RNA-Seq Assembly Discovers Many Splice Variants

Likit Preeyanon[1], Jerry B. Dodgson[1] Hans Cheng[2], C. Titus Brown[3,1*]

1 Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA.

2 Avian Disease and Oncology Laboratory, East Lansing, MI, USA.

3 Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA.

∗ **E-mail: ctb@msu.edu**

Abstract

Comparison of spliced reads and reference gene annotations has been successfully used to discover alternative splicing in model organisms. However, the method cannot be applied in organisms without high-quality reference genome and gene annotations. In this article, we introduce a pipeline, based on *de novo* assembly, that constructs gene models with splice variants. The pipeline includes a new technique called local assembly that enhances the sensitivity of alternative splicing detection. We demonstrates that the pipeline detects many novel splice variants in RNA-Seq data from chicken spleens. Some of those splice variants; however, only detected by our pipeline and not Cufflinks. This indicates that the pipeline can be used to facilitate splice variants detection from RNA-Seq.

Author Summary

Introduction

Until recently, studies of alternative splicing had been limited to a small number of genes and isoforms due to high-cost and low-throughput sequencing of expressed sequence tags (ESTs) and full-length cDNA libraries. RNA sequencing (RNA-Seq) using the next-generation sequencing (NGS) has been used successfully in many studies to gain unprecedented insight into a complexity of transcriptomes. Novel isoforms have been identified based on reads mapped across known splice junctions or putative splice junctions from annotations. It has been estimated that, in human, $92-94\%$ of multiexon genes undergo alternative

splicing and different isoforms are expressed in different tissues [1]. This suggests that even in human a large number of splice variants has not been explored.

Sequences from RNA-Seq are very short (75–250bp for Illumina reads); therefore, it is not feasible to assess gene and isoform expression without mapping reads to a reference annotation or construct a full-length transcript from *de novo* assembly. Several softwares employing different approaches have been developed to reconstruct transcripts from short reads. Tophat [2] maps reads to a reference genome to identify exons and splice sites *ab initio*. It first searches for known splie sites in a genome and then tries to map reads across splice juntions to identify exon junctions. Exon junctions can be used to build gene models by Cufflinks [3]. On the other hand, Trinity [4], has been developed to assemble short reads *de novo* to construct transcripts. Trans-Abyss [5] and Oases [6] are an extension of Abyss [7] and Velvet [8,9] assembler that are designed to work with RNA-Seq data. These programs are comparable at identifying alternative splicing with a slightly different sensitivity and specificity. However, Oases with Oases-M has been shown to be superior than other *de novo* assembler at discovering isoforms in human and mouse [6].

In this study, we present a pipeline that uses Velvet and Oases assembler to discover alternative splicing in RNA-Seq data from chickens. We use a new technique called a local assembly that enhances sensitivity of alternative splicing detection of Oases. The results show that the pipeline can detect more isoforms than Oases-M and can detect isoforms not found by Cufflinks.

Results

Transcripts from Global and Local Assembly

Without high-quality annotations, it is necessary to construct gene models with putative isoforms to identify alternative splicing in our samples. We used Velvet [8] and Oases [6] assembler to construct transcripts from short reads. We ran Velvet using default settings except a hash length or k-mer length, which has no default values. It has been shown that transcripts from a combination of multiple k-mer lengths is superior than a single k-mer length [6]. In addition, we found that different k-mer lengths detected different isoforms from the same dataset (Fig. 4). Therefore, it is necessary to cover a wide range of k-mers to discover as many isoforms as possible. For this study, we selected a range of kmers from $21 - 31$ since we found that k-mers outside this range are prone to introduce assembly errors.

Local Assembly Enhances Isoform Detection

Performing assembly with multiple k-mers is time and memory consuming. To facilitate this, we partitioned reads into groups based on their alignments to the chicken genome. We refer to this method as a local assembly and a conventional assembly method as a global assembly in this study (see materials and methods). The advantages of the local assembly method is that each group of reads can be assembled using small amount of memory and it can be assembled on different computers. Unexpectedly, we also found that local assembly detected alternative isoforms not found in global assembly. Therefore, we merged transcripts from both global and local assembly from multiple k-mers to obtain all possible alternative isoforms from our samples. Figure 5 shows an example of different isoforms detected from two assembly methods.

Using global and local assembly methods, we obtained a different number of contigs. Table 1 shows a number of contigs from four datasets and unique contigs from global and local assemblies. As expected, some sequences from local assembly (1,199–2,270) are not found in global assembly. Homology search in mouse proteins showed that 35.02–75.61% of unique regions from local assembly match protein coding sequences (Table 2). This suggests that some isoforms will be missing in the gene models based solely on global assembly. In addition, a significant number of transcripts from global assembly are unique, which indicates that approximately at least 3% of coding sequences are missing in a current chicken genome.

Transcripts from Oases-M contain fewer splice junctions

Transcripts from multiple k-mer lengths can be merged using Oases-M. It has been demonstrated that merged transcripts are more complete than those from any single kmer length [6]. However, sensitivity of Oases-M in detecting alternative isoforms has not been fully evaluated. To evaluate this, we compared transcripts from global assembly with k-mer lenght from 21–31 and merged transcripts from Oases-M using k-mer=27. We used alignments of ESTs as a control. The results show that Oases-M detect fewer splice junctions than unmerged global assembly. A number of total splice junctions and unique splice junctions detected by Oases-M and unmerged assembly is shown in Table 3. Only 391 unique splice junctions from Oases-M are supported by ESTs whereas unmerged assembly detects 1,752 unique splice junctions that also found in ESTs. This suggests that some genuine splice junctions are missing from merging transcripts from multiple k-mer lengths with Oases-M. Therefore, we did not use Oases-M in

our pipeline.

Gene models prediction using Gimme

We developed a program called, Gimme, to constructs a gene model from alignments of transcripts to the genome. The program can merge all transcripts from multiple-kmers and predict a structure of full-length isoforms and genes. As shown in figure 8, our gene models are comparable to other existing gene models. Total of 14,832 genes were obtained from global assembly and 15,297 genes from local assembly. However, the number of genes obtained from global + local assembly is only 15,934. The number of genes and isoforms slightly increased after combining transcripts from global and local assembly together (Table 4). This indicates that some transcripts from both methods were merged together.

Most transcripts include a large part of untranslated regions (UTRs), especially $3'$ UTR in our datasets. These regions are challenging to predict correctly solely from computational methods due to low degree of conservation [?]. RNA-Seq-derived gene models are useful for studying variations within UTR regions, which may be involved in regulation of isoform expression [] (Fig. 9).

An overestimated number of genes and transcripts

The number of genes from our pipeline might be overestimated due to a number of fragmented transcripts. In our datasets, read coverage is lower at the $5'$ end due to the method used to convert mRNA to cDNA in library preparation. A bias of read coverage toward $3'$ end and low expression level result in fragmented transcripts as shown in Fig. 7. In addition, overestimation of a number of isoforms is due to spurious splice junctions, which can result in more than 10,000 possible isoforms in one gene. This seems to occur in a few genome sequences such as chromosome E64_random. Figure 6 shows that a majority of genes from our gene models have only a few isoforms.

Validation of Gene Models

The results show that our pipeline can be used to detect alternative splicing and construct gene models from a large set of transcripts from assembly. However, many steps in our pipeline can introduce errors to the gene models. We performed several methods to validate the gene models as described below.

Reads Mapped to Gene Models

Single-end reads from the same datasets that we used to build the gene models were aligned to cDNA sequences derived from the gene models. We anticipated that a large number of reads could align to the gene models. We found that 63-78% of reads mapped to the gene models (Table 5). In addition, we mapped paired-end reads that we obtained later from the same samples to the gene models and found that 63-65% of reads mapped to the gene models (Table 6). The results suggest that the gene models are of high quality.

Reads mapped to splice junctions

To validate the splice junctions, we mapped reads to cDNA sequences derived from gene models using Bowtie [10]. Because the mapping algorithm implemented in Bowtie does not consider splice sites, reads mapped across splice junctions confirm that the splice junctions are in transcripts. Of 112,089 splice junctions from the gene models, only 1,341 (1.2%) junctions have no spliced reads and 3,447 (3%) junctions have 1–3 spliced reads (see materials and methods) (Fig. 11. The results indicate the method can detect splice junctions with a very high specificity. Some splice junctions with fewer than four spliced reads; however, may not be false junctions because of a bias of read coverage toward $3'$ end in our datasets as mentioned earlier. Reads are less abundant at the $5'$ end of transcripts; therefore, a relatively small number of spliced reads is mapped to splice junctions near the $5'$ end.

Comparison with ESTs and Genbank mRNAs

83,238 (79.7%) splice junctions from our gene models are supported by ESTs or Genbank mRNAs. This illustrates the efficiency of our pipeline in detecting genuine splice junctions. We also found that 46,132 junctions that are supported by both Genbank mRNAs and ESTs are also supported by at least four spliced reads. Therefore, we can assume that splice junctions with more than three spliced reads are likely to be real. Using this criterion, we can conclude that our gene models include 21,872 novel splice junctions that are not in ESTs.

Comparison with Cufflinks

To evaluate an efficiency of our pipeline in detecting splice junctions, we compared the number of splice junctions detected by our pipeline to those from Cufflinks [3], ESTs and Ensembl gene models version 64.

Cufflinks detected 93,756 (83%) splice junctions that are also detected by our pipeline. However, 46% (8,615) of junctions not detected by Cufflinks are supported by ESTs. This suggests that the assembly detects some splice junctions that are not detected by Cufflinks and they are genuine splice junctions.

We also aligned transcripts from Cufflinks and the pipeline to Ensembl transcripts to compare genes and transcripts detected by both methods. We selected only transcripts that match with more than 90% identity and cover more than 80% of an Ensembl transcript. Our pipeline detects 7,765 genes and Cufflinks detects 8,983 genes of total 17,934 Ensembl genes. 7,049 of those genes are both detectd by our pipeline and Cufflinks. In addition, we wanted to compare a number of isoforms detected by our pipeline and Cufflinks within the same gene; however, a transcript can match to multiple isoforms in the same gene, especially when it is not complete. Therefore, we compare a number of splice junctions detected by both methods within the same gene instead. From 7,049 genes, our pipeline detects 1,765 splice junctions, whereas Cufflinks detects 2,516 junctions.

However, because Ensembl gene models do not include all isoforms or splice junctions from ESTs, we investigated splice junctions that are not in Ensembl but are supported by ESTs in those 7,049 genes and found that our pipeline detects 2,039 splice junctions and Cufflinks only detects 999 splice junctions. Figure 12 shows an example of splice junctions not included in Ensembl gene models, which are detected by our pipeline.

To conclude, Cufflinks performs better than our pipeline for detecting both genes and splice junctions in Ensembl gene models. However, we found that our pipeline detects more splice junctions that are not in Ensembl but are supported by ESTs.

Homologous sequences in mouse

Splice variants must maintain the reading frame to be translated to functional proteins. Based on this assumption, we investigated isoforms in our gene models by searching for homologous protein sequences in other organism. To obtain a protein sequence of each isoform, we translated a cDNA sequence using ESTScan [11], which is designed to translate protein sequences from incomplete cDNA sequences such as ESTs. ESTScan succesfully translated 44,122 (94%) of all isoforms to protein sequences with 50 or longer amino acids. Then, all protein sequences were searched for homology in mouse *(Mus musculus)*, which is a related and well-studied animal. We found 27,770 (63%) isoforms from 4,272 genes match mouse proteins. As shown in Fig. 10, most genes and isoforms have a bit score/length ratio more than

1.0, which indicates a good agreement between chicken and mouse proteins. The results illustrate that isoforms constructed by our pipeline are translatable and may be functional. However, genes or isoforms with no match to mouse proteins are not necessarily artifacts. They can be genes or isoforms that are only found in non-mammal vertebrates or novel genes.

Discussion

Choice of k-mers greatly affects sensitivity of splice variants detection in RNA-Seq assembly using Velvet/Oases. Short k-mers increase sensitivity, but also introduce errors from misassembly. However, for a study focusing on alternative splicing, it is desirable to detect as many splice variant as possible. We have presented the local assembly technique that enhances the ability to detect splice variants of Velvet/Oases from the same k-mer. The mechanism behind this technique is to be investigated. Importantly, understanding of the mechanisms may lead to a filtering technique that can be applied to organisms that lack a reference genome.

We have also developed a pipeline that assembles reconstructed transcripts from multiple k-mers to build putative gene models. Overall, a majority of splice variants detected by our pipeline are also detected by Cufflinks. Both our pipeline and Cufflinks also detect some splice variants that are not overlapped. This suggests that combination of both *de novo* assembly and reference-based assembly is preferred to study alternative splicing in chicken.

Interestingly, Cufflinks detects more genes and isoforms included in Ensembl annotations than our pipeline. However, our pipeline detects many more splice junctions not included in Ensembl but supported by ESTs. As observed in previous study by Schulz *et al.* [6], Cufflinks also outperforms *de novo* assembly in detecting genes and isoforms in Ensembl annotations in mouse. We speculate that Cufflinks possesses a mechanism that is tailored to detect Ensembl gene models.

In this study, gene models are built from chicken genome version 2.1 (galGal3), which contains a large number of sequence duplications and missassemblies that supposed to be eliminated in the latest version of genome assembly (galGal4). Duplications and misassemblies lead to false splice junctions, which in turn produce a large number of splice variants as observed in some chromosomes. Conversely, this suggests that one might be able to use the pipeline to evaluate the quality of reference genome.

Materials and Methods

Quality trimming of reads

Both single- and paired-end reads in this study were trimmed using Condetri version 2.1 with default parameters. In addition, the first 10 bases of each reads were trimmed off due to an inconsistency of base-calling as shown in supplementary figure S?.

Mapping reads to The Reference Genome and Gene Models

Single and paired-end reads were mapped to chicken genome by Tophat2 [2] release 2.0.0 using default parameters without annotations. All reads were mapped to cDNA sequences derived from gene models by Bowtie2 [10] with default parameters `(n=2, l=28, e=70, k=1)`.

Global and Local Assembly

Reads from each dataset were first assembled separately in global assembly without using a reference genome. In contrast, reads from each dataset were first mapped to the chicken genome using Tophat2. Then only reads mapped to the genome were assembled by chromosomes in local assembly (Fig. 2). Global and local assembly was performed using Velvet version 1.2.03 [8] with default parameters except for hash length (k-mer). A range of k-mer length from 21-31 was used to assemble reads in both global and local assembly. Lastly, transcripts from both methods were assembled by Oases version 0.2.06 [6].

Transcript Reassembly

To reduce resources and time to reassemble transcripts, we first trimmed off poly-A tail and remove low complexity and short sequences using seqclean [12] with default parameters. Then we used cd-hit-est from CD-HIT suite [13] to remove all redundant transcripts. A large number of transcripts was removed at this step, which facilitates the reassembly process. In order to eliminate some sequence variations between samples. We mapped sequences to the chicken genome and then extracted all sequences from the genome. All transcripts from each chromosome were then assembled by CAP3 [14].

Gene Model Construction

Overall Pipeline

Figure 1 depicts an overall gene model construction pipeline. Transcripts of all datasets from local and global assembly were mapped to the chicken genome using BLAT [15] `(-t=dna -q=dna -noHead -out=psl -mask=lower -extendThroughN -dots=1000)`. Alignments and gaps from BLAT outputs are considered exons and introns respectively. Optionally, data from other sources (ESTs, RefGenes, etc.) can be incorporated with transcripts from assembly to improve gene models. All transcripts are then assembled using Gimme, a program that assembles transcripts based on their alignments to the reference genome. An algorithm for assembling transcripts is described below. A maximum set of transcripts obtained from Gimme are then reduced to only a minimum set of transcripts that contain all splice junctions and untranslated regions (UTRs). After that, transcripts that are identical are clustered and removed by CD-HIT version 4.5.6 [13]. Parameters for clustering are `-n 10, -r 0 and -c 1.0`. Only a representative of each cluster is kept in gene models.

Algorithm

A gene model can be represented as a splice graph composed of exons as nodes and introns as edges. However, transcripts of the same gene vary in size and structure depending on the expression level and a hash length number used in assembly. Furthermore, incomplete exons and fragmented transcripts complicate the construction of a splice graph. In this study, we developed an algorithm that handles incomplete exons and fragmented transcripts and constructs a maximum assembly of gene models.

The algorithm first builds an intron graph using introns as nodes. Each intron contains exons whose one of their splice sites perfectly match intron boundary. Exons are considered incomplete and eliminated if they locate at the $3'$ or $5'$ end of the transcripts and they are not the largest exons (exon 3a and 3b in Fig. 3). Transcripts were then grouped into the same gene if they have at least one intron or exon in common. Then, a splice graph composed of exons is created and structures of isoforms are derived from traversing paths in the splice graph. Gimme is open-source and available at `https://github.com/ged-lab/gimme`.

Protein sequence translation

We employed ESTScan version 2.1 to translate protein sequences from our gene models. The matrix used for building Hidden Markov model was built from chicken reference cDNA sequences using tools from ESTScan.

Finding unique sequences between datasets

To identify unique sequences from two datasets, a set of 20-mers is created for both datasets using khmer [16]. Then, 20-mers from a query dataset are compared with 20-mers from the target dataset. The sequence is considered unique if more than 90% of 20-mers in the query is unique. Any unique region shorter than 100 bp is ignored.

Sequence homology analysis

Protein sequences translated from each isoform using ESTScan were searched against mouse reference proteins by BLAST 2.2.25+ [17]. A bit score to a length ratio was calculated for each hit that had an e-value $\leq 10^{-20}$. Only the highest value of all isoforms from each gene was shown in the gene plot; whereas, values of all isoforms were shown in the isoform plot.

Spliced reads count

Reads from each dataset were mapped to transcripts from the gene models using Bowtie version 2.0-beta 5 with default parameters. Reads mapped across exon junctions from all datasets were counted using Samtools [18] and Pysam [19]

Sequence assembly using Cufflinks

Reads are mapped to a genome sequence using Tophat2. Gene models are built from each datasets by Cufflinks 2.0.0 [3]. All gene models are then merged together using Cuffmerge.

Expressed sequence tags and Genbank mRNA

Expressed sequence tags (ESTs) and mRNAs were downloaded from UCSC genome website on May 9th, 2012. The database last updated from GENBANK on April 29th, 2012. Sequences were alignd to a

chicken genome using BLAT.

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.

2. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England) 25: 1105–1111.

3. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28: 511–515.

4. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 29: 644–652.

5. Robertson G, Schein J, Chiu R, Corbett R, Field M, et al. (2010) De novo assembly and analysis of RNA-seq data. Nature methods 7: 909–912.

6. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics (Oxford, England) .

7. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome research 19: 1117–1123.

8. Zerbino D (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome research .

9. Zerbino DR, McEwen GK, Margulies EH, Birney E (2009) Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. PloS one 4: e8407.

10. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.

11. Iseli C, Jongeneel C, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol .

12. SeqClean. URL `http://compbio.dfci.harvard.edu/tgi/software/`.

13. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England) 22: 1658–1659.

14. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome research 9: 868–877.

15. Kent W (2002) BLAT—The BLAST-Like Alignment Tool. Genome research .

16. Khmer: In memory k-mer counting. URL `https://github.com/ged-lab/khmer`.

17. Tatusova T (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences - Tatusova - 2006 - FEMS Microbiology Letters - Wiley Online Library. FEMS microbiology letters .

18. Li H, Handsaker B, Wysoker A, Fennell T (2009) The Sequence Alignment/Map format and SAMtools. ... .

19. Pysam: Python interface for the SAM/BAM sequence alignment and mapping format. URL `http://code.google.com/p/pysam/`.

Tables

Figure Legends

Table 1. Unique sequences from global and local assembly

Dataset	Total Sequence		Unique Sequence	
	Global	Local	Global	Local
Line 6 uninfected	338,353	335,180	10,068 (2.97%)	1,258 (0.37%)
Line 6 infected	353,485	319,920	11,639 (3.29%)	2,270 (0.70%)
Line 7 uninfected	335,876	302,443	10,649 (3.17%)	1,570 (0.51%)
Line 7 infected	371,404	327,491	11,859 (3.19?%)	1,199 (0.36%)

Table 2. Unique regions from global and local assembly

Dataset	Unique Region		Matched with mouse proteins	
	Global	Local	Global	Local
Line 6 uninfected	2,978	117	1,956 (52.76%)	41 (35.04%)
Line 6 infected	3,524	236	1,991 (56.50%)	100 (42.37%)
Line 7 uninfected	3,334	163	1,933 (57.98%)	86 (52.76%)
Line 7 infected	3,235	82	1,858 (57.43%)	62 (75.61%)

Table 3. Number of total and unique splice junctions

Method	Total	Unique	Supported by ESTs
Oases-M assembly	114,336	8,510	391 (4.59%)
Unmerged assembly	116,573	10,747	1,752 (16.30%)

Table 4. Number of putative genes and isoforms

Method	Gene	Transcript
Global	14,832	32,311
Local	15,297	23,028
Global + Local	15,934	46,797
Cufflinks	30,235	37,967
Ensembl	17,934	23,392

Table 5. Single-end reads mapped to gene models (maximum)

Dataset	Mapped	Unmapped
Line 6 uninfected	18,375,966 (77.93%)	5,203,586 (22.07%)
Line 6 infected	17,160,695 (73.18%)	6,288,286 (26.82%)
Line 7 uninfected	18,130,072 (75.77%)	5,795,737 (24.22%)
Line 7 infected	19,912,046 (78.21%)	5,450,521 (21.49%)

Table 6. Paired-end reads mapped to gene models (maximum)

Dataset	Mapped	Unmapped
Line 6 uninfected	21,598,218 (64.16%)	12,065,659 (35.84%)
Line 6 infected	15,274,638 (63.89%)	15,274,638 (36.11%)
Line 7 uninfected	20,961,033 (63.67%)	20,961,033 (36.33%)
Line 7 infected	22,485,833 (65.22%)	22,485,833 (34.78%)

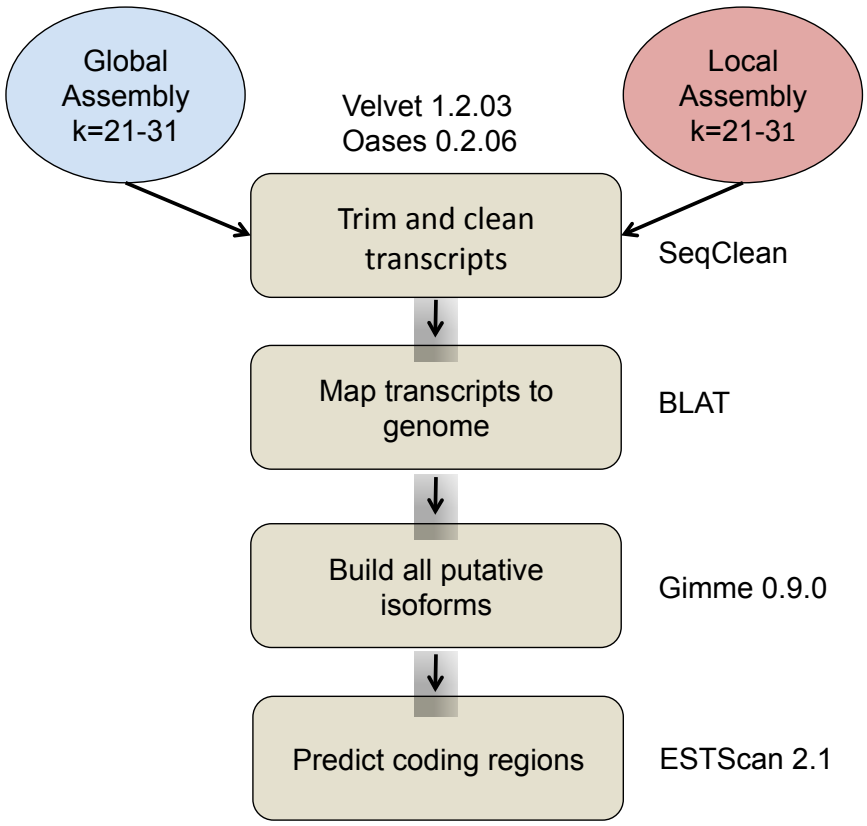

Figure 1. Gene model construction pipeline. Transcripts are obtained from two assembly methods – global and local assembly. Transcripts are aligned to a chicken genome by BLAT. Gimme then constructs gene models based on alignments of transcripts.

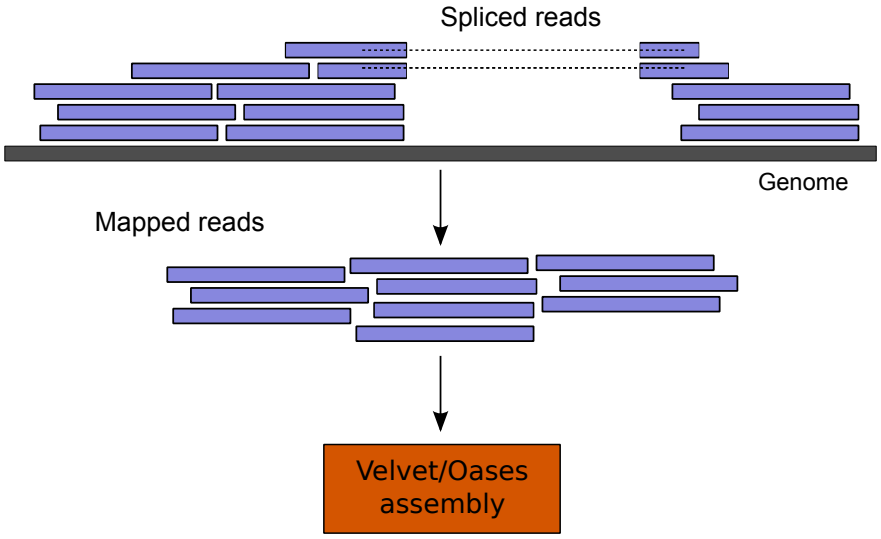

Figure 2. Local Assembly Pipeline. Reads are first mapped to a chicken genome. Then only mapped reads are assembled by Velvet and Oases.

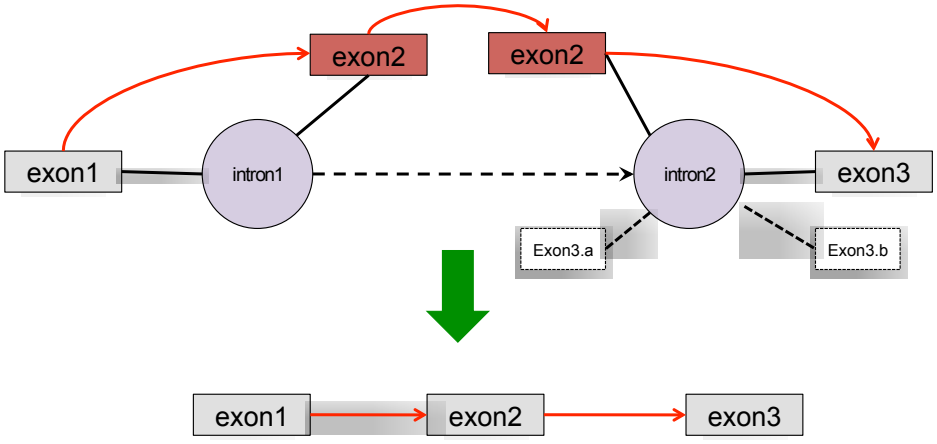

Figure 3. Intron and exon graphs. Each intron connects to exons whose splice junctions match it boundary. Some exons are excluded from the final gene model if they are incomplete (exon 3a,b). Introns sharing at least one exon are grouped together. Then an exon graph is made using exons as nodes.

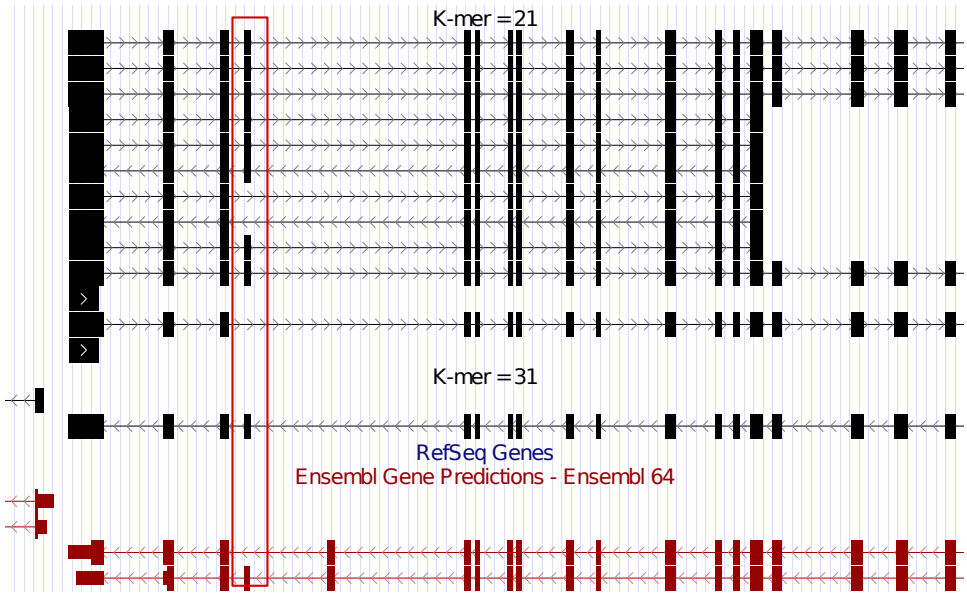

Figure 4. Different isoforms are detected by different k-mer lengths. K-mer=21 detects a skipped exon which is not detected by k-mer=31. The skipped exon is also annotated in Ensembl gene models.

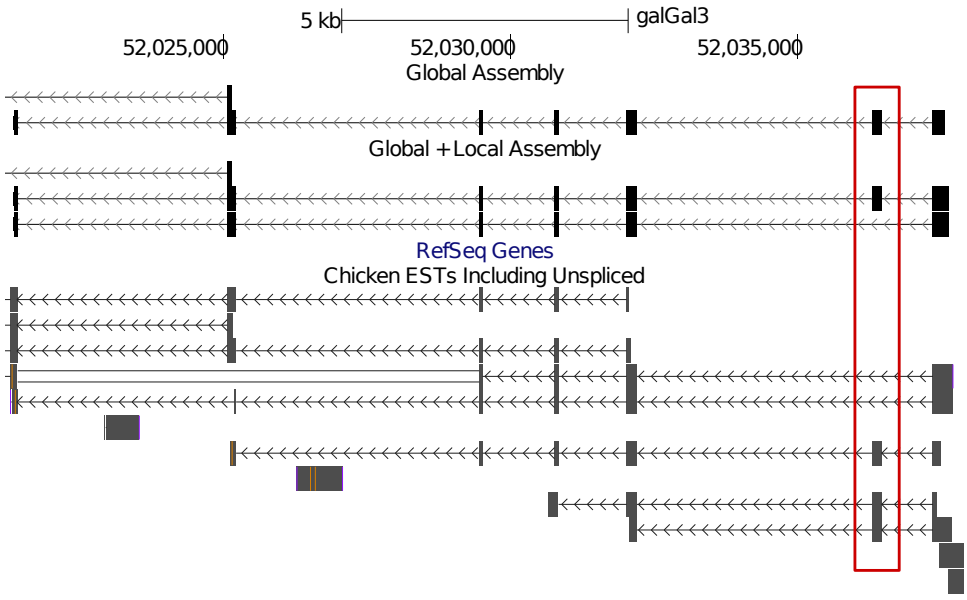

Figure 5. Global and local assembly detect different isoforms with the same k-mers.

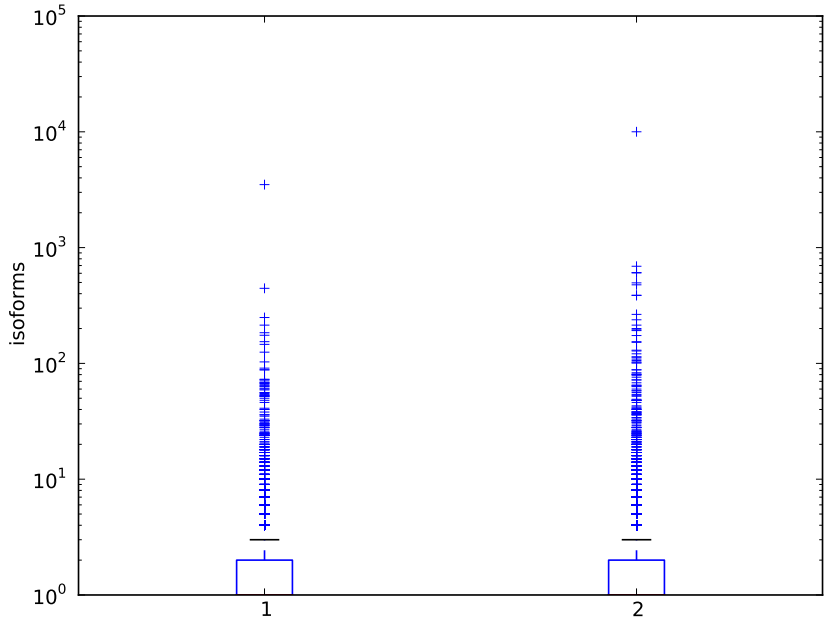

Figure 6. Distribution of a number of isoforms in each gene from 1) global and 2) local assembly

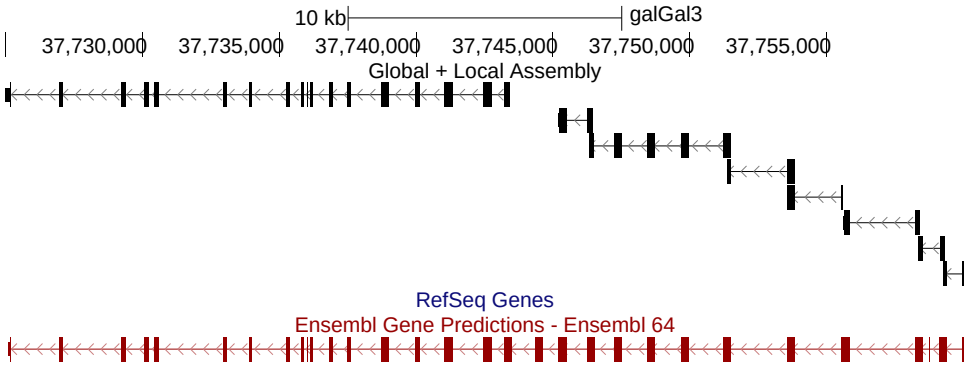

Figure 7. Example of fragmented transcripts near $5'$ end of a long transcript.

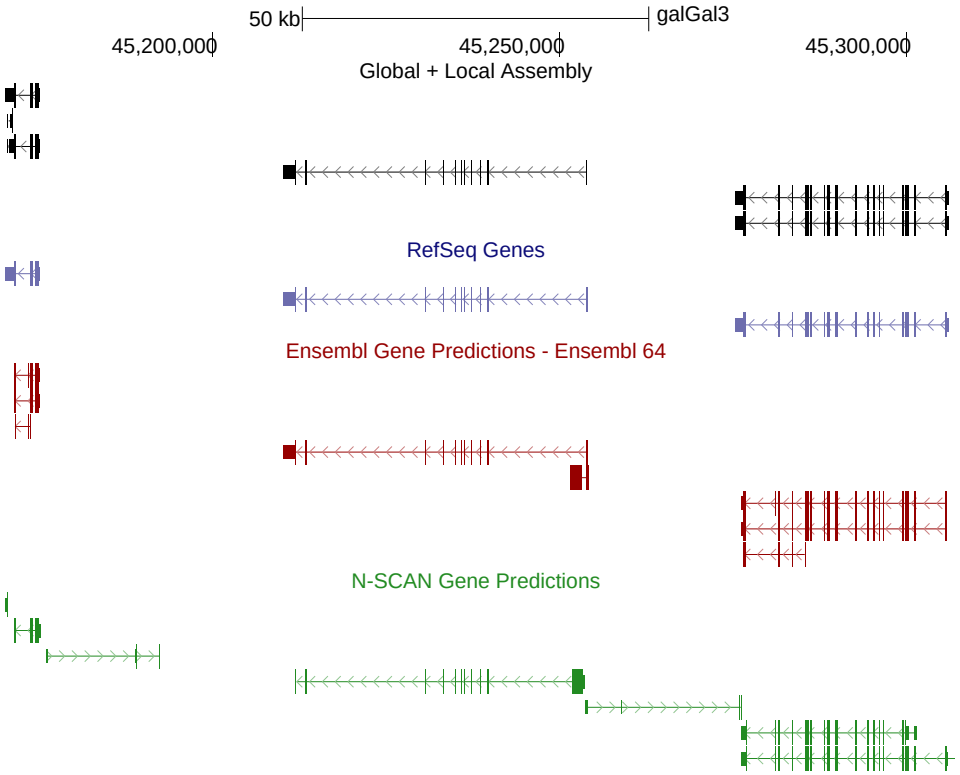

Figure 8. Comparison of gene models from our pipeline and other public gene models on UCSC genome browser.

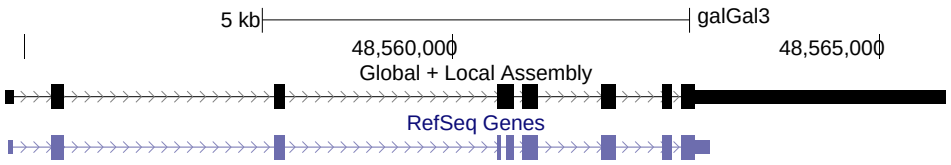

Figure 9. Examples of an extended $3'$ UTR detected in RNA-Seq gene models.

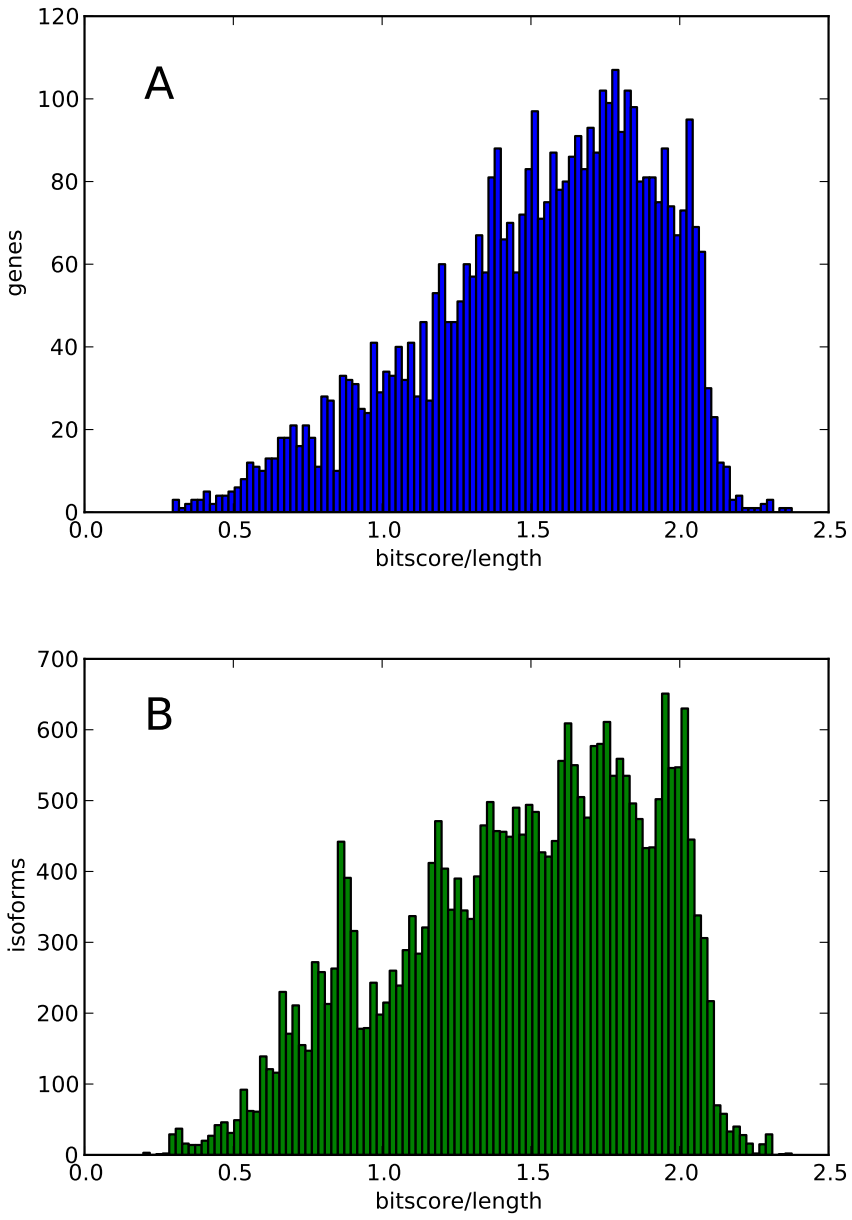

Figure 10. Histogram of bit score/length ratio of isoforms and genes that match mouse proteins. Genes in chromosome E64_random are not included.

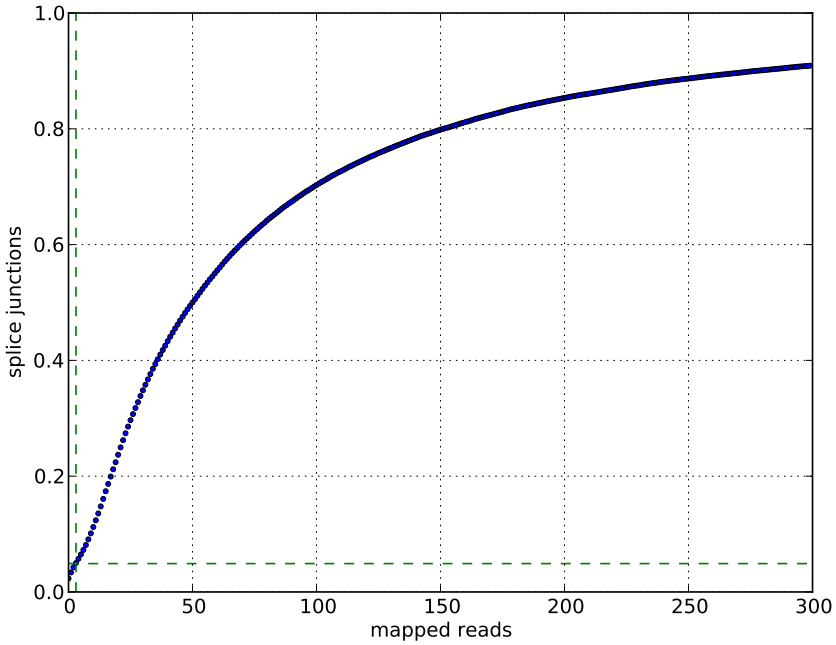

Figure 11. Cumulative counts of splice junctions with spliced reads.

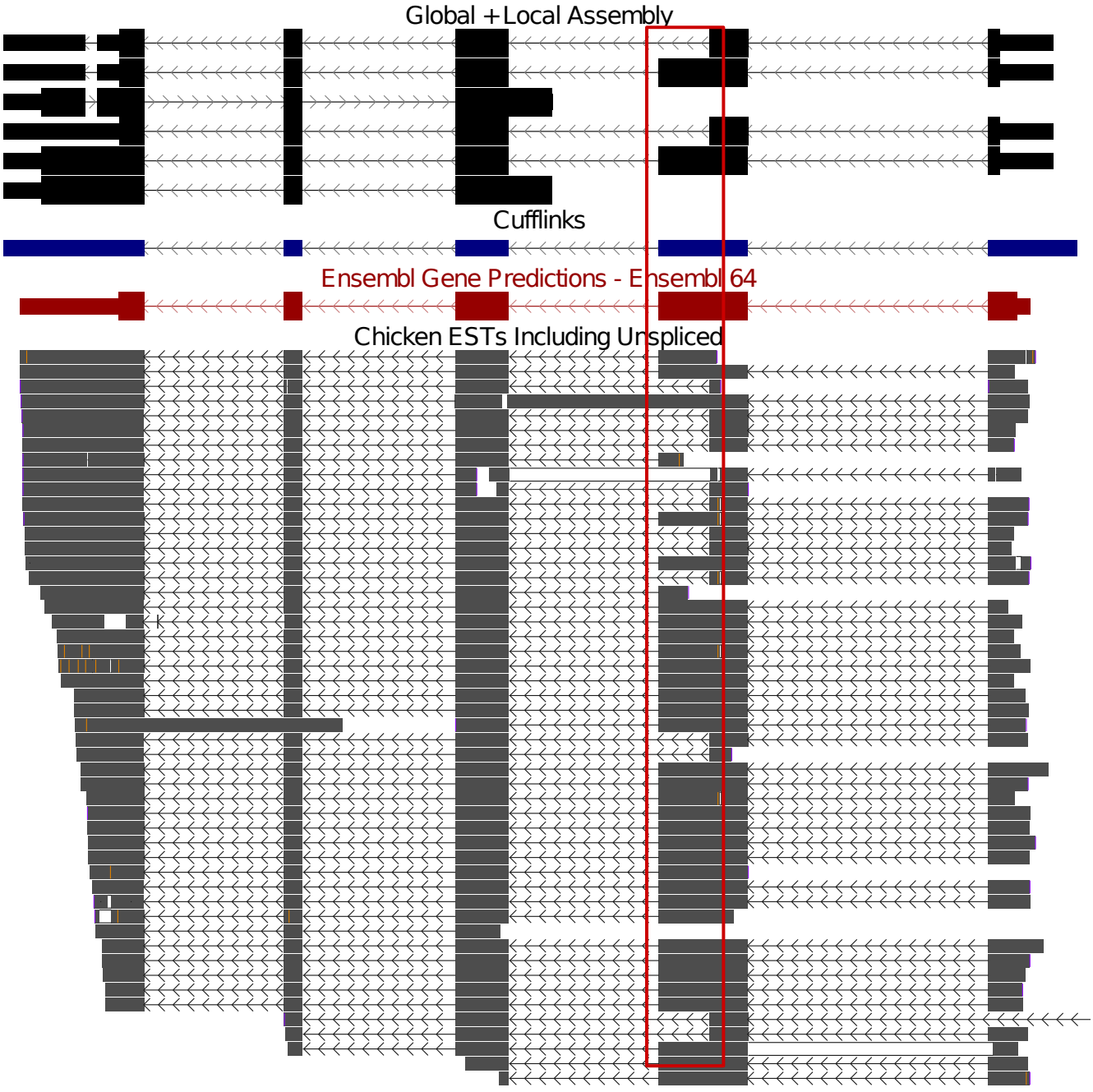

Figure 12. Unannotated alternative splice site. The pipeline can detect alternative splice site that is not annotated in Ensembl and Cufflinks.