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

Splice variants play an important role in biological systems, especially in the immune system and the central nervous system. An expression profile of splice variants has been shown to be a better signature of some diseases than an overall gene expression profile [1]. However, for most organisms, available gene models such as Ensembl may not include all splice variants expressed in RNA-Seq data. Reference-guided (e.g. Cufflinks [2]) and de novo assembly (e.g. Velvet/Oases [3]) methods are available for building gene models from RNA-Seq reads. However, we found that both types of method recovered distinct sets of splice variants. In this article, we introduce a pipeline that can combine transcripts from *de novo* assembly and other gene models to constructs gene models that include more splice variants. We also described a method called local assembly that we used to enhance the assembly of splice variants to recover splice variants not found by Cufflinks and *de novo* assembly. Combining Cufflinks gene models and transcripts from assembly yields more splice variants, increasing sensitivity of alternative splicing studies.

Author Summary

Introduction

Until recently, studies of alternative splicing have been limited to a small number of genes and isoforms due to the high cost and low throughput of sequencing 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 the complexity of transcriptomes (@cite). It has been estimated that, in human, $92-94\%$ of multiexon genes undergo alternative splicing and different isoforms are expressed in different tissues [4]. 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 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 falls 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) [5]. 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* [6] 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 [6].

Several reference-based 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 [2] relies on splice junctions detected from Tophat [7], a read aligner that can align reads across putative exon junctions, to reconstruct a full-length transcript. In mRNAseq from mouse myoblast cell lines, Cufflinks identified 12,712 novel isoforms, of which 7,395 (58%) contain novel splice junctions. Guttman *et al* used Scripture, a tool employing a similar mapping-based approach, to reconstruct 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 [8]. Although these mapping-based methods have been useful in detecting both splice junctions and isoforms, they rely heavily on a high-quality reference genome. Hence, it is not necessarily practical to apply these methods to organisms lacking such a reference.

The requirement for a high-quality reference genome 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 [9] 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 capable of increasing sensitivity of detecting alternative isoforms over a mapping-based method. Trans-Abyss [10] and Oases [3] are extensions of the Abyss [11] and Velvet [12, 13] 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 to other *de novo* assemblers at discovering isoforms in human and mouse [3].

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 approaches can be merged to build more complete gene models.

Results

Local Assembly Enhances Isoform Detection

We used the Velvet [12] and Oases [3] 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 [3]. We chose k-mers between 21 and 31, and recovered between 90,700 and 104,000 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 separately 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 these “local” assemblies were considerably more computationally efficient, they lacked several thousand unique regions that were present in the global assembly (Table 2,

and Table 3); 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 4 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 [3]. 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 (94%) of the predicted splice junctions, with 6% disjoint. Of these 6%, approximately 420 (6.1%) of the Oases-M-specific splice junctions are independently supported by ESTs, while 1,608 (19.4%) of the unmerged splice junctions are supported by ESTs (Table 4). 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 21,492 gene models containing 31,418 isoforms from our global assembly; 24,928 gene models containing 29,776 from our local assembly; and 22,587 gene models containing 34,800 isoforms from the merged global and local assemblies (Table 5).

Validation of Gene Models

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

The gene models include most reads

We used Bowtie to map single-end reads from the source datasets to the transcripts. More than $75-80\%$ 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 compared to the number of reads mapped to Ensembl gene models (Figure 5).

More importantly, we also mapped paired-end reads from technical replicates to the same gene models, and found that more than 74% of the paired-ends mapped concordantly the gene models. 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 [14]. 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 105,461 splice junctions from the gene models, 2,057 (2%) junctions have no spliced reads and only 4,626 (4.4%) junctions have fewer than 4 spliced reads (Fig. 6). Note that 710 junctions are in chrUn_random contigs, which may have a great number of genome missassemblies. The number of junctions outside random chromosomes with no reads is 1,347/99,986 (1.3%). More than 95% 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 105,461 splice junctions in our gene models, 83,560 (79.2%) are supported by ESTs or mRNAs from Genbank (Fig. 7). This is especially surprising since our mRNAseq data is from spleen, and most of the publicly available ESTs or mRNAs are from other tissues. 16,909 splice junctions are found by either Cufflinks or our gene models, but not both, suggesting a variation in sensitivity between two methods. Note that this cross-validation suggests that the 12,188 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 [2]. 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 92,077 splice junctions, and Gimme finds 105,461 splice junctions. 80,964 of them are in common (Figure 8). Both Cufflinks and Gimme find approximately 40-50% of the genes and 50-53% of the splice junctions present in the ENSEMBL gene models for chicken. The ENSEMBL pipeline does not, however, include a large number of splice junctions from ESTs (97,740) or mRNAs (13,987). Cufflinks and Gimme each recover about 18% 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 (Figure 9). 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 Gimme 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 incorporate the Cufflinks models to global and local assembly, Table. 5. 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 49.3% and 56.2% of splice junctions from ESTs and ENSEMBL respectively (Fig. 10), which is about 10% 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 [15]. ESTScan successfully translated 28,772 of 34,800 (82.6%) of our isoforms to protein sequences with 50 or more amino acids. We then searched for homologous sequences in mouse ENSEMBL, and found that 22,991 (79.9%) of our isoforms from 12,945 distinct genes match mouse proteins at a bit-score/length ≥ 1.0 (Fig. 11). A bit-score/length ratio greater than 1 indicates that approximately more than half of the protein sequence matches mouse proteins well. The results suggest that merging gene models from Cufflinks and *de novo* aseembly did not disrupt the structure of protein sequences.

Discussion

De novo assembly should be used to extend gene models from Cufflinks

For organisms with a reference genome and ENSEMBL annotation, Cufflinks with ENSEMBL gene models as a reference guide seems to be a preferable method for building gene models. However, we have shown that ENSEMBL models in chicken are missing a substantial number of splice variants found in ESTs and mRNAs that are also not recovered by Cufflinks. Moreover, results from mouse data suggest that Tophat's default parameters may have been tailored to maximize the finding of isoforms in ENSEMBL models (Fig. 9), which could lead to restricted detection of novel splice variants. Therefore, gene models built from RNA-Seq reads using Cufflinks + ENSEMBL gene models alone may not be suitable for genes and isoforms analyses that include finding novel genes and splice variants. *De novo* assembly is not tied to any specific gene model parameters and does not only detect canonical splice sites, thus we recommend that it is used to extend gene models built from Cufflinks + ENSEMBL gene models, especially in organisms with relatively incomplete ENSEMBL gene models. Besides *de novo* assembly, annotation from other sources can also be incorporated to extend Cufflinks models using our pipeline.

Reads with multiple alignments may be key to detection of unique splice variants from local assembly

We have shown that local assembly greatly increases sensitivity of splice variant detection; however, the mechanism is still not clearly understood. In local assembly, reads mapped to each chromosome are assembled separately and because Tophat reports multiple alignments for spliced reads, a single spliced read from an exon junction present in multiple chromosomes gets assembled multiple times. We speculate that this increases the coverage of exon junctions found in multiple chromosomes, which help improve the assembly of splice variants with low coverage. Further investigation of an inherent mechanism of local assembly may lead to a method that increases sensitivity of splice variants in organisms that lack a reference genome.

Spurious predicted isoforms are due to poor genome assembly

Both Cufflinks and our pipeline rely on a reference genome; therefore, the quality of the genome greatly affects the quality of the gene models. In this study, gene models were built from chicken genome version 2.1 (galGal3), which contains ~ 17 Mb of sequence duplications and missassemblies that were eliminated in the latest version of genome assembly (galGal4). Duplications and misassemblies lead to spurious splice junctions, which in turn produce spurious splice variants. Spurious junctions are non-canonical and a majority of them (17.8%) are in chrUn_random (sup.Fig.2). Figure 11 also shows that the median of bitscore/length ratios of isoforms is lower than that of genes, suggesting poorer alignments of isoforms to mouse protein sequences.

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 1.

Data

The mouse RNA-Seq dataset (SRX062280) was downloaded from the Short Read Archive. Chicken RNA-Seq datasets were obtained from sequencing of mRNAs from spleens of control and infected chicken (4 days post infection) lines 6 and 7 using standard Illumina protocol for non-strand specific single-end reads.

Mapping reads to The Genome and Gene Models

Single and paired-end reads were mapped to the chicken genome by Tophat [7] release 1.3.1 using default parameters without annotations. All reads were mapped to cDNA sequences derived from gene models by Bowtie [14] release 1.0.0 with default parameters Reads from the mouse dataset were mapped to the mouse genome (mm9) downloaded from Tophat website `http://tophat.cbcb.umd.edu`.

Global and Local Assembly

Reads from each dataset were first assembled separately by 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 chromosome in the local assembly (Fig. 2). Global and local assembly was performed using Velvet version 1.2.03 [12] with default parameters except for hash length (k-mer). A range of k-mer length from 21-31 was used to assemble reads from chicken data and k-mer length 27 was used to assemble reads from mouse data. Lastly, transcripts from both methods were assembled by Oases version 0.2.06 [3].

A poly-A tail, short transcripts and transcripts with low complexity are removed by SeqClean [16] with default parameters. Redundant transcripts are removed by `cd-hit-est` from the CD-HIT suite [17] release 4.5.4. A substantial number of transcripts are removed at this step, which facilitates gene model construction process. We obtained 339,199 transcripts, of which 315,998 transcripts (93.2%) 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 [18]. 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 the 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).

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 the 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 splice sites perfectly match intron boundaries. At the 5′ and the 3′ ends, some exons may be smaller than others due to incomplete assembly (supp. Fig.??). Only the largest exons are kept to ensure that the transcript has the most complete exons (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 3.0.3 to translate protein sequences from our gene models. The matrix used for building Hidden Markov models was built from chicken reference cDNA sequences using tools from ESTScan. Only protein sequences longer than 50 bp are included in the analysis.

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 300 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+ [19]. 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 1.0.0. The parameter is set for Bowtie to report up to 100 alignments per read. Reads mapped across exon junctions from all datasets were counted using Samtools [20] and Pysam [21].

Sequence assembly using Cufflinks

Reads are mapped to a genome sequence using Tophat. Gene models are built from each dataset by Cufflinks 2.0.0 [2]. All gene models are then merged together using Cuffmerge.

Expressed sequence tags and Genbank mRNA

Expressed sequence tags (ESTs) and mRNAs were downloaded from the UCSC genome website. The database was loaded from GENBANK on 1 January 2014. Sequences were aligned to the chicken genome using BLAT.

Pipeline for the study

The pipeline, scripts and instructions for reproducing this study are open source and available at `http://github.com/likit/gimme-paper`.

References

1. Zhang Z, Pal S, Bi Y, Tchou J, Davuluri RV (2013) Isoform level expression profiles provide better cancer signatures than gene level expression profiles. Genome medicine 5: 33.

2. 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.

3. 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) .

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

5. 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.

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

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

8. 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.

9. 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.

10. 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.

11. 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.

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

13. 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.

14. 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.

15. 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 .

16. Seqclean: a script for automated trimming and validation of ests or other dna sequences by screening for various contaminants, low quality and low complexity sequences. URL `http://compbio.dfci.harvard.edu/tgi/software/`.

17. 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.

18. Kent WJ (2002) Blatthe blast-like alignment tool. Genome research 12: 656–664.

19. Tatusova TA, Madden TL (1999) Blast 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS microbiology letters 174: 247–250.

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The sequence alignment/map format and samtools. Bioinformatics 25: 2078–2079.

21. Pysam: a python module for manipulating samfiles. URL `http://github.com/pysam-developers/pysam`.

Tables

Figure Legends

Table 1. Total unique sequences from global and local assembly (k=$21-31$)

Dataset	Total Sequence	
	Global	Local
Line 6 uninfected	90,705	68,845
Line 6 infected	104,785	70,191
Line 7 uninfected	90,125	63,302
Line 7 infected	92,192	67,097

Table 2. Unique sequences between global and local assembly

Dataset	Total size (bp)		Unique Sequence (bp)	
	Global	Local	Global	Local
Line 6 uninfected	77,454,439	36,662,830	3,686,835 (4.8%)	307,975 (0.8%)
Line 6 infected	86,622,623	37,877,766	4,157,541 (4.8%)	400,702 (1.0%)
Line 7 uninfected	76,566,717	33,571,348	4,180,202 (5.4%)	365,850 (1.1%)
Line 7 infected	74,957,624	33,824,849	4,242,922 (5.7%)	326,169 (9.6%)

Table 3. Unique regions from global and local assembly

Dataset	Unique Region		Matched with mouse proteins	
	Global	Local	Global	Local
Line 6 uninfected	1,285,929	96,830	260,321 (20.0%)	9,413 (9.7%)
Line 6 infected	1,631,356	59,813	312,849 (19.2%)	5,132 (8.6%)
Line 7 uninfected	1,800,634	104,229	349,346 (19.4%)	9,883 (9.5%)
Line 7 infected	1,611,354	125,640	296,915 (18.4%)	9,381 (7.5%)

Table 4. Number of total and unique splice junctions

Method	Total	Unique	Unique/supported by ESTs
Oases-M assembly	111,273	6,860	420 (6.1%)
Global assembly	112,708	8,295	1,608 (19.4%)

Table 5. Number of putative genes and isoforms

Method	Gene	Isoform
Ensembl	17,934	23,392
Global	21,492	31,418
Local	24,928	29,776
Global + Local	22,587	34,800
Cufflinks	31,073	38,307
Global + Local + Cufflinks	25,044	45,793
Global + Local + Cufflinks (w/Ensembl)	24,915	60,976

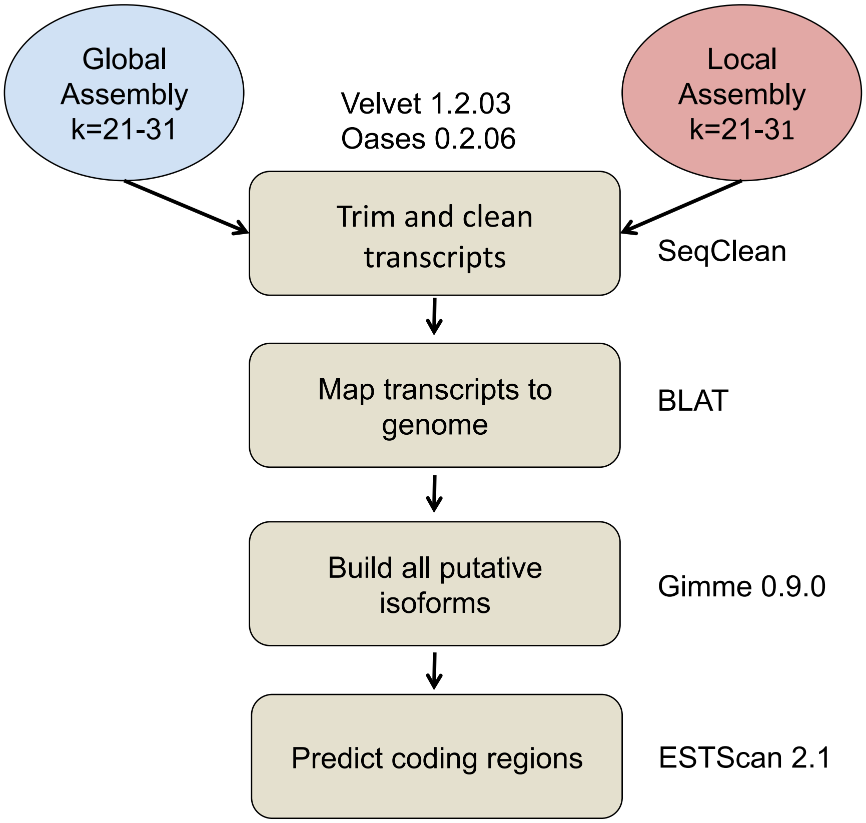

Figure 1. Gene model construction pipeline. Transcripts are obtained from two assembly methods – global and local assembly. Transcripts are aligned to the chicken genome by BLAT. Gimme then constructs gene models based on alignments of transcripts.

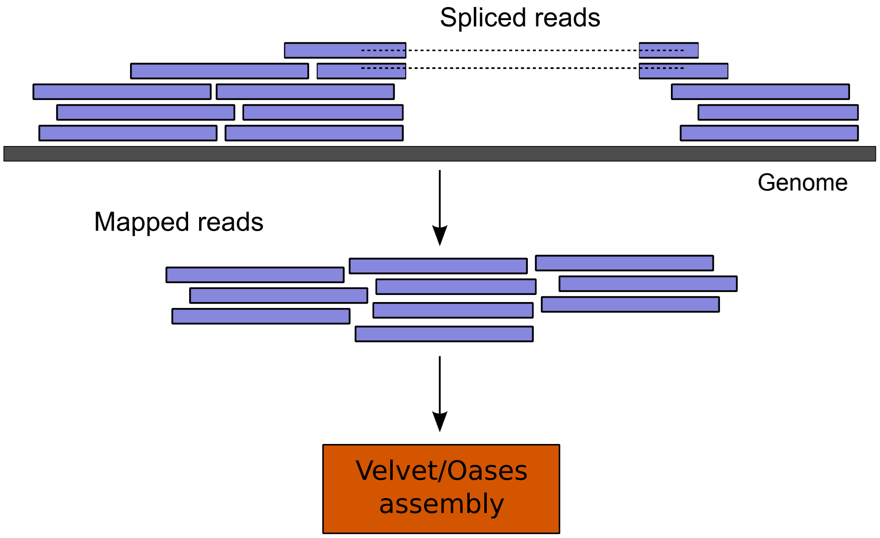

Figure 2. Local Assembly Pipeline. Reads are first mapped to a chicken genome. Then only mapped reads are assembled by Velvet and Oases.

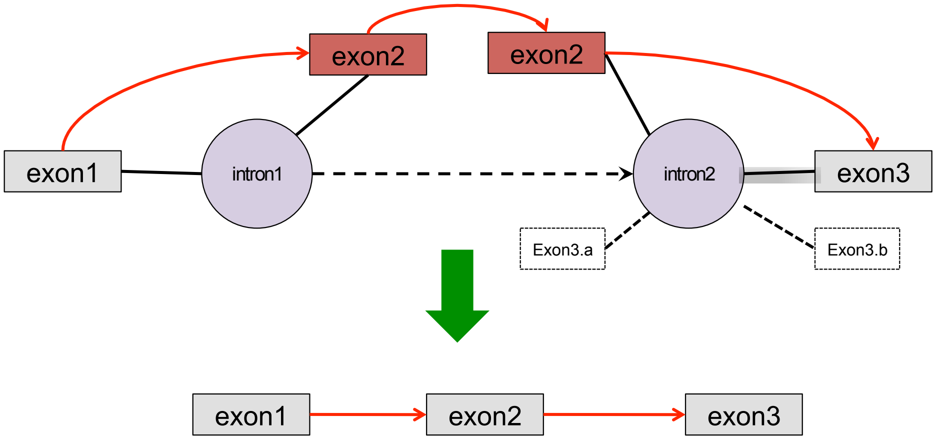

Figure 3. Intron and exon graphs. Each intron connects to exons whose splice junctions match it boundary. Some exons are excluded from the final gene model if they are incomplete (exon 3a,b). Introns sharing at least one exon are grouped together. Then an exon graph is made using exons as nodes.

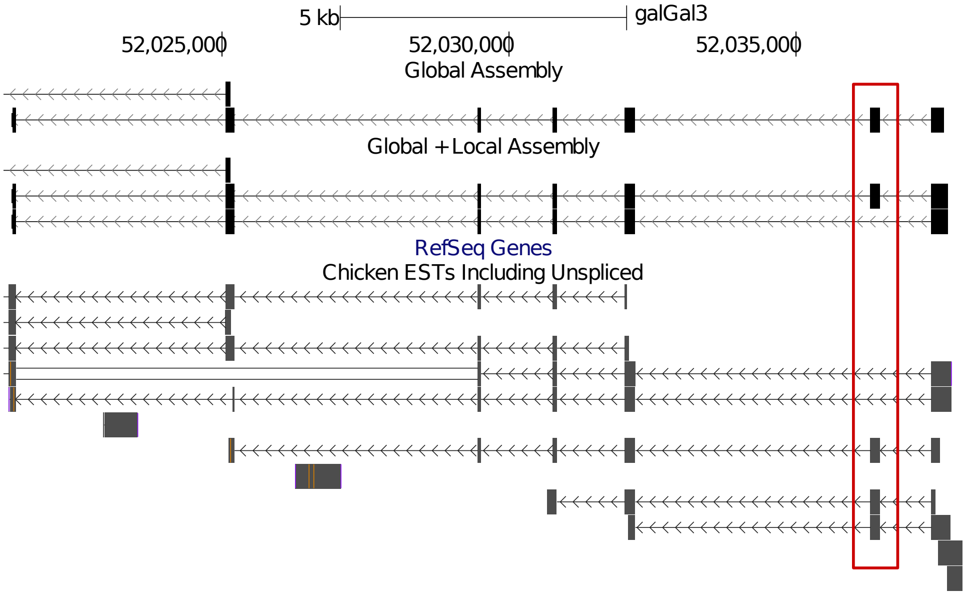

Figure 4. Global and local assembly detect different isoforms with the same k-mers.

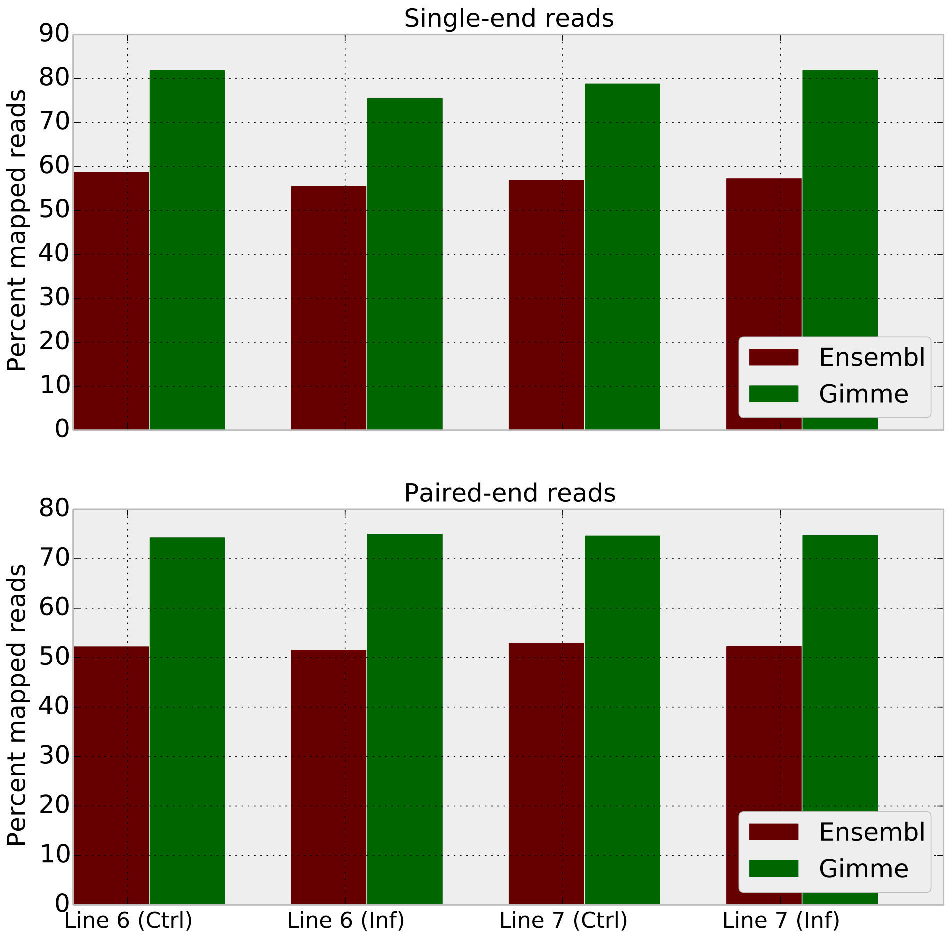

Figure 5. Cumulative counts of splice junctions with spliced reads from both single- and paired-end data from Cufflinks and Gimme models.

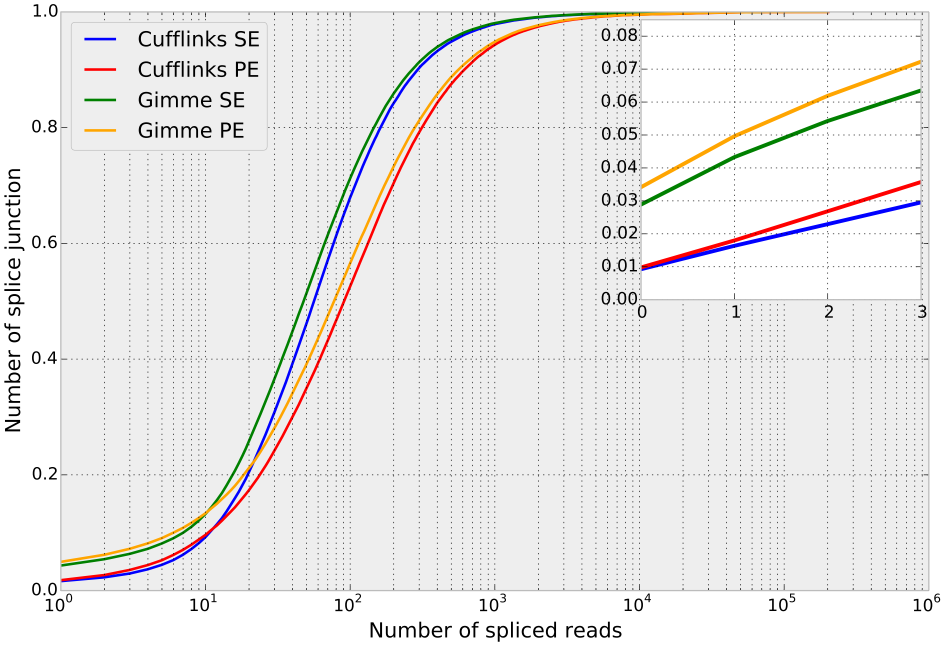

Figure 6. Cumulative counts of splice junctions with spliced reads from single- and paired-end data.

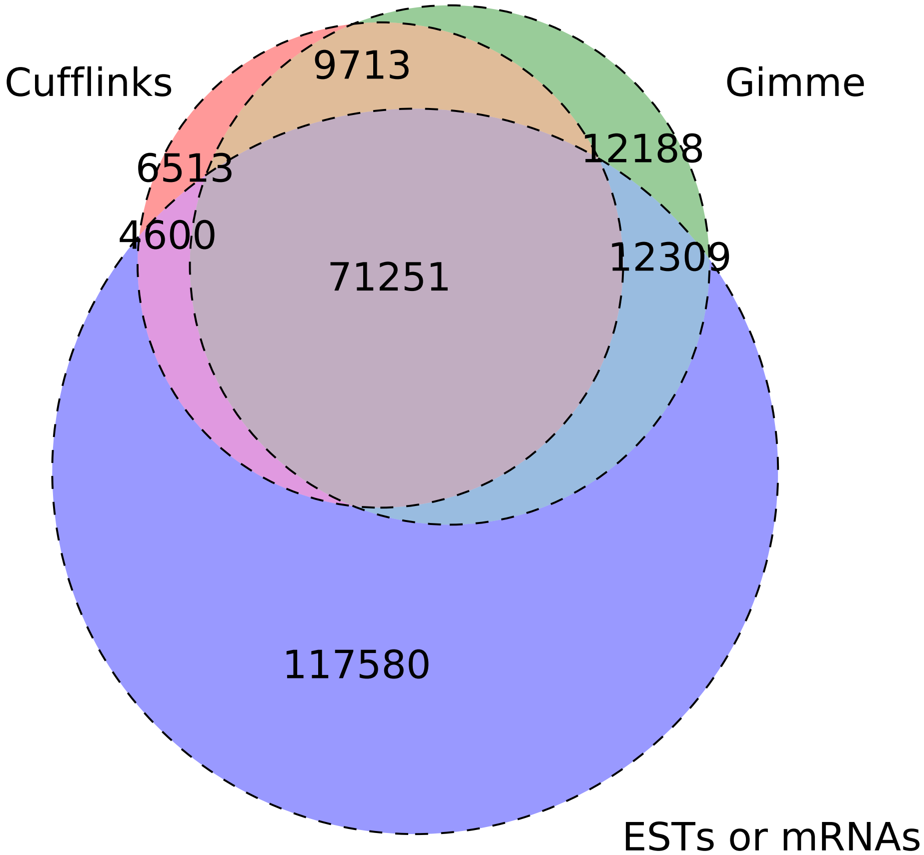

Figure 7. Splice junctions supported by ESTs or mRNAs.

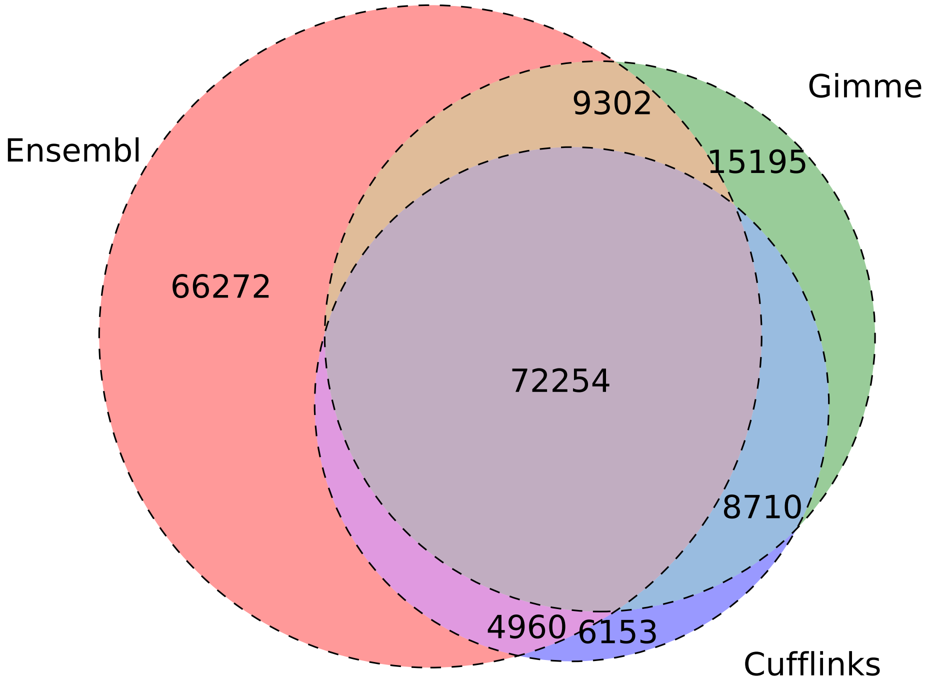

Figure 8. Splice sites in chicken Ensembl gene models detected by Cufflinks and the *de novo* assembly pipeline. Cufflinks detects many annotated isoforms that are not detected by the pipeline. The figure also shows that both methods detect a large number of unannotated splice junctions.

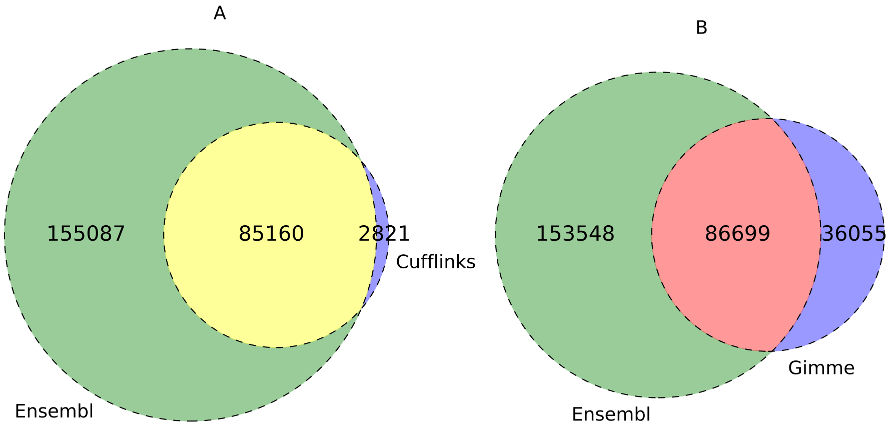

Figure 9. Splice sites in mouse Ensembl gene models, Cufflinks and the *de novo* assembly pipeline. In contrast to chicken datasets, 85,160 (96.8%) of splice junctions found by Cufflinks are annotated in mouse ENSEMBL models. The higher percentage of ENSEMBL splice junctions found by Cufflinks may be a result of more complete ENSEMBL gene models. While Gimme can detect the similar number of ENSEMBL splice junctions, it detects many more non-ENSEMBL splice junctions in mouse.

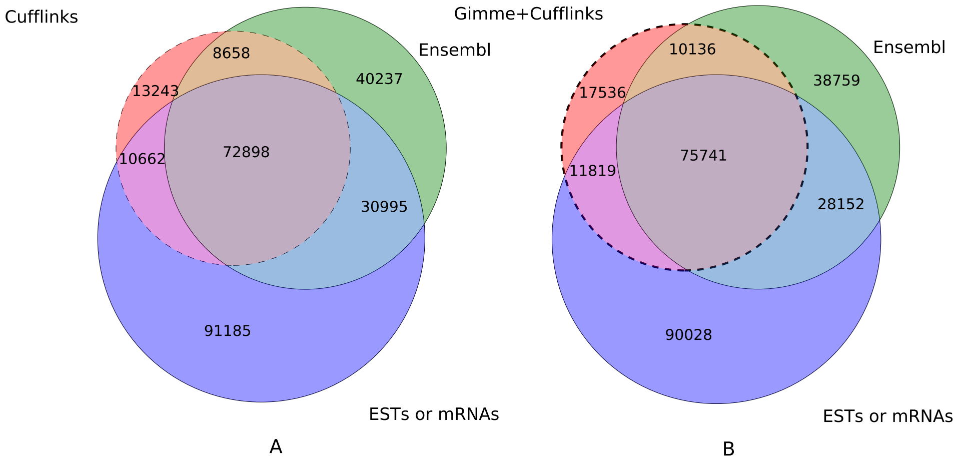

Figure 10. Splice junctions found in merged models. Merged models (B) find more splice junctions in ENSEMBL and ESTs + mRNAs than Cufflinks models only (A).

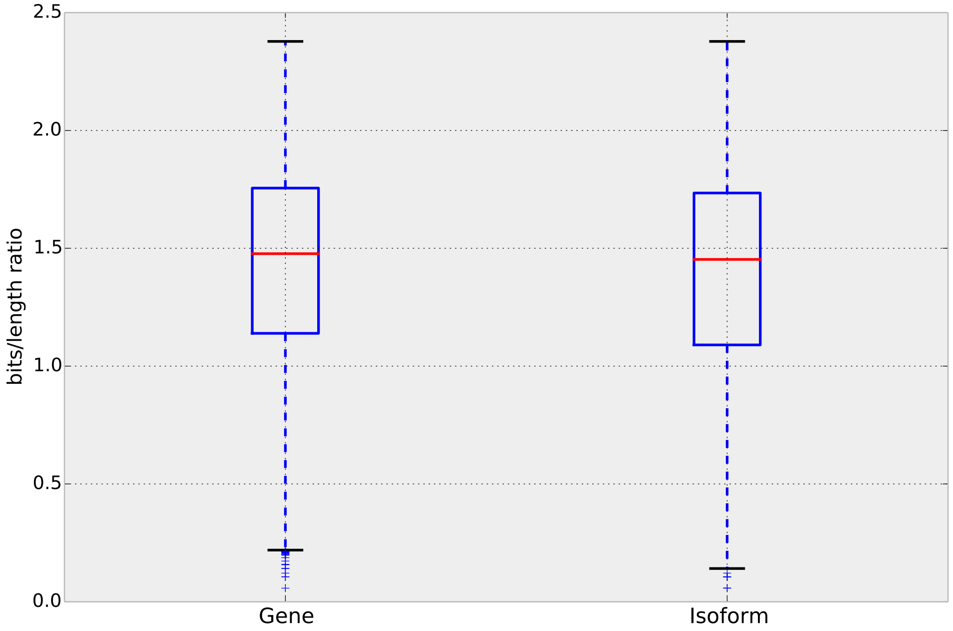

Figure 11. Box plots of bit score/length ratio of isoforms and genes that match mouse proteins. Only the greatest ratio of isoforms from the same gene is plotted for the gene plot. Ratios from every isoform are plotted for the isoform plot.