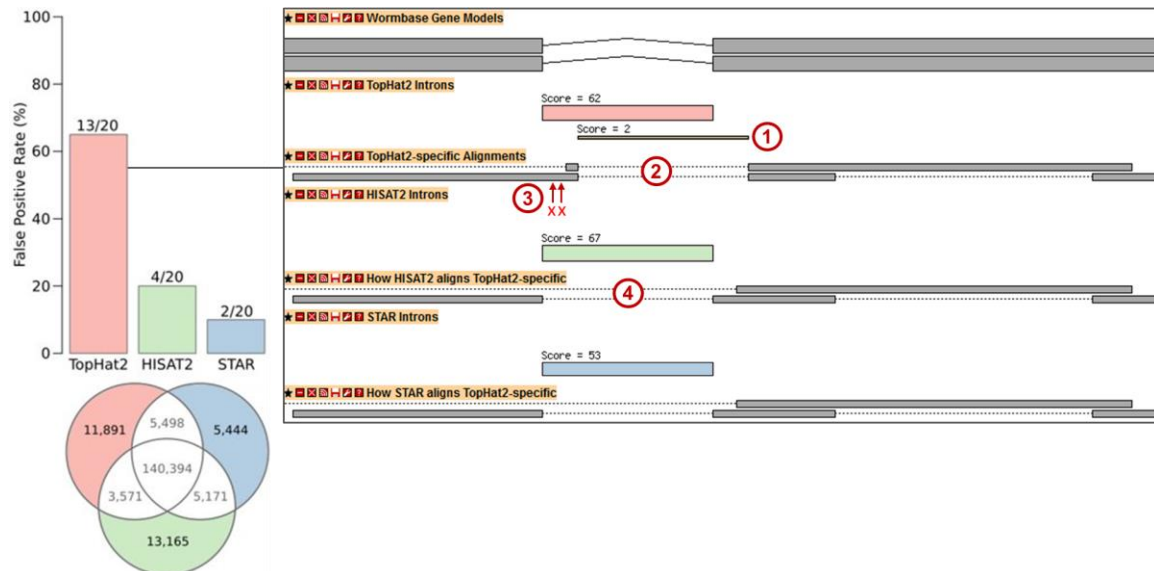


Figure 14. Twenty program-specific introns were randomly sampled for each program and manually called as true-positive or false-positive. (Left) False-positive rates for each program based on the randomly sampled program-specific introns. (Right) Example of an intron identified by TopHat2 that was called as a

false positive based on the following criteria: (1) The intron is located on the opposite strand to the gene model. (2) Only 2 reads support this intron. (3) At least half of the supporting reads are aligned (by TopHat) with multiple mismatches. (4) The other programs align the same reads to a different splice junction, without mismatches.

Based on this semi-quantitative analysis, STAR had the lowest false positive rate. While this analysis used a limited set of introns, the result agrees with other more comprehensive studies that found STAR outperformed TopHat in a variety of tests (Baruzzo *et al.*, 2017; Engström *et al.*, 2013; Williams *et al.*, 2016a).

We selected STAR to produce the alignments used in this thesis. STAR was run in paired-end mode with the following parameters: “--outSAMstrandField intronMotif --outFilterType BySJout --outSAMattributes NH HI NM MD --alignIntronMin 30 --alignIntronMax 5000” to produce a set of alignments in BAM format for each dataset. Alignment sorting and indexing was performed using “SAMtools” 1.5 (<http://samtools.sourceforge.net>).

Introns were identified using Python 3.6 and the module “pysam” 0.14.1 to count the number of alignments split across the same genomic region.

2.2.5. Selecting intron length thresholds

STAR, like most splice-aware alignment programs, allows the user to set thresholds for minimum and maximum intron length. For STAR, the minimum is 21 bp with no maximum value. The median intron length in humans is 1,023 bp (Lander *et al.*, 2001) yet thousands of introns exceed 50,000 bp with the longest more than 1 Mbp (Shepard *et al.*, 2009). Compare this to *C. elegans* where 56% of introns are under 100 bp long (with a sharp peak around 47 bp); the longest annotated WormBase intron is 133,736 bp. A high maximum threshold may be appropriately inclusive for humans, but in *C. elegans* it may only allow for false positives. However, setting too restrictive a threshold can exclude a large portion of valid introns. Here we define minimum and maximum intron length thresholds that include the majority of *C. elegans* introns while minimizing false positives.

We investigated the distribution of intron lengths in WormBase and from introns identified using long reads generated using Iso-Seq (see Supplemental Methods). A total of 603,652 Iso-Seq reads were aligned to the genome using GMAP (Wu and Watanabe, 2005) with default parameters allowing for introns 9-200,000 bp long. There were 57,710

introns identified and the distribution of lengths mirrored that of the WormBase introns. The exception being over a dozen introns much longer than any in WormBase – the largest being 499,799 bp. Examination of this intron showed it spanned dozens of genes, with the 3' end located within a known pseudogene (WB052.3). Only one Iso-Seq read supported this intron and we were unable to validate it using RNA-Seq data, which strongly indicated that this is an artifact introduced during read alignment. The smallest WormBase intron listed as confirmed (by EST sequence) is 30 bp. We were unable to validate any introns shorter than this using RNA-Seq data. Additionally, WormBase annotations also include several features 1-4 bp long labelled “intron” which are not true introns. Instead, they exist to correct the frame of a CDS when a genomic sequence error is strongly suspected (Spieth *et al.*, 2005).

We consider a threshold that captures 99% A modest range of 30 to 5,000 bp was enough to capture over 99.2% of both WormBase and Iso-Seq introns (Figure 15).

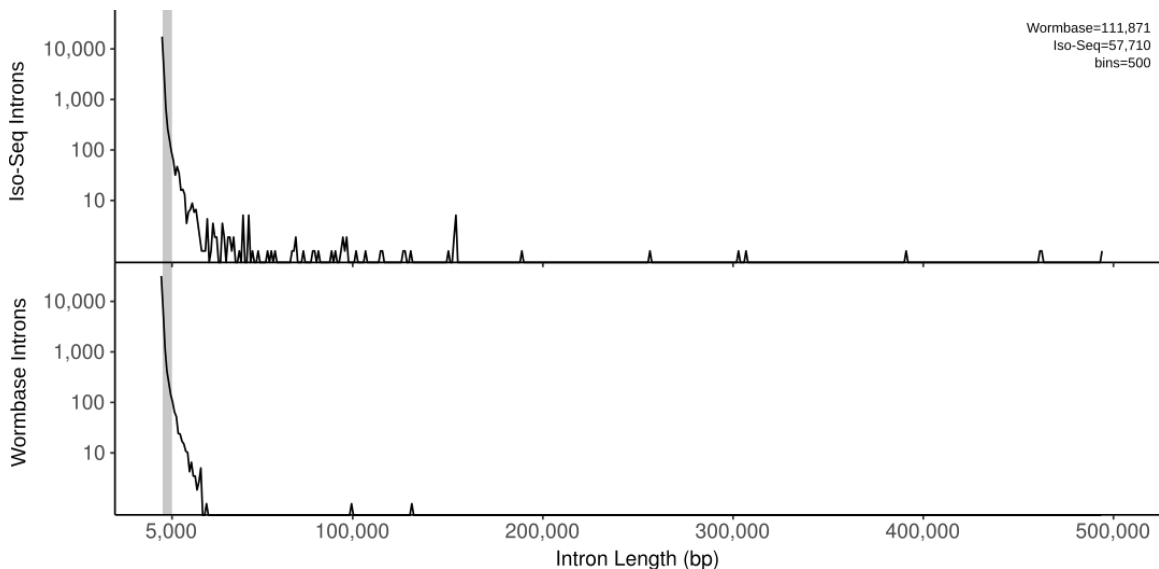


Figure 15. Distribution of intron lengths identified using Iso-Seq (top) and in WormBase (bottom). Y-axis is log10 scale. The shaded region indicates 30-5000 bp window containing >99% of Iso-Seq and WormBase introns, respectively.

2.2.6. Selecting a minimum support threshold for introns

A common criterion for accepting/rejecting introns is to set a minimum score threshold such that spurious splicing events (or technical artifacts) are removed, while rare but functionally important introns are retained. We predicted that many introns may be highly expressed in a few libraries, but lowly expressed in the majority. Therefore,

rather than imposing a minimum threshold on every library individually, we required that an intron have five or more supporting reads in at least one library to be accepted. A minimum threshold of five reads offers a compromise between minimizing spurious introns, while retaining the majority of WormBase introns (Figure 16).

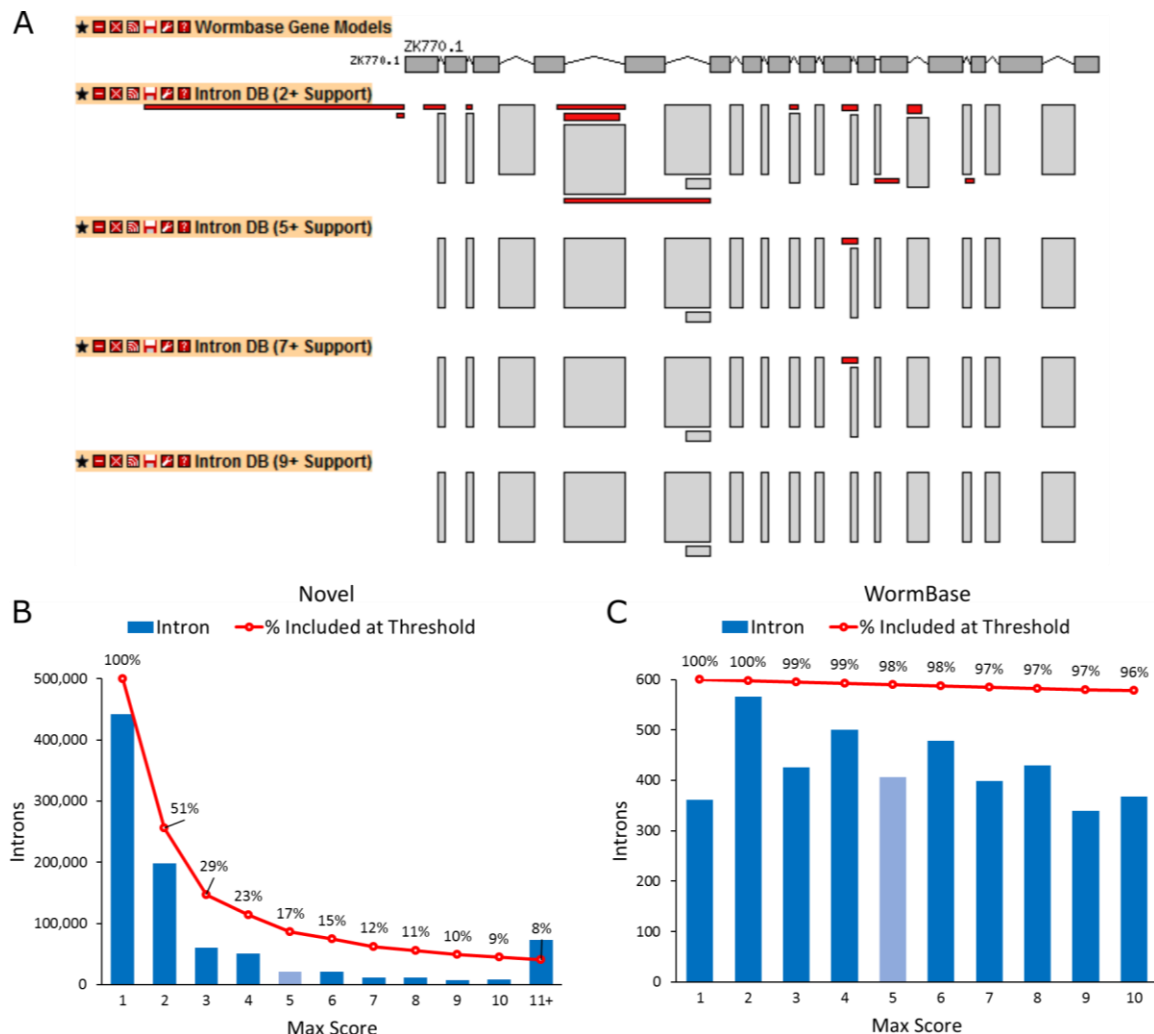


Figure 16. Effect of different minimum score thresholds on intron identification. (A) Example of the effect of setting various minimum scores on introns in *asic-1*. Introns in red are removed as filtering increases to a threshold of 10 reads. (B) Number of novel introns with the indicated maximum read support in any library. “% Included at threshold” (red line) denotes the number of introns retained when the minimum score is set to that value. A minimum score of 5 support was selected for use (light blue bar). (C) Introns present in WormBase that had 10 or less read support in any library.

2.3. Results

2.3.1. Validation of WormBase curated introns

Using 802 RNA-Seq libraries (1,075 runs) and the data selection and quality

filtration criteria described previously, we defined a database of 239,333 introns for *C. elegans*. Iso-Seq reads support 54,551 (22%) of these introns; 51,949 (46%) of these are annotated introns. Our intron database was used to validate introns in WormBase, both at the level of individual introns and the level of whole transcripts. Over 93% (*i.e.*, 104,384/111,871) of individual introns in WormBase are supported by our database (Figure 17B). Similarly, 87% (*i.e.*, 27,571/31,574) of WormBase transcripts of protein-coding genes had all introns represented in our database; almost all the remaining coding transcripts were partially supported.

In total, 7487 of WormBase introns were not supported by our database. Roughly 0.8% of WormBase introns were detected by RNA-Seq, but the amount of support fell below our cut-off of five reads in at least one library. Another 2% were excluded for being too long or short (<30 bp or >5,000 bp). Overall, 42% of WormBase introns that are not represented in our database were excluded based on these filtering criteria. Out of the remaining 4,345 undetected WormBase introns, many were either the first or last intron of the transcript (Figure 17D), reflecting the relative difficulty of obtaining complete transcript coverage of the terminal ends of genes (Hansen *et al.*, 2010). For 1,637 of these same introns, Iso-Seq supported a transcript that excludes them (Figure 17E). Unexpectedly, Iso-Seq supports 37 WormBase introns that were not detected with RNA-Seq. The majority of these (30/37) use non-GT/AG splice sites. STAR appears to have a bias against non-canonical splice sites; we observed that RNA-Seq reads at these near these non-canonical splice sites were mapped to nearby canonical GT/AG splice sites even when this introduced mismatches into the alignment that would not be present if the reads were mapped to the non-canonical split site. Iso-Seq reads were mapped using GMAP, which does not appear to share this bias. Of the 4,345 WormBase introns that were not supported by RNA-Seq, only 161 (2%) are listed as confirmed in WormBase (*i.e.* they have associated EST or cDNA support). Our reasoning for why these experimentally validated introns were not detected is generally the same as the whole population of unsupported WormBase introns: 45 were the 5'-most intron of the transcripts and 40 were the 3'-most intron, reflecting the difficulty of achieving end-to-end coverage of transcripts. Additionally, 55 use non-GT/AG splice sites, which may reflect a bias of STAR against non-canonical splice sites. The remaining introns (69 total) were internal to the transcripts, and five were from single-intron transcripts. One possible reason why these were missed is because their expression is exceptionally

rare. For example, *che-1* is predominantly expressed in ASE neurons (Uchida *et al.*, 2003) and contains one of these confirmed, but undetected introns. Another possibility is that the introns is only expressed under stress-response conditions. The undetected intron in *pdi-2* – a gene upregulated in response to endoplasmic reticulum stress (Glover-Cutter *et al.*, 2013) – may be such a case. A full list of confirmed undetected introns is listed in Supplemental Table 2.

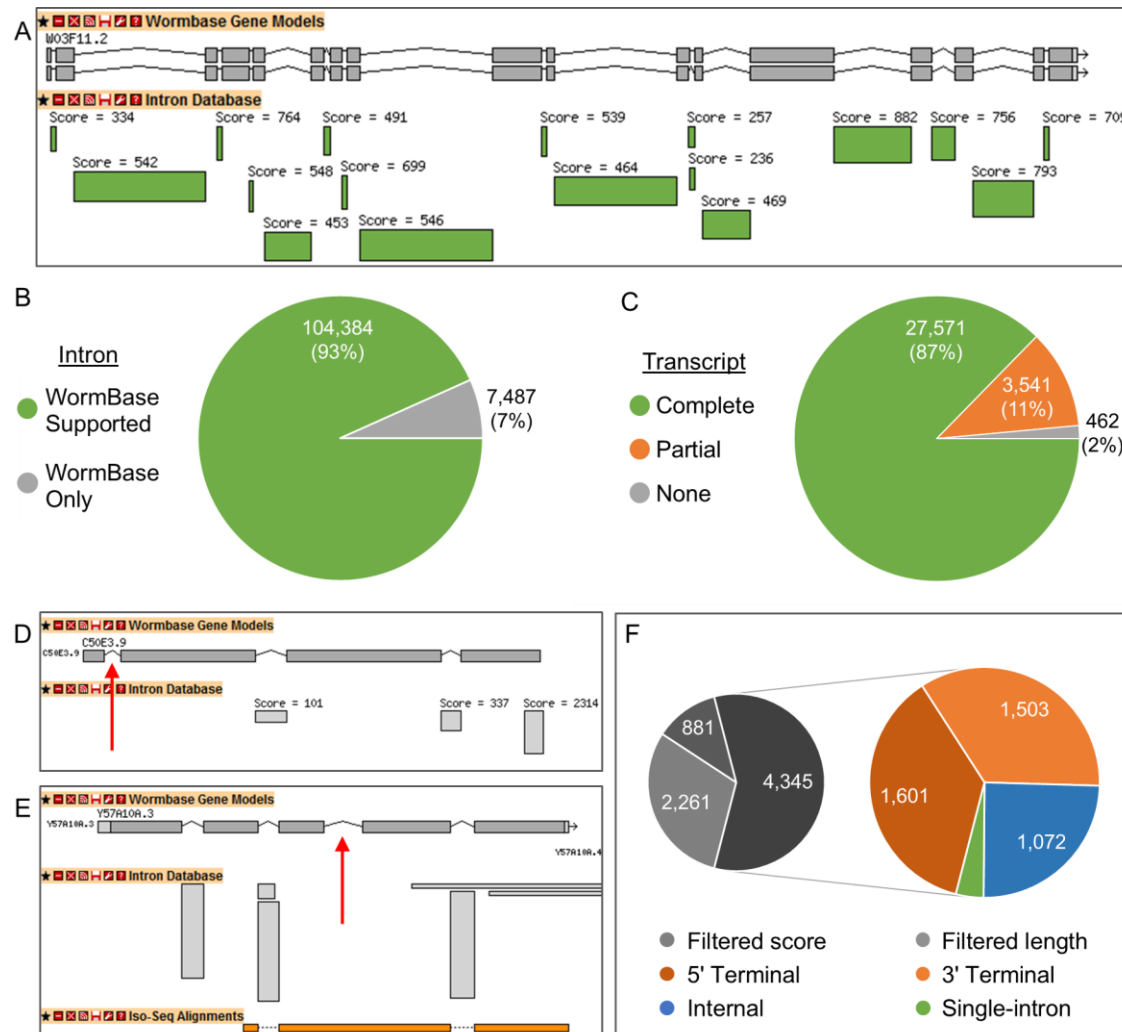


Figure 17. Introns identified from RNA-Seq was used to validate WormBase introns. (A) All introns in the gene models for *gcy-17* are represented in our intron database. (B) Individual introns in WormBase are either supported by our intron database or only found in WormBase. (C) Transcripts in protein-coding genes were validated at the level of introns. Transcripts were designated “complete” if all introns were represented in our database, “partial” if only some introns were represented, or “none” if no introns were represented. (D) Indicated WormBase intron (red arrow) at the 5’ end of the gene model for C50E3.9 is not represented in our database. (E) Indicated intron in WormBase gene Y57A10A.3 (red arrow) is not represented in our database. Iso-Seq data supports an alternative transcript that excludes this intron. (F) Breakdown of WormBase introns not represented in our database. Some introns fell below our minimum score threshold (“filtered score”), or outside our intron length thresholds (“filtered length”). Other introns are labelled according to their position within the transcript: “terminal” if they are the first or last intron, otherwise “internal,” or “single-intron” if the transcript only contains one.

2.3.2. Identification of novel introns

Our database includes 134,949 introns (over half the database) that could not be assigned to a WormBase transcript (Figure 18). A total of 2,602 (1.9%) of these putative novel introns are supported by Iso-Seq reads. At the level of individual splice sites: 9,551 (7.08%) of represent a novel combination of annotated splice sites; 49,056 (36.4%) include one novel splice site, and 76,342 (56.6%) have two novel splice sites.

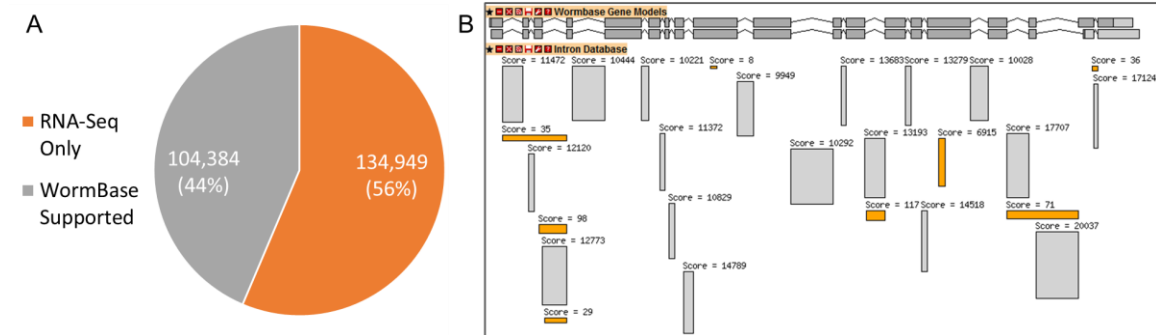


Figure 18. Identification of novel introns using RNA-Seq. (A) Breakdown of introns in our database that are either unique to our RNA-Seq analysis or represented in WormBase. (B) Pictured is the WormBase gene models for *aex-1* and our intron database. Novel introns are highlighted in orange

Novel introns suggest modifications to many of the protein-coding gene models in WormBase. The majority are entirely contained within the boundaries of a WormBase protein-coding gene model. Three-quarters of protein-coding genes contain at least one novel intron, representing a potential novel transcript. Novel introns also suggest necessary modifications to over half of these gene models – either extending the gene directly, extending the gene via a novel exon (see Chapter 3), or merging two separate gene models (Table 1). 16% of novel introns did not map to a known gene.

Table 1. Modifications to WormBase protein-coding gene models based on our intron database

Category	Novel introns	Protein-Coding Genes Affected
Internal novel intron	87,499 (64.8%)	15,395 (75.6%)
Directly extends gene	9,217 (6.83%)	4,901 (24.1%)
Extends gene via novel exon ¹	7,165 (5.31%)	2,444 (12.0%)
Merge genes	2,595 (1.92%)	2,613 (12.8%)
Pseudogene	5,876 (4.35%)	-
Non-coding gene	960 (0.71%)	-
Other	21,637 (16.0%)	-

¹Novel exons discussed in Chapter 3.

2.3.3. Globally rare vs. locally rare introns

We observed that many introns in our database, particularly novel introns, have a low number of supporting reads compared to other introns in the same gene. We investigated whether these introns that are comparatively rare in the context of our entire database, are non-rare under more-specific circumstances. We used the same evaluation method as Tourasse *et al.* (2017) to compare the relative levels of intron expression: Usage of a given intron is represented as the ratio of the number of reads supporting the intron divided by the number of reads supporting the most highly supported intron in the same gene. The ratio of read support for each intron was calculated for each individual library – the “local” ratio – and for each intron in the context of the entire database – the “global” ratio. In either context, ratios less than 0.01 were deemed “rare” and any above 0.01 “non-rare.” This approach revealed introns that are globally rare but become decidedly non-rare at specific intervals during embryo development (Figure 19).

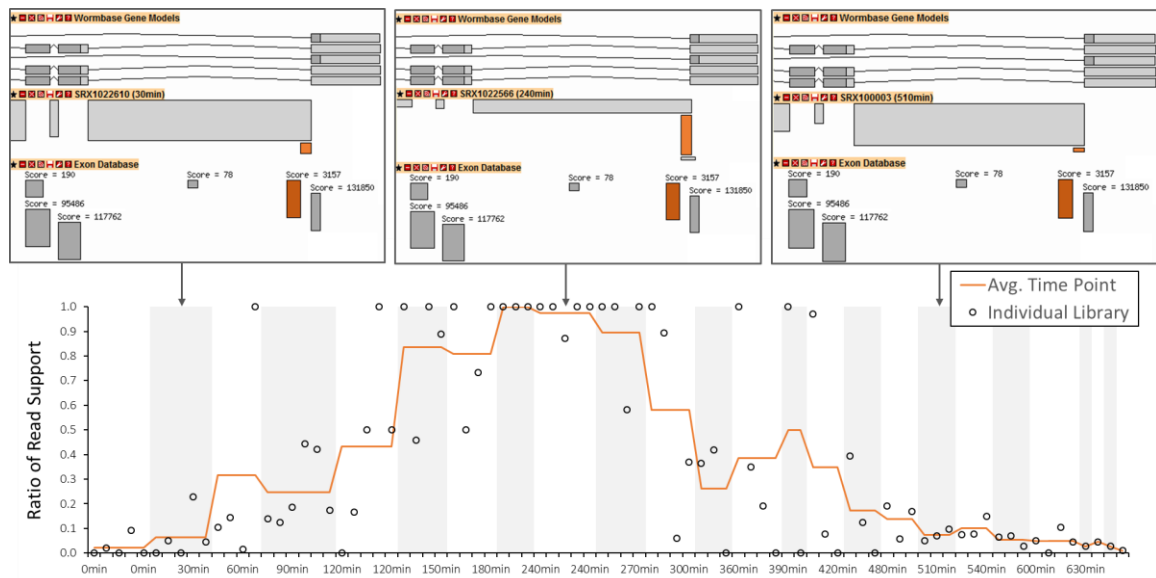


Figure 19. Relative usage of a novel intron and adjacent novel exon (orange) in *cki-2* (Cyclin-dependent Kinase Inhibitor) shown at 30, 240, and 510 minutes into embryo development. Height of the intron represent the relative ratio of read support compared to other introns in the gene in that library. Orange line (bottom) is the average ratio of read support between biological replicates (open circles) at that time point (shaded or non-shaded bars).

The majority (53%) of introns in our database were globally non-rare; almost all of these are represented in WormBase. Less than 1% of WormBase introns were rare either globally or locally, and almost all were the highest (or tied for highest) supported intron in the gene in at least one library. In contrast, only 16% of novel introns are globally non-rare, meaning that these could be interpreted as noise when they are only

evaluated in the context the entire database. However, 63% of novel introns are non-rare locally and 40% are expressed at least 10% of the level of the most highly expressed intron in the gene in at least one library. Over half of all introns in our database are non-rare globally, but 80% are non-rare locally (Figure 20).

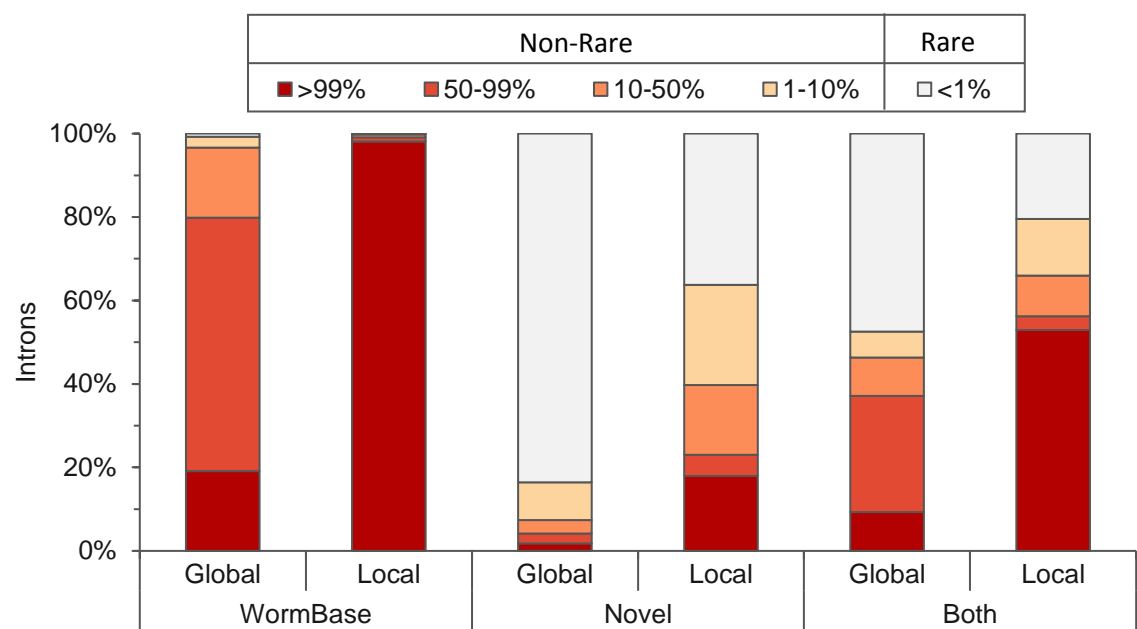


Figure 20. Number of WormBase introns and novel introns that are non-rare in the global context (ratio calculated for all libraries pooled together), and non-rare in the local context (ratio calculated for each library, individually).

2.3.4. Identification of embryo stage-specific introns

We were particularly interested in novel introns that show strong expression at a very specific time points during embryo development. We categorized the embryo time-series libraries, based on their sampling time, into “early” (0-330min), “mid” (360-600min), or “late” (630-720min) and searched for introns that have a high usage ratio (at least 0.7) in one stage and low usage ratios (less than 0.1) in the other two. 135 novel introns met these criteria; principally in early and late stages. We performed an over-representation enrichment analysis on genes that showed embryo stage-specific expression patterns using “DAVID” (Huang *et al.*, 2009) and found these genes are enriched for GO terms and functional annotations associated with transmembrane components (Figure 21).

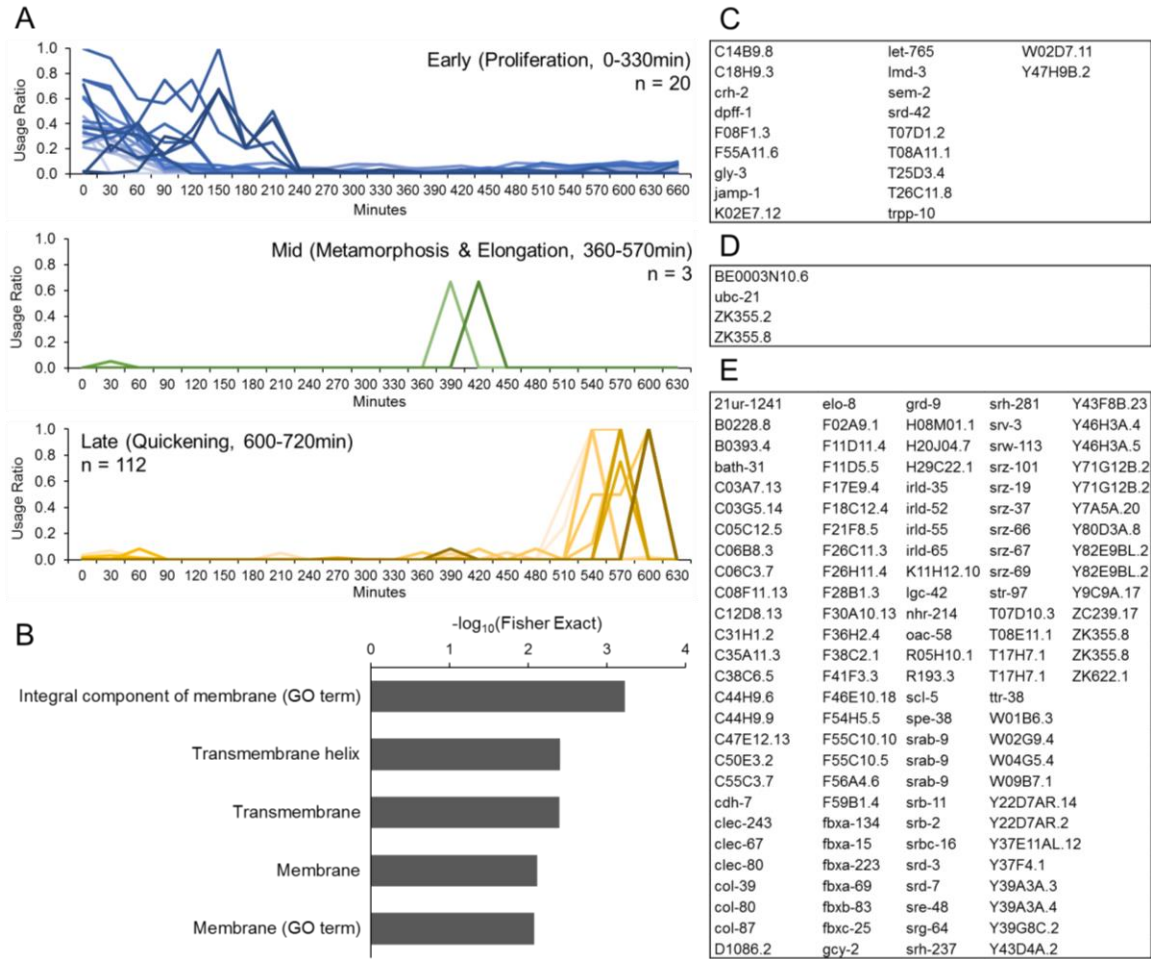


Figure 21. Set of novel introns that show strong expression in one stage of embryo development were found in a gene set enriched for transmembrane components. (A) 135 novel introns show very specific, strong expression at a certain time point during development (average usage ratio ≥ 0.7 in one stage and < 0.1 in the other two stages). (B) Over-representation analysis of the gene set containing the novel embryo-stage specific introns is enriched for GO terms and upregulated pathways involved in membrane components and transmembrane helices. Lists of genes containing the introns showing strong expression in (C) early, (D) mid, or (E) late embryo.

2.4. Discussion

In this chapter, we constructed a database of *C. elegans* introns. We employed a set of key quality control steps, including a novel approach for eliminating false positive introns introduced by multi-mapping reads, to minimize spurious splicing events and technical artifacts.

To demonstrate the utility of our database, we sought to validate introns in the WormBase gene models. We were able to validate over 93% of individual WormBase introns, and 87% of protein-coding transcripts at the intron level. Despite imposing

significant restrictions on which introns are accepted into our database – over 400,000 were excluded based on our filtering criteria – our approach remains sensitive enough to validate nearly all individual WormBase introns and coding transcript at the intron level. For the minority of WormBase introns that were not validated, a third were excluded based on our filtering criteria. We also found that some introns were missed due to imperfect transcript coverage; a third of unvalidated introns (not due to filtering criteria) were the 5'-most intron of the transcript, a region notorious for low coverage in poly-A selected RNA-Seq experiments. Only 2% of unvalidated introns have experimental evidence (e.g. EST) listed in WormBase, raising the possibility that some transcripts may be mispredictions. For over a thousand unvalidated introns, our database is supported by Iso-Seq evidence for an alternative transcript that excludes these introns. In such cases, our intron database can serve as a guide for follow-up for experimental validation.

Our database is not limited to only WormBase introns: Out of the 239,333 introns in our database, over half (134,949) are not represented in any WormBase transcript. Over three-quarters of WormBase protein-coding genes have at least one new intron. These novel introns suggest additional transcripts or modifications to the current WormBase gene models. The most common modification to a gene model is extension via a novel intron, suggesting a novel exon lying outside the boundaries of the current gene model. Some introns fell completely outside the boundaries of the current gene models but could be linked back a gene via a novel exon (identification of novel exons is discussed in Chapter 3). Other introns that fell outside the current gene models either represent a hitherto unknown intron-containing UTR, or an artifact introduced in a complex genomic region. The latter often manifesting as clusters of low-support, overlapping introns commonly found in repeat-containing intergenic regions. We identified over twenty-five hundred cases where an intron exists that spans between two separate gene models, suggesting the two genes should be merged. While this is an exciting possibility, it is complicated by the fact that roughly 15% of *C. elegans* genes are expressed as part of an operon (Spieth *et al.*, 1993), meaning that some of these cases may be caused by reads originating from polycistronic transcripts. Experimental validation is needed to distinguish between the two possibilities.

The finding of many novel introns agrees with the finding of similar large-scale RNA-Seq meta-analyses (Gerstein *et al.*, 2010; Hillier *et al.*, 2009; Ramani *et al.*, 2011; Boeck *et al.*, 2016), which suggest that there is far more alternative splicing occurring in

C. elegans than what is represented in the current gene models. An alternative explanation is that most introns detected by RNA-Seq are simply spurious splicing events, not functional introns. This is argued by Tourasse *et al.* (2017) who found that 88% of introns they identified using RNA-Seq are “rare” (have 100-fold less read support across the whole database than the highest supported intron of the parent gene) and more highly expressed genes had more rare introns (*i.e.* higher expression provides more chances for “misfiring” of the spliceosome). We found that only 47% of our database was rare using the same definition of “rare” and “non-rare”. The lower percentage of rare introns in our database compared to that of Tourasse *et al.* (2017) suggests that our comprehensive filtering and intron acceptance criteria is effective at minimizing biological noise. We note that most of non-rare introns in our database are those that are represented in WormBase. Only 16% of novel introns were non-rare globally, however 63% were non-rare locally. This finding shows the importance of considering biological context when evaluating introns; only considering the global context may understate the functional relevance of many introns. In total, 80% of our intron database is non-rare locally, with over half becoming the dominant transcript of the parent gene in at least one local context. Alternative splicing is often highly regulated. For a globally rare intron to undergo a dramatic increase in relative usage in one or more libraries is not only plausible, but also implies a functional role for the intron plays in those libraries. We identified 135 novel introns that were extreme cases of this – introns that are highly specific to a certain stage of embryo development. Overall, we expect that most, if not all introns in our database are functional under certain circumstances.

Chapter 3. Building a high-quality exon database

3.1. Introduction

In Chapter 2 we described the construction of a high-quality intron database from 802 RNA-Seq libraries, which were used to validate WormBase introns and identify novel introns. Not all introns necessarily contribute to the coding capacity of *C. elegans*; some instead have a regulatory role. Alternative splicing may only alter the UTR, or an intron can introduce a frame-shift into the transcript, triggering degradation via NMD (Soergel *et al.*, 2013). However, introns are only one component of eukaryotic protein-coding transcripts. Coding exons are the segments of a transcript that together encode a protein product. Therefore, identifying all coding exons provides us with a useful metric for evaluating the coding capacity of the *C. elegans* genome.

Here we present an algorithm for reconstructing exons with protein-coding potential named “ExonTrap” after the conceptually similar experimental technique for detecting exons using an intron-containing vector (Auch and Reth, 1990). This algorithm uses our intron database and “translation blocks” encoded in the genome. Using ExonTrap we constructed a high-quality exon database. We demonstrate its utility by validating coding exons in the WormBase gene models and identifying a sizeable number of novel exons.

3.2. Algorithm

ExonTrap is an algorithm for reconstructing exons using two lines of evidence: “Translation blocks” encoded in the genomic sequence, and our intron database from Chapter 2. Translation blocks are consecutive regions from one stop codon (TAA, TGA, or TAG) to the next stop codon in each of the six reading frames of the genome (referred to as +0, +1, +2, -0, -1, and -2, respectively). Translation will not continue through a stop codon, which means a translation block represents the longest stretch of DNA that can encode a protein-coding exon. Introns represent direct evidence for exon boundaries, so are used to refine exon boundaries within each translation block. Exons are defined as a unique pair of genomic coordinates corresponding to the first and last base of a contiguous coding region. Therefore, two exons can overlap and still be considered

distinct exons provided their first and/or last bases differ. Over 98% of annotated *C. elegans* protein-coding transcripts contain at least one intron, so this approach theoretically allows us to capture nearly all coding exons. At the time of writing, our exon reconstruction algorithm does not include a method for identifying single-exon genes. While several approaches relying on depth of read alignments were explored, but none were identified that could accurately identify the remaining 531 (2%) of protein-coding transcripts that lack an intron.

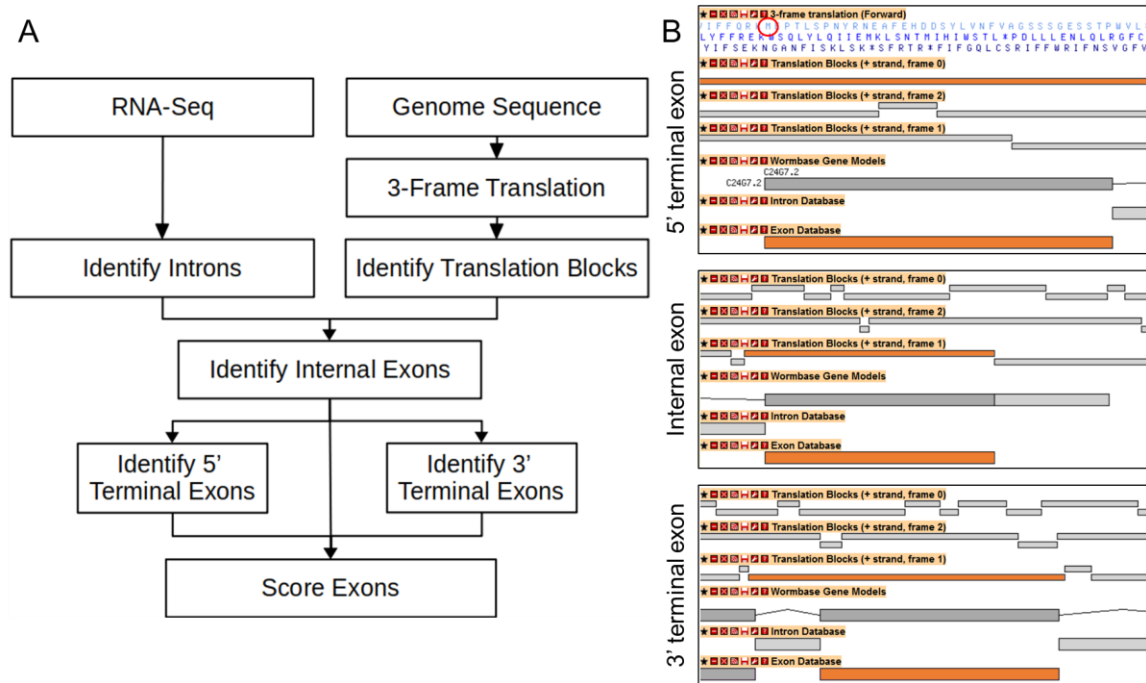


Figure 22. (A) Schematic representation of the exon reconstruction algorithm. (B) Examples of internal, 5' terminal, and 3' terminal exons with the relevant exon and translation block highlighted in orange.

Exons in a transcript can be categorized as either “terminal” or “internal” depending on their relative position within the transcript. Terminal exons are either the most 5' or most 3' exon of the transcript, which and are bounded by a splice site and a start codon – ATG in *C. elegans* (Riddle *et al.*, 1997b) – or a stop codon, respectively. Internal exons are the remaining exons that fall within the interior of the transcript.

The defining characteristic of an internal exon is that it is bounded at both ends by splice sites. Therefore, we can use our intron database to precisely determine internal exon boundaries. Internal exons are defined by the combinatorial pairing of each base immediately downstream of each splice acceptor to each base immediately upstream of each splice donor, where both the donor and acceptor sites are located within the same

translation block.

Cases where an intron is entirely contained within a translation block and flanked by additional introns on both sides are characteristic of an intron retention event. It is possible that such a case simply occurs by coincidence – long translation blocks may contain several introns. To minimize false positives, only cases where at least 80% of the intronic bases are covered by two or more aligned reads are accepted as an intron retention event.

Once all internal exons are defined, they are used to guide the identification of terminal exons. A cursory search of our dataset identified several million cases where both an ATG codon and a splice donor fall within the same translation block, which is the defining characteristics of a 5' terminal exon. Similarly, any translation block with a splice acceptor could also be considered to have a 3' terminal exon unless some other selection criteria is applied. We predicted most of these cases do not contain a real terminal exon. To narrow down the list of potential terminal exons, we enforce two rules: First, a terminal exon cannot overlap an internal exon. While this approach reduces sensitivity to alternative start codons and poly-A sites (Figure 23B) it does offer a much higher degree of accuracy than would be achieved if all of the potential terminal exons were accepted. Second, the terminal exon must be in a translation block such that it maintains the open-reading frame if both exons are translated together. The open-reading frame is maintained between exons according to the following equations:

$$X = (Y - Z) \text{ modulo } 3 \quad \text{if the terminal exon is a 5' terminal exon, or}$$

$$X = (Y + Z) \text{ modulo } 3 \quad \text{if the terminal exon is a 3' terminal exon.}$$

Where X is the reading frame of the terminal exon, Y is the reading frame of the internal exon, and Z is the length of the intervening intron (in bp) modulo 3. For example, if an internal exon is located on the +1 reading-frame, and the intervening intron is 31 bp long, then an adjacent 3' terminal exon would be located on the +2 reading-frame.

Precisely defining transcripts ends is a challenging task. In particular, obtaining adequate sequence coverage of the 5' ends of transcripts is notoriously difficult – so much so that alternative protocols have been developed to specifically target these regions (Hwang *et al.*, 2004; Yeku and Frohman, 2011). To accurately identify 5'

terminal exons we used the following logic: In *C. elegans*, the coding region always starts at an ATG codon (Riddle *et al.*, 1997b). Therefore, we select an ATG codon as the start of the 5' terminal exon and the last base before the downstream splice donor as the 3' end of the exon, provided both the ATG codon and the splice donor appear in the same translation block. If multiple ATG codons are present, the one closest to the 5' end of the translation block is selected (Figure 23A). We observe that over two-thirds of 5' terminal exons in WormBase have an exact match to a 5' terminal exon in our database, and most of the remaining 5' terminal exons in WormBase have a partial match (Figure 23C), indicating that our approach produces accurate results.

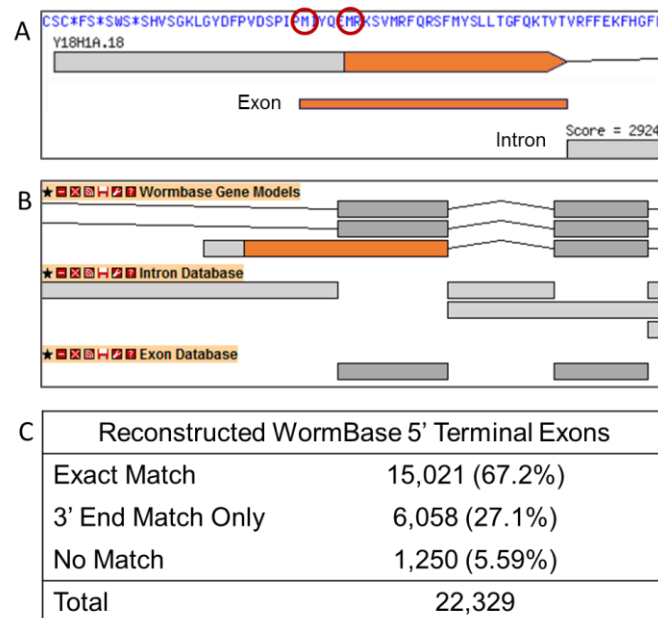


Figure 23. Approaches taken to resolving 5' terminal exon boundaries. (A) When multiple putative start codons (methionines) are in frame, the one most distal from the splice donor is selected as the start of the 5' terminal exon. (B) Alternative start sites overlapping a longer transcript are partially missed because internal exons are identified first and are mutually exclusive with terminal exons. Cases like this are considered a partial match if the 3' end of the exon matches a 3' end of one of our exons. (C) Number of WormBase 5' terminal exons that have an exact match, a partial match, or no match in our exon database.

3' terminal exons are bounded by a splice acceptor at the 5' end and the end of the translation block (*i.e.* a stop codon) at the 3' end. Translation will not proceed past a stop codon, so the 3' end of every 3' terminal exon is always the first incidence of a UAA, UAG, or UGA codon. To minimize false positives during the reconstruction of 3' terminal exons, we only look for potential 3' terminal exons that do not overlap internal exons.

Once all exons are defined, they are assigned a score that represents the

amount of read support for their existence. This score is a conservative estimate based on the number of reads supporting flanking introns. To obtain this score, first we count the number of reads supporting all introns flanking the 5' end of the exon. Next, we count the number of reads supporting flanking introns at the 3' end. To avoid over-estimating the relative expression of a given exon, we use the lower of these two values as the exon score.

3.3. Results

3.3.1. Evaluating the accuracy of exon reconstruction

To test the utility of ExonTrap and our newly constructed exon database, we compared the exons in our database to the coding exons in the WormBase gene models (those labelled "CDS"). A WormBase exon was considered to have an exact match in our database if the position of the first and last bases of the exon match those of an exon in our database. Some WormBase terminal exons only had a partial match, which means that one end of the exon matched one end of an exon in our database but differed at their other end. For example, a 5' terminal exon may match at the side adjacent to the splice donor but differ at its choice of start codon. Only terminal exons may have a partial match; internal exons had an exact match or no match at all. Due to the way introns are used to guide exon reconstruction, the quality of our exon database is intrinsically tied to the quality of our intron database. Since we were able to validate nearly all WormBase introns with our intron database, we predicted that we would be able to reconstruct nearly all WormBase coding exons. Indeed, the boundaries of 85% (*i.e.* 111,378/131,674) of individual WormBase coding exons are reconstructed exactly using our algorithm (Figure 24); another 9% (12,022) are a partial match. The 6% (8,274) of WormBase exons that have no match in our database are those where an adjacent intron is not represented in our intron database.

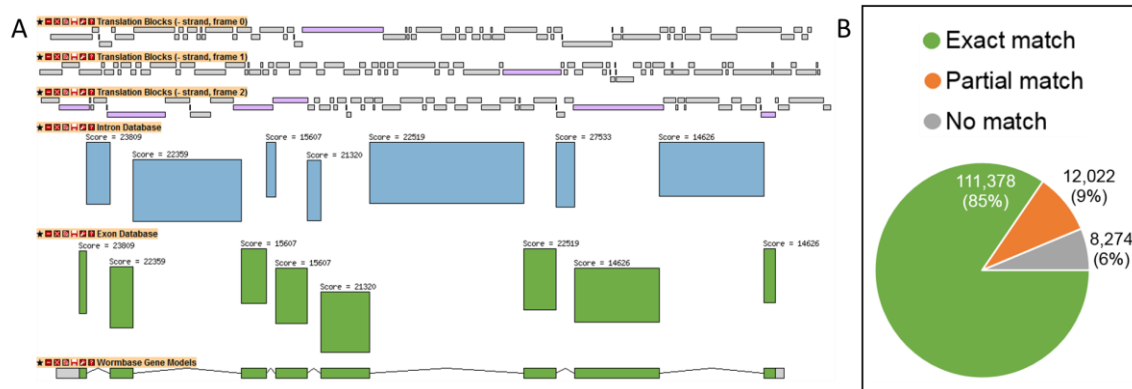


Figure 24. (A) Example illustrating how translation blocks (relevant blocks in purple) and introns (blue) are used to define the boundaries of coding exons (green). The exons defined in this example match those of the gene model for *ric-19* exactly. (B) Breakdown of WormBase coding exons that are represented in our database. Terminal exons that differs only at the terminal end are deemed a partial match. WormBase coding exons not represented in our database are “no match.”

3.3.2. Validation of the WormBase transcripts

In previous sections we described the creation of two high-quality databases for introns and exons. Here, we used both databases to validate WormBase protein-coding transcripts. Most transcripts in their current state have extensive experimental support – the product of over 20 years of work by WormBase staff and independent investigators alike. Our expectation was that most (if not all) protein-coding transcripts should be completely supported by our intron and exons. To test this hypothesis, we compared the individual introns and exons of each protein-coding transcript to those in our databases before categorizing the transcript as follows:

- *Complete* – All introns and exons in the transcript are included in our databases.
- *Partial* – Some (but not all) introns and exons are included.
- *None* – None of the introns or exons are included.

A given WormBase transcript was considered validated if all the introns and exons that make up the transcript are represented in our databases. We checked for exact matches between our databases and WormBase with some leeway given to the ends of the transcripts. Due to the challenge of accurately defining transcript ends, we allowed the 5' end of 5' terminal exons and the 3' ends of 3' terminal to differ from those in our database and still be considered a match provided the other end (the end adjacent to the splice site) matched an exon in our database.

We found that 84% (*i.e.* 26,433/31,574) of WormBase coding transcripts are completely validated by our databases (Figure 25). Another 12% (3,854) were partially validated; 25% of partially validated transcripts had an intron that was too long/short based on our filtering criteria, so we did not expect to completely validate these. The remaining partially validated transcripts generally had incomplete coverage, indicating a universally low level of expression across all 802 libraries. Only 4% (1,287) of protein-coding transcripts had no supporting evidence in our databases; 894 of these are single-exon transcripts which we could not reliably reconstruct using the methods described. In addition, 532 of the non-validated transcripts are listed as “predicted” in WormBase; lacking experimental evidence.

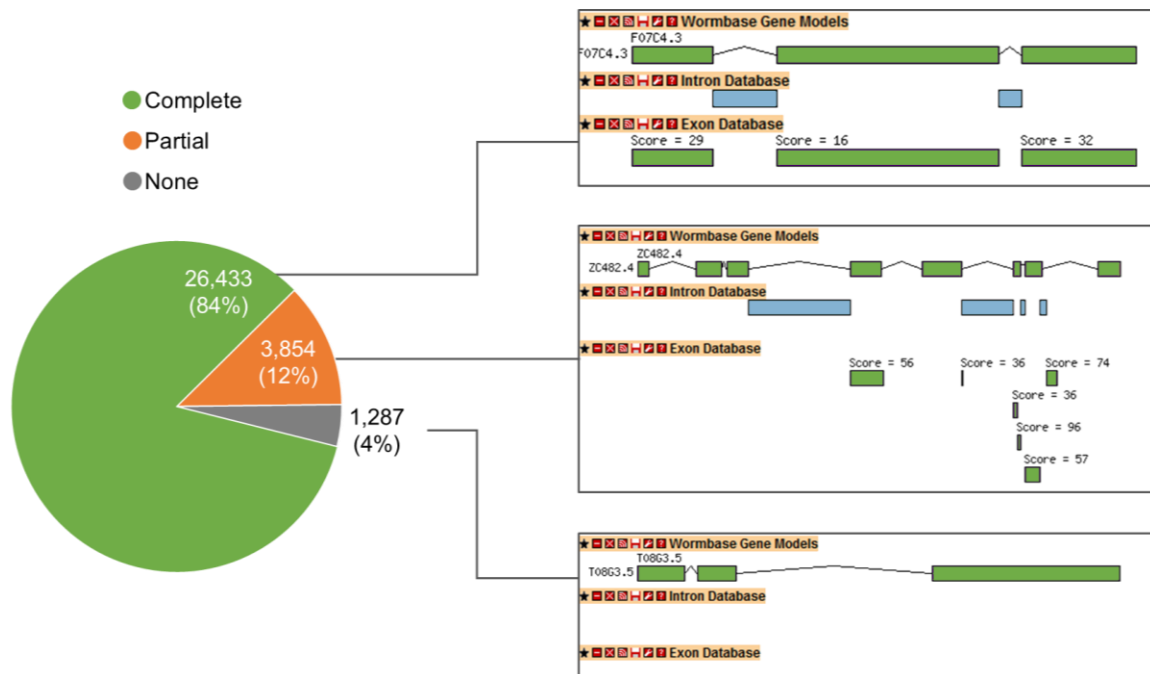


Figure 25. Percentage of WormBase protein-coding transcripts that are completely, partially, or not supported by our intron and exon databases. Example gene models, from top to bottom, are *srd-29*, *irld-61*, and *srh-141*.

3.3.3. Identification of novel exons

Our exon database contains almost all exons present in WormBase protein-coding transcripts, yet we identified 204,812 exons (62% of our database) that do not match any known WormBase coding exon (Figure 26).

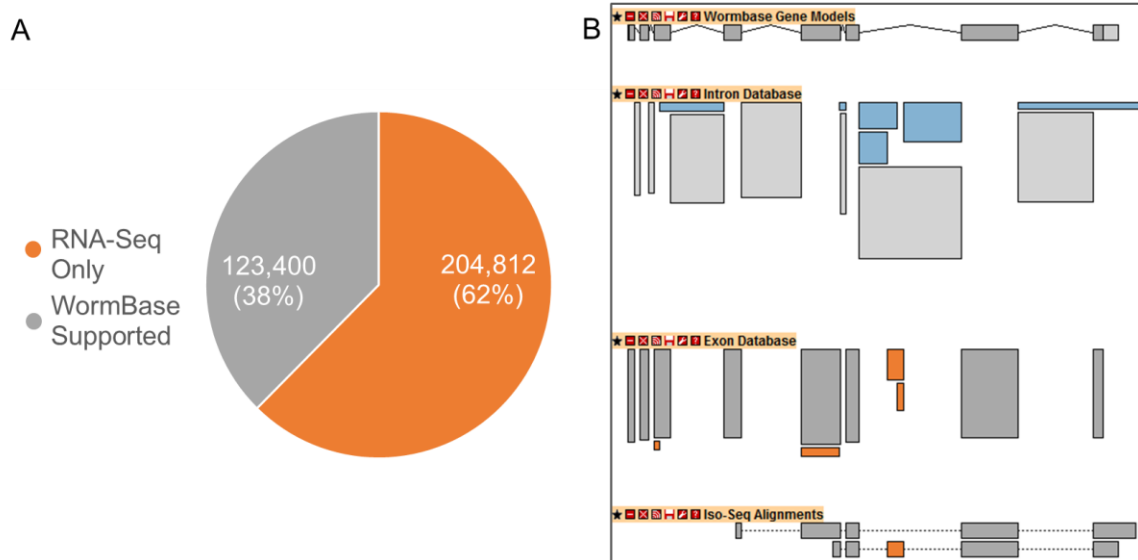


Figure 26. Identification of novel exons using RNA-Seq. (A) Breakdown of exons in our database that are either unique to our RNA-Seq analysis or represented in WormBase. (B) Novel exons (orange) identified for *ceh-93* from novel introns (blue). The highest scored novel exon is supported by Iso-Seq data.

While most of the novel exons fall within annotated gene boundaries, there are thousands that suggest modifications to many of protein-coding gene models in WormBase (Table 2). Over 20% of gene models are extended by a novel exon directly or by a novel exon linked to an extant gene model via a novel intron in our database. Just under 7% of novel exons could not be linked to a WormBase protein-coding gene model: 4.6% (9,392) of novel exons overlap a non-coding gene, with most of these (7,419) being pseudogenes genes where the intron-exon structure mimics that of protein-coding genes well enough that it was falsely picked up by our algorithm.

Table 2. Modifications to WormBase protein-coding gene models based on our exon database

Category	Exons	Protein-Coding Genes Affected
Internal novel exon	150,913 (46.0%)	16,715 (82.1%)
Directly extends gene	5,433 (1.65%)	4,025 (19.8%)
Extends gene via novel intron	20,391 (6.21%)	4,385 (21.5%)
Links multiple genes	684 (0.21%)	444 (2.18%)
Overlaps non-coding gene	9,392 (2.86%)	-
Overlaps pseudogene	7,419 (2.26%)	-
Other	13,955 (4.25%)	-

3.3.4. Globally rare vs. locally rare exons

The relative usage of exons in our database was evaluated at both the global level and local level, using the same approach as for introns described in Section 2.3.3. The results here echo those for our intron database, with some variation: 58% of exons in our database were globally non-rare, with most of these being exons that are represented in WormBase. A slightly lower percentage of novel exons were globally rare compared to novel introns. This is because many rare introns overlap a WormBase 5' UTR where coverage is lower (which affects our calculation of the usage ratio for that intron), whereas exons are limited to the coding portion of the gene. Most novel exons (72%) are non-rare locally with half expressed at least 10% of the level of the most highly expressed exon in the gene in at least one library. In total, almost all exons (82%) are non-rare locally (Figure 27). Like our intron database, the relative usage of many exons increases under specific circumstances.

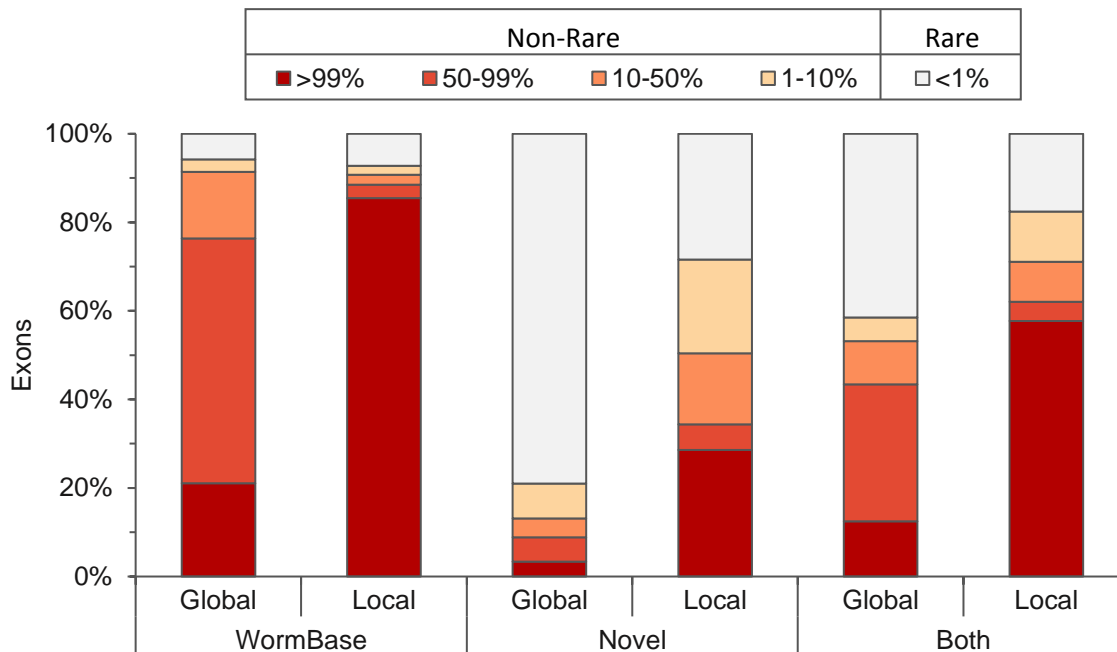


Figure 27. Relative usage of WormBase exons and novel exons in the global context (ratio calculated for all libraries pooled together), and the local context (ratio calculated for each library, individually).

3.4. Discussion

In this chapter, we introduced an algorithm, ExonTrap, for reconstructing protein-coding exons using a genome sequence and a set of introns. ExonTrap takes advantage of the defining characteristics of eukaryotic coding exons to accurately identify exon boundaries. Specifically, translation halts at a stop codon, which allows us to define

“translation blocks” which contain all protein-coding exons. While this approach is not able to reconstruct non-coding exons (including UTRs) it does allow us to define exon boundaries more precisely than an approach based on read coverage alone. We demonstrated the accuracy of this approach by precisely reconstructing 85% of protein-coding exons in the WormBase transcripts, and partially reconstructing additional 9%.

Our hypothesis was that much of the coding capacity of *C. elegans* is not represented in the WormBase gene models. Therefore, we expected that not all reconstructed exons could be assigned to a WormBase transcript. A surprising two-thirds of our exon database (204,812 exons) do not match any WormBase transcript. While a minority of these exons map to non-coding genes (particularly pseudogenes), most novel exons represent evidence of potentially novel protein-coding transcripts. Nearly all novel exons directly overlap or can be linked to a protein-coding gene in WormBase. These exons suggest modifications to thousands of gene models; the most notable finding is that the coding sequence of over 20% of protein-coding genes appears to be larger than the annotated coding sequence, *i.e.* these gene models should be extended at one or both ends. One reason these exons are absent in the current gene models may be that they are comparatively rare in the context of the entire database. Nearly all WormBase coding exons that are represented in our database are non-rare globally. In contrast, less than 20% of novel exons are non-rare globally. The finding that most exons in our database are globally rare could imply that they are artifacts introduced by spurious splicing events that were included in our intron database. However, when we look at the local context of each exon, we find that 70% are non-rare under specific circumstances and 30% are the highest supported exon of their parent gene in at least one library. While most novel exons are, on average, lowly expressed across the whole of *C. elegans* life cycle, they are highly expressed in specific stages or tissues. Our exon database is evidence that much of the coding capacity of the *C. elegans* genome is missing from the current transcript annotations.

Chapter 4. Assembling transcripts and evaluating coding capacity of *C. elegans*

4.1. Introduction

Our hypothesis that the full coding capacity of *C. elegans* is not represented in the current WormBase gene models is supported by our finding that 134,949 introns, and 204,812 exons in our databases could not be assigned to any WormBase transcript. Our expectation was that these introns and exons belong to transcripts that are not represented from the gene models. Therefore, to quantify the full coding capacity of *C. elegans* we had to use additional methods to capture the full set of protein-coding transcripts encoded by the *C. elegans* genome.

In this chapter, we discuss the construction of a set of transcripts in *C. elegans* that are fully supported by our intron and exon databases. Transcripts were assembled using multiple publicly available transcript assembly programs and the same set of publicly available RNA-Seq libraries that we used to construct our intron database. Supported transcripts were evaluated for protein-coding potential. Finally, we used our supported transcripts to evaluate the coding capacity of *C. elegans*.

4.2. Constructing a fully-supported set of transcripts

4.2.1. Assembling transcripts from RNA-Seq data

Three publicly available transcript assembly programs were used to assemble transcripts from the same 802 RNA-Seq publicly available libraries used to construct our intron database. Our rationale is that using multiple programs will compensate for potential limitations that any individual program may have and result in a more complete set of transcripts. Two of the programs – Cufflinks version 2.2.1 (Trapnell *et al.*, 2012) and Stringtie version 1.3.4d (Pertea *et al.*, 2015) – use a reference-based approach where reads are mapped to a reference genome prior to assembly. The third program – *Trans*-ABYSS version 1.5.5 (Robertson *et al.*, 2010) – assembles reads *de novo* into transcripts that were mapped to the genome using GMAP (median percentage of mapped transcripts per library was >99.9%). Each program was run with default parameters. Transcripts were assembled for each library individually using all three

programs and then merged in a subsequent step.

4.2.2. Selecting transcripts that are fully-supported by our intron and exon databases

To minimize the effects of biological and technical noise on transcript assembly, we only accepted transcripts where all introns and all exons in the transcript are represented in our intron and exon databases, respectively. Internal exons must match those in our database exactly. Due to the difficulty of precisely defining start and poly-A sites within a transcript, we allowed terminal exons to differ at their terminal end so long as the other end of the exon (adjacent to the splice site) matched that of an exon in our exon database. On average, 60-74% of transcripts per library were fully supported by our databases depending on the program used to assemble them (Figure 28). 92% of exons in our exon database were assigned to an assembled transcript.

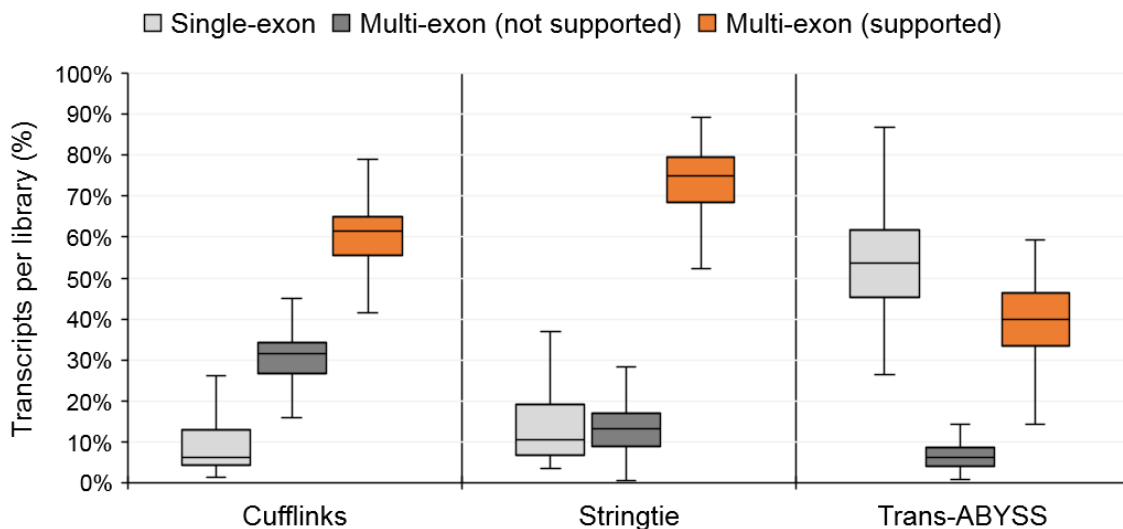


Figure 28. Percentage of transcripts per library, from 802 RNA-Seq libraries, that are fully supported by our intron and exon databases. Box boundaries denote first and third quartile, horizontal line denotes the median, whiskers denote minima and maxima. Note: single-exon transcripts are not supported by our databases.

4.2.3. Selecting full-length supported transcripts

Our goal was to identify full-length transcripts in *C. elegans*. Libraries with low sequencing depth may only achieve partial coverage of transcripts which results in the program only assembling a fragment of the full-length transcript (a “transfrag”) to be assembled. To preferentially select for full-length transcripts, we merged the supported

transcripts identified for each library into a unified set. We used GffCompare version 0.10.1 (<http://ccb.jhu.edu/software/stringtie/gffcompare.shtml>) to merge the transcripts. GffCompare was run with the “-X” parameter which specifies that transcripts will be merged if they share the same introns, or if a transcript contains a set of introns that are a subset of those of a longer transcript. Merging occurs even if the transfrag ends stick out into the intron of a longer transfrag (illustrated in Figure 29A). The advantage of this approach is that it favours the inclusion of full-length transcripts, however the downside is that it potentially discards valid transcripts with alternative start and/or poly-A sites. The total number of supported transcripts after merging was 99,627 (Figure 29B).

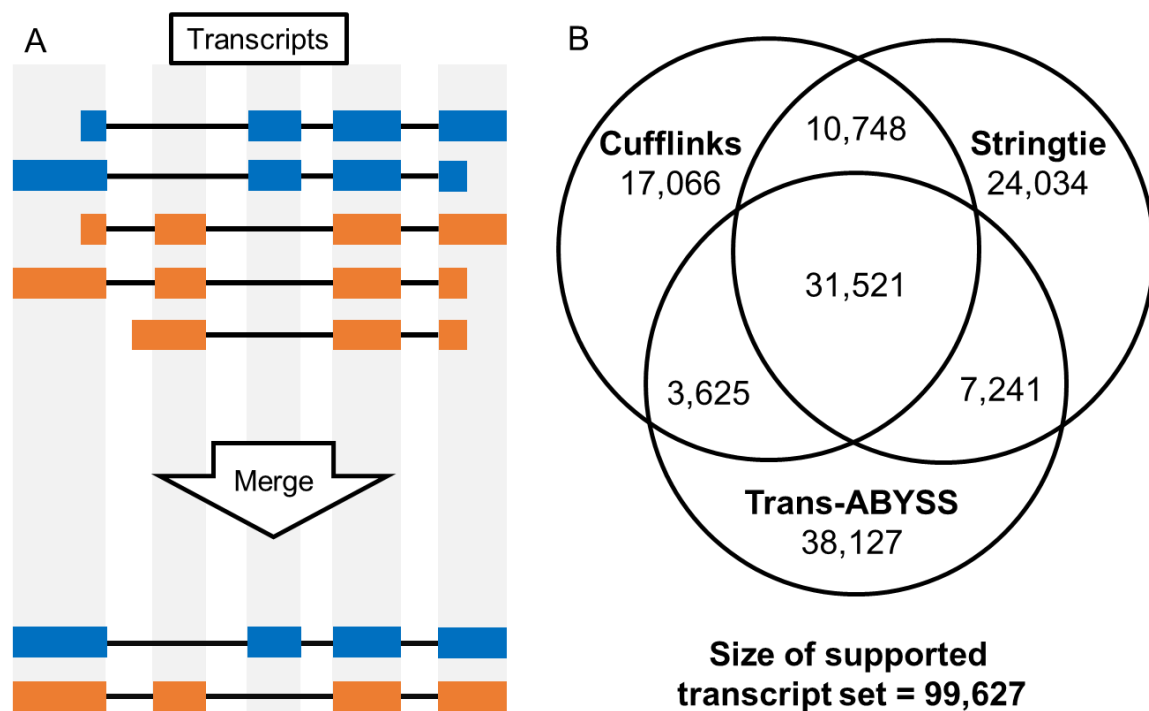


Figure 29. (A) Illustration of how assembled transcripts with the same introns, or a subset of introns of a longer series are merged. (B) Number of transcripts after merging that are fully supported by our intron and exon databases.

4.2.4. Assessing supported transcripts for coding potential

Not all alternative splicing events result in a protein product. One function of alternative splicing is to regulate gene expression; shifting the reading frame of a transcript through an alternative splicing event can result in pre-mature translation termination, ultimately triggering nonsense-mediated mRNA decay. To evaluate which supported transcripts potentially encode a functional protein product, we used TransDecoder version 5.0.1 (<http://github.com/TransDecoder/TransDecoder>) to identify

candidate coding regions. Out of the 99,627 supported transcripts, 74,581 (74.9%) had a candidate coding region, together encoding 72,274 unique coding sequences (Figure 30). 60% of exons in our exon database were assigned to a supported protein-coding transcript.

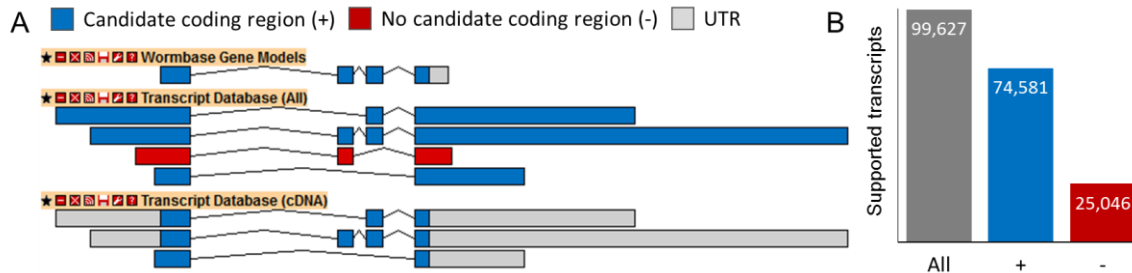


Figure 30. Identifying candidate coding regions in assembled transcripts. (A) Candidate coding regions identified within supported transcripts mapping to *sptrf-2*. (B) Total number of supported transcripts with, or without, a candidate coding region.

4.2.5. Evaluating the accuracy of supported transcripts

To assess the accuracy of the structure of our set of supported coding transcripts, we compared them against both WormBase protein-coding transcripts and Iso-Seq transcripts (Figure 31). A transcript was considered a match to either a WormBase or Iso-Seq transcript if they contained the same series of introns. This metric allows the 5' and 3' terminal exons to vary in length and still be considered a match. Transcripts were considered to extend a WormBase or Iso-Seq transcript if they contained the same series of introns plus additional introns extending beyond the boundaries of the WormBase or Iso-Seq transcript. Transcripts that were a subset of a longer WormBase or Iso-Seq transcript were treated as a novel transcript, as we are not able to distinguish whether they represent a modification to the current transcript model or are merely a fragment of the full-length transcript caused by lack of coverage.

Nearly all (>99%) of protein-coding genes in WormBase were overlapped by one or more supported coding transcripts. In total, 86% (27,055/31,574) of WormBase protein-coding transcripts were an exact match, or extended by, one of our transcripts; 14% (4,519/31,574) of WormBase protein-coding transcripts had no match to any of our transcripts; 7.2% (2,262) of WormBase protein-coding transcripts were missed because they contain introns excluded by our filtering parameters (Figure 31).

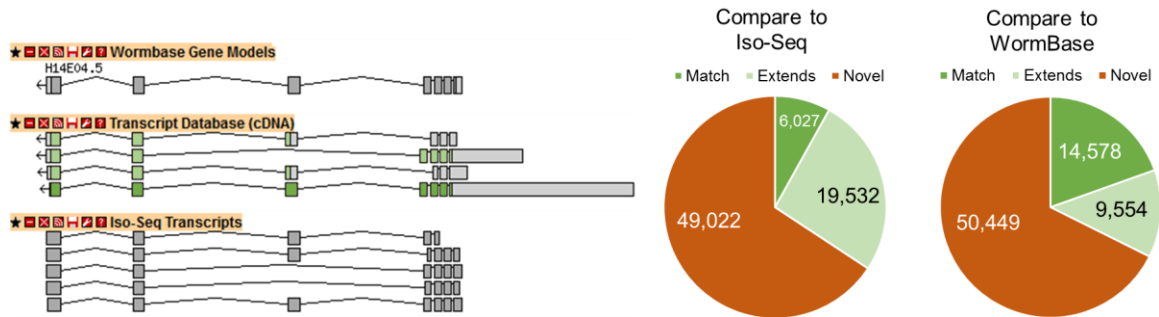


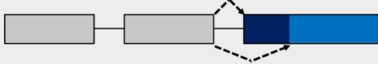


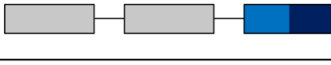


Figure 31. Comparison of supported RNA-Seq transcripts to WormBase transcripts and Iso-Seq reads. Comparison is based on the series of introns in each transcript: (Left) The WormBase transcript model for *cic-1*, the supported RNA-Seq transcripts, and the Iso-Seq transcripts overlapping the gene. (Right) Percentage of supported RNA-Seq transcripts that match or extend a WormBase or Iso-Seq transcript.

4.3. Evaluating the Coding Capacity of *C. elegans*

We compared our set of supported coding transcripts to the set of WormBase transcripts to identify any novel transcripts. Novel transcripts are defined as those that contain a series of introns not represented in any WormBase transcript. In total our supported transcript set contains 50,449 novel protein-coding transcripts (70% of supported coding transcripts) - almost three times as many transcripts as there are annotated in WormBase.

Our supported transcript set represents thousands of novel alternative splicing events. We categorized both the type and number of alternative splicing events in both the WormBase gene models and our supported transcripts. ASTALAVISTA (Foissac and Sammeth, 2007) was used to identify exon skipping events, alternative splice donors, alternative splice acceptors, and intron retention events. Python was used to count the number of unique alternative transcript start and poly-A sites (Figure 32).

Type of alternative splicing	WormBase	RNA-Seq
 Exon skipping ¹	313	10,768
 Alternative splice donor ¹	208	15,749
 Alternative splice acceptor ¹	350	19,028
 Intron retention ¹	132	19,666
 Alternative start ²	5,387	11,027
 Alternative poly-A ²	5,374	11,006

¹Identified using ASTALAVISTA (Foissac et al., 2007)

²Automated counting of unique transcript start and end positions

Figure 32. Number and types of alternative splicing in WormBase gene models and our supported transcript set (RNA-Seq).

We investigated how the supported transcripts are distributed across different genes and counted the number of cases where multiple supported transcripts overlap a gene not previously known to encode multiple transcripts. We compared the number of transcripts in WormBase from each protein-coding gene to the number of our supported transcripts that overlap that gene. We found that 77% of protein-coding genes in WormBase had multiple transcripts in our database; 73% had multiple transcripts with multiple distinct coding sequences, 70% of these transcripts are supported by locally non-rare introns and exons. We identified multiple transcripts for over 44% of protein-coding genes that are not currently known to undergo alternative splicing (Figure 33).

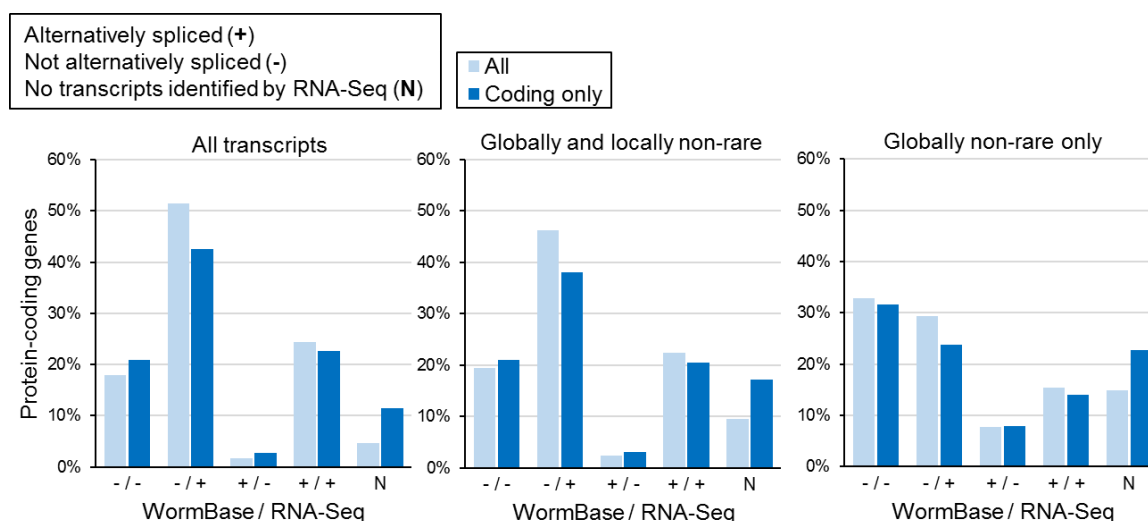


Figure 33. Number of genes with one (-) or multiple (+) transcripts in WormBase compared to the number of transcripts identified from RNA-Seq. Only transcripts that are fully supported by our intron and exon databases are counted. "Coding only" refers to supported transcripts with a predicted coding sequence.

Not all our supported transcripts mapped to a known gene. We did not use any genome annotation to guide transcript assembly, not just to avoid any possible bias against the identification of novel transcripts, but also to explore the possibility of identifying potentially novel protein-coding genes. We identified 466 candidate coding transcripts that are fully supported by our intron and exon databases, that do not overlap any WormBase gene (coding or non-coding). While we cannot rule out the possibility of artifacts introduced by repetitive or otherwise complex intergenic regions, we identified 50 transcripts at 27 loci that have a multi-intron structure with robust read support (all introns have >1000 supporting reads; example Figure 34).

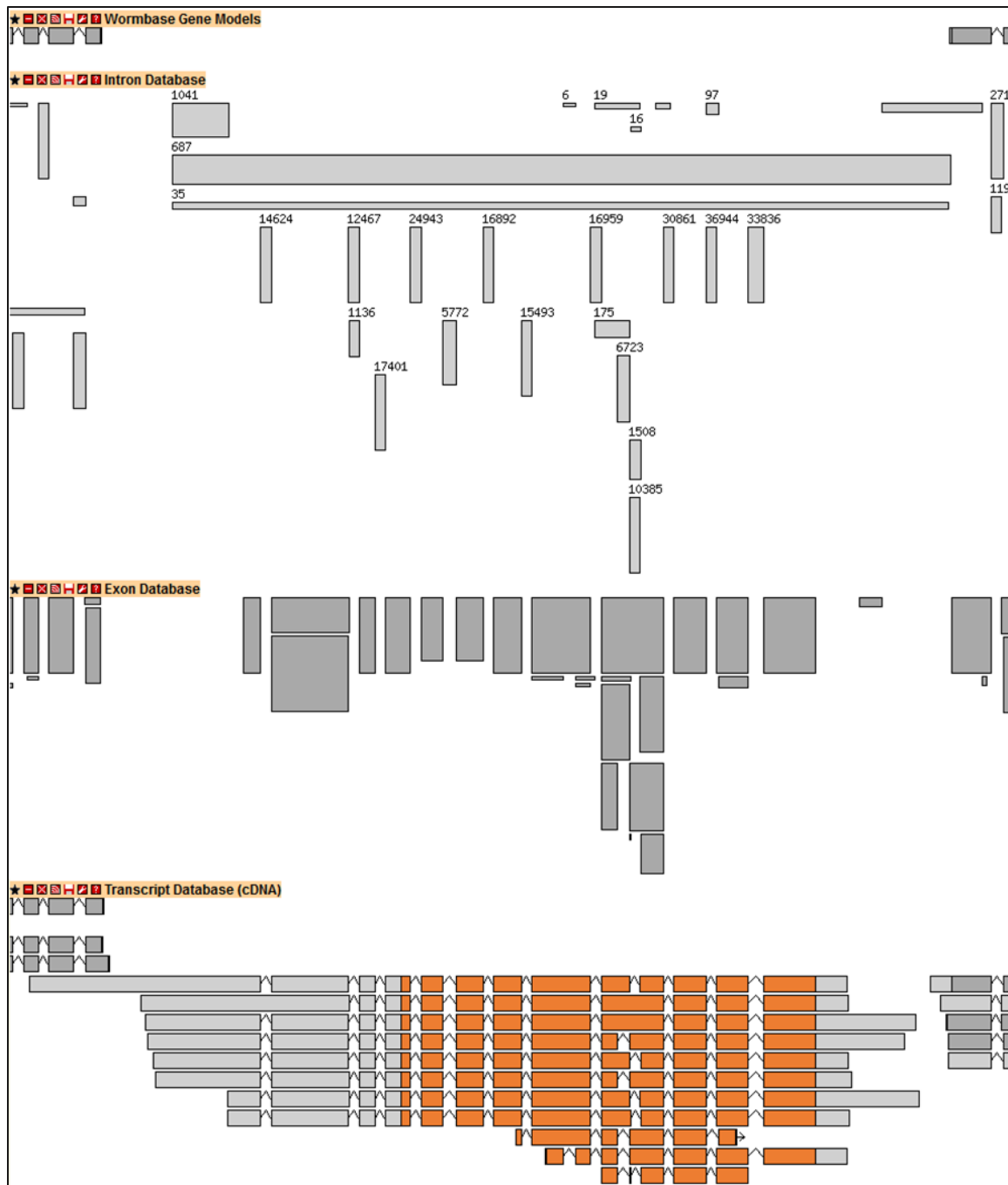


Figure 34. A potentially novel protein-coding gene with multiple transcripts identified between genomic coordinates 13,292,958 and 13,296,957 on chromosome III (- strand). Shown (from top to bottom) are the nearby WormBase gene models, our intron database, our exon database, and supported transcripts with candidate coding regions highlighted in orange.

4.4. Discussion

Alternative splicing is essential for *C. elegans* development, playing a key role in many processes including apoptosis (Shaham and Horvitz, 1996), cell differentiation and

proliferation (Mantina *et al.*, 2009). However, based on the WormBase gene models, it appears to be underutilized as a mechanism for expanding coding capacity in *C. elegans* compared to “more complex” organisms. In the current gene models, only 25% of protein-coding genes have multiple transcripts; compared to humans where an estimated ~95% of multi-exon genes encode multiple transcripts (Pan *et al.*, 2008). Despite humans and *C. elegans* having approximately the same number of protein-coding genes, the coding capacity of *C. elegans* appears to be much lower. However, in our genome-wide investigation of alternative splicing we identified a surprising number of introns and exons that could not be attributed to any WormBase transcript, indicating that a significant portion the coding capacity of *C. elegans* may be missing from the current gene models. Because the set of transcripts in WormBase may be incomplete, obtaining a quantitative measure of the full coding capacity of *C. elegans* necessitated the construction of a high-quality set of *C. elegans* transcripts from RNA-Seq.

At the time of writing, several publicly available transcript assembly programs exist that are able to detect alternative splicing transcripts. These generally fall into two categories: reference-based and *de novo* assembly. Cufflinks and Stringtie, used here, use the reference-based strategy where reads are aligned to the genome prior to being assembled together into transcripts. This strategy is generally less computationally intensive than *de novo* assembly (clusters of aligned reads can be assembled independently, rather than needing to sort through the entire pool of reads first like *de novo*). It also has the advantage of a reference genome to guide assembly. Reference-based assembly is generally preferred where a high-quality reference is available. Programs using this approach can assemble alternative transcripts, even at low sequencing depth (Martin and Wang, 2011). Sequencing artifacts of contaminating reads do not generally affect assembly as they would not align to the reference genome – though this is dependent on the quality of the reference genome. Fortunately, over twenty-years of sequencing and assembly has produced an extremely high-quality *C. elegans* genome, but genome quality is something to consider for non-model organisms with varying degrees of genome quality. The reference genome is also used to fill in small gaps in assemblies caused by lack of read coverage. However, as a side-effect of this last point reference-based assemblers often generate long UTRs (Figure 31 as an example) where read coverage is low. A long UTR may also overlap an adjacent gene model if the two genes are in close proximity in the genomic sequence. Reference-

based assemblers also struggle with *trans*-spliced transcripts. Reads originating from the 5' end of *trans*-spliced transcripts in *C. elegans* contain a SL sequence. The 22 nucleotide SL sequence does not match the genomic sequence of the 5' end of the transcripts, meaning that the reads can not be contiguously aligned back to the genome (though certain steps, such as soft-clipping, can allow the non-SL portion of the read to align). *Trans*-splicing affects over 70% of *C. elegans* protein-coding genes. This makes precise identification of the 5' ends of *C. elegans* transcripts a challenging task. Long introns can also be a challenge for reference-based assemblers, though we limited intron identification to those under a conservative threshold of 5000 bp which somewhat avoids this problem. *De novo* assembly largely avoids these issues. *De novo* assembly programs do not rely on an aligned set of reads, or a reference genome, which avoids errors introduced from misaligned or ambiguously aligned reads propagating into the assembled transcripts. As an overview, *de novo* assembly programs look for overlaps between reads and assemble them into transcripts and various algorithms have been developed to accomplish this. *De novo* transcript assembly does not rely on intron detection, so transcripts with novel introns or long introns do not pose a problem for assembly. Similarly, *trans*-spliced transcripts with an SL sequence at the 5' end do not pose an issue as the sequence is not compared against a reference genome prior to assembly. One major drawback of *de novo* assembly is that low-quality reads and contaminants are readily assembled into transcripts. However, we took steps to remove the former (see Section 2.2.2) and for the latter we only accepted transcripts that could be aligned back to the *C. elegans* genome, which should exclude any exogenous transcripts. A major drawback we were not able to compensate for is that *de novo* assembly programs often struggle with transcripts where sequence coverage is low (Martin *et al.*, 2010). This is a likely explanation for why *Trans*-ABYSS generated so many single-exon transcripts (50-60% of transcripts per library, on average). These single-exon transcripts generally overlapped longer, multi-exon transcripts, meaning that they may be incomplete fragments of a longer transcripts, mis-identified as distinct isoforms. We excluded these single-exon transcripts from our analysis.

While reference-based transcript assembly is often preferred over *de novo* when a high-quality reference is available, to our knowledge there has been no systematic evaluation that showed one approach is universally better than the other. Therefore, we used both approaches to overcome the weakness of either.

One challenge we could not avoid is in the inherent difficulty of assembling long transcripts from short RNA-Seq reads. As part of the assembly workflow, we only accepted transcripts where all introns and exons were fully supported by our intron and exon databases, respectively. However, this only gives us confidence that the individual components of each transcript are correct. The exact structure of each transcript – *i.e.* that all the introns and exons are pieced together correctly – requires further validation. Iso-Seq reads provide a valuable tool for validating transcripts; the long read-length achievable with this technology allows for end-to-end sequencing of even the longest transcripts. However, sequencing depth of our Iso-Seq dataset is limited and not all transcripts are covered. Nevertheless, we were able to validate many of our assembled transcripts with both Iso-Seq reads and WormBase transcripts showing that the transcript assembly process, and subsequent selection of fully-supported transcripts, produces a largely accurate set of transcripts.

We identified 50,449 novel protein-coding transcripts indicating that alternative splicing is far more ubiquitous in *C. elegans* than what is represented in the current transcript models. Our supported transcript set contains tens-of-thousands of each type of alternative splicing event – at least 30 times more than the number of events represented in the WormBase gene models. The exception is for alternative start and poly-A sites. Alternative start and poly-A sites are by far the most common type of alternative splicing event in the WormBase gene models (~5,400 of each). While our supported transcript set contains just over 11,000 each of alternative start and poly-A sites – almost two-times more than the WormBase gene models – we expect these numbers are an underestimate. To construct our transcript set, we used GffCompare to merge transfrags with the same series of introns. The advantage of this approach is that we preferentially select full-length transcripts. However, the drawback is that transcripts with alternative start/poly-A sites may be discarded if they contain the same series, or subset, of introns as a longer transcript.

Where 28% of WormBase genes currently have multiple isoforms, we identified multiple distinct protein-coding transcripts for 68% of WormBase coding genes. The WormBase gene models contain 27,876 distinct protein-coding sequences. In total, we identified 72,274 distinct protein-coding transcripts, most of which are supported by locally non-rare introns and exons, showing that the *C. elegans* genome can encode many more proteins than what is represented in the WormBase gene models.

Chapter 5. Conclusion

The coding capacity of a eukaryote – the set of distinct protein-coding transcripts encoded in its genome – is expanded through the mechanism of alternative splicing. Curiously, alternative splicing appears to be used more extensively in eukaryotes that are perceived to be “more complex.” Comparing the current human genome annotation to the annotation of the relatively simple nematode *C. elegans* reveals they have broadly the same number of protein-coding genes (~20,000 each). Yet, 90-95% of human multi-exon protein-coding genes are estimated to undergo alternative splicing compared to the 25% of *C. elegans* protein-coding genes. The implication is that humans have a much greater coding capacity than *C. elegans*, suggesting a correlation between coding capacity and organismal complexity. However, to fully explore this possibility we must first know the full coding capacity of the organisms being compared.

Extensive experimental and computational efforts over the last twenty years have gone into annotating the genome of *C. elegans*. More recently, RNA-Seq has been used to modify the *C. elegans* transcript models. The unparalleled depth of coverage offered by RNA-Seq has allowed investigators to reliably detect the rarest transcripts and identify thousands of novel introns. However, there exists uncertainty about whether these novel introns, particularly the rare ones, contribute to the coding capacity of *C. elegans* or represent non-coding (or spurious) transcripts. Therefore, additional methods are required to get an accurate measure of coding capacity.

For this thesis project, we used *C. elegans* as a model to develop methods for evaluating completeness of coding capacity at the genome-scale. This project relied on a curated set of 802 publicly available RNA-Seq libraries to achieve ultra-deep coverage of the *C. elegans* transcriptome. Using these libraries and an empirically derived set of filtering criteria, we constructed a high-quality intron database and a high-quality exon database which served as two metrics for evaluating coding capacity completeness. We found that over 93% (104,384) of introns and 85% (111,376) of coding exons in the WormBase gene models are represented in our database, indicating that our approach is accurate. In addition, we identified 134,949 introns and 204,812 exons that are completely novel (56% and 62% of our databases, respectively). Most novel introns and exons in our databases were globally rare, giving the initial impression that these may be

spurious rather than the result of a functional splicing event. However, our investigation showed that almost all introns and exons (80% and 82%, respectively) showed non-rare levels of expression locally, implying a functional role under specific conditions.

Based on the novel introns and exons we identified, we predicted that the coding capacity of *C. elegans* is far from complete. To quantify the full coding capacity of *C. elegans*, we developed a protocol for constructing a high-quality database of coding transcripts. We applied multiple publicly available transcript assembly programs on the same 802 RNA-Seq libraries used to construct the intron and exon databases; discarding any assembled transcripts that were not fully supported by both databases. The result was a high-quality set of 99,627 transcripts, which fully support 83% (26,167) of WormBase coding transcripts and partially support nearly all remaining coding transcripts. Our database contains 50,449 novel transcripts (83.7% of the database), many of which are novel transcripts of protein-coding genes that up now had only one annotated transcript. Currently, only 25% of WormBase coding genes have multiple transcripts annotated. In this study we identified multiple transcripts for 77% of coding genes, showing that alternative splicing is far more extensive in *C. elegans* than what is represented in the current gene models. In total, we identified 72,274 distinct protein-coding sequences within our supported transcripts. By comparison, there are 27,876 distinct protein-coding sequences annotated in WormBase, meaning as little as a third of the full coding capacity of *C. elegans* is represented in the current genome annotations.

Chapter 6. Future Directions

The methods we developed in this study for evaluating completeness of coding capacity can be applied to any organism with a reference genome and a sufficient amount of RNA-Seq data. The next logical step would be to apply these methods to evaluate the coding capacity other organisms, including humans. Our finding that the coding capacity of *C. elegans* is much greater than previously thought suggest that the same may hold true of many eukaryotes.

The set of novel supported transcripts identified in this thesis may be of use for improving gene prediction algorithms. Many of the current gene prediction algorithms rely on a Hidden Markov Model, or other statistical model, trained on a known set of transcripts. Our transcripts offer a more comprehensive set of exons, splice sites, and other signals that may help better train these prediction algorithms.

The major finding of this thesis is that tens-of-thousands of novel introns and exons – along with the transcripts they make up – are rare in the context of whole *C. elegans* across all developmental stages yet are non-rare under specific circumstances. Upregulation of these transcripts implies a functional role in specific developmental stages/tissue types/cells. However, the functional importance of these transcripts is still unknown. 50,449 novel transcripts were predicted to have a coding region, though not all of these may ultimately produce a functional protein. Whether there exists a regulatory mechanism that prevents their translation, or they have simply been mis-identified as protein-coding by the TransDecoder scoring algorithm, the protein-product of these transcripts remains to be validated. Homology to other species may shed light on the function of novel transcripts. *C. elegans* is commonly used as a model to gain biological insights about human biological processes, but in this case the reverse may be possible. About 41% of *C. elegans* protein-coding genes have functional orthologs in the human genome (Kim *et al.*, 2018). It is possible that novel alternative isoforms of these genes have been found in the human genome. Novel transcript without orthologs may require isoform-specific knockout experiments to determine their function.

Despite our careful approach to identifying introns, exons, and transcripts, our databases may still contain some false positives (or valid features may have been missed). The advantage of RNA-Seq is unparalleled depth of coverage, but short reads

mean that transcripts must be assembled computationally – a technically challenging task. While we ensured that the individual introns and exons in our assembled transcripts are supported by our databases, the exact structure of each transcript (*i.e.* the sequence of introns and exons that make up the transcript) requires further validation. We were able to support a subset of assembled transcripts using our pool of Iso-Seq reads. However, the coverage offered by these reads was limited so we were not able to validate the full set of transcripts – particularly rare transcripts. Ultimately, extensive sequencing of the *C. elegans* transcriptome using Iso-Seq or other long-read sequencing technologies may be necessary to precisely define all full-length transcripts.

References

- Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., and Reddy, A.S.N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications* 7, 11706.
- Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., *et al.* (2016). The Ensembl gene annotation system. *Database (Oxford)* 2016.
- Auch, D., and Reth, M. (1990). Exon trap cloning: using PCR to rapidly detect and clone exons from genomic DNA fragments. *Nucleic Acids Res* 18, 6743–6744.
- Baruzzo, G., Hayer, K.E., Kim, E.J., Camillo, B.D., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods* 14, 135–139.
- Boeck, M.E., Huynh, C., Gevirtzman, L., Thompson, O.A., Wang, G., Kasper, D.M., Reinke, V., Hillier, L.W., and Waterston, R.H. (2016). The time-resolved transcriptome of *C. elegans*. *Genome Res.* 26, 1441–1450.
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Brenner, S. (1974). The Genetics of *Caenorhabditis Elegans*. *Genetics* 77, 71–94.
- Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Edited by F. E. Cohen. *Journal of Molecular Biology* 268, 78–94.
- Chen, L., Bush, S.J., Tovar-Corona, J.M., Castillo-Morales, A., and Urrutia, A.O. (2014). Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol Biol Evol* 31, 1402–1413.

Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C.-K., *et al.* (2005). WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res* 33, D383–D389.

Chhangawala, S., Rudy, G., Mason, C.E., and Rosenfeld, J.A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol* 16.

Dillman, A.A., Hauser, D.N., Gibbs, J.R., Nalls, M.A., McCoy, M.K., Rudenko, I.N., Galter, D., and Cookson, M.R. (2013). mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nature Neuroscience* 16, 499–506.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Consortium, T.R., Alioto, T., Behr, J., Bertone, P., Bohnert, R., *et al.* (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* 10, 1185–1191.

Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet* 23, 5866–5878.

Flicek, P. (2007). Gene prediction: compare and CONTRAST. *Genome Biology* 8, 233.

Frenkel-Morgenstern, M., Gorohovski, A., Lacroix, V., Rogers, M., Ibanez, K., Boullosa, C., Andres Leon, E., Ben-Hur, A., and Valencia, A. (2013). ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res* 41, D142–D151.

Gabut, M., Samavarchi-Tehrani, P., Wang, X., Slobodeniuc, V., O’Hanlon, D., Sung, H.-K., Alvarez, M., Talukder, S., Pan, Q., Mazzoni, E.O., *et al.* (2011). An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell* 147, 132–146.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., *et al.* (2010). Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project. *Science* 330, 1775–1787.

Gilmour, L.M.R., Macleod, K.G., McCaig, A., Gullick, W.J., Smyth, J.F., and Langdon, S.P. (2001). Expression of erbB-4/HER-4 Growth Factor Receptor Isoforms in Ovarian Cancer. *Cancer Res* 61, 2169–2176.

Glover-Cutter, K.M., Lin, S., and Blackwell, T.K. (2013). Integration of the Unfolded Protein and Oxidative Stress Responses through SKN-1/Nrf. *PLOS Genetics* 9, e1003701.

- Gramates, L.S., Marygold, S.J., Santos, G. dos, Urbano, J.-M., Antonazzo, G., Matthews, B.B., Rey, A.J., Tabone, C.J., Crosby, M.A., Emmert, D.B., *et al.* (2017). FlyBase at 25: looking to the future. *Nucleic Acids Res* 45, D663–D671.
- Guigó, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., *et al.* (2006). EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biology* 7, S2.
- Hannon, G.J., Maroney, P.A., and Nilsen, T.W. (1991). U small nuclear ribonucleoprotein requirements for nematode cis- and trans-splicing in vitro. *J. Biol. Chem.* 266, 22792–22795.
- Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 38, e131.
- Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., *et al.* (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774.
- Hashimoto, T., Hoon, D., J.I, M., Grimmond, S.M., Daub, C.O., Hayashizaki, Y., and Faulkner, G.J. (2009). Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics* 25, 2613–2614.
- Herai, R.H., and Yamagishi, M.E.B. (2010). Detection of human interchromosomal trans-splicing in sequence databanks. *Brief Bioinform* 11, 198–209.
- Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res.* 19, 657–666.
- Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44–57.
- Hwang, B.J., Müller, H.-M., and Sternberg, P.W. (2004). Genome annotation by high-throughput 5' RNA end determination. *Proc Natl Acad Sci U S A* 101, 1650–1655.
- Kahles, A., Behr, J., and Räscher, G. (2016). MMR: a tool for read multi-mapper resolution. *Bioinformatics* 32, 770–772.
- Kalsotra, A., Xiao, X., Ward, A.J., Castle, J.C., Johnson, J.M., Burge, C.B., and Cooper, T.A. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *PNAS* 105, 20333–20338.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14, R36.
- Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 12, 357–360.

- Kim, W., Underwood, R.S., Greenwald, I., and Shaye, D.D. (2018). OrthoList 2: A New Comparative Genomic Analysis of Human and *Caenorhabditis elegans* Genes. *Genetics* 210, 445–461.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Lee, R.Y.N., Howe, K.L., Harris, T.W., Arnaboldi, V., Cain, S., Chan, J., Chen, W.J., Davis, P., Gao, S., Grove, C., *et al.* (2018). WormBase 2017: molting into a new stage. *Nucleic Acids Res* 46, D869–D874.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Li, S., Tighe, S.W., Nicolet, C.M., Grove, D., Levy, S., Farmerie, W., Viale, A., Wright, C., Schweitzer, P.A., Gao, Y., *et al.* (2014). Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF Next-Generation Sequencing Study. *Nat Biotechnol* 32, 915–925.
- Mantina, P., MacDonald, L., Kulaga, A., Zhao, L., and Hansen, D. (2009). A mutation in *teg-4*, which encodes a protein homologous to the SAP130 pre-mRNA splicing factor, disrupts the balance between proliferation and differentiation in the *C. elegans* germ line. *Mechanisms of Development* 126, 417–429.
- Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics* 12, 671–682.
- Martin, J., Bruno, V.M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* 11, 663.
- Mullen, G.P., Rogalski, T.M., Bush, J.A., Gorji, P.R., Moerman, D.G., and Kimble, J. (1999). Complex Patterns of Alternative Splicing Mediate the Spatial and Temporal Distribution of Perlecan/UNC-52 in *Caenorhabditis elegans*. *MBoC* 10, 3205–3221.
- Nellore, A., Jaffe, A.E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., Phillips III, R.A., Karbhari, N., Hansen, K.D., Langmead, B., *et al.* (2016). Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology* 17, 266.
- Neves, G., Zucker, J., Daly, M., and Chess, A. (2004). Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nature Genetics* 36, 240–246.
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40, 1413–1415.

- Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33, 290–295.
- Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., *et al.* (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 42, D756–D763.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.
- Ramani, A.K., Calarco, J.A., Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M., *et al.* (2011). Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*. *Genome Res.* 21, 342–348.
- Reboul, J., Vaglio, P., Rual, J.-F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., *et al.* (2003). *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* 34, 35–41.
- Revil, T., Gaffney, D., Dias, C., Majewski, J., and Jerome-Majewska, L.A. (2010). Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* 11, 399.
- Riddle, D.L., Blumenthal, T., Meyer, B.J., and Priess, J.R. (1997a). Frequency of Operons and Trans -Splicing (Cold Spring Harbor Laboratory Press).
- Riddle, D.L., Blumenthal, T., Meyer, B.J., and Priess, J.R. (1997b). Translation Initiation and Termination Signals (Cold Spring Harbor Laboratory Press).
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., *et al.* (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods* 7, 909–912.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nature Biotechnology* 29, 24–26.
- Ruzanov, P., Jones, S.J., and Riddle, D.L. (2007). Discovery of novel alternatively spliced *C. elegans* transcripts by computational analysis of SAGE data. *BMC Genomics* 8, 447.
- Ruzo, A., Ismailoglu, I., Popowski, M., Haremaki, T., Croft, G.F., Deglincerti, A., and Brivanlou, A.H. (2015). Discovery of novel isoforms of huntingtin reveals a new hominid-specific exon. *PLoS ONE* 10, e0127687.
- Salehi-Ashtiani, K., Lin, C., Hao, T., Shen, Y., Szeto, D., Yang, X., Ghamsari, L., Lee, H., Fan, C., Murray, R.R., *et al.* (2009). Large-scale RACE approach for proactive experimental definition of *C. elegans* ORFeome. *Genome Res.* 19, 2334–2342.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). *Drosophila* Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *Cell* 101, 671–684.

Shaham, S., and Horvitz, H.R. (1996). An Alternatively Spliced *C. elegans* ced-4 RNA Encodes a Novel Cell Death Inhibitor. *Cell* 86, 201–208.

Shepard, S., McCreary, M., and Fedorov, A. (2009). The Peculiarities of Large Intron Splicing in Animals. *PLOS ONE* 4, e7853.

Shin, H., Hirst, M., Bainbridge, M.N., Magrini, V., Mardis, E., Moerman, D.G., Marra, M.A., Baillie, D.L., and Jones, S.J. (2008). Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biology* 6, 30.

Soergel, D.A.W., Lareau, L.F., and Brenner, S.E. (2013). Regulation of Gene Expression by Coupling of Alternative Splicing and NMD (Landes Bioscience).

Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1993). Operons in *C. elegans*: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* 73, 521–532.

Spieth, J., Lawson, D., Davis, P., Williams, G., and Howe, K. (2005). Overview of gene structure in *C. elegans* (WormBook).

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001). WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res* 29, 82–86.

The *C. elegans* Sequencing Consortium (1998). Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science* 282, 2012–2018.

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M., and Kitts, P. (2013). Eukaryotic Genome Annotation Pipeline (National Center for Biotechnology Information (US)).

Tourasse, N.J., Millet, J.R.M., and Dupuy, D. (2017). Quantitative RNA-seq meta-analysis of alternative exon usage in *C. elegans*. *Genome Res.* 27, 2120–2128.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Ólason, P. Ísólfur, Albrecht, M., Hegyi, H., Giorgetti, A., *et al.* (2007). The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci U S A* 104, 5495–5500.

Uchida, O., Nakano, H., Koga, M., and Ohshima, Y. (2003). The *C. elegans* che-1 gene encodes a zinc finger transcription factor required for specification of the ASE chemosensory neurons. *Development* 130, 1215–1224.

Veikkolainen, V., Vaparanta, K., Halkilahti, K., Iljin, K., Sundvall, M., and Elenius, K. (2011). Function of ERBB4 is determined by alternative splicing. *Cell Cycle* 10, 2647–2657.

Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.

Wang, S.M., Fears, S.C., Zhang, L., Chen, J.-J., and Rowley, J.D. (2000). Screening poly(dA/dT)- cDNAs for gene identification. *PNAS* 97, 4162–4167.

Wang, Y., Liu, J., Huang, B., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., *et al.* (2015). Mechanism of alternative splicing and its regulation. *Biomed Rep* 3, 152–158.

Wang, Z., Chen, Y., and Li, Y. (2004). A Brief Review of Computational Gene Prediction Methods. *Genomics Proteomics Bioinformatics* 2, 216–221.

Williams, A.G., Thomas, S., Wyman, S.K., and Holloway, A.K. (2016a). RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. *Current Protocols in Human Genetics* 83, 11.13.1-11.13.20.

Williams, C.R., Baccarella, A., Parrish, J.Z., and Kim, C.C. (2016b). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* 17, 103.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875.

Yeku, O., and Frohman, M.A. (2011). Rapid Amplification of cDNA Ends (RACE). In *RNA*, (Humana Press), pp. 107–122.

Zhang, Z., Huang, S., Wang, J., Zhang, X., Pardo Manuel de Villena, F., McMillan, L., and Wang, W. (2013). GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference owing to RNA-seq reads misalignment. *Bioinformatics* 29, i291–i299.

Appendix A. Supplemental Materials and Methods

6.1. *C. elegans* reference annotation

The *C. elegans* genome and genome annotations used in this thesis were from WormBase release WS250 (data freeze 31-Jul-2015). All sections that refer to “reference,” “annotated,” or “WormBase” transcript models or transcript features refer to this particular release. This release was used to maintain consistency with other analyses performed by our lab that also use release WS250. Release notes can be found at: https://wormbase.org/about/wormbase_release_WS250

6.2. Iso-Seq

A PacBio RSII “Iso-Seq” sequencer was used to sequence a mixed population of *C. elegans*. A total of 603,652 Iso-Seq long reads were obtained (median read length is 1033 bp, minimum is 51 bp, maximum is 36387 bp). Iso-Seq reads were mapped to the *C. elegans* genome using GMAP (Wu and Watanabe, 2005).

A set of putative transcripts was generated from the Iso-Seq data by Jiarui Li. This involved grouping reads that contained the same series of introns. Reads in each group were collapsed into a single read representing one putative transcript. Reads that contained a series of introns that were a subset of introns of another read were treated as a separate group.

6.3. Programs used

Table 3. List of external programs used in this thesis.

Program	Version/Release	Reference
BBDuk	Last modified June 1, 2017 (BBDuk 37.36)	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/
Cufflinks	2.2.1 (uses SAMtools version 1.20)	Trapnell <i>et al.</i> , 2012
fastq-dump	2.8.2	https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/
GBrowse	2.54	http://gmod.org/wiki/GBrowse
GMAP	version 2018-07-04	Wu and Watanabe, 2005
STAR	2.6.0c	Dobin <i>et al.</i> , 2013
SAMtools	1.6 (uses htlib 1.6)	Li <i>et al.</i> , 2009
Stringtie	1.3.4d	Pertea <i>et al.</i> , 2015
Trans-ABYSS	1.5.5	Robertson <i>et al.</i> , 2010
Trimmomatic	0.36	Bolger <i>et al.</i> , 2014

SAMtools 1.6 (Li *et al.*, 2009) was used to parse SAM/BAM formatted alignments. Generic Genome Browser (“GBrowse”) 2.54 was used to display transcripts and sequence features in GFF format. The relative height of displayed introns and exons represents their relative levels of read support.

6.4. ExonTrap availability

ExonTrap is implemented in Python 3 and uses the modules “BioPython” (<https://biopython.org/>) and “pysam” (<https://pysam.readthedocs.io/en/latest/api.html>). ExonTrap can be obtained from: https://github.com/mattdoug604/exon_trap

Appendix B. Supplemental Tables

Supplemental Table 1. RNA-Seq libraries selected from SRA.

Library ID	Run ID	Read Pairs	Avg. Length	Notes
ERX1545801	ERR1474664	9881439	200	
ERX1545802	ERR1474665	11853884	200	
ERX1545803	ERR1474666	12486224	170	
ERX1545804	ERR1474667	16144455	171	
ERX1545805	ERR1474668	10878316	200	
ERX1545806	ERR1474669	15624009	200	
ERX1545807	ERR1474670	15048370	169	
ERX1545808	ERR1474671	17461197	180	
ERX1545809	ERR1474672	10560564	200	
ERX1545810	ERR1474673	8274123	200	
ERX1545811	ERR1474674	14609930	172	
ERX1545812	ERR1474675	12530368	169	
ERX1545813	ERR1474676	6487959	200	
ERX1545814	ERR1474677	8489882	200	
ERX1545815	ERR1474678	13387369	172	
ERX1545816	ERR1474679	14140905	164	
ERX1545817	ERR1474680	30081203	202	
ERX1545818	ERR1474681	21864318	202	
ERX1545819	ERR1474682	30695500	202	
ERX1545820	ERR1474683	29143207	202	
ERX1545821	ERR1474684	39019471	202	
ERX1545822	ERR1474685	23269462	202	
ERX1545823	ERR1474686	14303135	156	
ERX1545824	ERR1474687	11787423	169	
ERX1545825	ERR1474688	20530746	169	
ERX1545826	ERR1474689	13833868	168	
ERX1545827	ERR1474690	13901720	200	
ERX1545828	ERR1474691	10315699	200	
ERX1545829	ERR1474692	16826291	175	
ERX1545830	ERR1474693	12026560	174	
ERX1545831	ERR1474694	11002475	200	
ERX1545832	ERR1474695	8745584	200	
ERX1545833	ERR1474696	13402132	174	
ERX1545834	ERR1474697	15759579	183	
ERX1545835	ERR1474698	9208505	200	
ERX1545836	ERR1474699	9369717	200	
ERX1545837	ERR1474700	16851571	170	
ERX1545838	ERR1474701	9842274	176	
ERX1545839	ERR1474702	7988742	200	
ERX1545840	ERR1474703	9919656	200	
ERX1545841	ERR1474704	11217713	172	
ERX1545842	ERR1474705	11484086	174	
ERX278737	ERR305394	24145595	150	
ERX278739	ERR305400	23246239	150	
ERX278740	ERR305390	25953543	150	
SRX026728	SRR065719	1.11E+10	152	
SRX026729	SRR065717	4.43E+09	152	
SRX085111	SRR317083	13037975	200	Mislabeled as EST in SRA
SRX085112	SRR316196	2510020	152	Mislabeled as EST in SRA

SRX085217	SRR316753	3384227	152	Mislabeled as EST in SRA
SRX085218	SRR317082	11015944	200	Mislabeled as EST in SRA
SRX085218	SRR350977	11284046	200	Mislabeled as EST in SRA
SRX085286	SRR316929	36457803	200	Mislabeled as EST in SRA
SRX085287	SRR316928	2434193	152	Mislabeled as EST in SRA
SRX092371	SRR332923	2602842	152	Mislabeled as EST in SRA
SRX092372	SRR332924	6284520	200	Mislabeled as EST in SRA
SRX092477	SRR332921	2355433	152	Mislabeled as EST in SRA
SRX092478	SRR478539	9422317	200	Mislabeled as EST in SRA
SRX092478	SRR332922	6407647	200	Mislabeled as EST in SRA
SRX092479	SRR332926	5079434	200	Mislabeled as EST in SRA
SRX092479	SRR332927	518118	200	Mislabeled as EST in SRA
SRX092480	SRR332925	3830232	152	Mislabeled as EST in SRA
SRX099973	SRR350988	1756213	200	Mislabeled as EST in SRA
SRX099973	SRR350987	8955548	200	Mislabeled as EST in SRA
SRX099975	SRR350991	1388523	200	Mislabeled as EST in SRA
SRX099975	SRR350990	5180559	200	Mislabeled as EST in SRA
SRX099978	SRR350995	1234504	200	Mislabeled as EST in SRA
SRX099978	SRR350996	10606303	200	Mislabeled as EST in SRA
SRX099978	SRR350994	6238336	200	Mislabeled as EST in SRA
SRX099979	SRR350999	9316827	200	Mislabeled as EST in SRA
SRX099979	SRR350998	945964	200	Mislabeled as EST in SRA
SRX099979	SRR350997	6810106	200	Mislabeled as EST in SRA
SRX099980	SRR351000	944004	200	Mislabeled as EST in SRA
SRX099980	SRR351001	11450241	200	Mislabeled as EST in SRA
SRX099981	SRR351003	1039662	200	Mislabeled as EST in SRA
SRX099981	SRR351002	10637975	200	Mislabeled as EST in SRA
SRX099982	SRR351005	1320403	200	Mislabeled as EST in SRA
SRX099982	SRR478540	2815108	200	Mislabeled as EST in SRA
SRX099982	SRR351004	11998870	200	Mislabeled as EST in SRA
SRX099984	SRR351008	1461561	200	Mislabeled as EST in SRA
SRX099984	SRR351007	12200911	200	Mislabeled as EST in SRA
SRX099985	SRR351009	26344268	200	Mislabeled as EST in SRA
SRX099986	SRR351010	2747776	200	Mislabeled as EST in SRA
SRX099986	SRR351011	11513195	200	Mislabeled as EST in SRA
SRX099988	SRR351014	1718342	200	Mislabeled as EST in SRA
SRX099988	SRR351013	4478988	200	Mislabeled as EST in SRA
SRX099991	SRR351017	3636534	200	Mislabeled as EST in SRA
SRX099991	SRR351018	1585913	200	Mislabeled as EST in SRA
SRX099994	SRR351022	1561066	200	Mislabeled as EST in SRA
SRX099994	SRR351021	12689344	200	Mislabeled as EST in SRA
SRX099995	SRR351024	11781115	200	Mislabeled as EST in SRA
SRX099995	SRR351023	1843764	200	Mislabeled as EST in SRA
SRX099996	SRR351025	1175757	200	Mislabeled as EST in SRA
SRX099996	SRR351026	645714	200	Mislabeled as EST in SRA
SRX099996	SRR351027	10916635	200	Mislabeled as EST in SRA
SRX099999	SRR351030	1562554	200	Mislabeled as EST in SRA
SRX099999	SRR351031	11378571	200	Mislabeled as EST in SRA
SRX100002	SRR351034	18101136	200	Mislabeled as EST in SRA
SRX100002	SRR351035	1608435	200	Mislabeled as EST in SRA
SRX100003	SRR351036	2430232	200	Mislabeled as EST in SRA
SRX100003	SRR351037	1914188	200	Mislabeled as EST in SRA
SRX100006	SRR351041	1719081	200	Mislabeled as EST in SRA
SRX100006	SRR351040	14599081	200	Mislabeled as EST in SRA
SRX100631	SRR351936	11437708	200	Mislabeled as EST in SRA
SRX100631	SRR351935	1218523	200	Mislabeled as EST in SRA

SRX100633	SRR351938	1241771	200	Mislabeled as EST in SRA
SRX100633	SRR351939	12151987	200	Mislabeled as EST in SRA
SRX100819	SRR352279	1223765	200	Mislabeled as EST in SRA
SRX100819	SRR352280	11792221	200	Mislabeled as EST in SRA
SRX1020630	SRR2012769	7320704	200	
SRX1020630	SRR2012770	875530	200	
SRX1020630	SRR2012771	11352979	200	
SRX1020632	SRR2012777	32082886	200	
SRX1020632	SRR2012775	13906283	200	
SRX1020632	SRR2012776	1688594	200	
SRX1020634	SRR2012779	7633553	200	
SRX1020634	SRR2012781	14372593	200	
SRX1020634	SRR2012780	915894	200	
SRX1020636	SRR2012785	12567755	200	
SRX1020636	SRR2012783	7497090	200	
SRX1020636	SRR2012784	912122	200	
SRX1020638	SRR2012789	10266101	200	
SRX1020638	SRR2012787	10129253	200	
SRX1020638	SRR2012788	1228038	200	
SRX1020640	SRR2012791	7694975	200	
SRX1020640	SRR2012792	934680	200	
SRX1020640	SRR2012793	16409506	200	
SRX1022566	SRR2015249	3237432	200	
SRX1022566	SRR2015247	10355172	200	
SRX1022566	SRR2015248	1262426	200	
SRX1022568	SRR2015253	9331437	200	
SRX1022568	SRR2015252	617297	200	
SRX1022568	SRR2015251	5248666	200	
SRX1022570	SRR2015256	874465	200	
SRX1022570	SRR2015255	7252206	200	
SRX1022570	SRR2015257	11721613	200	
SRX1022572	SRR2015262	14428202	200	
SRX1022572	SRR2015260	9181419	200	
SRX1022572	SRR2015261	1170367	200	
SRX1022574	SRR2015265	822234	200	
SRX1022574	SRR2015266	11329322	200	
SRX1022574	SRR2015264	6617764	200	
SRX1022576	SRR2015270	13918016	200	
SRX1022576	SRR2015269	2183831	200	
SRX1022576	SRR2015268	17759621	200	
SRX1022578	SRR2015273	1777687	200	
SRX1022578	SRR2015272	14861344	200	
SRX1022578	SRR2015274	32888149	200	
SRX1022580	SRR2015276	5716606	200	
SRX1022580	SRR2015278	10103707	200	
SRX1022580	SRR2015277	687796	200	
SRX1022582	SRR2015281	1210863	200	
SRX1022582	SRR2015282	16991077	200	
SRX1022582	SRR2015280	9714828	200	
SRX1022584	SRR2015285	916654	200	
SRX1022584	SRR2015286	10482683	200	
SRX1022584	SRR2015284	7669279	200	
SRX1022586	SRR2015289	1704308	200	
SRX1022586	SRR2015290	22686262	200	
SRX1022586	SRR2015288	13881447	200	
SRX1022588	SRR2015293	2406027	200	

SRX1022588	SRR2015294	9650969	200	
SRX1022588	SRR2015292	19872904	200	
SRX1022592	SRR2015298	12873192	200	
SRX1022592	SRR2015299	1562534	200	
SRX1022592	SRR2015300	13797106	200	
SRX1022595	SRR2015304	876149	200	
SRX1022595	SRR2015305	9443151	200	
SRX1022595	SRR2015303	7080949	200	
SRX1022597	SRR2015308	930714	200	
SRX1022597	SRR2015307	7280594	200	
SRX1022597	SRR2015309	18582488	200	
SRX1022599	SRR2015311	11626153	200	
SRX1022599	SRR2015310	628279	200	
SRX1022600	SRR2015312	552683	200	
SRX1022600	SRR2015313	11166310	200	
SRX1022601	SRR2015314	440293	200	
SRX1022601	SRR2015315	9109188	200	
SRX1022602	SRR2015316	565816	200	
SRX1022602	SRR2015317	14953746	200	
SRX1022603	SRR2015318	489054	200	
SRX1022603	SRR2015319	10413482	200	
SRX1022604	SRR2015320	883219	200	
SRX1022604	SRR2015321	11571591	200	
SRX1022605	SRR2015322	686401	200	
SRX1022605	SRR2015323	9345351	200	
SRX1022607	SRR2015325	11653619	200	
SRX1022607	SRR2015324	988500	200	
SRX1022608	SRR2015326	777831	200	
SRX1022608	SRR2015327	15794109	200	
SRX1022609	SRR2015328	746121	200	
SRX1022609	SRR2015329	14113809	200	
SRX1022610	SRR2015330	981694	200	
SRX1022610	SRR2015331	8584678	200	
SRX1022611	SRR2015333	10807933	200	
SRX1022611	SRR2015332	1196349	200	
SRX1022645	SRR2015474	11899175	200	
SRX1022645	SRR2015473	1309440	200	
SRX1022646	SRR2015475	1291146	200	
SRX1022646	SRR2015476	7038914	200	
SRX1022647	SRR2015478	10648984	200	
SRX1022647	SRR2015477	1068515	200	
SRX1022648	SRR2015479	1023742	200	
SRX1022648	SRR2015480	17151212	200	
SRX1022649	SRR2015482	7947023	200	
SRX1022649	SRR2015481	1141427	200	
SRX1022650	SRR2015484	8300244	200	
SRX1022650	SRR2015483	1633122	200	
SRX1022651	SRR2015485	1554449	200	
SRX1022651	SRR2015486	16650473	200	
SRX1022652	SRR2015487	1396874	200	
SRX1022652	SRR2015488	10279486	200	
SRX1022653	SRR2015490	9915915	200	
SRX1022653	SRR2015489	815516	200	
SRX1022654	SRR2015492	11385995	200	
SRX1022654	SRR2015491	865367	200	
SRX103269	SRR358684	1770190	200	Mislabeled as EST in SRA

SRX103269	SRR358683	6771115	200	Mislabeled as EST in SRA
SRX103271	SRR358686	3242416	202	Mislabeled as EST in SRA
SRX103273	SRR358688	4530294	200	Mislabeled as EST in SRA
SRX103273	SRR358689	1036325	200	Mislabeled as EST in SRA
SRX103275	SRR358691	2589808	202	Mislabeled as EST in SRA
SRX103277	SRR358694	1512001	200	Mislabeled as EST in SRA
SRX103277	SRR358693	17135321	200	Mislabeled as EST in SRA
SRX103278	SRR358695	4581280	200	Mislabeled as EST in SRA
SRX103278	SRR358696	3484793	200	Mislabeled as EST in SRA
SRX103280	SRR358698	8997917	202	Mislabeled as EST in SRA
SRX103649	SRR359063	930982	202	Mislabeled as EST in SRA
SRX103650	SRR359065	2822764	202	Mislabeled as EST in SRA
SRX103651	SRR359066	12240044	202	Mislabeled as EST in SRA
SRX103652	SRR359067	6154044	202	Mislabeled as EST in SRA
SRX103653	SRR359069	185285	202	Mislabeled as EST in SRA
SRX103669	SRR359087	9817424	202	Mislabeled as EST in SRA
SRX103670	SRR359088	6420527	202	Mislabeled as EST in SRA
SRX103671	SRR359089	5135775	202	Mislabeled as EST in SRA
SRX103672	SRR359090	8756285	202	Mislabeled as EST in SRA
SRX103673	SRR359091	10724305	202	Mislabeled as EST in SRA
SRX103677	SRR359094	969656	202	Mislabeled as EST in SRA
SRX1037996	SRR2039609	1258773	200	
SRX1037996	SRR2039610	9257950	200	
SRX1037997	SRR2039613	1207560	200	
SRX1037997	SRR2039614	540435	200	
SRX1037997	SRR2039612	4034039	200	
SRX1037999	SRR2039616	11185266	200	
SRX1037999	SRR2039615	1045366	200	
SRX1038000	SRR2039619	3239605	200	
SRX1038000	SRR2039620	2795439	200	
SRX1038000	SRR2039618	5111065	200	
SRX103983	SRR360121	1742664	200	Mislabeled as EST in SRA
SRX103983	SRR360120	29239498	200	Mislabeled as EST in SRA
SRX103985	SRR360124	37767206	200	Mislabeled as EST in SRA
SRX103985	SRR360125	1571622	200	Mislabeled as EST in SRA
SRX103986	SRR473297	7272287	200	
SRX103986	SRR360127	1877179	200	
SRX103986	SRR473298	19942840	200	
SRX103986	SRR360126	2304676	200	
SRX103988	SRR360129	5093861	202	
SRX1041553	SRR2043247	5461809	200	
SRX1041553	SRR2043249	2267252	200	
SRX1041553	SRR2043248	5374000	200	
SRX1041556	SRR2043251	4471486	200	
SRX1041556	SRR2043252	35520	200	
SRX1041556	SRR2043253	10111	200	
SRX1041558	SRR2043256	529646	200	
SRX1041558	SRR2043257	420590	200	
SRX1041558	SRR2043255	7078675	200	
SRX1041561	SRR2043261	4543216	200	
SRX1041561	SRR2043259	5442168	200	
SRX1041561	SRR2043260	5803426	200	
SRX1041563	SRR2043263	4368566	200	
SRX1041563	SRR2043265	404467	200	
SRX1041563	SRR2043264	1410992	200	
SRX1041571	SRR2043273	4631033	200	

SRX1041571	SRR2043274	704955	200
SRX1041571	SRR2043275	116563	200
SRX1041638	SRR2043334	7657543	200
SRX1041638	SRR2043333	5051610	200
SRX1041641	SRR2043337	5517925	200
SRX1041642	SRR2043341	144625	200
SRX1041642	SRR2043340	450886	200
SRX1041642	SRR2043339	5491344	200
SRX1041644	SRR2043343	14658677	200
SRX1041644	SRR2043342	2286893	200
SRX1041646	SRR2043347	256805	200
SRX1041646	SRR2043348	163493	200
SRX1041646	SRR2043346	4508601	200
SRX1041666	SRR2043368	9203312	200
SRX1041666	SRR2043369	6541386	200
SRX1041748	SRR2043471	24912399	200
SRX1051870	SRR2054400	1774384	200
SRX1051870	SRR2054401	14716165	200
SRX1051871	SRR2054403	31166336	200
SRX1051872	SRR2054404	1202508	200
SRX1051872	SRR2054405	11507725	200
SRX1051873	SRR2054408	693175	200
SRX1051873	SRR2054410	435401	200
SRX1051873	SRR2054409	419242	200
SRX1051873	SRR2054407	5145661	200
SRX1051875	SRR2054411	1053269	200
SRX1051875	SRR2054412	5673188	200
SRX1051877	SRR2054416	174181	200
SRX1051877	SRR2054417	269805	200
SRX1051877	SRR2054415	5175130	200
SRX1051879	SRR2054419	3539985	200
SRX1051879	SRR2054421	152554	200
SRX1051879	SRR2054420	401705	200
SRX1051881	SRR2054422	1081040	200
SRX1051881	SRR2054423	10666724	200
SRX1051882	SRR2054424	2067458	200
SRX1051882	SRR2054425	16269118	200
SRX1051883	SRR2054429	72528	200
SRX1051883	SRR2054428	85884	200
SRX1051883	SRR2054427	4200105	200
SRX1051885	SRR2054430	4637186	200
SRX1051885	SRR2054431	3740153	200
SRX1051886	SRR2054432	3933901	200
SRX1051886	SRR2054433	4561769	200
SRX1051887	SRR2054437	195169	200
SRX1051887	SRR2054435	4716728	200
SRX1051887	SRR2054436	2675713	200
SRX1051887	SRR2054438	294396	200
SRX1051889	SRR2054442	21027	200
SRX1051889	SRR2054440	2697649	200
SRX1051889	SRR2054441	35126	200
SRX1051889	SRR2054443	8745	200
SRX1051891	SRR2054444	2576993	200
SRX1051891	SRR2054445	18056534	200
SRX1051892	SRR2054446	20420848	200
SRX1051893	SRR2054448	9772614	200

SRX1051893	SRR2054447	1065949	200
SRX1051894	SRR2054449	2792814	200
SRX1051894	SRR2054450	23152650	200
SRX1051895	SRR2054451	1049152	200
SRX1051895	SRR2054452	10212079	200
SRX1051896	SRR2054456	9215340	200
SRX1051896	SRR2054457	679107	200
SRX1051896	SRR2054455	2205069	200
SRX1051899	SRR2054460	4718077	200
SRX1051899	SRR2054461	2013974	200
SRX1051899	SRR2054462	751864	200
SRX1051901	SRR2054464	4799160	200
SRX1051901	SRR2054465	5744961	200
SRX1051901	SRR2054466	1472520	200
SRX1051903	SRR2054469	652646	200
SRX1051903	SRR2054468	3596334	200
SRX1051903	SRR2054470	168180	200
SRX1051905	SRR2054472	4478600	200
SRX1051905	SRR2054474	129085	200
SRX1051905	SRR2054473	497559	200
SRX1051907	SRR2054477	573630	200
SRX1051907	SRR2054478	558940	200
SRX1051907	SRR2054476	4232867	200
SRX1051910	SRR2054480	5974315	200
SRX1051912	SRR2054484	36538290	200
SRX1051912	SRR2054483	1619341	200
SRX1051912	SRR2054482	6164102	200
SRX1051914	SRR2054486	5023481	200
SRX1051914	SRR2054487	786050	200
SRX1051916	SRR2054489	5824933	200
SRX1051916	SRR2054490	896499	200
SRX1051918	SRR2054494	844154	200
SRX1051918	SRR2054493	847770	200
SRX1051918	SRR2054492	1474075	200
SRX105293	SRR363980	86776631	200
SRX105294	SRR363981	76081136	200
SRX1067755	SRR2072653	17800675	200
SRX1067756	SRR2072654	21029644	200
SRX1067757	SRR2072655	22529895	200
SRX1067758	SRR2072656	19830222	200
SRX1067759	SRR2072657	17900135	200
SRX1067760	SRR2072658	22618460	200
SRX1067761	SRR2072659	24244884	200
SRX1067762	SRR2072660	21599084	200
SRX1067763	SRR2072661	19693226	200
SRX1067764	SRR2072662	19393455	200
SRX1080515	SRR2086428	24343892	202
SRX1080516	SRR2086430	30312849	202
SRX1098674	SRR2104395	16147174	202
SRX1098675	SRR2104396	13948832	202
SRX1098676	SRR2104397	15306394	202
SRX1098677	SRR2104398	13913910	202
SRX1130124	SRR2142254	49743412	202
SRX1130126	SRR2142255	38836876	202
SRX1165475	SRR2185654	57999404	202
SRX1165476	SRR2185655	40607164	202

SRX1165477	SRR2185656	49090847	202	
SRX1165478	SRR2185657	43990735	202	
SRX1165479	SRR2185658	37593429	202	
SRX1165480	SRR2185659	40730514	202	
SRX1165481	SRR2185660	49349324	202	
SRX1165482	SRR2185661	47066389	202	
SRX1165483	SRR2185662	37660281	202	
SRX1165484	SRR2185663	38830218	202	
SRX1165485	SRR2185664	37177623	202	
SRX1165486	SRR2185665	44012793	202	
SRX1165487	SRR2185666	44930736	202	
SRX1165488	SRR2185667	45738402	202	
SRX1165489	SRR2185668	44785529	202	
SRX1165490	SRR2185669	55212586	202	
SRX1165491	SRR2185670	43496211	202	
SRX1165492	SRR2185671	48640150	202	
SRX1165493	SRR2185672	52210533	202	
SRX1165494	SRR2185673	47968711	202	
SRX1165495	SRR2185674	41452404	202	
SRX1225069	SRR2352993	28350034	152	
SRX1225070	SRR2352994	34383409	152	
SRX1225071	SRR2352995	35977669	152	
SRX1225072	SRR2352996	47402823	152	
SRX1225073	SRR2352997	45518505	152	
SRX1225074	SRR2352998	38022847	152	
SRX1225075	SRR2352999	46053382	152	
SRX1225076	SRR2353000	38788451	152	
SRX1225077	SRR2353001	36832518	152	
SRX1225078	SRR2353002	39699232	152	
SRX1225079	SRR2353003	35302422	152	
SRX1308270	SRR2566273	10313961	150	Used for aligner comparison
SRX139566	SRR473085	18285547	200	
SRX139567	SRR473086	8253048	200	
SRX139591	SRR473299	11446900	200	
SRX139592	SRR473300	9265571	200	
SRX139602	SRR474827	88503754	200	
SRX139603	SRR474828	10299069	200	
SRX1433703	SRR2969230	35763023	191	Used for aligner comparison
SRX1433714	SRR2969231	12439998	191	Used for aligner comparison
SRX1433717	SRR2969232	42549042	181	Used for aligner comparison
SRX1433860	SRR2969236	32346851	191	Used for aligner comparison
SRX1447175	SRR2954656	12482548	200	
SRX1447176	SRR2954657	12521419	200	
SRX1447177	SRR2954658	12506974	200	
SRX1447178	SRR2954659	12355228	200	
SRX145443	SRR493075	1.63E+08	154	
SRX145444	SRR493076	1.11E+08	154	
SRX145445	SRR493077	1.42E+08	152	
SRX145446	SRR493078	1.93E+08	154	
SRX145447	SRR493079	1.92E+08	154	
SRX145480	SRR493100	4314228	200	
SRX145480	SRR493099	5036711	200	
SRX145482	SRR493103	1376993	200	Mislabeled as EST in SRA
SRX145482	SRR493102	9596653	200	Mislabeled as EST in SRA
SRX145486	SRR493105	3447018	200	
SRX145486	SRR554453	46797752	200	

SRX145660	SRR493358	11320149	200
SRX145660	SRR493361	1515413	200
SRX145660	SRR493360	2398451	200
SRX145660	SRR493359	1264450	200
SRX145661	SRR493363	1145716	200
SRX145661	SRR493364	2025795	200
SRX145661	SRR493365	1190809	200
SRX145661	SRR493362	12649217	200
SRX1463040	SRR2973733	30936129	152
SRX1463041	SRR2973734	48391833	152
SRX1463042	SRR2973735	48699685	152
SRX151597	SRR504316	17402820	200
SRX151598	SRR504317	12345830	200
SRX151598	SRR504318	5404094	200
SRX151599	SRR504319	11363982	200
SRX151602	SRR504323	20962044	200
SRX151602	SRR504322	28010592	200
SRX151607	SRR504324	10784658	200
SRX151607	SRR504325	4661761	200
SRX151617	SRR504337	6743390	200
SRX151617	SRR504336	16103184	200
SRX151618	SRR504339	1729285	200
SRX151618	SRR504338	6327128	200
SRX1588742	SRR3173732	4000000	150
SRX1588742	SRR3173729	4000000	150
SRX1588742	SRR3173733	4000000	150
SRX1588742	SRR3173730	4000000	150
SRX1588742	SRR3173731	4000000	150
SRX1588742	SRR3173722	4000000	150
SRX1588742	SRR3173726	267076	150
SRX1588742	SRR3173725	4000000	150
SRX1588742	SRR3173723	4000000	150
SRX1588742	SRR3173724	4000000	150
SRX1588742	SRR3173728	4000000	150
SRX1588742	SRR3173727	4000000	150
SRX1588742	SRR3173720	4000000	150
SRX1588742	SRR3173734	4000000	150
SRX1588743	SRR3173745	4000000	150
SRX1588743	SRR3173735	4000000	150
SRX1588743	SRR3173736	4000000	150
SRX1588743	SRR3173737	4000000	150
SRX1588743	SRR3173739	4000000	150
SRX1588743	SRR3173746	4000000	150
SRX1588743	SRR3173738	3071585	150
SRX1588743	SRR3173743	4000000	150
SRX1588743	SRR3173741	4000000	150
SRX1588743	SRR3173742	4000000	150
SRX1588743	SRR3173744	4000000	150
SRX1588743	SRR3173740	4000000	150
SRX1588744	SRR3173748	4000000	150
SRX1588744	SRR3173749	4000000	150
SRX1588744	SRR3173758	4000000	150
SRX1588744	SRR3173747	4000000	150
SRX1588744	SRR3173750	4000000	150
SRX1588744	SRR3173760	4000000	150
SRX1588744	SRR3173753	4000000	150

SRX1588744	SRR3173752	1046635	150	
SRX1588744	SRR3173757	4000000	150	
SRX1588744	SRR3173754	4000000	150	
SRX1588744	SRR3173756	4000000	150	
SRX1588744	SRR3173759	4000000	150	
SRX1588744	SRR3173755	4000000	150	
SRX1588744	SRR3173751	4000000	150	
SRX1588745	SRR3173773	4000000	150	
SRX1588745	SRR3173764	4000000	150	
SRX1588745	SRR3173761	4000000	150	
SRX1588745	SRR3173769	4000000	150	
SRX1588745	SRR3173770	4000000	150	
SRX1588745	SRR3173771	4000000	150	
SRX1588745	SRR3173772	4000000	150	
SRX1588745	SRR3173766	4000000	150	
SRX1588745	SRR3173765	1895595	150	
SRX1588745	SRR3173767	4000000	150	
SRX1588745	SRR3173762	4000000	150	
SRX1588745	SRR3173763	4000000	150	
SRX1588745	SRR3173768	4000000	150	
SRX1613096	SRR3203635	2532461	487	
SRX1613097	SRR3203636	2542014	490	
SRX1613098	SRR3203637	3545611	490	
SRX1613099	SRR3203638	2941221	488	
SRX1659621	SRR3289718	9769976	148	
SRX1659627	SRR3289724	33833794	148	
SRX1659633	SRR3289731	9628178	148	
SRX1659639	SRR3289737	19771745	148	
SRX1674082	SRR3320128	58239349	160	
SRX1674083	SRR3320129	28865873	160	
SRX1674084	SRR3320130	1.02E+08	200	
SRX1674085	SRR3320131	88158798	200	
SRX1787479	SRR3560827	31266730	148	
SRX1787480	SRR3560828	26304787	148	
SRX1787481	SRR3560829	51690113	148	
SRX1787482	SRR3560830	53063970	148	
SRX181515	SRR548309	24364676	152	
SRX181516	SRR548310	26040051	152	
SRX181517	SRR548311	32472553	160	
SRX181518	SRR548312	23760240	160	
SRX2011752	SRR4017994	24294352	202	Used for aligner comparison
SRX2011753	SRR4017995	25025505	202	Used for aligner comparison
SRX2011754	SRR4017996	20929433	202	Used for aligner comparison
SRX2011755	SRR4017997	22988560	202	Used for aligner comparison
SRX2169993	SRR4252579	28198392	302	
SRX2169994	SRR4252580	26184995	302	
SRX2169995	SRR4252585	34305521	302	
SRX2169996	SRR4252590	44129079	302	
SRX2169997	SRR4252591	29786370	302	
SRX2169998	SRR4252592	26487596	302	
SRX2169999	SRR4252593	32450572	302	
SRX2170000	SRR4252594	31698842	302	
SRX2170001	SRR4252595	30386890	302	
SRX2170003	SRR4252596	29551539	302	
SRX2170004	SRR4252559	23249311	302	
SRX2170005	SRR4252560	22220737	302	

SRX2170006	SRR4252561	22977496	302
SRX2170007	SRR4252562	23042146	302
SRX2170008	SRR4252563	21127939	302
SRX2170009	SRR4252564	19671513	302
SRX2170010	SRR4252565	22183509	302
SRX2170011	SRR4252576	18432484	302
SRX2170012	SRR4252577	23122980	302
SRX2170013	SRR4252578	17943557	302
SRX2173094	SRR4253131	28198392	302
SRX2173095	SRR4253132	26184995	302
SRX2173096	SRR4253133	23249311	302
SRX2173097	SRR4253134	22220737	302
SRX2173098	SRR4253135	22977496	302
SRX2173099	SRR4253136	23042146	302
SRX2173100	SRR4253137	21127939	302
SRX2173101	SRR4253138	19671513	302
SRX2173102	SRR4253139	22183509	302
SRX2173103	SRR4253140	18432484	302
SRX2173104	SRR4253141	23122980	302
SRX2173105	SRR4253142	17943557	302
SRX2173106	SRR4253143	34305521	302
SRX2173107	SRR4253144	44129079	302
SRX2173108	SRR4253145	29786370	302
SRX2173109	SRR4253146	26487596	302
SRX2173110	SRR4253147	32450572	302
SRX2173111	SRR4253148	31698842	302
SRX2173112	SRR4253149	30386890	302
SRX2173113	SRR4253150	29551539	302
SRX2281970	SRR4478608	34896192	252
SRX2281971	SRR4478609	33943487	252
SRX2281972	SRR4478610	38173847	252
SRX2281973	SRR4478611	37959093	252
SRX2281974	SRR4478612	44214888	252
SRX2281975	SRR4478613	29253002	252
SRX2281976	SRR4478614	31310718	252
SRX2281977	SRR4478615	35110164	252
SRX2281978	SRR4478616	22940466	252
SRX2281979	SRR4478617	35592156	252
SRX2281980	SRR4478618	41255806	252
SRX2281981	SRR4478619	33412109	252
SRX2281982	SRR4478620	32523963	252
SRX2281983	SRR4478621	29189118	252
SRX2281984	SRR4478622	34994511	252
SRX2281985	SRR4478623	39003245	252
SRX2281986	SRR4478624	25407424	252
SRX2281987	SRR4478625	32702552	252
SRX2281988	SRR4478626	41908091	252
SRX2281989	SRR4478627	39311099	252
SRX2281990	SRR4478628	34240786	252
SRX2281991	SRR4478629	36959467	252
SRX2281992	SRR4478630	36742056	252
SRX2281993	SRR4478631	31081780	252
SRX2281994	SRR4478632	26927889	252
SRX2281995	SRR4478633	29261744	252
SRX2281996	SRR4478634	29047490	252
SRX2281997	SRR4478635	30641348	252

SRX2281998	SRR4478636	40564313	252
SRX2281999	SRR4478637	35494546	252
SRX2282000	SRR4478638	40304048	252
SRX2282001	SRR4478639	35175197	252
SRX2282002	SRR4478640	36538216	252
SRX2282003	SRR4478641	44156696	252
SRX2282004	SRR4478642	46066579	252
SRX2408830	SRR5091919	33919823	150
SRX2408831	SRR5091920	35674269	150
SRX2408832	SRR5091921	36425062	150
SRX2408833	SRR5091922	35728223	150
SRX2408834	SRR5091923	35577417	150
SRX2408835	SRR5091924	27923103	150
SRX2408836	SRR5091925	36907210	150
SRX2408837	SRR5091926	34143852	150
SRX2408838	SRR5091927	39360104	150
SRX2408839	SRR5091928	38500653	150
SRX2438632	SRR5123640	78431941	250
SRX2438633	SRR5123641	86368707	250
SRX2438634	SRR5123642	71376891	250
SRX2438635	SRR5123643	81240428	250
SRX2438636	SRR5123644	58624753	250
SRX2438637	SRR5123645	76009729	250
SRX2438638	SRR5123646	1E+08	250
SRX2438639	SRR5123647	78662555	250
SRX2438640	SRR5123648	71446557	250
SRX2438641	SRR5123649	1E+08	250
SRX2438642	SRR5123650	67862183	250
SRX2438643	SRR5123651	74406801	250
SRX2486694	SRR5170241	7306124	202
SRX2486695	SRR5170242	14338625	202
SRX2486696	SRR5170243	8828538	202
SRX2486697	SRR5170244	13342375	202
SRX2486698	SRR5170245	11783000	202
SRX2486699	SRR5170246	7546013	202
SRX2486700	SRR5170247	5263046	202
SRX2486701	SRR5170248	9341269	202
SRX2486702	SRR5170249	7964197	202
SRX2486703	SRR5170250	8234192	202
SRX2486704	SRR5170251	8829724	202
SRX2486705	SRR5170252	8457290	202
SRX2486706	SRR5170253	12633259	202
SRX2486707	SRR5170254	8616879	202
SRX2511392	SRR5195771	4624723	182
SRX2511393	SRR5195772	4890049	180
SRX2511394	SRR5195773	4123650	189
SRX2511395	SRR5195774	4904399	190
SRX2511396	SRR5195775	5495501	187
SRX2511397	SRR5195776	4523919	187
SRX2511398	SRR5195777	3255826	186
SRX2511399	SRR5195778	2752940	188
SRX2511400	SRR5195779	3830159	184
SRX2511401	SRR5195780	1883640	188
SRX2511402	SRR5195781	5692114	187
SRX2511403	SRR5195782	5043256	189
SRX2511404	SRR5195783	9292274	190

SRX2511405	SRR5195784	4723270	190
SRX2516749	SRR5202807	22502092	302
SRX2516750	SRR5202808	22486843	302
SRX2516751	SRR5202809	22136675	302
SRX2516752	SRR5202810	21396144	302
SRX2516753	SRR5202811	19596259	302
SRX2516754	SRR5202812	23206149	302
SRX2516755	SRR5202813	22977682	302
SRX2516756	SRR5202814	20702070	302
SRX2536729	SRR5227665	13678183	300
SRX2536730	SRR5227666	11291806	300
SRX2559213	SRR5253683	2543925	250
SRX2559214	SRR5253684	4181836	250
SRX2559215	SRR5253685	3105000	250
SRX2559216	SRR5253686	9735382	250
SRX2559217	SRR5253687	5420509	200
SRX2559218	SRR5253688	15027655	250
SRX2559219	SRR5253689	1569603	202
SRX2559220	SRR5253690	4256424	200
SRX2559221	SRR5253691	4506026	200
SRX2559222	SRR5253692	4394140	250
SRX2559223	SRR5253693	3429771	200
SRX2559224	SRR5253694	5387688	200
SRX2559225	SRR5253695	4750445	250
SRX2559226	SRR5253696	3977449	200
SRX2559227	SRR5253697	3009898	200
SRX2622492	SRR5322181	61497489	200
SRX2622493	SRR5322182	53665921	200
SRX2622494	SRR5322183	57461969	200
SRX2622495	SRR5322184	56904518	200
SRX2622496	SRR5322185	65279613	200
SRX2622497	SRR5322186	60731620	200
SRX2728030	SRR5438096	35881605	300
SRX2728031	SRR5438097	18383701	300
SRX2728032	SRR5438098	14474155	300
SRX2728033	SRR5438099	12460135	300
SRX2744284	SRR5456157	34250907	202
SRX2744285	SRR5456158	34242415	202
SRX2744286	SRR5456159	34428444	202
SRX2744287	SRR5456160	31781088	202
SRX2744288	SRR5456161	27637922	202
SRX2744289	SRR5456162	36523034	202
SRX2744290	SRR5456163	33052551	202
SRX2744291	SRR5456164	29141829	202
SRX2744292	SRR5456165	29427154	202
SRX2744293	SRR5456166	26980711	202
SRX2744294	SRR5456167	37502518	202
SRX2744295	SRR5456168	27930284	202
SRX2795681	SRR5526359	21364193	200
SRX2795682	SRR5526358	17792489	200
SRX2795683	SRR5526357	30334498	200
SRX2795684	SRR5526356	21811751	200
SRX2795685	SRR5526355	25699239	200
SRX2795686	SRR5526354	29354855	200
SRX2795687	SRR5526353	22087467	200
SRX2795688	SRR5526352	21012716	200

SRX2826535	SRR5564855	34113492	300	
SRX2826536	SRR5564856	35553613	300	
SRX2826537	SRR5564857	30414544	300	
SRX2826538	SRR5564858	31620926	300	
SRX2826539	SRR5564859	24120107	300	
SRX2826540	SRR5564860	24507949	300	
SRX2826541	SRR5564861	25639020	300	
SRX2826542	SRR5564862	29097209	300	
SRX2826543	SRR5564863	35008262	300	
SRX2826544	SRR5564864	44102929	300	
SRX2826545	SRR5564865	42883270	300	
SRX2826546	SRR5564866	37342641	300	
SRX2826547	SRR5564867	38049324	300	
SRX2826548	SRR5564868	27463001	300	
SRX2826549	SRR5564869	27752003	300	
SRX2859382	SRR5606856	20912780	202	
SRX2859383	SRR5606855	21486458	202	
SRX2859384	SRR5606852	19244019	202	
SRX2859385	SRR5606851	19993393	202	
SRX2859386	SRR5606853	18303252	202	
SRX2859387	SRR5606850	19820749	202	
SRX2859388	SRR5606854	19733741	202	
SRX2859389	SRR5606849	19924092	202	
SRX2859390	SRR5606848	18418220	202	
SRX2859391	SRR5606847	17930302	202	
SRX286929	SRR868958	53017572	200	
SRX286930	SRR868932	35210683	200	
SRX286931	SRR868957	35383355	200	
SRX286932	SRR868939	30383493	200	
SRX286933	SRR868942	35202128	200	
SRX2953321	SRR5753106	11907096	250	Used for aligner comparison
SRX2953322	SRR5753105	11659943	250	Used for aligner comparison
SRX2953323	SRR5753104	12882453	250	Used for aligner comparison
SRX2953324	SRR5753103	12426588	250	Used for aligner comparison
SRX2953325	SRR5753102	11716965	250	Used for aligner comparison
SRX2953326	SRR5753101	11768527	250	Used for aligner comparison
SRX3009489	SRR5832182	5357669	202	
SRX3009490	SRR5832183	5707329	202	
SRX3009491	SRR5832184	4980094	202	
SRX3009492	SRR5832185	5027165	202	
SRX3009493	SRR5832186	5378513	202	
SRX3009494	SRR5832187	5369530	202	
SRX3009495	SRR5832188	5212740	202	
SRX3009496	SRR5832189	5508363	202	
SRX3009497	SRR5832190	4247703	202	
SRX3009498	SRR5832191	4763495	202	
SRX3009499	SRR5832192	4987388	202	
SRX3009500	SRR5832193	4374610	202	
SRX3009501	SRR5832194	4059608	202	
SRX3009502	SRR5832195	4503587	202	
SRX3009503	SRR5832196	4141000	202	
SRX3009504	SRR5832197	4525879	202	
SRX3009505	SRR5832198	4000456	202	
SRX3009506	SRR5832199	2858919	202	
SRX3020076	SRR5849892	24921719	200	
SRX3020076	SRR5849891	20361844	200	

SRX3020077	SRR5849893	19705507	200
SRX3020077	SRR5849894	24023791	200
SRX3020078	SRR5849895	19791784	200
SRX3020078	SRR5849896	24335388	200
SRX3020079	SRR5849898	20854815	200
SRX3020079	SRR5849897	17163283	200
SRX3020080	SRR5849900	23885134	200
SRX3020080	SRR5849899	19576294	200
SRX3020081	SRR5849902	17940756	200
SRX3020081	SRR5849901	14786549	200
SRX3020082	SRR5849903	16074666	200
SRX3020082	SRR5849904	19450184	200
SRX3020083	SRR5849905	17602306	200
SRX3020083	SRR5849906	21330660	200
SRX3020084	SRR5849907	17023708	200
SRX3020084	SRR5849908	20847614	200
SRX3020085	SRR5849909	21823815	200
SRX3020085	SRR5849910	26553198	200
SRX3020086	SRR5849912	18093028	200
SRX3020086	SRR5849911	14891075	200
SRX3020087	SRR5849914	22295545	200
SRX3020087	SRR5849913	18211053	200
SRX3020088	SRR5849916	10252757	200
SRX3020088	SRR5849915	9926463	200
SRX3020089	SRR5849918	8563425	200
SRX3020089	SRR5849917	8322764	200
SRX3020090	SRR5849919	11119330	200
SRX3020090	SRR5849920	11482481	200
SRX3020091	SRR5849921	7850118	200
SRX3020091	SRR5849922	8089330	200
SRX3020092	SRR5849923	13753207	200
SRX3020092	SRR5849924	14119483	200
SRX3020093	SRR5849926	10995200	200
SRX3020093	SRR5849925	10631396	200
SRX3020094	SRR5849927	9795620	200
SRX3020094	SRR5849928	10090040	200
SRX3020095	SRR5849930	11173086	200
SRX3020095	SRR5849929	10823285	200
SRX3020096	SRR5849932	10629582	200
SRX3020096	SRR5849931	10304572	200
SRX3020097	SRR5849934	9677453	200
SRX3020097	SRR5849933	9415382	200
SRX3020098	SRR5849936	8846655	200
SRX3020098	SRR5849935	8598262	200
SRX3020099	SRR5849937	9077957	200
SRX3020099	SRR5849938	9361753	200
SRX3020100	SRR5849939	11357671	200
SRX3020100	SRR5849940	11719879	200
SRX3020101	SRR5849942	11971946	200
SRX3020101	SRR5849941	11624879	200
SRX3020102	SRR5849944	12013434	200
SRX3020102	SRR5849943	11642372	200
SRX3020103	SRR5849946	19977972	200
SRX3020103	SRR5849945	19328833	200
SRX3020104	SRR5849947	14934967	200
SRX3020104	SRR5849948	15409180	200

SRX3020105	SRR5849950	9649852	200
SRX3020105	SRR5849949	9368883	200
SRX3020106	SRR5849952	11241573	200
SRX3020106	SRR5849951	10911305	200
SRX3020107	SRR5849954	9816200	200
SRX3020107	SRR5849953	9510548	200
SRX3020108	SRR5849955	8446794	200
SRX3020108	SRR5849956	8710722	200
SRX3020109	SRR5849957	10241515	200
SRX3020109	SRR5849958	10613339	200
SRX3020110	SRR5849959	12482542	200
SRX3020110	SRR5849960	12887607	200
SRX3020111	SRR5849962	10736112	200
SRX3020111	SRR5849961	10418983	200
SRX3020952	SRR5851344	5186519	170
SRX3020953	SRR5851343	7982585	170
SRX3020954	SRR5851342	6061543	170
SRX3020955	SRR5851341	9185791	170
SRX3020956	SRR5851340	8232172	170
SRX3020957	SRR5851339	4969176	170
SRX3020958	SRR5851338	1849196	170
SRX3020959	SRR5851337	7356011	170
SRX3020960	SRR5851336	11254160	170
SRX3165815	SRR6012260	69039219	200
SRX3165817	SRR6012258	69305221	200
SRX3165819	SRR6012256	75272434	200
SRX3165823	SRR6012252	73278584	200
SRX3165825	SRR6012250	83132676	200
SRX3229756	SRR6117023	18366023	152
SRX3229757	SRR6117022	18749183	152
SRX3229758	SRR6117021	21219414	152
SRX3229759	SRR6117020	19248855	152
SRX3241954	SRR6129524	1.92E+08	200
SRX3241955	SRR6129523	2.07E+08	200
SRX3241956	SRR6129522	1.7E+08	200
SRX3241957	SRR6129521	1.8E+08	200
SRX3241958	SRR6129520	1.98E+08	200
SRX3241959	SRR6129519	1.95E+08	200
SRX3346330	SRR6238092	56458162	202
SRX3346331	SRR6238093	1.3E+08	200
SRX3346332	SRR6238094	83132811	250
SRX3346333	SRR6238095	1.89E+08	200
SRX3346334	SRR6238096	1.8E+08	200
SRX3346335	SRR6238097	58776433	202
SRX3346336	SRR6238098	1.72E+08	300
SRX3346337	SRR6238099	2.37E+08	250
SRX3346338	SRR6238100	1.7E+08	250
SRX3346339	SRR6238101	1.94E+08	250
SRX3346340	SRR6238102	55646159	202
SRX3346341	SRR6238103	1.69E+08	200
SRX3346342	SRR6238104	48548249	250
SRX3346343	SRR6238105	1.28E+08	200
SRX3346344	SRR6238106	1.76E+08	200
SRX3346345	SRR6238107	60816434	202
SRX3346346	SRR6238108	14420643	200
SRX3346347	SRR6238109	1.07E+08	250

SRX3346348	SRR6238110	3.19E+08	250
SRX3346349	SRR6238111	1.48E+08	250
SRX335720	SRR953117	57236939	200
SRX335721	SRR953118	55206405	200
SRX335722	SRR953119	55089686	200
SRX335723	SRR953120	55763048	200
SRX335724	SRR953121	58547396	200
SRX335725	SRR953122	55407700	200
SRX335726	SRR953123	61374457	180
SRX335727	SRR953124	58488606	200
SRX335728	SRR953125	58319935	200
SRX335729	SRR953126	56900370	200
SRX335730	SRR953127	64860192	180
SRX335731	SRR953128	56004565	200
SRX335732	SRR953129	58704248	200
SRX335733	SRR953130	57326625	200
SRX335734	SRR953131	63041961	180
SRX335735	SRR953132	57592182	200
SRX335736	SRR953133	56244723	200
SRX360655	SRR1003113	81103258	202
SRX360655	SRR1003098	48225301	202
SRX360860	SRR1003268	34426904	202
SRX360860	SRR1003270	1.39E+08	202
SRX362654	SRR1006135	11056166	202
SRX362654	SRR1006136	11056166	202
SRX362655	SRR1006196	1.11E+08	202
SRX362686	SRR1006197	92914953	202
SRX362687	SRR1006198	21746278	202
SRX362860	SRR1006413	22229137	202
SRX362868	SRR1006417	72565372	202
SRX362869	SRR1010357	1.28E+08	202
SRX392694	SRR1050769	46271268	146
SRX392695	SRR1050770	51202173	146
SRX392697	SRR1050772	46213054	146
SRX392698	SRR1050773	48101699	146
SRX392703	SRR1050778	13760189	152
SRX392705	SRR1050780	15312869	152
SRX437618	SRR1125001	32809423	200
SRX475894	SRR1176664	16340848	202
SRX475895	SRR1176665	10286620	202
SRX475896	SRR1176666	17322035	202
SRX514835	SRR1233915	1.7E+08	200
SRX514835	SRR1261335	1.22E+08	202
SRX533796	SRR1272308	70549834	200
SRX533797	SRR1272309	66884309	200
SRX533798	SRR1272310	76650688	202
SRX533799	SRR1272311	82290914	200
SRX533800	SRR1272312	81330815	200
SRX533801	SRR1272313	69744890	202
SRX533802	SRR1272314	26288552	200
SRX533803	SRR1272315	26201075	200
SRX533804	SRR1272316	1.04E+08	202
SRX533805	SRR1272317	20298431	152
SRX533806	SRR1272318	20820923	152
SRX533807	SRR1272319	25386247	200
SRX533808	SRR1272320	24798076	200

SRX533809	SRR1272321	88575503	202	
SRX533810	SRR1272322	21512129	152	
SRX559705	SRR1313054	42072663	202	
SRX559706	SRR1313055	42072663	202	
SRX559707	SRR1313056	35380262	202	
SRX559708	SRR1313057	35380262	202	
SRX559709	SRR1313058	1.62E+08	202	
SRX559710	SRR1313059	1.64E+08	202	
SRX559711	SRR1313060	41266636	152	
SRX559712	SRR1313061	41266636	152	
SRX659943	SRR1523361	38073370	152	
SRX659944	SRR1523362	23290441	152	
SRX659945	SRR1523363	19930170	152	
SRX659946	SRR1523364	15928789	152	
SRX659947	SRR1523365	2.29E+08	200	Used for aligner comparison
SRX659948	SRR1523366	2.28E+08	200	Used for aligner comparison
SRX659949	SRR1523367	1.85E+08	200	Used for aligner comparison
SRX659950	SRR1523368	2.07E+08	200	Used for aligner comparison
SRX669136	SRR1536002	25744308	202	Used for aligner comparison
SRX669137	SRR1536003	26310691	202	Used for aligner comparison
SRX669138	SRR1536004	28104498	202	Used for aligner comparison
SRX669139	SRR1536005	33133494	202	Used for aligner comparison
SRX669140	SRR1536006	25387888	202	Used for aligner comparison
SRX669141	SRR1536007	26261998	202	Used for aligner comparison
SRX669142	SRR1536008	29374210	202	Used for aligner comparison
SRX669143	SRR1536009	38583681	202	Used for aligner comparison
SRX669144	SRR1536010	29605768	202	Used for aligner comparison
SRX669145	SRR1536011	30257678	202	Used for aligner comparison
SRX669146	SRR1536012	24794639	202	Used for aligner comparison
SRX669147	SRR1536013	37052793	202	Used for aligner comparison
SRX669148	SRR1536014	36188562	202	Used for aligner comparison
SRX669149	SRR1536015	34360409	202	Used for aligner comparison
SRX669150	SRR1536016	31846472	202	Used for aligner comparison
SRX669151	SRR1536017	24786881	202	Used for aligner comparison
SRX669152	SRR1536018	28808906	202	Used for aligner comparison
SRX669153	SRR1536019	27506259	202	Used for aligner comparison
SRX669154	SRR1536020	25561158	202	Used for aligner comparison
SRX669155	SRR1536021	24287434	202	Used for aligner comparison
SRX669156	SRR1536022	21095958	202	Used for aligner comparison
SRX669157	SRR1536023	26068309	202	Used for aligner comparison
SRX669158	SRR1536024	35199828	202	Used for aligner comparison
SRX669159	SRR1536025	65032209	202	Used for aligner comparison
SRX669160	SRR1536026	28622952	202	Used for aligner comparison
SRX669161	SRR1536027	24339934	202	Used for aligner comparison
SRX669162	SRR1536028	63406758	202	Used for aligner comparison
SRX669163	SRR1536029	28689617	202	Used for aligner comparison
SRX669164	SRR1536030	29581396	202	Used for aligner comparison
SRX669165	SRR1536031	26965651	202	Used for aligner comparison
SRX669166	SRR1536032	24245422	202	Used for aligner comparison
SRX669167	SRR1536033	25172351	202	Used for aligner comparison
SRX669168	SRR1536034	36146088	202	Used for aligner comparison
SRX669169	SRR1536035	32614319	202	Used for aligner comparison
SRX669170	SRR1536036	29137208	202	Used for aligner comparison
SRX669171	SRR1536037	27185773	202	Used for aligner comparison
SRX669172	SRR1536038	26622238	202	Used for aligner comparison
SRX669173	SRR1536039	25532094	202	Used for aligner comparison

SRX669174	SRR1536040	33968209	202	Used for aligner comparison
SRX669175	SRR1536041	35362797	202	Used for aligner comparison
SRX669176	SRR1536042	25268183	202	Used for aligner comparison
SRX669177	SRR1536043	28347371	202	Used for aligner comparison
SRX669178	SRR1536044	25251775	202	Used for aligner comparison
SRX669179	SRR1536045	27027231	202	Used for aligner comparison
SRX669180	SRR1536046	27180524	202	Used for aligner comparison
SRX669181	SRR1536047	28541781	202	Used for aligner comparison
SRX669182	SRR1536048	31679354	202	Used for aligner comparison
SRX669183	SRR1536049	21834608	202	Used for aligner comparison
SRX669184	SRR1536050	35931029	202	Used for aligner comparison
SRX669185	SRR1536051	37469456	202	Used for aligner comparison
SRX669186	SRR1536052	31830914	202	Used for aligner comparison
SRX669187	SRR1536053	52135271	202	Used for aligner comparison
SRX669188	SRR1536054	28795671	202	Used for aligner comparison
SRX669189	SRR1536055	27602564	202	Used for aligner comparison
SRX669190	SRR1536056	54077673	202	Used for aligner comparison
SRX669191	SRR1536057	31637764	202	Used for aligner comparison
SRX688584	SRR1560104	40286177	172	Used for aligner comparison
SRX688585	SRR1560105	40242431	172	
SRX688586	SRR1560106	34844871	172	
SRX688587	SRR1560107	45066804	172	
SRX704261	SRR1578745	51093813	200	
SRX704262	SRR1578746	61237752	200	
SRX704263	SRR1578747	54013714	200	
SRX707276	SRR1582059	1.44E+08	152	
SRX707279	SRR1582062	1.45E+08	152	
SRX707290	SRR1582073	87211981	202	
SRX707291	SRR1582074	85146261	202	
SRX707292	SRR1582075	43331188	200	
SRX707293	SRR1582076	89090759	200	
SRX707294	SRR1582077	82759535	202	
SRX707295	SRR1582078	37065178	200	
SRX707296	SRR1582079	24212675	200	
SRX709649	SRR1585277	40302838	152	
SRX709650	SRR1585278	50516835	152	
SRX709651	SRR1585279	46094277	152	
SRX709652	SRR1585280	47059209	152	
SRX732432	SRR1611854	5982820	200	
SRX763579	SRR1657113	15557274	202	
SRX763580	SRR1657114	17686675	202	
SRX763581	SRR1657115	18470906	202	
SRX819627	SRR1727796	23296814	202	Used for aligner comparison
SRX819628	SRR1727797	19246188	202	Used for aligner comparison
SRX819629	SRR1727798	33874974	202	Used for aligner comparison
SRX819630	SRR1727799	34898434	202	Used for aligner comparison
SRX819631	SRR1727800	28871680	202	Used for aligner comparison
SRX819632	SRR1727801	24833518	202	Used for aligner comparison
SRX819633	SRR1727802	22660816	202	Used for aligner comparison
SRX819634	SRR1727803	26985230	202	Used for aligner comparison
SRX819635	SRR1727804	35086157	202	Used for aligner comparison
SRX819636	SRR1727805	31832957	202	Used for aligner comparison
SRX819637	SRR1727806	35455676	202	Used for aligner comparison
SRX819638	SRR1727807	24201482	202	Used for aligner comparison
SRX819639	SRR1727808	27036883	202	Used for aligner comparison
SRX819640	SRR1727809	21922246	202	Used for aligner comparison

SRX819641	SRR1727810	34038890	202	Used for aligner comparison
SRX819642	SRR1727811	28942941	202	Used for aligner comparison
SRX819643	SRR1727812	30843145	202	Used for aligner comparison
SRX819644	SRR1727813	31542819	202	Used for aligner comparison
SRX819645	SRR1727814	20513599	202	Used for aligner comparison
SRX819646	SRR1727815	24276604	202	Used for aligner comparison
SRX819647	SRR1727816	38994766	202	Used for aligner comparison
SRX819648	SRR1727817	21437496	202	Used for aligner comparison
SRX819649	SRR1727818	32306353	202	Used for aligner comparison
SRX819650	SRR1727819	31873013	202	Used for aligner comparison
SRX833896	SRR1746094	18147994	199	
SRX834375	SRR1525432	14793166	199	
SRX834376	SRR1746748	15093948	200	
SRX834377	SRR1525433	14798098	200	
SRX834378	SRR1746752	13048576	200	
SRX834379	SRR1746751	13069714	200	
SRX834380	SRR1746750	11999998	199	
SRX834381	SRR1746749	12786352	199	

Supplemental Table 2. Introns listed as "confirmed" in WormBase that were not detected.

Gene	Intron	Strand	Relative Position	Splice Junction
shc-1	I:971951-976480	-	5' terminal	GT/AG
atg-5	I:1709119-1709586	+	5' terminal	GT/AG
W10C8.4	I:2852386-2852498	-	3' terminal	CT/GC
nol-5	I:3269114-3269143	+	3' terminal	GC/AG
C50F2.4	I:3888467-3889071	-	3' terminal	CT/AC
mes-3	I:5001650-5001811	+	Internal	GT/AG
mat-1	I:5125911-5126035	-	Internal	CT/AC
rpl-19	I:5485745-5486050	-	5' terminal	CT/AC
gpa-14	I:5942612-5942826	+	5' terminal	GT/AG
ech-1.2	I:6209940-6211081	-	5' terminal	GT/AG
acdh-3	I:6465710-6466801	-	5' terminal	GT/AG
che-1	I:6518633-6519990	-	Internal	GT/AG
try-6	I:6586709-6586759	+	Internal	GC/AG
smg-1	I:6902351-6903649	-	Internal	GT/AG
pck-2	I:7871485-7871821	+	3' terminal	GT/AG
ngp-1	I:8398111-8398233	+	Internal	GC/AG
madd-4	I:8940278-8940561	-	Internal	GT/AG
usp-48	I:9507905-9508000	+	Internal	GC/AG
K02A11.4	I:9748301-9748515	-	3' terminal	GT/AG
eif-3.C	I:9976644-9976688	-	3' terminal	CT/GC
pab-1	I:10434982-10435170	-	Internal	GC/AG
lrk-1	I:10894092-10894138	-	5' terminal	GC/AG
vab-10	I:11774792-11774835	-	Internal	GT/AG
glct-1	I:12338051-12338099	+	5' terminal	GT/AG
glct-3	I:12385876-12387848	+	5' terminal	GT/AG
H28O16.1	I:12652934-12653038	+	Internal	CT/GC
acox-1	I:12938902-12939302	+	5' terminal	GC/AG
F08A8.5	I:12957416-12957517	-	Internal	GT/AG
Y26D4A.8	I:13085220-13087942	-	3' terminal	GT/AG
C17H1.2	I:13085220-13087942	-	3' terminal	GT/AG

unc-122	I:14873345-14873820	+	Internal	GT/AG
K10B4.3	II:120330-120376	-	Internal	GC/AG
F48A11.4	II:216385-216493	-	Internal	GT/AG
math-42	II:2099038-2099931	+	5' terminal	GT/AG
Y46D2A.3	II:3326811-3327031	-	5' terminal	GT/AG
F53C3.13	II:3920688-3920730	-	3' terminal	GC/AG
W06A11.1	II:4069562-4069737	+	3' terminal	GT/AG
nhr-109	II:4435170-4435358	+	3' terminal	GT/AG
nhr-273	II:4435170-4435358	+	3' terminal	GT/AG
C04G6.13	II:5094406-5094582	+	5' terminal	GT/AG
cdc-14	II:5589011-5590397	-	Internal	GT/AG
tat-4	II:6235506-6236237	-	5' terminal	GC/AG
C56C10.7	II:6586171-6586264	-	Internal	GT/AG
syd-1	II:7586163-7586634	-	3' terminal	GT/AG
klp-3	II:7843482-7844033	+	3' terminal	GT/AG
chil-9	II:9845339-9846927	-	5' terminal	GT/AG
ZK938.8	II:9845339-9846927	-	5' terminal	GT/AG
mpz-1	II:10759610-10759675	-	Internal	GT/AG
clh-2	II:11368100-11368137	+	Internal	GT/AG
W03C9.6	II:11953459-11954507	+	5' terminal	GC/AG
clec-146	II:13590565-13590642	-	3' terminal	CT/AC
nurf-1	II:14405831-14406656	+	5' terminal	CT/GC
eif-3.B	II:14795984-14796289	+	3' terminal	GC/AG
C24A1.3	III:692743-692936	-	3' terminal	CT/AC
Y71D11A.3	III:1140097-1140364	+	Internal	GT/AG
Y71D11A.3	III:1140452-1141817	+	Internal	GT/AG
gop-3	III:5263838-5263942	+	3' terminal	GC/AG
pqe-1	III:5312138-5312613	+	5' terminal	GT/AG
szy-2	III:5371275-5372277	-	5' terminal	GT/AG
mig-21	III:5878039-5878275	-	5' terminal	GT/AG
C16A3.10	III:6394992-6395225	-	Internal	GC/AG
dig-1	III:6757591-6757746	+	Internal	GT/AG
kap-1	III:7339071-7339176	+	5' terminal	GT/AG
lig-4	III:7523791-7523840	+	Internal	GC/AG
pcp-5	III:7907609-7907818	-	5' terminal	GT/AG
plk-1	III:8101615-8101658	+	5' terminal	GT/AG
mig-39	III:8468340-8472502	+	5' terminal	GT/AG
ceh-16	III:8622717-8622885	+	5' terminal	GT/AG
mig-22	III:8763305-8763759	-	3' terminal	GT/AG
tpk-1	III:8913048-8913136	+	Internal	GT/AG
T16H12.3	III:10081832-10082077	-	Internal	GT/AG
atx-2	III:10466583-10467037	+	Internal	GC/AG
arrd-16	III:12486099-12488136	+	3' terminal	GC/AG
ant-1.1	III:13463651-13463821	+	3' terminal	GC/AG
ant-1.1	III:13463890-13463992	+	3' terminal	GC/AG
lit-1	III:13714185-13714557	+	5' terminal	GT/AG
grld-1	IV:124853-124921	-	Internal	GT/AG
T21D12.7	IV:290595-290638	-	Internal	GC/AG
nog-1	IV:394549-396188	+	3' terminal	CT/AC
Y77E11A.7	IV:1419457-1419507	-	Internal	GC/AG

Y69A2AR.18	IV:2496427-2496549	+	Internal	CT/AC
srw-95	IV:4812230-4812276	+	5' terminal	GC/AG
Y4C6B.3	IV:5334010-5334078	+	5' terminal	GT/AG
tin-9.1	IV:7043975-7044166	-	3' terminal	GT/AG
srx-50	IV:7058455-7058585	+	5' terminal	GT/AG
unc-8	IV:7198531-7198903	+	5' terminal	GT/AG
unc-8	IV:7198962-7199531	+	Internal	GT/AG
F55G1.6	IV:7474110-7474154	+	Internal	GC/AG
klp-11	IV:8756137-8756240	-	Internal	GT/AG
klp-11	IV:8756808-8757090	-	Internal	GT/AG
klp-11	IV:8760133-8760174	-	Internal	GT/AG
exc-5	IV:8799821-8799881	-	Internal	GT/AG
ZC410.5	IV:9090436-9091845	+	3' terminal	GT/AG
ugt-21	IV:9501436-9501483	+	Internal	GC/AG
pyp-1	IV:9996792-9997555	-	Internal	GC/AG
mboa-4	IV:11157227-11157454	-	3' terminal	GT/AG
mboa-4	IV:11158576-11158632	-	Internal	GT/AG
sru-20	IV:12497623-12497667	+	3' terminal	GC/AG
Y37A1A.4	IV:13938543-13938639	+	5' terminal	GT/AG
ntl-11	IV:15467733-15468014	-	5' terminal	GC/AG
gcy-27	IV:17435975-17436025	-	Internal	GT/AG
srbc-18	V:2169929-2169981	+	Internal	GC/AG
C29G2.2	V:2590674-2590805	+	Single-intron	GT/AG
nuo-5	V:2701028-2702444	+	3' terminal	GT/AG
ketn-1	V:2797798-2797869	-	Internal	GT/AG
ketn-1	V:2798831-2799118	-	Internal	GT/AG
ketn-1	V:2799538-2799771	-	5' terminal	GT/AG
Y73C8B.5	V:3197361-3197942	+	Internal	GT/AG
F32D1.8	V:4376629-4376700	-	3' terminal	GT/AG
glb-5	V:5561939-5563228	+	Internal	GT/AG
F44E7.12	V:5790365-5790410	-	Single-intron	CT/AC
clik-1	V:6766779-6767194	-	Internal	CT/AC
frpr-18	V:6872424-6872902	+	Internal	GT/AG
F41E6.1	V:8622664-8622763	+	5' terminal	GT/AG
hum-2	V:9386287-9386326	+	5' terminal	GT/AG
hum-2	V:9386335-9386383	+	Internal	GT/AG
str-118	V:10053039-10053715	+	Internal	GT/AG
egl-3	V:10171512-10173475	-	5' terminal	GT/AG
aqp-6	V:10648892-10650048	+	Single-intron	GT/AG
mig-17	V:11446740-11446790	+	5' terminal	GC/AG
tre-3	V:11714753-11714812	-	Internal	GT/AG
twk-24	V:12266070-12266711	-	3' terminal	GT/AG
cdr-7	V:12411693-12411733	-	Internal	GT/AG
gpa-13	V:12754541-12754863	-	5' terminal	GT/AG
ret-1	V:14830318-14830526	-	5' terminal	GC/AG
T26E4.9	V:15799522-15802626	-	5' terminal	GT/AG
nhr-233	V:16569730-16569774	-	3' terminal	GT/AG
F59A1.11	V:17661984-17662349	-	3' terminal	GT/AG
Y43F8B.10	V:19474799-19476618	+	Single-intron	GT/AG
sri-67	V:19989967-19990019	+	3' terminal	GT/AG

dct-16	V:20498239-20498276	-	3' terminal	CT/GC
mrp-1	X:579388-579515	-	Internal	GT/AG
R160.5	X:4370159-4370535	-	3' terminal	GT/AG
pdi-2	X:4525173-4525340	+	Internal	CT/AC
acn-1	X:5093785-5096548	-	Internal	GT/AG
syx-3	X:5351136-5351809	+	3' terminal	GC/AG
klp-13	X:5987936-5988658	+	Internal	GT/AG
got-2.2	X:6241738-6242076	+	3' terminal	GC/AG
C03B1.6	X:6350681-6351082	+	3' terminal	GT/AG
abts-4	X:6748475-6751456	-	Internal	GT/AG
21ur-10165	X:6748475-6751456	-	Internal	GT/AG
C55B6.1	X:7201497-7201864	+	Internal	GC/AG
eef-1A.2	X:7823869-7825225	+	Single-intron	GC/AG
cca-1	X:7854202-7854305	-	Internal	GT/AG
F16F9.3	X:8458275-8458316	+	3' terminal	GT/AG
chup-1	X:8794504-8795662	+	3' terminal	GT/AG
gly-13	X:9295930-9295980	-	Internal	GT/AG
gap-2	X:9513025-9513130	+	5' terminal	GT/AG
nhr-214	X:12700963-12701267	-	3' terminal	GT/AG
T14G8.3	X:12861301-12861351	+	5' terminal	GC/AG
F11C1.5	X:12993598-12998465	+	Internal	GT/AG
odr-1	X:13550489-13550656	-	5' terminal	GT/AG
unc-84	X:13588998-13589153	+	Internal	GT/AG
prx-1	X:14317682-14317754	+	Internal	GC/AG
C11H1.9	X:14319231-14319303	-	Internal	GC/AG
ram-5	X:14556017-14556064	+	3' terminal	GC/AG
tag-53	X:14710529-14711578	+	5' terminal	GT/AG
tag-53	X:14715239-14715498	+	Internal	GT/AG
mbl-1	X:17006924-17006972	+	Internal	GT/AG
C08A9.3	X:17091435-17093205	-	5' terminal	GT/AG
vap-1	X:17394825-17394867	-	Internal	GT/AG