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 uses a technique called local assembly that enhances the sensitivity of alternative splicing detection. We demonstrate that the pipeline detects many novel splice variants in RNA-Seq data from chicken spleen. Many of these splice variants; however, are only detected by our pipeline and not Cufflinks. This indicates that the pipeline can be used to facilitate splice variant detection from RNA-Seq.

Author Summary

Introduction

Until recently, studies of alternative splicing have 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 deep short-read sequencing has been used successfully in many studies to gain unprecedented insight into a complexity of transcriptomes.

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 have not been explored.

Despite the small size of sequencing reads, several studies have detected novel splice junctions based on alignment of reads spanning across putative exon junctions. To map reads across exon junctions, reads are split into two parts and each part is mapped to the genome independently (@cite). A splice junction is then identified based on alignments of each half of a read that fall between two exons at exon-intron boundaries. With this approach, Wang *et al* have identified a large number of splice junctions that are not annotated from human cell lines (HUVEC and NHEK) [2]. These novel splice sites include both canonical and non-canonical splice sites. Approximately, $46\% - 75\%$ of canonical splice sites are supported by ESTs. Novel splice junctions have different levels of read coverage suggesting that both high- and low-expressed isoforms are unannotated.

Using a similar approach, Pickrell *et al* identified more than 150,000 novel canonical splice junctions in lymphoblastoid cells. The study also shows that the number of unannotated splice junctions varies among cells from different human tissues, which suggests tissue-specific expression of isoforms [3]. A majority of unannotated isoforms found in both studies are from alternative splicing of annotated exons. Only a small fraction of isoforms contain previously unknown exons.

Several tools have been developed not only to detect novel splice junctions but also to reconstruct full-length isoforms from short reads without using prior gene annotations. These tools are especially useful for transcriptome analysis of organims with incomplete gene annotations. Cufflinks [4] relies on splice junctions detected from Tophat [5], a read aligner that can align reads across putative exon junctions, to reconstruct a full-length transcript. Cufflinks identified 12,712 novel isoforms, of which 7,395 (58%) contain novel splice junctions in mouse myoblast cell lines. Guttman *et al* used Scripture, a tool employing similar mapping-based approach, to reconstruct a full-length transcripts from mouse RNA-Seq data and discovered approximately 490 novel alternative isoforms in lincRNA loci, which are expressed in any of the three different cell types [6].

Although these mapping-based methods have been useful in detecting both splice junctions and isoforms, they rely heavily on a reference genome. Hence, it is not necessarily practical to apply these methods to organisms lacking a high-quality reference genome. This limitation can be overcome by *de novo* assembly of short reads.

A number of *de novo* assemblers have been used to reconstruct transcripts from RNA-Seq data in many studies. Trinity [7], was successfully used to reconstruct transcripts from yeast and mouse datasets.

It was also shown that Trinity detects a unique set of novel splice junctions not detected by Cufflinks or Scripture. This suggests that a *de novo* assembly approach is capabale of increasing sensitivity of detecting alternative isoforms of a mapping-based method. Trans-Abyss [8] and Oases [9] are extensions of the Abyss [10] and Velvet [11, 12] genome assemblers that are tuned to work with RNA-Seq data. These assemblers are comparable at reconstructing existing and novel alternative isoforms with a slightly different sensitivity and specificity. However, Oases with Oases-M has been shown to be superior than other *de novo* assemblers at discovering isoforms in human and mouse [9].

In this study, we present a pipeline that uses *de novo* assembly to reconstruct alternative isoforms in RNA-Seq data from chickens. We apply a technique we call "local assembly" that enhances the sensitivity of alternative isoform detection by Oases. The results show that the pipeline can detect more isoforms than Oases-M and can detect isoforms not found by Cufflinks. We also showed that transcripts reconstructed from *de novo* assembly and mapping-based approach can be merged to build more complete gene models.

Results

Local Assembly Enhances Isoform Detection

We used the Velvet [11] and Oases [9] assemblers to construct transcript fragments from four entire Illumina GAII mRNAseq data sets sequenced from chicken spleen (see Methods and Materials). In the assembly, we used multiple distinct k-mer values for Velvet to sensitively recover as many different isoforms as possible [9]. We chose k-mers between 21 and 31, and recovered between 340,000 and 370,000 unique sequences from each data set (see Table 1, and Materials and Methods). These unique sequences represented an unknown number of true genes, due to fragmentation from low coverage and incomplete assembly.

We next used Tophat to align mRNAseq reads to the genome and partition reads by chromosome; we then assembled the partitioned reads with Velvet and Oases using the same range of parameters as the global assembly, above. While the local assemblies were considerably more computationally efficient, they lacked several thousand unique regions that were present in the global assembly (Table 2); this is probably due to the incomplete nature of the current chicken genome assembly, which is lacking approximately 5% of its true gene content (@cite). Interestingly, over a hundred regions were present *only* in the

local assemblies, suggesting that the local assemblies might be recovering additional exons. Significant numbers of unique regions from both global and local assemblies showed homology to the mouse genome, indicating that at least some of these unique sequences represented real sequence content. Figure 5 shows an example of different isoforms detected by the two assembly methods.

Oases-M discards splice variants

The above approaches recovered transcript fragments, but not entire genes (figure in suppl?) To construct a more comprehensive gene set containing all of the assembled contigs, we tried using Oases-M to merge the assemblies from multiple k values [9]. While it has been demonstrated that merged transcripts from multiple-k assemblies contain more isoforms than those from any single k, the sensitivity of Oases-M for recovering splice variants has not been fully evaluated. We merged transcripts from our global assembly, above, with Oases-M using a k-mer size of 27, and compared them with the unmerged transcripts. We then cross-validated using publicly available ESTs, which were not used in our assembly. The results show that Oases-M and the unmerged assembly share about 104,413 (95%) of the predicted splice junctions, with 5% disjoint. Of these 5%, approximately 421 (6%) of the Oases-M-specific splice junctions are independently supported by ESTs, while 1,607 (19%) of the unmerged splice junctions are supported by ESTs. This suggests that Oases-M probably discards a number of real splice variants, although the unmerged assembly is also missing some found by Oases-M.

Exon graphs can reconstruct putative splice variants

We used an exon graph approach to construct gene models from alignments of transcripts against the genome. Our approach, implemented in a software package called Gimme, merges transcripts and gene models based on overlapping exons using an exon-graph approach (see Materials and Methods; Figure 3). We used Gimme to obtain 20,439 gene models containing 25,859 isoforms from our global assembly; 21,963 gene models containing 25,859 from our local assembly; and 20,922 gene models containing 27,732 isoforms from the merged global and local assemblies (Table 4).

Validation of Gene Models

Our pipeline predicts many gene models and isoforms after assembly. We validated these gene models in several different ways. XX more!

The gene models include most reads

We used bowtie to map single-end reads from the source datasets to the transcripts. More than 80% of the original reads could be aligned to the transcripts, demonstrating that we did not lose a significant amount of information during the merge process (Table 5). More importantly, we also mapped paired-end reads from technical replicates to the same gene models, and found that more than 85% of the paired-ends mapped concordantly the gene models (Table 6). Most of the reads that did not map were either highly erroneous or contained low-complexity artifactual sequence that probably originated from sample processing and reverse transcription (@cite). Thus the merged gene models produced by Gimme represent the significant majority of the assemblable data.

Almost all splice junctions have high coverage

To validate the splice junctions reconstructed by the Gimme pipeline, we used Bowtie to map mRNAseq reads directly to the transcript sequences derived from gene models [13]. Because the Velvet/Oases assembly pipeline does not make use of the reference genome, reads that map across a splice junction constitute independent verification of a splice junction's presence in a transcript.

Of 100,411 splice junctions from the gene models, 95 (0.09%) junctions have no spliced reads and only 448 (0.45%) junctions have fewer than 4 spliced reads (Fig. 11). More than 99% of our predicted splice junctions have a coverage of 4 or higher in our combined mRNAseq data sets, suggesting that they are real splice junctions.

Most splice junctions are independently supported

Of the 100,411 splice junctions in our gene models, 81,838 (81.50%) are supported by ESTs or mRNAs from Genbank. This is especially surprising since our mRNAseq data is from spleen, and most of the publicly available ESTs or mRNAs are from other tissues. Note that this cross-validation suggests that the 18,573 novel splice junctions *not* seen in publicly available ESTs and mRNAs are also likely to be real splice junctions from spleen.

Our pipeline improves on existing reference-based approaches

We next compared the Gimme gene models to those produced by Cufflinks, another reference-based approach to building gene models from mRNAseq data [4]. We also compared the results from both methods to the ENSEMBL gene annotations, which are produced by a pipeline that incorporates de novo gene prediction and homology-based approaches as well as expression data (@cite).

Cufflinks finds 109,641 splice junctions, and Gimme finds 100,411 splice junctions. 88,289 of them are in common. (Venn diagram.)

Both Cufflinks and Gimme find approximately 40-50% of the genes and 80-90% of the splice junctions present in the ENSEMBL gene models for chicken. Cufflinks performs about 10% better in both cases than Gimme, demonstrating that Cufflinks has higher sensitivity in recovering ENSEMBL gene models.

The ENSEMBL pipeline does not, however, include a large number of splice junctions from ESTs (99,755) or mRNAs (13,641). Cufflinks and Gimme each recover about 10% of these, with more than 2/3 of these recovered by both Cufflinks and Gimme. This indicates that both Gimme and Cufflinks are equally adept at recovering novel splice junctions.

When we apply Gimme and Cufflinks to a publicly available mouse mRNAseq data set, Gimme and Cufflinks recover approximately the same number of splice junctions already known from ENSEMBL. However, Gimme recovers a substantial number of additional splice variants beyond Cufflinks and ENSEMBL both.

Gimme can iteratively merge sets of gene models

As shown above, Cufflinks and the assembly method detect a number of distinct but equally valid splice junctions, which suggests that we could obtain greater sensitivity to exon-exon junctions in our gene models by merging both sets of predictions. We therefore used Gimme to merge the Cufflinks and assembly gene models Table. 4. This resulted in a decreased number of total genes, suggesting that some fragmented genes were merged together to form more complete gene structures (e.g. see Fig. **??**). The merged gene models recover 44.19% and 58.57% of splice junctions from ESTs and ENSEMBL respectively, which is 10-15% greater than that from corresponding unmerged gene models.

Validating chicken sequences by using mouse homologs

To validate our predicted isoforms, we extracted putative coding sequences from our gene models with ESTScan [14]. ESTScan successfully translated 22,488 of 27,732 (81.1%) of our isoforms to protein sequences with 50 or more amino acids. We then searched for homologous sequences in mouse ENSEMBL, and found that 15,399 (68.47%) of our isoforms from 12,399 distinct genes match mouse proteins at a bit score $\geq$ 1.0. These matches have a bit-score/length ratio greater than 1, which indicates a good agreement between chicken and mouse proteins.

Leftover Results text

The program can merge all transcripts from multiple-kmers and predict a structure of full-length isoforms and genes. Total of 20,439 genes were obtained from global assembly and 21,963 genes from local assembly. However, the number of genes obtained from global + local assembly is only 20,922 genes. The number of isoforms slightly increased after combining transcripts from global and local assembly together (Table 4).

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 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. 6. Moreover, spurious splice junctions due to duplication in genome sequence can lead to a large number of dubious isoforms. This seems to occur in a few genome sequences such as chromosome E64_random. Figure **??** shows that a majority of genes from our gene models have only a few isoforms.

In addition, parameters used in building gene models such as minimum exon/intron size determine how transcripts are merged together, which can increases or decreases the number of isoforms. In this study, we adjust the parameters to enhance sensitivity of splice junctions detection that may result in

overestimation of the number of isoforms.

Most transcripts from RNA-Seq assembly include 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 (cite). RNA-Seq-derived gene models are useful for studying variations within UTR regions, which may be involved in regulation of isoform expression (cite). Figure 8 and figure 9 shows an example of alternative 5′UTRs and extended 3′UTR respectively.

Cufflinks

To evaluate an efficiency of our pipeline in detecting splice junctions, we compared the number of splice junctions detected by our pipeline to that from , ESTs and Ensembl gene models version 64. Cufflinks detected 88,289 (87.93%) splice junctions that are also detected by our pipeline. However, 42.43% (5,143) of junctions not detected by Cufflinks are supported by ESTs or mRNAs. 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 7,298 genes and Cufflinks detects 8,341 genes of total 17,934 Ensembl genes. 6,798 of those genes are both detectd by our pipeline and Cufflinks. In addition, we wanted to compare the 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 the number of splice junctions detected by both methods within the same gene instead. From 6,798 genes, our pipeline detects 56,332 splice junctions, whereas Cufflinks detects 62,150 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 6,798 genes and found that our pipeline detects 1,997 splice junctions and Cufflinks only detects 1,860 splice junctions. Figure 15 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.

isoforms and mouse

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. Combination of both methods results in more complete gene models, which suggests that combination of *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.* [9], 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 considerable 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?.

Data

Mouse RNA-Seq dataset (SRX062280) is downloaded from Short Read Archives (SRA). Chicken RNA-Seq datasets were obtained from sequencing of mRNAs from spleen of chicken line 6 and 7.

Mapping reads to The Reference Genome and Gene Models

Single and paired-end reads were mapped to chicken genome by Tophat2 [5] release 2.0.0 using default parameters without annotations. All reads were mapped to cDNA sequences derived from gene models by Bowtie2 [13] with default parameters `(n=2, l=28, e=70, k=1)`. Reads from mouse dataset were mapped to mouse genome (mm9) downloaded from Tophat website `http://tophat.cbcb.umd.edu`.

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 [11] 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 [9].

A poly-A tail, short transcripts and transcripts with low complexity are removed by seqclean [**?**] with default parameters. Redundant transcripts are removed by cd-hit-est from CD-HIT suite [15]. A large number of transcripts are removed at this step, which facilitates gene models construction process.

We obtained 334,475 transcripts, of which 295,588 transcripts (88.37%) mapped to chicken genome. Only transcripts mapped to chicken genome are used to build gene models.

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 [16] `(-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 highly similar ($> 99\%$) are clustered and removed by CD-HIT version 4.5.6 [15]. Parameters for clustering are `-n 10, -r 0 and -c 0.99`. 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 boundaries. 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 [**?**]. 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 [**?**]

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 [4]. 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. The database was updated from GENBANK on 1 January 2014. Sequences were aligned 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. Wang L, Wang X, Wang X, Liang Y, Zhang X (2011) Observations on novel splice junctions from RNA sequencing data. Biochemical and biophysical research communications 409: 299–303.

3. Pickrell JK, Pai AA, Gilad Y, Pritchard JK (2010) Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet 6: e1001236.

4. 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.

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

6. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology 28: 503–510.

7. 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.

8. 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.

9. 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) .

10. 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.

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

12. 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.

13. 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.

14. 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 .

15. 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.

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

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. ... .

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	2,132	104	1,322 (62.01%)	39 (37.50%)
Line 6 infected	2,499	104	1,514 (60.58%)	40 (38.46%)
Line 7 uninfected	2,633	136	1,560 (59.25%)	52 (38.24%)
Line 7 infected	2,409	152	1,390 (57.70%)	50 (32.89%)

Table 3. Number of total and unique splice junctions

Method	Total	Unique	Unique/supported by ESTs
Oases-M assembly	111,237	6,860	421 (6.1%)
Unmerged assembly	112,708	8,295	1,607 (19.4%)

Table 4. Number of putative genes and isoforms

Method	Gene	Isoform
Global	20,439	26,035
Local	21,963	25,859
Global + Local	20,922	27,732
Cufflinks	25,318	34,959
Global + Local + Cufflinks	21,734	32,855
Global + Local + Cufflinks + mRNAs	26,726	43,074
Ensembl	17,934	23,392

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

Dataset	Mapped	Unmapped
Line 6 uninfected	20,169,993 (85.54%)	3,409,559 (14.46%)
Line 6 infected	18,790,831 (80.13%)	4,658,150 (19.87%)
Line 7 uninfected	19,844,293 (82.94%)	4,081,516 (17.06%)
Line 7 infected	21,772,102 (85.84%)	3,590,465 (14.16%)

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

Dataset	Mapped	Unmapped
Line 6 uninfected	28,696,112 (85.36%)	4,967,765 (14.76%)
Line 6 infected	20,514,438 (85.93%)	3,394,055 (14.20%)
Line 7 uninfected	28,159,776 (85.65%)	4,761,556 (14.46%)
Line 7 infected	29,564,592 (85.88%)	4,913,243 (14.25%)

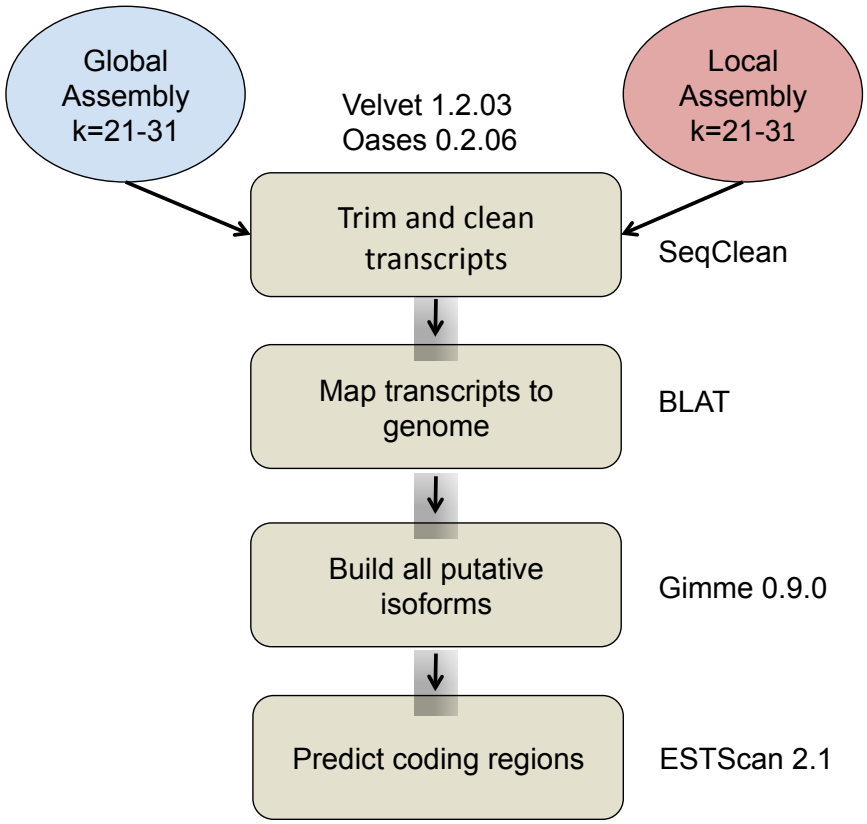

Figure 1. Gene model construction pipeline. Transcripts are obtained from two assembly methods – global and local assembly. Transcripts are aligned to a chicken genome by BLAT. Gimme then constructs gene models based on alignments of transcripts.

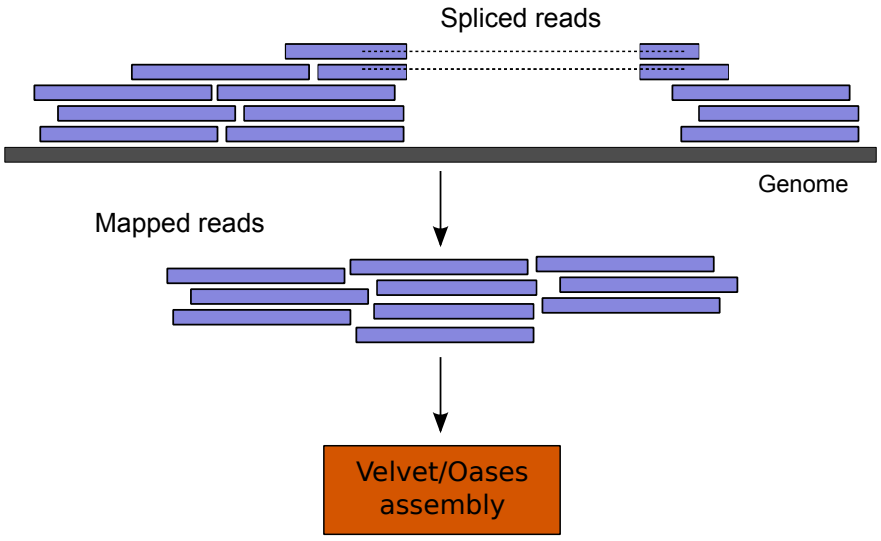

Figure 2. Local Assembly Pipeline. Reads are first mapped to a chicken genome. Then only mapped reads are assembled by Velvet and Oases.

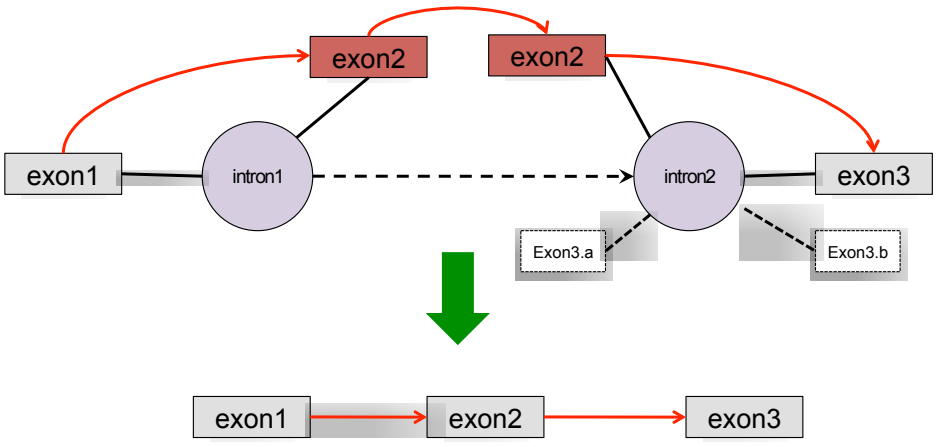

Figure 3. Intron and exon graphs. Each intron connects to exons whose splice junctions match it boundary. Some exons are excluded from the final gene model if they are incomplete (exon 3a,b). Introns sharing at least one exon are grouped together. Then an exon graph is made using exons as nodes.

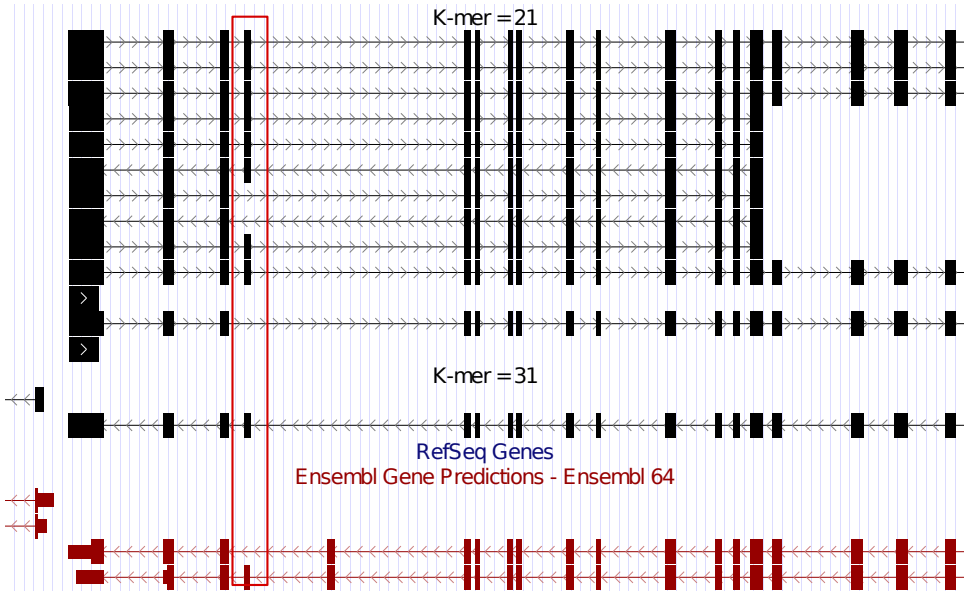

Figure 4. Different isoforms are detected by different k-mer lengths. K-mer=21 detects a skipped exon which is not detected by k-mer=31. The skipped exon is also annotated in Ensembl gene models.

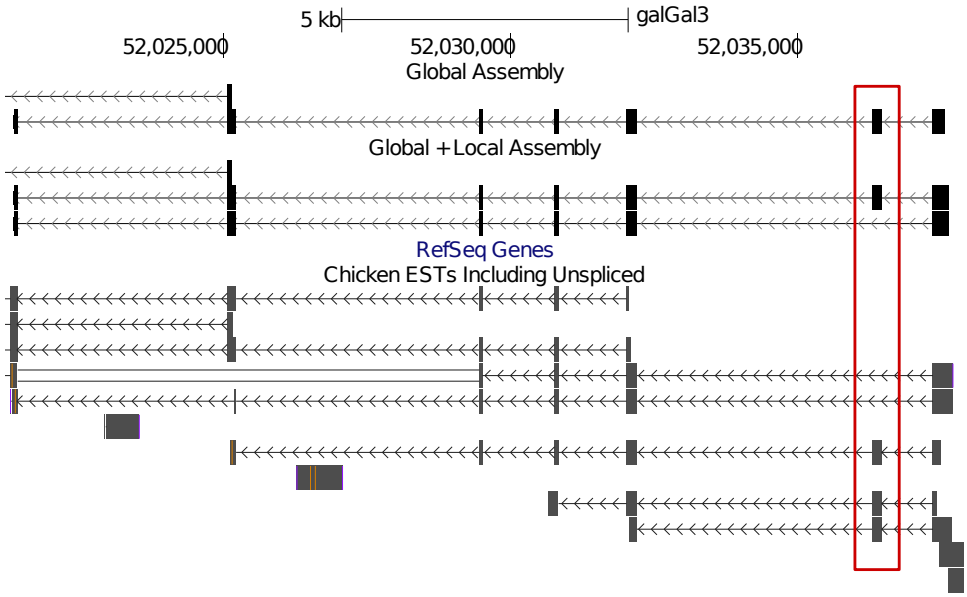

Figure 5. Global and local assembly detect different isoforms with the same k-mers.

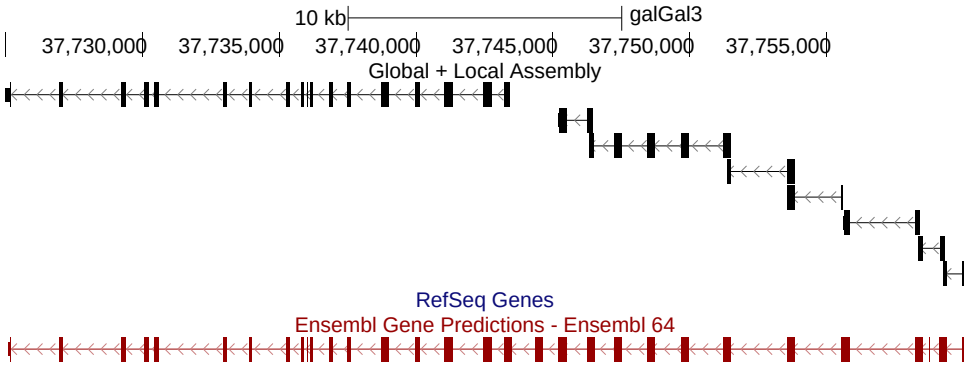

Figure 6. Example of fragmented transcripts near $5'$ end of a long transcript.

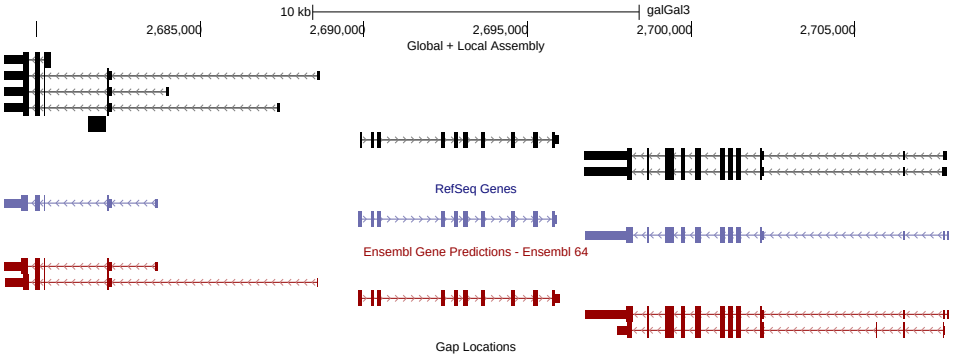

Figure 7. Comparison of gene models from the *de novo* assembly pipeline with reference and Ensembl gene models on UCSC genome browser.

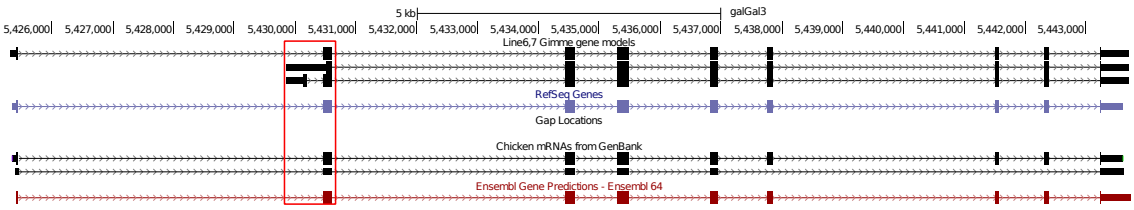

Figure 8. Examples of alternative $5'$ UTRs in RNA-Seq gene models.

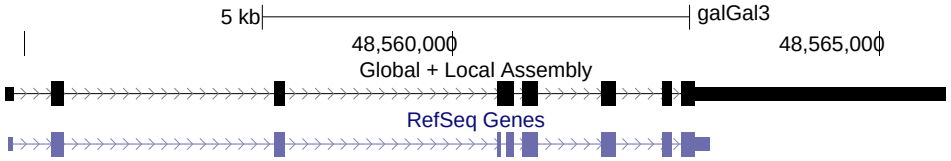

Figure 9. Examples of an extended $3'$ UTR in RNA-Seq gene models.

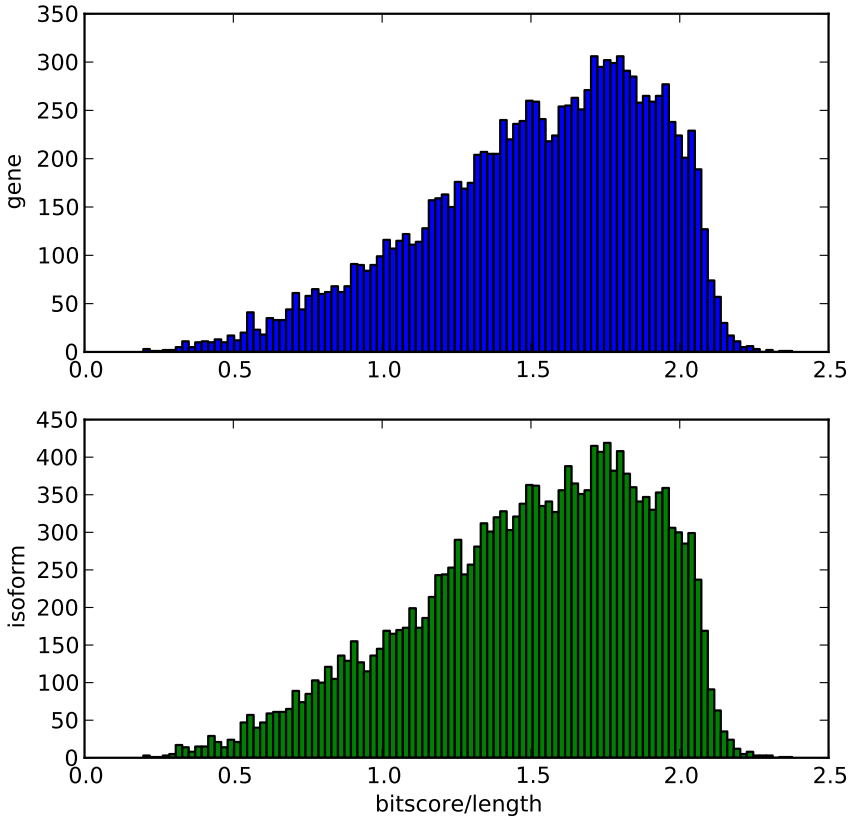

Figure 10. Histogram of bit score/length ratio of isoforms and genes that match mouse proteins.

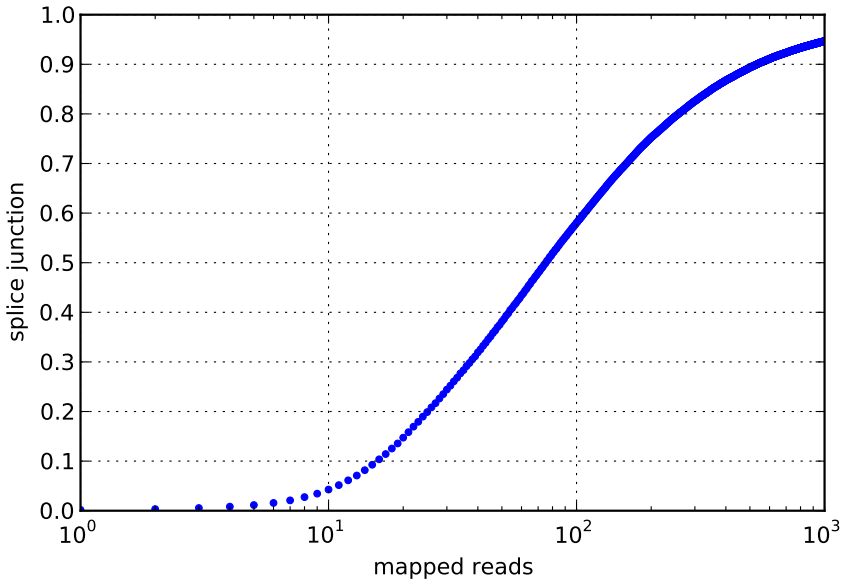

Figure 11. Cumulative counts of splice junctions with spliced reads up to 1000 reads.

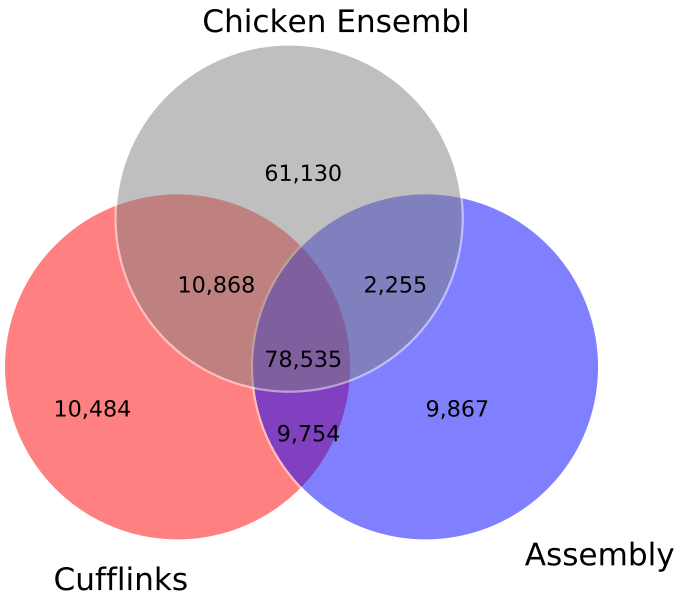

Figure 12. Splice sites in chicken Ensembl gene models detected by Cufflinks and the *de novo* assembly pipeline. Cufflinks detects many annotated isoforms that are not detected by the pipeline. The figure also shows that both methods detect a large number of unannotated splice junctions, which suggests that those junctions may be genuine.

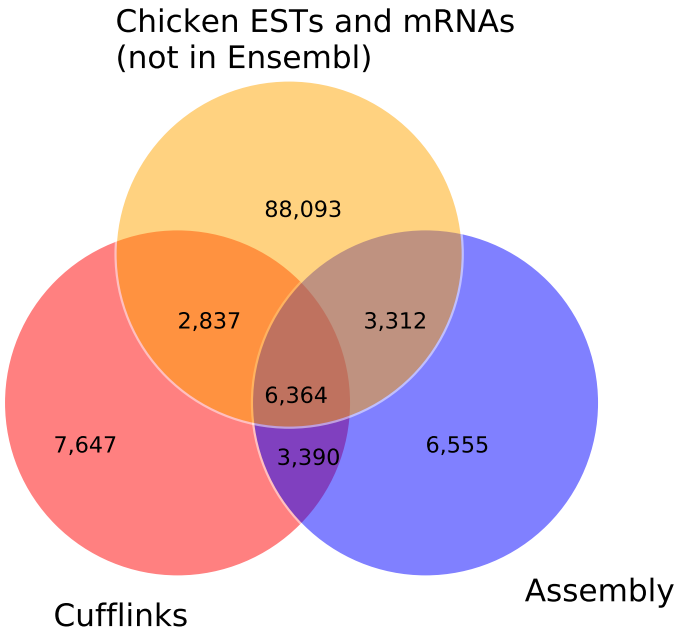

Figure 13. Splice sites in Chicken ESTs detected by Cufflinks and the *de novo* assembly pipeline. In contrast to Ensembl gene models, the pipeline and Cufflinks detects the similar number of splice junctions from ESTs and mRNAs. This indicates that the pipeline is as efficient as Cufflinks at detecting non-annotated splice junctions.

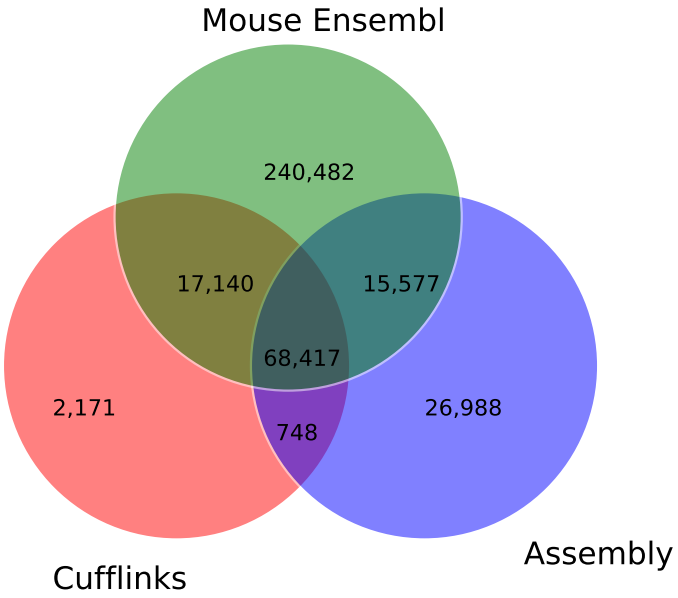

Figure 14. Splice sites in mouse Ensembl gene models, Cufflinks and the *de novo* assembly pipeline. In contrast to chicken datasets, Cufflinks and the pipeline detects the similar amount of non-overlapped annotated junctions. This may due to a higher quality of gene annotations in mouse.

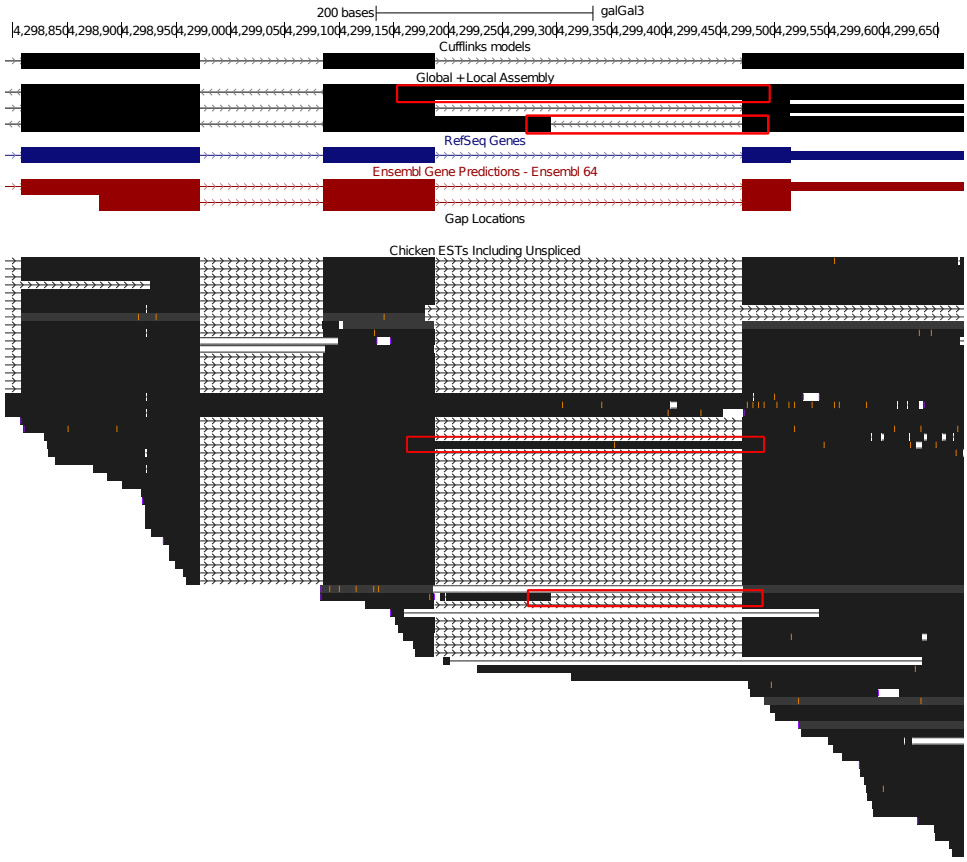

Figure 15. Unannotated alternative splice site. The pipeline detects alternative splices site not annotated in Ensembl and Cufflinks but are supported by ESTs.

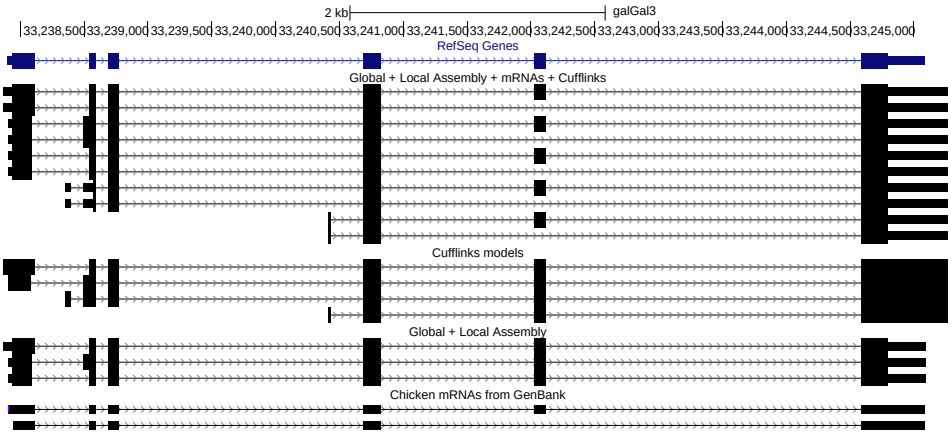

Figure 16. mRNAs + Cufflinks + Assembly gene models. Gimme can combine transcripts from different sources to build gene models. In this figure, the final gene model includes several isoforms not annotated in the reference gene model.