C. briggsae genome annotation and comparative analysis with *C. elegans* using RNA-Seq data

by

Shinta Thio

B.Sc., University of Surabaya, 2013

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

SIMON FRASER UNIVERSITY
Spring 2020

Approval

Name:	**Shinta Thio**
Degree:	**Master of Science**
Title:	***C. briggsae* genome annotation and comparative analysis with *C. elegans* using RNA-Seq data**
Examining Committee:	**Chair:** **Mark Paetzel** Professor
	Jack Chen Senior Supervisor Professor
	Fiona Brinkman Supervisor Professor
	Ryan Morin Supervisor Associate Professor
	Christopher Beh Internal Examiner Professor
Date Defended/Approved:	April 6, 2020

Abstract

Complete genome annotations are essential for comparative genomics. Currently, the *C. briggsae* genome annotation is incomplete that limits its utility as a comparative platform for *C. elegans*. Using RNA-Seq data, we have generated a more complete *C. briggsae* genome annotation. We identified 20,660 novel introns, 35,635 novel exons, and 5,654 novel protein-coding transcripts, and generated improved databases consisting of 123,974 introns, 150,690 exons, and 28,129 protein-coding transcripts, respectively. The improved *C. briggsae* annotation together with comparative analyses revealed 132 novel ortholog relationships (between *C. briggsae* and *C. elegans*) and 2 novel *C. elegans* protein-coding genes. This has shown that despite limited data available for *C. briggsae*, the improved annotation has enhanced the utility of *C. briggsae* as a comparative platform for *C. elegans*. As more RNA-Seq data becomes available, this method can be used to further refine not only *C. briggsae* annotation but also *C. elegans* annotation.

Keywords: *Caenorhabditis briggsae*, *Caenorhabditis elegans*, Comparative genomics, RNA-Seq, Transcriptome, Improved annotations

This work is dedicated to my parents and siblings for their unconditional love, endless support, encouragement, prayers, and sacrifices.

Karya ini saya persembahkan untuk Mama, Papa, Koko, Dajie, Erjie tercinta atas segala pengorbanan, cinta kasih, doa, nasehat dan semangat yang selalu diberikan.

Acknowledgements

Firstly, I would like to express my sincere gratitude towards my supervisor Dr. Jack Chen for the opportunity to work and learn in his lab, thirteen thousand kilometers away from home. This project could not have been done without his continuous support, patience, motivation, encouragement, and knowledge. He has also taught me invaluable lessons not only on how to be a better scientist, but also to be a better person in life.

I would also like to extend my gratitude to my committee members, Dr. Fiona Brinkman and Dr. Ryan Morin for their guidance, feedback, and inspiration throughout my graduate degree. I am grateful to Fiona for her positivity and spirit during my research project. I am thankful for Ryan for his essential advice. Thanks also go to my examining committee, Dr. Christopher Beh, thank you for taking the time to read my thesis and attend my defence, and Dr. Mark Paetzel, for chairing my defence.

I would also like to thank the past and present graduate students of the Chen lab, especially Matthew Douglas, Marija Jovanovic, and Kate Gibson. I am so grateful for our journey together, all your support, encouragement, friendship, inside and outside the lab. Thank you to Dr. Jiarui Li and Dr. Michelle Hu, for the friendliness, knowledge, mature perspectives, and advice over the past few years.

Special thanks to my parents, sisters, and brother, who have always been there for me, supporting me to pursue my dream studying abroad, and for always believing in me. You are my everything.

Finally, thanks to Adair family, Uplands family, Indo friends, CG friends, MBB friends, SFU Omics friends, LYMT friends, and everyone for your continuous love, support, and encouragement throughout my graduate school journey. I would not be who I am today without you. Thank you all.

Table of Contents

List of Tables

List of Figures

List of Acronyms

BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA
CDS	Coding Sequence
E-value	Expect value
GBrowse	Generic genome Browser
GeMoMa	Gene Model Mapper
GFF	General Feature Format
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
rRNA	Ribosomal RNA
UTR	Untranslated Region

Glossary

Acceptor or 3’ splice site	Intron-exon boundary (the 3’ side of an intron and the 5’ side of an exon)
Bit score	A normalized alignment score. Represents a statistical significance of an alignment
BLASTP	Program that searches protein databases using a protein query
cDNA	DNA synthesized from an mRNA template via reverse transcription
CDS	A segment of genomic DNA or mRNA that can be divided into codons and translated into a protein. In this thesis, CDS is used interchangeably with exon
Donor or 5’ splice site	Exon-intron boundary (the 3’ side of an exon and the 5’ side of an intron)
E-value	Describes the number of hits one can expect to see by chance during database searching (i.e., the probability that two sequences are similar to each other by chance)
Exon	Parts of the pre-mRNA that are spliced and joined together to form mature mRNA
GBrowse	A web-based genome browser to visualize genomic features
General Feature Format	A tab-separated file format for describing the location of genomic features. The file consists of one line per genomic feature, each containing 9 columns of data (chromosome/scaffold, source, feature type, start, end, score, strand, frame, attribute).
Homologs	Genes that have common ancestral origins
Intron	Parts of the pre-mRNA that are spliced out and not included in the mature mRNA
Nanopore	Long-read or third generation sequencing technology. This technology generates long-reads that provides opportunities to sequence full-length cDNAs
Orthologs	Homologs produced by speciation. These are genes in different species that evolved from a common ancestral gene by speciation. They tend to have similar functions
Paralogs	Homologs produced by gene duplication. These are genes that duplicated within a species and diverged subsequently. They tend to have different functions
Python	A programming language

RNA-Seq	Next-generation sequencing technology, a method that allows us to investigate and discover transcriptome
Transcriptome	The total set of transcripts in a given cell, tissue, or organism
UTR	Portion of mRNA that is not translated. Usually found on the 5’ and 3’ ends of mature mRNA
WormBase	An online resource for gene, genome, and other biological information about *C. elegans* and other related nematodes

Chapter 1. Introduction

Caenorhabditis elegans (*C. elegans*) has been an ideal model organism to address many important biological questions. Much effort has been invested towards annotating the *C. elegans* genome, one approach is performing comparative genomics with *Caenorhabditis briggsae* (*C. briggsae*). However, much less research has been performed in annotating *C. briggsae* itself that limits the benefit of *C. briggsae* as a comparative platform as accurate annotations are essential for comparative studies. We hypothesize that the genome annotation of *C. briggsae* can be improved using RNA-Seq by finding additional *C. briggsae* introns, exons, and protein-coding transcripts. A more complete *C. briggsae* annotation will enhance its utility as a comparative platform. In Chapter 2 of this thesis, we aim to improve the *C. briggsae* genome annotation at the intron and exon level. In Chapter 3, we aim to improve the *C. briggsae* annotation at the transcript level, which we used to discover additional *C. briggsae-C. elegans* ortholog pairs and improve *C. elegans* protein-coding transcript annotation in Chapter 4.

1.1. *Caenorhabditis elegans* as a model organism

Model organisms, such as yeast (*Saccharomyces cerevisiae*), nematode worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), plant (*Arabidopsis thaliana*), and ciliated protozoan (*Tetrahymena thermophila*) are often used to understand human biology because of their small genome size, short generation time, rich genetic resources, and their ease of manipulation in genetic experiments.

In 1963, Sydney Brenner proposed *C. elegans*, a small free-living soil nematode as a suitable model organism for studying the nervous system (Brenner, 1974). *C. elegans* is ideal for genetic study due to its rapid life cycle, simple reproductive cycle, small cell numbers, and transparent body. *C. elegans* can be grown on agar plates or in liquid cultures in approximately 3 days at 25 °C from egg to egg-laying adult (Corsi et al., 2015). The adults are approximately 1 mm in length. The life cycle of *C. elegans* is comprised of the embryonic stage, four larval stages (L1-L4), and adulthood (Altun and Hall, 2009). This species exists primarily as a self-fertilizing hermaphrodite (XX) with a low frequency of male (XO, <0.2%), which allows for homozygous worms to generate

genetically identical progeny (Altun and Hall, 2009; Corsi et al., 2015). The adult hermaphrodite has 959 cells, while the adult male has 1031 cells (Altun and Hall, 2009). This organism is transparent, which makes observation of its cellular structure and biological processes possible by microscopy (Ankeny, 2001). Finally, this organism has a small genome size of 100 Mbp consisting of 6 nuclear chromosomes (5 autosomes and 1 sex chromosome) and a mitochondrion genome (WormBase, 2019). It is the first multicellular organism to have its genome sequenced (*C. elegans* Sequencing Consortium, 1998).

The use of *C. elegans* as a model organism has led to numerous important discoveries. MicroRNAs (miRNAs, small regulatory RNAs) were first found in *C. elegans*. In 1993, the 22nt *lin-4* was found to regulate the timing of *C. elegans* development (Ambros, 2004; Lee et al., 1993). Seven years later, the second miRNA, *let-7*, was also discovered in *C. elegans* (Reinhart et al., 2000). miRNAs have since been found to be produced naturally in cells, control the expression of cellular genes, and are widespread not only in *C. elegans* but also in insects, plants, and mammals (Horvitz, 2003). The study of miRNA in *C. elegans* has helped reveal gene functions and facilitated studies in human cancer research (Poulin et al., 2004). In the 1970s and 1980s, Bob Horvitz and colleagues discovered the mechanisms that regulate programmed cell death (PCD) or apoptosis by studying *C. elegans* cell lineage (Ellis and Horvitz, 1986; Ellis et al., 1991; Sulston and Horvitz, 1977). During worm development, a fully formed worm has 1090 somatic cells, of which 131 cells undergo apoptosis resulting in a 959-celled adult worm. The discovery of apoptosis pathway also revealed that *ced-9* encodes a protein similar to that of the human proto-oncogene *Bcl-2* (Horvitz, 2003). In addition, in 1993, Green Fluorescence Protein (GFP) was first used as a reporter of gene expression and protein localization in *C. elegans* (Chalfie et al., 1993, 1994). GFP is now widely used in cell biology and other biological disciplines. Thus, *C. elegans* has become an ideal model organism to address many important biological questions.

1.2. *Caenorhabditis briggsae* as a comparative tool to improve the understanding of *C. elegans*

Caenorhabditis briggsae (*C. briggsae*) is another small free-living soil nematode and the most extensively studied sister species of *C. elegans* (Hillier et al., 2005). It was first obtained from rich garden soil at Stanford University in 1944 by Margaret Briggs

Gochnauer, and was classified as *Rhabditis* sp. (Gochnauer and McCoy, 1954). The nematode was then described as *Rhabditis briggsae* (Nigon and Dougherty, 1949) and later as *Caenorhabditis briggsae*. Sydney Brenner considered *C. briggsae* as a possible model system for studying the genetic basis of cellular development, but his final choice was *C. elegans* (Ankeny, 2001; Ross et al., 2011).

Although *C. briggsae* and *C. elegans* diverged from their common ancestor ~80-100 million years ago (Stein et al., 2003), *C. briggsae* has many features similar to *C. elegans*, such as being a self-fertilizing hermaphrodite, having a small percentage of males, similar morphology, and life cycle. Both species co-occur in rotting plant material (Félix and Duveau, 2012). The *C. briggsae* genome sequence was published in 2003 (Stein et al., 2003) and was originally sequenced to facilitate *C. elegans* genome annotation (Hillier et al., 2005; Stein et al., 2003). Both *C. elegans* and *C. briggsae* have 6 chromosomes, similar genome sizes, and similar numbers of protein-coding and non protein-coding genes (Gupta et al., 2007; Stein et al., 2003). The *C. briggsae* draft sequence was produced by using a 'hybrid' strategy, which combined a 10X-coverage whole-genome shotgun sequencing (WGS) and physical map sequences (Stein et al., 2003). *C. briggsae* was chosen to be sequenced because it was the closest known species to *C. elegans* that shares hermaphroditic mode of reproduction (Figure 1, Figure 4). Alignments using WABA algorithm (Kent and Zahler, 2000) on the *C. briggsae* and *C. elegans* genomes suggested that *C. briggsae* genome covers 52.3% of the *C. elegans* genome (52.4/100.2 Mbp) and *C. elegans* genome covers 50.1% of the *C. briggsae* genome (52.9/105.6 Mbp) (Stein et al., 2003).

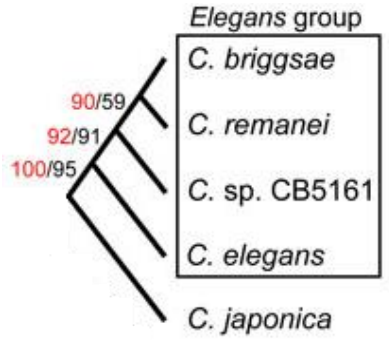

Figure 1. Phylogenetic relationships between *C. briggsae* and *C. elegans.* Adapted from Kiontke, 2005 with permission.

C. briggsae has been used in comparative studies and is an attractive model system to facilitate *C. elegans* research. First, regarding the biology of these two species, as they are nearly identical morphologically, it was presumed that they also share much in cell composition, development, and behaviour (Félix, 2004). A study conducted by Zhao et al. (2008) suggested that the embryonic cell lineage patterns of both species are nearly identical up to the 350-cell stage of embryogenesis. They produce an identical number of progeny and develop with similar timing and cell positions. 113 out of the 671 of the embryonic cells of both species also undergo programmed cell death (Sulston et al., 1983; Zhao et al., 2008). A recent study showed that the early induction and tissue development between the two species are maintained during evolution (Memar et al., 2019). The conservation of embryonic development of these two species could shed light on the evolution of the Rhabditid nematodes (Zhao et al., 2008).

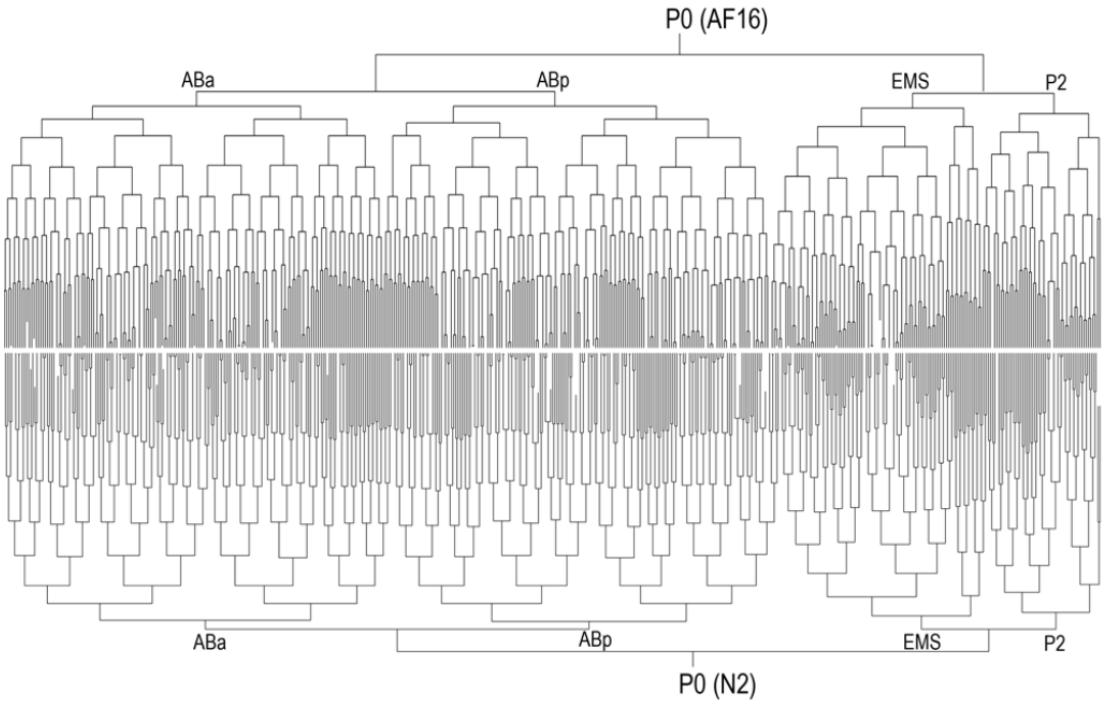

Figure 2. The embryonic cell lineage between *C. briggsae* AF16 (top) and *C. elegans* N2 (bottom) when the embryos contained roughly 450 cells. Reprinted from Zhao et al., 2010 with permission.

Second, on the molecular perspective, the presence of *C. briggsae* genome has resulted in a better *C. elegans* genome annotation, through sequence comparison with its ortholog in *C. briggsae*. Homologs are genes or proteins with similar sequences in two species shared by a common ancestor (Bhasin and Raghava, 2006; Coghlan et al.,

2006; Koonin, 2005). Orthologs are homologs in different species that tend to have similar functions (see Section 1.4). Sixty-two percent of the *C. briggsae* genes (12,155 out of 19,500 genes) have one-to-one orthologs with *C. elegans*. Those orthologs included approximately 60-65% of the *C. elegans* and *C. briggsae* gene sets (Stein et al., 2003). Further study revealed that there are 15,108 ortholog relationships between *C. briggsae* and *C. elegans* (Uyar et al., 2012). Sequences of functional features, such as protein-coding exons and *cis*-regulatory regions, should be conserved among closely related species. For instance, sequence comparison between *C. briggsae dpy-20*-like gene and putative novel *C. elegans dpy-20* gene was able to not only identify and confirm the start codon of that novel *dpy-20* gene, but also identify four homology regions that suggest regulatory elements of the gene (Clark et al., 1995). Comparison of promoter elements of orthologous genes has also led to the identification of shared regulatory elements. For example, screening for potential X-box motifs in promoters of *C. briggsae* genes helped identify 93 (out of 4,291) valid candidates of ciliary genes in *C. elegans*. The candidates were further experimentally validated by subsequent studies to prove that the shared elements are functional (Chen et al., 2006). Another study that performed alignment of upstream sequences of *pha-4* orthologs was successful in finding novel conserved upstream *cis*-regulatory regions that control *C. elegans* temporal pharyngeal development (Figure 3, Gaudet et al., 2004). The upstream regions of *egl-17, zmp-1,* and *cdh-3* orthologs were also found to be similar and correspond to promoting the expression of those genes in vulval cells and the anchor cells of *C. elegans* expression in other studies (Kirouac and Sternberg, 2003). These findings were also experimentally validated by observing their expressions using GFP.

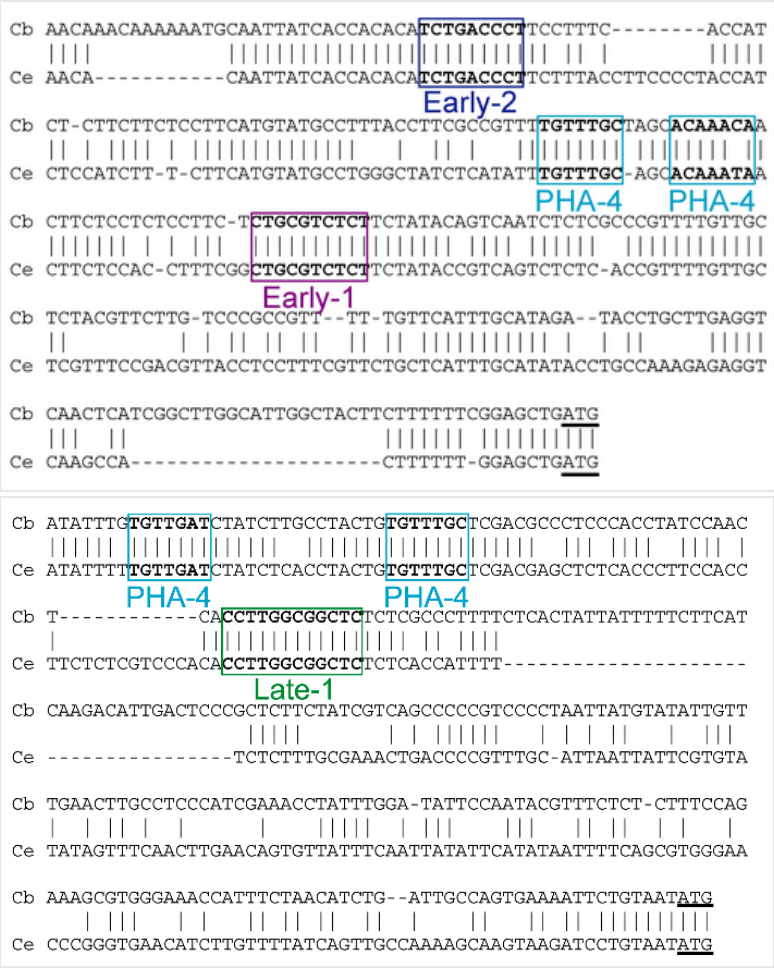

Figure 3. Illustrations of transcription factor binding sites (top: Early-1, Early-2, bottom: Late-1) controlling transcription of *pha-4* during early (top) and late (bottom) stages of *C. elegans* pharyngeal development. Reprinted from Gaudet et al., 2004 with permission.

Third, despite their similar developmental programs, their transition to hermaphroditism evolved independently. In evolutionary biology, different organisms that occupy similar ecological niches often independently evolve similar traits, this process is defined as convergent evolution (Stern, 2013). Only *C. briggsae* and *C. elegans*, out of 11 *Caenorhabditis* nematodes, are hermaphrodite while the rest develop to females and males (Figure 4, Braendle and Félix, 2006). Through pairwise comparisons of orthologous genes, it was found that hermaphroditism has evolved independently in *C. elegans* and *C. briggsae* (Kiontke et al., 2004). *C. briggsae* could therefore facilitate speciation research.

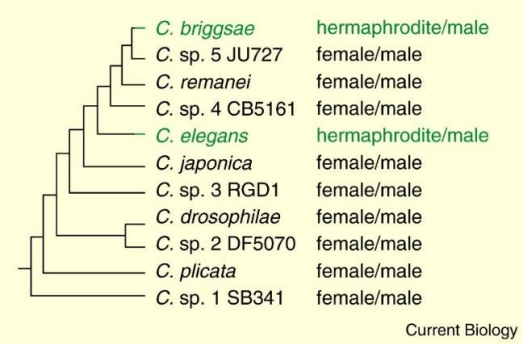

Figure 4. Phylogenetic relationships of *Caenorhabditis* species and their reproductive modes. Reprinted from Félix, 2004 with permission.

Hence, *C. briggsae* can serve as a comparative tool for *C. elegans* and *C. briggsae* annotation can be beneficial in improving the *C. elegans* annotation.

1.3. *C. briggsae* genome annotation effort and status

The *C. elegans* genome has been extensively annotated through combined approaches ranging from bioinformatics *ab initio* gene prediction to experiment-based transcriptomics and proteomics. The *C. elegans* genome annotation was further refined using the *C. briggsae* genome annotation. In details, the efforts include the use of computational *ab initio* gene finding tool Genefinder (GreenP, unpublished), homology-based gene prediction (GeneWise, ORFgene2), and experimental methods such as expressed sequence tags/ESTs (Kohara, 1996; Shin et al., 2008), open reading frame sequence tags/OSTs (Lamesch et al., 2004; Reboul et al., 2003; Wei et al., 2005), serial analysis of gene expression/SAGE (Nesbitt et al., 2010; Ruzanov and Riddle, 2010; Ruzanov et al., 2007), rapid amplification of complementary DNA ends/RACE (Salehi-Ashtiani et al., 2009), trans-spliced exon coupled RNA end determination/TEC-RED (Hwang et al., 2004), RNA-Sequencing/RNA-Seq (Allen et al., 2011; Boeck et al., 2016; Douglas, 2018; Gerstein et al., 2010; Hillier et al., 2009; Tourasse et al., 2017), and translational expression evidence (Shim and Paik, 2010).

In contrast, the *C. briggsae* genome annotations have been limited to mostly bioinformatics studies with a few transcriptome studies. The *C. briggsae* annotation efforts include computational *ab initio* (Genefinder and FGENESH), sequence conservation gene finding (TWINSCAN), and experimental ESTs and protein-based comparisons (Ensembl annotation pipeline) that were done as a part of the publication of

the genome (Stein et al., 2003). The genome was then re-annotated by the nematode genome annotation assessment project or nGASP project (Coghlan et al., 2008), revised based on their homology to *C. elegans* genes using genBlastG (She et al., 2011), and further refined using RNA-Sequencing data (Uyar et al., 2012).

Despite the similar number of protein-coding genes between these two organisms, *C. elegans* has two or threefold more annotated protein-coding transcripts, introns, exons, and spliced leader (SL) trans-splicing acceptor sites than *C. briggsae*, according to current annotations (Table 1).

Table 1. Comparison of the *C. briggsae* and *C. elegans* genome annotations

	C. briggsae	***C. elegans***
Protein-coding genes	21,827[a]	20,359[c]
Protein-coding transcripts	21,863[a]	72,274[d]
Introns	107,848[a]	239,333[d]
Exons	121,849[a]	328,212[d]
SL trans-splicing acceptor sites (*genes*)	11,617 (*8,555*)[b]	28,249 (*11,387*)[e]

Sources: [a]Wormbase release WS254; [b]Uyar et al., 2012; [c]Wormbase release WS250; [d]Douglas, 2018; [e]Allen et al., 2011

1.4. Genome annotation and comparative genomics

Genome annotation is the process of finding functional elements in a genome of interest (Armstrong et al., 2019). Genome annotation can be classified into three levels: the nucleotide, protein, and process level. At the nucleotide level, it involves identifying genes and their intron-exon structures using a combination of *ab initio* and homology-based computational pipelines along with the use of experimental pipelines to validate the predictions (Stein, 2001). This is also called structural annotation. Accurate structural genome annotation is important because it may affect the quality of downstream analysis (König et al., 2018). At the protein level, it involves assigning functions to the products of the genome. This process uses databases of protein sequences and, possibly, functional domains and motifs (Stein, 2001). At the process level, it involves assigning gene ontology terms to the structural annotations in the context of cellular and organismal physiology (König et al., 2018; Stein, 2001).

Comparative genomics is one of the approaches used in distinguishing and assigning roles to functional DNAs as functional sequences are subject to evolutionary selection. This approach uses information obtained from one genome to make inferences about any information, such as the map positions and functions of genes in a second genome (Brown, 2006).

Having a high-quality structural annotation is important for comparative genomics. Protein-coding exons and *cis*-acting sequences regulating expression are usually highly conserved (i.e., under purifying selection), while the untranslated regions are less well conserved. A higher quality genome annotation is therefore can be beneficial not only for the species itself, but also for other species. Defective annotations may lead to incorrect predictions in comparative studies between genomes of interest (for example, in protein family evolution study). When a gene model in one species is inaccurate or can be improved, the corrected annotation can increase the quality of the annotation of its close relatives as well. Observing gene models that were previously unobserved can also be beneficial for both species. Thus, a more complete and accurate genome representation can benefit comparative genomic analyses.

The identification of homologous genes, specifically orthologous genes, is an initial pivotal step in comparative genomics. As mentioned in Section 1.2, homologs are genes or proteins with similar sequences that have common ancestral origins. Orthologs and paralogs are types of homologs. Orthologs are genes in different species that evolved from a common ancestral gene by speciation. These genes tend to have similar functions although they sometimes have functionally different functions due to gene fusions or protein domain rearrangements (Koonin, 2005). On the other hand, paralogs are genes that originated by duplication within a genome followed by a subsequent divergence. Paralogs tend to have different functions. Detection of orthologs is crucial for reliable functional annotation and evolutionary analyses of genes and species (Tekaia, 2016).

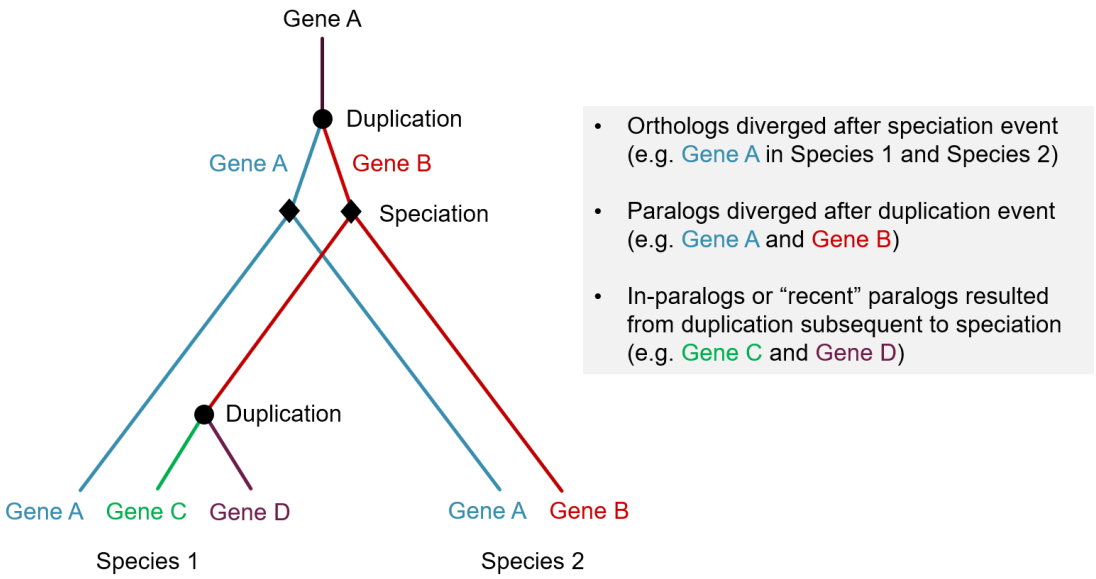

Figure 5. Illustration of orthologs and paralogs.

Ortholog prediction methods are classified into two categories, tree-based methods that infer orthologous relationships using phylogenetic trees and network/graph-based methods that rely on pairwise sequence similarities computed across all sequences involved to predict orthologs (Kuzniar et al., 2008).

Tree-based methods involve detecting putative homologs, performing multiple-sequence alignment, constructing phylogenetic tree(s), and evaluating the ortholog relationships against a reference species tree. These methods are computationally intensive for large datasets, not easily automated due to the need to choose appropriate outgroup species, and depend on the pre-defined protein families (Kuzniar et al., 2008).

Graph-based methods rely on the assumption that orthologous genes or proteins are reciprocally most similar to each other than to any other genes or proteins from their respective genomes. In terms of speed, a heuristic algorithm BLAST (Basic Local Alignment Search Tool) is the most commonly used method than an exhaustive process using the dynamic programming algorithm (Smith-Waterman) that could be time consuming. In general, BLAST is applied twice on two sets of protein sequences from two genomes, with each set used as query and subject to find the reciprocal best BLAST hits (RBBH or RBH). Specifically, BLAST is performed on all proteins from genome A (query) against protein database from genome B (subject), and second BLAST is performed using all proteins from genome B (query) against protein database from

genome A (subject). The top BLAST hits are recorded for each protein in genomes A and B, and a reciprocal best BLAST hit is found when the proteins encoded by two genes find each other as the best scoring match in the other genome. Those pairs of genes that are reciprocally most similar to each other are predicted to be orthologs. Some tools also incorporate clustering techniques to cluster ortholog groups. Graph-based approaches are computationally less-intensive and more efficient for large datasets than tree-based approaches (Kuzniar et al., 2008; Li et al., 2003).

In this study, we use OrthoMCL (Fischer et al., 2011; Li et al., 2003) to identify ortholog relationships between *C. elegans* and *C. briggsae*. This tool was demonstrated to be one of the two best performing ortholog detection tools for eukaryote genomes including in *C. elegans* (Chen et al., 2007). OrthoMCL uses the reciprocal best BLAST hit approach and makes an adjustment for species distance (normalization) to distinguish orthologs from in-paralogs (i.e., “recent” paralogs resulted from a lineage-specific duplication subsequent to speciation. They are likely to have similar functions within species; Figure 5). This tool integrates a Markov Cluster algorithm (MCL) to cluster proteins into ortholog groups using the normalized BLAST scores between the corresponding proteins (Li et al., 2003). InParanoid, is another graph-based tool that also uses reciprocal best BLAST hit and accommodates in-paralog but does not incorporate the additional clustering step like OrthoMCL. Ortholuge is similar to InParanoid but uses phylogenetic distance ratios instead of BLAST similarities. InParanoid and Ortholuge are limited to two organisms, whereas OrthoMCL can be applied for two or more organisms (Kuzniar et al., 2008).

1.5. Alternative splicing, coding capacity, and organism complexity

Unlike bacterial genes, the precursor messenger RNA (pre-mRNA) from most eukaryotic genes undergo post-transcriptional modifications before it becomes a mature messenger RNA (mRNA) that can be translated into proteins. Post-transcriptional modifications occur in the nucleus and include 5’ capping, RNA splicing, and 3’ polyadenylation (Figure 6, Bhagavan and Ha, 2015).

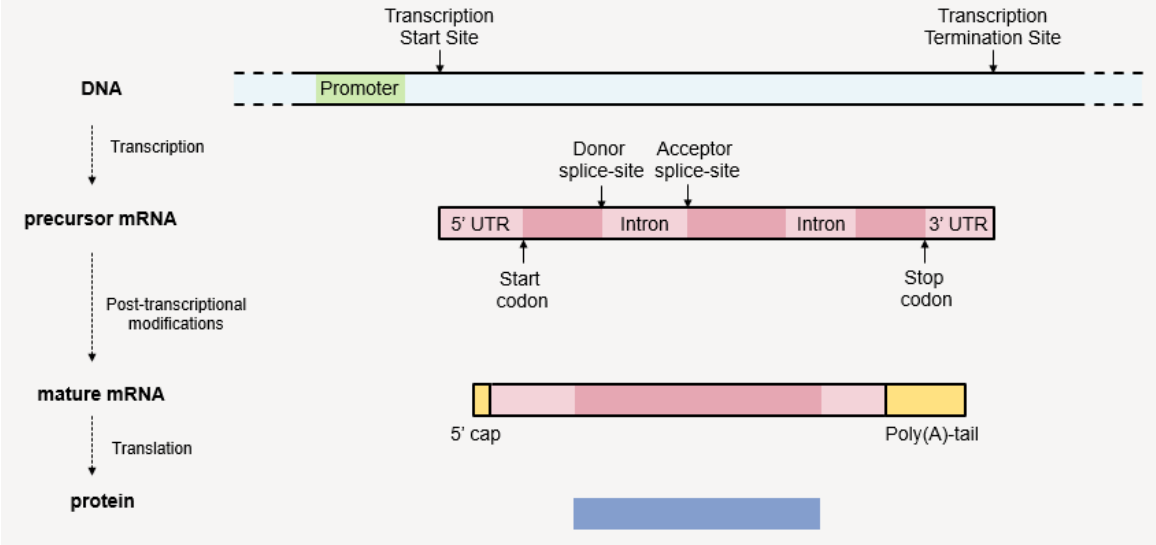

Figure 6. Illustration of transcription, post-transcriptional modifications, and translation in eukaryotes. Dark pink boxes denote coding exons, dark blue box denotes protein.

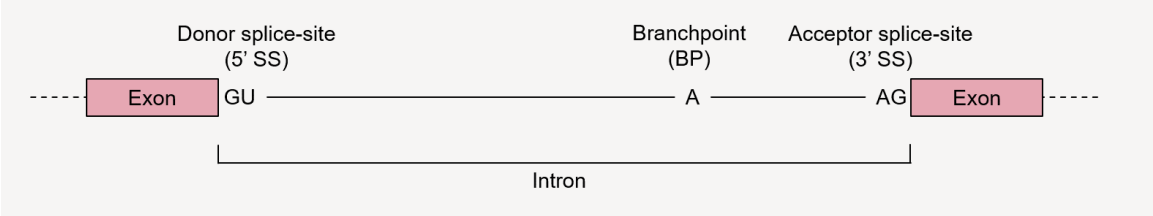

Figure 7. Illustration of general sequence features of pre-mRNA that undergoes splicing in eukaryotes. The splicing signals in most pre-mRNAs are consensus GU in 5' splice site (exon/intron boundary), A in branchpoint close to the 3' splice site (intron/exon boundary), and AG in 3' splice site (intron/exon boundary).

RNA splicing is catalyzed by the splicing machinery (spliceosome), a complex of hundreds of interacting proteins and small nuclear RNAs (snRNAs) including the five small nuclear ribonucleoproteins (snRNPs) U1, U2, U4, U5, and U6. To initiate splicing, splicing signals were detected that include the consensus GU in 5' splice site, A in branchpoint close to the 3' splice site, and AG in 3' splice site. The branchpoint sequence is nearly invariant in yeast (UACUAAC) and more variable in higher eukaryotes (YNCURAC). Y stands for either pyrimidine (U or C), N stands for any nucleotide, and R stands for either purine (A or G). In all cases, the branchpoint sequence contains an Adenine (Weaver, 2012). Additionally, *C. elegans* introns have an extended, very highly conserved 3' splice site consensus sequence, UUUUCAG, whereas introns from most mammalians have polypyrimidine (Y) tract between branchpoint and 3' splice site AG. Polypyrimidine tract between the branchpoint and 3' splice sites is also absent from yeast introns (Krämer, 1996). Furthermore, the yeast (*S.*

cerevisiae) genome only has 295 introns from 280 genes (~5%), significantly lower than most eukaryotic genomes (Parenteau et al., 2019).

The spliceosome assembly is initiated by the recognition and base pairing of 5' splice site by the U1 snRNP and 3' splice site by the U2 snRNP auxiliary factor (U2AF), generating the E (early) complex. The U2 snRNP is then recruited to the branchpoint Adenosine to generate the pre-splicing complex A that also involves polypyrimidine tract in higher eukaryotes. This and the subsequent steps are ATP-dependent mechanisms. The U4-U6-U5 tri-snRNP joins complex A to generate complex B, which is then converted into the catalytically active complex C. The first transesterification reaction occurs and U1 and U4 snRNPs are released. In the first transesterification reaction, the 2'-hydroxyl group of Adenine in branchpoint attacks and breaks the phosphodiester bond linking the first exon to the 5' splice site, yielding a 'free' first exon and the lariat-shaped intron-second exon intermediate. The second transesterification reaction follows from the 'free' 3-hydroxyl group on the first exon that attacks the phosphodiester bond between the 3' splice site and the second exon. This process generates spliced exon-exon product (i.e., the final mature mRNA) and releases the lariat-shaped intron (Figure 8, Krämer, 1996; Weaver, 2012). This type of splicing occurs within the same pre-mRNA molecule and therefore is called *cis*-splicing (Figure 8, Figure 9A).

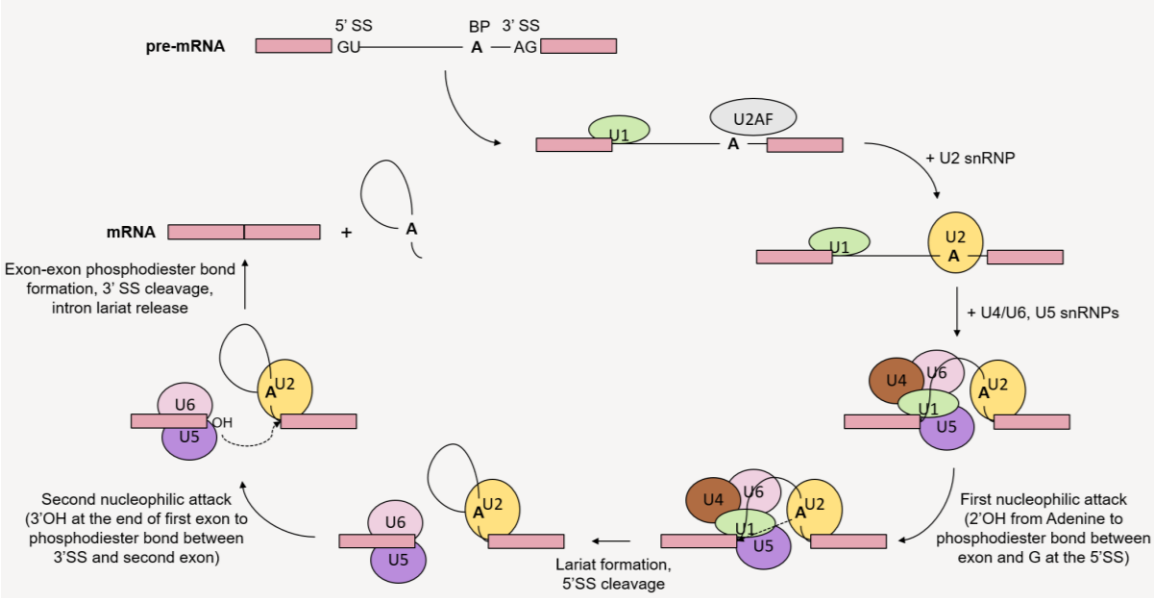

Figure 8. The molecular mechanism of pre-mRNA splicing that is catalyzed by the spliceosome, an assembly of snRNPs (U1, U2, U4, U5, U6) and interacting proteins (most of them are not shown). The spliceosome recognizes the splicing signals in the pre-mRNA molecule, catalyzes the two-step transesterification reaction, and joins two exons together.

In addition to *cis*-splicing, *trans*-splicing occurs in some eukaryotes (Lei et al., 2016). While *cis*-splicing occurs within the same pre-mRNA, *trans*-splicing occurs between two different mRNA molecules (Figure 9B). *Trans*-splicing plays important roles in many physiological and pathological processes, although it occurs at a low frequency in humans. Both *cis*- and *trans*- splicing occurs in nematodes (Lei et al., 2016). One specific type of *trans*-splicing, spliced leader (SL) *trans*-splicing, is when a 22-nucleotide SL sequence donated by a 100-nucleotide SL RNA, is *trans*-spliced to the 3' splice site of a pre-mRNA molecule. This process replaces the outron, which is the 5' end of pre-mRNA intron-like region. The same pre-mRNAs that receive the spliced leader is also processed by conventional *cis*-splicing. Both processes are catalyzed by the spliceosome. SL1 and SL2 are the two types of SL sequences. SL1 is *trans*-spliced to non-operon genes and to first genes in operons, while SL2 is *trans*-spliced to downstream genes in operons. Operons are gene clusters that can be up to eight genes long and are transcribed into polycistronic pre-mRNAs controlled by a single promoter (Allen et al., 2011; Spieth et al., 1993). In this thesis project, I will focus on analyzing *cis*-splicing in *C. briggsae*.

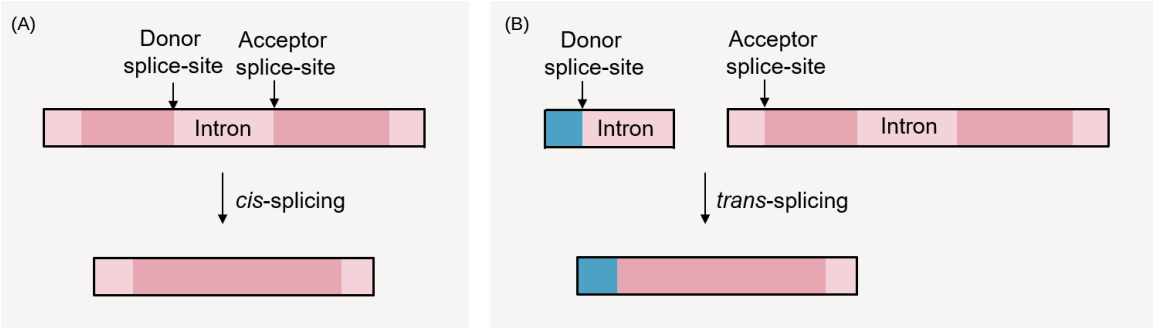

Figure 9. Illustrations of (A) *cis*- and (B) *trans*-splicing.

In eukaryotes (including nematodes), instead of having a single splicing pathway for each pre-mRNA, genes can follow alternative *cis*-splicing pathways to process pre-mRNAs into two or more mature transcripts that encode different proteins. This process is called alternative splicing (Figure 10) and is an essential cellular process for regulating the transcriptome and plays a central role in cellular homeostasis (Gamazon, 2016; Wang et al., 2009). Several diseases, such as cystic fibrosis, have been linked with mutations or variations that lead to aberrant splicing and abnormal protein production (Garcia-Blanco et al., 2004). Compared to nematodes, there are only a few examples of alternative splicing in yeast (Parenteau et al., 2008).

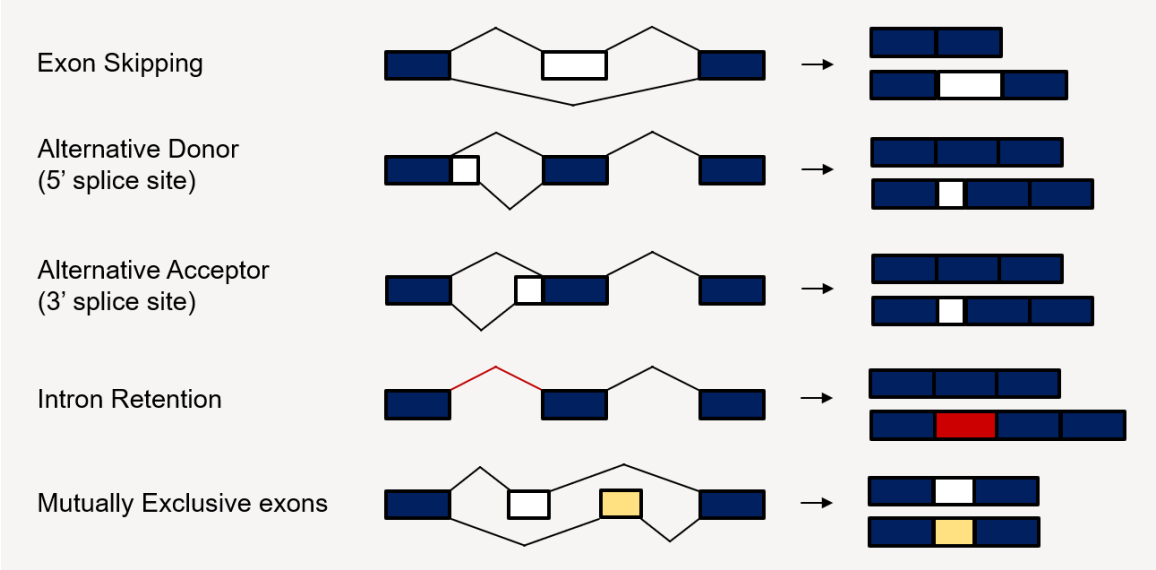

Figure 10. Several types of alternative splicing. Boxes denote exons, black and red lines denote introns.

One of the mechanisms that regulate alternative splicing is the interactions between RNA-binding proteins and *cis*-regulatory sequences (found in the exons and introns). Exons can contain sequences known as exonic splicing enhancers (ESEs), which stimulate splicing, and exonic splicing silencers (ESSs), which inhibit splicing. Intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs) are also found in introns. Serine and arginine-rich proteins (SR proteins) and heterogeneous nuclear ribonucleoproteins (hnRNP proteins) are RNA-binding proteins. SR proteins tend to bind to ESEs, while hnRNP proteins, such as hnRNP A1, bind to ESSs and intronic silencing elements. Such bindings can lead to splicing activation or repression at nearby splice sites leading to alternative splicing (Weaver, 2012).

Alternative splicing is prevalent in metazoan genomes. It was estimated that at least 42% of *Drosophila* genes (Stolc et al., 2004) and over two-thirds of mouse and human genes (Johnson et al., 2003) encode alternatively spliced pre-mRNAs. These numbers have been increasing at a brisk pace over the past several years and are likely to still be underestimates because many low abundance, tissue- or stage- specific isoforms likely remain to be characterized (Park and Graveley, 2007). For instance, 58% of genes in *Drosophila* are estimated to encode multiple transcript isoforms, ~16% more than a decade ago (Brown et al., 2014) and 92-94% human genes undergo alternative splicing, ~30% more identified in 5 years (Wang et al., 2008). As the annotation and our understanding of the entire repertoire of mRNAs expressed by the genome improves,

the number of genes that express multiple isoforms and the number of isoforms per gene increases (Blencowe and Graveley, 2008). Furthermore, some transcripts are rarely expressed. More transcripts will likely be detected and their quantification will be more precise with deeper sequencing (Mortazavi et al., 2008).

Organismal complexity is difficult to define or measure. The number of distinct cell types in an organism can be used to define complexity (Carroll, 2001; McShea, 1996; Valentine et al., 1994). Morphological features can also be used as a measure of complexity, for example, change in the number of limb-pair types in free-living aquatic anthropods (Cisne, 1974; McShea, 1996). It was also suggested that physiological differences due to changing environments could also be used to approximately measure organism complexity (Adami et al., 2000). The total amount of DNA contained in a single (haploid) set of its chromosome in the genome (C-value) was also used to measure organism complexity. However, it was found that the physical size of genomes are unrelated to organism complexity (Cavalier-Smith, 1978).

Alternative splicing of protein-coding genes was proposed to explain organismal complexity (Chen et al., 2014). Because alternative splicing generates multiple distinct transcripts from a single gene, it has the potential to increase the total number of coding-transcripts encoded in a genome in the absence of increases in gene number. While the number of protein-coding genes may not necessarily reflect the perceived complexity of those organisms, the number of protein-coding transcripts may (Table 2, Brown, 2006; Elliott, 2011).

Table 2. Protein-coding genes and transcripts in various eukaryotes

Organism	Protein-coding genes	Protein-coding transcripts
Nematode (*C. elegans*) [a]	20,359	31,574
Fruit fly (*D. melanogaster*) [b]	13,947	34,920
Mouse (*M. musculus*) [c]	21,856	59,252
Human (*H. sapiens*) [d]	19,957	84,107

Sources: [a]Wormbase release WS250; [b]Ensembl version 99 (BDGP6.28); [c]GENCODE version M24; [d]GENCODE version 33

The coding capacity of a protein-coding gene is defined as the complete set of protein-coding transcripts from all combinations of exons of the gene (Douglas, 2018). At the genome level, the coding capacity is the complete set of all transcripts derived from

all protein-coding genes (Abdel-Ghany et al., 2016; Douglas, 2018). Because alternative splicing contributes to the number of protein-coding transcripts, it is therefore essential to define the complete set of protein-coding transcripts of a genome (i.e., coding capacity) to understand an organism's complexity. If the coding capacity of a genome has been fully defined, then the genome annotation of that genome is complete.

Various technologies have been developed to annotate transcripts, including hybridization-based and sequence-based approaches. Hybridization-based approaches that include several types of microarrays are high-throughput and relatively inexpensive except for high-resolution tilling arrays to study large genomes. However, these methods require previous knowledge of the genome sequence of study. On the other hand, sequence-based approaches can directly determine the cDNA sequence. The effort started with Sanger Sequencing of cDNA or expressed sequence tag (EST) libraries (Boguski et al., 1993, 1994; Gerhard et al., 2004). This method is relatively low throughput, expensive, and generally not quantitative. Tag-based methods including serial analysis of gene expression or SAGE (Harbers and Carninci, 2005; Velculescu et al., 1995) and cap analysis of gene expression or CAGE (Kodzius et al., 2006; Shiraki et al., 2003) were then developed. These methods are high throughput and can provide expression levels, but they are expensive and only a portion of transcript are analyzed. Isoforms are also generally indistinguishable from each other. Finally, the development of high-throughput transcriptome sequencing, RNA Sequencing (RNA-Seq) has provided a method to map and quantify transcripts (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2009).

RNA-Seq uses massively parallel sequencing to allow transcriptome analyses of genomes at a far higher resolution than is available with Sanger Sequencing and microarray-based methods (Nagalakshmi et al., 2010). This method starts with the generation of cDNA libraries from the RNAs of interest with adapters attached on one or both ends and are directly sequenced using high-throughput next-generation sequencing technologies to obtain short sequences from one end (single-end sequencing) or both ends (paired-end sequencing). Following sequencing, the reads obtained are aligned to a reference genome or assembled *de novo* without the genome sequence, to construct a whole-genome transcriptome map that contains the transcriptional structure and/or level of expression for each gene (Nagalakshmi et al., 2010; Wang et al., 2009). RNA-Seq offers several key advantages over tilling microarray and EST sequencing. For instance,

it is not limited to detecting transcripts that correspond to existing genomic sequence. Moreover, RNA-Seq reads do not contain introns, and will be split at the splicing junctions during alignment, providing evidence for exons. Furthermore, compared to DNA microarray, RNA-Seq has a very low background signal because DNA sequences can be mapped to genomic regions directly. In addition, RNA-Seq could identify genes expressed at low or very high levels depending on the number of sequences on hand. However, this method does have challenges regarding library construction (fragmentation step that could lead to bias in the outcome), bioinformatics analysis (efficient pipeline development, error base-calling, low-quality reads, full-length transcript recovery), and coverage (the more sequencing depth required for adequate coverage, the higher the sequencing cost).

1.6. Thesis aims

Here, we are using RNA-Seq to facilitate *C. briggsae* genome annotation with the goal of improving the utility of *C. briggsae* as a comparative platform for *C. elegans*.

Unlike *C. elegans* that has been extensively annotated using computational and experimental approaches, the experimental annotations in *C. briggsae* are limited. *C. elegans* has two or threefold more annotated genomic features than *C. briggsae*, according to current annotations. Our hypothesis is that the *C. briggsae* genome annotation is currently incomplete, hence the difference in numbers of annotated elements between the two species, such as introns, exons, and protein-coding transcripts. By pooling RNA-Seq libraries, from both publicly available datasets and datasets generated by our lab, we expect to generate a more complete *C. briggsae* genome annotation, by finding additional *C. briggsae* introns, exons, and protein-coding transcripts.

With a more complete *C. briggsae* genome annotation, comparative analyses between the two species will be performed. We hypothesize that with the more complete annotation, we will not only identify additional orthologous relationships between *C. briggsae* and *C. elegans* that were previously missed, but also improve the *C. elegans* genome annotation.

This thesis is organized as follows. Chapter 2 describes our bioinformatics pipeline and results of identifying RNA-Seq introns and exons and improving *C. briggsae* genome annotation at the intron and exon level. Chapter 3 presents our bioinformatics pipeline and results of assembling RNA-Seq transcripts and evaluating them to improve *C. briggsae* genome annotation at the transcript level. Chapter 4 describes the approaches and results of comparative analysis between *C. briggsae* and *C. elegans*. And finally, Chapter 5 presents our conclusion and future directions. Bioinformatics tools used in this thesis are listed in Appendix A (Table 13).

Chapter 2. Improving *C. briggsae* intron and exon databases

2.1. Introduction

In this section, we annotated (i.e., built more complete) *C. briggsae* intron and exon databases using RNA-Seq data and WormBase WS254 annotated introns and exons. Firstly, RNA-Seq libraries were pre-processed, which include the removal of ribosomal RNA (rRNA) reads, adapters, and low-quality bases. Pre-processed reads were aligned to the *C. briggsae* reference genome WS254, and the multi-mapped reads were further filtered. Introns were defined and intron threshold was applied to generate a high-quality RNA-Seq intron database. Exons were reconstructed using RNA-Seq introns defined and *C. briggsae* reference genome. The coding capacity of *C. briggsae* was then evaluated using introns and exons. When the coding capacity could still be improved (i.e., not all RNA-Seq introns and exons are represented in the annotated protein-coding transcripts), the RNA-Seq-specific introns and exons were integrated into WormBase databases generating integrated intron and exon databases. These databases served as the more complete *C. briggsae* annotation at the intron and exon level and used in the next chapter to generate an improved *C. briggsae* protein-coding transcript set.

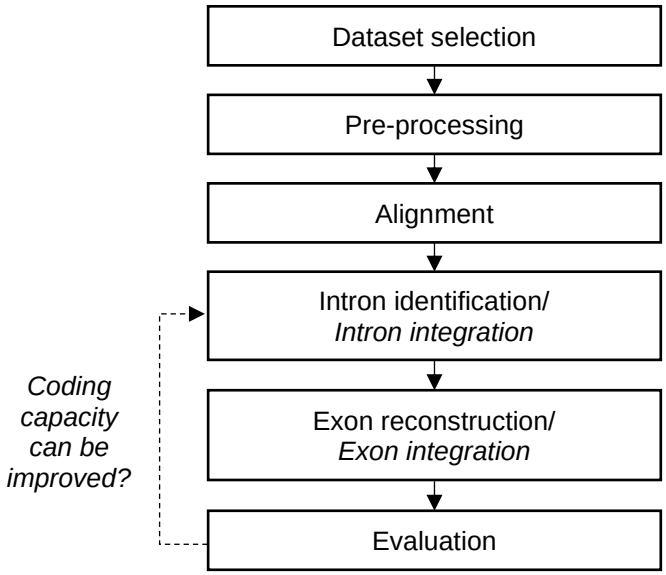

Figure 11. Workflow for building (*integrated*) intron and exon databases.

2.2. Data set selection and pre-processing

2.2.1. Data set selection

RNA-Seq libraries are from publicly available libraries stored in the Sequence Read Archive (SRA) in NCBI database and in-house generated libraries. A pool of 60 libraries from NCBI was obtained by using the SRA Advanced search builder with query "("caenorhabditis briggsae"[Organism]) AND "transcriptomic"[Source]". Results were sent to SRA Run Selector, and "rna-seq" was applied on the Assay Type option. Out of those, 47 were single-end and 11 were paired-end. Raw reads from the 11 paired-end libraries in FASTQ format were selected and downloaded using "fastq-dump" from the SRA Toolkit version 2.8.2 (http://ncbi.github.io/sra-tools/fastq-dump.html). After including the 2 in-house libraries, 13 paired-end RNA-Seq libraries, consisting of 174 million read pairs (29.5 Gigabasepairs) were used for downstream analyses (Table 3).

Table 3. RNA-Seq libraries selected from SRA NCBI and from our laboratory

Library ID	Run ID	Read Pairs	Total Length	Developmental Stage
SRX392707	SRR1050782	29,058,339	152	Embryo
SRX392708	SRR1050783	38,526,175	152	Embryo
SRX392710	SRR1050785	7,757,872	152	L1
SRX392711	SRR1050786	2,105,586	152	L1
SRX392716	SRR1050791	10,120,026	152	Young Adult
SRX392718	SRR1050793	12,719,780	152	Mixed-stage
SRX1500344	SRR3052000	16,602,099	263	Young Adult
SRX1500345	SRR3052001	15,370,016	262	Young Adult
SRX1500346	SRR3052002	6,380,997	266	Young Adult
SRX127748	SRR440441	6,495,173	84	L1
SRX127749	SRR440557	5,054,921	84	Mixed-stage
in-house	inhouse_L1	11,441,628	202	L1
in-house	inhouse_Mixed	12,643,306	202	Mixed-stage

2.2.2. Ribosomal RNA reads removal

Typically, wet-lab protocols for extracting mRNA for RNA-Seq include a step to deplete rRNA in the sample, however some rRNA carryover is commonly observed. We

applied rRNA filtering because rRNA reads in these libraries will produce valid alignments (Delhomme et al., 2014). The program BBDuk (BBMap/BBTools version 37.36, https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide) was used to perform this filtering. Sequences of the rRNA genes were given as an input, and BBDuk detects the given rRNA sequences as a type of contaminant and removes them. In total, 1,115,942 read pairs (0.64%) were filtered out in this step.

2.2.3. Adapters and low-quality reads removal

A quality control step is commonly performed to observe the overall quality of the reads. Reads are scanned for low confidence bases, biased nucleotide composition, adapters, duplicates, etc. This step helps guide the preprocessing decisions (Korpelainen et al., 2014). FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to perform quality check. FastQC results of 13 libraries showed that libraries contain low-quality base calling scores (Figure 12, top) and adapters (Figure 12, bottom) that need to be trimmed.

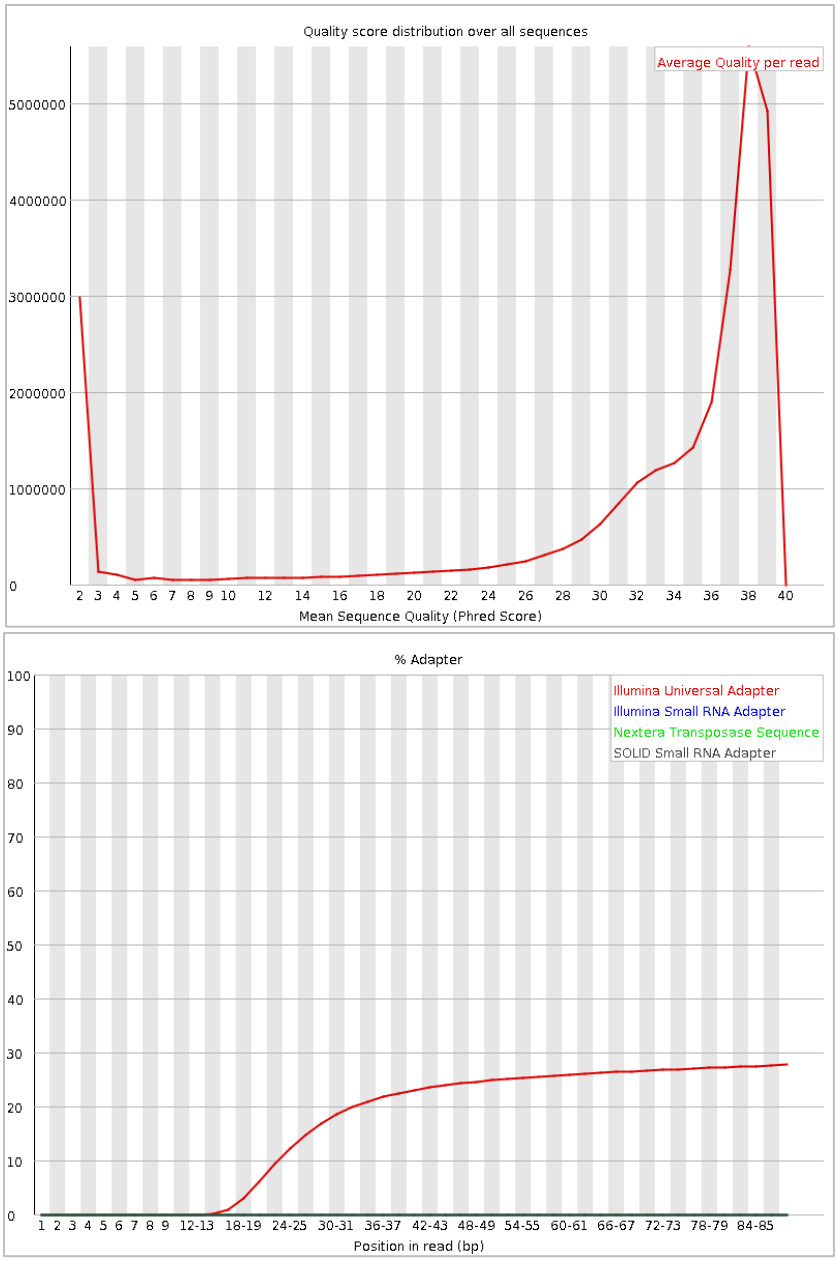

Figure 12. Representative FastQC results from in-house L1 library. (top) Quality score distribution over all sequences showing peaks at Phred quality score 2 and score 37 and (bottom) Adapter content graph showing Illumina Universal Adapter content.

Trimmomatic version 0.36 (Bolger et al., 2014) was used to trim low-quality bases and adapter. Minimum read length was set to 50 bp, with the exception for SRR440441 and SRR440557 whose read lengths are 42 bp to begin with (minimum read length was set to 21 bp). The minimum base quality score for trimming bases at the beginning (LEADING) or end (TRAILING) was set to 5, and the average base quality score within 4-base window (SLIDINGWINDOW) was set to 7.5. Setting the thresholds

too low may retain low-quality reads, however setting it too high can negatively impact intron detection. Parameters were set to minimize the number of artifacts introduced by low-quality reads, while maintaining sensitivity of intron detection. Base quality threshold of LEADING/TRAILING:5 and SLIDINGWINDOW:4:7.5 were sufficient to maintain the sensitivity of intron detection (Douglas, 2018, Appendix B1). 54,471,950 read pairs were removed and 118,687,926 read pairs with both reads survived the quality filtering were used for the next step.

2.3. Alignment of pre-processed reads to the *C. briggsae* genome

Pre-processed reads were aligned against the *C. briggsae* WS254 reference genome using the splice-aware alignment program Spliced Transcripts Alignment to a Reference or STAR version 2.5.3a (Dobin et al., 2013), generating 124,319,101 alignments. STAR was found to have the lowest false positive rate for intron identification (10%) compared to the other splice-aware alignment programs TopHat2 version 2.1.0 (65%) and HISAT2 version 2.1.0 (20%) (Douglas, 2018).

Additionally, because sequence similarity of tandem duplications introduces multiple ambiguous alignment of read pairs, multi-mapped reads were filtered out to reduce false positives of intron identification in downstream analysis. When the multimappings are of the same length, one alignment is randomly selected; when multimappings are of different lengths, the alignment with the shortest distance between reads is kept (Douglas, 2018). Out of 124,319,101 alignments, 12,458,519 (10.02%) and 424,946 (0.34%) alignments were filtered out by random selection and shortest alignment selection, respectively, keeping 111,435,636 alignments for the next step.

2.4. Building RNA-Seq intron and exon databases

2.4.1. Intron identification and filtration to generate a high-quality intron database

Since mature RNAs from RNA-Seq do not contain introns in their sequences, the alignment of RNA-Seq reads to a reference genome sequence will result in the splitting

of reads at the splicing junctions (Figure 13). In the alignment result, CIGAR string ‘N’ is described as the skipped region from the reference that represents an intron.

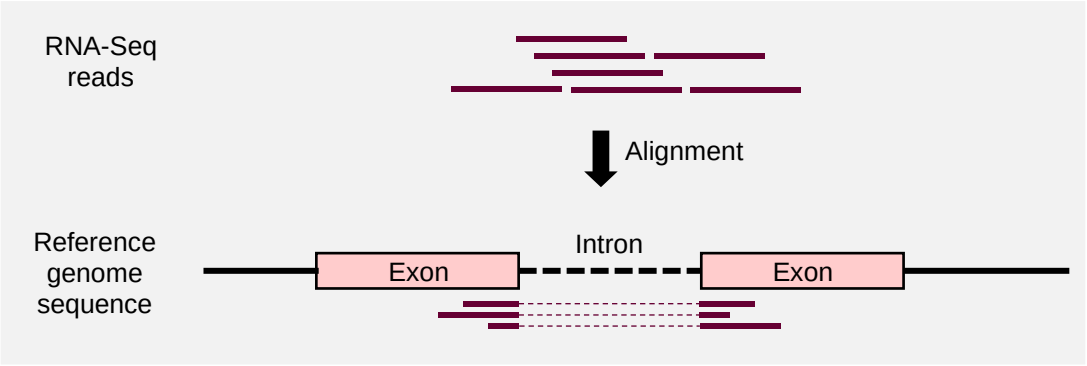

Figure 13. Splitting of intron-less RNA-Seq reads across splice junctions during alignment against a reference genome defining an intron.

During the alignment using STAR, the minimum and maximum intron length parameters were set to 30 and 5000 bp. Using this range of intron length, STAR captured 99.99% of WormBase introns (62 out of 103,314 introns were not captured, Appendix B2) and captured nearly all of RNA-Seq introns (Figure 14). We also calculated the number of reads supporting each intron and assign those numbers to the defined RNA-Seq introns (hereafter, read support).

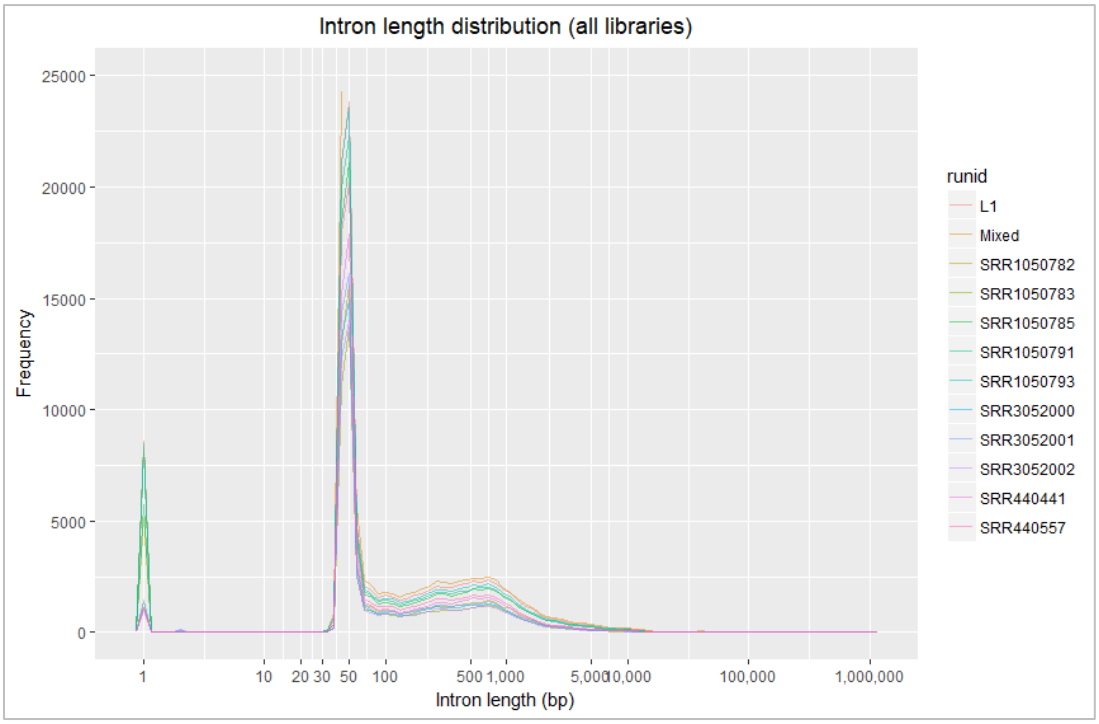

Figure 14. Intron length distribution in 13 *C. briggsae* RNA-Seq libraries.

To build a high-quality intron set, a minimum intron threshold of read support was set to identify true introns (true positives) while keeping the number of false positives low to allow novel introns to be detected. Diagnostic tests including sensitivity and specificity on various minimum intron thresholds were performed using randomly selected 50 RNA-Seq introns from each minimum intron thresholds and WormBase introns. Sensitivity was calculated as the proportion of WormBase introns that were correctly identified by RNA-Seq [TP/(TP+FN)], and sensitivity was calculated as the proportion of WormBase introns that were correctly not identified by RNA-Seq [TN/(TN+FP)]. Based on a previous study (Douglas, 2018) and the diagnostic tests, a minimum intron threshold of five reads in at least one library was found to compromise between minimizing spurious introns while retaining the majority of WormBase introns (Figure 15, Appendix B3-4).

Figure 15. Introns detected in *Cbr-let-2* using various minimum intron thresholds. GBrowse track: (1) WormBase WS254 gene models, (2-4) Intron database when applying threshold 1, 2, and 5, respectively. Lower thresholds for intron support contain many spurious introns.

After applying several pre-processing and filtration steps on the selected 13 RNA-Seq libraries, an RNA-Seq intron dataset in General Feature Format (GFF3) composed of 95,632 introns was generated. The RNA-Seq intron dataset was uploaded to MySQL database (i.e., intron database) for visualization. All visualizations of genomic features in this thesis were performed using Generic Genome Browser or GBrowse (Stein et al., 2002).

2.4.2. Exon reconstruction

Using the introns defined in the previous section and 6-frame translation blocks from *C. briggsae* WS254 reference genome, we reconstructed RNA-Seq exons using ExonTrap (https://github.com/mattdoug604/exon_trap; Douglas, 2018). Briefly, ExonTrap uses translation blocks of the six reading frames of the genome and the intron boundaries to build exons. This tool does not reconstruct exons from single-exon genes. ExonTrap categorizes exons into 2 categories, internal exons if the exons bounded by splice sites at both ends are located within the transcript and terminal exons if located in the 5' end or 3' end of the transcript that are bounded by splice site and a start codon (ATG) or a stop codon (TAG, TAA, or TGA), respectively. RNA-Seq exon dataset (GFF3 format) composed of 115,689 exons was generated and stored in GBrowse MySQL database.

2.5. Evaluating intron and exon databases and WormBase annotated introns, exons, and transcripts

2.5.1. Evaluation of intron database

C. briggsae RNA-Seq intron database composed of 95,632 introns defined in Section 2.4.1 was used to validate WormBase introns to observe the quality of this database. 74,972 introns or 73% of WormBase introns are supported by introns in our database (Figure 16, Figure 17) which shows the value of our database.

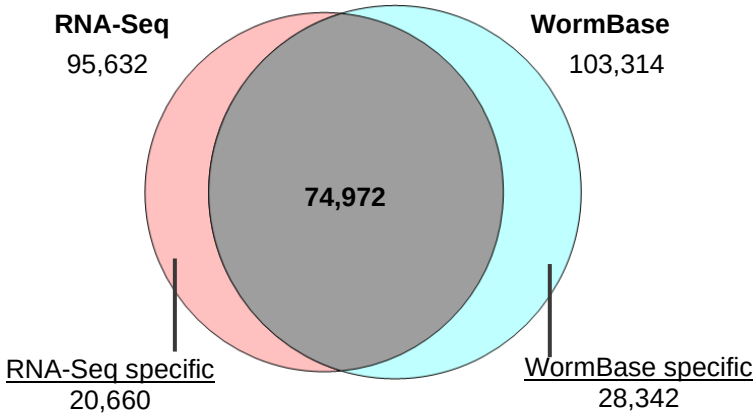

Figure 16. Venn diagram of RNA-Seq introns (left circle) and WormBase annotated introns (right circle). From WormBase introns' perspective, 73% of them are present in RNA-Seq introns, while 27% of them are not. From RNA-Seq introns' perspective, 78% of RNA-Seq introns are present in WormBase introns, and 22% of the introns suggest novel introns.

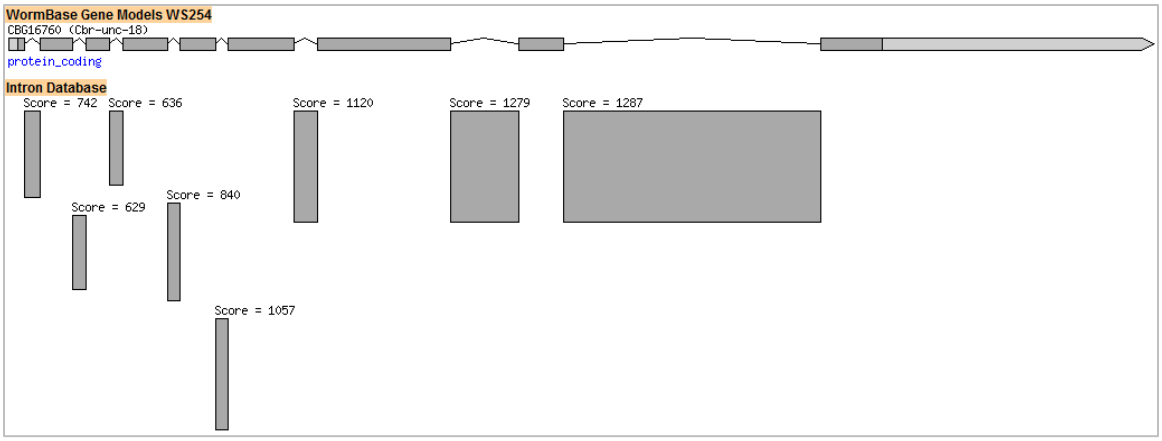

Figure 17. A representative *C. briggsae Cbr-unc-18* (uncoordinated) gene model whose introns are validated by RNA-Seq introns.

We also identified 20,660 novel introns that are not annotated in WormBase yet. 959 (4.64%) introns were observed to have a combination of two different current annotated splice sites, 9,956 (48.19%) introns observed to have one novel splice site, and 9,745 (47.17%) introns were observed to have two novel splice sites. Those novel introns indicated some modifications to 9,516 protein-coding genes in WormBase annotation (Table 4). 2,999 (14.51%) of novel introns did not map to existing protein-coding genes.

Table 4. Modifications to WormBase protein-coding gene models by RNA-Seq intron database

Category	Novel introns	Protein-coding genes affected
Internal novel intron	14,997	7,270
Directly extends gene	1,198	1,054
Extends gene via novel exon	1,292	905
Links multiple genes	164	287
Pseudogene	8	-
Non-coding	2	-
Other	2,999	-

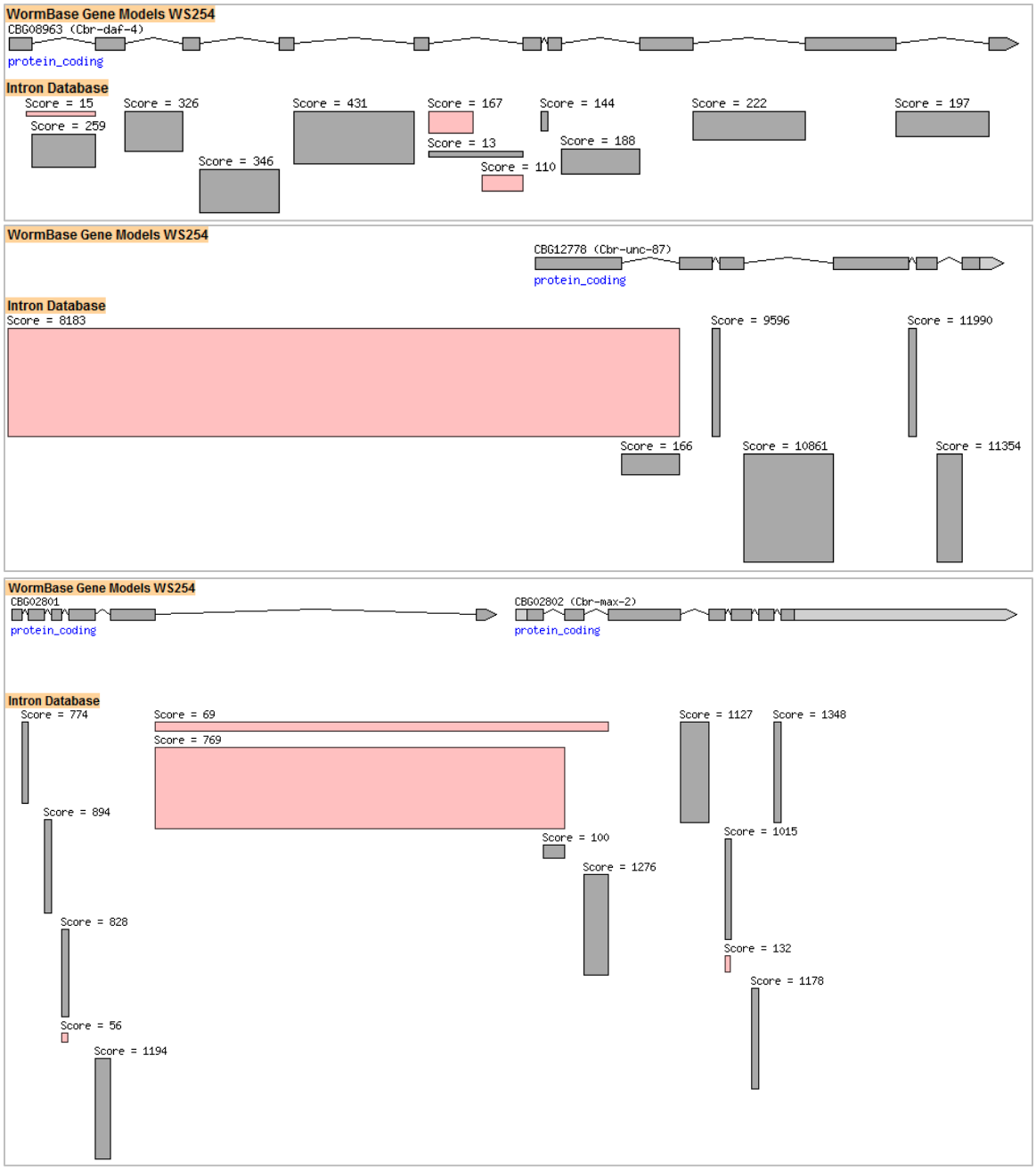

Figure 18. Representatives of *C. briggsae* gene models suggesting novel introns (top) in existing gene, (middle) extending existing gene, and (bottom) merging existing genes. Novel introns and existing introns are denoted in pink and grey, respectively. Genes pictured are (top) *Cbr-daf-4* (abnormal dauer formation), an ortholog of *C. elegans daf-4*, (middle) *Cbr-unc-87* (uncoordinated), an ortholog of *C. elegans unc-87*, (bottom) *Cbr-max-2* (motor axon guidance), an ortholog of *C. elegans max-2.*

Out of 103,314 WormBase introns, 28,343 (27%) are not supported by RNA-Seq introns. These introns were not captured due to their intron sizes (introns that are shorter than 30 bp or longer than 5000 bp are excluded during alignment using STAR, Figure 14), lack of support (introns that are not defined with or without RNA-Seq coverage for the introns), and inadequate support (introns that are defined but were filtered out after the minimum intron read support of 5 was applied) (Figure 16, Figure 19).

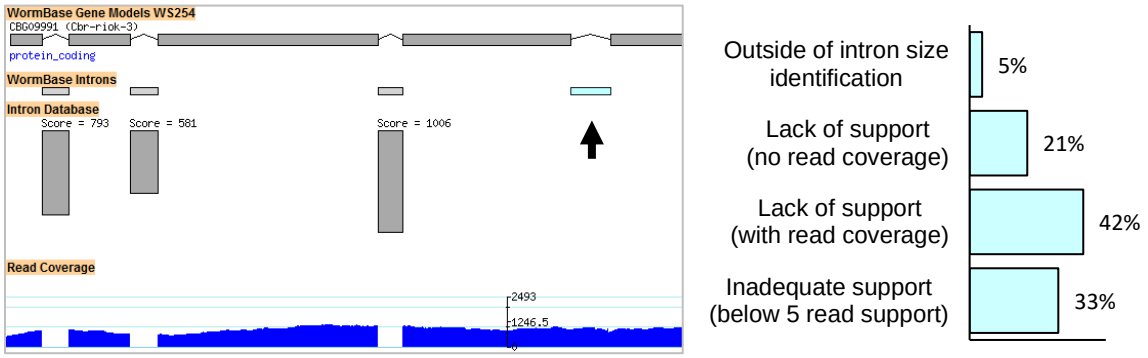

Figure 19. (left) Representative of *C. briggsae* gene model showing a WormBase intron is absent in RNA-Seq intron database (arrow). Gene pictured is *Cbr-riok-3* (rio kinase homolog), an ortholog of *C. elegans riok-3*, (right) Reasons WormBase introns were absent in RNA-Seq introns.

2.5.2. Evaluation of exon database

The RNA-Seq exon database composed of 115,689 exons generated in Section 2.4.2 was compared to the WormBase coding exons (feature type: CDS). 80,054 exons or 66% of WormBase coding exons are supported by our exons (Figure 20, Figure 21).

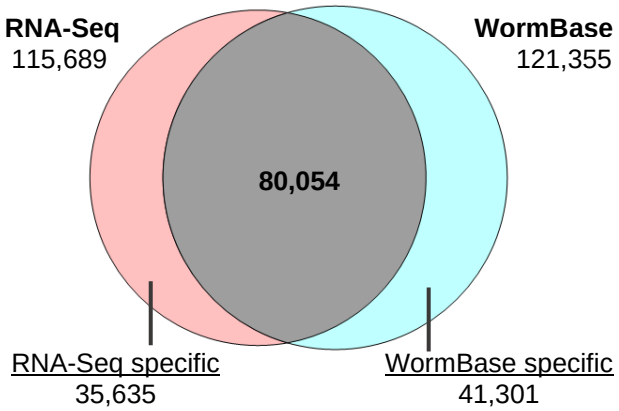

Figure 20. Venn diagram of RNA-Seq exons (left circle) and WormBase annotated exons (right circle). From WormBase exons' perspective, 66% of them are present in RNA-Seq result, while 34% of them are not. From RNA-Seq exons' perspective, 73% of RNA-Seq exons are present in WormBase, and 27% of the introns suggest novel introns.

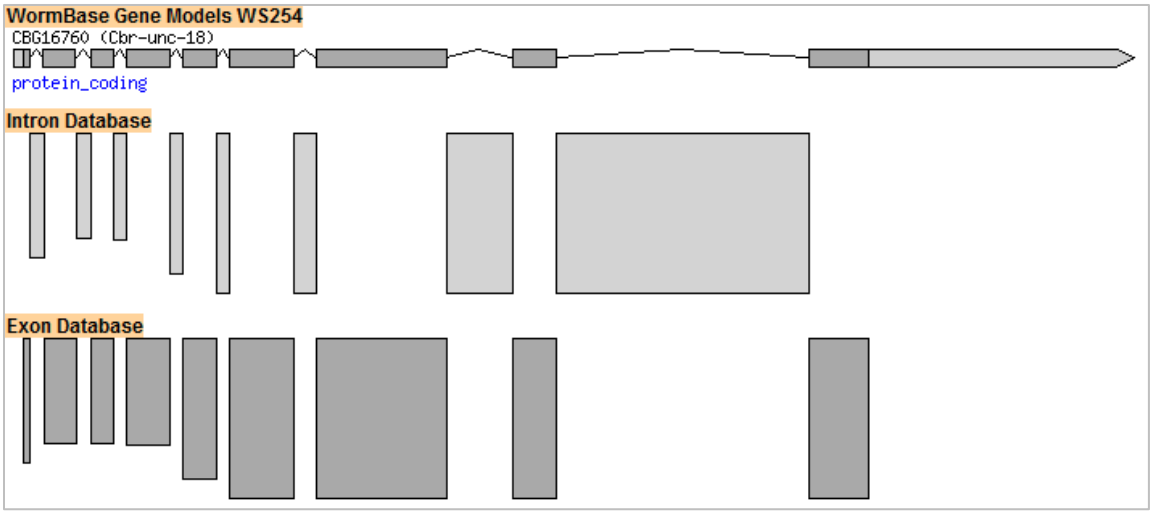

Figure 21. A representative *C. briggsae Cbr-unc-18* (uncoordinated) gene model whose exons completely match our exons in the database. GBrowse track: (1) WormBase WS254 gene models, (2) RNA-Seq intron database, (3) RNA-Seq exon database.

Out of our 115,689 exons, 35,635 exons (31% of RNA-Seq exons) were only present in RNA-Seq result (Figure 20). 1,177 exons (3.30%) were observed to have two different current annotated exon boundaries, 21,562 exons (60.51%) were observed to have one novel exon boundary, and 12,896 exons (36.19%) were observed to have two novel exon boundaries. All 35,635 exons were identified as novel exons, and they suggest modifications to 15,187 of gene models (Table 5). 3,200 (8.98%) of novel exons did not map to existing genes.

Table 5. Modifications to WormBase protein-coding gene models

Category	Novel exons	Protein-coding genes affected
Internal novel exon	27,197	10,733
Directly extends gene	3,007	2,803
Extends gene via novel intron	2,148	1,515
Links multiple genes	71	136
Overlaps non-coding gene	2	-
Overlaps pseudogene	10	-
Other	3,200	-

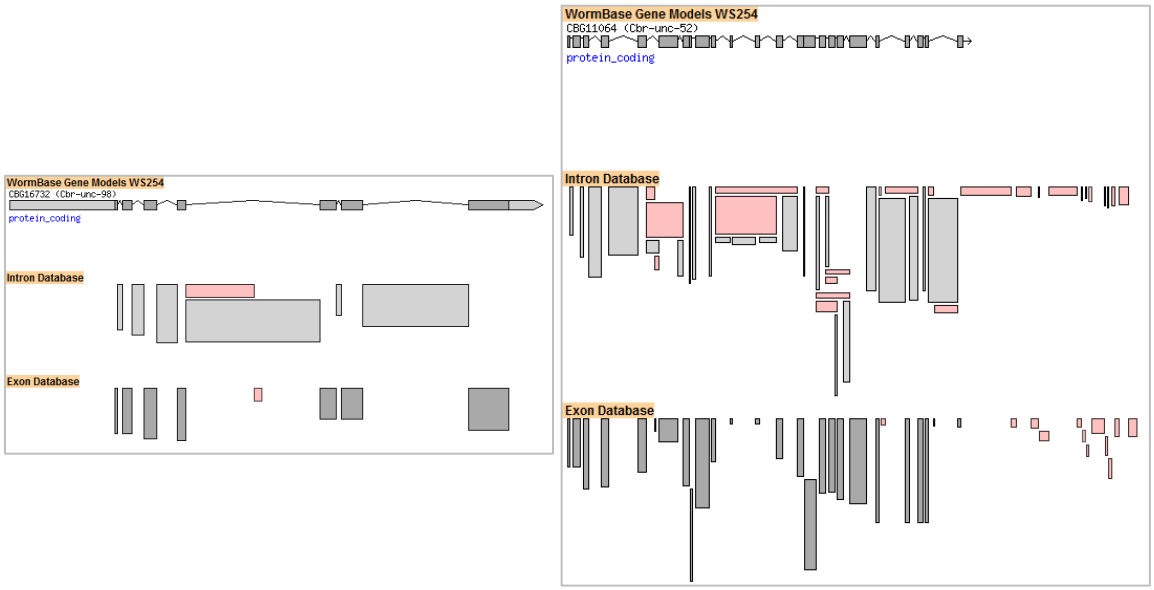

Figure 22. Representatives of *C. briggsae* gene models suggesting novel exons (left) in an existing gene, and (right) extending an existing gene. Genes pictured are *Cbr-unc-98* and *Cbr-unc-52* (uncoordinated), orthologs of *C. elegans unc-98* and *unc-52*. Novel introns and exons are denoted in pink.

Out of 121,355 WormBase exons, 4,476 (4%) partially match RNA-Seq exons and 36,825 (30%) have no match due to no adjacent introns in our intron database to reconstruct exon with (Figure 23).

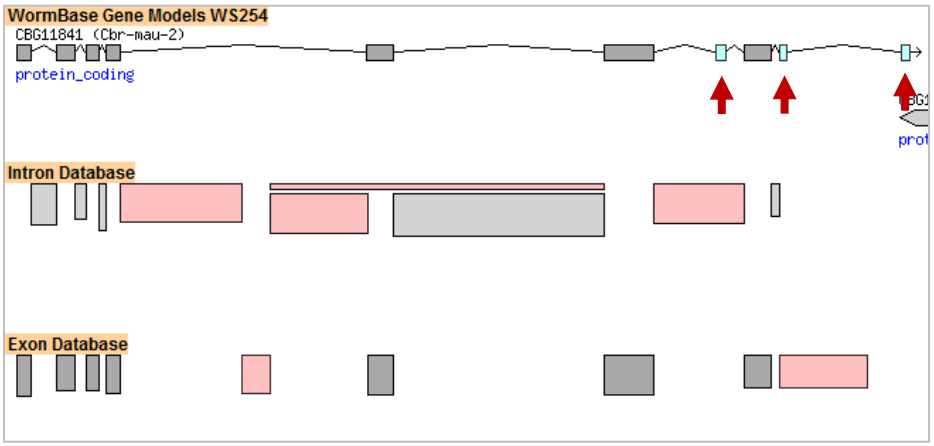

Figure 23. A representative of *C. briggsae* gene model that has a partial match (blue, second arrow) and no match (blue, first and third arrows) with RNA-Seq exons. Pictured is *Cbr-mau-2* (maternally affected uncoordination).

2.5.3. Evaluation of WormBase transcripts using the intron and exon databases

Evaluation of WormBase transcripts was done by using both high-quality databases defined previously. The results were categorized as complete (if all the introns and exons in the WormBase transcripts exist in our databases), partial (if not all introns and exons exist in our databases), and none (if none exists in our databases). 46% of the WormBase coding transcripts are validated by our databases, 31% of them are partially validated, while 23% of them do not exist in our databases (Figure 24).

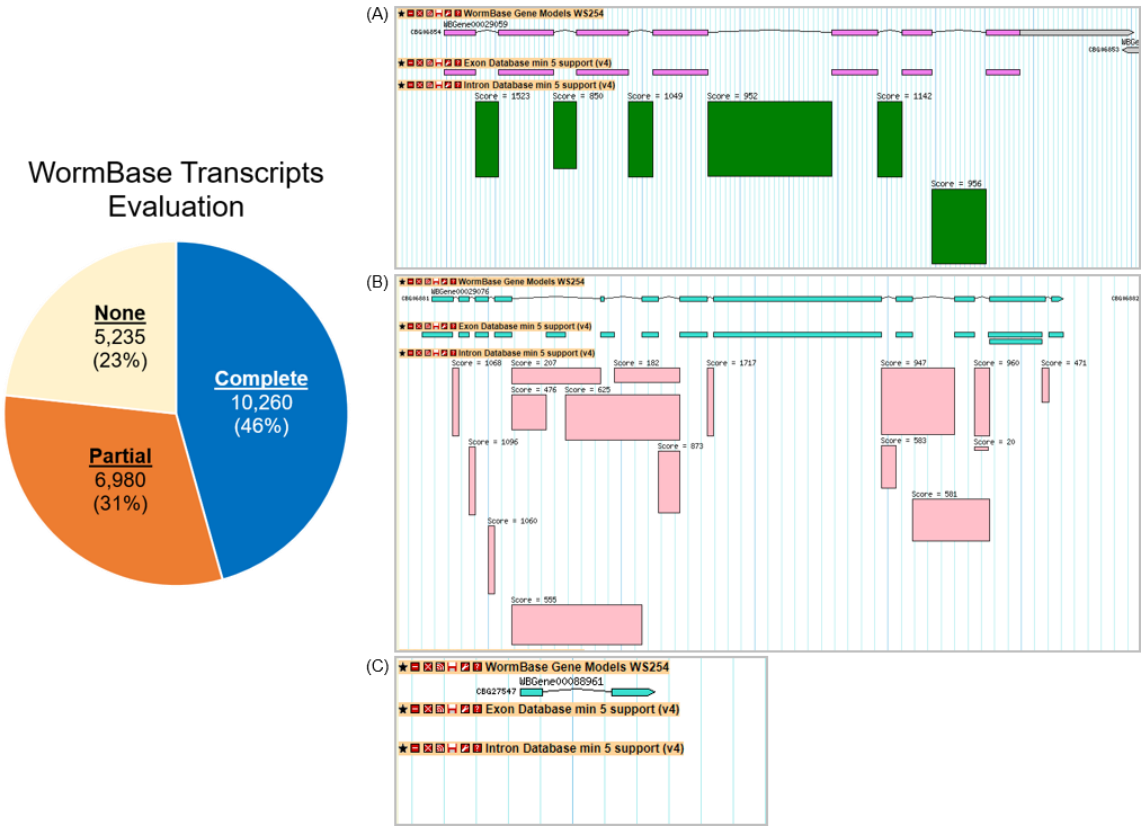

Figure 24. (left) Pie chart showing the proportion of WormBase transcripts whose introns and exons are completely, partially, or not represented by our databases; (right) Representatives of *C. briggsae* gene models showing WormBase coding transcripts where (A) all introns and exons present (complete), (B) not all introns and exons present (partial), (C) none of introns and exons present in our databases (none). Genes pictured are *Cbr-rab-6.1,* an ortholog of *C. elegans rab-6.1* involved in cortical granule exocytosis, *Cbr-unc-32*, an ortholog of *C. elegans' unc-32* involved in larval development, and *CBG27547*.

2.6. Building integrated intron and exon databases

Since we observed additional introns and exons and our goal was to improve the *C. briggsae* genome annotation, we further integrated our RNA-Seq-specific introns with the WormBase annotated introns to obtain a more complete *C. briggsae* intron set. The integrated intron database (n=123,974) was then used to reconstruct exons (n=147,648). The RNA-Seq-specific exons were further integrated with the WormBase annotated exons generating a more complete *C. briggsae* exon set (n=150,690).

2.7. Discussion

A more complete and accurate structural annotation of a sequenced genome is essential for downstream analyses. One approach to achieve that is to identify evidence of the individual gene components (features) such as introns and exons. In this chapter, we improved the completeness of *C. briggsae* annotation at the intron and exon level using a pool of 13 paired-end RNA-Seq libraries.

Paired-end libraries were selected because single-end reads were found to detect fewer splice junctions compared with paired-end reads (Chhangawala et al., 2015). Paired-end RNA-Seq is thought to be critical for alternative splicing studies because it increases the probability of observing fragments that connect exon junctions compared to single-end RNA-Seq (Rossell et al., 2014), helps to reduce false discoveries of splicing junctions (Au et al., 2010), and increases the number of reads that could be uniquely mapped. Longer paired-end reads are also significantly better for detecting splice junctions and alternative splicing events (Chhangawala et al., 2015).

In the process of building a high-quality intron database, we applied a threshold for selecting introns generated from RNA-Seq. In general, there is a trade-off between sensitivity (True Positive Rate) and specificity (True Negative Rate). High sensitivity, in the case of allowing any read support for intron identification, would allow our analysis to detect all introns that will not only include those that are real (true positives) but also include a lot of spurious introns (false positives), therefore reducing specificity. On the other hand, high specificity (stringent filtering with high minimum intron support) would allow our analysis to reduce the number of spurious introns (false positives) but would consequently reduce the number of real introns (true positives), leading to lower

sensitivity. Minimum intron support of five reads in at least one library was chosen to compromise between sensitivity (79%) and specificity (52%) for intron detection. The sensitivity calculated is an approximate due to lack of ground truth. Nevertheless, in addition to having minimum 5 read support, a set of criteria leading up to the intron definition step were applied to ensure the quality of our RNA-Seq introns being defined. The criteria include filtering out multimapping alignments, selecting a splice-aware alignment tool that was found to have the lowest false positive rate for intron identification, filtering out introns with non-canonical splice sites, and selecting introns with lengths between 30 and 5000 bp.

We generated RNA-Seq intron and exon databases consisting of 95,632 introns and 115,689 exons that validated 73% and 66% of WormBase annotated introns and exons, showing the value of both databases. If the *C. briggsae* annotation was complete, we predict that our RNA-Seq introns and exons would all be represented in annotated transcripts. On the other hand, if an annotation can still be improved, we would observe unannotated RNA-Seq introns and exons. Our RNA-Seq introns and exons revealed that many gene model annotations can be improved. Specifically, 22% (20,660) introns and 27% (35,635) exons are currently absent from the *C. briggsae* WS254 annotation with the majority of them suggesting additional introns and exons in the existing genes, and the minority of them suggesting extension of existing genes and merging of multiple genes. The finding of novel features within known and previously unknown genes from RNA-Seq studies is consistent with other studies in other organisms (Bruno et al., 2010; Hillier et al., 2009; Loraine et al., 2013; Mortazavi et al., 2008; Uyar et al., 2012). The identification of novel introns and exons along with the observation of gene model modifications is the first evidence supporting our hypothesis that *C. briggsae* genome annotation can still be improved.

On the other hand, 27% and 34% of WormBase introns and exons are not present in our databases. One reason those introns and exons were not covered was due to limited RNA-Seq libraries that are currently available for use. We further support this argument by performing the same pipeline on a limited number of *C. elegans* RNA-Seq libraries (13 libraries, see Section 3.4.2, additional analysis), which resulted in a significantly lower number of introns and exons compared to when hundreds of libraries were used. This analysis suggested that feature discovery power seems to be proportional to RNA-Seq data quantity and it showed a promising potential for identifying

more *C. briggsae* features by producing more RNA-seq data in the future. Additionally, some of the introns were not covered due to the parameters we applied during the process of building high-quality databases, which are outside of the intron size limit (30-5000bp) and below read support threshold (minimum 5 read support in at least one of the libraries). The latter was also affected by the limited number of RNA-Seq libraries available that is likely to be solved by having more libraries in future studies. Moreover, 42% of WormBase-specific introns were potentially mispredictions because there were no introns defined by our method despite read covering the corresponding intron regions, however, more evidence is needed to confirm this. Furthermore, the partially represented and completely not represented exons could also be due to mispredictions (i.e., false positives, those that should be non-existent and corrected) or false negatives caused by missing adjacent introns to reconstruct those exons affected by limited data. False positive features in WormBase are possible because the *C. briggsae* genome annotations are limited to mostly computational studies with some experimental studies.

Despite the limited number of RNA-Seq libraries selected in this thesis, our RNA-Seq databases are powerful to validate almost half of the annotated transcripts. Additionally, thousands of RNA-Seq novel introns and exons cannot be assigned to WormBase transcripts suggesting that the current *C. briggsae* coding capacity is lower than it actually is. Therefore, we incorporated our RNA-Seq-specific features into the WormBase annotated features. This approach is beneficial to reduce false negatives (by 27% to 34% from possible missing annotations) due to limited RNA-Seq libraries available with the consequence of keeping possible false positives (~0.1% from WormBase potential errors). The resulting integrated features serve as the more complete *C. briggsae* annotation at the intron and exon level.

Another evidence that can support the idea of incomplete *C. briggsae* annotation is the number of SL *trans*-splicing acceptor sites in *C. briggsae* and *C. elegans*. As shown in Table 1, 11,617 sites were found in 8,555 *C. briggsae* genes (Uyar et al., 2012), while 28,249 sites were found in 11,387 *C. elegans* genes. Total sites are more than genes because 2,130 of *C. elegans* genes were found to have more than one site (Allen et al., 2011). Through a preliminary analysis using 13 RNA-Seq libraries, we identified 12,516 SL *trans*-splicing acceptor sites in *C. briggsae* with at least 2 read support. Out of those, 9,750 (69.91%) sites are not annotated in WormBase WS254 (see

Section 3.4.2, preliminary analysis). The study of *trans*-splicing in *C. briggsae* can be followed up in the future.

Chapter 3. Improving *C. briggsae* protein-coding transcript set

3.1. Introduction

In this section, we built an improved *C. briggsae* protein-coding transcript set. RNA-Seq transcripts were assembled and only transcripts supported by introns and exons (in integrated databases) were kept. The supported protein-coding transcripts were further computationally filtered prior to integration with the current WormBase (WB) WS254 protein-coding transcript set. The integrated *C. briggsae* protein-coding transcripts were then used for comparative analysis with *C. elegans* in the next chapter.

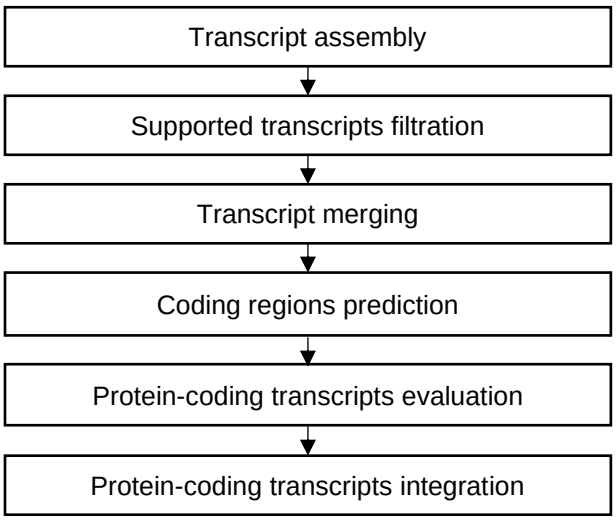

Figure 25. Workflow for building an improved set of *C. briggsae* protein-coding transcripts

3.2. Assembling transcripts using *de novo* and genome-guided methods and filtering transcripts supported by integrated introns and exons

Cufflinks version 2.2.1 (Trapnell et al., 2012), StringTie version 1.3.4d (Pertea et al., 2015), and Trans-AbySs version 1.5.5 (Robertson et al., 2010) were used to assemble transcripts from RNA-Seq. Cufflinks and StringTie are genome-guided transcript assemblers, and Trans-AbySs is a *de-novo* transcript assembler. Input for Cufflinks and StringTie was filtered mapped reads used for intron identification, while input for Trans-AbySs was pre-processed raw reads. Trans-AbySs used GMAP version

2017-04-13 (Wu et al., 2016) to align the merged assembly to the reference genome. Parameters were set as default for all assemblers.

The assembled transcripts were filtered by the presence of introns and exons in the improved databases. 43-76% of transcripts per library were fully supported by our database depending on the transcript assembler used (Appendix B5). Filtered transcripts from 3 programs were merged using GffCompare version 0.10.4 (https://github.com/gpertea/gffcompare), resulting in 29,352 transcripts (11,648 redundant transcripts were discarded).

3.3. Predicting coding-regions of the supported transcripts

The coding regions of merged transcripts (n=29,352) were identified using the TransDecoder version 5.5.0 (http://github.com/TransDecoder/TransDecoder). TransDecoder predicts likely coding regions within the transcript sequences by identifying open reading frames (ORFs) and scoring them according to their sequence composition. 24,705 (84.17%) of the assembled transcripts were found to have a candidate coding region.

3.4. Evaluating the assembled protein-coding transcripts and generating an improved protein-coding transcript set

The assembled protein-coding transcripts (n=24,705) were further evaluated for their potential to improve the current WB protein-coding transcript set (release WS254). Comparison between the assembled protein-coding transcripts and WB protein-coding transcripts were performed. Intron chains in the assembled and WB transcripts were parsed. An intron chain is a set of introns flanked by two CDSs in a transcript. For each transcript, we compare their intron chains against each other and categorize the assembled transcripts into 13 specific categories (Figure 26).

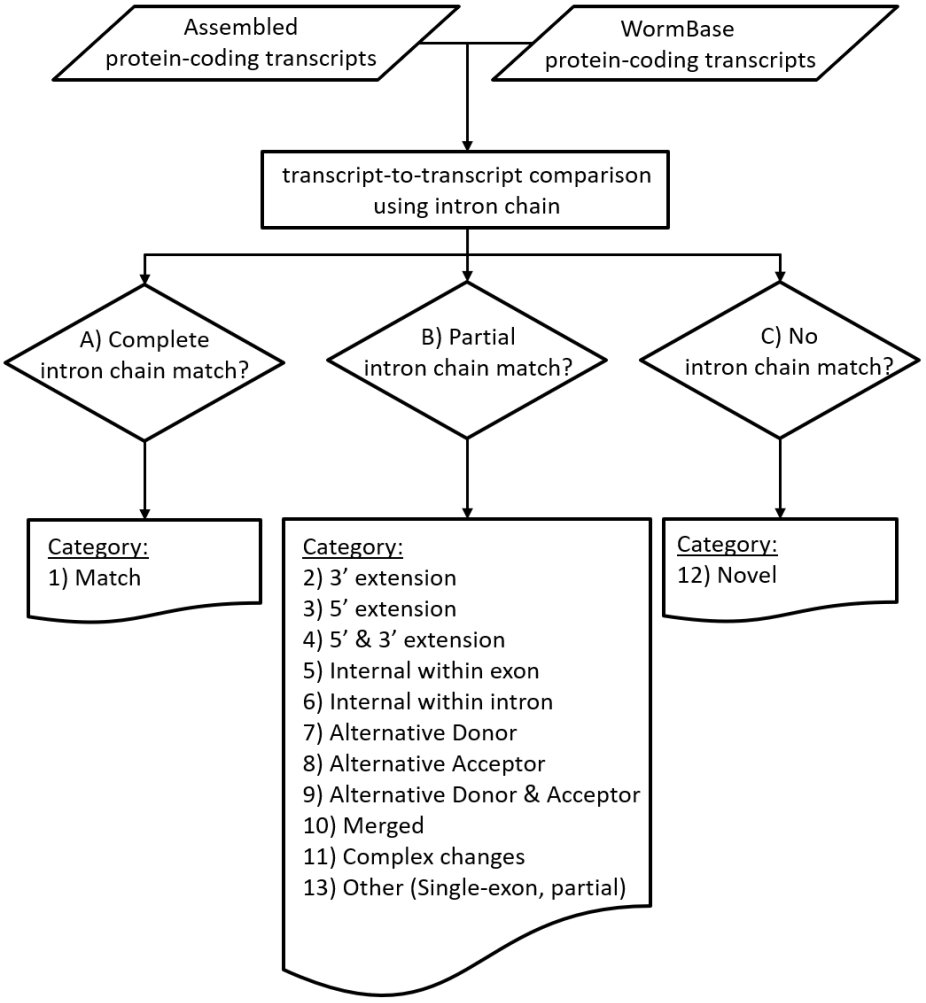

Figure 26. Protein-coding transcripts evaluation pipeline.

3.4.1. Algorithm and representative figures

Strand-specific transcript-to-transcript comparison and categorization were performed using a custom Python algorithm. There are 3 general categories−match, overlap, and novel. The first two categories are for assembled transcripts that have their intron chains completely match or partially match (i.e., overlapping) those of existing WormBase transcripts. The last category is for assembled transcripts that are not present (i.e., not overlapping) WormBase transcripts. Below are criteria to capture those instances and representative figures displayed using GBrowse showing WormBase transcript (track 1), assembled transcript (track 2), and integrated intron database (track 3), respectively.

1. Match WB transcripts

When both assembled transcript and WB transcript have identical intron chains, the assembled transcript falls into this category. Transcripts in this category are transcripts confirming WB annotated transcripts.

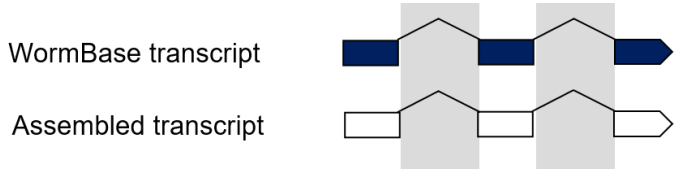

Figure 27. An illustration of match category

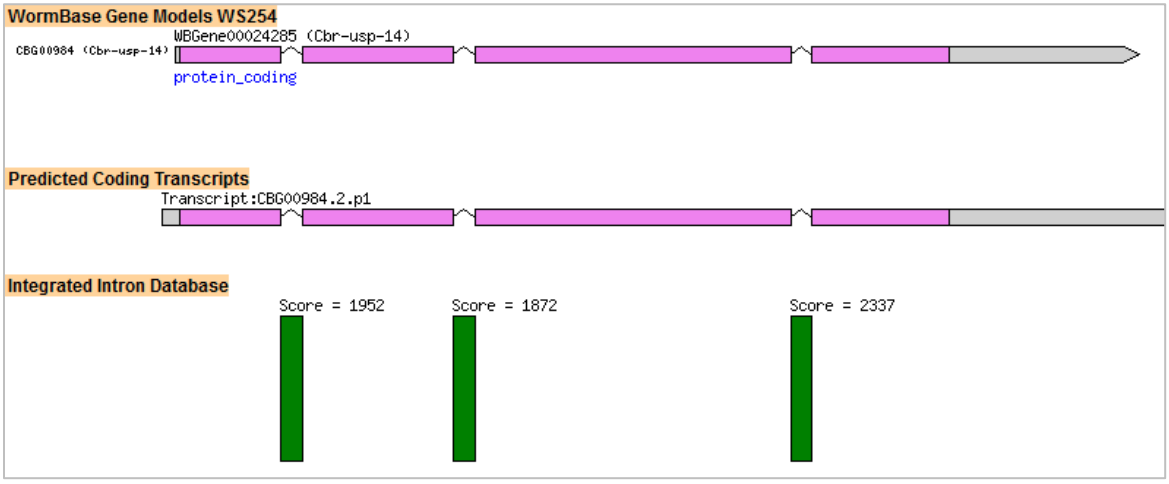

Figure 28. An example of an assembled transcript (CBG00984.2) with all introns match those of WB transcript *Cbr-usp-14* (ubiquitin specific protease), an ortholog of *C. elegans' usp-14*.

2. Extending 3' of WB transcripts

When the intron chain in a WB transcript is a subset of the intron chain in an assembled transcript (i.e., all the WB introns exist in the assembled transcript), and one or more additional introns are found at the 3' end of the assembled transcript, this assembled transcript falls into this category. In some cases, transcripts that fall into this category also have additional intron(s) overlapping annotated CDS (see category 5). This category will also contain linked 3' genes where one of the two transcripts/genes is a single-exon gene.

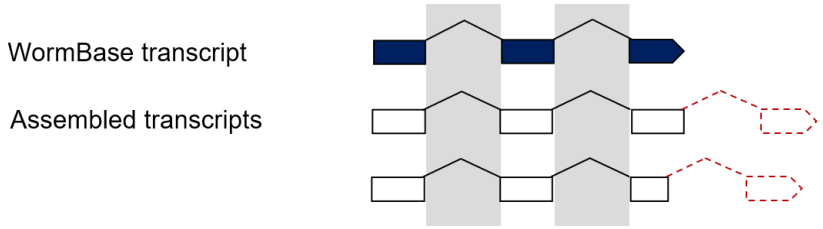

Figure 29. An illustration of 3’ extension category.

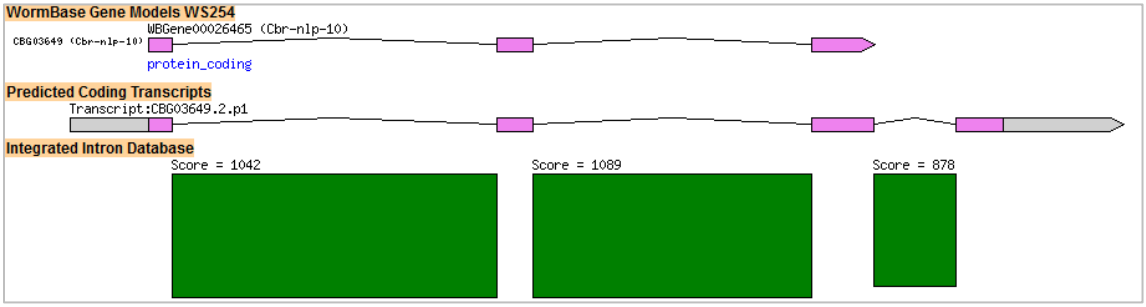

Figure 30. An example of an assembled transcript (CBG0649.2) with one additional intron with high support extending 3’ of the WB transcript *Cbr-nlp-10* (neuropeptide-like protein), an ortholog of *C. elegans’ nlp-10*.

3. Extending 5’ of WB transcripts

When the intron chain in a WB transcript is a subset of the intron chain in an assembled transcript (i.e., all the WB introns exist in the assembled transcript), and one or more additional introns are found in the 5’ end of the assembled transcript, this assembled transcript falls into this category. In some cases, transcripts that fall into this category also have additional intron(s) overlapping annotated CDS (see category 5).

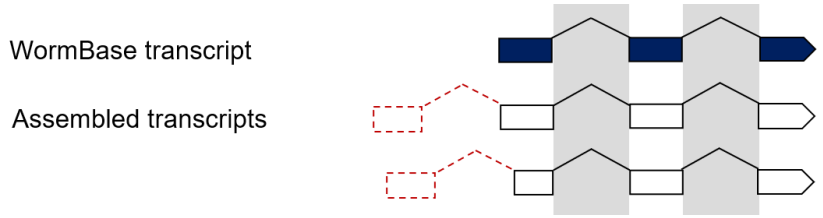

Figure 31. An illustration of 5’ extension category.

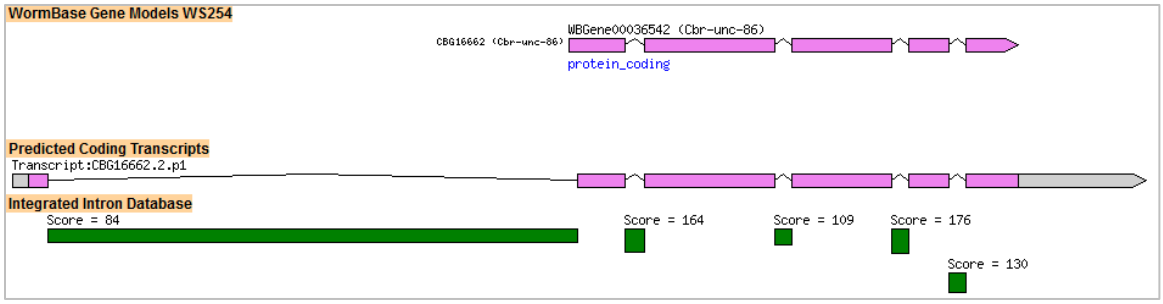

Figure 32. An example of an assembled transcript (CBG16662.2) with one additional intron extending 5' of the WB transcript *Cbr-unc-86* (uncoordinated), an ortholog of *C. elegans' unc-86*.

4. Both 5' and 3' extension

When the intron chain in a WB transcript is a subset of the intron chain in an assembled transcript (i.e., all the WB introns exist in the assembled transcript), and exactly two more additional introns are found in the 5' and 3' end of the assembled transcript, this assembled transcript falls into this category. Assembled transcripts with more than two additional introns will fall into complex changes category (both 5' and 3' extension combined with internal within exon, see category 11A).

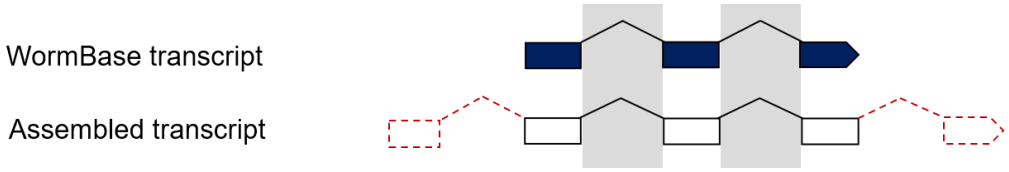

Figure 33. An illustration of 5' and 3' extension category.

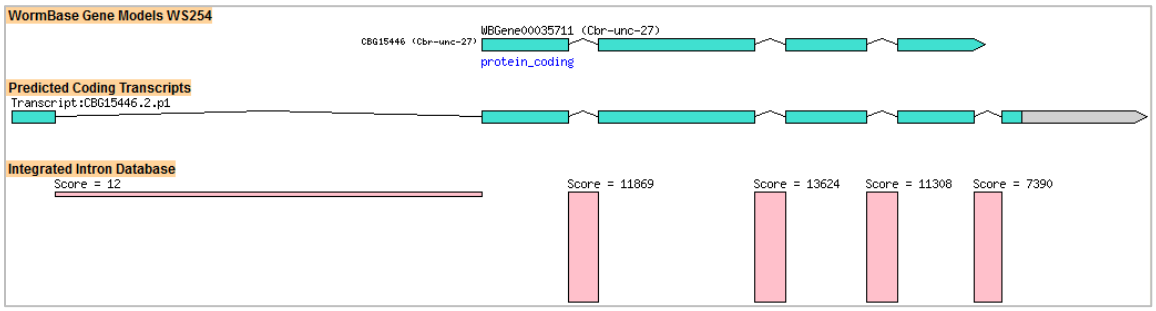

Figure 34. An example of an assembled transcript (CBG15446.2) with two additional introns extending 5' and 3' of the WB transcript *Cbr-unc-27* (uncoordinated), an ortholog of *C. elegans' unc-27*.

5. Introns overlapping WB internal exon (CDS)

An assembled transcript will fall into this category when the intron chain in a WB transcript is a subset of intron chain in an assembled transcript (i.e., all the WB introns exist in the assembled transcript), the terminal intron boundaries of assembled transcript are the same as those of WB transcript, and one or more additional introns are found in the assembled transcript, those introns are overlapping/located within an annotated exon.

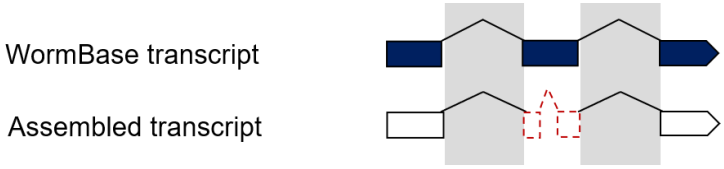

Figure 35. An illustration of intron overlapping internal exon category.

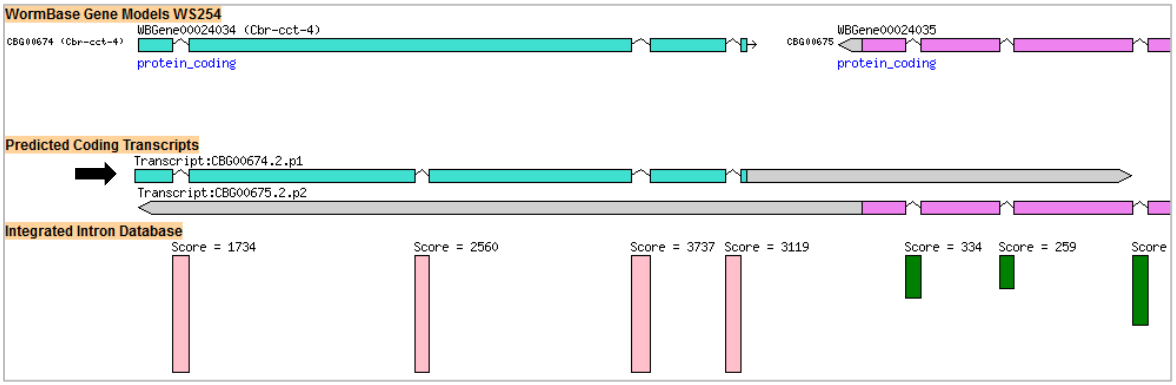

Figure 36. An example of an assembled transcript (CBG00674.2, arrow) with one additional intron internal of exon of the WB transcript *Cbr-cct-4* (chaperonin containing TCP-1), an ortholog of *C. elegans' cct-4.*

6. Introns overlapping WB internal intron

An assembled transcript will fall into this category when some but not all introns in a WB transcript are found in an assembled transcript, and two or more additional introns are found and overlapping/located within an annotated intron. Changes are always internal of the leftmost and rightmost intron boundaries of the transcript.

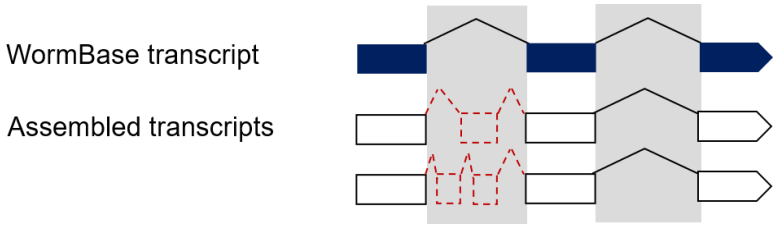

Figure 37. An illustration of intron overlapping internal intron category.

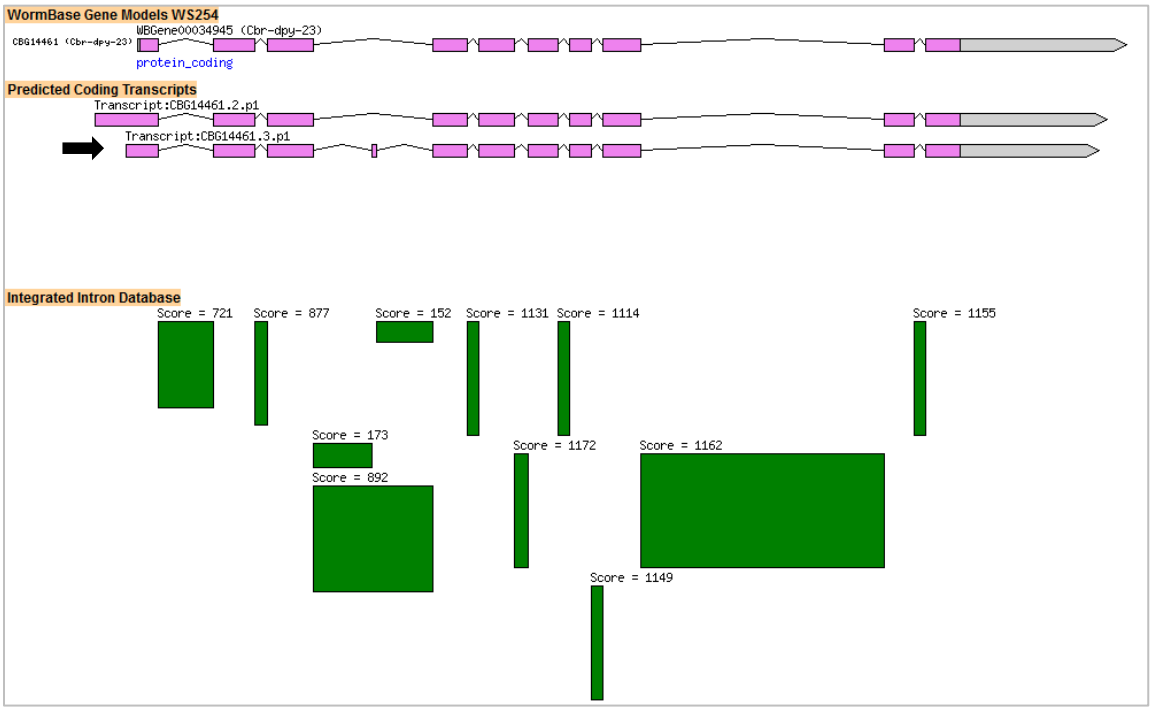

Figure 38. An example of an assembled transcript (CBG14461.3, arrow) with one additional intron internal of WB intron of the WB transcript *Cbr-dpy-23* (dumpy: shorter than wildtype), an ortholog of *C. elegans' dpy-23*.

7. Alternative donor

An assembled transcript will fall into this category when one intron in the assembled transcript is not in the WB transcript, and the non-existent intron has a different start but the same end position (i.e., different 5' splice site but the same 3' splice site). If multiple events are observed, will be categorized as complex changes (category 11).

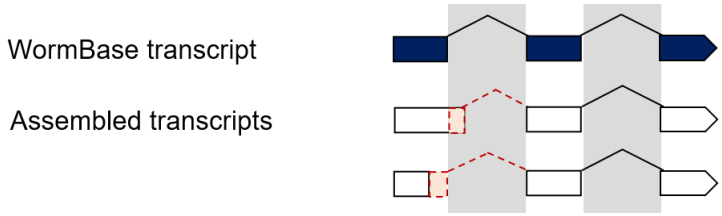

Figure 39. An illustration of alternative donor category.

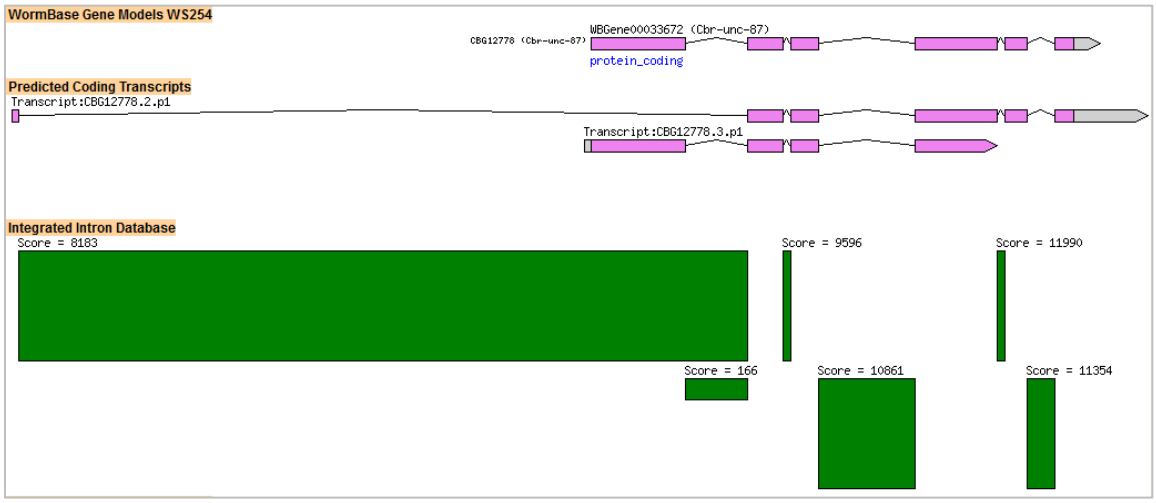

Figure 40. An example of an assembled transcript (CBG12778.2, Predicted Coding Transcripts track, top transcript) with a different 5' splice site compared to the WB transcript CBG12778.2 of *Cbr-unc-87* (uncoordinated), an ortholog of *C. elegans' unc-87*.

8. Alternative acceptor

An assembled transcript will fall into this category when one intron in the assembled transcript is not in the WB transcript, and the non-existent intron has a different end but the same start position (i.e., different 3' splice site, but the same 5' splice site). If multiple events are observed, will be categorized as complex changes (category 11).

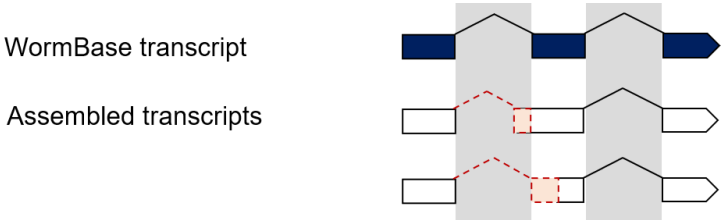

Figure 41. An illustration of alternative acceptor category.

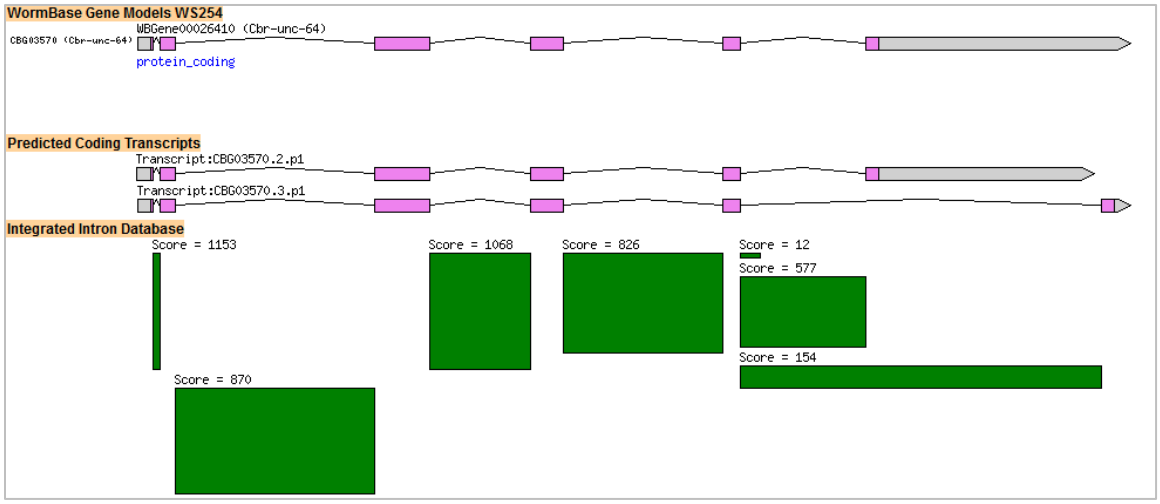

Figure 42. An example of an assembled transcript (CBG03570.3, Predicted Coding Transcripts track, bottom transcript) with a different 3' splice site compared to the WB transcript *Cbr-unc-64* (uncoordinated), an ortholog of *C. elegans' unc-64*.

9. Alternative donor and acceptor

An assembled transcript will fall into this category when the total number of introns in the assembled transcript is equal to those in the WB transcript, and only one intron in the assembled transcript is not in the WB transcript. The non-existent intron has different start and end positions (i.e., different 5' & 3' splice sites). If multiple events are observed, will be categorized as complex changes (category 11)

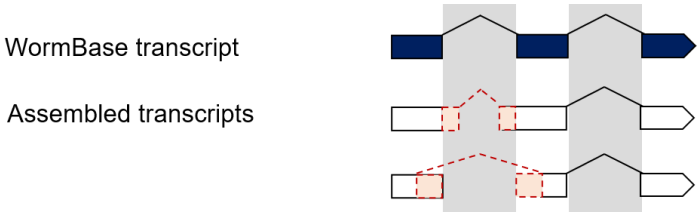

Figure 43. An illustration of alternative donor and acceptor category.

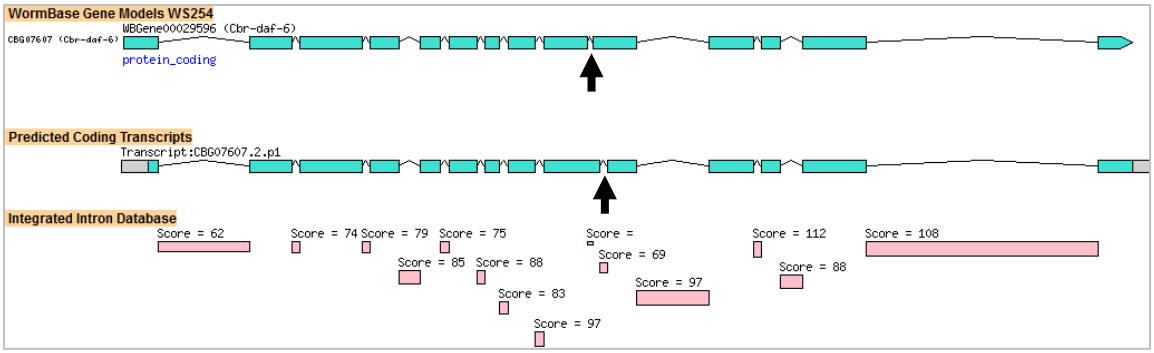

Figure 44. An example of an assembled transcript (CBG07607.2) with different 5' and 3' splice sites compared to WB transcript *Cbr-daf-6* (abnormal dauer formation), an ortholog of *C. elegans' daf-6.*

10. Merged

An assembled transcript will fall into this category when intron chain in a WB transcript is a subset of intron chain in the assembled transcript (i.e., all WB introns exist in the assembled transcript), and one or more additional introns that are not in the corresponding WB transcripts are found in another WB transcript (in the case of merging 2 genes) or other WB transcripts (in the case of merging 3 genes).

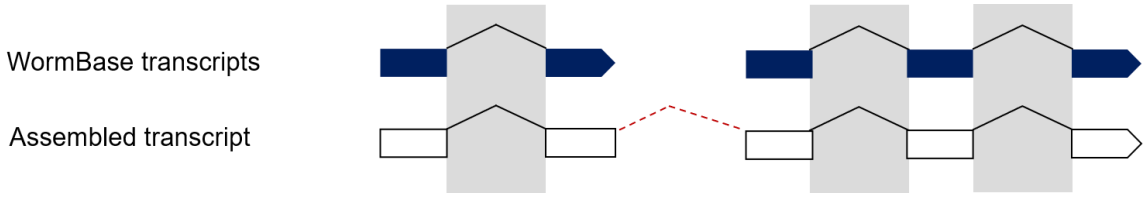

Figure 45. An illustration of merging genes category.

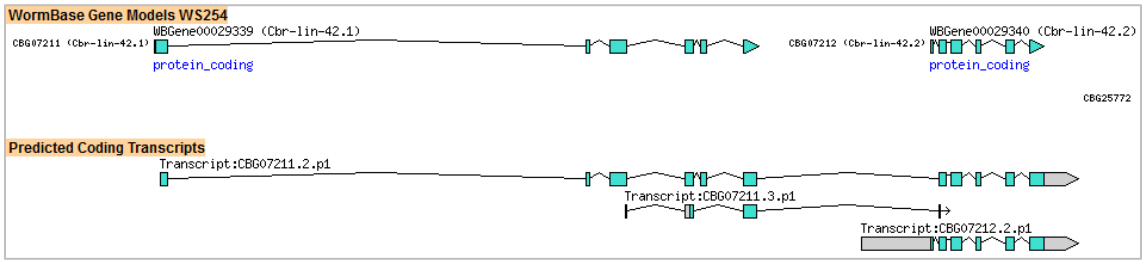

Figure 46. An example of an assembled transcript (CBG07211.2, Predicted Coding Transcripts track, top transcript) merging two WB transcripts *Cbr-lin-42.1* and *Cbr-lin-42.2* (abnormal cell lineage), an ortholog of *C. elegans' lin-42.*

11. Complex changes

A. Both 5’ and 3’ extension & internal within exon

An assembled transcript will fall into this category when the intron chain in WB transcript is a subset of intron chain in the assembled transcript (i.e., all the WB introns exist in the assembled transcript), and the assembled transcript contains two more additional introns at the 5’ and 3’ end and one or more introns overlapping annotated exon (CDS).

B. Multiple alternative splicing events

An assembled transcript will fall into this category when the number of introns in the assembled transcript is equal to those in the WB transcript, and more than one intron in the assembled transcript are not in the WB transcript and those introns do not share any intron boundaries with introns in the WB transcript.

C. Modifications internal of terminal intron boundaries

An assembled transcript will fall into this category when the 5’ and 3’ terminal intron boundaries of the assembled transcript are the same as those in the WB transcript, and the assembled transcript has a combination of many types of modifications (for example, internal within exon and alternative donor, internal within introns and intron retention, internal within exon and alternative acceptor, internal within intron and alternative donor and internal within exon) in one transcript.

D. Modifications unrelated to terminal intron boundaries

An assembled transcript will fall into this category when the 5’ and 3’ terminal intron boundaries of the assembled transcript are different to those in the WB transcript, and the assembled transcript has a combination of many types of modifications (for example, 5’ and 3’ extension and alternative acceptor, 3’ extension and alternative acceptor, alternative acceptor and internal within exon, 3’ extension and alternative acceptor).

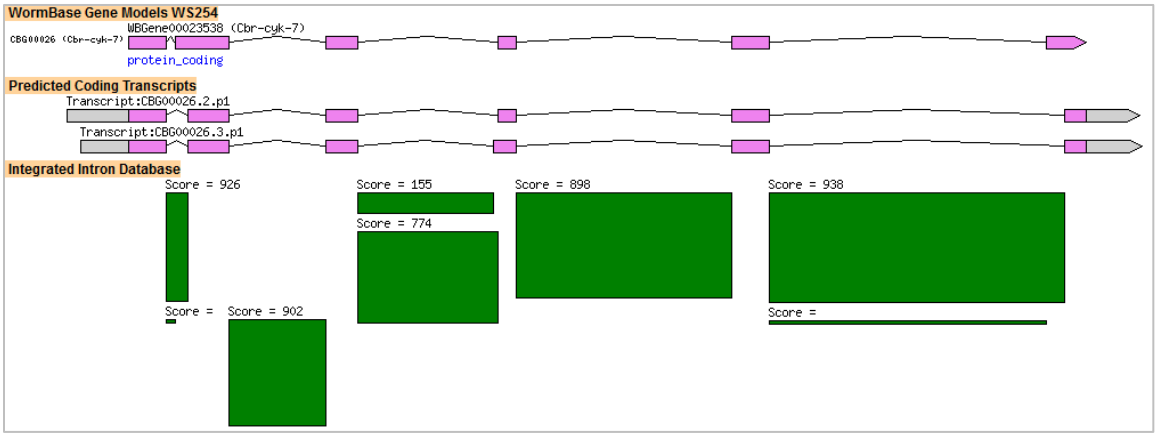

Figure 47. An example of two assembled transcripts (CBG00026.2, CBG00026.3) with multiple alternative splicing events compared to the WB transcript *Cbr-cyk-7* (cytokinesis defect), an ortholog of *C. elegans' cyk-7.*

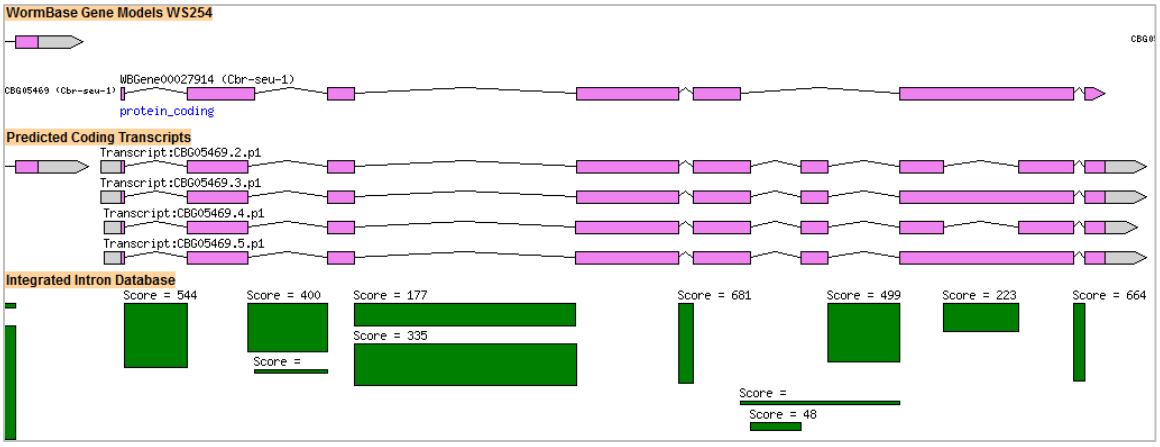

Figure 48. An example of four assembled transcripts (CBG05469.2-5) with additional introns suggesting a combination of multiple modifications compared to the WB transcript *Cbr-seu-1* (suppressor of ectopic unc-5), an ortholog of *C. elegans' seu-1*. Modifications include alternative donor, alternative acceptor, and internal intron within exon.

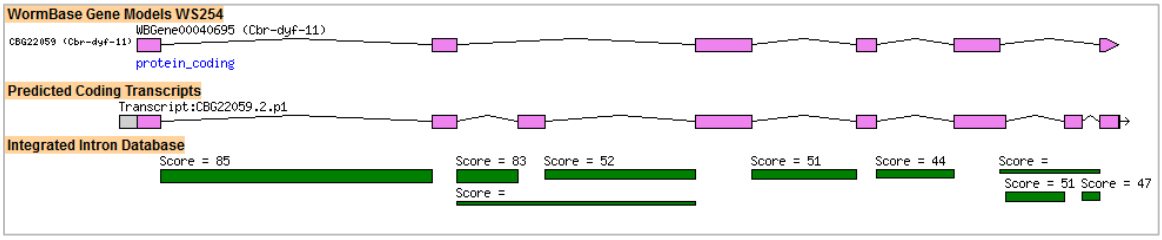

Figure 49. An example of an assembled transcript (CBG22059.2) with additional introns suggesting a combination of multiple modifications compared to the WB transcript *Cbr-dyf-11* (abnormal dye filling), an ortholog of *C. elegans' dyf-11.* Modifications include internal intron within intron and alternative donor usage.

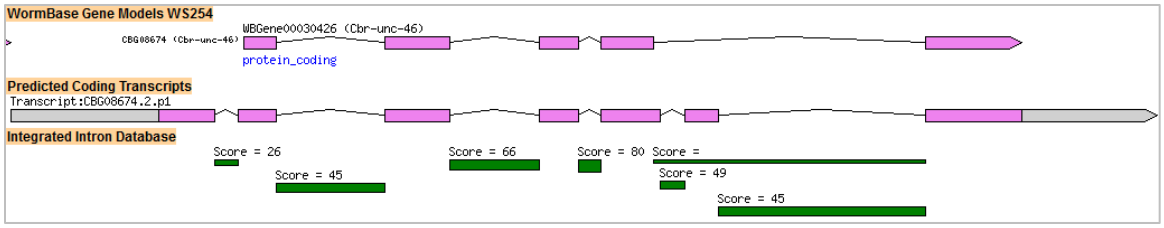

Figure 50. An example of an assembled transcript (CBG08764.2) with additional introns suggesting a combination of multiple modifications compared to the WB transcript *Cbr-unc-46* (uncoordinated), an ortholog of *C. elegans' unc-46.* Modifications include 5' extension, alternative donor usage, and additional intron within intron.

12. Novel

When the assembled transcripts do not exist in WB annotated transcripts. These transcripts were labeled as nCBG00001 to nCBG00159.

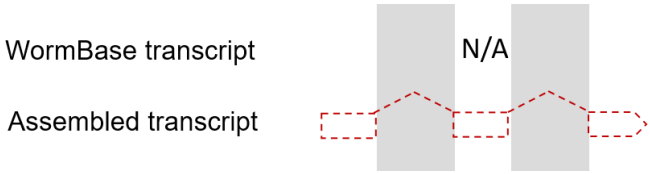

Figure 51. An illustration of novel transcript category.

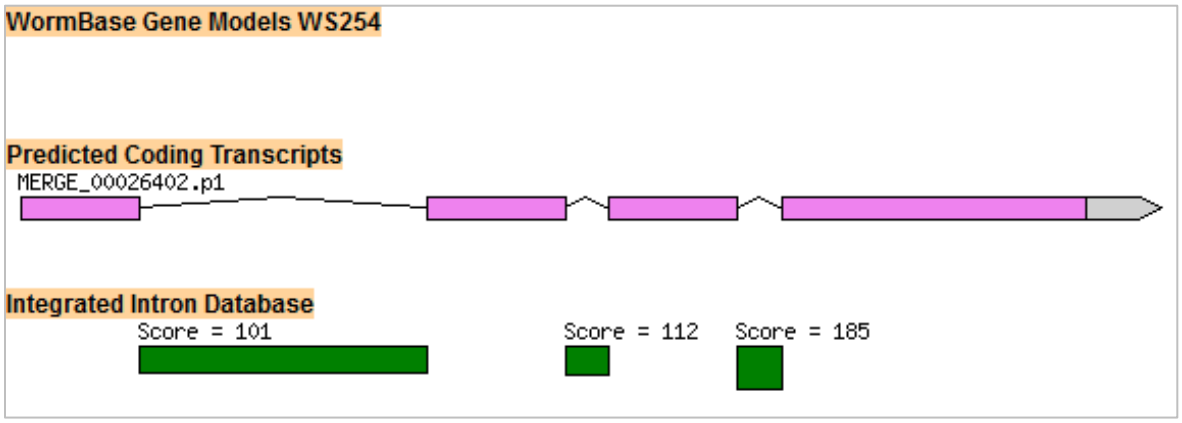

Figure 52. An example of an assembled transcript (MERGE_00026402.p1 or nCBG00109) suggesting a novel transcript and gene in this region (V:10,295,768..10,297,767).

13. Other

A. Single-exon transcripts
   a. Case 1: When no intron identified in both assembled and WB transcripts,
   b. Case 2: When there are introns in assembled transcripts but none in WB transcripts.

Figure 53. An illustration of single-exon transcript category.

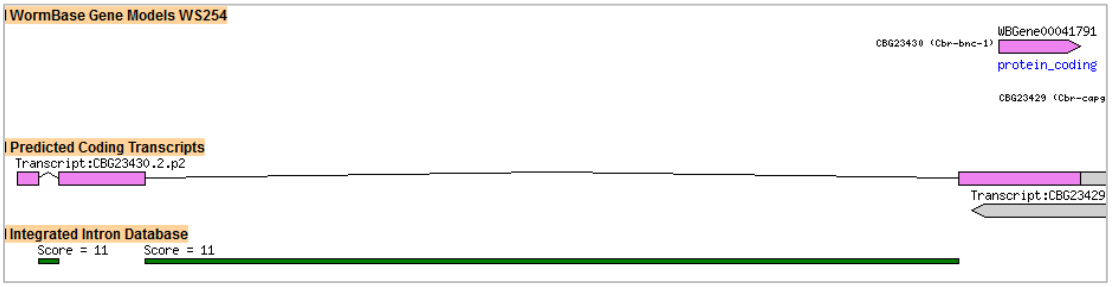

Figure 54. An example of an assembled transcript (CBG23430.2) with introns in comparison to WB annotated single-exon transcript *Cbr-bnc-1* (basonuclin-1 zinc finger protein homolog), an ortholog of *C. elegans' bnc-1.*

B. Partial transcripts

When the number of introns in assembled transcripts are less than in corresponding WB transcripts, these could be fragmented transcripts or real partial transcripts that need to be corrected.

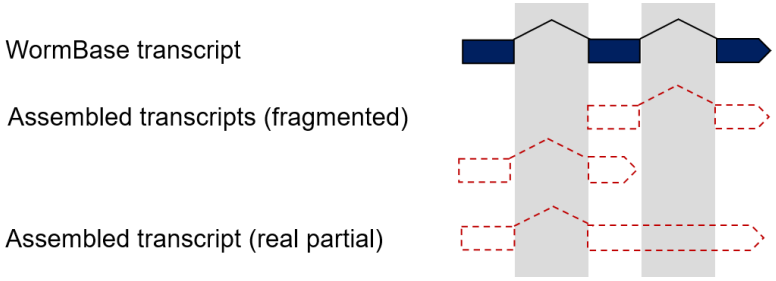

Figure 55. An illustration of partial transcript category.

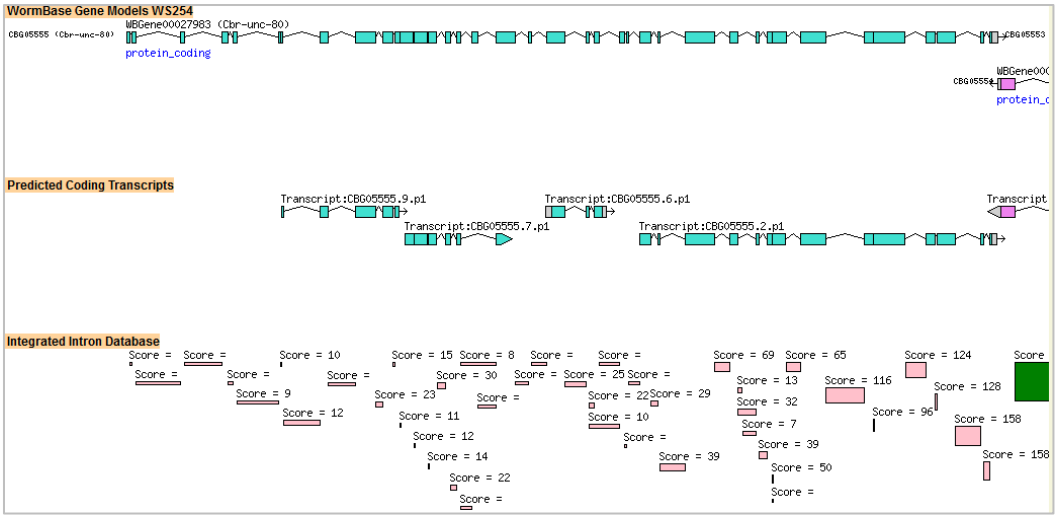

Figure 56. An example of fragmented assembled transcripts (CBG05555.X) in comparison to one long WB annotated transcript *Cbr-unc-80* (uncoordinated), an ortholog of *C. elegans' unc-80.*

3.4.2. Summary of evaluation and protein-coding transcripts integration

Table 6. Assembled protein-coding transcripts in 13 categories

No.	Category	Protein-coding transcripts	Protein-coding genes
1.	Complete match (WB confirmed)	8,080	8,055
2.	3' extension	316	287
3.	5' extension	753	687
4.	5' & 3' extension	26	25
5.	Intron overlapping internal exon	358	332
6.	Introns overlapping intron	217	205
7.	Alternative donor (5'ss)	777	746
8.	Alternative acceptor (3'ss)	882	810
9.	Alternative donor & acceptor	346	327
10.	Merging 2 or more genes	206	116
11.	Complex changes	2,245	1,517
12.	Novel	159	159
13.	Other – Single-exon (no intron)	120	95
	Other – Partial	10,220	5,304

6,285 transcripts from 11 categories (Category 2 to 12, excluding those in 'match' and 'other' categories, Table 6) were chosen as candidates to be incorporated into WB

protein-coding transcript set. Out of those, there are cases where the first CDSs generated by TransDecoder do not start with ATG (i.e., peptide does not start with Methionine). Specifically, 4,134 protein-coding transcripts (66%) start with ATG, while 2,151 of them (34%) do not. We assigned a correct start codon for each transcript that does not start with a proper start codon. We initially scanned for start codon upstream within the translation block of the current first start codon in the transcript. If no start codon was found upstream prior to hitting an upstream stop codon, a downstream search within the same translation block was performed. We hypothesized that most of them would get assigned to a proper start codon. Start codon was found in 6,262 (99.6%) of them (28% upstream, 71% downstream), and only 23 (0.4%) of them could not be assigned to a proper start codon.

Out of those that start with ATG, excluding those that do not end with a stop codon (TAA/TAG/TGA), 5,654 transcripts (89.96% of the candidate transcripts, Table 7) were integrated into the WormBase protein-coding transcript set generating an improved set of protein-coding transcripts composed of 28,129 transcripts. Both selected candidate transcript set (n=5,654) and improved transcript set (n=28,129) are used for comparative analysis with *C. elegans* in the next chapter.

Table 7. Candidate protein-coding transcripts with proper start and stop codons

No.	Category	Protein-coding transcripts	With proper start & stop codons
2.	3’ extension	316	284
3.	5’ extension	753	692
4.	5’ & 3’ extension	26	23
5.	Intron overlapping internal exon	358	341
6.	Introns overlapping intron	217	197
7.	Alternative donor (5’ss)	777	717
8.	Alternative acceptor (3’ss)	882	818
9.	Alternative donor & acceptor	346	284
10.	Merging 2 or more genes	206	179
11.	Complex changes	2,245	2,015
12.	Novel	159	104
	Total	6,285	5,654 (89.96%)

Additional analysis: Data availability limits introns, exons, and transcripts discovery but shows the potential of RNA-Seq to boost genome annotation

Due to the limited number of *C. briggsae* libraries compared to *C. elegans*, we performed an additional analysis to test the effect of data availability on results. Using only 13 *C. briggsae* RNA-Seq libraries, we defined a 22,209 fully supported protein-coding transcripts (i.e., supported by our intron and exon RNA-Seq databases) encoding 22,110 distinct ORFs including novel and improvable transcripts. For a comparison, we randomly picked 13 *C. elegans* RNA-Seq libraries and performed the same pipeline to obtain *C. elegans* RNA-Seq introns, exons, and supported protein-coding transcripts. In comparison to when 802 RNA-Seq libraries were used, the number of protein-coding transcripts from 13 *C. elegans* libraries identified was around the same as in *C. briggsae* (Figure 57).

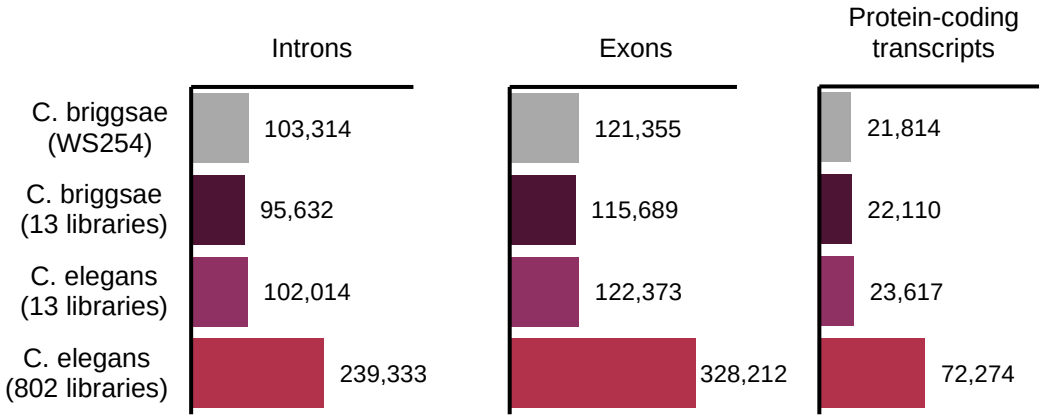

Figure 57. Comparison of introns, exons, and protein-coding transcripts identified using few (13) and many (802) RNA-Seq libraries in *C. briggsae* and *C. elegans.* The result of *C. elegans* 802 libraries was adopted from (Douglas, 2018).

Preliminary analysis: Spliced leader *trans*-splicing in *C. briggsae*

A preliminary analysis of *C. briggsae* SL *trans*-splicing was also performed in *C. briggsae*. We first predicted the putative SL *trans*-splicing acceptor sites using the *C. briggsae* reference genome and gene model release WS254. 96,420 putative acceptor sites were identified by finding AG sequences in the 100 bp window upstream of the start codon (ATG) of protein-coding sequences (Figure 58).

Focusing on sites that have at least 2 read support from the 13 *C. briggsae* RNA-Seq libraries, we identified 12,516 sites in 11,239 genes (10,352 sites in 9,279 genes

have support for SL1 and 2,164 sites in 1,960 genes have support for SL2). Out of the 12,516 sites, 9,750 (69.91%) sites are not annotated in WormBase WS254. At the site level, 8,457 sites have support for SL1, 1,895 sites have support for both SL1 and SL2, and 269 sites have support for SL2.

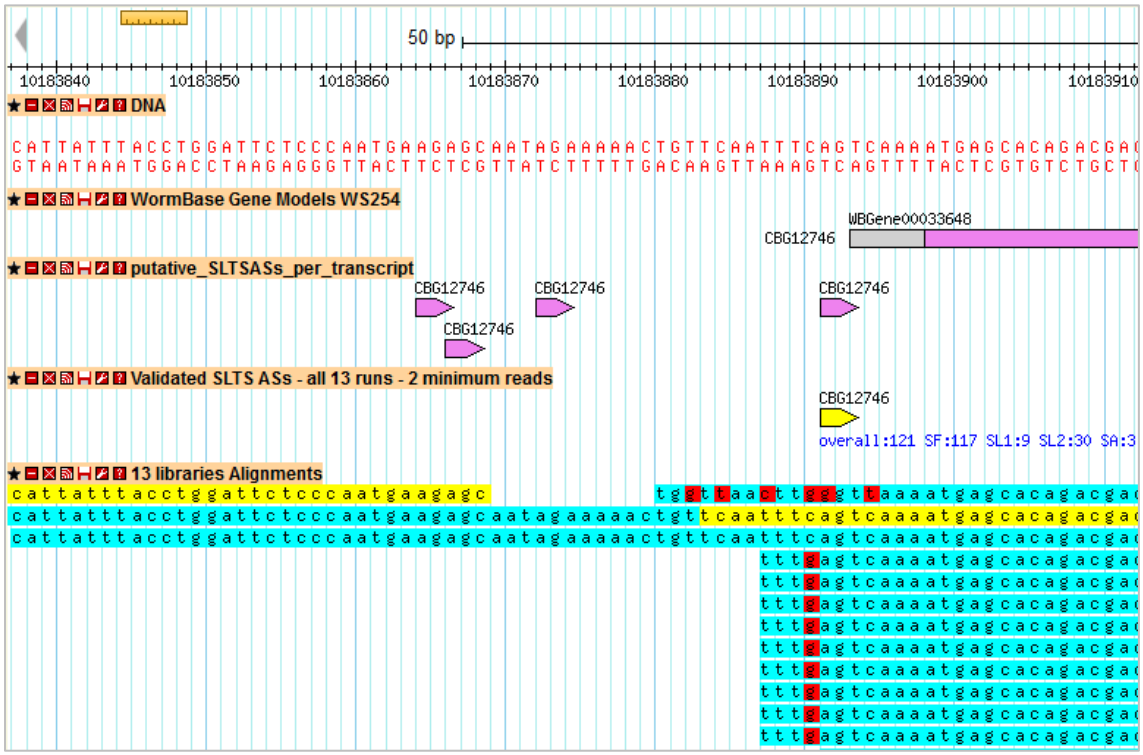

Figure 58. Putative SL *trans*-splicing acceptor sites were predicted using the *C. briggsae* reference genome and gene models WS254 (sites are located in the AGs in the 100 bp window upstream of ATG). Putative SLTS ASs were validated using the aligned RNA-Seq reads and the *C. briggsae* SL sequences (Table 8 below). Region: I:10,183,880-10,183,980. Overall represents overall read support, SF represents support for *trans*-splicing, SL1 represents support for SL1 *trans*-splicing, SL2 represents support for SL2 *trans*-splicing, and SA represents support against *trans*-splicing.

Table 8. *C. briggsae* spliced-leader *trans*-splicing sequences

Spliced Leader		Sequence
SL1		GGTTTAATTACCCAAGTTTGAG
SL2	Cb_SL2	GGTTTTAACCCAGTTACTCAAG
	Cb_SL3	GGTTTTAACCCAGTTAACCAAG
	Cb_SL4	GGTTTTAACCCAGTTTAACCAAG
	Cb_SL10	GGTTTTAACCCAAGTTAACCAAG
	Cb_SL13	GGATTTATCCCAGATAACCAAG
	Cb_SL14	GGTTTTTACCCTGATAACCAAG

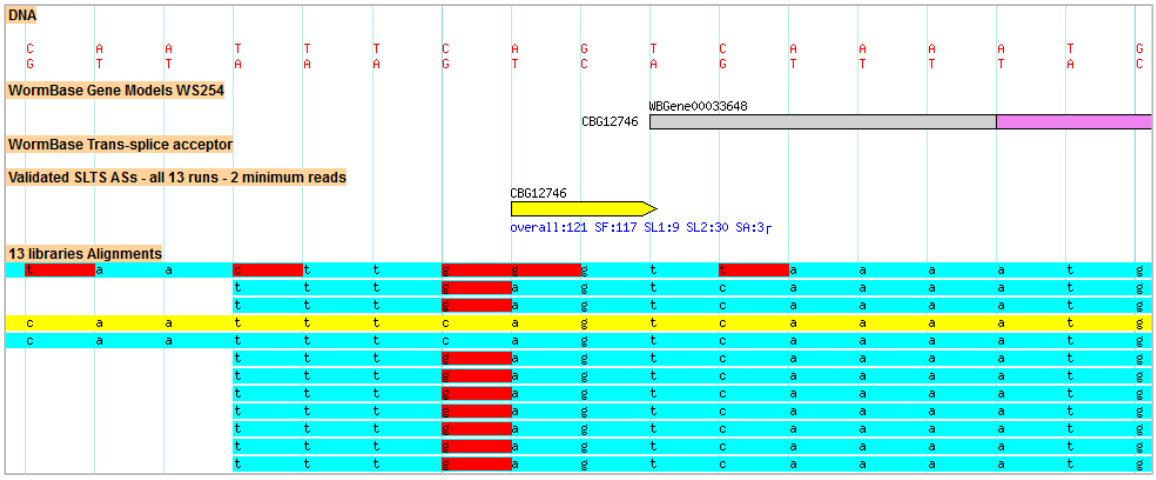

Figure 59. Illustration of SL *trans*-splicing acceptor sites that are not annotated in WormBase WS254.

3.5. Discussion

Having a more complete genome representation is necessary for many studies to enable more thorough genomic analyses. Inaccurate and/or absent gene models can impact both non-comparative and comparative studies. In this chapter, we improved *C. briggsae* annotation at the transcript level. To ensure the quality of assembled RNA-Seq protein-coding transcripts, we applied methods including filtering transcripts supported by gene features (introns, exons) and evaluating those that have coding potentials against WormBase coding transcripts. Candidate assembled RNA-Seq protein-coding transcripts were selected to be incorporated to the WormBase transcript set generating an improved *C. briggsae* transcript set. This improved transcript annotation, together with intron and exon annotations are useful for downstream analyses and are valuable resources for future studies by the scientific community.

We first reconstructed RNA-Seq transcripts using the combination of both genome-guided and *de novo* transcript assembly methods. Genome-guided transcript assembly method assembles transcripts from overlapping read alignments after alignment to the reference genome, while *de novo* transcript assembly creates short contigs from the overlapping reads independent of the reference genome. Both strategies offer their own advantages. *De novo* method is sensitive to sequencing artifacts and can recover previously unannotated transcripts that were missing from the genome assembly. Genome-guided assembly is very sensitive and can assemble transcripts of low abundance to fill gaps of previously annotated transcripts that can

result in full-length transcripts discovery. However, *de novo* method is time and resource-intensive, while genome-guided method result could be affected by the quality of the reference genome itself (i.e., a not-so-well annotated reference genome could contain misassemblies that could affect transcript reconstruction). Using a combination of both methods, we could enhance the recovery of transcripts and detection of novel transcripts (Lu et al., 2013; Martin and Wang, 2011). Transcripts assembled using both methods were filtered based on the presence of introns and exons in the databases to minimize assembly errors. This allowed us to identify 29,352 supported transcripts, including 24,705 (84%) that have coding potentials.

One limitation in performing transcriptome assembly from short RNA-Seq reads is that short reads rarely span across several splice junctions and thus it is challenging to accurately reconstruct full-length transcripts. During transcript assembly (mentioned above), fragments of transcripts from short RNA-Seq reads are computationally assembled to recover full-length isoforms. We applied a filtration step to select transcripts that are supported by introns and exons to provide reassurance that all the features in the transcripts are real. However, further validation of the correct combination of features (i.e., transcript structure) is essential and can be done in the future. Validation can be performed using long-read sequencing data (e.g., Iso-Seq data) that offers an advantage to identify full-length transcripts. At the time of writing, we were not able to incorporate long-read sequencing data due to the unavailability of such data for *C. briggsae*. Incorporating long-read sequencing data can positively impact the quality of the transcripts assembled (Amarasinghe et al., 2020).

In the transcript-to-transcript evaluation, we consider a gene model to be improvable when the RNA-Seq transcript partially matches the current annotated transcript and contains additional intron(s) relative to the annotated transcript (for instance, a novel intron that leads to a transcript extension). In an effort to select candidate protein-coding transcripts that may improve the current annotation, we developed a method to categorize transcripts that completely match and partially match the annotated WB transcripts and only keep those that are improvable (partially match and contain additional intron(s) relative to the annotated WB transcripts). Additionally, we consider a gene model to be novel when the RNA-Seq transcript is currently unannotated. By applying this approach, we were able to obtain 6,285 transcripts as candidates to improve *C. briggsae* annotation.

One limitation of the intron chain comparison approach is that single-exon transcripts were automatically filtered out based on the approach we used (details in Section 3.4.1). To overcome the limitation of single-exon transcripts identification, an additional check based on coding sequence regions could be performed in future studies to identify WB transcripts that should be revised to be multi-exon transcripts.

A functional protein-coding transcript should contain an Open Reading Frame (ORF) capable of being translated into a functional protein. An ORF begins with a start codon (AUG) and ends with a stop codon (UGA, UAA, UAG) in the same reading frame (Majoros et al., 2014). Due to this reason, we further reassured that all the candidates start and stop with the start codon and one of the stop codons, respectively. We assigned a proper start codon for those that do not start with one. This step is necessary as TransDecoder does not have a start-codon finding function and will include transcript from the beginning if there is no upstream in-frame stop codon at the beginning of the transcript (Haas, 2014, 2018). We also only keep the corrected transcripts that also end with a stop codon. After the start and stop codon assessment, 5,654 transcripts were integrated into the current annotated protein-coding transcript set. This integration resulted in an improved *C. briggsae* protein-coding transcript set composed of 28,129 transcripts (25.2% higher than current annotation) that offers refined gene structures and new gene models. Overall, this reveals a higher complexity of *C. briggsae* or higher coding capacity of *C. briggsae* genome compared to the current annotation.

Of our 5,654 protein-coding transcripts, those that overlap existing genes we identified may reflect errors in the WB annotated transcripts, or they may indicate unannotated alternative isoforms. The extensions categories (category 2-4, 17.7%) and merging category (category 10, 3.2%) suggest misannotations, while the other categories indicate alternative isoforms (category 5 to 9 and category 11, 77.3%). Misannotations are possible because most *C. briggsae* gene structures are based on computational predictions that often unsupported by experimental evidence (35.5% *C. briggsae* protein-coding transcripts are partially confirmed and 17.2% have no mRNA or EST evidence). The percentage of rare transcripts, length of transcripts, number of introns per transcripts that are misannotated were not being observed in this study. In either case, our results suggest that the existing gene model set can be improved.

In general, gene models can be improved in many ways. First, increasing RNA-Seq depth can support the identification of more splicing junctions. Second, long-read sequencing technology can reduce positional ambiguity during alignment, capture higher confidence splicing junctions, and provide full-length transcript structures. Third, sequencing on more comprehensive developmental stages and tissues can capture splicing events and transcripts that are rare globally (relative to total transcripts) but non-rare locally (relative to transcripts in individual stage or tissue).

Chapter 4. Homology and RNA-Seq based comparative analysis using the improved *C. briggsae* genome annotation

4.1. Introduction

The main goal of this thesis is to improve *C. briggsae* as a comparative tool for *C. elegans.* In the previous chapter, RNA-Seq data has allowed us to generate an improved annotation of gene models in *C. briggsae* genome. In this chapter, we perform comparative analyses to discover additional *C. briggsae-C. elegans* ortholog pairs and improve the *C. elegans* protein-coding transcript annotation.

4.2. Orthology analysis between *C. briggsae* and *C. elegans*

We assigned ortholog relationships between the two species before and after *C. briggsae* improvement using OrthoMCL version 2.0.9 (Li et al., 2003). The procedure was followed as outlined in the user guide (Fischer et al., 2011). In brief, OrthoMCL uses all-against-all BLASTP to calculate pairwise protein sequence similarities and obtain pairs of orthologous proteins. The tool further clusters the pairs into groups by using the MCL program. E-value cutoff of 1e-5 was used in the BLASTP step. Inputs were *C. briggsae* and *C. elegans* peptide sequences. For both species, the longest peptide (i.e., longest translated protein-coding transcript in a gene) was chosen when multiple transcript isoforms representing the same gene were found.

We identified 16,748 ortholog pairs in the first comparison prior to improvement, while the number increased slightly to 16,880 pairs after improvement (Table 9). We identified 14,778 ortholog pairs that were shared between the original and improved comparison. Out of 16,880 pairs from improved comparison, 1,894 pairs were modified, 132 pairs were novel, including 32 pairs that belong to novel *C. briggsae* genes/transcripts were identified (Table 10).

Table 9. Ortholog assignment results between *C. briggsae* and *C. elegans*

	***C. briggsae* original transcript set (longest, n=21,814)**	***C. briggsae* revised transcript set (longest, n=21,913)**
C. elegans transcript set (longest, n=20,254)	Ortholog pairs = 16,748	Ortholog pairs = 16,880

Table 10. New ortholog pairs from novel transcripts

C. briggsae	*C. elegans*	E-value	Percent identity[1]	Percent match[2]
nCBG00005	K04G2.12	2.00E-66	82.2	100
nCBG00013	W04A4.3	1.00E-44	50.4	89.9
nCBG00018	F56H1.10	6.00E-09	30.9	69.4
nCBG00024	Y53H1A.2b	5.00E-17	38.7	59
nCBG00026	F47H4.2b	1.00E-23	28.4	93.2
nCBG00031	ZK1127.13	3.00E-74	71.9	100
nCBG00033	ZK1127.3	2.00E-52	40.5	96.6
nCBG00035	F41E6.1	3.00E-08	28.1	93.6
nCBG00041	W09B6.2a	0	73.8	99.7
nCBG00044	C15F1.4	0	78	100
nCBG00047	C09D8.1j	0	44.6	91.9
nCBG00050	F44F4.9	2.00E-31	46.2	98.3
nCBG00054	T01H3.5	8.00E-47	52.9	100
nCBG00060	C35D10.17	3.00E-70	89.2	100
nCBG00062	T03F6.9	3.00E-64	72.3	88.8
nCBG00075	Y75B8A.19	8.00E-30	26.2	84.4
nCBG00076	Y6D1A.1	1.00E-20	30.6	50.7
nCBG00085	Y105C5B.18	3.00E-11	27	88.7
nCBG00092	C43F9.11a	7.00E-90	80.7	99.5
nCBG00103	H35N09.1	3.00E-12	30.2	86.6
nCBG00107	F53F4.18	5.00E-18	54	74.7
nCBG00112	Y6G8.14	2.00E-07	22.4	93.2
nCBG00117	F21F8.5	1.00E-25	45.7	91.3
nCBG00118	F19F10.4	1.00E-39	42.8	98.7
nCBG00121	C45B11.8	1.00E-18	38.5	99.1
nCBG00132	W04G3.13	1.00E-50	61.3	100
nCBG00138	F09A5.9	2.00E-55	57.3	99.2
nCBG00146	C07A12.18a	1.00E-74	74.3	99.3
nCBG00147	K02G10.15	5.00E-56	88	100
nCBG00151	ZK662.6	3.00E-102	70.1	100
nCBG00153	Y51A2B.4	1.00E-82	78	83.5
nCBG00158	F31E3.12	3.00E-39	50.4	98.4

[1] Percentage of identical characters in each sequence.

[2] Fraction of aligned regions, based on the shorter sequence.

4.3. *C. elegans* gene models improvement using the improved *C. briggsae* genome annotation

4.3.1. Predicting *C. elegans* gene models

Using the selected candidate *C. briggsae* transcripts (n=5,654), we predicted 4,313 *C. elegans* protein-coding transcripts using Gene Model Mapper (GeMoMa) version 1.6.1 (Keilwagen et al., 2018). All of the predicted transcripts begin with the start codon and end with a stop codon. We assigned *C. elegans* gene and transcript names to the GeMoMa predicted transcripts due to lack of information about the transcript origin (i.e., which transcripts/genes they overlap with) prior to transcripts evaluation.

Figure 60. Illustration of GeMoMa predicted *C. elegans* protein-coding transcripts (*C. elegans'* C49D10.2 or *nhr-166* predicted from *C. briggsae* CBG23578). GBrowse track: (1) Annotated *C. elegans* WormBase WS254 gene models, (2) GeMoMa predicted *C. elegans* transcript.

4.3.2. Evaluating the predicted *C. elegans* transcripts

We further evaluate the tagged predicted coding-transcripts by applying the same custom Python algorithm as described in Section 3.4.1, that is, comparing every predicted coding-transcript to their corresponding annotated WormBase *C. elegans* protein-coding transcript (release WS254). For *C. elegans* genes that are alternatively spliced, we picked the longest protein-coding transcript (longest annotated CDS chain) to represent an annotated gene for the comparison. The evaluation result can be found in Table 11.

Table 11. Predicted GeMoMa *C. elegans* protein-coding transcripts in 13 categories

No.	Category	Protein-coding transcripts
1.	Complete match (WB confirmed)	1,425
2.	3' extension	80
3.	5' extension	127
4.	5' & 3' extension	1
5.	Intron overlapping internal exon	186

6.	Introns overlapping intron	43
7.	Alternative donor (5'ss)	229
8.	Alternative acceptor (3'ss)	233
9.	Alternative donor & acceptor	138
10.	Merging 2 or more genes	27
11.	Complex changes	549
12.	Novel	85
13.	Other – Single-exon (no intron)	30
	Other – Partial	1,160

The evaluation resulted in 1,698 predicted GeMoMa transcripts as candidate new isoforms for *C. elegans* (Category 2 to 12). These candidates were further validated using an independent set of introns from 802 *C. elegans* RNA-Seq libraries. This intron database consists of 239,334 introns with minimum intron support of 5 in at least one library (Douglas, 2018). 281 of the candidate transcripts have all their introns supported by RNA-Seq introns (Table 12, Figure 61 to Figure 65). Random and manual sampling of those 281 cases suggested that some of them are also supported by long-read alignments from WormBase.

Table 12. Predicted GeMoMa *C. elegans* protein-coding transcripts in Category 2 to 12 that are supported by RNA-Seq introns from 802 libraries

No.	Category	Protein-coding transcripts	Supported by RNA-Seq introns from 802 libraries
2.	3' extension	80	13
3.	5' extension	127	35
4.	5' & 3' extension	1	0
5.	Intron overlapping internal exon	186	9
6.	Introns overlapping intron	43	14
7.	Alternative donor (5'ss)	229	48
8.	Alternative acceptor (3'ss)	233	83
9.	Alternative donor & acceptor	138	10
10.	Merging 2 or more genes	27	8
11.	Complex changes	549	59
12.	Novel	85	2
	Total	1,698	281 (16.6%)

Below are some representatives of well-supported *C. elegans* transcripts: 3' transcript extension (Figure 61), transcript with a novel intron internal of an annotated CDS (Figure 62), transcript merging (Figure 63), transcript with a combination of multiple modifications (multiple alternative splicing events, Figure 64), and novel genes (Figure 65). All representative figures were displayed using GBrowse showing WormBase *C. elegans* protein-coding transcript, predicted *C. elegans* transcripts, and *C. elegans* introns from 802 libraries, respectively.

Figure 61. (top) An example of a predicted transcript (CBG23739.2_R0) with one additional intron extending 3' of the *C. elegans* annotated transcript (*bro-1,* brother (drosophila tx factor partner) homolog). All introns in WormBase F56A3.5 (*bro-1*) transcript are observed in the predicted transcript. One more intron with high support was observed, suggesting an extension of the gene model at the 3' end; (bottom) The introns are also supported by long-read alignments (source: WormBase Jbrowse, as of April 2020).

Figure 62. (top) An example of a predicted transcript (CBG17297.2_R0) with one additional intron internal of exon compared to the *C. elegans* annotated transcript (*trpp-9,* transport protein particle). All introns in WormBase C35C5.6 transcript are observed in the predicted transcript. One more intron with 20,385 support was observed (arrow), suggesting internal intron overlapping internal exon; (bottom) The intron is also supported by long-read alignments (source: WormBase Jbrowse, as of April 2020).

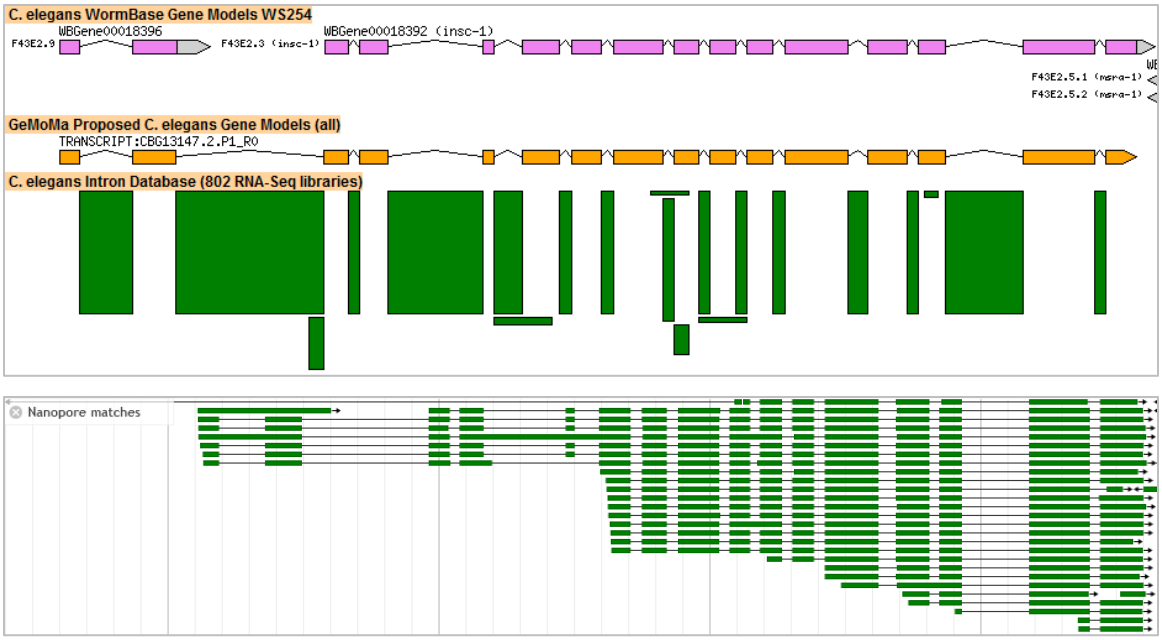

Figure 63. (top) An example of a predicted transcript (CBG13147.2.P1_R0) merging two annotated *C. elegans* transcripts (*F43E2.9* and *insc-1,* inscuteable “drosophila asymmetric cell division protein” homolog). One intron in between the two genes with 14,530 support was observed; (bottom) That second intron is also supported by long-read alignments (source: WormBase Jbrowse, as of December 2019).

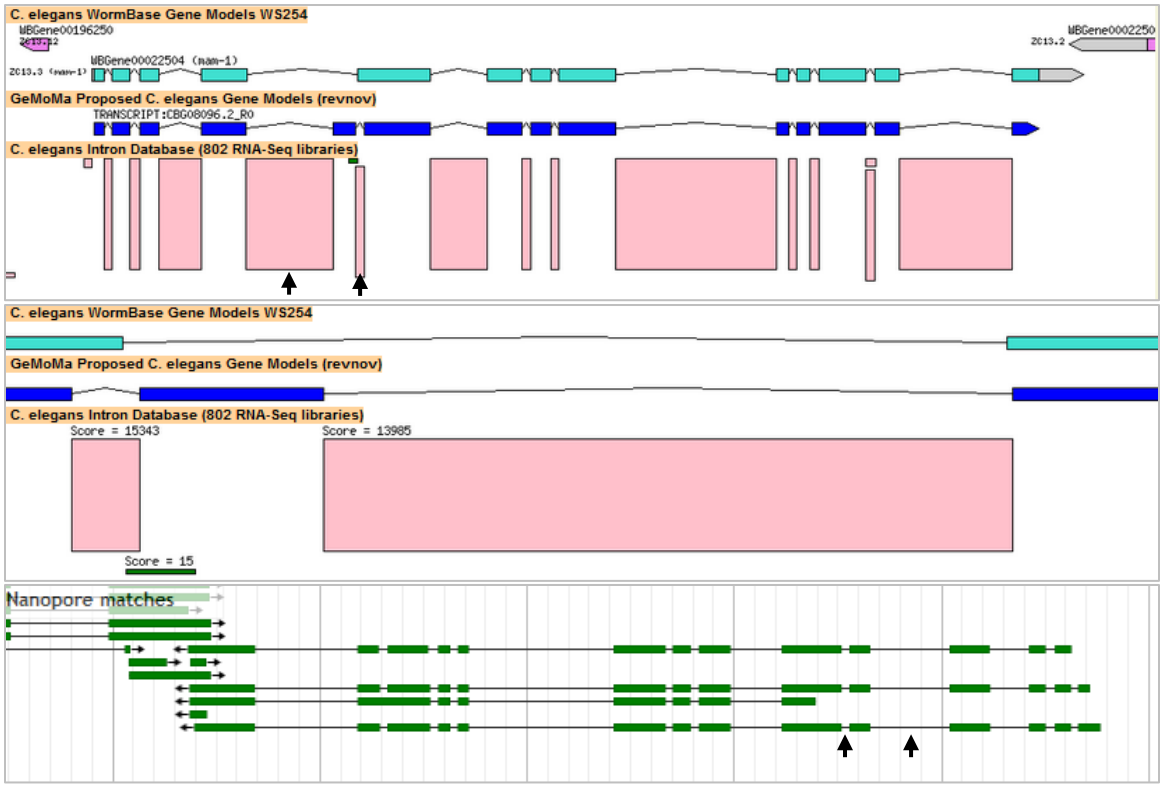

Figure 64. (top) An example of a predicted transcript (CBG08096.2_R0) with additional introns in comparison to the annotated *C. elegans* transcript (*mam-1,* mam (meprin, A5-protein, PTPmu) domain protein) suggesting a combination of multiple modifications (arrows); (middle) zoomed region of the transcript that contain the modifications; (bottom) The introns are also supported by long-read alignments (source: WormBase Jbrowse, unflipped strand, as of April 2020).

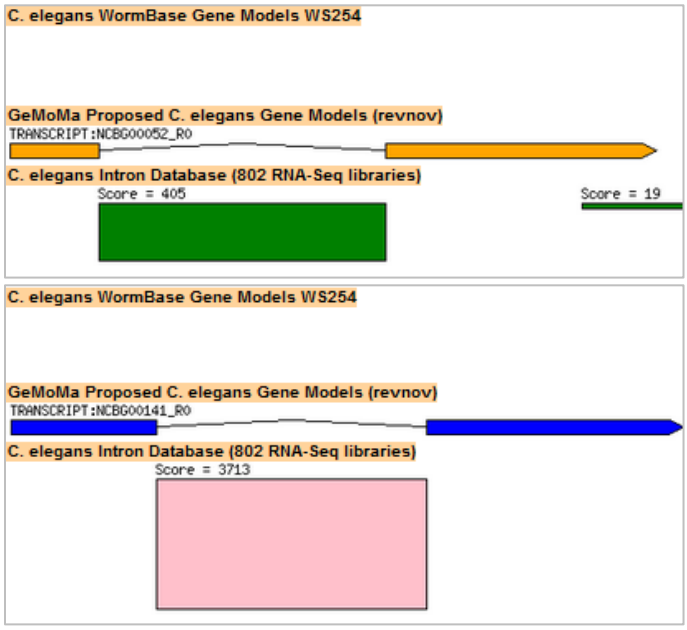

Figure 65. Predicted transcripts suggesting novel genes. (top) NCBG00052_R0, region II:5788171-5788381+; (bottom) NCBG00141_R0, region X:17566822-17566984-.

4.4. Discussion

When the genome sequence of a closely related organism is available, comparison of the two sequences can be a very powerful tool for that species. To assess the impacts of the *C. briggsae* improved annotation on *C. elegans*, we performed ortholog assignment using OrthoMCL and predicted the *C. elegans* gene models using GeMoMa.

At the time of writing, several tools have been developed to assign ortholog relationships in eukaryotes. OrthoMCL was chosen because it was one of the best performing graph-based ortholog detection tools for eukaryotes including *C. elegans* (Chen et al., 2007; Li et al., 2003). This tool has also been used to identify orthologous gene clusters in *Caenorhabditis*, including *C. briggsae* and *C. elegans* (Fierst et al., 2015; Kraus et al., 2017). An E-value cut-off of 1e-5 was used in the BLASTP step of OrthoMCL to avoid assigning paralogs to ortholog pairs and therefore generating a high-confidence set of ortholog pairs (Stein et al., 2003; Strange, 2006). In BLAST result, E-value is defined as the number of distinct alignments with scores greater than or equal to a given value expected to occur in a search against a database of known size, based solely on chance, not homology. Large E-values suggest that the query sequence and retrieved sequence similarities are due to chance, while small E-values suggest that the sequence similarities are due to shared ancestry (or potentially convergent evolution) (Kerfeld and Scott, 2011). Bit score represents a statistical significance of an alignment. Therefore, the lower the E-value is, the more significant the score and the alignment is; the higher the bit-score, the more similar the two sequences are.

Our OrthoMCL result demonstrated that modified and novel transcripts identified by RNA-Seq method led to the identification of additional ortholog pairs between the two species. Specifically, 132 more ortholog pairs are found after revision (Table 9), 32 of which belong to new *C. briggsae* transcripts suggesting new ortholog assignments between the two species (Table 10). The rest of the novel transcripts that are not assigned to *C. elegans* may predict new genes or transcripts in *C. elegans.* Furthermore, 1,894 pairs were modified. For instance, due to an alternatively 5’ splice usage in transcript CBG03308, we previously identified another transcript for *C. briggsae*, named CBG03308.2 which was incorporated into the improved transcript set. Transcript CBG03308.2 was chosen to represent the gene because it is longer than CBG03308.

From OrthoMCL result, CBG03308:mdt-10 pair has lower E-value (3e-85) and higher bit score (241) compared to those of CBG03308.2:mdt-10 pair (4e-86 and 244, respectively) suggesting a better ortholog pair between the two species.

The use of multiple ortholog prediction tools, including those that are more recent (for example, OrthoFinder) can be considered in future studies. OrthoFinder (Emms and Kelly, 2019), was recently updated and found to have the highest ortholog inference accuracy across 66 species including eukaryotes. This tool involves phylogenetic tree to clarify orthology relationships that may contain false positives or false negatives that are identified using pairwise sequence similarity approach applied in OrthoMCL. OrthoFinder also requires minimum computation by using only a single-command pipeline compared to the 13-step OrthoMCL pipeline. Using multiple algorithms (meta-methods), orthologs could be more confidently declared as true orthologs, for example, by finding intersections from all tools or setting a threshold of acceptance (when three out of the five tools agree with each other). These decisions are dependent on the purpose of the research (Glover et al., 2019).

One approach to improve the understanding of a genome of interest is through computational gene prediction (i.e., finding the location of protein-coding regions) using a genome from a closely-related organism. *Ab initio* and homology-based searches are two categories of gene prediction methods. The former is a method based on gene structure and signals, such as splice sites, branchpoint, polypyrimidine tract, start and stop codons. The latter is an approach based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome, which relies on the assumption that functional regions (exons) are more conserved evolutionarily than non-functional regions (intergenic or intronic regions) (Wang et al., 2004). The *ab initio* methods are usually sensitive in finding genes in novel genomes but often produce many false positives. The homology-based methods usually have higher specificity and produce fewer false positives than *ab initio* methods because they are based on biological evidence (homology) to existing genes (She, 2010).

We used GeMoMa (Keilwagen et al., 2018), a homology-based gene prediction program that uses the annotation of protein-coding genes in a reference genome to infer the annotation of protein-coding genes in a target genome. Specifically, GeMoMa uses amino acid sequence and intron position conservations to predict gene models in

species of interest (in this thesis, *C. elegans*). Out of other gene prediction tools, GeMoMa is able to predict transcripts with many exons and lower PID more accurately compared to other tools that do not exploit intron position conservation such as genBlastG and exonerate (Keilwagen et al., 2016). The predictions were further evaluated by comparing them to the annotated *C. elegans* gene models to obtain candidate *C. elegans* gene models. Similar to the previous approach, we categorized transcripts that completely match the annotated WB transcripts, partially match the annotated WB transcripts, do not contain an intron (single-exon transcript), and novel. Our result suggested that the improved set of transcripts could identify thousands of additional *C. elegans* gene models. However, incorrect predictions/annotations could negatively impact downstream analysis. We therefore applied RNA-Seq evidence validation by selecting only *C. elegans* candidate protein-coding transcripts whose introns are all represented in the high-quality intron database from 802 RNA-Seq libraries (Douglas, 2018) to reduce false positives. This enabled us to identify a set of well-supported *C. elegans* predicted transcripts. Random sampling of these transcripts suggested that some of them are also supported by long-read alignments.

The approach we took, selecting only *C. elegans* candidate protein-coding transcripts that are fully supported by *C. elegans* introns, was very strict and highly specific towards selecting those that are well-supported by introns (i.e, real or true positives). High specificity is beneficial because it would filter those that are falsely predicted (false positives), however, this offers a trade-off to the correctly predicted transcripts (true positives). Additionally, it is challenging to judge whether those transcripts that did not meet the criteria are false positives or true positives. Because GeMoMa is based on homology, we expected that those that are predicted have a certain level of conservation between the two species. Observing the level of conservation of novel introns or exons can be considered for future studies to reduce false negatives. Another solution that could be applied in the future is to perform gene predictions using additional gene prediction tools and obtain consensus predicted transcripts (Allen et al., 2004).

Overall, current *C. briggsae* annotation could still be improved and we demonstrated that the improved *C. briggsae* gene models from limited RNA-Seq data is valuable for improving *C. elegans* annotation. With more RNA-Seq data in the future (not limited to short-read RNA-Seq data), we could produce more improvement for not only

C. briggsae but also for *C. elegans*. For instance, long-read sequencing technology can provide a more comprehensive evidence of gene structures because most reads likely correspond to full-length transcripts. The use of long-read sequencing data (for example, PacBio Iso-Seq data) with or without integration with short-read sequencing data have shown significant improvements on genome annotations (Beiki et al., 2019; Magrini et al., 2018; Wang et al., 2016, 2019).

Chapter 5. Conclusion and Future Directions

5.1. Conclusion

In this thesis, RNA-Seq provided evidence that the *C. briggsae* genome annotation is currently incomplete. Despite limited RNA-Seq data, we have revealed thousands of introns, exons, and protein-coding transcripts that suggest gene model corrections and additional transcript isoforms in *C. briggsae*. By integrating those features to the current annotated features, we have improved the *C. briggsae* annotation at the intron, exon, and transcript level. Because the feature discovery power is proportional to RNA-Seq data quantity, additional *C. briggsae* features are likely to be observed as additional RNA-Seq data being generated in the future.

This study has also improved the utility of *C. briggsae* as a comparative platform for *C. elegans*. We have demonstrated that the improved *C. briggsae* annotation together with comparative analyses is beneficial for improving *C. elegans* annotation. Specifically, we have found new ortholog relationships between the two species and identified gene models that could be improved in addition to new gene models in *C. elegans*.

Taken together, it can be concluded that our RNA-Seq based annotation has improved both *C. briggsae* and *C. elegans* annotations.

5.2. Future directions

Despite the findings of thousands of introns, exons, and protein-coding transcripts in *C. briggsae*, we believe that the annotation of this organism can be further improved by more experimental evidence (not limited to RNA-Seq data) being generated. Performing tissue-specific and stage-specific sequencing analyses can be beneficial to discover more or confirm features that are only expressed at certain stages or tissues. Integrating data from short and long-sequencing technologies or utilizing long-sequencing technology alone on different developmental stages and tissues could lead to a more comprehensive genome annotation. Moreover, analysis of SL *trans*-splicing acceptor sites can be further explored.

RNA-Seq is powerful for annotating transcripts, but transcript fragments from short RNA-Seq reads need to be computationally assembled to recover full-length transcripts. Although we reassured that our RNA-Seq assembled transcripts are supported by introns and exons in our databases, a validation step using long-read sequencing data can be performed in the future to confirm the correct combination of features assembled by our assemblers.

In the pipeline where we evaluated those RNA-Seq assembled transcripts, our intron chain comparison method is not applicable for single-exon transcripts. An additional approach to compare the coding sequence chains would be beneficial for capturing and/or improving the annotation of such transcripts. By doing so, we could provide a more comprehensive revision for single-exon transcripts that are supposed to be multi-exon transcripts.

Regarding ortholog prediction, although OrthoMCL is one of the most commonly used tools, the use of a more recent tool or a combination of tools (meta-methods) can be considered in future studies. The approach of selecting consensus ortholog pairs or filtering results that are found by a specific number of tools can increase the robustness of the predictions by compensating for deficiencies that each individual method might have.

In the *C. elegans* candidate protein-coding transcripts evaluation, we only select those that are fully supported by *C. elegans* introns. Further approach to use multiple gene prediction tools to obtain consensus predicted transcript structures or to observe the conservation of novel features involved can be considered to compromise our ‘strict’ filtering method.

Even though *C. elegans* genome has been extensively annotated using a number of computational and experimental approaches, our study could still improve its annotation. We believe that the methods we applied in this study can be applied to other organisms that have reference genome and RNA-Seq data available, including those that have been extensively annotated, such as mouse and human.

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. Nat. Commun. *7*, 11706.

Adami, C., Ofria, C., and Collier, T.C. (2000). Evolution of biological complexity. Proc. Natl. Acad. Sci. U. S. A. *97*, 4463–4468.

Allen, J.E., Pertea, M., and Salzberg, S.L. (2004). Computational Gene Prediction Using Multiple Sources of Evidence. Genome Res. *14*, 142–148.

Allen, M.A., Hillier, L.W., Waterston, R.H., and Blumenthal, T. (2011). A global analysis of C. elegans trans-splicing. Genome Res. *21*, 255–264.

Altun, Z.F., and Hall, D.H. (2009). Hermaphrodite Introduction.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biol. *21*, 30.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.

Ankeny, R.A. (2001). The natural history of Caenorhabditis elegans research. Nat. Rev. Genet. *2*, 474–479.

Armstrong, J., Fiddes, I.T., Diekhans, M., and Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. Annu. Rev. Anim. Biosci. *7*, 41–64.

Au, K.F., Jiang, H., Lin, L., Xing, Y., and Wong, W.H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. *38*, 4570–4578.

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T.P.L., Reecy, J.M., and Tuggle, C.K. (2019). Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. BMC Genomics *20*, 344.

Bhagavan, N.V., and Ha, C.-E. (2015). Chapter 24 - Regulation of Gene Expression. In Essentials of Medical Biochemistry (Second Edition), N.V. Bhagavan, and C.-E. Ha, eds. (San Diego: Academic Press), pp. 447–464.

Bhasin, M., and Raghava, G.P.S. (2006). 8 - Computational Methods in Genome Research. In Applied Mycology and Biotechnology, D.K. Arora, R.M. Berka, and G.B. Singh, eds. (Elsevier), pp. 179–207.

Blencowe, B.J., and Graveley, B.R. (2008). Alternative Splicing in the Postgenomic Era (Springer Science & Business Media).

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.

Boguski, M.S., Lowe, T.M.J., and Tolstoshev, C.M. (1993). dbEST — database for "expressed sequence tags." Nat. Genet. *4*, 332–333.

Boguski, M.S., Tolstoshev, C.M., and Bassett, D.E. (1994). Gene Discovery in dbEST. Science *265*, 1993–1994.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Braendle, C., and Félix, M.-A. (2006). Sex Determination: Ways to Evolve a Hermaphrodite. Curr. Biol. *16*, R468–R471.

Brenner, S. (1974). The Genetics of Caenorhabditis Elegans. Genetics *77*, 71–94.

Brown, T.A. (2006). Genomes (Garland Science).

Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M., et al. (2014). Diversity and dynamics of the Drosophila transcriptome. Nature *512*, 393–399.

Bruno, V.M., Wang, Z., Marjani, S.L., Euskirchen, G.M., Martin, J., Sherlock, G., and Snyder, M. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. Genome Res. *20*, 1451–1458.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Carroll, S.B. (2001). Chance and necessity: the evolution of morphological complexity and diversity. Nature *409*, 1102–1109.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. *34*, 247–278.

Chalfie, M., Tu, Y., and Prasher, D. (1993). Glow Worms - A New Method of Looking at C. elegans Gene Expression.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., and Prasher, D.C. (1994). Green Fluorescent Protein as a Marker for Gene Expression. Science *263*, 802–805.

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. PLOS ONE *2*, e383.

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., Mah, A., Blacque, O.E., Chu, J., Phgora, K., Bakhoum, M.W., Newbury, C.R.H., Khattra, J., Chan, S., Go, A., et al. (2006). Identification of ciliary and ciliopathy genes in Caenorhabditis elegans through comparative genomics. Genome Biol. *7*, R126.

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*, 131.

Cisne, J.L. (1974). Evolution of the World Fauna of Aquatic Free-Living Arthropods. Evolution *28*, 337–366.

Clark, D.V., Suleman, D.S., Beckenbach, K.A., Gilchrist, E.J., and Baillie, D.L. (1995). Molecular cloning and characterization of the dpy-20 gene of Caenorhabditis elegans. Mol. Gen. Genet. MGG *247*, 367–378.

Coghlan, A., Stajich, J.E., and Harris, T.W. (2006). Comparative genomics in C. elegans, C. briggsae, and other Caenorhabditis species. Methods Mol. Biol. Clifton NJ *351*, 13–29.

Coghlan, A., Fiedler, T.J., McKay, S.J., Flicek, P., Harris, T.W., Blasiar, D., nGASP Consortium, and Stein, L.D. (2008). nGASP--the nematode genome annotation assessment project. BMC Bioinformatics *9*, 549.

Corsi, A.K., Wightman, B., and Chalfie, M. (2015). A Transparent Window into Biology: A Primer on Caenorhabditis elegans. Genetics *200*, 387–407.

Delhomme, N., Mähler, N., Schiffthaler, B., Sundell, D., Hvidsten, T.R., and Street, N.R. (2014). Guidelines for RNA-Seq data analysis. Epigenesys Protoc. *67*, 24.

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. Bioinforma. Oxf. Engl. *29*, 15–21.

Douglas, M. (2018). Quantitative analysis of the coding capacity of C. elegans using RNA-Seq data. Thesis. Science: Department of Molecular Biology and Biochemistry.

Elliott, D. (2011). Molecular biology of RNA / David Elliott, Michael Ladomery. (Oxford ; New York: Oxford University Press).

Ellis, H.M., and Horvitz, H.R. (1986). Genetic control of programmed cell death in the nematode C. elegans. Cell *44*, 817–829.

Ellis, R.E., Yuan, J.Y., and Horvitz, H.R. (1991). Mechanisms and functions of cell death. Annu. Rev. Cell Biol. *7*, 663–698.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

Félix, M.-A. (2004). Genomes: A Helpful Cousin for Our Favourite Worm. Curr. Biol. *14*, R75–R77.

Félix, M.-A., and Duveau, F. (2012). Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae. BMC Biol. *10*, 59.

Fierst, J.L., Willis, J.H., Thomas, C.G., Wang, W., Reynolds, R.M., Ahearne, T.E., Cutter, A.D., and Phillips, P.C. (2015). Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. PLoS Genet. *11*.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., and Stoeckert, C.J. (2011). Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. Curr. Protoc. Bioinforma. *35*, 6.12.1-6.12.19.

Gamazon, E.R. (2016). Alternative Splicing and Genome Evolution. In ELS, (American Cancer Society), pp. 1–6.

Garcia-Blanco, M.A., Baraniak, A.P., and Lasda, E.L. (2004). Alternative splicing in disease and therapy. Nat. Biotechnol. *22*, 535.

Gaudet, J., Muttumu, S., Horner, M., and Mango, S.E. (2004). Whole-genome analysis of temporal gene expression during foregut development. PLoS Biol. *2*, e352.

Gerhard, D.S., Wagner, L., Feingold, E.A., Shenmen, C.M., Grouse, L.H., Schuler, G., Klein, S.L., Old, S., Rasooly, R., Good, P., et al. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res. *14*, 2121–2127.

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.

Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S.K., Gabaldón, T., Huerta-Cepas, J., Martin, M.-J., Muffato, M., Patricio, M., Pereira, C., et al. (2019). Advances and Applications in the Quest for Orthologs. Mol. Biol. Evol. *36*, 2157–2164.

Gochnauer, M.B., and McCoy, E. (1954). Response of a soil nematode, Rhabditis briggsae, to antibiotics.

Gupta, B.P., Johnsen, R., and Chen, N. (2007). Genomics and biology of the nematode Caenorhabditis briggsae (WormBook).

Haas, B. (2014). TranscriptDecoder, Extracting likely coding regions from transcript sequences.

Haas, B. (2018). sequence stretch before start codon · Issue #66 · TransDecoder/TransDecoder.

Harbers, M., and Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. Nat. Methods *2*, 495–502.

Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E., and Waterston, R.H. (2005). Genomics in C. elegans: So many genes, such a little worm. Genome Res. *15*, 1651–1660.

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.

Horvitz, H.R. (2003). Worms, Life, and Death (Nobel Lecture). ChemBioChem *4*, 697–711.

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.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141–2144.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. *44*, e89.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics *19*.

Kent, W.J., and Zahler, A.M. (2000). Conservation, Regulation, Synteny, and Introns in a Large-scale C. briggsae–C. elegans Genomic Alignment. Genome Res. *10*, 1115–1125.

Kerfeld, C.A., and Scott, K.M. (2011). Using BLAST to Teach “E-value-tionary” Concepts. PLOS Biol. *9*, e1001014.

Kiontke, K. (2005). The phylogenetic relationships of Caenorhabditis and other rhabditids. WormBook.

Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H.A. (2004). Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc. Natl. Acad. Sci. U. S. A. *101*, 9003–9008.

Kirouac, M., and Sternberg, P.W. (2003). cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of Caenorhabditis elegans and C. briggsae. Dev. Biol. *257*, 85–103.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. (2006). CAGE: cap analysis of gene expression. Nat. Methods *3*, 211–222.

Kohara, Y. (1996). [Large scale analysis of C. elegans cDNA]. Tanpakushitsu Kakusan Koso. *41*, 715–720.

König, S., Romoth, L., and Stanke, M. (2018). Comparative Genome Annotation. In Comparative Genomics: Methods and Protocols, J.C. Setubal, J. Stoye, and P.F. Stadler, eds. (New York, NY: Springer New York), pp. 189–212.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. *39*, 309–338.

Korpelainen, E., Tuimala, J., Somervuo, P., Huss, M., and Wong, G. (2014). RNA-seq Data Analysis: A Practical Approach (Chapman and Hall/CRC).

Krämer, A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. Annu. Rev. Biochem. *65*, 367–409.

Kraus, C., Schiffer, P.H., Kagoshima, H., Hiraki, H., Vogt, T., Kroiher, M., Kohara, Y., and Schierenberg, E. (2017). Differences in the genetic control of early egg development and reproduction between C. elegans and its parthenogenetic relative D. coronatus. EvoDevo *8*, 16.

Kuzniar, A., van Ham, R.C.H.J., Pongor, S., and Leunissen, J.A.M. (2008). The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. TIG *24*, 539–551.

Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stamm, L., Vandenhaute, J., Hill, D.E., et al. (2004). C. elegans ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. Genome Res. *14*, 2064–2069.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843–854.

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., and Zhou, R. (2016). Evolutionary Insights into RNA trans-Splicing in Vertebrates. Genome Biol. Evol. *8*, 562–577.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. *13*, 2178–2189.

Loraine, A.E., McCormick, S., Estrada, A., Patel, K., and Qin, P. (2013). RNA-Seq of Arabidopsis Pollen Uncovers Novel Transcription and Alternative Splicing. Plant Physiol. *162*, 1092–1109.

Lu, B., Zeng, Z., and Shi, T. (2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci. China Life Sci. *56*, 143–155.

Magrini, V., Gao, X., Rosa, B.A., McGrath, S., Zhang, X., Hallsworth-Pepin, K., Martin, J., Hawdon, J., Wilson, R.K., and Mitreva, M. (2018). Improving eukaryotic genome annotation using single molecule mRNA sequencing. BMC Genomics *19*, 172.

Majoros, W.H., Lebeck, N., Ohler, U., and Li, S. (2014). Improved transcript isoform discovery using ORF graphs. Bioinformatics *30*, 1958–1964.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. *12*, 671–682.

McShea, D.W. (1996). Perspective: Metazoan Complexity and Evolution: Is There a Trend? Evolution *50*, 477–492.

Memar, N., Schiemann, S., Hennig, C., Findeis, D., Conradt, B., and Schnabel, R. (2019). Twenty million years of evolution: The embryogenesis of four Caenorhabditis species are indistinguishable despite extensive genome divergence. Dev. Biol. *447*, 182–199.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science *320*, 1344–1349.

Nagalakshmi, U., Waern, K., and Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. Curr. Protoc. Mol. Biol. *Chapter 4*, Unit 4.11.1-13.

Nesbitt, M.J., Moerman, D.G., and Chen, N. (2010). Identifying novel genes in C. elegans using SAGE tags. BMC Mol. Biol. *11*, 96.

Nigon, V., and Dougherty, E.C. (1949). Reproductive patterns and attempts at reciprocal crossing of Rhabditis elegans maupas, 1900, and Rhabditis briggsae Dougherty and nigon, 1949 (Nematoda: Rhabditidae). J. Exp. Zool. *112*, 485–503.

Parenteau, J., Durand, M., Véronneau, S., Lacombe, A.-A., Morin, G., Guérin, V., Cecez, B., Gervais-Bird, J., Koh, C.-S., Brunelle, D., et al. (2008). Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function. Mol. Biol. Cell *19*, 1932–1941.

Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V., and Elela, S.A. (2019). Introns are mediators of cell response to starvation. Nature *565*, 612–617.

Park, J.W., and Graveley, B.R. (2007). Complex alternative splicing. Adv. Exp. Med. Biol. *623*, 50–63.

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. Nat. Biotechnol. *33*, 290–295.

Poulin, G., Nandakumar, R., and Ahringer, J. (2004). Genome-wide RNAi screens in Caenorhabditis elegans: impact on cancer research. Oncogene *23*, 8340–8345.

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.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901–906.

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. Nat. Methods *7*, 909–912.

Ross, J.A., Koboldt, D.C., Staisch, J.E., Chamberlin, H.M., Gupta, B.P., Miller, R.D., Baird, S.E., and Haag, E.S. (2011). Caenorhabditis briggsae Recombinant Inbred Line Genotypes Reveal Inter-Strain Incompatibility and the Evolution of Recombination. PLOS Genet. *7*, e1002174.

Rossell, D., Stephan-Otto Attolini, C., Kroiss, M., and Stöcker, A. (2014). QUANTIFYING ALTERNATIVE SPLICING FROM PAIRED-END RNA-SEQUENCING DATA. Ann. Appl. Stat. *8*, 309–330.

Ruzanov, P., and Riddle, D.L. (2010). Deep SAGE analysis of the Caenorhabditis elegans transcriptome. Nucleic Acids Res. *38*, 3252–3262.

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.

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.

She, R. (2010). Fast and accurate gene prediction by protein homology. Thesis. School of Computing Science - Simon Fraser University.

She, R., Chu, J.S.-C., Uyar, B., Wang, J., Wang, K., and Chen, N. (2011). genBlastG: using BLAST searches to build homologous gene models. Bioinforma. Oxf. Engl. *27*, 2141–2143.

Shim, Y.-H., and Paik, Y.-K. (2010). Caenorhabditis elegans proteomics comes of age. Proteomics *10*, 846–857.

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

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. *100*, 15776–15781.

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.

Stein, L. (2001). Genome annotation: from sequence to biology. Nat. Rev. Genet. *2*, 493.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., et al. (2002). The Generic Genome Browser: A Building Block for a Model Organism System Database. Genome Res. *12*, 1599–1610.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. (2003). The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. PLoS Biol. *1*.

Stern, D.L. (2013). The genetic causes of convergent evolution. Nat. Rev. Genet. *14*, 751–764.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., et al. (2004). A gene expression map for the euchromatic genome of Drosophila melanogaster. Science *306*, 655–660.

Strange, K. (2006). C. Elegans: Methods and Applications - Google Books.

Sulston, J.E., and Horvitz, H.R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev. Biol. *56*, 110–156.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.

Tekaia, F. (2016). Inferring Orthologs: Open Questions and Perspectives. Genomics Insights *9*, 17–28.

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., 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.

Uyar, B., Chu, J.S.C., Vergara, I.A., Chua, S.Y., Jones, M.R., Wong, T., Baillie, D.L., and Chen, N. (2012). RNA-seq analysis of the C. briggsae transcriptome., RNA-seq analysis of the C. briggsae transcriptome. Genome Res. Genome Res. *22, 22*, 1567, 1567–1580.

Valentine, J.W., Collins, A.G., and Meyer, C.P. (1994). Morphological complexity increase in metazoans. Paleobiology *20*, 131–142.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial Analysis of Gene Expression. Science *270*, 484–487.

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C., and Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat. Commun. *7*, 1–13.

Wang, B., Kumar, V., Olson, A., and Ware, D. (2019). Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing. Front. Genet. *10*.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, 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, Z., Chen, Y., and Li, Y. (2004). A Brief Review of Computational Gene Prediction Methods. Genomics Proteomics Bioinformatics *2*, 216–221.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Weaver, R.F. (2012). Molecular biology / Robert F. Weaver. (New York: McGraw-Hill).

Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M.R. (2005). Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions. Genome Res. *15*, 577–582.

WormBase (2019). C. elegans - WormBase : Nematode Information Resource.

Wormbase release WS250 Wormbase release WS250 - WormBase : Nematode Information Resource.

Wormbase release WS254 Wormbase release WS254 - WormBase : Nematode Information Resource.

Wu, T.D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M.J. (2016). GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Methods Mol. Biol. Clifton NJ *1418*, 283–334.

Zhao, Z., Boyle, T.J., Bao, Z., Murray, J.I., Mericle, B., and Waterston, R.H. (2008). Comparative analysis of embryonic cell lineage between Caenorhabditis briggsae and Caenorhabditis elegans. Dev. Biol. *314*, 93–99.

Zhao, Z., Flibotte, S., Murray, J.I., Blick, D., Boyle, T.J., Gupta, B., Moerman, D.G., and Waterston, R.H. (2010). New tools for investigating the comparative biology of Caenorhabditis briggsae and C. elegans. Genetics *184*, 853–863.

Appendix A. Bioinformatics Tools

Table 13. List of open-source bioinformatics tools used in this thesis

Tools (version)	Brief descriptions	Usage in this thesis	References
SRA Toolkit (ver. 2.8.2)	SRA (Short Read Archive) manipulation tool	To download RNA-Seq libraries (fastq format) from NCBI using fastq-dump function	http://ncbi.github.io/sra-tools/fastq-dump.html
BBDuk (BBMap/BBTools ver. 37.36)	Data decontamination (using kmers) tool	To filter out reads mapping to rRNA genes	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide
FastQC (ver. 0.11.5)	Quality control tool for high throughput sequence data	To perform quality check (low confidence bases and adapters)	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic (ver. 0.36)	Reads and adapters trimming tool	To trim low-quality bases and adapters	Bolger et al., 2014
STAR (ver. 2.5.3a)	Splice-aware RNA-Seq aligner	To align the reads to the *C. briggsae* WS254 reference genome	Dobin et al., 2013
ExonTrap	Putative protein-coding exons identification tool	To reconstruct exons from RNA-Seq introns and integrated intron database	https://github.com/mattdoug604/exon_trap; Douglas, 2018
Cufflinks (ver. 2.2.1)	Genome-guided transcript assembler	To assemble transcripts from filtered RNA-Seq alignments	Trapnell et al., 2012
StringTie (ver. 1.3.4d)	Genome-guided transcript assembler	To assemble transcripts from filtered RNA-Seq alignments	Pertea et al., 2015
Trans-ABySs (ver. 1.5.5)	De-novo transcript assembler	To assemble transcripts from pre-processed RNA-Seq reads	Robertson et al., 2010
GffCompare (ver. 0.10.4)	Transcript datasets (GFF-format) comparison tool	To merge (collapse) assembled transcripts from 13 libraries for each assembler and further merge transcripts from 3 assemblers	https://github.com/gpertea/gffcompare
TransDecoder (ver. 5.5.0)	Candidate coding regions prediction tool	To predict and select transcripts that are protein-coding	http://github.com/TransDecoder/TransDecoder
OrthoMCL (ver. 2.0.9, with	Genome-scale algorithm for	To assign ortholog relationships between the *C. briggsae* and *C.*	Li et al., 2003

BLASTP ver. 2.5.0+)	grouping orthologous protein sequences	*elegans* before and after *C. briggsae* improvement	
GeMoMa (ver. 1.6.1)	Homology-based gene prediction program	To predict *C. elegans* protein-coding transcripts using *C. briggsae* protein-coding transcripts as the input	Keilwagen et al., 2018

Appendix B. Supplemental materials

1. Analysis of various quality thresholds for Trimmomatic

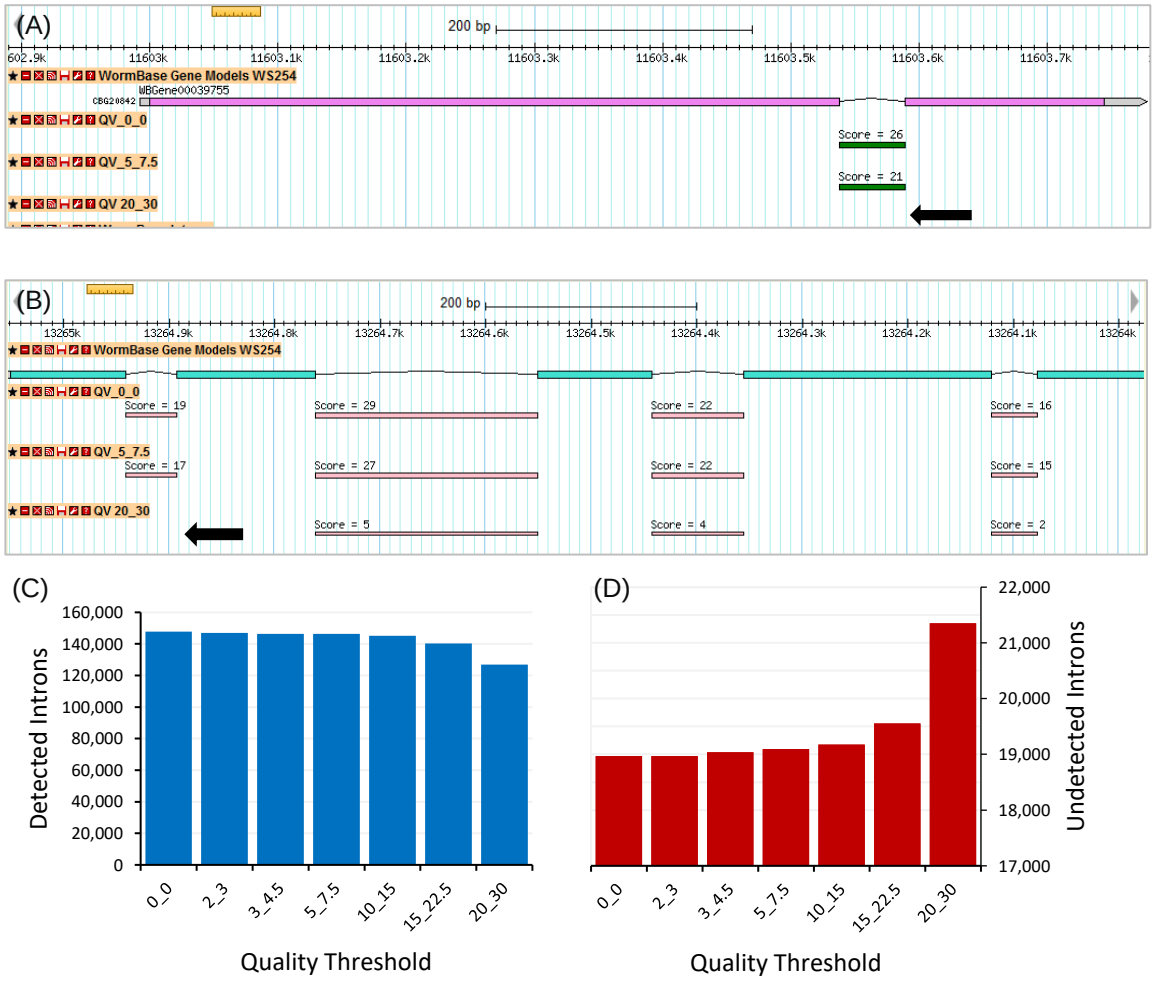

Figure B1. (A, B) Representatives showing the effect of stringent filtering using various quality thresholds; (C) The number of detected introns; (D) The number of WormBase introns undetected due to stringent filtering.

2. WormBase WS254 intron length distribution

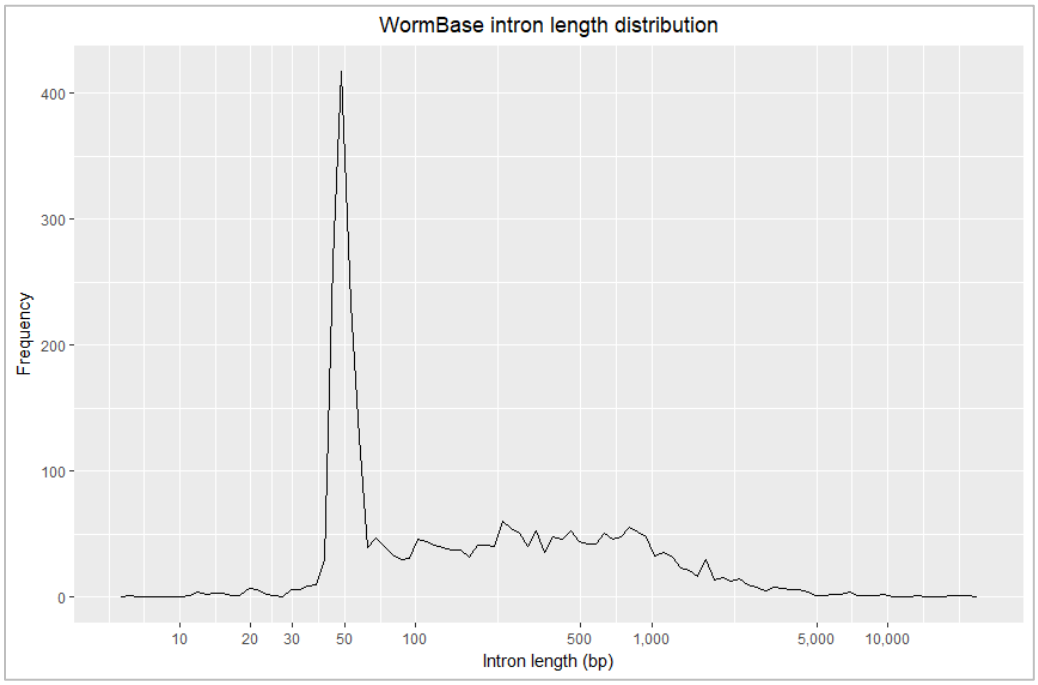

Figure B2. Distribution of intron lengths of WormBase introns.

3. Introns detection on various minimum intron thresholds

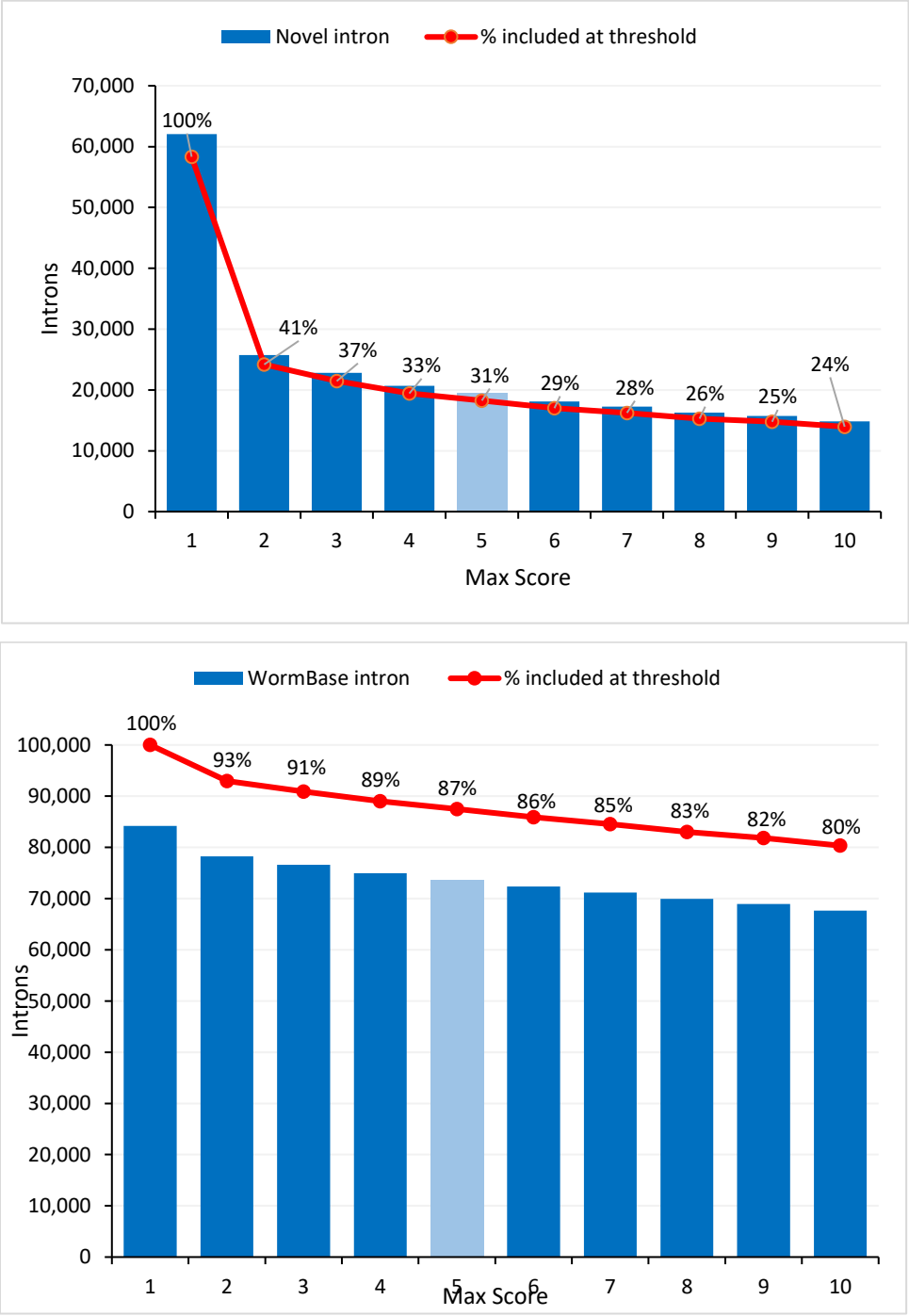

Figure B3. Different minimum intron thresholds affect intron detection. (top) Total of novel introns with the corresponding maximum read support in any library; (bottom) WormBase introns that have 10 or less read support in any library.

4. Diagnostic tests for various minimum intron thresholds

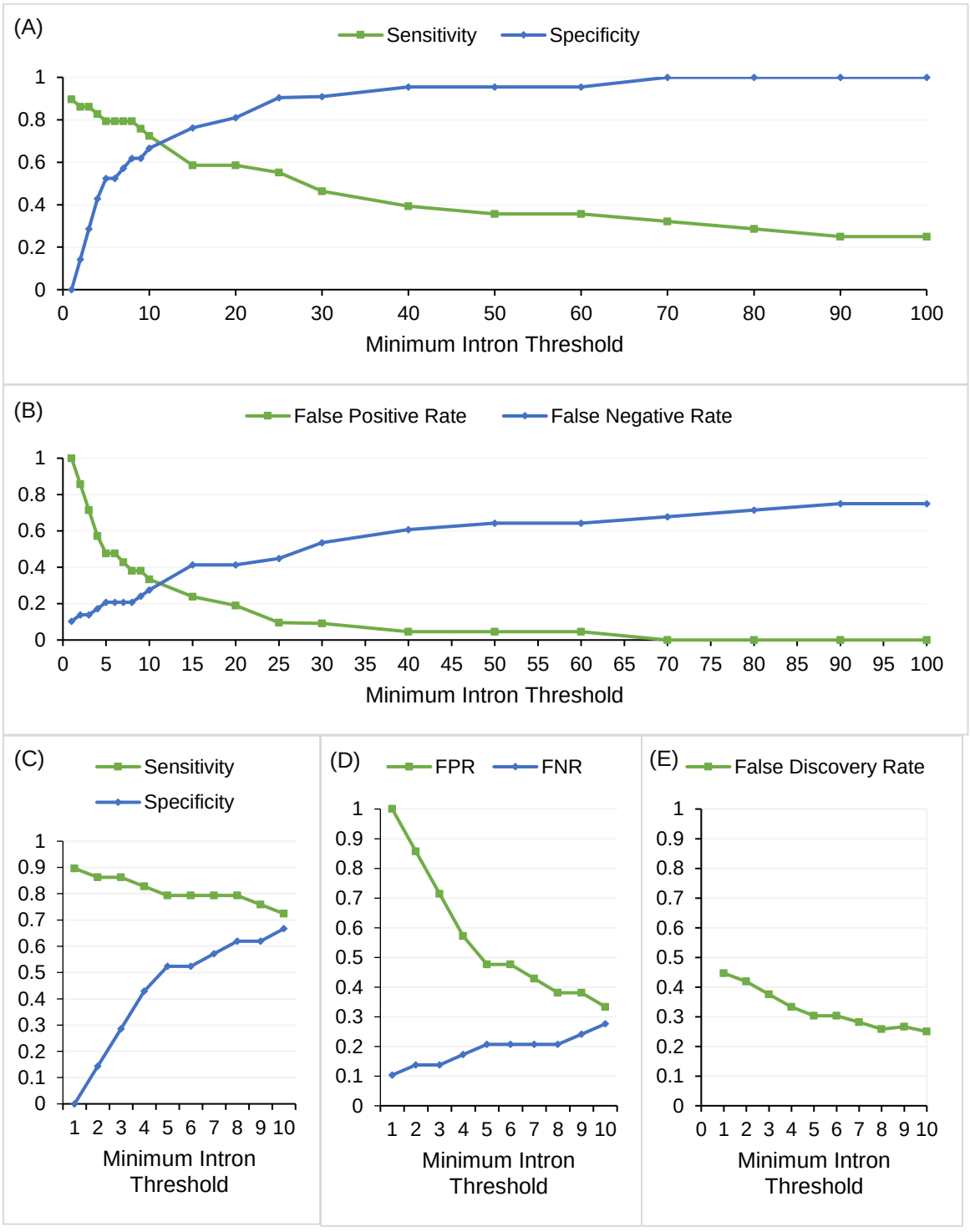

Figure B4. Diagnostic tests of various minimum intron support thresholds. (A-B) for minimum 1 to 100+ introns; and (C-E) for minimum 1 to 10+ introns. Sensitivity = TP/Total WB present = TP/(TP+FN), Specificity = TN/Total WB absent = TN/(TN+FP), False Positive Rate (FPR) = FP/Total WB absent = FP/(FP+TN), False Negative Rate (FNR) = FN/Total WB present = FN/(FN+TP), False Discovery Rate (FDR) = FP/(TP+FP).

5. Total transcripts per library supported by our intron and exon databases

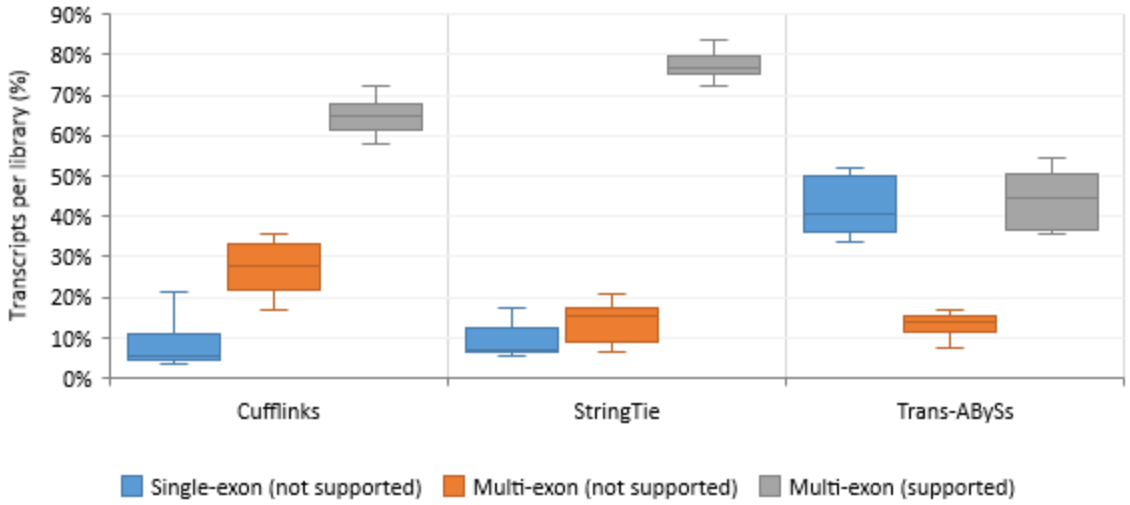

Figure B5. Percentage of transcripts per library that are supported or not supported by our databases.