HOMOLOGY AND EVIDENCE-BASED GENOME ANNOTATION OF *CAENORHABDITIS* SPECIES

by

BORA UYAR
B.Sc., Sabanci University, 2008

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Fall 2010

APPROVAL

Name: **Bora Uyar**

Degree: **M.Sc.**

Title of Thesis: **Homology and Evidence-based Genome Annotation of *Caenorhabditis* species**

Examining Committee:

Chair: **Dr. Marinko V. Sarunic,** Assistant Professor
School of Engineering Science

_____

Dr. Jack Chen, Associate Professor
Senior Supervisor, Department of Molecular Biology & Biochemistry

_____

Dr. David Baillie, Professor
Supervisor, Department of Molecular Biology & Biochemistry

_____

Dr. Cenk Sahinalp, Professor
Supervisor, School of Computing Science

_____

Dr. Fiona Brinkman, Professor
Internal Examiner, Department of Molecular Biology & Biochemistry

Date Defended/Approved: December 2 2010

Declaration of
Partial Copyright Licence

ABSTRACT

I have re-annotated the genomes of four sister species of the model organism *Caenorhabditis elegans: C. briggsae, C. remanei, C. brenneri, and C. japonica*, using a homology-based gene predictor genBlastG. Compared to the existing gene sets of these four species, genes of the revised gene sets demonstrate higher conservation with their orthologs in *C. elegans* as well as more ortholog pairs. I have validated and further revised the *C. briggsae* gene set, through next-generation short read sequencing of the transcriptome. In the revised *C. briggsae* gene set (23,159 genes), 7,347 genes (33.9% of all genes with introns) have all introns confirmed. Of all introns in the revised gene set (103,083), 62,727 (60.9%) are confirmed. Additionally, I have identified 1,034 operons in *C. briggsae*, with 532 (51.4%) perfectly conserved in *C. elegans*. This study sets up a solid platform for comparative genomics analysis and evolutionary studies of *Caenorhabditis* species.

Keywords: *Caenorhabditis* species, Comparative genomics, RNA-seq, transcriptome, gene models, synteny, operons.

To my family

"We do not consider our principles as dogmas contained in books that are said to come from heaven. We derive our inspiration, not from heaven, or from an unseen world, but directly from life".

Mustafa Kemal Atatürk

ACKNOWLEDGEMENTS

I would like to thank my senior supervisor Dr. Jack Chen for his utmost patience and encouraging support throughout my thesis project and helping me mature as a researcher. I would also like to present my gratitude to my thesis supervisors Dr. David Baillie and Dr. Cenk Sahinalp for their valuable comments and support for my research.

I sincerely thank Shu Yi Chua and Dr. Martin Jones, who prepared the L1 and mixed stage *C. briggsae* cDNA library; Dr. David Baillie for making the sequenced reads available for my study and Dr. Heesun Shin for helping the transfer of reads; Jeffrey Chu, Dr. Rong She, and Dr. Ke Wang for making genBlastG available for this project before publication and for numerous requested revisions of the program; Jeffrey Chu, Christian Frech, and Ismael Vergara for discussions on developing methods for building hybrid gene set and many other topics; Ismael Vergara for advices on using OrthoCluster in synteny analysis and for working together on operon identification and analysis; Ata Roudgar and Martin Siegert (SFU Research Computing, WestGrid site) as well as Christian Frech and Tammy Wong for working with computing resources at WestGrid; and all Chen lab members for their friendship and scientific discussions. I thank the CIHR/MSFHR Strategic Training Program in Bioinformatics; the Greater Vancouver *C. elegans* community and the monthly VanWoRM meeting organizers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Boundary Solexa intron	A Solexa intron, which is found at a genomic region where it overlaps with the 3' or 5' end of an existing gene model.
Hexamer frequency	Frequency of the occurence of 6 nucleotide long DNA subsequences in a genomic segment.
Intragenic Solexa intron	A Solexa intron, which is found at a genomic region where it is fully covered by an existing gene model.
Operon	Cluster of closely spaced genes, regulated from a single promoter, transcribed into poly-cistronic pre-mature mRNA molecules. The closely spaced cluster of genes have identical orientation and all downstream genes are *trans*-spliced by spliced leader type 2.
Perfect synteny block	A genomic segment of perfectly conserved gene content, order, and strandedness (Coghlan and Wolfe 2002)
Solexa exon	A genomic segment in which all bases are supported by >= 2 Solexa reads.
Solexa intron	A putative intron with canonical splice junctions found by spliced alignment of a Solexa read to a genomic sequence.
Spliced leader	A ~22 nucleotide exonic sequence which is donated by a 100 nucleotide small nuclear ribonucleoprotein particle and trans-spliced to the 5' end of pre-mRNAs (Blumenthal 2005).
***trans*-splicing**	A pre-mRNA maturation process, which involves removal of an outron and ligation of a spliced leader (SL) sequence with 5' most exon of a transcript (Bonen 1993).

1: INTRODUCTION

1.1 Motivation

1.1.1 Importance of high quality gene annotations for comparative genomics

Although the genome sequence contains all the necessary information regarding the characteristics of an organism, it needs to be characterized further in order to detect the functional segments that determine the phenotype of an organism (Hardison 2003). For a genome sequencing project, an important step is the annotation of a high quality gene sets, which is critical for subsequent analysis of the genome and the comparison between this genome and other genomes, for the ultimate understanding of the relationship between genotype and phenotype. Having a good understanding of transcriptional the start sites is critical for downstream analysis of the regulatory sequences such as promoters located upstream of the gene models. Similarly, accurate identification of stop codons would be important for subsequent analysis of post-transcriptional modification sites (such as poly-adenylation sites, microRNA binding motifs etc) located downstream of 3' ends of gene models.

The first step in comparative genomic studies is ortholog assignment between different species, using annotated gene sets for each species as the starting point. Many programs have been designed for this purpose, including OrthoMCL (Li, Stoeckert and Roos 2003) and InParanoid (Remm, Storm and

Sonnhammer 2001). In this project, we use InParanoid, which has been demonstrated to be one of the best performing programs for identifying orthologs (Chen, et al. 2007), to identify orthologs between *C. elegans* and *C. briggsae*.

Synteny analyses provide an overview of conservation between two genomes. The identification of synteny blocks can be based on similarity of DNA sequences, the similarity of protein-coding genes, or the correspondence of other genetic markers.

Defective gene models may lead to incorrect deductions about the evolutionary differences between genomes of compared species. Studies aiming to understand the evolution of genes, or gene families, would be impacted by the quality of the annotated genes models.

A higher quality annotation of genes in a species can be utilized to increase the quality of the annotation of its close relatives. Some homology-based gene prediction algorithms take advantage of the protein sequences of one species to predict gene models in a closely related species. If a gene model of the reference species is incorrect, an incorrect prediction will be made in the target species.

1.1.2 Phylogenetic relationships between five *Caenorhabditis* species

C. elegans has been very well studied as a model organism and it has a complete genome sequence. Recently, the genomes of four additional *Caenorhabditis* species were also sequenced to provide a platform for comparative genomics analyses. Among these, *C. remanei* and *C. briggsae* are

the closest relatives. *C. brenneri, C. briggsae,* and *C. remanei* form a group of sister species for *C. elegans. C. japonica* is an outgroup species for these 4 *Caenorhabditis* species (Kiontke and Fitch 2005).

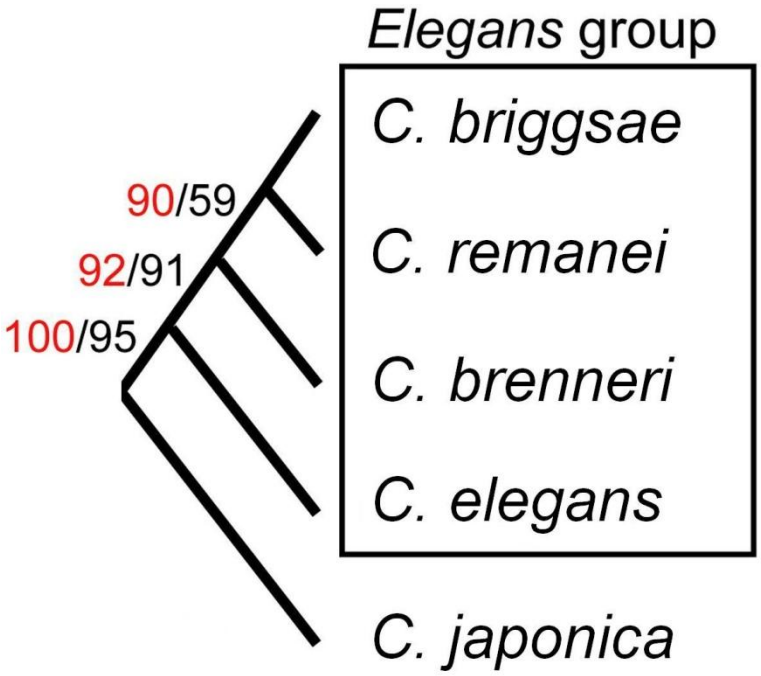

Figure 1-1: Phylogenetic relationships between *C. elegans, C. briggsae, C. remanei, C. brenneri*, and *C. japonica.* Plylogenetic relationships were computed by analyzing small subunit and large subunit rRNA genes and largest subunit of RNA Pol II. Black and red numbers denote the percent jackknife support computed by two different analysis involving 2,000 and 5,000 replicates respectively (Modified from WormBook).

1.1.3 Genome sequences of five *Caenorhabditis* species

Genome sequence of the model organism *C. elegans* was first published in 1998 (The C. elegans Sequencing Consortium 1998). A finished quality genome of *C. elegans* was completed in 2002. Thus, *C. elegans* is the first animal with all of its bases known.

C. briggsae is a soil-dwelling nematode, a close relative of *C. elegans* and needs an expert eye to distinguish the two. *C. briggsae* and *C. elegans* are predicted to have diverged from a common ancestor less than 30 million years ago (Cutter 2008). Its genome sequence was published in 2003 (Stein, et al. 2003) and is being prepared as another nematode model organism for comparative genomics analyses, especially studies of evolution (Baird and Chamberlin 2006). Like *C. elegans, C. briggsae* has 6 chromosomes and comparable size of genome (108 Mb).

In order to facilitate further comparative genomics analyses, three additional *Caenorhabditis* species (*C. remanei, C. brenneri,* and *C. japonica*) have been sequenced at the Genome Sequence Center of Washington University (http://genome.wustl.edu/genomes/list/invertebrates). However, their genome sequences are not assembled at chromosomal level, yet. They are still not finished and remain as contig level assemblies (ftp://ftp.wormbase.org/pub/wormbase/genomes).

1.1.4 nGASP gene predictions in Wormbase

In 2008, the nematode genome annotation assessment project (nGASP) was initiated in order to assess the existing gene prediction algorithms using a set of well annotated *C. elegans* protein coding genes and then use this assessment results to predict gene models in all four sister species of *C. elegans* (Coghlan, Fiedler, et al. 2008). In the competition, 47 algorithms from 17 research groups worldwide participated. The scientists were provided with 10% of *C.elegans* genome as training and test sets to run their algorithms. Results

were compared in four different categories: 1) *ab initio* gene predictors 2) predictors utilizing multi-genome alignments 3) predictors utilizing expressed sequences such as cDNA, EST sequences 4) combiner algorithms which may incorporate different algorithms and evidence from different categories. As a result of the study, combiner algorithms were found to provide the highest sensitivity and specificity. nGASP is the basis for the current annotation of gene models of nematodes. Currently, the genome annotations of the five *Caenorhabditis* species are maintained by WormBase (http://wormbase.org/) (Chen, Harris, et al. 2005).

1.1.5 Large scale comparison of *C. elegans* and *C. briggsae* genomes

Initial aim of sequencing and annotating additional *Caenorhabditis* species was to create a platform for comparative genomics. Therefore, comparison of *C. elegans* genome with *C. briggsae,* which has a chromosomal level genome assembly, has been carried out by various research groups who have analysed features such as orthologs, synteny, and conservation of operons between these species. Such studies have shown that large scale similarities exist between *C. elegans* and *C. briggsae*.

1.1.5.1 Orthologs and synteny blocks

Kent and Zahler developed a program called WABA (Kent and Zahler 2000) which aligned ~8 million bp of available *C. briggsae* genomic sequence against *C. elegans* genome. They found that 59% of the sequences were

conserved between the two species and the conserved sequences were found in coding regions (55%), introns (20%), and intergenic regions (25%).

Using BLASTX (Altschul, Madden, et al. 1997), Coghlan and Wolfe predicted 1784 genes in *C. briggsae* and found 252 long conserved segments of size 1 to 19 genes (Coghlan and Wolfe 2002). Segments were disrupted due to genomic rearrangements such as duplications, transpositions, and inversions (Coghlan and Wolfe 2002).

Stein and colleagues, who published a high quality draft of *C. briggsae* genome, reported that 12,200 out of 19,500 *C. briggsae* genes were found to have clear orthologs with *C. elegans* while 800 genes had no homology to any *C. elegans* genes (Stein, et al. 2003).

Hillier et al also reported that *C. elegans* and *C. briggsae* chromosomes are organized very similarly and more than 95% of 1-to-1 orthologs were found to be on the same chromosomes (Hillier, Miller, et al. 2007).

Finally, our lab recently reported that 51.1 % of *C. elegans* genomic sequence was covered by perfect synteny blocks which include collinear orthologous genes preserving order and strandedness (Vergara and Chen 2010).

1.1.5.2 Operons

Operons in nematodes are closely spaced gene clusters, which are transcribed into a poly-cistronic pre-mRNA and regulated from a single promoter upstream of the cluster while downstream genes are trans-spliced to SL2 (Spieth, et al. 1993). Zorio and colleagues have found that 70% of the *C. elegans*

genes are trans-spliced and around a quarter of them are in operons (Zorio, et al. 1994). The first evidence-based genome-wide annotation of operons in *C. elegans* was carried out using micro-array data, which detected ~1000 operons (Blumenthal, Evans, et al. 2002). Stein and colleagues found that 96% of the *C. elegans* operons are conserved in *C. briggsae* (Stein, et al. 2003). However, Qian and Zhang reported a lower level of conservation of operons (93%) between *C. elegans* and *C. briggsae* (Qian and Zhang 2008).

1.2 Methods

1.2.1 Overview of gene prediction methods

With the advent of next generation sequencing technologies, whole genome sequences of many species are being made available with an exponential decrease in the cost of time and money. Therefore, genome-wide computational prediction of protein coding genes has become an immediate problem. Predicting gene models involves detecting exons, splicing sites, 5' and 3' ends, 5' and 3' UTR regions. In eukaryotes, the problem of gene prediction is more complex due to increased number of introns and decreased percentage of coding regions in the genome (Mathe, et al. 2002). Methods developed to tackle this problem in eukaryotes have been reviewed by a number of review papers: (Mathe, et al. 2002), (Do and Choi 2006), (Brent 2007), (Flicek 2007), and (Sleator 2010).

Gene prediction methods have been categorized mainly into two: *ab initio* gene predictors and homology-based gene predictors. There are also 'combiners', which combine multiple types of algorithms or evidence. In the

scientific literature, *ab initio* gene predictors have also been referred to as *de novo* or intrinsic gene predictors. Similarly, homology-based gene predictors have also been referred to as expression-based, evidence-based, sequence similarity-based, empirical, or extrinsic gene predictors.

1.2.1.1 *Ab initio* gene prediction algorithms

Ab initio gene predictors rely on the intrinsic properties of dna sequences that signal existence of a gene. By utilizing intrinsic content sensors such as hexamer frequency (Mathe, et al. 2002), nucleotide composition, codon usage, base occurrence frequency (Sleator 2010), such predictors measure the probability of a genomic sequence to be protein-coding or not. In order to further tune the gene structure, intrinsic signal sensors such as splicing sites, transcriptional start sites and poly-adenylation signals are utilized. The statistical properties of the protein-coding regions are modelled by different statistical techniques such as neural networks, Markov models, and Fourier Transforms (Do and Choi 2006).

GENSCAN (Burge and Karlin 1997), TWINSCAN (Flicek, Keibler, et al. 2003), N-SCAN (Gross and Brent 2006) are among the most popular *ab initio* gene predictors.

Ab initio gene predictors depend on a experimentally verified set of known genes in order to train their algorithms. The accuracy of the gene finders depend on the strength of the signal and content sensors. Therefore, gene finders may perform poorly with genes with e.g. unusual hexamer content, exons/introns of

extreme number or lengths, and especially isoforms (Coghlan, Fiedler, et al. 2008).

1.2.1.2 Homology-based gene prediction algorithms

Homology-based gene predictors work by alignment of expressed sequences such as cDNA, EST, or protein sequences to genome sequences. Such alignments may be intra or inter species alignments. Assuming that the coding sequences are more conserved than non-coding sequences (Mathe, et al. 2002), highly conserved exonic regions are found. This comparison enables accurate identification of exon-intron boundaries (Brent 2007). However the predictions in this approach have some limitations. Firstly, cDNA sequences are not always available for all the genes, especially for the genes with low expression levels. Furthermore, when inter-species comparisons are made, the predictions are impacted by the phylogenetic distances between species (Sleator 2010). Existence of homologs in the databases is also a pre-condition for prediction of a gene.

Homology based gene prediction tools include AAT (Huang, et al. 1997), GeneSeqer (Usuka, Zhu and Brendel 2000), SIM4 (Florea, et al. 1998), GeneWise (Birney, Clamp and Durbin 2004), and genBlastG (R. She, J. S.-C. Chu, et al. Submitted).

1.2.1.3 Combiner algorithms

Some gene prediction algorithms take advantage of the type of information used by *ab initio* and homology-based gene predictors. These

methods are shown to produce more accurate and sensitive results in both nGASP (Coghlan, Fiedler, et al. 2008) and EGASP (Guigo, et al. 2006) gene prediction accuracy assessment projects.

1.2.2 Next generation sequencing technologies

The advent of next generation sequencing (NGS) technologies has revolutionized the biomedical research by making it possible to produce large amount of genome sequencing data in an unparalleled scale with low costs of time and money (Metzker 2010).

The general pipeline of NGS includes template preparation, sequencing, and imaging. Various platforms have been developed that differ in the way they implement these steps (Metzker 2010). DNA is sheared into smaller pieces and each small piece is used to create a fragment template or mate-pair template. These templates are sequenced and images captured during sequencing are post-processed to produce the nucleotide sequence of the fragments.

In Illumina/Solexa sequencing, templates are clonally amplified by solid phase amplification (Adessi, et al. 2000) and it is followed by sequencing by synthesis with cyclic reversible termination.

Roche/454 applies emulsion PCR to clonally amplify the templates and automates pyrosequencing (Ronaghi, Uhlén and Nyrén 1998).

Life/APG also clonally amplifies templates with emulsion PCR and sequences by ligation using support oligonucleotide ligation detection (SOLiD).

Helicos BioSciences is the first to implement single molecule sequencing, i.e the templates aren't amplified, which reduces the sequencing errors introduced during PCR amplification. This technology uses sequencing with cyclic virtual terminators.

Pacific Biosciences also uses single template and is the first to implement real time sequencing, which captures the images of the dye-labeled nucleotides being added to the DNA in real time.

Most of these methods produce giga-bytes of data per machine run and each machine run takes from hours to a week.

1.2.2.1 Transcriptome analysis using RNA-Seq

One application of NGS is the sequencing of transcriptome in order to catalogue all the transcript species, determine the transcriptional structure of genes in terms of 5' and 3' ends, post-translational modifications, and quantify expression levels under different conditions (Wang, Gerstein and Snyder 2009). Sequencing the transcriptome by NGS (RNA-seq) has been shown to be effective in gene expression profiling, small non-coding RNA discovery, detection of aberrant transcription, and protein-coding gene annotation (Morozova and Marra 2008).

RNA-Seq has been utilized for the study of the transcriptomes various species including humans (Morin, et al. 2008), mice (Mortavazi, et al. 2008), yeast (Nagalakshmi, et al. 2008).

For *C. elegans,* RNA-Seq has been applied for various aims such as profiling the transcriptome in its first larval stage (Shin, et al. 2008), for creating a high-resolution transcriptome map (Ramani, et al. 2009), generating a high quality genome annotation (Hillier, Rienke, et al. 2009) , and characterization of 3'UTRome (Mangone, et al. 2010).

For *C. briggsae,* RNA-Seq has never been utilized before. This is the first study using high-throughput sequencing of the transcriptome of *C. briggsae.*

1.2.2.2 Bioinformatics Tools

1.2.2.2.1 Short read mappers

An immediate problem after obtaining the output of an NGS run, millions of short sequence reads, is the alignment of these reads to a target sequence, usually a reference genome sequence. Due to the amount of data and short length of each read, mapping has required algorithms that handle both time and memory efficiency. Various algorithms have been developed to tackle this problem. What is common between various mapping algorithms is that they all do indexing of short reads or reference sequences. Depending on the data structures used for indexing, most of the fast mapping algorithms can be mainly divided into two: algorithms utilizing 1) hash tables and 2) suffix trees (Li and Homer 2010).

Algorithms that utilize hash tables for indexing are based on the seed-and-extend idea which was first used by BLAST (Altschul, Gish, et al. 1990). This idea involves finding matches for seeds of fixed length without allowing for gaps

or substitutions and then extending and refining the alignments (Li and Homer, 2010). Mappers that utilize seed-and-extend idea are different from each other in the way they implement the seeding. Spaced-seeding, i.e using seeds that allow for mismatches, and various versions of spaced seeding, which increase the sensitivity of alignments, was implemented by different mappers such as Eland(A. J. Cox, unpublished software), SOAP (Li, et al. 2008), SeqMap (Jiang and Wong 2008), MAQ (Li, Ruan and Durbin 2008), RMAP (Smith, Xuan and Zhang 2008), ZOOM (Lin, et al. 2008). Further algorithms such as SHRiMP (Rumble, et al. 2009), mrsFAST (Hach, et al. 2010) have used q-gram filtering that allows for gaps in the seeds (Li and Homer 2010).

Algorithms that utilize suffix trees are found to be faster than hash-based mappers. However, the speed comes with the cost of memory (Li and Homer 2010)

Base quality is also an important parameter that needs to be taken into account by short read mappers. The accuracy of Solexa read mapping is shown to be increased by incorporating the base quality information (Smith, Xuan and Zhang 2008).

1.2.2.2.2 Spliced Aligners

As stated before, RNA-seq data can be used to annotate protein-coding genes. To determine the structure of genes, determination of introns/splicing sites has crucial importance.

Identification of *cis*-splicing sites is an important problem to determine the intronic regions of genes to obtain a well-defined gene model structure. Recently developed splicing site detection algorithms include Q-PALMA (De Bona, et al. 2008), TopHat (Trapnell, Pachter and Salzberg 2009), GSNAP (Wu and Nacu 2010), Scripture (Guttman, et al. 2010), SpliceMap (Au, et al. 2010), Split-Seek (Ameur, et al. 2010), and SuperSplat (Bryant Jr., et al. 2010). Q-PALMA is a machine learning approach which is trained by a set of known splice junction sequences to predict the splice junctions. TopHat also relies on canonical splice junctions and it also depends on exonic coverage islands which mean that there has to be enough number of reads mapping to an exonic region in order to be able to confidently distinguish such a region from intronic regions. However, the rest of more recently developed methods don't rely on either canonical splice junctions or depth of sequence reads and they are particularly developed to tackle the problem of spliced alignment of short reads generated by RNA-seq, making them favourable to have an unbiased detection of introns.

1.3 Analysis overview

In this study, firstly I re-annotated the genomes of *C. briggsae, C. remanei, C. brenneri, and C. japonica* using a homology-based gene predictor, genBlastG. genBlastG defines protein-coding gene models by annotating *cis*-splicing sites, 5' and 3' ends from start to stop codons. Using *C. elegans* protein sequences as query, genBlastG predicted gene models in the target genomes by maximizing the sequence similarity to the query sequences. In order to produce a higher quality gene set, we compared the genBlastG predictions with WormBase

annotations and created a hybrid set of gene models using both sets. We replaced WormBase annotations with genBlastG predictions if the genBlastG models improved sequence similarity to the orthologous *C. elegans* genes. 6,715, 7,676, 7,904, and 7,385 gene models in *C. briggsae, C. remanei, C. brenneri, and C. japonica* were replaced by genBlastG predictions, respectively. Furthermore, genBlastG predicted 1,091, 640, 4,447, 855 novel gene models in these four species, respectively. In all four species, homology-based gene model predictions improved the number of orthologous *C. elegans* genes, increased the amount of perfect synteny coverage and synteny block sizes.

For *C. briggsae* strain AF16*,* we sequenced the transcriptomes of L1 and mixed developmental stages using Illumina Solexa Genome Analyzer II platform. We aimed to use RNA-seq data to confirm or further improve computationally predicted protein-coding gene models. Currently, very small number of *C. briggsae* gene models is supported by expression evidence. We used MAQ to map Solexa reads to the genome sequence of *C. briggsae* to find the exonic regions. Then, we annotated the *cis*-splicing sites and introns by running SuperSplat. Comparison of these introns with predicted set of introns, we confirmed 59,137 (or 57.5%) of introns. By revising existing introns/exons and incorporating novel introns/exons in 2,346 gene models, we obtained 62,727 (or 60.9%) introns confirmed by RNA-seq data. 14,812 (or 68.3%) of gene models (with >= 1 intron) had at least 1 intron confirmed and 7,347 (or 33.9%) gene models (with >=1 intron) had all of their introns confirmed. Also, 47% of the genes had >95% of their cDNA sequences supported by Solexa reads.

Furthermore, we used a local alignment tool, cross_match, to identify trans-splicing sites. We found 7,871 genes trans-spliced by SL1 and 2,287 genes trans-spliced by SL2 type trans-splice leaders. By finding closely spaced gene clusters with SL2 trans-spliced downstream genes, we annotated operons in *C. briggsae.* We identified 1,034 operons in *C. briggsae.* Comparison of these operons with annotated *C. elegans* operons revealed that 532 (or 51.5%) of operons are perfectly conserved,349 (or 33.8%) of them were found to be divergent, and 153 of them were found to be *C.briggsae* specific. .

Detection of trans-splicing sites and Solexa read alignments in the flanking regions of the gene models gave us the opportunity to annotate 5' and 3' UTR regions. For 16,408 genes (or 70.85%), either 5'UTR or 3'UTR regions were defined. 11,770 genes have both 5' and 3' UTR regions defined. 2,319 genes have only 5' UTR regions and 2,319 genes have only 3'UTR regions defined.

Thus, we produced the first evidence-based genome-wide set of protein-coding gene models for *C. briggsae.* Based on RNA-seq data, we annotated *cis* and *trans*-splicing sites, 5' and 3' UTRs, start and stop codons. Using the information we annotated operons and provided a preliminary comparison of operons between *C. elegans* and *C. briggsae.*

2: HOMOLOGY-BASED GENE MODEL REVISION OF *CAENORHABDITIS* SPECIES

2.1 Genome sequencing of *Caenorhabditis* species

2.1.1 Genome sequencing and assembly

The *C. elegans* genome is the first metazoan genome subject to whole genome sequencing (The C. elegans Sequencing Consortium 1998) and is the only metazoan genome that is completely sequenced, without any sequencing gaps. In contrast, all other metazoan genomes contain certain numbers of gaps. For example, the current reference human genome assembly (GRCh37) has 357 major sequencing gaps (Genome Reference Consortium 2010). Genomes of *Caenorhabditis* species related to *C. elegans*, including *C. briggsae*, *C. remanei*, *C. brenneri*, *C. japonica*, have been recently sequenced and assembled. This availability of sequenced genomes of multiple species present as a valuable resource for comparative genomes, which can reveal functional elements in *C. elegans* and facilitate understanding of *C. elegans* genome conservation, architecture and its impact in gene expression and function (Chen and Stein 2006).

2.1.2 Gene predictions in *Caenorhabditis* species

The *C. elegans* genome is arguably the best annotated metazoan genome, making it valuable as a reference for comparative genomic analysis (Hillier, Coulson and Murray 2005). Annotated *C. elegans* gene models, together

with gene expression and function data, as well as information of *C. elegans* resources, are all available in WormBase (http://www.wormbase.org/) (Chen, Harris, et al. 2005) , making it very convenient for fast access of relevant information. In contrast to the high quality annotation of the *C. elegans* genome, the genome annotations of its sister species have not as well annotated. These four genomes have been annotated by the nematode genome annotation assessment project (nGASP) (Coghlan, Fiedler, et al. 2008). Based on the current WormBase release note (WS215), very few (<1%) of these predicted gene sets have been experimentally validated. We have found that a large proportion of these predicted gene models might be defective, which is described below.

2.1.3 Problems with current predictions

Accurate annotation of genomes is critical for comparative genomics analyses of these genomes. Flawed gene models will lead to false negative identification of orthologs and false positive identification of divergent genes. To evaluate the quality of the predicted genes sets for these four species, we compared predicted genes in each of these four gene sets with their orthologous genes in *C. elegans*, assuming that orthologous genes should show overall similar gene models for most genes.

Our comparison revealed that many gene models in these genes sets need to revised because they likely miss exons and introns and some exons and introns need revision (Figure 2-1), adjacent gene models should be merged to

form one gene model (Figure 2-2), gene models should be split into two or more separate gene models (Figure 2-3). Additionally, some gene models may be entirely missing from the predicted gene sets (Figure 2-4). Because of these problematic gene models, the number of predicted orthologous genes between each of these four species and *C. elegans* is low, especially for the three species. Using InParanoid (Remm, Storm and Sonnhammer 2001), I was able to identify 14,755 *C. remanei-C. elelgans* orthologs, 13,481 *C. brenneri-C. elegans* orthologs, and 11,638 *C. japonica-C. elegans* ortholgs, whose gene sets have not been published. In contrast, the *C. briggsae* genome, whose annotation was published recently (Stein, et al. 2003), was reported to have 12200 orthologs, similar to what we find. We found 14,167 *C. briggsae-C. elegan* orthologs using InParanoid.

In this part of my thesis, I aim to examine all predicted gene models in these four gene sets by comparing these gene models with their corresponding orthologs in *C. elegans*, applying genBlastG**,** a homology-based gene model prediction program that was recently developed in my laboratory (R. She, J. S.-C. Chu, et al. Submitted) The predicted gene sets were obtained in WormBase (release WS215). The revised gene set for each species contains all revised gene models, novel gene models, and the predicted gene models that do not need revision.

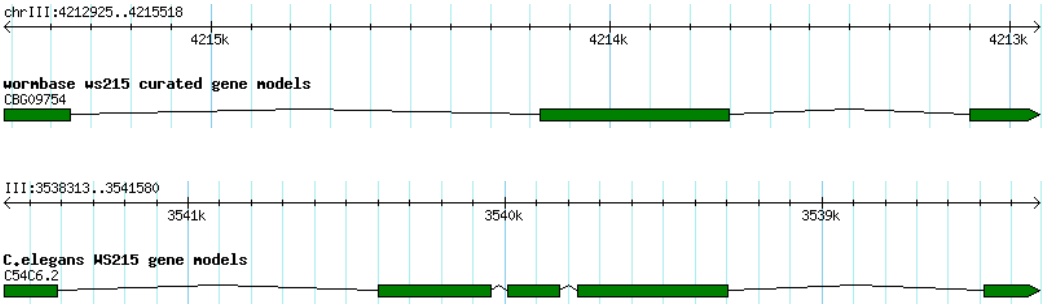

Figure 2-1: A representative *C. briggsae* gene model that may need modification. Comparison between the predicted *C. briggsae* gene model CBG09754 (Wormbase) (upper panel) and its ortholog in *C. elegans* C54D6.2 (lower panel) suggests that the predicted *C. briggsae* gene model may miss internal coding exons.

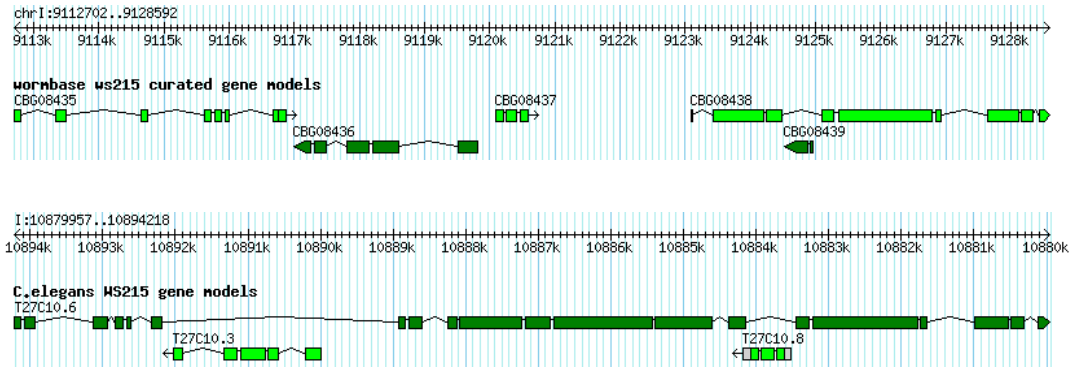

Figure 2-2: A representative case in which multiple *C. briggsae* gene models need to be merged. Three *C. briggsae* gene models CBG08435, CBG08437, and CBG08438 (upper panel) are orthologous to fragments of one single gene T27C10.6 in *C. elegans* (lower panel). Likely, these three *C. briggsae* gene models need to be merged into a single gene model.

.

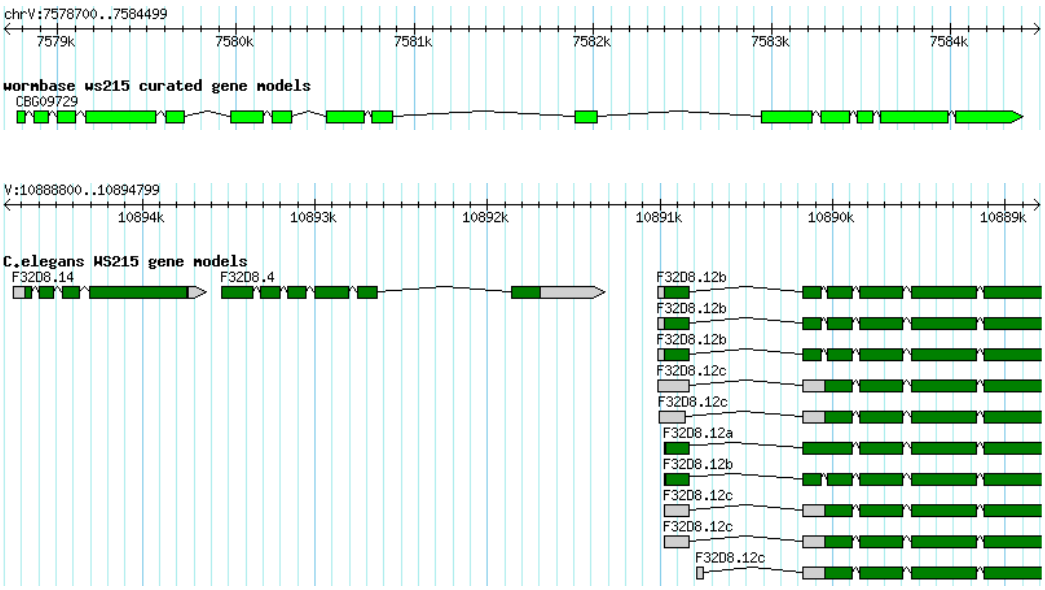

Figure 2-3: A representative *C. briggsae* gene model that may need to be split into two independent gene models. The *C. briggsae* gene model CBG09729 (upper panel) is orthologous to three *C. elegans* genes (F32D8.14, F32D8.4, and F32D8.12) in tandem (lower panel), which suggests that *C. briggsae* gene model likely needs to be split.

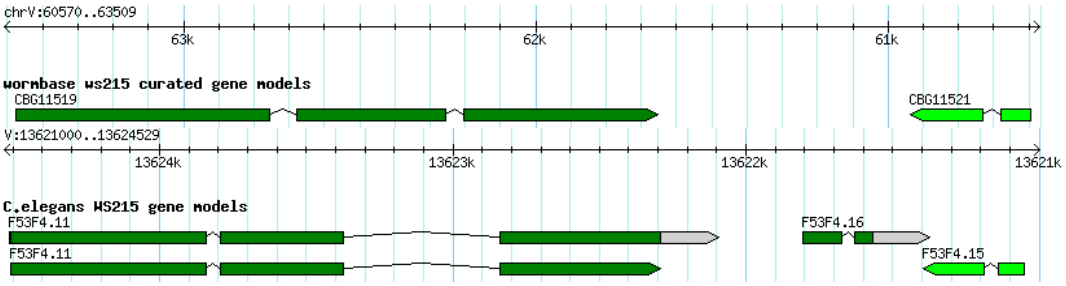

Figure 2-4: A representative novel gene model in *C. briggsae.* The *C. briggsae* gene models CBG11519 and CBG11521 (upper panel) are orthologous to *C. elegans* gene models F53F4.11 and F53F4.15 (lower panel), respectively. However in the same region of *C. elegans,* there's an extra gene model F53F4.16, which is a potentially novel gene that is missed in *C. briggsae* genome annotation.

2.2 genBlastG and revision procedure

2.2.1 Description of genBlastG

genBlastG is a homology based gene prediction algorithm recently developed in my laboratory. It builds on a previously published method called genBlastA (R. She, J. S. Chu, et al. 2009). Compared to GeneWise, a widely used homology-based gene prediction program, genBlastG is up to 1,000-fold faster. More importantly, genBlastG is more accurate. The input to genBlastG is the peptide sequences of a query in the reference species and the genome sequence of a target species. In the target genomic sequence, locally similar regions to the query peptide sequences are found and gene models are defined in those regions. genGlastG identifies both gene start and stop codons. It detects introns by finding in-frame splicing donor and acceptor sites. The gene model defined by genBlastG is optimized for the percent sequence identity (PID) between predicted gene models' exonic sequences and the query peptide sequences.

2.2.2 Filtration of genBlastG output

For each query protein, genBlastG can predict one or more candidate gene models (that are potentially paralogous). For each candidate gene model, genBlastG calculates the percentage identity (PID) between the query protein and the predicted peptide sequence. Users can modify the parameters to limit the number of candidate paralogous for each query.

When genBlastG is used to predict homologous genes in a genome, the resulting predicted gene set may contain many redundant gene models predicted

by paralogous query genes. This situation is especially severe for multi-gene families whose members are highly similar, such as the histone gene families. Each histone query can predict almost all paralogous histone genes. As such, the gene set predicted using genBlastG should be filtered for redundancy. Thus, among gene models that overlap with each other, only the one with the highest PID value is selected.

In the homology-based gene model improvement pipeline, *C. elegans* gene set was chosen to be the reference because it is the most widely studied and well curated. The goal of this project is to predict one isoform for each protein-coding gene. Thus I selected only one isoform (the longest one) for each *C. elegans* gene as a query. I used altogether 20,335 *C. elegans* peptide sequences (in WS215) for gene revision. I obtained 256,367, 292,613, 412,462, and 165,602 raw genBlastG genes models in the genomes of *C. briggsae, C. remanei, C. brenneri,* and *C. japonica*, respectively. For each set of these predictions, the following filtering criteria were applied:

a) All the predictions are sorted in decreasing order with respect to the PID values computed by genBlastG.

b) Let two gene models G1 and G2 with coding sequence length of L1 and L2 with PID values p1 and p2 (p1 >= p2) have a coding sequence overlap of L. If L/L1 or L/L2 is bigger than 5%, then G2 is filtered out and G1 is kept.

c) From the remaining set of gene models, only gene models with PID>=40% are kept and the rest is filtered out.

After filteration, I predicted 16,285, 17,565, 23,007, and 13,157 gene models in the genomes of *C. briggsae, C. remanei, C. brenneri,* and *C. japonica*, respectively. These predicted gene models will be compared against the nGASP-predicted gene models for revision.

2.2.3 Merging of genBlastG predictions with previously predicted gene set

Although the five *Caenorhabditis* species analyzed in this project are related and are very similar in gross morphology, they show much difference at the genetics and genomics level. For example, the two best studied species *C. elegans* and *C. briggsae* might have split up to 100 million years ago (Coghlan and Wolfe 2002), (Stein, et al. 2003). Expectedly, each of these species harbours a large number of species-specific genes, which cannot be found using homology-based gene prediction programs. Each of nGASP-predicted gene models in the genomes of *C. briggsae, C. remanei, C. japonica, and C. brenneri* was examined for its relationship with the genBlastG-predicted gene model. For each species, I generated a “hybrid gene set”, which contains all nGASP-predicted gene models that do not overlap with genBlastG-predicted gene models, as well as nGASP-predicted gene models show higher or slightly lower PID to their corresponding *C. elegans* orthologs. I decided to use a genBlastG-predicted gene model to replace an nGASP-predicted gene model only when the PID between genBlastG-predicted gene model and its *C. elegans* ortholog is 2% higher than the PID between nGASP-predicted gene model and its *C. elegans* ortholog. Maximal overlap (5% of coding sequences) between two adjacent gene

models is allowed. In this analysis, I used genome and gene annotation data from WormBase release WS215 and genBlastG (v135) (Table 1-1).

Table 1: Summary of revised gene set.

	C .briggsae	***C. brenneri***	***C. remanei***	***C. japonica***
Total	23,276	34,887	31,648	25,574
WS215*	15,470	22,536	23,332	17,334
Replaced	6,715 (28.8%)	7,904 (22.6%)	7,676 (24.2%)	7,385 (28.9%)
Novel	1,091 (4.7%)	4,447 (12.7%)	640 (2%)	855 (3.3%)

*Predicted gene models kept.

2.3 Analysis of revised gene set

2.3.1 Types of revisions done

Examination of the revised gene models in the hybrid gene set reveals three types of revisions, as suggested in the above illustrations (Figure 2-1, Figure 2-2, Figure 2-3, and Figure 2-4): (1) gene model revision (such as extension, truncation, and addition or removal of internal exons); (2) split of predicted models into multiple gene models; (3) adjacent predicted gene models are merged into a single gene model (see Table 2).

Table 2 : Types of revisions made by replacing WS215 models with genBlastG models. Each cell in the table denotes the number of genes that are subject to the corresponding type of revision in the genome of the corresponding species.

	C. briggsae	***C. brenneri***	***C. remanei***	***C. japonica***
Modified	5,387	6,501	6,363	5,477
Merged	282	549	668	1,251
Split	954	794	558	540
Merged & split	92	60	87	117
Total	6,715	7,904	7,676	7,385

2.3.2 Evaluation of genBlastG-predicted novel genes

Using our pipeline, I have detected thousands of novel genes in all four *Caenorhabditis* species (see Table 1). All of these gene models have predicted open reading frames with start/stop codons and identified exons interleaved by introns, which are defined by splicing signals. Moreover, all of these models have >=40 PID to a *C. elegans* query peptide sequence suggesting that these are conserved genes but are missing in WormBase annotations. See Figure 2-5 for successfully repaired gene models.

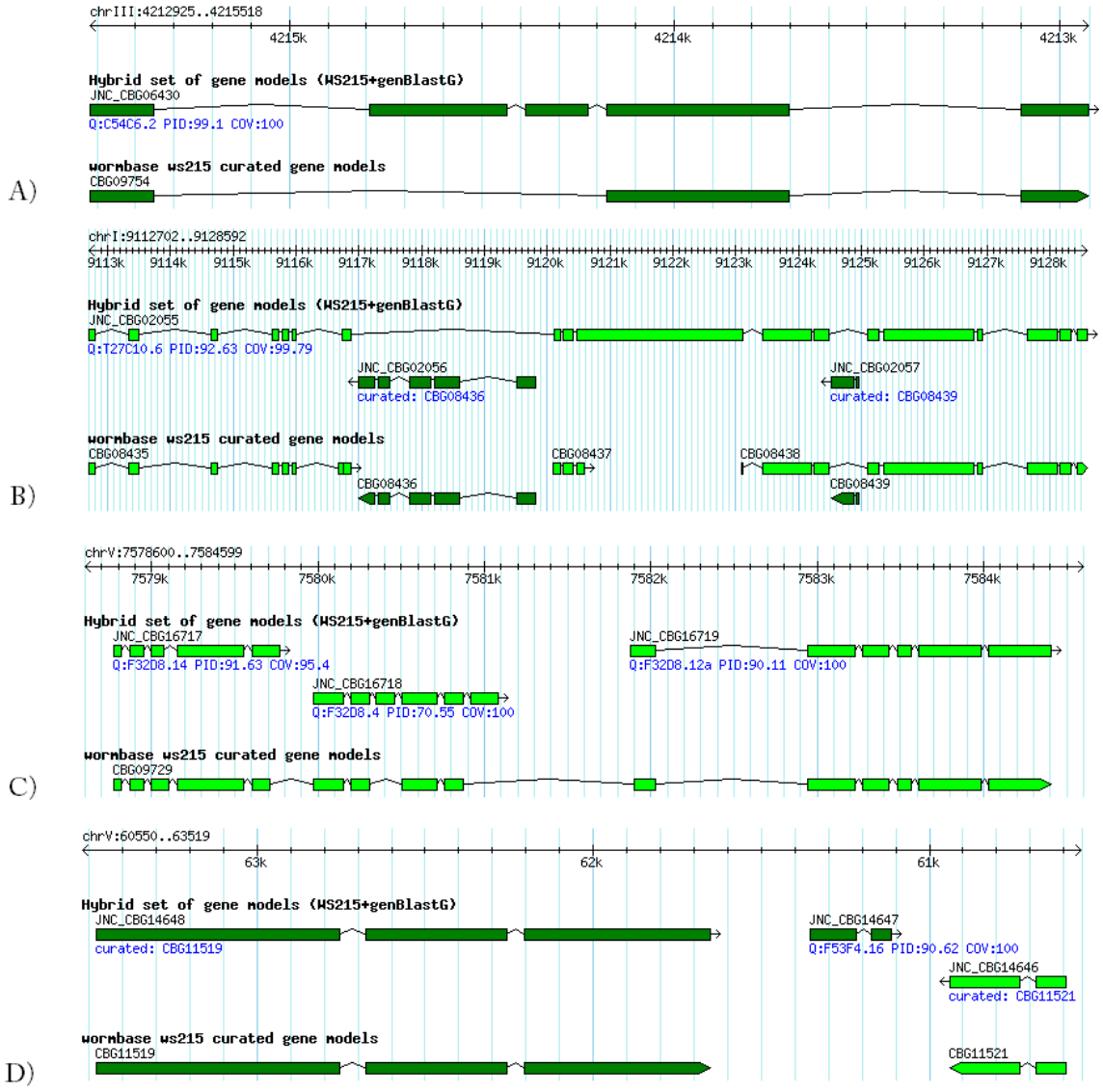

Figure 2-5: Examples of revised gene models by genBlastG in *C. briggsae.* Gene models in forward strand are dark-green coloured and those in negative strand are light-green coloured. A) CBG09754 is modified. Two extra exons are added and the revised gene model shows 99% PID to its ortholog *C. elegans* gene C54C6.2. B) CBG08435, CBG08437, and CBG08438 are merged into a single gene model JNC_CBG02055, which has a 92% PID to its orthologous *C. elegans* gene T27C10.6. C) CBG09729 is split into 3 gene models JNC_CBG16717, JNC_CBG16718, and JNC_CBG16719, which have 91%, 70%, and 90% PID to their orthologous *C. elegans* genes F32D8.14, F32D8.4, and F32D8.12, respectively. D) A novel gene JNC_CBG14647, which is orthologous to *C. elegans* gene F53F4.16, is found between two *C. briggsae* genes CBG11519 and CBG11521, which are orthologous to *C. elegans* genes F53F4.11 and F53F4.9, respectively.

2.3.2.1 Many novel genes are single copy genes

Members of multi-gene families tend to be missed in gene annotation. I would like to examine whether the genBlastG-predicted novel genes are enriched with members of large multi-gene families such as the chemosensory gene families or transcription factor families, while all single-copy genes, which tend to be more conserved across different species, tend to be better annotated. To examine this idea, I examined whether all novel genes are members of multi-gene family members. Based on the number of predictions generated per *C. elegans* query gene, we have observed that many novel gene models are single copy genes (Figure 2-5). In other words, for the *C. elegans* queries of many of the novel gene models, a single gene model is predicted in the genomes in the genomes of all four *Caenorhabditis* species.

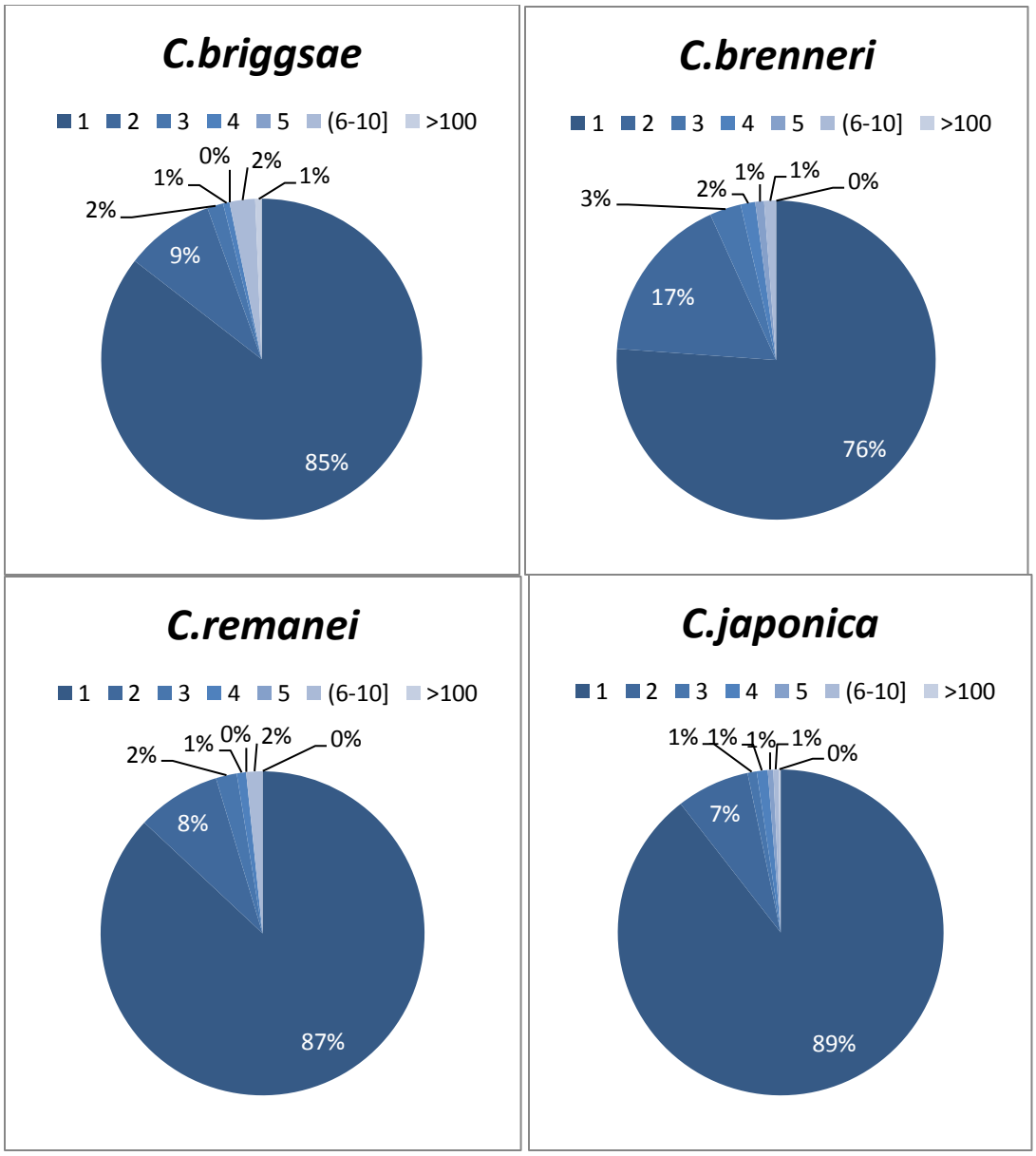

Figure 2-6: Number of Predictions per *C. elegans* query gene. Each *C. elegans* gene can be used to predict one or more paralogous gene models in the target genomes. Most of these genes are used to predict single copies of genes while some of them have led to generation of multiple gene predictions.

2.3.2.2 Many novel genes encode long proteins

I also wanted to observe if the novel gene models that we've found are mostly non-specific hits. We've seen that many of the missing genes encode for long proteins, which decreases the probability of finding a gene model by chance. See Figure 2-7 for the distribution of the novel genes with respect to their length.

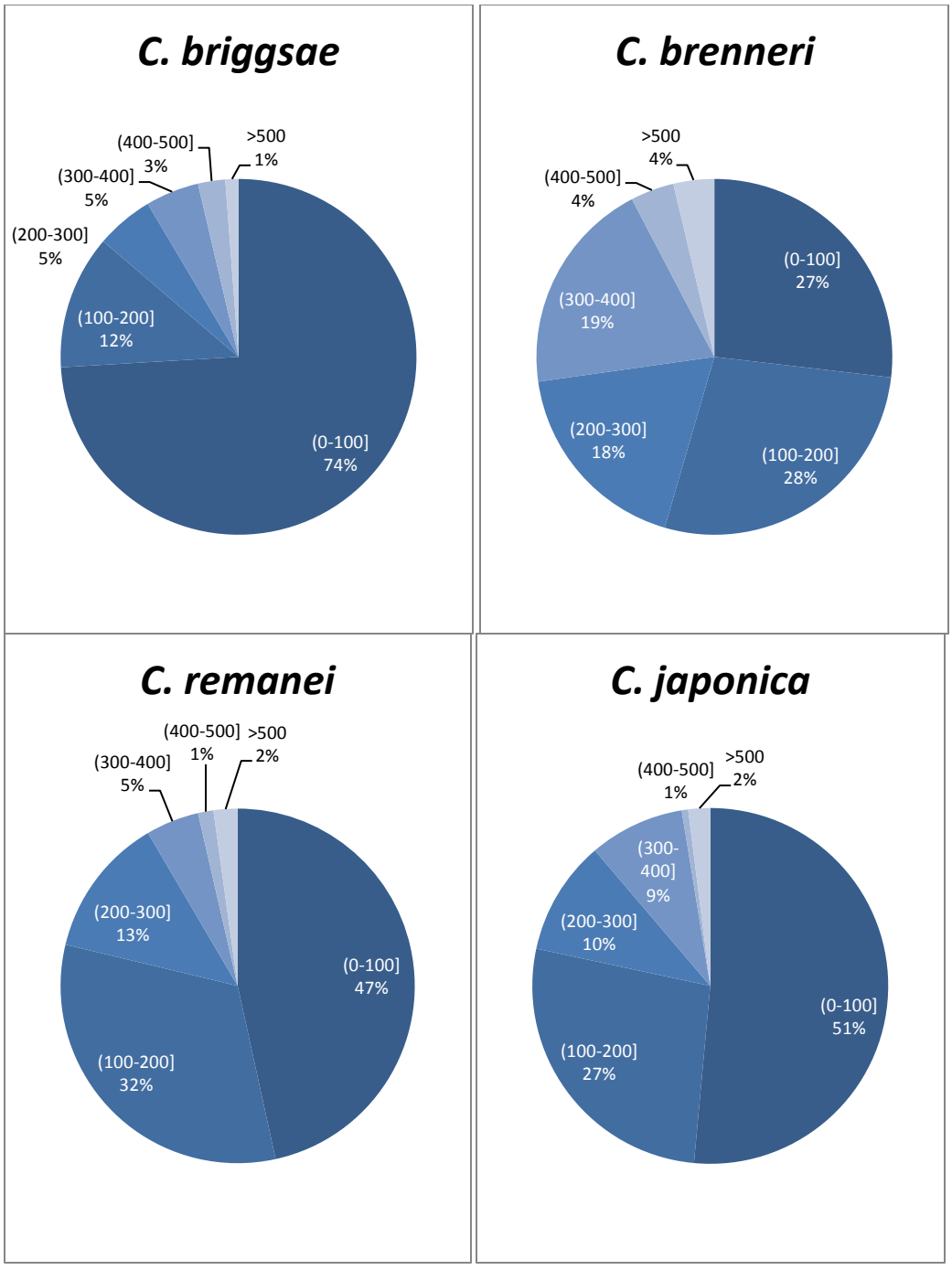

Figure 2-7: Length (in number of amino-acids in the peptide) distribution of novel genes. A significantly large proportion of the genes encode peptides longer than 100 amino-acids in all four species.

2.4 Improvement of ortholog assignment

The value of homology-based gene model improvement is demonstrated in the improvement of ortholog assignment between *C. elegans* and each of the four sister species. Before homology-based gene model improvement, *C. elegans* gene set has 14,167, 14,755, 11,638, and 13,481 genes that have clear orthologs in *C. briggsae*, *C. remanei*, *C. japonica*, and *C. brenneri*, respectively. After homology-based gene model improvement using genBlastG, *C. elegans* gene set has 15,108, 15,256, 12,953, and 14,319 genes that have clear orthologs in *C. briggsae*, *C. remanei*, *C. japonica*, and *C. brenneri*, respectively.

2.5 Synteny analysis

In order to evaluate the quality of the revised annotations, synteny improvement between *C. elegans* and the corresponding sister species was quantified. A synteny block is a block of genes which is conserved in chromosomes of species that are related. In our analysis we quantified perfect synteny blocks (see Figure 2-8: Perfect Synteny Blocks), in which, corresponding genes in the block are orthologs and order and strandedness of the orthologs are conserved.

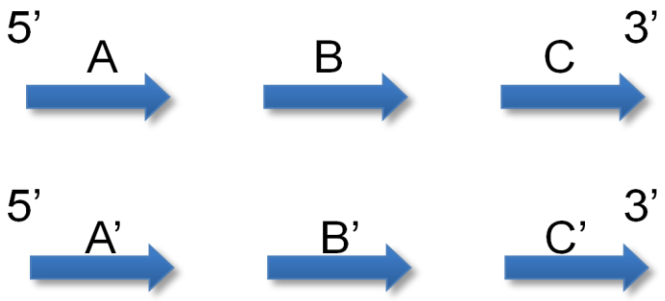

Figure 2-8: Illustration of a perfect synteny block. A, B, C represent genes in one genome, while A’, B’, and C’ represent their corresponding orthologs in a second genome. Collectively, these orthologs define a perfect synteny block because the order and strandedness are conserved.

Pairs of orthologous genes in the compared gene sets were found by InParanoid (v4.1). Perfect synteny blocks between species pairs were detected using OrthoCluster (Zeng, et al. 2008). As expected, compared to the amount of synteny computed using the WormBase annotations, the level of synteny was increased with our improved annotations for all four sister species of *C. elegans*. The annotation improvement pipeline, by fixing defective gene models and finding missing gene models, aided reconstruction of synteny blocks that would otherwise be broken (See Figure 2-9). As a result, the coverage of the perfectly syntenic regions of the genome (See Figure 2-10) and largest synteny block sizes (See Figure 2-11) were increased.

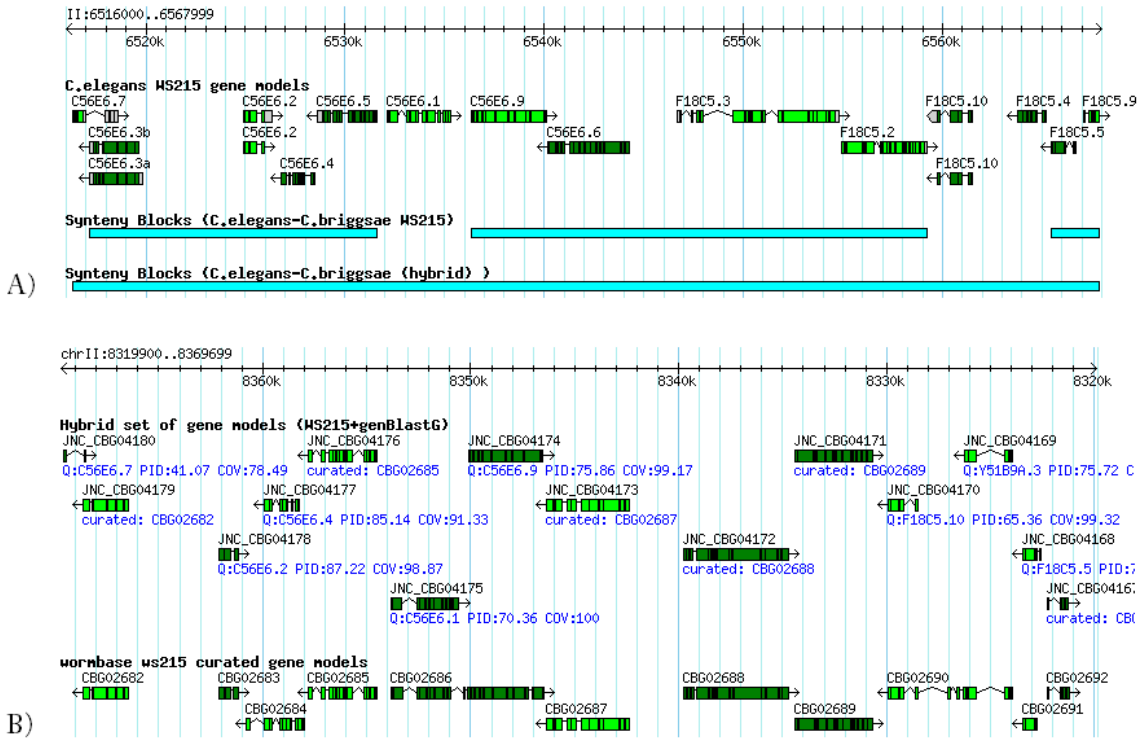

Figure 2-9: Larger perfect synteny block is observed after defective gene models are fixed. A) The genomic region (II:6516000-6568000) of *C. elegans* contains 14 genes. Based on the predicted *C. briggsae* gene models (WS215), this region contains 3 separate synteny blocks (See the track "Synteny Blocks (*C. elegans-C. briggsae* WS215)"). However, based on the hybrid set of gene models obtained by merging genBlastG predictions with WS215 gene models of *C. briggsae,* this region is found to be completely syntenic between *C. elegans* and *C. briggsae*(See the single synteny block in track "Synteny Blocks (*C. elegans-C. briggsae* (hybrid)"). B) *C. briggsae* gene models from the "hybrid set" and the WS215 set are shown in tracks 1 and 2 respectively. By predicting a "novel" gene JNC_CBG04180 using C56E6.7 gene as query, splitting both of the two *C. briggsae* genes CBG02686 (using *C. elegans* genes C56E6.1 and C56E6.9 as queries) and CBG02690 (using *C. elegans* genes F18C5.10 and Y51B9A.3 as queries), the 3 synteny blocks are reconstructed into a single synteny block.

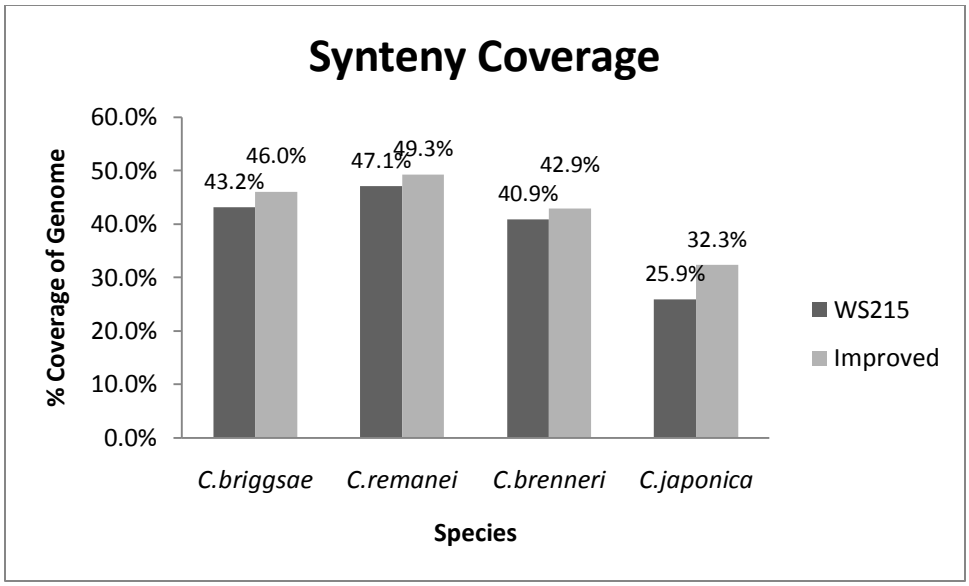

Figure 2-10: *C. briggsae* gene model revision resulted in higher synteny coverage. Synteny coverage is measured as the ratio of the genomic region covered by the perfect synteny blocks to the size of the *C. elegans* genome. Gene set obtained by our pipeline shows significant improvement of synteny coverage compared to that obtained using the WS215 gene set.

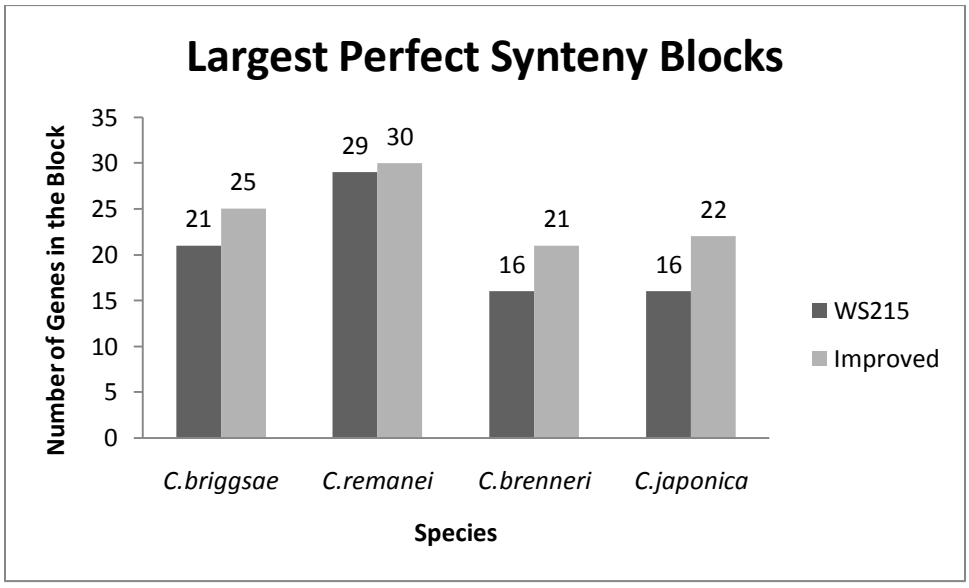

Figure 2-11: *C. briggsae* gene model revision resulted in larger perfect synteny blocks. The largest perfect synteny blocks found between *C. elegans* and the corresponding species has been compared between two sets of gene models. Our improved annotation has increased the number of genes in the largest perfect synteny blocks in all four *Caenorhabditis* species.

2.6 Summary of homology-based gene model revision

In summary, genBlastG was applied to re-annotate four sister species of *C. elegans.* Many defective gene models were fixed and many novel gene models were found, improving the synteny coverage and size. The revised annotations were submitted to Five Worm Genome Analysis Consortium. Although the computationally obtained gene models look promising in terms of genome annotation improvement, they need to be experimentally validated. In the next section, based on these homology-based computationally predicted gene models, our evidence based gene model validation and improvement is described.

3: EVIDENCE-BASED REVISION OF *CAENORHABDITIS* GENE ANNOTATION

3.1 Next-generation sequencing of *C. briggsae* transcriptome and alignment of sequence reads

3.1.1 Transcriptome library preparation and Solexa-sequencing

In this project, two transcriptomes of *C. briggsae* are sampled. One is a L1 stage transcriptome, and one is a mixed stage transcriptome. Both transcriptomes are prepared using the *C. briggsae* AF16 strain and both are sequenced using the Illumina Solexa Genome Analyzer II platform. After sequencing, we obtained 15,469,937 paired-end reads for the L1 transcriptome, 17,486,915 paired-end reads for the mixed-stage transcriptome. Each read is 42 bp in size. Altogether, we have 33,956,852 paired-end reads for validating and revising *C. briggsae* gene models.

3.1.2 Alignment of Solexa paired-end reads to the *C. briggsae* virtual transcriptome

Solexa reads are aligned to the C. briggsae virtual transcriptome built on the hybrid gene set using MAQ. Of 33,956,852 Solexa read pairs, 14,905,545 (or 43.89%) are aligned. I found that 20,181 (or 86.7%) transcripts are supported (read depth >= 2), 1,748 of which are fully (100%) supported.

3.1.3 Alignment of Solexa paired-end reads

3.1.3.1 Alignment of paired-end reads to the *C. briggsae* genome

Solexa reads are aligned to the genome DNA sequence using the program MAQ (Li, Ruan and Durbin 2008) . All parameters are in default except for the following parameters: (a=700, the maximum insert size allowed for correct pairing of reads; n=3, the maximum number of mismatches allowed in the first 28 bp of the Solexa read alignment). The default maximum insert size (a=200) was not used in order to accommodate the correct pairing of reads that align to neighboring exons, which are separated by introns. Both ends of most (13,959,652, or 42.35%) read pairs are mapped to the reference (WormBase, WS215) genome, suggesting that they are either mapped to a same exon, or adjacent exons. However, for many (3,101,911 or 9.13%) read pairs, only one end of the read pair is mapped while its mate is not. MAQ codes such read pairs as “64/192” read pairs. Read ends with code 64 are mappable and their mates with code 192 are unmappable. Together, 48.42% of the Solexa read pairs are not mapped to the genome.

The Solexa reads may be unmappable to the genome because of the following reasons. During the cDNA preparation step, there may have been bacterial sequence contamination. Secondly, there may be sequencing errors producing low quality reads. Thirdly, the read sequences may be low complexity, i.e. repetitive sequences, which reduces the mapping quality by increasing the probability of a mapping location being incorrect. Furthermore, the *C. briggsae* genome assembly currently contains gaps. These gaps may contain highly

expressed genes, which would produce many read sequences that cannot be mapped to the genome because the assembly is incomplete. Finally, there may be some paired-end reads such that both ends map to exon-exon junctions. Previous studies that align short transcriptome reads to the genome report that big portions of the short reads are not mapped to the genome either. Percentage of unmapped reads Berger *et al* report that 32% of the reads from melanoma transcriptome weren't mapped to human genome (Berger, et al. 2010). In a study analysing the HeLa S3 transcriptome this percentage was reported to be 34% (Morin, et al. 2008). Furthermore, Nagalakshmi *et al* report that 44% of the short reads from the yeast transcriptome couldn't be mapped to the yeast genome (Nagalakshmi, et al. 2008).

The collection of all the reads that are mapped by MAQ is useful for confirmation or improvement of predicted gene models. Confirmation or addition of new coding exons to the predicted gene models depend on existence of Solexa reads alignment to the genomic region of interest. Genomic segments in which all bases are covered by Solexa read alignments, which hereby is termed "Solexa exons", can be obtained by first running MAQ's "assemble" function to get the consensus sequence from the reads mapped to the genome. Then the Solexa exons are parsed from the consensus sequence.

3.1.3.2 Further alignment of code 192 reads using cross_match

Read pairs with one end not mappable to the genome suggest the existence of cis-splicing or trans-splicing events (Blumenthal 2005). The ends that are not aligned using MAQ, which are annotated as code 192 ends, can be

further aligned to the reference genome using a local alignment tool such as cross_match (Green 1993). Such local alignments can be used to define cis-splicing sites (described below) and trans-splicing sites (which is described in Section 3.4).

3.1.4 Intron prediction using Supersplat

Local alignment of code 192 read ends can be used to define splicing sites. In fact, many programs have been designed to exploit such local alignments to define introns. Here we will apply SuperSplat, which uses Solexa reads as input (Bryant Jr., et al. 2010), to define introns. After initial mapping by MAQ, I found 3,101,911 code 192 reads and use them, together with their corresponding target genomic regions, as input to SuperSplat. The target genomic region is found by extending the alignment of region of the mapped end by 2,500 bp towards both 5' and 3' directions. The unmapped reads are subject to spliced alignment by SuperSplat. SuperSplat splits these reads into 2 segments by finding a match for each segment separated by a distance defined by "maximum/minimum intron length" parameters. The region between the two segments defines a putative intron. As SuperSplat doesn't depend on canonical splice sites, from the putative introns reported by SuperSplat, we select only those introns which have canonical splice sites. Because a putative introns can be supported by independent Solexa reads, each intron is annotated with the number of supporting Solexa reads.

For convenience, an intron defined by Solexa reads, as described above, is called a "Solexa intron". Based on their relationship with predicted protein-coding

gene models, Solexa introns can be categorized as intragenic intron if it overlaps with a gene and is entirely nested within a predicted gene model. Otherwise, it is categorized as boundary or intergenic introns.

Altogether, 78,252 Solexa introns are detected, 59,560 of which are supported by two or more Solexa reads.

3.2 Internal revision of gene models

3.2.1 Problem statement

Protein-coding genes in eukaryotic genomes consist of exons, introns, and untranslated regions (5' and 3' UTRs). In this section, I will focus on the internal components of genes, in particular, internal introns. Once internal introns are defined, exons are readily defined. For gene models predicted using computer programs, exons and introns may be correct, incomplete or missing. Because exons and introns are complementary, once introns of a protein-coding gene are defined, its internal exons are readily defined. The goal of this section is to apply Solexa reads to validate the introns and revise them if they are not correct. When compared to Solexa introns, introns in predicted gene models can be confirmed, modified, or removed. Additionally, novel introns may be introduced to the gene models. Furthermore, introns in predicted gene models can be spurious and will be removed if their existence is in conflict with transcript reads. Finally, alternative introns that overlap with each other can be identified as well.

3.2.2 Algorithm and flowchart

As illustrated in Figure 3-1, Solexa reads are aligned to the genome DNA sequence using the program MAQ (Li, Ruan and Durbin 2008)

In this section, we only use intragenic Solexa introns for validating and revising predicted gene models. An intragenic Solexa intron can be a perfect match or a partial match to a predicted intron, or a novel intron.

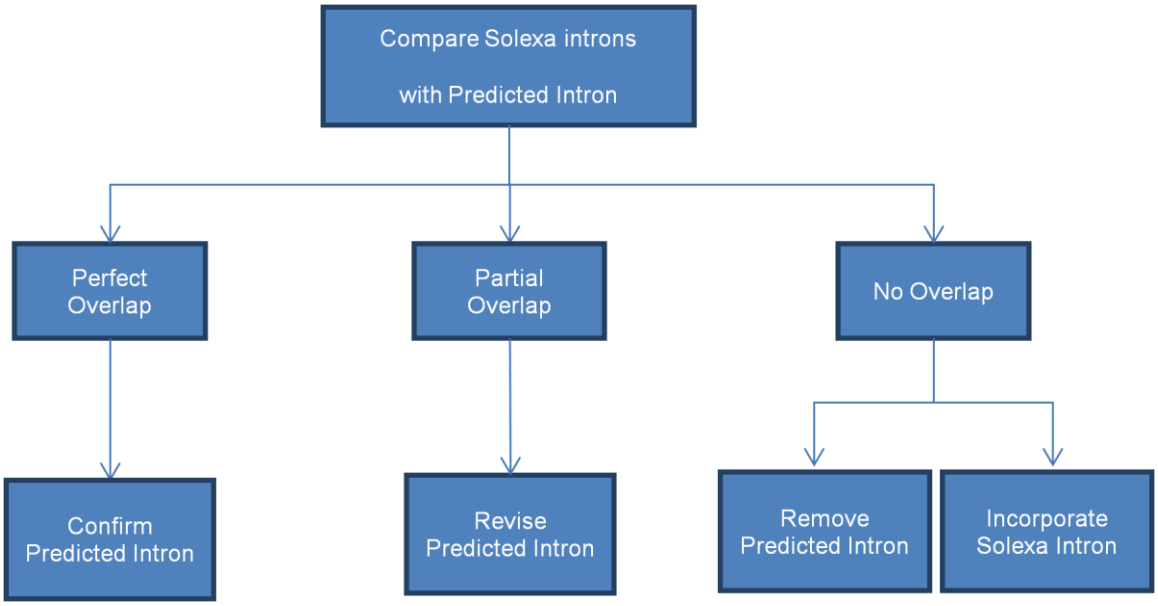

Figure 3-1: Internal revision pipeline. Solexa introns are annotated by spliced alignment of reads. For internal revision of gene models, the genomic positions of such Solexa reads are compared with those of predicted introns. If the compared genomic positions perfectly overlap, the predicted intron is confirmed; if there is a partial overlap, the predicted intron is revised by the Solexa intron; if there is no corresponding predicted intron for Solexa introns, novel Solexa introns are introduced into the gene model; if the predicted intron doesn't have any corresponding Solexa introns and the genomic region spanned by the predicted intron is fully covered by Solexa read alignments, the predicted intron is removed from the gene model.

3.2.2.1 Intron confirmation

If a predicted intron is identical to a Solexa intron that is supported by one or more independent Solexa reads, we conclude that this predicted intron is

confirmed (or validated). If a validated intron overlaps with one or more different Solexa introns (Figure 3-2), these introns are recorded as alternative Solexa introns. Alternative Solexa introns suggest alternative isoforms of a same gene. The goal of this study is to identify one transcript per gene, hence the presence of alternative isoforms is recorded but the isoforms are not reconstructed. These alternative Solexa introns will be valuable for further defining of full-length isoforms in the future.

Among 102,406 predicted introns in the *C. briggsae* hybrid gene set (based on WS215 and genBlastG version 135), 59,137 (or 57.5 %) predicted introns are confirmed by Solexa introns. These validated introns fall into 14,703 protein-coding genes. Therefore, we have detected expression of 63.2% of the predicted genes in the *C. briggsae* hybrid gene set. Out of 21,768 genes with at least 1 intron, all introns in 6,546 (or 30.07%) genes are fully confirmed (See Table 4). In contrast, 10,079 out of 59,560 (or 16.92%) Solexa introns that are supported by two or more Solexa reads do not match fully with predicted introns, suggesting that many introns in the hybrid gene set need to be revised or introduced. We expect that genes whose entire introns are supported by Solexa introns will be even higher (than 30.07%) when these Solexa introns are used for revising the hybrid gene models.

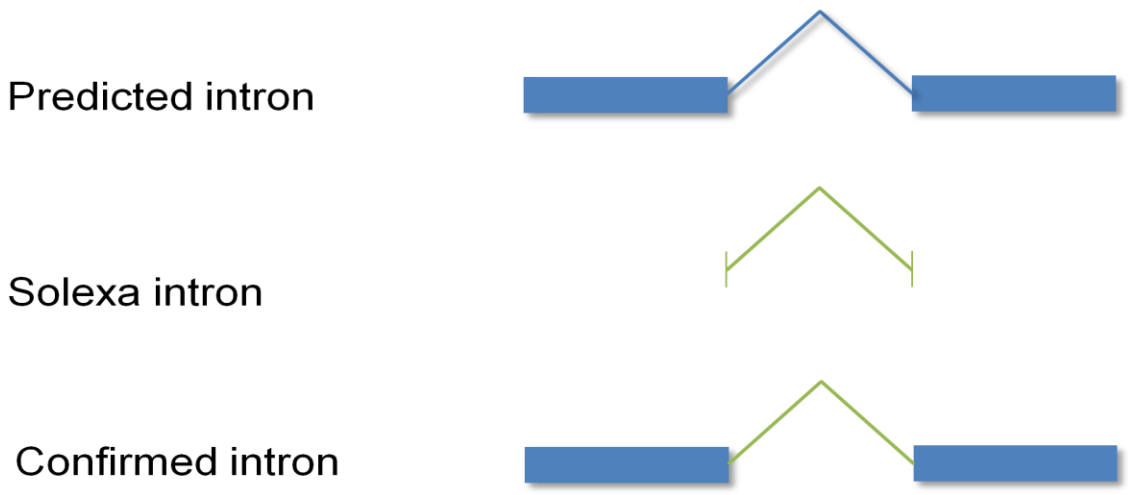

Figure 3-2: Confirmation of a predicted intron. If one Solexa intron (read support >= 1) perfectly overlaps with an predicted intron, predicted intron is defined as confirmed. If this intron overlaps with other Solexa introns, these Solexa introns are annotated as alternative introns.

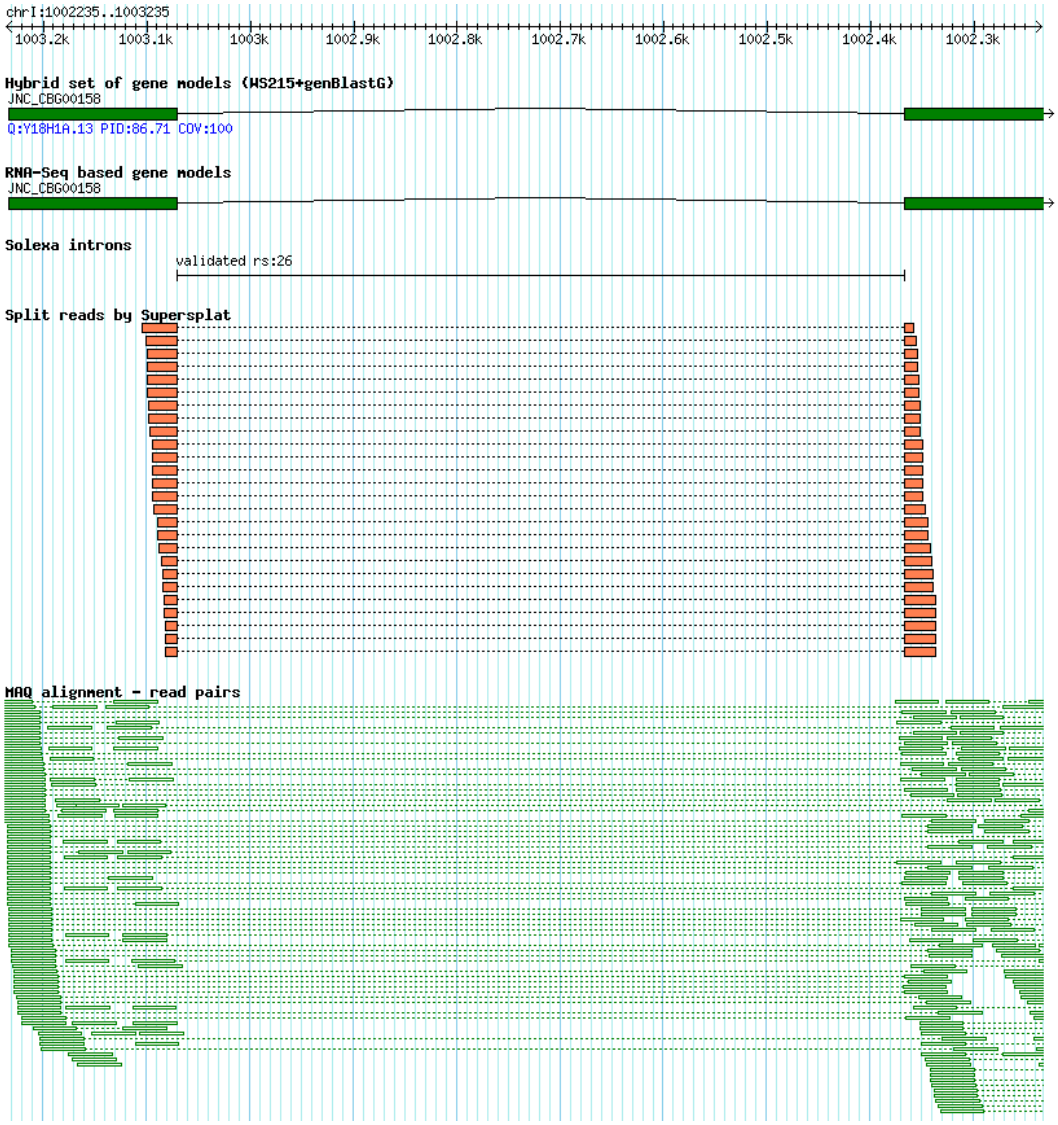

Figure 3-3: An example of a confirmed predicted intron. This intron in JNC_CBG00158 matches perfectly with a Solexa intron that is supported by 26 independent split Solexa reads

3.2.2.2 Intron revision

If a predicted intron overlaps with one Solexa intron that is supported by two or more independent Solexa reads but is not identical to any Solexa inton, this predicted intron will be revised. If the predicted intron overlaps with multiple overlapping Solexa introns, the one with the highest read support is selected for

the revising process. As shown in Figure 3-4, when a Solexa intron is used to replace the predicted intron, the flanking exons will be altered to create splice-junctions for the revised gene model. Here we enforce that the length difference between the predicted intron (which is to be replaced) and the Solexa intron must be a multiple of 3 so that the reading frame is not shifted and the introduced coding region does not contain stop codons. Furthermore, we enforce that the newly introduced coding regions must be supported by Solexa reads (at least 90% of the length of the new coding regions must have read support).

We find 6,617 Solexa introns that overlap with but not identical to one or more predicted introns in the *C. briggsae* hybrid gene set. 2,111 (or 31.90%) of these Solexa introns successfully replaced 2,244 predicted introns. On the other hand 2,301 (or 34.77%) of these introns were not incorporated because they overlapped with other Solexa introns, which had higher support. Such intron revision has been done in 2,077 (or 8.9%) predicted *C. briggsae* gene models in the hybrid gene set.

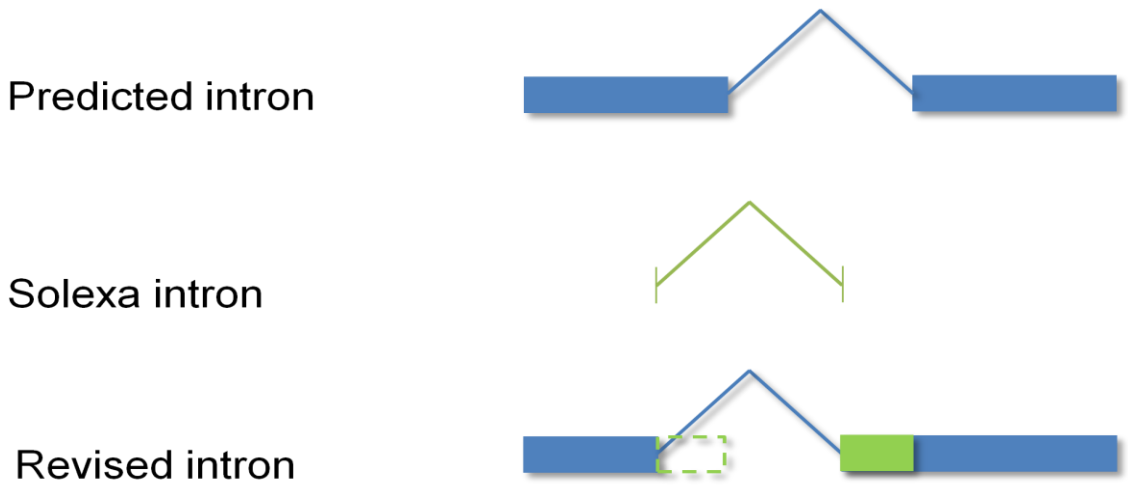

Figure 3-4: Revising predicted introns. If one Solexa intron partially overlaps with existing predicted intron and the predicted intron is not fully supported by any Solexa read, the predicted intron is revised. The revision of the intron should not alter reading frame. If multiple overlapping Solexa introns exist, the one with the highest number of support is selected.

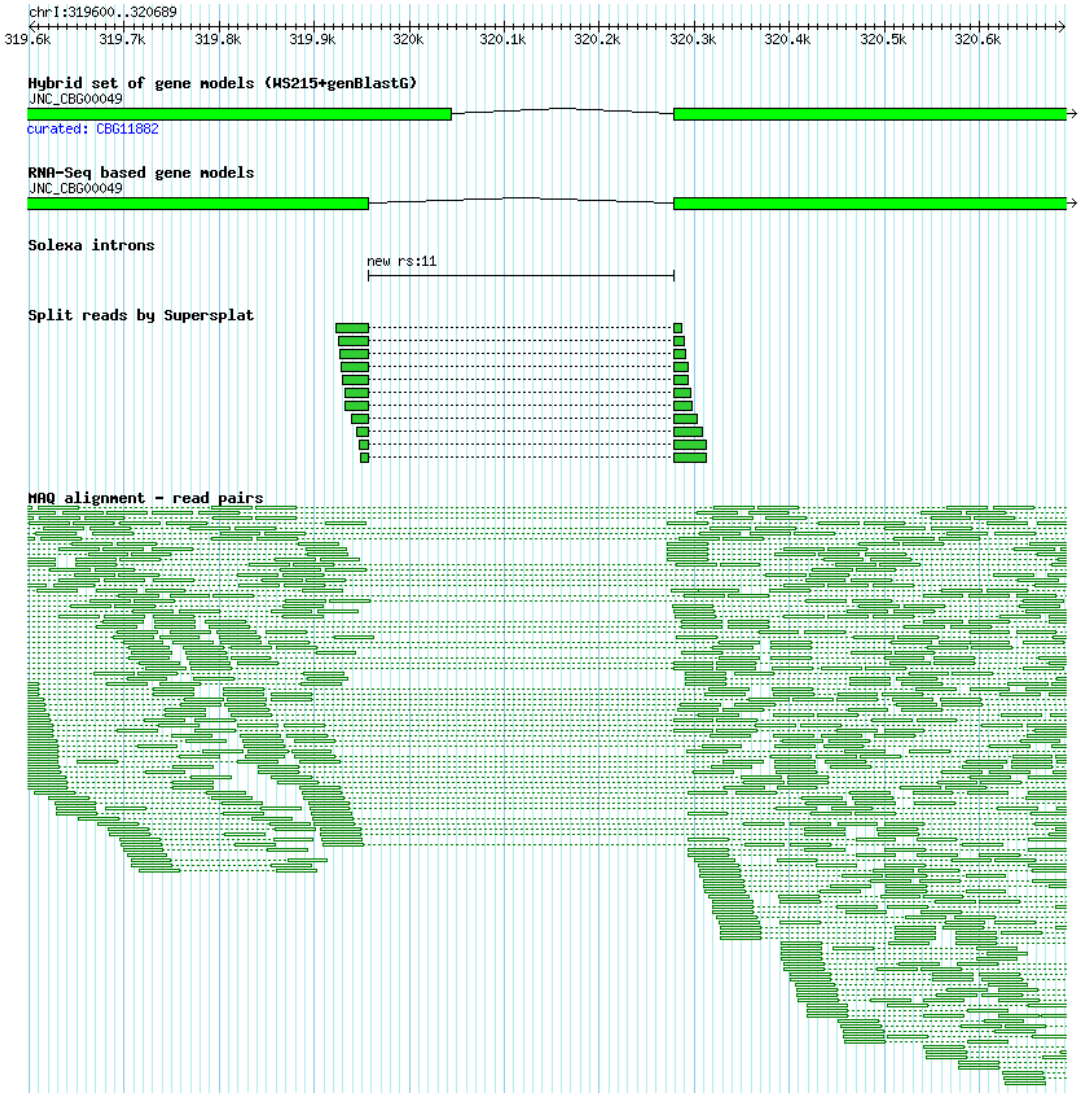

Figure 3-5: An example revised intron. This Solexa intron, which is supported by 11 independent Solexa reads, overlaps with a predicted intron. However, the predicted intron is not fully supported. The predicted intron is thus replaced by the Solexa intron. This replacement does not cause reading frame shift.

3.2.2.3 Novel intron incorporation

Among 59,560 Solexa introns that are supported by two or more independent Solexa reads in *C. briggsae*, 878 (or 14.72%) overlap with predicted coding exons (but not predicted introns). Such Solexa introns are introduced as novel introns if their incorporation does not shift reading frame (Figure 3-6). In

other words, the lengths of such introns should be multiples of 3. If such a novel Solexa intron overlaps with other Solexa introns, only the one supported by the largest number of independent Solexa reads is incorporated.

We incorporated 716 out of 878 novel introns in 638 genes. These novel introns range from 39 bp to 927 bp (average=51.7bp) in size. The rest of these novel introns (162 out of 878) weren't incorporated in any gene models because they either cause frame-shift or have lower support.

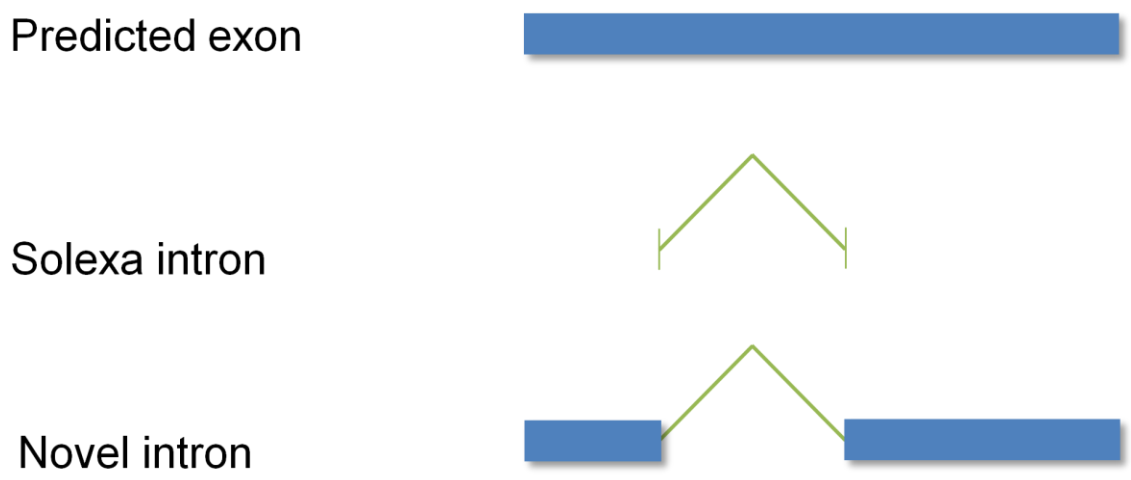

Figure 3-6: Incorporation of a novel Solexa intron into a gene model. If a Solexa intron overlaps with a predicted coding exon and if the intron length is a multiple of 3, the Solexa intron is incorporated into the gene model.

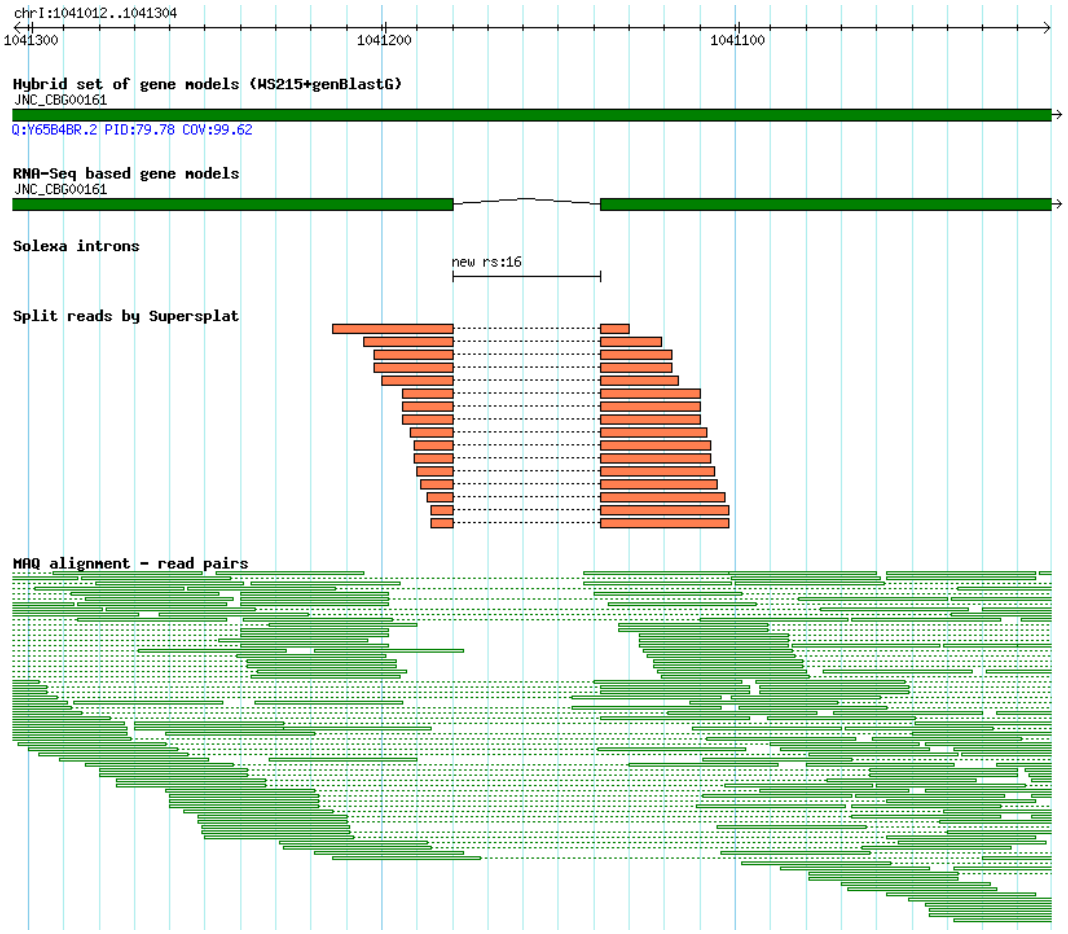

Figure 3-7: An example of novel intron incorporation into a predicted gene model. A Solexa intron, which is supported by 16 independent Solexa reads, is incorporated into the gene model JNC_CBG00161. Also note that the corresponding genomic region is not supported by any Solexa reads.

3.2.2.4 Intron removal

If a predicted intron does not have any supporting evidence (i.e., it is not identical to any Solexa intron, and it does not overlap with any Solexa intron) and the corresponding genomic region is covered by Solexa reads alignment, the predicted intron is removed from the gene model. In the meantime, the corresponding genomic region is converted into a coding exon. For an intron to be removed, we enforce that the length of the intron is a multiple of 3 so that the

incorporation of the new coding region doesn't shift the reading frame. Also, new coding region must not contain any stop codons.

In *C. briggsae*, we removed 461 such predicted introns in 392 gene models. These introns range from 18 bp to 1059 bp (average=106.05bp) in size.

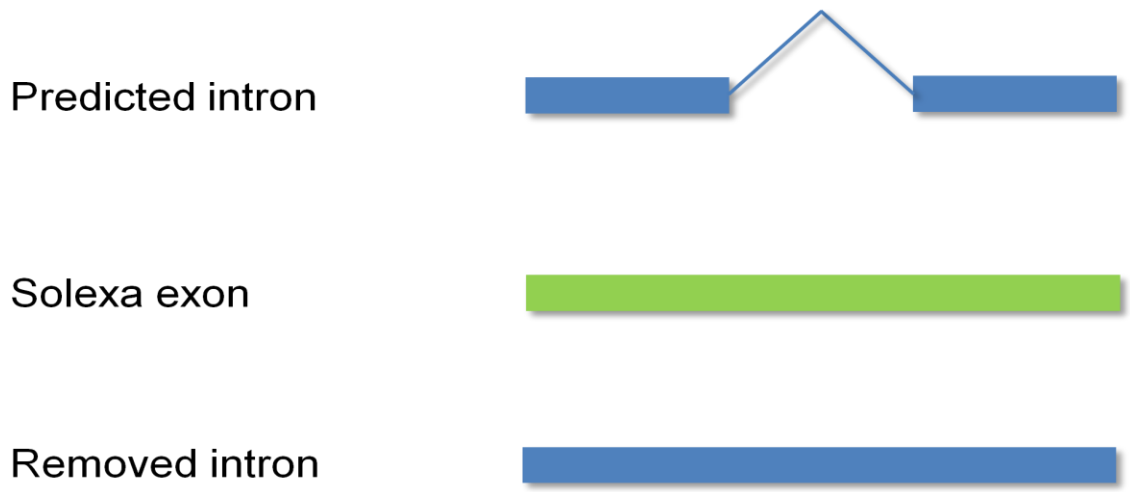

Figure 3-8: If a predicted intron doesn't overlap with any solexa introns and solexa reads cover the region spanned by the intron, predicted ntron is converted to a coding exon if the length of the intron is a multiple of 3.

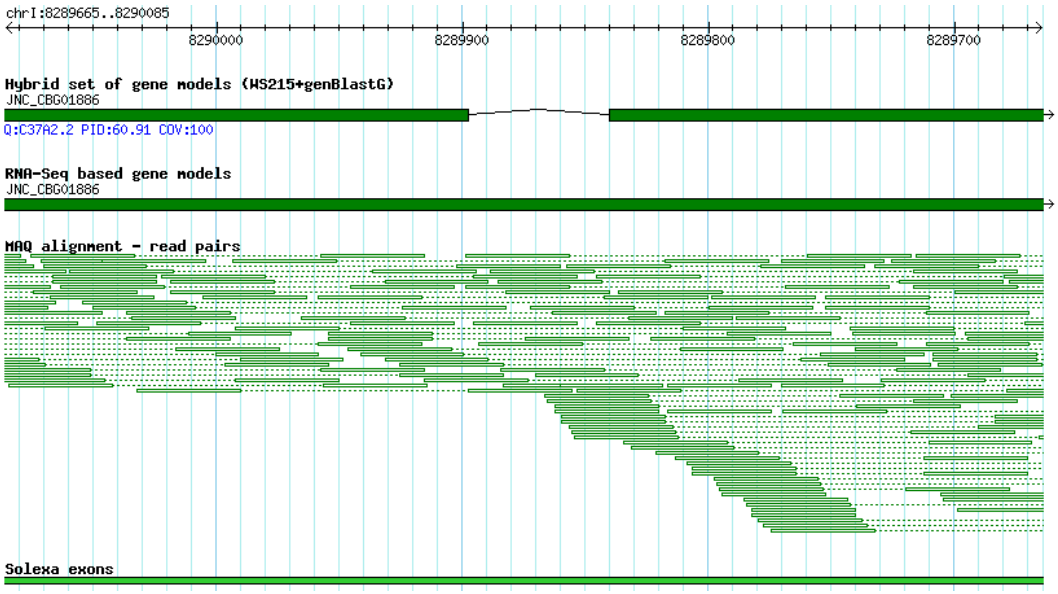

Figure 3-9: An example of a removed spurious intron. This predicted intron is not supported by any Solexa intron. On the contrary, the corresponding genomic region is covered entirely by Solexa reads. The removal of this intron does not alter the reading frame.

3.2.3 Summary of revisions

When the Solexa introns (78,252) were compared to the predicted introns in the hybrid *C. briggsae* gene set, 15,200 (65.3%) gene models are found to overlap with at least one Solexa inton. Among these gene models, introns of 10,335 gene models that overlap with Solexa introns are identical to their corresponding Solexa introns, thus require no revision. On the other hand, 4,865 gene models contain introns that need to be revised (3,689), or can be alternatively transcribed (1,176) because they possess at least one alternatively transcribed intron.

Of 3,689 gene models that need resiving, 3,194 of these gene models only overlapped with intragenic Solexa introns, thus required internal revision,

while 495 gene models only overlapped with Solexa introns at the 3' or 5' boundaries and their revision will be described in the following section.

These 3,194 gene models that need revising overlap with 5,453 Solexa introns that suggest revision and are non-overlapping with each other. These Solexa introns have three fates: (1) 2,827 of these introns were successfully inserted into 2,077 different models. Out of the 2,827 introns, 716 of them were inserted into coding exons, thus they were novel introns, while 2,111 were used to replace 2,244 predicted introns: (2) 2,018 of them failed to be incorporated into gene models because of lack of read support for the newly introduced coding regions; (3) 636 Solexa introns were not incorporated into gene models because their incorporation would cause frame shift.

Additionally, 461 predicted introns were removed and converted into coding exons in *C.briggsae.*

Table 3: Summary of the internal revision (Introns)

	Confirmed introns	Replaced introns	Novel introns	Removed introns
Introns	59,137 (57.5%)	2,244	716	461

Table 4: Summary of the internal revision (Genes)

	At least 1 intron confirmed	All introns confirmed	Genes with replaced introns	Genes with Novel introns	Genes With Removed introns
Genes	14,307 (67.5%)	6,546 (30.1%)	1,635	638	398

3.2.4 Discussion

Throughout the internal revision process, a conservative approach was taken not to introduce new features that could introduce more defects into the gene model than the fixed ones. The changes made to the gene model for each insertion was recorded and insertion was not allowed if the reading frame is disrupted, newly introduced coding regions were not supported, or a premature stop codon is added to the gene model.

The decision to insert a Solexa intron into a gene model is made independently from the decisions for other Solexa introns. In other words, if there are multiple Solexa introns that suggest a revision is required at different parts of the gene model, the decision for each part is made independently from other parts. For example, the insertion of 3 different introns all of which shift the reading frame by 1 bp would not change the overall reading frame when all of them are inserted. However, we make the insertions one intron at a time. Thus, none of these 3 introns would be allowed to be inserted according to our pipeline.

There were further modifications, which could have been done for internal revisions of gene models but weren't done due to the time constraints. These changes are summarized below.

Multiple non-overlapping Solexa introns that overlap with the same predicted intron can be inserted into the gene model all at once. The current procedure is not suited to multiple insertions at a time.

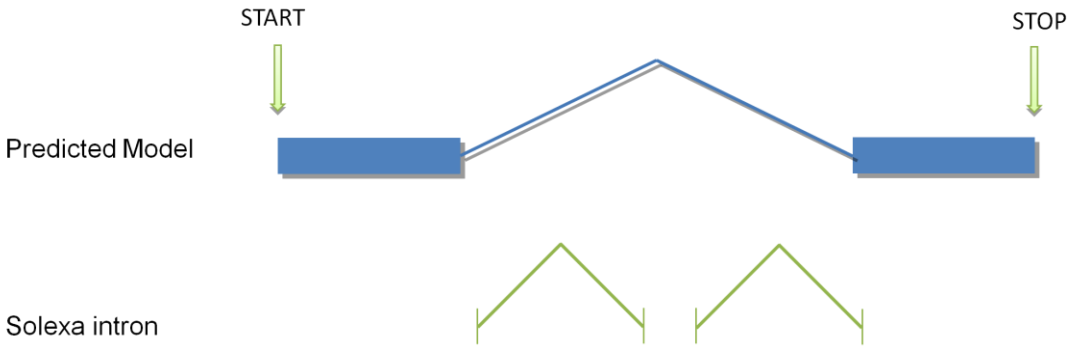

Figure 3-10: If there are multiple non-overlapping Solexa introns that overlap with the same predicted intron, then they must replace the predicted intron all at once if the change in the coding sequence (|B+C-A|) is a multiple of 3 and the added coding sequences (Y and Z) are supported by Solexa reads. In this case, insertion of introns I1 and I2 one at a time fails for both introns because of lack of read support for the flanking region where the next Solexa intron (I1 or I2) is located.

When there are 2 or more overlapping Solexa introns which are used to introduce novel introns or revise predicted introns, the intron with the highest read support among the overlapping introns is selected to make the revision. However, if that intron fails to be inserted into the gene model, the insertion of the introns with fewer read support is not further considered.

There are many cases where Solexa introns suggest alternative splicing. However, alternative isoforms are not created and only 1 isoform is provided for each gene.

3.3 Boundary revision/extension of gene models

3.3.1 Problem statement

A gene model is extended at 3' end if the following conditions are met:

The gene model lacks a stop codon and Solexa exons at 3' end either contains a stop codon or connects the gene model to a neighboring Solexa intron or predicted intron. The gene models that are predicted by genBlastG always contain stop codons. However we've observed that some models in the current annotation of WormBase don't have stop codons. In this project we attempt to identify a stop codon for all gene models.

When a Solexa intron overlaps with the 5'/3' terminal (i.e., boundary) exon of a predicted gene model, which suggests that the predicted gene model should be extended at 5'/3', the Solexa intron will be incorporated into the predicted gene model. Thus the overlapping terminal exon will be reduced, and the predicted start/stop codons will be redefined.

When a Solexa intron is found in the neighbourhood of the gene model, and a Solexa exon is found to bridge the Solexa intron to the predicted gene model, the Solexa intron will also be incorporated into the predicted gene model. Accordingly, the terminal exon will be extended and the predicted start/stop codons will be redefined.

In a further scenario, a Solexa exon can connect a predicted gene model to a predicted intron or an adjacent predicted gene model suggesting that the two adjacent gene models should be merged into a single gene model. Thus, the Solexa exon is incorporated as an exon (or part of an exon) of the new gene model.

3.3.2 Algorithm and flowchart

3.3.2.1 3’ end extension

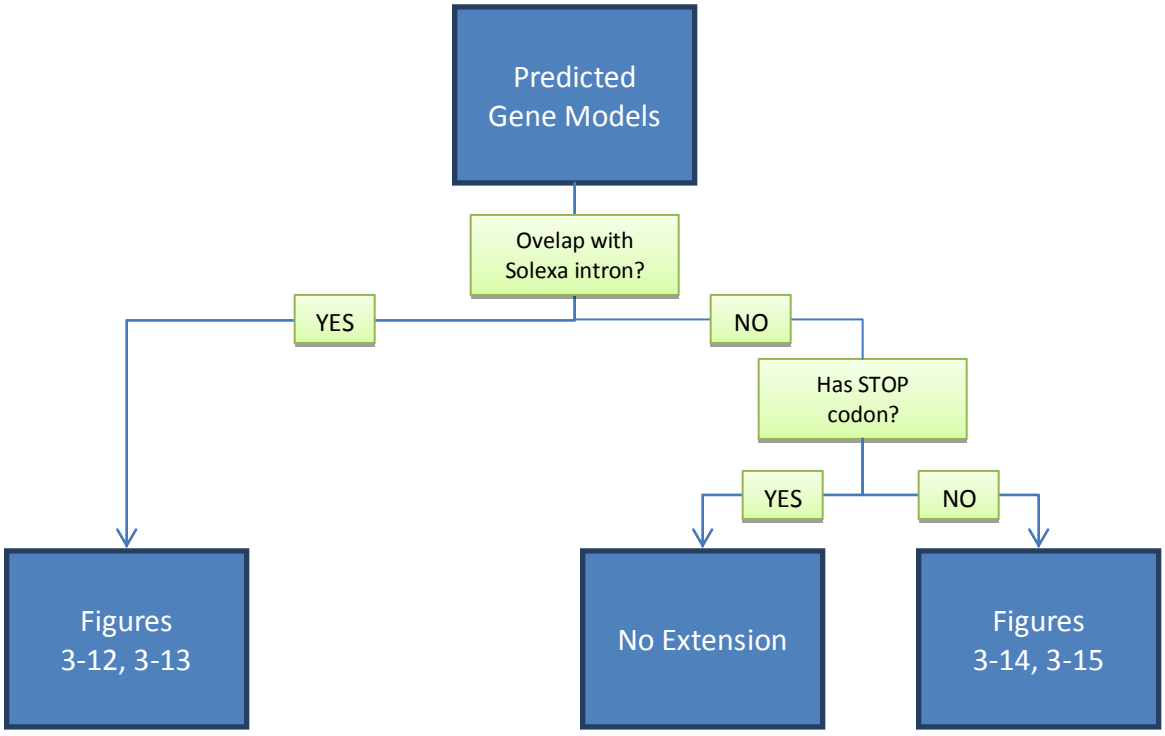

Figure 3-11: Pipeline for 3' Extension of Gene Models

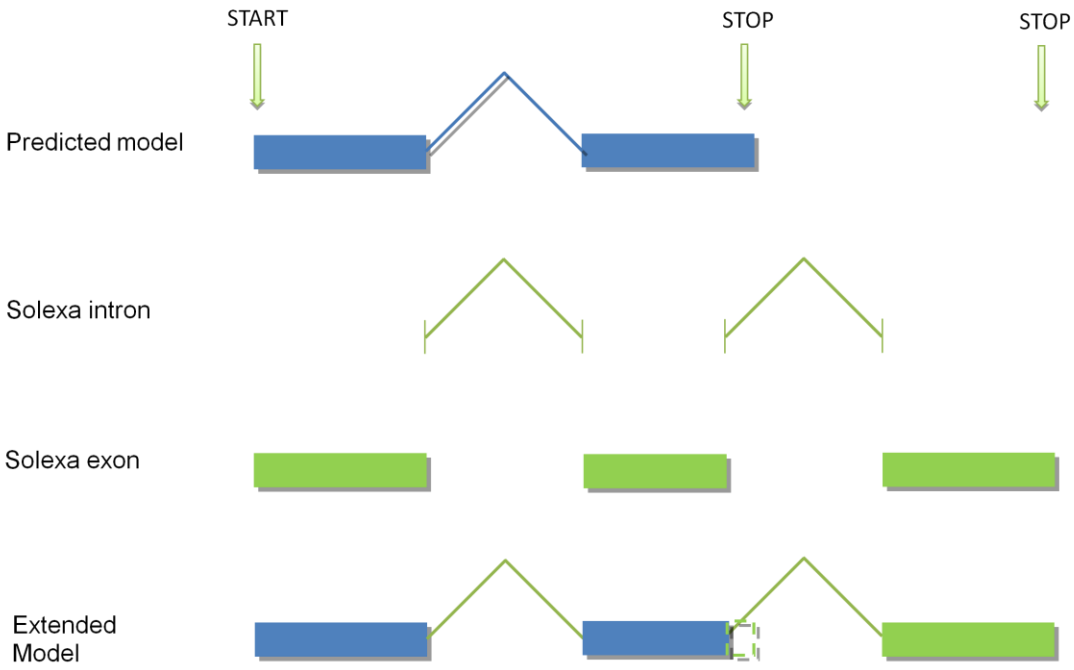

Figure 3-12: Extension of gene models at 3' end with a boundary intron. The gene model overlaps with a Solexa intron at 3' end, suggesting that the predicted stop codon should be removed. In this case, the Solexa intron is incorporated to the predicted gene model and the extension starts from the edge of the Solexa intron.

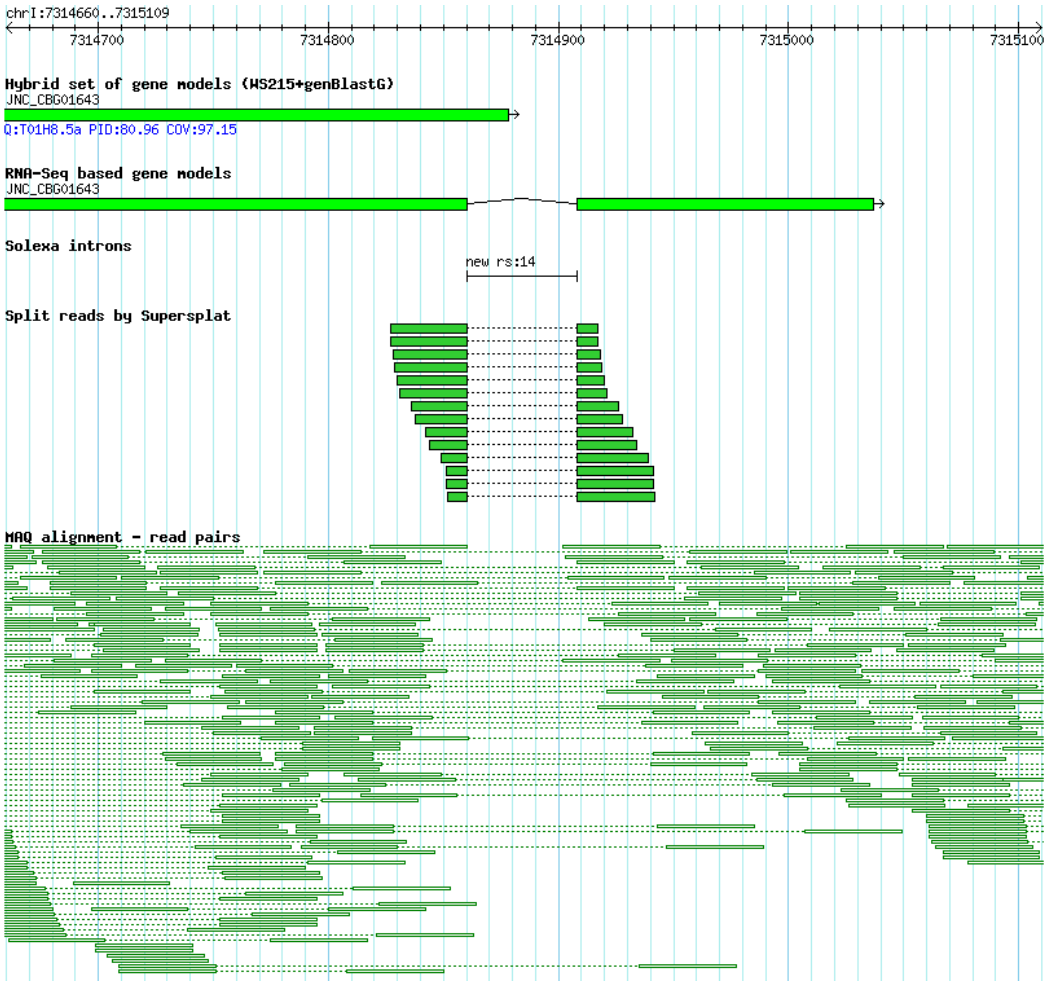

Figure 3-13: Extension of a gene model at 3' end with a boundary Solexa intron. The Solexa intron,which is supported by 14 independent Solexa reads, overlaps with the 3' terminal exon of the predicted gene model JNC_CBG01643. Therefore, the Solexa intron is incorporated as the last intron of the gene model and a new exon is created based on read alignments by MAQ.

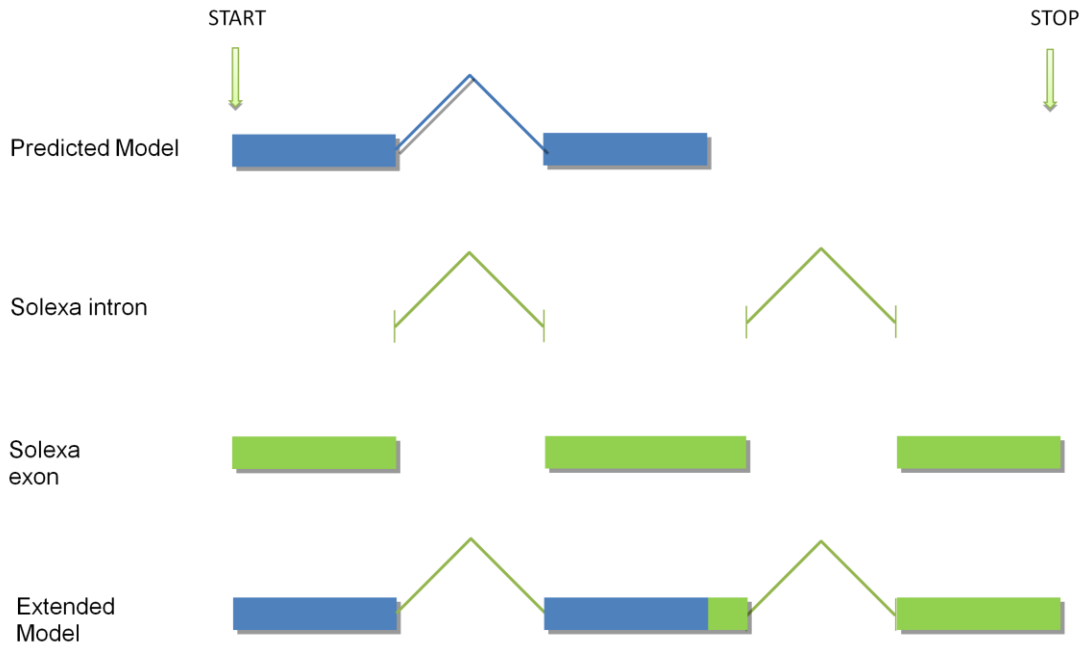

Figure 3-14: Extension of gene models at 3' end without a boundary intron. This case exists only when the predicted gene model lacks a predicted stop codon.

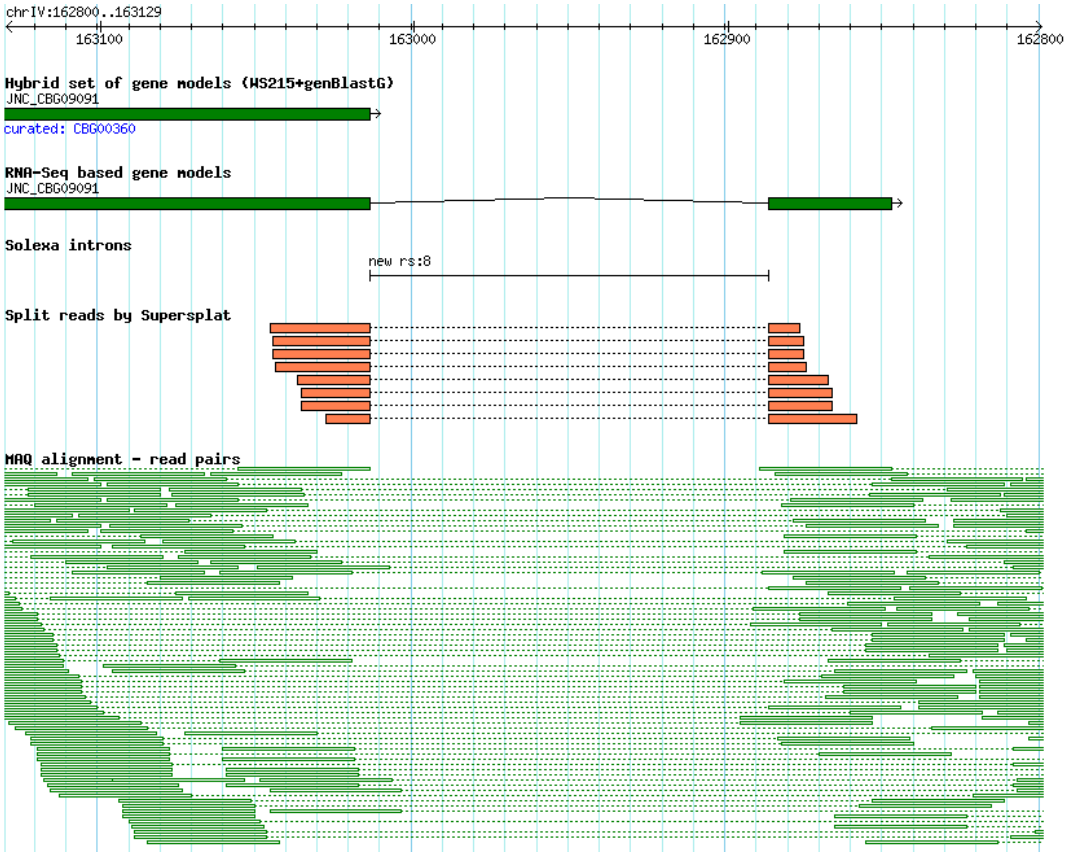

Figure 3-15: The predicted gene model JNC_CBG09091 lacks a stop codon at 3' end. The Solexa intron, which is supported by 8 independent Solexa reads, is located immediately downstream of this gene model and it is incorporated into the gene model. Additionally, a novel exon is added which contains a stop codon.

Solexa introns or predicted introns of neighboring genes are added to the gene model as the extension proceeds. Any Solexa intron that's incorporated must be supported by 2 or more reads. The extension stops after the first in-frame stop codon that's found. Extension is accepted only if the coding exons introduced to the gene model until the new stop codon is supported by Solexa reads.

In *C. briggsae,* 78 gene models without boundary introns were extended at 3' end incorporating 12 introns. 72 of these gene models were only extended

to find a nearby stop codon. Furthermore, 321 gene models with boundary introns were extended at 3' end incorporating 709 introns.

3.3.2.2 5' end extension

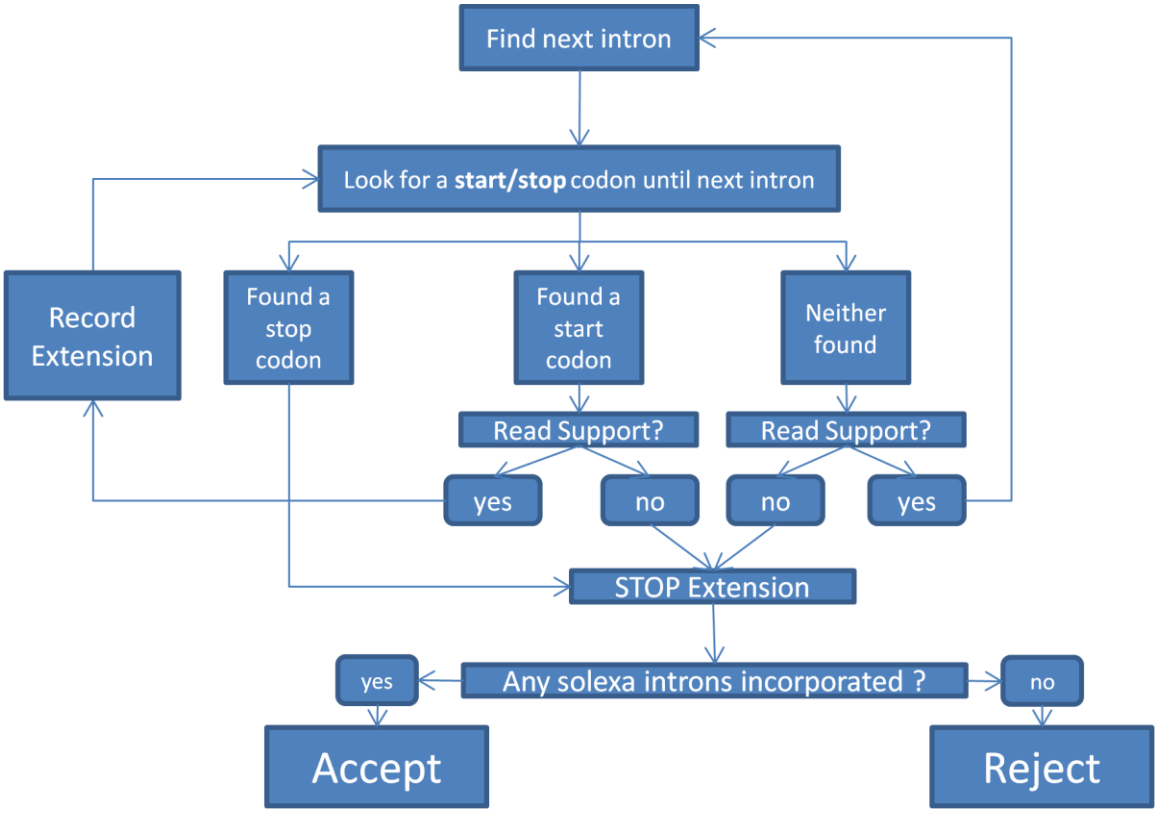

Figure 3-16: Pipeline for 5' Extension of Gene Models

The gene model extension at 5' end is implemented based on a similar idea as the extensions at 3' end. The main difference for 5' extensions is that both "START" and "STOP" codons need to be considered simultaneously. "START" codons are looked for to find the 5' end of the new gene model while "STOP" codons are looked for to ensure that a premature stop codon is not incorporated into the revised gene model.

If there's a Solexa intron that overlaps with the gene model at 5' end or the Solexa exons at 5' end connect the gene model to an upstream Solexa intron or a predicted intron of an upstream neighbouring gene model, then the gene model can be extended at 5' end.

The gene model is extended and the "START" codons found upstream are recorded. Extension proceeds until:

1) A stop codon is found.
2) Solexa exons or existing coding exons of neighbouring gene models don't support the extension.

Thus, the upstream-most “START” codon in the genomic region supported by Solexa reads before the first encountered “STOP” codon is annotated as the new “START” codon.

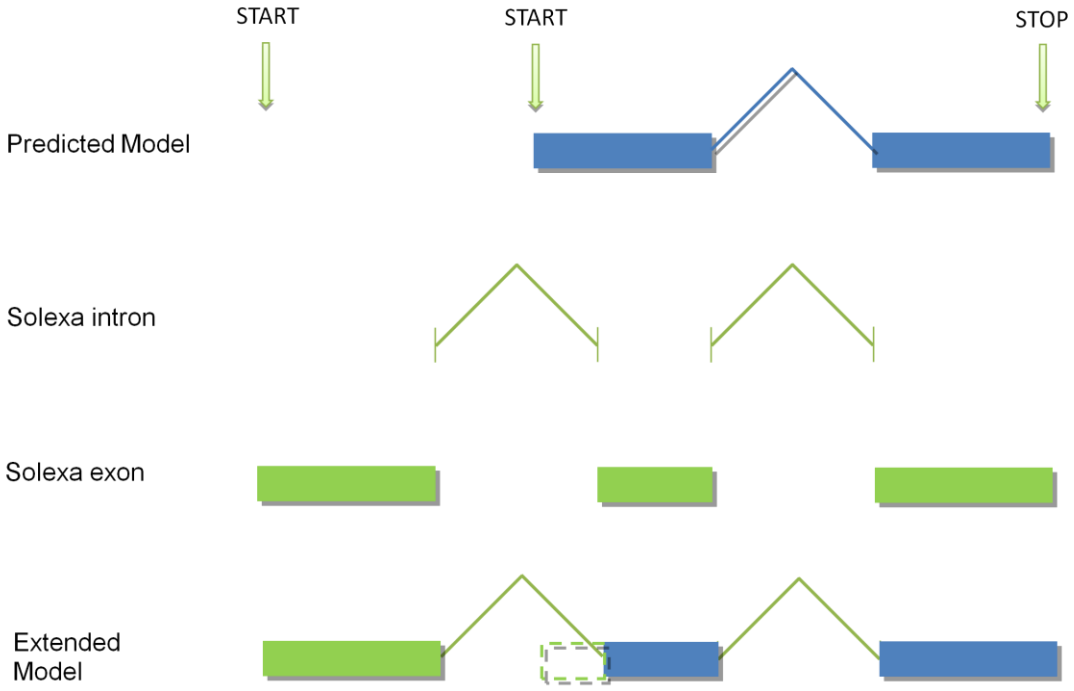

Figure 3-17: Extension of a gene model at 5'end with a boundary intron.

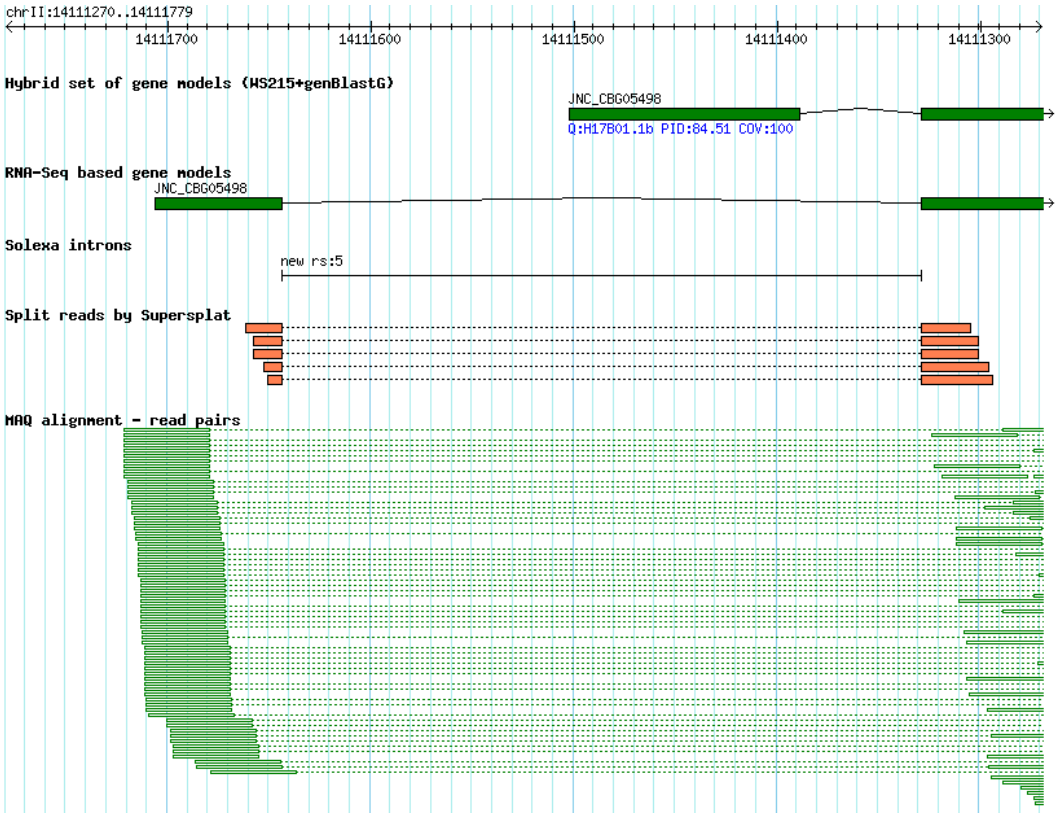

Figure 3-18: The terminal exon of the predicted gene model JNC_CBG05498 overlaps with a Solexa intron, which is supported by 5 independent Solexa reads. The Solexa intron is incorporated into the gene model and a novel exon which contains a start codon is added to the gene model at 5' end.

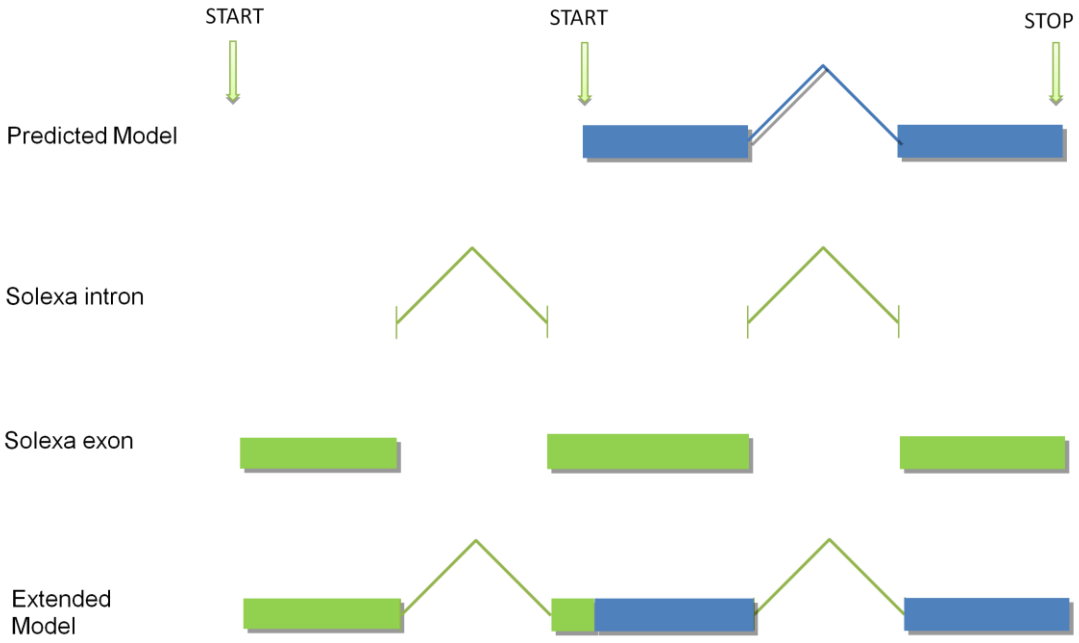

Figure 3-19: Extension of gene models at 5' end with 5' non-boundary Solexa intron.

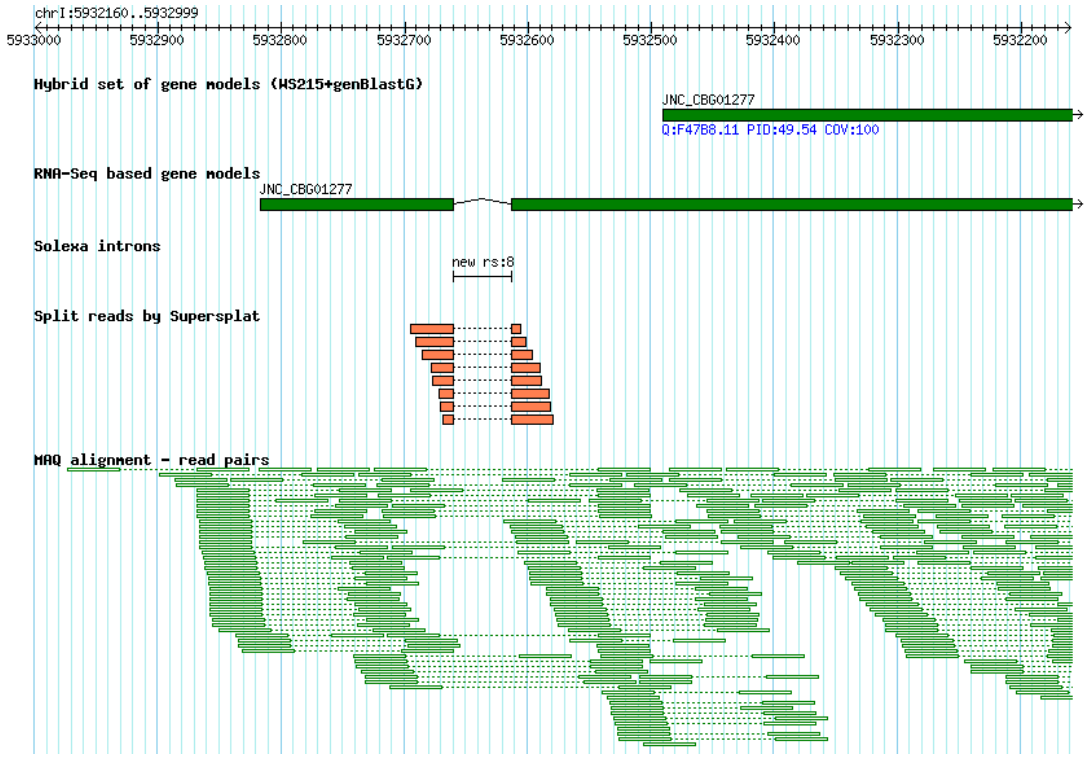

Figure 3-20: The predicted gene model JNC_CBG01277 overlaps with a Solexa exon that links the gene model to an upstream Solexa intron, which is supported by 8 independent reads. Therefore, the gene model is extended at 5' end and the Solexa intron is incorporated into the gene model. A new exon is also added to the gene model that contains a start codon and is supported by Solexa exons.

3.3.2.3 3'UTR

At the 3' end of the gene model, the Solexa exon that covers the "STOP" codon contains the 3'UTR, which starts immediately after the "STOP" codon and ends with the end of the Solexa exon. 3'UTR regions were found for 14,089 genes.

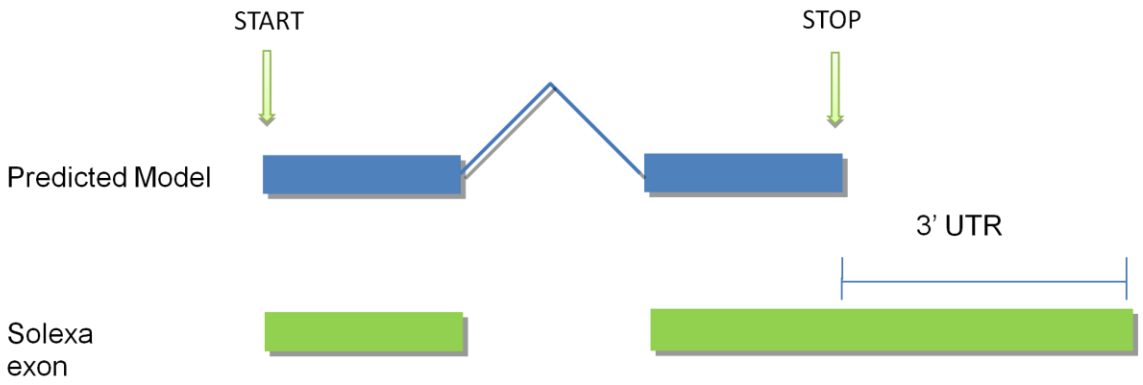

Figure 3-21: 3' UTR regions of genes.

3.3.2.4 5'UTR

The Solexa exon immediately upstream of the 5' end of the gene models contain the 5'UTRs. For trans-spliced genes, the 5'UTR region starts immediately after the trans-splicing acceptor site and ends before the "START" codon. For the genes that are not trans-spliced, the 5'UTR is the region between "START" codon and the start of the Solexa exon that covers the "START" codon. 5'UTR regions were found for 14,089 genes.

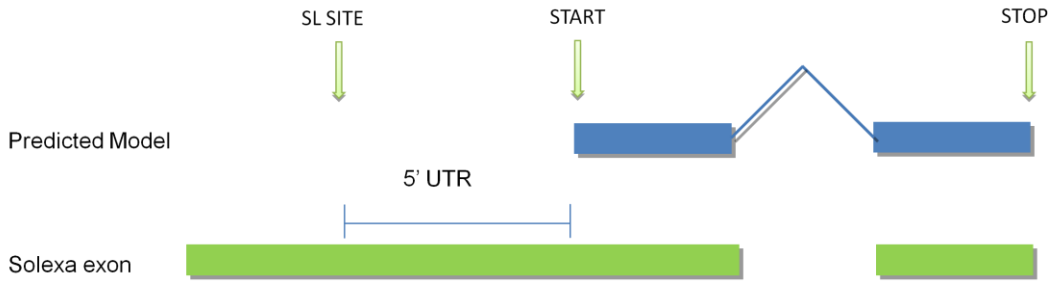

Figure 3-22: 5' UTR regions of genes.

3.3.3 Summary of revisions

As a result of the application of extension procedures to 23,276 gene models in the "hybrid" set for *C. briggsae,* 762 gene models were extended at either 3' or 5' end. 399 gene models were extended at 3' end and 386 gene models were extended at 5' end. 23 gene models were extended at both 3' and 5' ends.

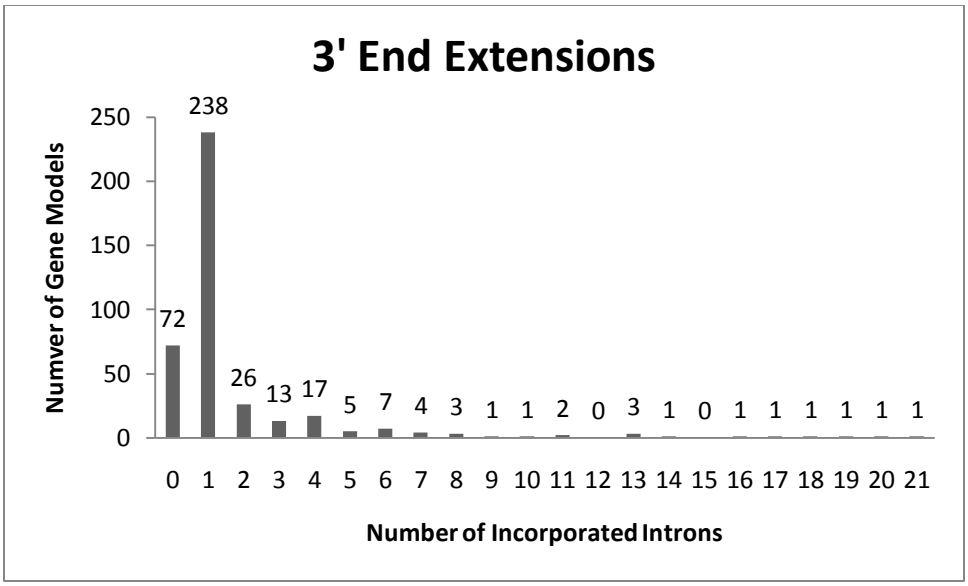

Figure 3-23: 3' Extension of gene models. X axis denotes the number of introns added to a gene model. Y axis denotes the number of models that were extended incorporating that many introns.

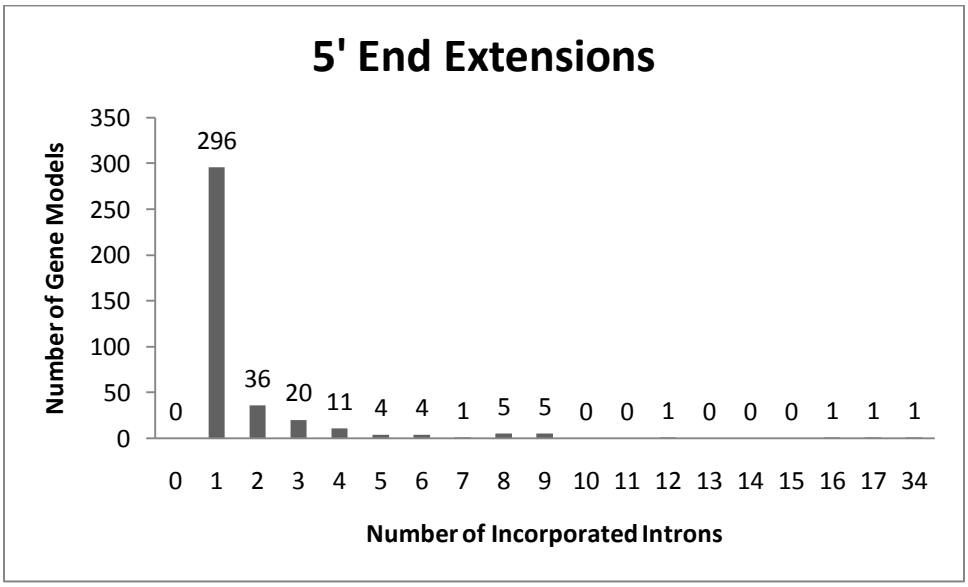

Figure 3-24: 5' Extension of gene models. X axis denotes the number of introns added to a gene model. Y axis denotes the number of models that were extended incorporating that many introns.

3.3.4 Discussion

Throughout the gene model extension pipeline, a conservative approach is taken. New coding sequences are not introduced unless they're found to be supported by Solexa read alignments. Pre-mature stop codon introduction is avoided. For the 5'extensions, if no new introns are added to the gene models, extensions are not allowed. While allowing the usage of predicted introns during the extension process in order to be able to merge gene models where there's enough evidence (e.g Solexa exons connecting the gene models), gene models without boundary Solexa introns are not extended if the 5' end of the gene model overlaps with another gene model. This helps to avoid over-extension of gene models by mere usage of predicted introns.

Similarly to the internal revision process, extension procedure doesn't produce alternative isoforms for a given model. When an intron is looked for adding to the 5' or 3' end of a gene model, the closest intron is selected for extension.

5' UTR and/or 3' UTR regions were defined for 16,408 genes. These UTR regions were defined as the genomic region upstream/downstream of 5'/3' ends, which is completely supported by Solexa read alignments. However, this procedure doesn't fully define UTR regions that may contain introns. As a result of all gene model revisions including homology and RNA-seq based improvements, 60.9 % of all introns were confirmed, 68.3% of all the genes (with >= 1 introns) had at least 1 intron confirmed and 33.9% of all the genes (with >=

1 introns) had all their introns confirmed by RNA-seq data (See Table 5). Furthermore, 10,235 (or 47%) genes were found to have >=95 % of their cDNA sequences supported by Solexa reads alignments (See Figure 3-25).

Table 5: Summary of All Revisions

Gene Set	Genes with Introns	Confirmed Introns	Gene models with at least 1 intron confirmed	Genes with all introns confirmed	Number of Revised Gene Models
Hybrid	21,768	59,137 (57.5%)	14,703 (67.5%)	6,546 (30.1%)	7,806
Solexa Improved	21,683	62,727 (60.9%)	14,812 (68.3%)	7,347 (33.9%)	2,346

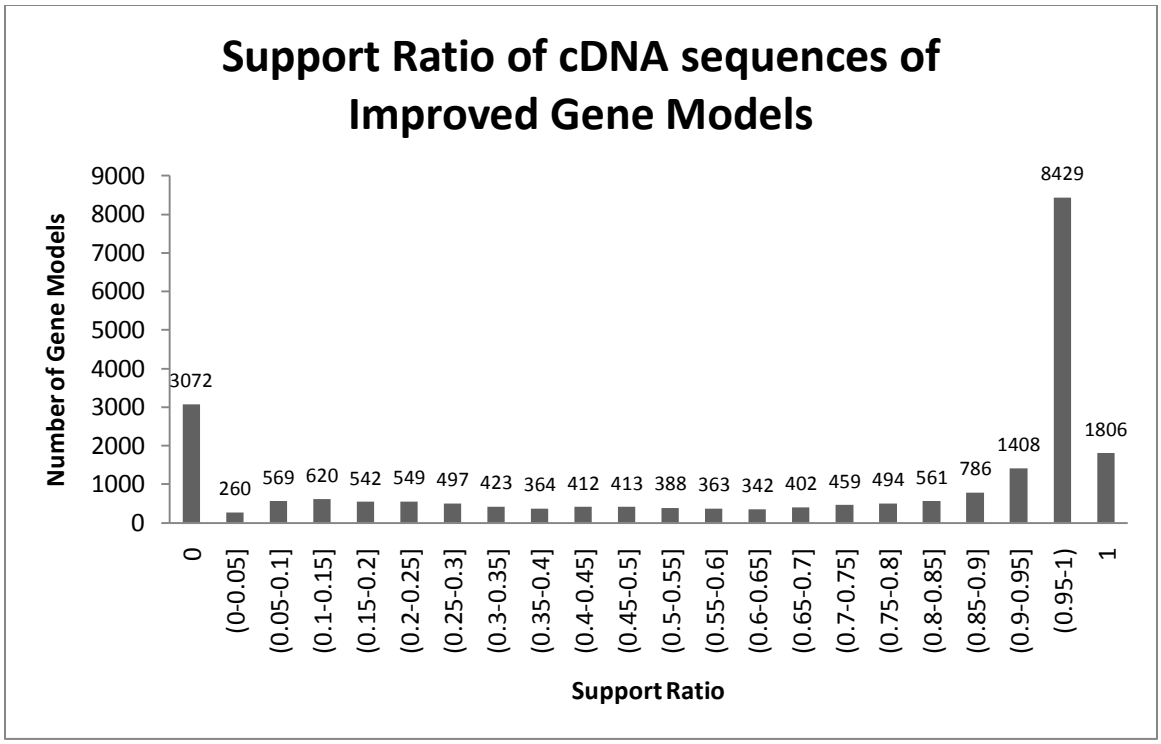

Figure 3-25: Support ratio of cDNA sequences of improved gene model set.

3.4 Trans-splicing site detection & prediction

3.4.1 Problem statement

Maturation process of eukaryotic pre-mRNAs involves removal of introns within splicing complexes. Splicing may occur on a single pre-mRNA molecule (cis-splicing) or two different RNA molecules may contribute to the maturation process by donating sequences for removal of introns (trans-splicing) (Bonen 1993). In *C. elegans,* one of two classes of spliced leader (SL) RNA is added to about 70% of the pre-mRNA molecules to generate a mature mRNA molecule via trans-splicing (Graber, et al. 2007). Trans-spliced mRNA molecules are trimmed off at 5' ends and capped with SL1 or SL2 type sequences (Blumenthal, Trans-splicing and operons 2005).

C. briggsae is among the closest relative of *C .elegans* and trans-splicing is expected to occur in *C. briggsae,* too. However, trans-splicing sites haven't been annotated in *C. briggsae* genome, yet. High throughput sequencing of the transcriptome of *C. briggsae* is expected to produce short reads that contain full or partial spliced leader sequences.

The genomic sequence upstream of the 5'end of the gene models doesn't contain the SL sequences. Therefore, Solexa reads (42 bp long) that contain full SL sequences, which are ~22 bp long, won't be directly alignable to the genome by MAQ because we do not allow so many mismatches for direct mapping. Such reads need to be remapped by a local aligner such as cross_match (Green 1993), a fast implementation of Smith-Waterman local alignment algorithm (Smith and Waterman 1981), to the genome in order to find the ~20 bp hits between the read and the genome. However, reads that contain partial SL sequences will be alignable by MAQ by allowing some mismatches.

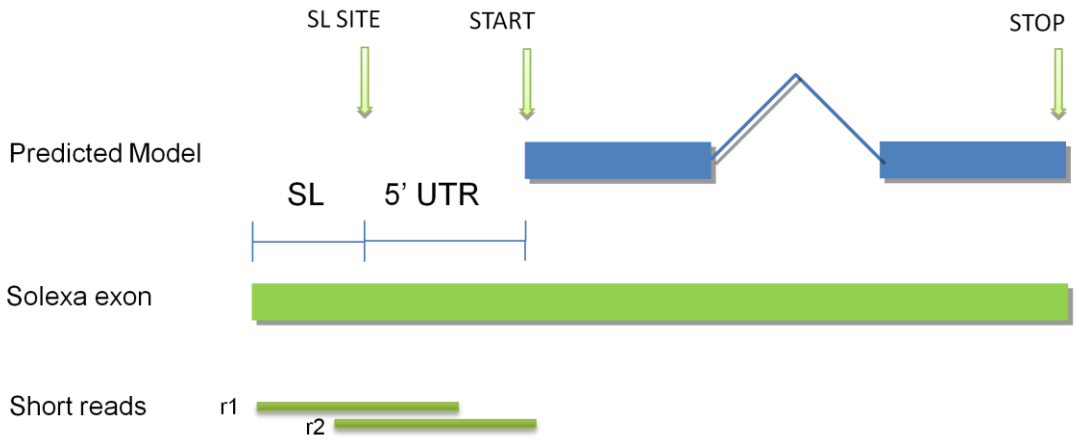

Figure 3-26: Trans-spliced genes. Solexa reads that contain full SL sequences (eg. r1) won't be directly alignable by MAQ. Thus they need to be aligned to the genome by a local alignment algorithm such as cross_match. Solexa reads that contain partial SL sequences (eg. r2) will be alignable by MAQ by allowing for some mismatches.

Solexa reads that are mapped by MAQ/remapped by cross_match to the upstream regions of the gene models are candidates to contain SL sequences. Thus, the problem is to find the candidate Solexa reads and then verify if they contain full/partial SL sequences in order to detect trans-spliced genes. In this work, we detect trans-splicing sites and categorize each site according to the type of SL sequence that's found at the 5'end of the gene models. Also, we provide the number of reads that support each trans-splicing site.

Detecting the trans-spliced genes based on Solexa reads may not help discover all trans-spliced genes because this approach depends on the sequencing depth. It has been shown that *C. elegans* genomic sequences contain motifs that may be used as signals for trans-splicing (Graber, et al. 2007). Thus, even if some genes are not found to be trans-spliced because of

the low read depth, such signals can be defined to predict genes whose mRNA products are subject to trans-splicing. In this work, we define trans-splicing signal motifs based on the detected trans-splicing sites and then use these motifs to predict trans-splicing sites across the genome of *C. briggsae.*

3.4.2 Algorithm

3.4.2.1 Detection of trans-splicing sites

1) Map the paired-end reads to the genome by MAQ.
2) Find unmapped reads whose mates are mapped by MAQ to the genome. Using cross_match, remap those reads to the flanking region where their mates are mapped.
3) For each gene model, check the region 100 bp upstream of 5'end and find reads that are mapped by MAQ/remapped by cross_match to this region.
4) Align the 5' end of these read sequences to the 3' end of the known SL sequences to detect and categorize trans-splicing sites (Guiliano and Blaxter 2006).
5) For each trans-splicing site, group the reads together that support the same site.

3.4.2.2 Prediction of trans-splicing sites

1) Randomly select and manually confirm well supported (>=5 reads) 50 trans-splicing sites.

2) For each such trans-splicing site, grab 8bp length genomic region starting from the trans-splicing site.

3) Generate a positional weight matrix from these 8bp segments.

4) Use the positional weight matrix to scan the100 bp upstream region of all gene models in order to detect genomic segments that match the profile defined by the positional weight matrix. The profile search is done by Motif Occurrence Detection Suite (MOODS) (Korhonen, et al. 2009).

3.4.3 Summary of revisions

11,617 *trans*-splicing sites were detected in the 100 bp upstream region of 8,555 gene models in *C. briggsae* based on Solexa read alignments.

8,856 of these were SL1 type *trans*-splicing sites (7,871 genes) and 2,761 of these were SL2-like (including all SLs from SL3 to SL12) *trans*-splicing sites (2,287 genes).

Table 6 : Trans-Splicing sites detected based on Solexa reads.

SL type	Sites	Genes
SL1*	8,856	7,871
SL2	1,505	1,386
SL3	21	21
SL4	795	762
SL6	35	34
SL7*	1,603	1,567
SL8	15	14
SL9	988	930
SL10	1,295	1,198
SL11	66	65
SL12	1,810	1,655

*SL1 and SL7 have identical sequence at the last five nucleotides. Ambiguous SL7 sequences were not used for calling operons.

50 of the detected SL1 trans-splicing sites were randomly selected. 8bp windows starting from these sites were grabbed from the genomic sequence and a positional weight matrix was generated which denotes the frequency of the occurrence of the types of nucleotides on each position from 1 to 8. This matrix was used to do a profile scan against the genomic regions 100 bp upstream of the gene models. Using an e-value threshold of 0.001, 11,900 SL1 trans-splicing sites were predicted for a total of 10,018 genes and 11,923 SL2-like trans-splicing sites were predicted for a total of 10,158 genes.

3.4.4 Representative figures

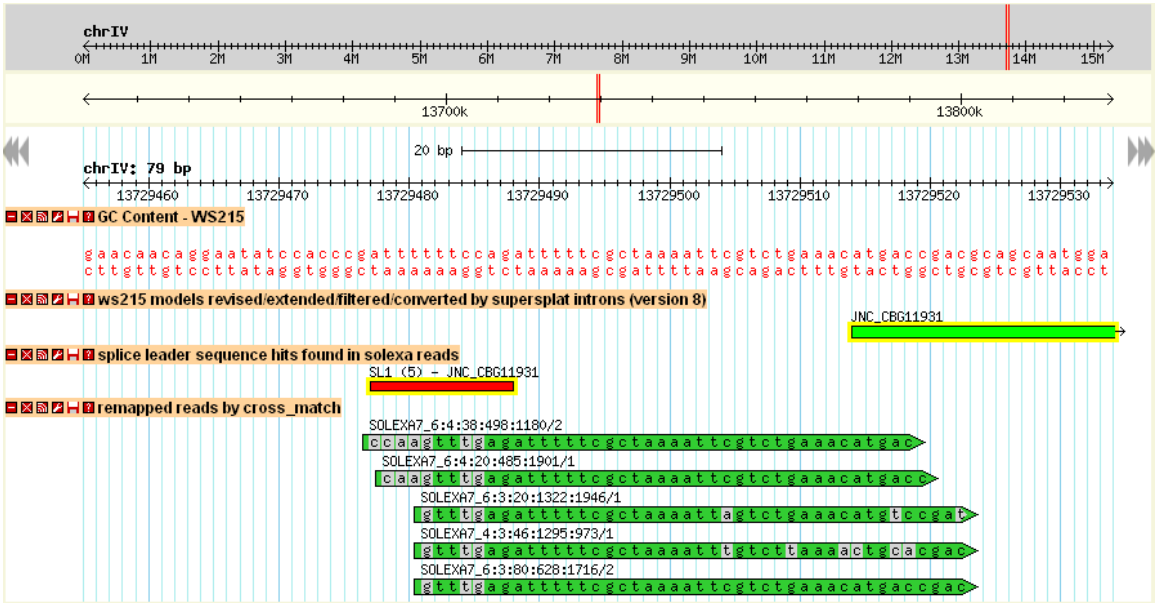

Figure 3-27: Partial SL1 sequence detected upstream of JNC_CBG11931 gene. Varying lengths of SL1 subsequences (7 to 11 nucleotides) are found in 5 Solexa reads that were remapped by cross_match.

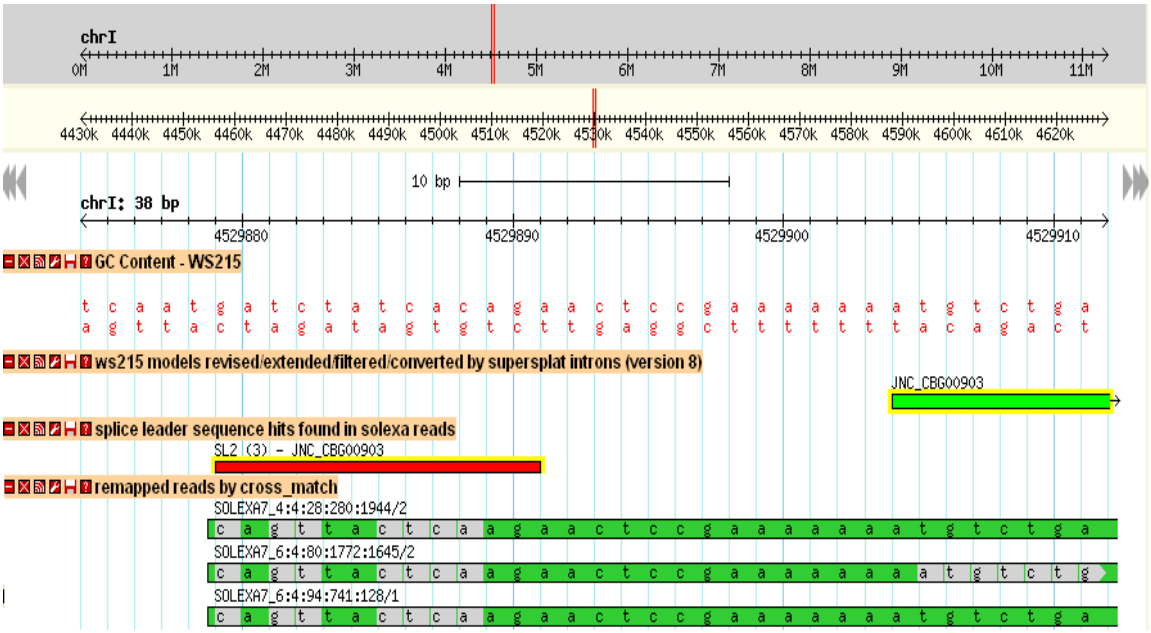

Figure 3-28: Partial SL2 sequence (12 nucleotides) detected in the upstream region of JNC_CBG00903 gene.

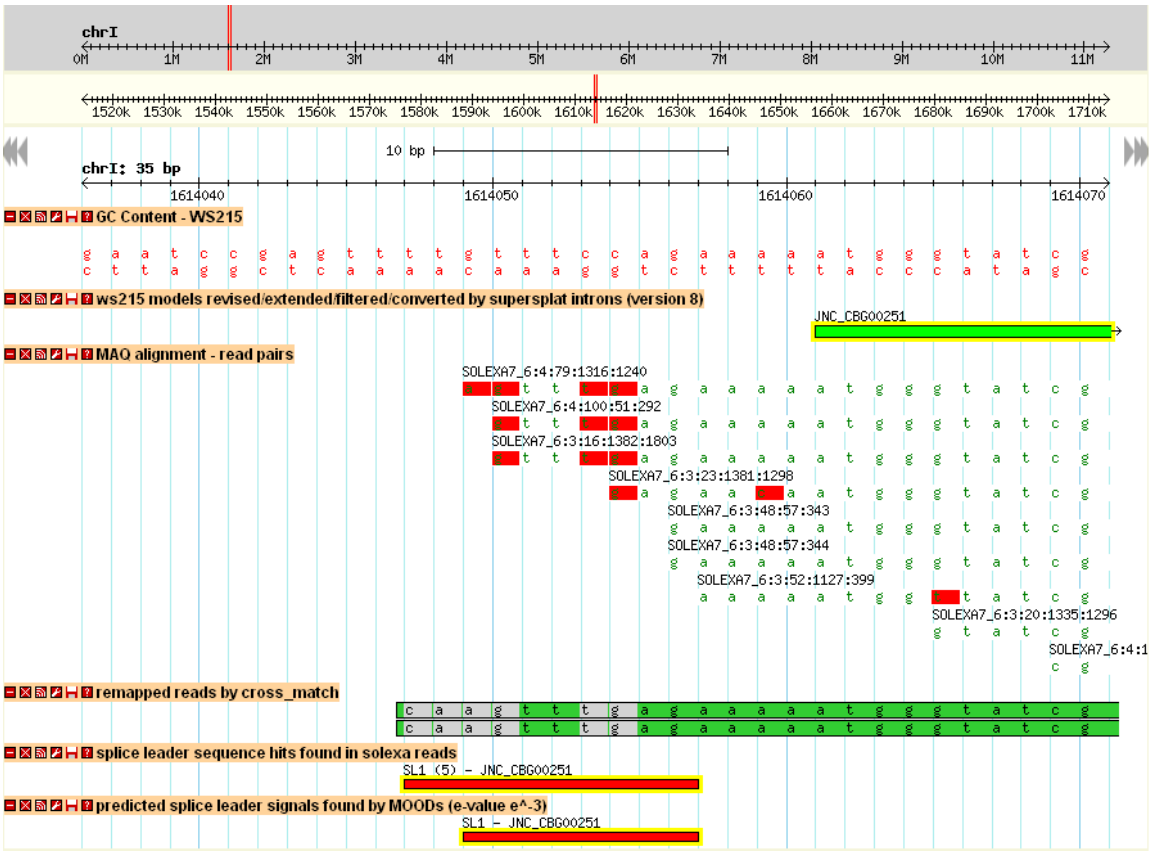

Figure 3-29: Predicted SL1 trans-splicing site. The SL1 site predicted from the genomic sequence, is also confirmed by SL1 site which is detected from the Solexa read sequences.

3.4.5 Discussion

For detection and prediction, the genomic region 100 bp upstream of each gene model is checked. This number is selected based on a manually selected set of trans-splicing sites where we've observed that ~90% of the sites were within 100 bp region upstream of the gene models. However, a more comprehensive statistical analysis needs to be done to clarify the length of this region.

Prediction is only done for SL1. Randomly selected SL1 sites supported by 5 or more reads. SL1 sites that overlap with other SL2-like sites (SL2 to SL12) were discarded in order to remove potential noises. Based on the e-value criteria used (0.001), around half of the genes in *C. briggsae* were predicted to be trans-spliced, which is also similar in *C. elegans* (Blumenthal, Trans-splicing and operons 2005). However, a statistical analysis needs to be done in order to compute sensitivity/specificity values for the predicted splicing sites.

We've observed that most of the SL2 type trans-splicing sites overlap with each other. At the exact same site, different reads have common subsequences with SL2 type splice leader sequences. Furthermore, many genes were found to contain multiple SL1/SL2 type trans-splicing sites suggesting that these genes may be subject to alternative trans-splicing.

Reads containing full SL sequences are not frequent. In our analysis, 5' ends of the reads that were found to contain partial SL sequences contained at least 5 nucleotide matches to the 3' end of corresponding SL sequences. This short segment matches sometimes cause ambiguities in assigning reads to different SL types because of the common subsequences between different SL types. For instance, SL1 and SL7 share the last 5 nucleotides. Additionally, SL5 and SL6 share the last 15 nucleotides. Those hits may be ambiguous. Ambiguity between SL5 and SL6 sites wouldn't cause a problem because they are different variations of SL2 type trans-splicing sites. However, the ambiguity between SL1 and SL7 should be resolved.

The analysis relies on the known SL sequences which are found in *C. elegans.* Therefore, this approach isn't open to discovering novel SL sequences in *C. briggsae.*

4: COMPARATIVE ANALYSIS OF REVISED *C. BRIGGSAE* GENE SET

4.1 Synteny analysis between *C. elegans* and revised *C. briggsae*

In Chapter 2.5, I provided a synteny analysis to evaluate the quality of the revised annotations in terms of synteny improvement between Wormbase predicted gene models and our hybrid set of gene models. The hybrid set of gene models consist of a merged set of genBlastG and Wormbase computationally predicted gene models.

In this chapter, I will describe a similar analysis between our hybrid set of gene models and the final set of gene models which incorporates improvements based on Solexa reads.

For both sets of gene models, I detected the orthologous genes using InParanoid (v4.1) and detected synteny blocks between *C. briggsae* and *C.elegans* using OrthoCluster. I compared the level of synteny found for both gene model sets in terms of number of orthologs in *C. elegans,* percentage of *C. elegans* genome covered by perfect synteny blocks, largest and average synteny blocks sizes in “number of genes” and “number of base pairs” spanned by blocks.

In Table 7, we can observe that RNA-Seq based gene model improvement has impacted the orthology relationships between gene models of *C. briggsae* and *C. elegans.* This procedure has decreased the number of *C.*

elegans orthologs from 15,108 to 15,013. This change in the orthologous relationships has also impacted the synteny blocks. Genome-wide percent coverage of synteny blocks in *C. elegans* has a minor decrease from 45.97% to 45.58%. This minor decrease can also be observed based on the average size of perfect synteny blocks. Number of genes in an average synteny block has decreased from 3.67 to 3.66 genes and the average size in bp decreased from 15,892 bp to 15,842 bp. However, the largest perfect synteny block hasn't been affected. It's the same perfect synteny block that spans a 152,869 bp region (V:10107907-10199687) consisting of 25 genes in *C. elegans.*

Table 7: Comparison of perfect synteny blocks between *C. briggsae* and *C. elegans.*

Perfect	*C.elegans* Orthologs	Synteny Coverage %	# of Non-Nested Blocks	Largest block (genes)	Average (genes)	Largest block (bp)	Average (bp)
Hybrid	15,108.00	45.97	2,877.00	25.00	3.67	152,869.00	15,892.45
Solexa Improved	15,013.00	45.58	2,859.00	25.00	3.66	152,869.00	15,842.55

The minor decrease of the level of perfect synteny between the hybrid set of gene models and the RNA-seq based gene models is due to the revisions made to the gene models as described in Chapter 3. In some cases the revised gene models may disrupt synteny blocks because an extended gene model may be merged into a neighboring gene (See Figure 4-1). Merging of two neighboring genes can disrupt a perfect synteny (See Figure 4-2).

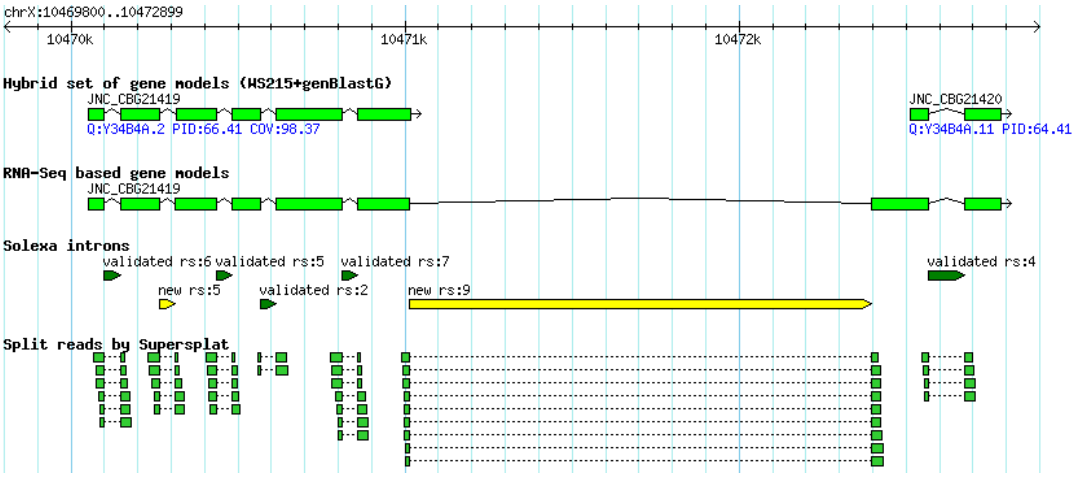

Figure 4-1: Merged models. Two neighboring gene models from the hybrid set, JNC_CBG21419 and JNC_CBG21420, are merged into a single gene model by extension of JNC_CBG21419 at 3' end because of a boundary Solexa intron.

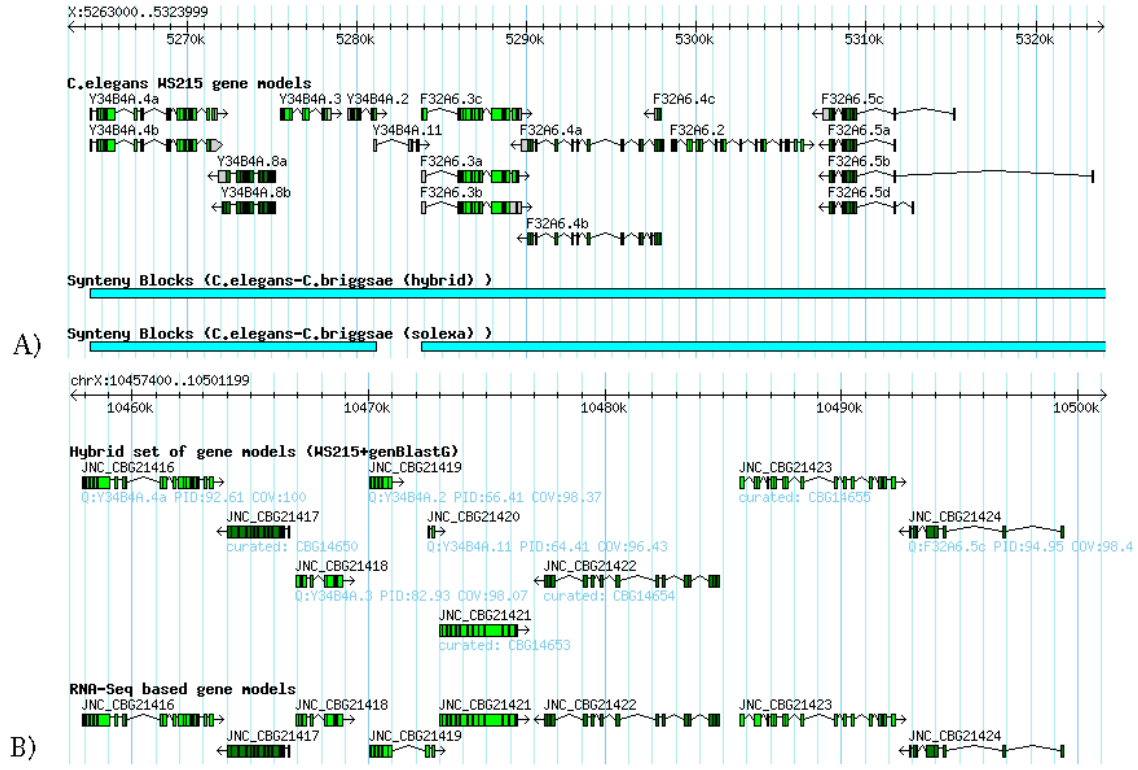

Figure 4-2: Perfect synteny block disruption by merging two neighboring genes. A) *C.elegans* genomic region contains a single perfect synteny block compared to the "hybrid" set and two synteny blocks compared to the RNA-seq based gene set. B) The *C.briggsae* gene models from the hybrid set ,JNC_CBG21419 and JNC_CBG21420, which are orthologous to *C.elegans* genes Y34B4A.2 and Y34B4A.11, are merged and this leads to disruption of the perfect synteny block.

4.2 Operons in *C. briggsae*

Operons in nematodes, are closely spaced gene clusters, which are co-transcribed producing a poly-cistronic mRNA precursor like bacterial operons. These mRNA precursors are further processed by 3'end formation and trans-splicing by SL2 (Blumenthal, Evans, et al. 2002). ~20% of the mRNAs of *C. elegans* are trans-spliced by SL2 and these genes are dominantly found to be downstream genes in *C. elegans* operons ((Spieth, et al. 1993), (Zorio, et al.

1994)). Blumenthal and colleagues have provided a genome-wide analysis of operons in *C. elegans* by a microarray based study, by which the authors identify more than 1000 operons in *C. elegans,* which comprises 15% of the genome.

In Chapter 3.4, RNA-seq based detection of trans-splicing sites in *C. briggsae* is described. The types of SL sequences except SL1 are variations of SL2 that signal trans-splicing of downstream genes in closely spaced gene clusters.

In this chapter, I identify candidate operons in *C. briggsae.* I follow the criteria defined by (Blumenthal, Evans, et al. 2002). Firstly I identify closely spaced gene clusters in *C. briggsae*. Secondly, among these clusters, I detect candidate operons based on the existence of downstream genes that are trans-spliced by SL2. Finally, I compare the genes in candidate *C. briggsae* operons with their orthologs in *C. elegans* in order to observe if the orthologs are also operonic genes in *C. elegans.*

4.2.1 Identification of operons in *C. briggsae*

There are three basic criteria that I have followed in identification of operons in *C. briggsae* (See Figure 4-5).

1) The gene clusters must be closely spaced. The distance between the stop codon of the upstream gene and the start codon of the downstream gene must be smaller than 1 kb.

2) The genes in the clusters must be identically oriented, i.e they must be all on the same strand.

3) All the downstream genes of the closely spaced gene cluster must be SL2 trans-spliced.

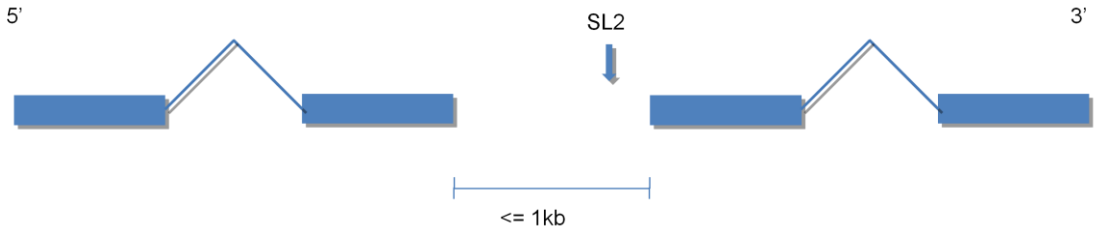

Figure 4-3: Identification of operons in *C. briggsae.*

Following these criteria, 1,034 operons were identified in *C. briggsae,* which are closely spaced gene clusters, in which all genes are identically oriented and SL2 trans-spliced (See Figure 4-6). The operons (391 bp to 58,633 bp long) contain 2 to 9 genes (total 2,408 genes) and have a median size of 2 genes (4,903 bp) (See Figure 4-7).

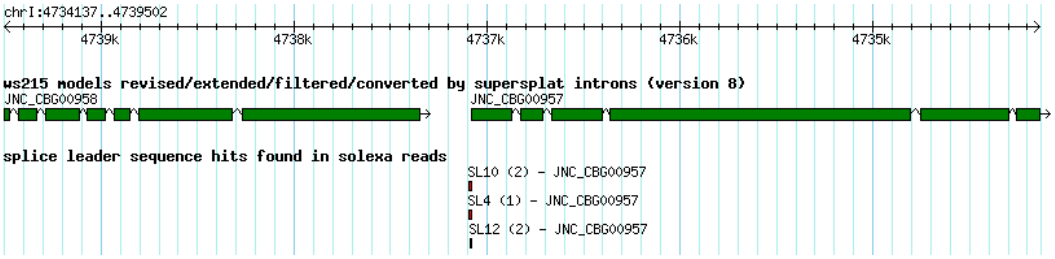

Figure 4-4: An operon identified in *C. briggsae* chromosome I. The cluster consists of 2 genes, JNC_CBG00958 and JNC_CBG00957, which are separated by ~200 bp. The downstream gene JNC_CBG00957 contains SL2 type trans-splice sites upstream of the 5' end of the gene model.

Thus these 1,034 gene clusters, are identified as the first evidence-based annotation of operons in *C. briggsae.*

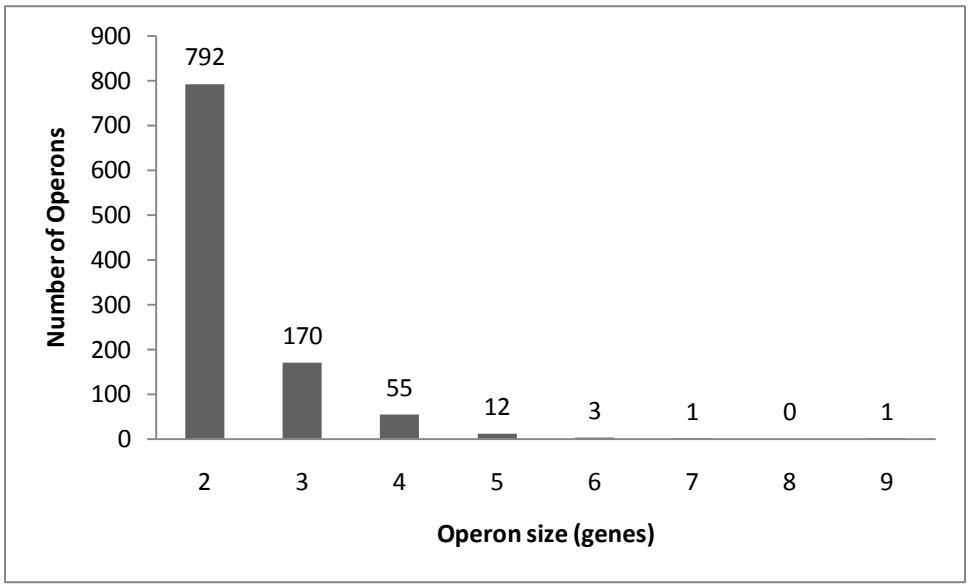

Figure 4-5: Size distribution of detected operons in *C. briggsae*

4.2.2 Comparison of *C. briggsae* operons with their orthologous operons in *C. elegans*

Previous studies ((Stein, et al. 2003), (Qian and Zhang 2008)) have reported that operons are extremely conserved between *C. elegans* and *C. briggsae.* For example, Stein et al (2003) reported that 96% of operons in *C. elegans* are conserved in *C. briggsae*. Therefore, if we've correctly identified operons in *C. briggsae,* we predict that most of the operonic genes in *C. briggsae* are expected to have orthologous operonic genes in *C. elegans.*

Detected operons in *C. briggsae* genome were compared with operons annotated in *C. elegans* genome. Based on this comparison, operons were

classified as 1) “conserved” if all the genes of a *C. briggsae* operon have orthologs in a *C. elegans* operon and vice versa. 2) “species specific” if none of the operonic genes have an orthologous operonic gene 3) “divergent”: any operon that is neither “conserved” nor “species specific”.

Of the 1,034 *C. briggsae* operons detected in this project, only 532 (or 51.45%) were perfectly conserved; 349 (or 33.75%) were divergent; and 153 (or 14.80%) were entirely *C. briggsae* specific operons. My analysis suggests that operons in *Caenorhabditis* species may not be as conserved as previously reported.

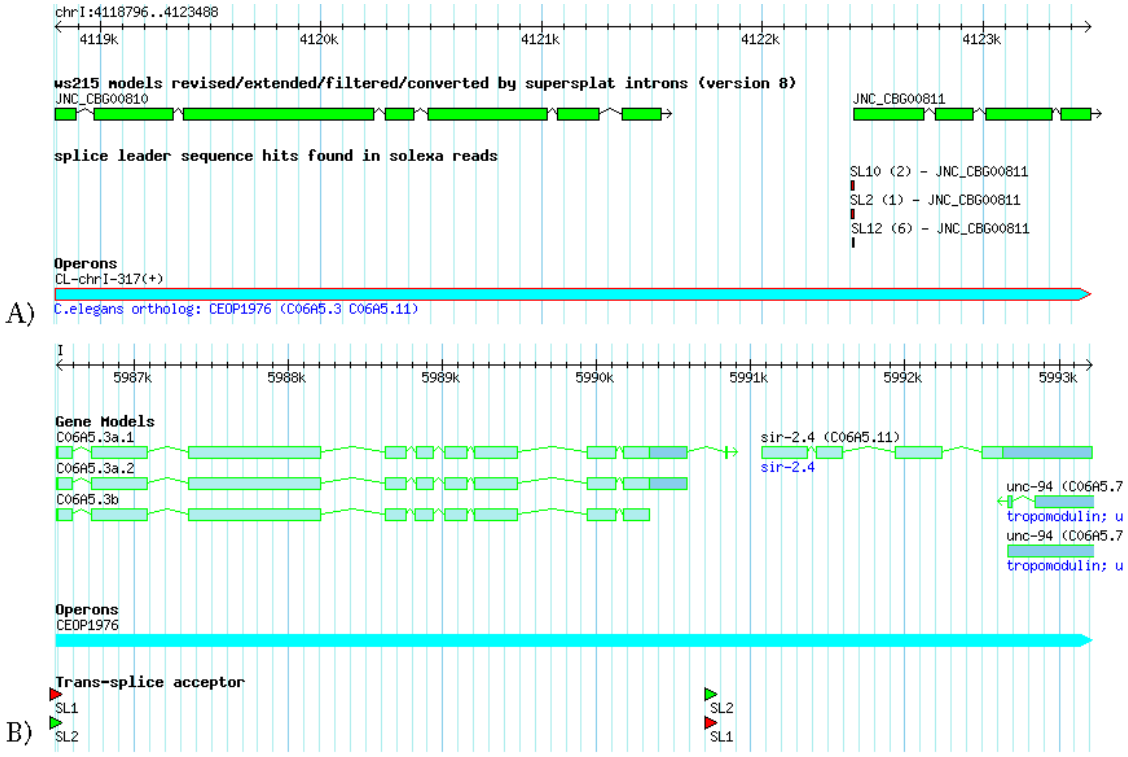

Figure 4-6: Conserved operon. A *C. briggsae* operon with two genes JNC_CBG00810 and JNC_CBG00811 is shown in A) and a *C. elegans* operon with two genes C06A5.3 and C06A5.11, which are orthologs of JNC_CBG00810 and JNC_CBG00811 is shown in B).

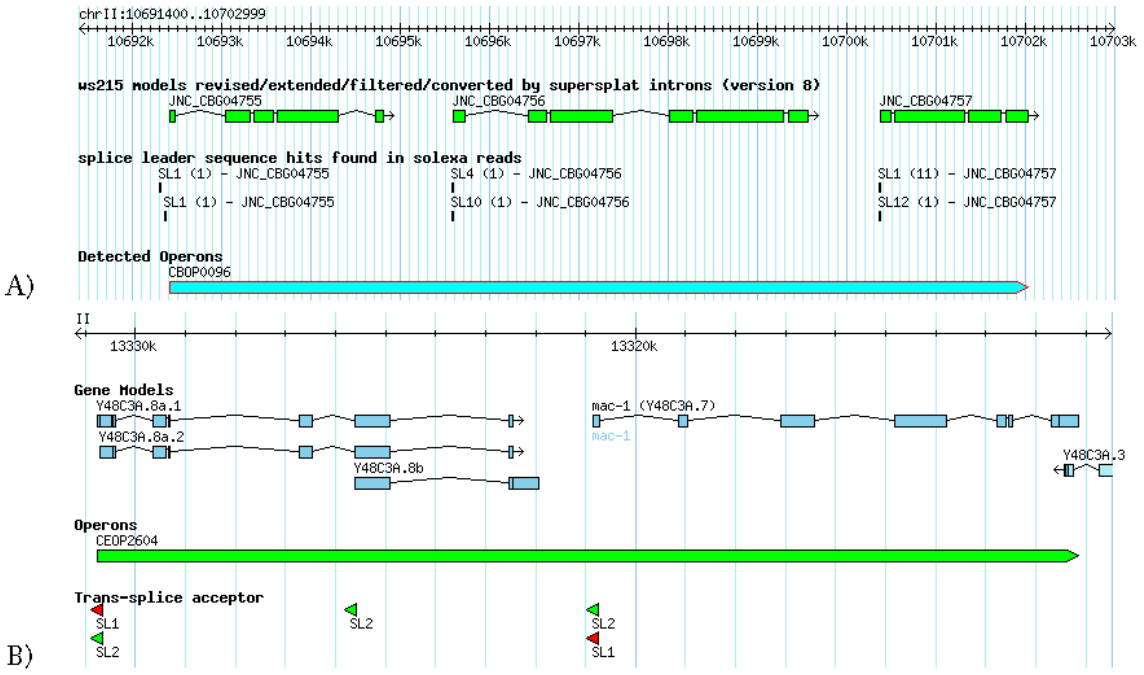

Figure 4-7: Divergent operon. A *C. briggsae* operon with three genes is shown in A) and a *C. elegans* operon with two genes is shown in B). *C. briggsae* operon contains three genes. *C. briggsae* genes JNC_CBG04755 and JNC_CBG04746 are orthologs of *C. elegans* genes Y48C3A.8 and Y48C3A.7, respectively. The third gene in *C. briggsae* operon, JNC_CBG04757, doesn't have any orthologs in *C. elegans.* Therefore, these operons have diverged by either loss of a gene in *C. elegans* or a gain of a gene in *C. briggsae*.

4.2.3 Genome-wide prediction of SL2 trans-splicing sites and operons

Detection of operons depends on the existence of SL2 type *trans*-splicing sites. SL2 *trans*-splicing sites are found based on Solexa read alignments, therefore detection is significantly impacted by the sequencing depth. There may be genes that are not found to be *trans*-spliced just because of lack of read coverage. To overcome this issue, SL2 *trans*-splicing sites were detected as described in Chapter 3.4.2.1. 11,923 SL2 *trans*-splicing sites were predicted for a total of 10,158 genes in the collection of 100 bp upstream regions of all *C.*

briggsae genes. These predicted SL2 *trans*-splicing sites were used to predict operons in *C. briggsae* genome.

In total, 1,722 operons (2 to 6 genes, median = 2) were predicted. These operons contain 3,727 genes. Operons vary in sizes between 521 bp to 58,633 bp with a median length of 4,676 bp.

4.3 Comparison of *trans*-spliced and alternatively *cis*-spliced genes in *C. elegans* and *C. briggsae*

4.3.1 Comparison of *trans*-spliced genes in *C. elegans* and *C. briggsae*

We compared *trans*-spliced genes in *C. briggsae* with their orthologs in *C. elegans.* We wanted to observe if *trans*-splicing patterns are conserved between the orthologs of the two species.

As summarized in Table 7, out of 14,675 *C. briggsae* genes which have 15,013 orthologs in *C. elegans;* 6,378 *C. briggsae* genes which are not *trans*-spliced by either of SL1 or SL2, have *C. elegans* orthologs which aren't *trans*-spliced, either. 61 genes *trans*-spliced by only SL2, 1,978 *trans*-spliced by only SL1, and 906 genes *trans*-spliced by both SL1 and SL2, have orthologs with the same *trans*-splicing patterns in *C. elegans.* Overall, there is a high conservation of trans-splicing between these two nematode species between orthologous pairs.

Table 8: Comparison of orthologs in terms of *trans*-splicing.

C. briggsae		***C. elegans***		
SL1	SL2	SL1	SL2	*C. briggsae* genes
N	N	N	N	6,378
N	Y	N	Y	61
Y	N	Y	N	1,978
Y	Y	Y	Y	906

4.3.2 Comparison of alternatively spliced genes in *C. briggsae* and *C. elegans*

We wanted to observe if a *C. briggsae* gene which has alternative transcription variants, i.e. isoforms, has an alternatively spliced orthologous gene in *C. elegans.*

Isoforms of *C. elegans* genes are annotated in WormBase but isoforms of *C. briggsae* genes haven't been annotated in the current WormBase release. We haven't annotated the *C. briggsae* isoforms in this study, either. However, based on the existence of multiple overlapping Solexa introns, we can deduce which genes may contain isoforms. Thus, we produced a list of *C. briggsae* genes, which may be alternatively spliced. We used this set of genes to compare with annotated *C. elegans* alternatively spliced genes.

As shown in Table 9, it is surprising that only 39.6% of the genes that have isoforms in *C. briggsae* were found to have *C. elegans* orthologs with isoforms. There are two alternative reasons explaining this situation. First, alternative splicing is species-specific. Second, alternative isoforms for many genes have yet to be discovered.

Table 9: Comparison of alternatively spliced orthologous genes of *C. briggase* and *C. elegans.*

	C. briggsae vs C. elegans	***C. elegans vs C. briggsae***
Alternatively Spliced Genes	1,897	2,715
Alternatively spliced genes with orthologs	1,702	2,494
Alternatively spliced genes with alternatively spliced orthologs	752 (39.6%)	745 (27,4 %)

5: DISCUSSION

5.1 Impact of the study

In this study, a higher quality annotation of the genomes four sister species of *C. elegans* has been generated and the first genome-wide set of *C. briggsae* genes supported by evidence are established. My analysis sets up a higher quality platform for comparative genomics analyses.

By applying a hybrid approach of homology-based gene model improvement pipeline, I have replaced thousands of gene models which have better sequence similarity to *C. elegans* orthologs and discovered thousands of novel genes. These changes has increased the number of orthologous relationships between *C. elegans* and its sister species. Fixing defective gene models and introducing novel gene models has reconstructed many synteny blocks, which would otherwise be broken. For all four sister species of *C. elegans,* the percent coverage of the genome by synteny blocks has increased, suggesting the higher quality of the revised gene models.

Furthermore, utilization of next generation sequencing technology and sequencing the transcriptome of *C.briggsae* has allowed us to produce to first genome-wide set of *C. briggsae* genes supported by expression data. Detection of introns from Solexa reads has allowed us to confirm or modify existing introns. Thus, thousands of gene models were internally revised or extended based on RNA-seq data. We were also able to detect trans-spliced genes in *C. briggsae.*

Detection of trans-spliced genes allowed us to annotate the first evidence-based annotation of operons in *C. briggsae* genome.

5.2 Future directions

In this study, although many gene models have been fixed, there are still further improvements that can be done. In our current RNA-Seq based gene model improvement pipeline, we have only internally revised or extended gene models. However, there may be additional evidence about completely novel genes which may have been missed by computational predictions. Annotation of such novel gene models suggested by Solexa reads is a worthwhile experiment to do. Furthermore, in our current approach, a conservative and simplistic approach is taken for the revision of existing introns. Each intron revision is made independently from our revisions. However, in some cases, multiple changes may need to be considered together. For instance when individual intron revisions cause frame-shifts, they wouldn't be allowed. However, revision of multiple introns at the same time may not cause a frame-shift, which would be allowed.

Another important future study would be the annotation of alternative isoforms. In our current pipeline, we produce only one model for each gene. However, the transcriptome data reveals that in many gene models, there are alternative selection of splice sites, which produces alternative transcription variants.

A further useful analysis would be the interpretation of gene evolution by comparison of highly supported *C.briggsae* gene models with their orthologs in *C.elegans* in order to observe evolutionary changes in gene models and their functions.

The experiments suggested above can all be implemented with the available transcriptome sequencing data. However, according to our analysis, the depth of sequencing is limited for some gene models. For instance, we have observed that 3,072 gene models don't have any read support. There are also genes with very low expression levels, which didn't allow us to have enough evidence to modify or confirm some gene models. This suggests that with deeper sequencing of transcriptomes from different cell types, stages, or environmental conditions, we can obtain a larger spectrum of genes with higher expression levels. This would allow us to confirm/modify more gene models, which would produce a much higher quality set of gene models.

It is also important to note that we have produced evidence-based gene models for only *C. briggsae* because we didn't have any experimental data for *C. remanei, C. brenneri,* and *C. japonica.* Should the transcriptomes of these species are sequenced, our pipeline can be applied to improve genome annotations of these species, too.

REFERENCE LIST

Adessi, Celine, et al. “Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms.” *Nucleic Acids Research*, 2000.

Altschul, S. F., et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” *Nucleic Acids Research*, 1997: 3389-3402.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. “Basic local alignment search tool.” *Journal of Molecular Biology*, 1990: 403-410.

Ameur, Adam, Anna Wetterbom, Lars Feuk, and Ulf Gyllensten. “Global and unbiased detection of splice junctions from RNA-seq data.” *Genome Biology*, 2010.

Au, Kin Fai, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. “Detection of splice junctions from paired-end RNA-seq data by SpliceMap.” *Nucleic Acids Research*, 2010.

Baird, S. E., and H. M. Chamberlin. “Caenorhabditis briggsae methods.” *WormBook.* 18 December 2006. http://www.wormbook.org/chapters/www_Cbriggsaemethods/Cbriggsaemethods.html.

Birney, Ewan, Michael Clamp, and Richard Durbin. “GeneWise and Genomewise.” *Genome Research*, 2004: 988-995.

Blumenthal, Thomas. “Trans-splicing and operons.” *Wormbook.* 25 June 2005. www.wormbook.org (accessed 2010).

Blumenthal, Thomas, et al. “A global analysis of Caenorhabditis elegans operons.” *Nature*, 2002: 851-854.

Bonen, Linda. “Trans-splicing of pre-mRNA in plants, animals, and protists.” *FASEB Journal*, 1993: 40-46.

Brent, Michael R. “How does eukaryotic gene prediction work?” *Nature Biotechnology*, 2007: 883-885.

Bryant Jr., Douglas W., Rongkun Shen, Henry D. Priest, Weng-Keen Wong, and Todd C. Mockler. “Supersplat - spliced RNA-seq alignment.” *Bioinformatics*, 2010.

Burge, Chris, and Samuel Karlin. “Prediction of complete gene structures in human genomic DNA.” *Journal of Molecular Biology*, 1997.

Chen, Feng, Aaron J Mackey, Jeroen K Vermunt, and David S Roos. “Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes.” *PLoS ONE*, 2007: e383.

Chen, Nansheng, and Lincoln D. Stein. “Conservation and functional significance of gene topology in the genome of Caenorhabditis elegans.” *Genome Research*, 2006: 606-617.

Chen, Nansheng, et al. “WormBase: a comprehensive data resource for Caenorhabditis biology and genomics.” *Nucleic Acids Research*, 2005: 383-389.

Coghlan, Avril, and Kenneth H. Wolfe. “Fourfold Faster Rate of Genome Rearrangement in Nematodes Than in Drosophila.” *Genome Research*, 2002: 857-867.
Coghlan, Avril, et al. “nGASP – the nematode genome annotation assessment project.” *BMC Bioinformatics*, 2008: 549-561.
Cutter, Asher D. “Divergence Times in Caenorhabditis and Drosophila Inferred from Direct Estimates of the Neutral Mutation Rate.” *Molecular Biology and Evolution, 25 (4)*, 2008: 778-786.
De Bona, Fabio, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. “Optimal spliced alignments of short sequence reads.” *Bioinformatics*, 2008: 174-180.
Do, Jin Hwan, and Dong-Kug Choi. “Computational Approaches to Gene Prediction.” *The Journal of Microbiology*, 2006: 137-144.
Flicek, Paul. “Gene prediction: compare and CONTRAST.” *Genome Biology*, 2007.
Flicek, Paul, Evan Keibler, Ping Hu, Ian Korf, and Michael R. Brent. “Leveraging the Mouse Genome for Gene Prediction in Human: From Whole-Genome Shotgun Reads to a Global Synteny Map.” *Genome Research*, 2003: 46-54.
Florea, Liliana, George Hartzell, Zheng Zhang, Gerald M. Rubin, and Webb Miller. “A computer program for aligning a cDNA sequence with a genomic DNA sequence.” *Genome Research*, 1998: 967-974.
Genome Reference Consortium. *Genome Reference Consortium.* 2010. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/data.shtml (accessed November 11, 2010).
Graber, Graber H., Jesse Salisbury, Lucie N. Hutchins, and Thomas Blumenthal. “*C.elegans* sequences that control trans-splicing and operon pre-mRNA processing.” *Bioinformatics*, 2007: 1409-1426.
Green, Phil. 1993. http://www.phrap.org/phredphrap/general.html (accessed 2010).
Gross, Samuel S., and Michael R. Brent. “Using Multiple Alignments to Improve Gene Prediction.” *Journal of Computational biology*, 2006: 379-393.
Guigo, Roderic, et al. “EGASP: the human ENCODE Genome Annotation Assessment Project.” *Genome Biology*, 2006.
Guttman, Mitchell, et al. “Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.” *Nature Biotechnology*, 2010: 503-514.
Hach, Faraz, et al. “mrsFAST : a cache-oblivious algorithm for short-read mapping.” *Nature Methods*, 2010: 576-577.
Hardison, Ross C. “Comparative Genomics.” *PLoS Biol 1(2)*, 2003: 156-160.
Hillier, LaDeana W., Alan Coulson, and John I. Murray. “Genomics in *C. elegans*: So many genes, such a little worm.” *Genome Research*, 2005: 1651-1660.
Hillier, LaDeana W., et al. “Comparison of C. elegans and C. briggsae Genome Sequences Reveals Extensive Conservation of Chromosome Organization and Synteny.” *PLoS Biology*, 2007: 1603-1616.

Hillier, LaDeana W., Valerie Rienke, Philip Green, Martin Hirst, and Marco A. Marra. “Massively parallel sequencing of the polyadenylated transcriptome of C.elegans.” *Genome Research*, 2009: 657-666.
Huang, X., M. D. Adams, H. Zhou, and A. R. Kerlavage. “A tool for analyzing and annotating genomic sequences.” *Genomics*, 1997: 37-45.
Jiang, Hui, and Wing Hung Wong. “SeqMap: mapping massive amount of oligonucleotides.” *Bioinformatics*, 2008: 2395-2396.
Kent, James W., and Alan M. Zahler. “Conservation, Regulation, Synteny, and Introns in a Large-scale C. briggsae–C. elegans Genomic Alignment.” *Genome Research*, 2000: 1115-1125.
Kiontke, K., and D.H.A. Fitch. “The Phylogenetic relationships of Caenorhabditis and other rhabditids.” *WormBook.* 11 August 2005. http://www.wormbook.org/chapters/www_phylogrhabditids/phylorhab.html#d0e328.
Korhonen, Janne, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. “MOODS: fast search for position weight matrix matches in DNA sequences.” *Bioinformatics*, 2009: 3181-3182.
Li, Heng, and Nils Homer. “A survey of sequence alignment algorithms for next-generation sequencing.” *Briefings in Bioinformatics*, 2010.
Li, Heng, Jue Ruan, and Richard Durbin. “Mapping short DNA sequencing reads and calling variants using mapping quality scores.” *Genome Research*, 2008: 1851-1858.
Li, Li, Christian J. Stoeckert, and David S. Roos. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” *Genome Research 13*, 2003: 2178-2189.
Li, Ruiqiang, Yingrui Li, Karsten Kristiansen, and Jun Wang. “SOAP: short oligonucleotide alignment program.” *Bioinformatics*, 2008: 713-714.
Lin, Hao, Zefeng Zhang, Michael Q. Zhang, Bin Ma, and Ming Li. “ZOOM! Zillions of oligos mapped.” *Bioinformatics*, 2008: 2431-2437.
Mathe, Catherine, Marie-France Sagot, Thomas Schiex, and Pierre Rouze. “Current methods of gene prediction, their strengths and weaknesses.” *Nucleic Acids Research*, 2002: 4103-4117.
Metzker, Michael L. “Sequencing technologies - the next generation.” *Nature Reviews - Genetics*, 2010: 31-45.
Morozova, Olena, and Marco A. Marra. “Applications of next-generation sequencing technologies in functional genomics.” *Genomics*, 2008: 255-264.
Qian, Wenfeng, and Jianzhi Zhang. “Evolutionary dynamics of nematode operons: Easy come, slow go.” *Genome Research*, 2008.
Remm, Maido, Christian E.V. Storm, and Erik L.L. Sonnhammer. “Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.” *Journal of Molecular Biology*, 2001: 1041-1052.
Ronaghi, Mostafa, Mathias Uhlén, and Pål Nyrén. “DNA SEQUENCING: A Sequencing Method Based on Real-Time Pyrophosphate.” *Science*, 1998: 363-365.

Rumble, Stephen M., Phil Lacroute, Adrian V. Dalca, Mark Fiume, Arend Sidow, and Michael Brudno. “SHRiMP: Accurate Mapping of Short Color-space Reads.” *PLoS Computational Biology*, 2009.
She, Rong, Jeffrey S.C. Chu, Ke Wang, Jian Pei, and Nansheng Chen. “genBlastA: Enabling BLAST to identify homologous gene sequences.” *Genome Research*, 2009: 143-149.
She, Rong, Jeffrey Shi-Chieh Chu, Bora Uyar, Ke Wang, and Nansheng Chen. “genBlastG: a Fast and Accurate Homology-Gene Prediction Program.” Submitted.
Sleator, Roy D. “An overview of the current status of eukaryote gene prediction strategies.” *Gene*, 2010.
Smith, Andrew D., Zhenyu Xuan, and Michael Q. Zhang. “Using quality scores and longer reads improves accuracy of Solexa read mapping.” *BMC Bioinformatics*, 2008.
Smith, Temple F., and Michael S. Waterman. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology*, 1981: 195-197.
Spieth, John, Glen Brooke, Scott Kuersten, Kristi Lea, and Thomas Blumenthal. “Operons in C. elegans: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions.” *Cell 73 (3)*, 1993: 521-532.
Stein, Lincoln D., et al. “The Genome Sequence of Caenorhabditis briggase: A Platform for Comparative Genomics.” *PLoS Biology 1 (2)*, 2003: 166-192.
The C. elegans Sequencing Consortium. “Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology.” *Science 282, (2012)*, 1998: 2012-2017.
Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. “TopHat: discovering splice junctions with RNA-Seq.” *Bioinformatics*, 2009: 1105-1111.
Usuka, Jonathan, Wei Zhu, and Volker Brendel. “Optimal spliced alignment of homologous cDNA to a genomic DNA template.” *Bioinformatics*, 2000: 203-211.
Vergara, Ismael A., and Nansheng Chen. “Large synteny blocks revealed between Caenorhabditis elegans and Caenorhabditis briggsae genomes using OrthoCluster.” *BMC Genomics*, 2010: 516-568.
Wang, Zhong, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics.” *Nature Reviews - Genetics*, 2009: 57-63.
Wu, Thomas D., and Serban Nacu. “Fast and SNP-tolerant detection of complex variants and splicing in short reads.” *Bioinformatics*, 2010: 873-881.
Zeng, Xinghuo, Jian Pei, Ismael A. Vergara, Matthew J. Nesbitt, Ke Wang, and Nansheng Chen. “OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics.” *EDBT.* Nantes, France, 2008.
Zorio, Diego A. R., Niansheng Nick Cheng, Thomas Blumenthal, and John Spieth. “Operons as a common form of chromosomal organization in C. elegans.” *Nature*, 1994: 270-272.