IDENTIFICATION OF CANDIDATE GENES AND ISOFORMS ASSOCIATED WITH GENETIC RESISTANCE TO MAREK'S DISEASE FROM RNA-SEQ DATA

By

Likit Preeyanon

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics – Doctor of Philosophy

2014

ABSTRACT

IDENTIFICATION OF CANDIDATE GENES AND ISOFORMS ASSOCIATED WITH GENETIC RESISTANCE TO MAREK'S DISEASE FROM RNA-SEQ DATA

By

Likit Preeyanon

Marek's disease (MD) in chickens is characterized by T cell lymphomas caused by the Marek¿s disease virus (MDV), an α-herpesvirus. MD is a major economical problem for the poultry industry as it causes approximately $2 billion in worldwide losses annually. Although vaccination has been effective at preventing tumor formation, it has not been able to prevent MDV infection or replication. Consequently, more virulent field strains have emerged over the past decades following the introduction of new vaccines. Poor practices of vaccination and incomplete immunity have been speculated to play a role in driving the evolution of the virus with greater virulence. Therefore, it is critically important to develop more sustainable control measures to the disease in the long run.

Development of genetically resistant chickens has been an alternative approach to control the virus and a number of studies have been conducted to identify specific genes that contribute to MD resistance. The major histocompatibility (MHC) locus has been found to be strongly associated with resistance or susceptibility to MD, and several alleles have been well characterized. Non-MHC genes also play a major role in resistance to MD. Two inbred lines (line 6 and line 7) maintained at Avian and Oncology Laboratory share the same MHC allele (B^2), yet line 6 is resistant and line 7 is susceptible to MD, respectively. These two lines have been used as a model to study non-MHC genes that contribute to resistance and susceptibility to the disease.

To identify non-MHC genes contributing resistance to MD, a computational pipeline was developed to integrate gene models from Ensembl, *de novo* assembly, and reference-based assembly (Cufflinks) of sequencing reads to construct a more complete set of gene models that include more complete untranslated regions (UTRs) and isoforms predicted from RNA-Seq data. The results

from expression analysis suggest that the immune response in line 7 is more active at the early stage of infection (4 days post-infection) compared to line 6. Differentially expressed genes are enriched in pathways involved in both the innate and the adaptive immune response in line 7, whereas, only genes involved in the innate immune response are significantly enriched in line 6. Due to the cell-associated nature of MDV and the current model of MDV infection, the virus is thought to transfer from B cells and antigen presenting cells (APCs) to activated T cells during the lytic infection. Therefore, repressed or delayed activation of the adaptive immune response in line 6 may be a key mechanism conferring MD resistance.

Investigation of differential exon usage suggests that genes involved in the cytoskeleton pathway may play a role in repressing the activation of the adaptive immune response. For instance, the *ITGB2* gene encodes integrin $\beta 2$, a component of several molecules including the lymphocyte function-associated antigen 1 (LFA-1). LFA-1 is exclusively expressed on the surface of leukocytes and plays an important role in cell-to-cell contact and antigen presentation. It could be speculated that an alternative isoform of *ITGB2* affects a function of LFA-1 and prevents T cells from being activated by APCs or B cells resulting in the delayed activation of the adaptive immune response or the lower number of activated T cells, the target of MDV.

The results from this study show that many genes not identified as differentially expressed at a gene level are differentially expressed at an isoform level; therefore, they will not be identified by gene expression analysis alone. Using the pipeline developed in this study, one can iteratively incorporate ENSEMBL models and RNA-seq data to construct better gene models that include genes and isoforms expressed in all samples and perform differential gene and isoform expression analysis to identify genes and isoforms that are responsible for resistance to MD.

Although functions of most isoforms are not fully annotated, we have shown that methods, such as protein prediction and pathway analysis, can be used to predict the putative functions of the isoforms and their potential roles in MD resistance, which could open up a new direction for MD research. Moreover, prediction of causative *cis*-regulatory elements in those genes will lead to identification of precise genetic factors contributing to MD resistance.

To scientists who encourage open science.

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my mentor, Dr. C. Titus Brown, for your guidance on how to be a good scientist and how to lead an awesome life. You have positively influenced me both inside and outside academia and your advice on both research and career have been invaluable. I would also like to thank my current and emeritus committees, Dr. Hans H. Cheng, Dr. Wenjiang Fu, Dr. Jerry Dodgson, Dr. Kefei Yu, Dr. Ian York, and Dr. Walter J. Esselman for serving as my guidance committee even at hardship. I also thank for your brilliant comments and suggestions. I would also like to thank Institute for Cyber-Enabled Research (iCER) at MSU for a superb service that enabled my research work. I also want to thank all members of Genomics, Evolution, and Development (GED) lab and Dr. Cheng's lab at Avian Disease and Oncology Laboratory (ADOL), and all my collaborators for providing a friendly working environment, friendship and all research support.

Last but not least, I would like to thank the Royal Thai government for a five-year financial support and staff at The Office Of Educational Affairs (OEADC) for all help and support during my graduate study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Marek's disease (MD) is a lymphoproliferative disease of chickens caused by a highly oncogenic Marek's disease virus (MDV). Vaccination has been effective in reducing tumor formation and loss from MD; however, it does not confer protective immunity against infection and shedding of MDV [13]. Continuous shedding of viruses from feather follicles of vaccinated chickens allows field viruses to circulate in the flock [41]. As a consequence, more virulent strains that overcome vaccinal protection have emerged [78, 13] leading to outbreaks that have cost an estimate of $2 billion [41]. Selective breeding of resistant birds is an alternative control measure against MD. In this study, we use transcriptome data from next-generation sequencing to compare gene and isoform expression in response to MDV infection between MD resistant (line 6) and susceptible (line 7) chickens in order to identify candidate genes and isoforms that contribute to resistance to MD.

1.2 Background

1.2.1 Genetic Resistance to MD

Genetic resistance to MD can be categorized into *MHC* and *non-MHC* associated forms. The *MHC* locus (B system haplotypes) is strongly associated with MD resistance and the phenotypes of different haplotypes have been well characterized. For example, B^{21}, B^{2} and B^{6} alleles are usually associated with MD resistance, on the other hand, B^{5}, B^{13} and B^{19} are associated with susceptibility. Chickens with homozygous B^{21} were most resistant to MD (0%), whereas, chicken heterozygous with other alleles and B^{21} developed 40 – 93% MD [5]. However, in some populations, the association of B haplotypes does not follow the aforementioned patterns. For example, the Cornell chicken strains C and K were selected for resistance to MD, yet strain C possesses B^6, B^{13} and B^{15} and strain K possesses B^{15} [1, 5]. Similarly, inbred chicken lines 6 and 7 selected for resistance and susceptible to MD are both $B^{2,2}$ homozygous. This suggests that genes outside the *MHC* locus also contribute to MD resistance.

It has been postulated that different mechanisms are controlled by *MHC* and *non-MHC* loci [20]. For instance, Kaiser et al. [20] reported that the level of MDV viral load also differed between lines 6 and N after 10 and 14 days post infection (dpi) – the onset of latency – and decreased to the same level by 21 dpi. Although both lines 6 and N are resistant to MD, the level of MDV viral load in line 6 was significantly higher than in line N at 10 and 14 dpi ($P< 0.05$). Resistance line N is associated with the B^{21} haplotype, whereas, the resistance of line 6 is associated with the MDV1 locus.

1.2.2 Genome-wide scan for genes conferring resistance to MD

Identification of precise genes is essential for developing resistant breeds using a marker-assisted selection (MAS) . A genetic-driven method has been used to identify quantitative trait loci (QTL) associated with MD resistance or MD infection using lines 6 and 7 as parents. Seven significant and seven suggestive QTLs were identified by Vallejo [69] and Yonash [79]; whereas, Bumstead [6] identified a significant locus on chromosome 1, which is referred to as MDV1. MDV1 region has a strong association with resistance in terms of reduced viral level and tumour incidence. Based on comparative mapping analysis, this locus has shared synteny with the mouse region that includes the *CMV1* and *Ly49* genes. *CMV1* controls resistance to murine cytomegalovirus by affecting viral proliferaton and *Ly49* serves as an inhibitor of cell killing by NK cells. Correlation between functions of *CMV1* and *Ly49* and resistance associated to this region is notable.

The drawback of QTL analysis of complex traits is it requires a large population size (> 500) to achieve reliable results. Ideally, a population size of > 1000 is needed for fine-mapping suitable for successful MAS (< 1 cM) [81]. Furthermore, only a few true QTL of major effect can be detected in any given study [10]. Therefore, an alternative approach that could facilitate the identification of precise genes at a genome-wide scale is warranted.

Differential gene expression is thought to contribute to phenotypic variation and difference in gene expression can be controlled by genetic factors [40]. Furthermore, functional analysis of differential-expressed (DE) genes can provide an insight into pathways and biological mechanisms that control phenotypes. Data from RNA-Seq can also be used to investigate mutations such as SNPs and indels that might be responsible for phenotypic differences.

Recently, high-throughput technologies including microarray and next-generation sequencing (NGS) have been used to compare gene or EST expression between MDV-infected birds of selected resistant and susceptible lines to identify candidate genes that contribute to MD resistance. Sarson [54] used a low-density immune-specific cDNA microarray to compare expression of immune genes from splenocytes between resistant and susceptible chickens, with B^{21} and the B^{19} haplotype respectively, and found that the percentage of differential-expressed genes between B^{19} control and infected birds was highest. Furthermore, B cell surface markers such as *Bu-1* and *IgM* levels were downregulated, which might contribute to the decrease in B cells in the lytic phase of infection. Morgan et al [39] also used a DNA microarray designed for chicken to compare the differential expression between lines 6 and 7 at multiple time points. Of $\sim$ 1200 genes and ESTs, a few genes including growth hormone (*GH1*) and lymphotactin (*SCYC1*) were found to be differentially expressed ($>$ 2 fold) and located near the QTL region on chromosome 1 (MDV1) identified by Bumstead [6]. Growth hormone binds to MDV *SORF2* and is associated with MD resistance [34] and lymphotactin serves as a chemoattractant of CD4+ and CD8+ T cells.

The findings from these studies are interesting, yet very limited due to the small scale of the analysis. A genome-wide scale microarray study was, therefore, conducted by Smith et al [61] to compare gene expression between virus-infected lines 6 and 7 from spleen and thymus at a larger scale. In control groups, 395 genes in spleen and 177 genes in thymus were differentially expressed between lines 6 and 7. Genes highly expressed in line 6 and involved in innate immune responses include *DNAJC3, DDT, NMU, GSTO1, VIP, HPS5, MMP7, FGFR3, HSCB, E2F4, SFTPA2* and

GNG12. In infected groups, 593 genes in spleen and 156 genes in thymus were differentially expressed including *IgG-H, AMIGO2, MMP13*, and *CLEC3B* that were highly expressed in line 6 and *AVD, IRG1, HSP25, ART1, IL-18, NOS2A, CXCL13, CCLi2, MX1, SOCS1*, and *IL-6* that were highly expressed in line 7. Approximately $26-30\%$ of DE genes were located in the previously reported QTL regions, which could be potential candidate genes.

To conclude, high-throughput technologies together with genetic approaches have been successfully used to identify many candidate genes associated with MD resistance, which confirms the complex nature of MD genetic resistance. However, the interaction of these candidate genes and their roles in disease resistance have not been fully defined. Furthermore, only global gene expression has been used for identification of genes conferring MD resistance, whereas, alternative isoforms have been shown to play a significant role in immune responses and contribute to disease susceptibility [38, 73]. This dissertation aims to identify both candidate genes and isoforms that are involved in resistance to MD.

1.3 Problem Statement

In the immune system, many genes can express different isoforms with a distinctive, synergistic or even opposing function [38, 73]. Expression of isoforms is regulated in part by *cis*-regulatory sequences within an exon or intron of a pre-mRNA [4]. Although expressions of various genes have been examined across stages of infection and various genetic background [61, 54, 20, 39], the investigation of genome-wide isoform expression has not been conducted. The gap in the knowledge is, therefore, the identification of candidate isoforms that contribute to MD resistance.

Many tools are currently available for estimating gene and isoform expression from RNA-Seq data as well as comparing their expressions between samples (reviewed in [67]). However, they are all reliant on completeness of provided gene models. Available gene models such as Ensembl annotation do not include all genes and isoforms; therefore, they are not ideal for identification of candidate genes and isoforms. In addition, some annotated gene models do not included unstranslated regions (UTRs) that can have important biological functions as well as significant sizes. Methods for expression estimation typically infer gene expression from the number of reads mapped to transcripts or gene models. Without complete gene models, the expression will be inaccurately estimated resulting in errors in differential expression (DE) prediction. Consequently, biological pathway prediction will be affected because the prediction is based solely on results from DE prediction.

1.4 Significance of Research

The study will create a pipeline to integrate RNA-Seq data with existing gene models to extend the models so that they include genes, isoforms, and UTRs expressed in a sample. The extended genes models will allow a better estimate of genes and isoforms expression, which will be used to predict biological pathways perturbed in responses to MDV infection. Comparison of perturbed biological pathways between resistant and susceptible birds will provide an insight into genes and mechanisms that contribute to MD resistance. Additionally, in depth investigation of isoform expression could lead to identification of exonic SNPs that regulate alternative splicing patterns, which could reveal unprecedented level of molecular and genetic mechanisms that impart resistance to MD.

1.5 Outline of Dissertation

In the first chapter, we describe a method developed to construct gene models from different sources including *de novo* assembly of short reads from RNA-Seq data, a reference-guided assembly (Cufflinks) and Ensembl gene models. We show that our method can be used to combine gene models from different sources to build gene models that include more splice variants. We also describe a local assembly method that can enhance sensitivity of splice variant detection. In the second chapter, we compare results of gene expression and KEGG pathway analysis from different gene models. We demonstrate that different gene models give different results from pathway enrichment analysis. In addition, we discuss the use of combined annotation from chicken and mouse to increase sensitivity of pathway prediction. In the last chapter, we report differentially expressed genes and isoforms and discuss a potential role of differentially expressed isoforms in MD resistance.

CHAPTER 2

RNA-SEQ ASSEMBLY DISCOVERS MANY SPLICE VARIANTS

2.1 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 [72]. 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. 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) [75]. 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* [47] 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 [47].

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 [68] relies on splice junctions detected from Tophat [66], 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 a 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 [15].

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 [14] 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 [53] and Oases [57] are extensions of the Abyss [59] and Velvet [82, 83] 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 [57].

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

Table 2.1: **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

showed that transcripts reconstructed from *de novo* assembly and mapping-based approaches can be merged to build more complete gene models.

2.2 Results

2.2.1 Local assembly enhances isoform detection

We used the Velvet [82] and Oases [57] 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 [57]. We chose k-mers between 21 and 31, and recovered between 90,700 and 104,000 unique sequences from each data set (see Table 2.1, 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 2.2, and Table 2.3); this is probably due to the incomplete nature of the current chicken genome assembly, which is lacking approximately 5% of its true gene content. 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

Table 2.2: **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 2.3: **Unique regions from global and local assembly**

Dataset	Unique Region		Matched with mouse proteins	
	Global	Local	Global	Local
Line 6 uninfected	1285929	96830	260321 (20.0%)	9413 (9.7%)
Line 6 infected	1631356	59813	312849 (19.2%)	5132 (8.6%)
Line 7 uninfected	1800634	104229	349346 (19.4%)	9883 (9.5%)
Line 7 infected	1611354	125640	296915 (18.4%)	9381 (7.5%)

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 2.1 shows an example of different isoforms detected by the two assembly methods.

2.2.2 Oases-M discards splice variants

The above approaches recovered transcript fragments, but not entire genes. 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 [57]. 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 indepen-

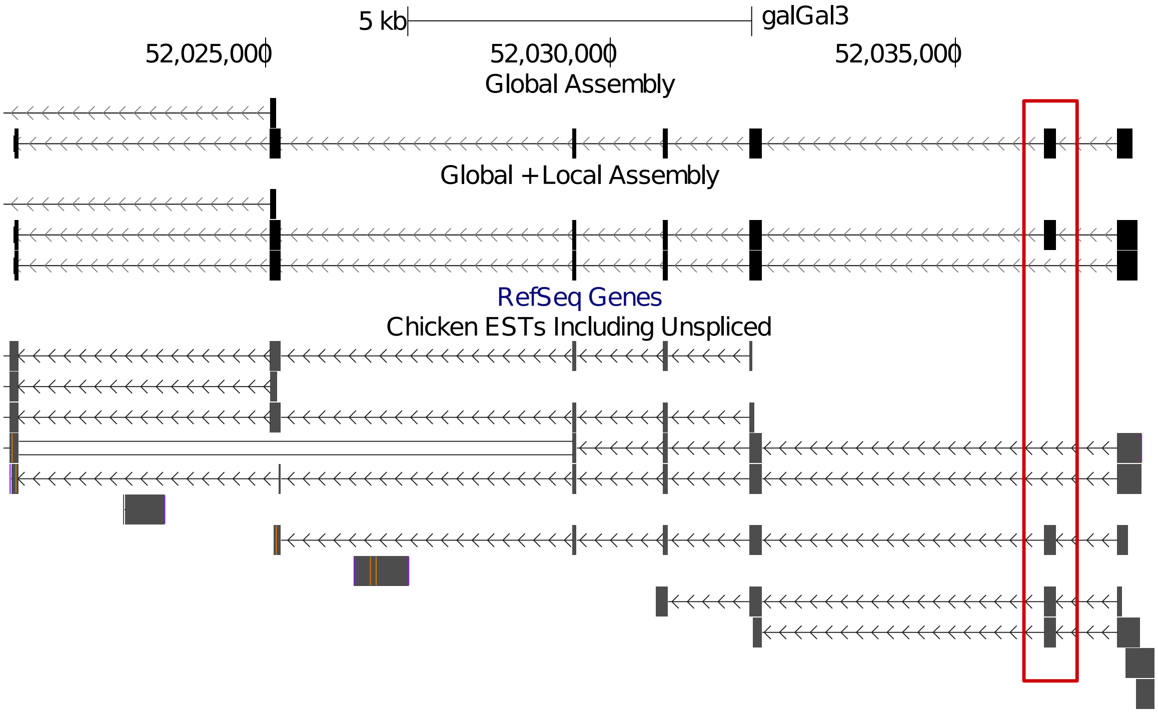

Figure 2.1: **Global and local assembly detect different isoforms with the same k-mers.** For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Table 2.4: **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%)

dently supported by ESTs, while 1,608 (19.4%) of the unmerged splice junctions are supported by ESTs (Table 2.4). 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.

2.2.3 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 2.2). 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 2.5).

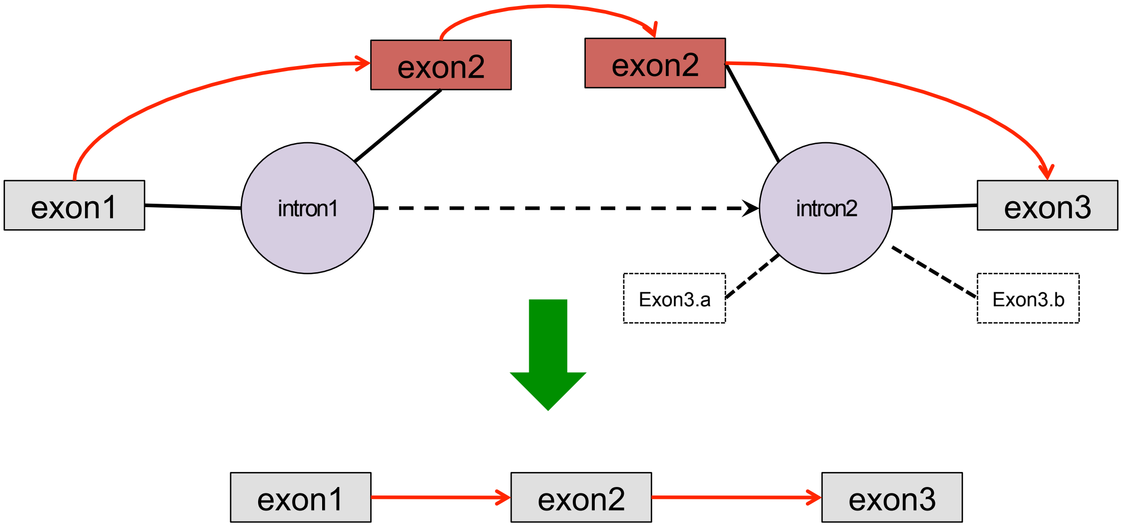

Figure 2.2: **Intron and exon graphs.** Each intron connects to exons whose splice junctions match it boundary. Some exons are excluded from the final gene model if they are incomplete (exon 3a,b). Introns sharing at least one exon are grouped together. Then an exon graph is made using exons as nodes.

2.2.4 Validation of gene models

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

Table 2.5: **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	25044	45793
Global + Local + Cufflinks (w/Ensembl)	24915	60976

2.2.4.1 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 2.3).

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. Thus the merged gene models produced by Gimme represent the significant majority of the assemblable data.

2.2.4.2 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 [26]. 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 (Figure 2.4). Note that 710 junc-

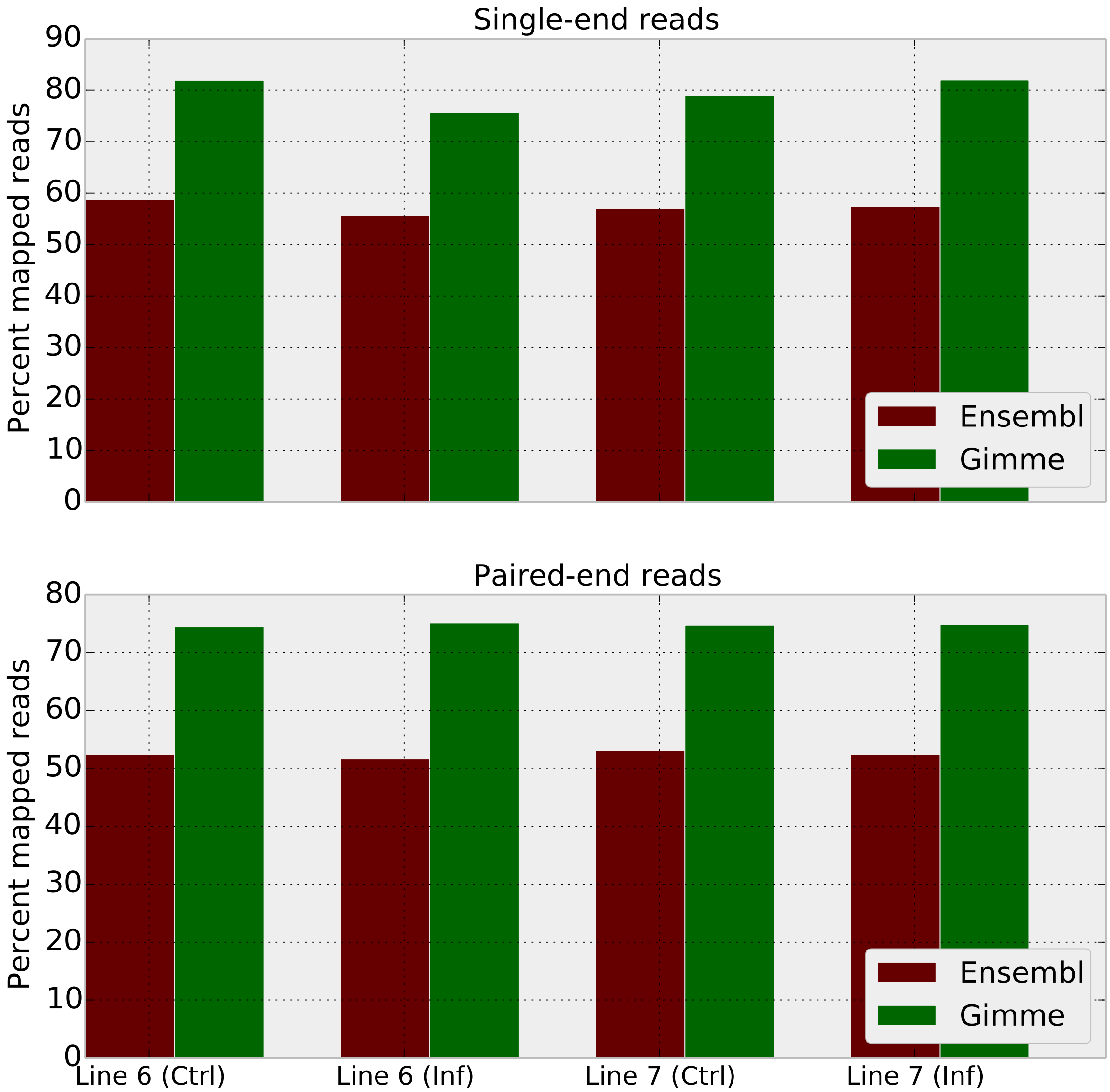

Figure 2.3: **Percentages of reads mapped to the gene models from Ensembl and Gimme.** Reads were mapped to transcripts derived from the gene models using Bowtie.

tions 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.

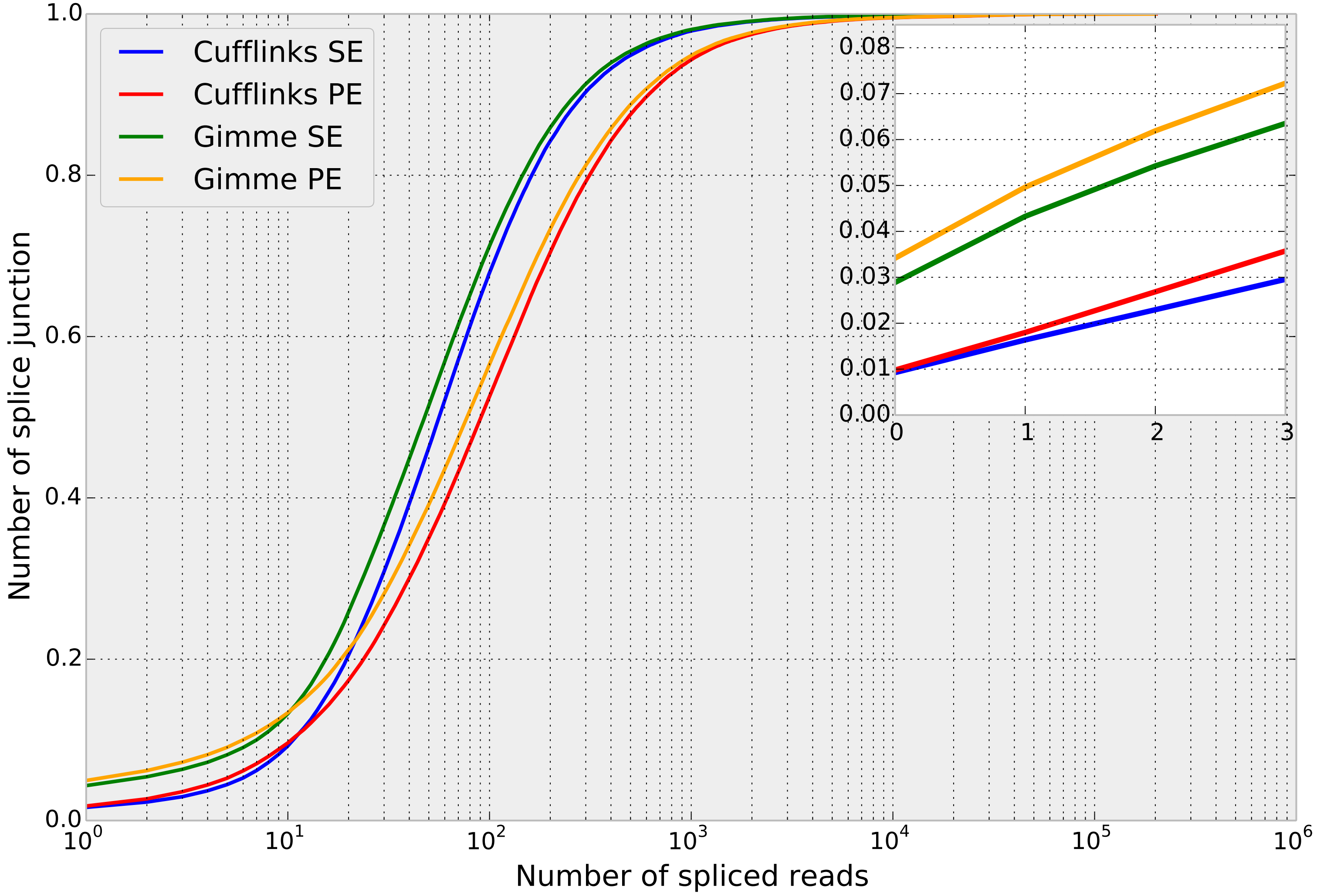

Figure 2.4: **Cumulative counts of splice junctions with spliced reads from both single- and paired-end data.** Reads were mapped to transcripts derived from the gene models using Bowtie.

2.2.4.3 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 (Figure 2.5). 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 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.

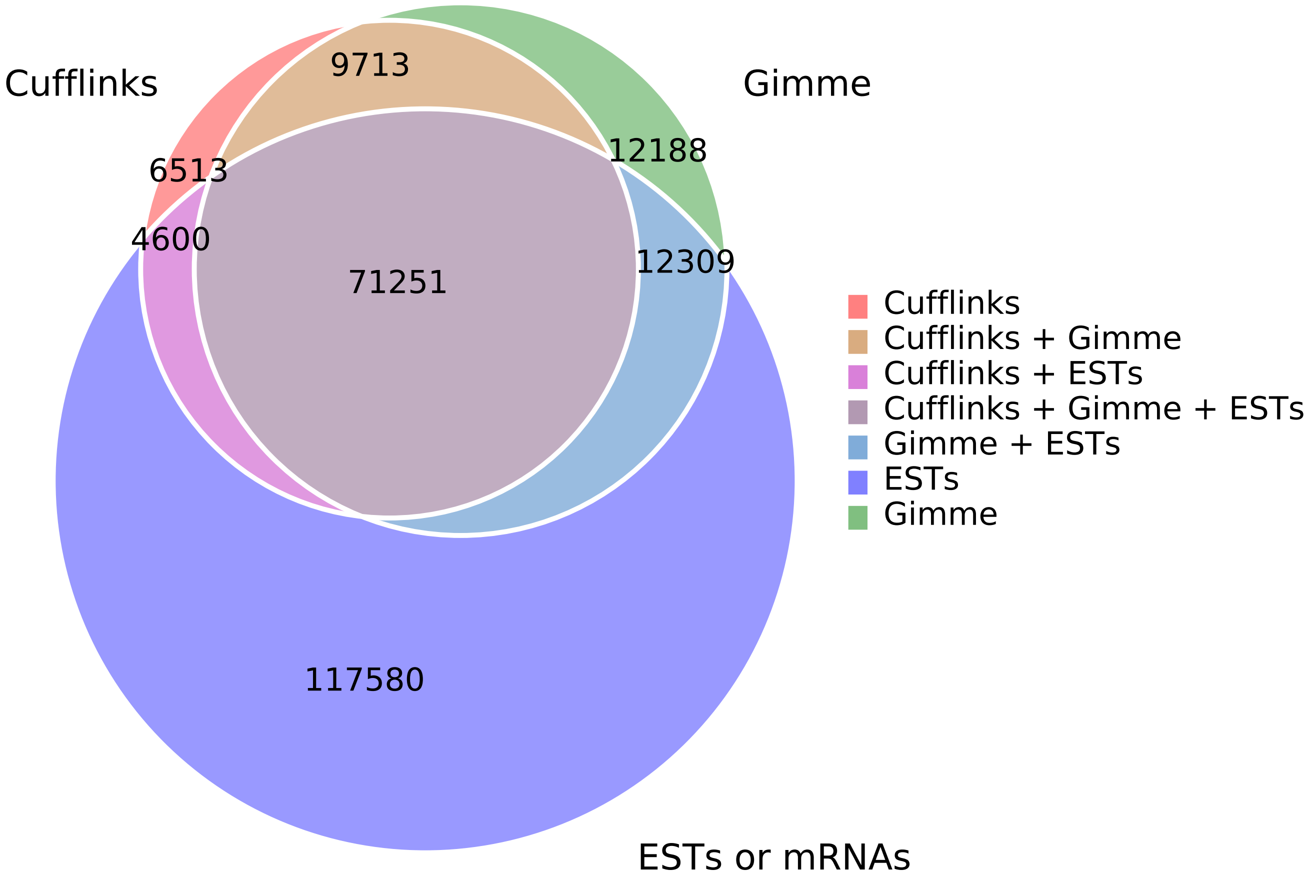

Figure 2.5: **Splice junctions supported by ESTs or mRNAs.** Both Cufflinks and Gimme detect unique splice junctions supported by ESTs + mRNAs. A large number of splice junctions in ESTs + mRNAs are not found in either Gimme or Cufflinks because ESTs + mRNAs contain transcripts for other tissues.

2.2.4.4 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 dataciteTrapnell:2010kd. 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.

Cufflinks finds 92,077 splice junctions, and Gimme finds 105,461 splice junctions. 80,964 of them are in common (Figure 2.6). 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 2.7). However, Gimme recovers a substantial number of additional splice variants beyond Cufflinks and Ensembl both.

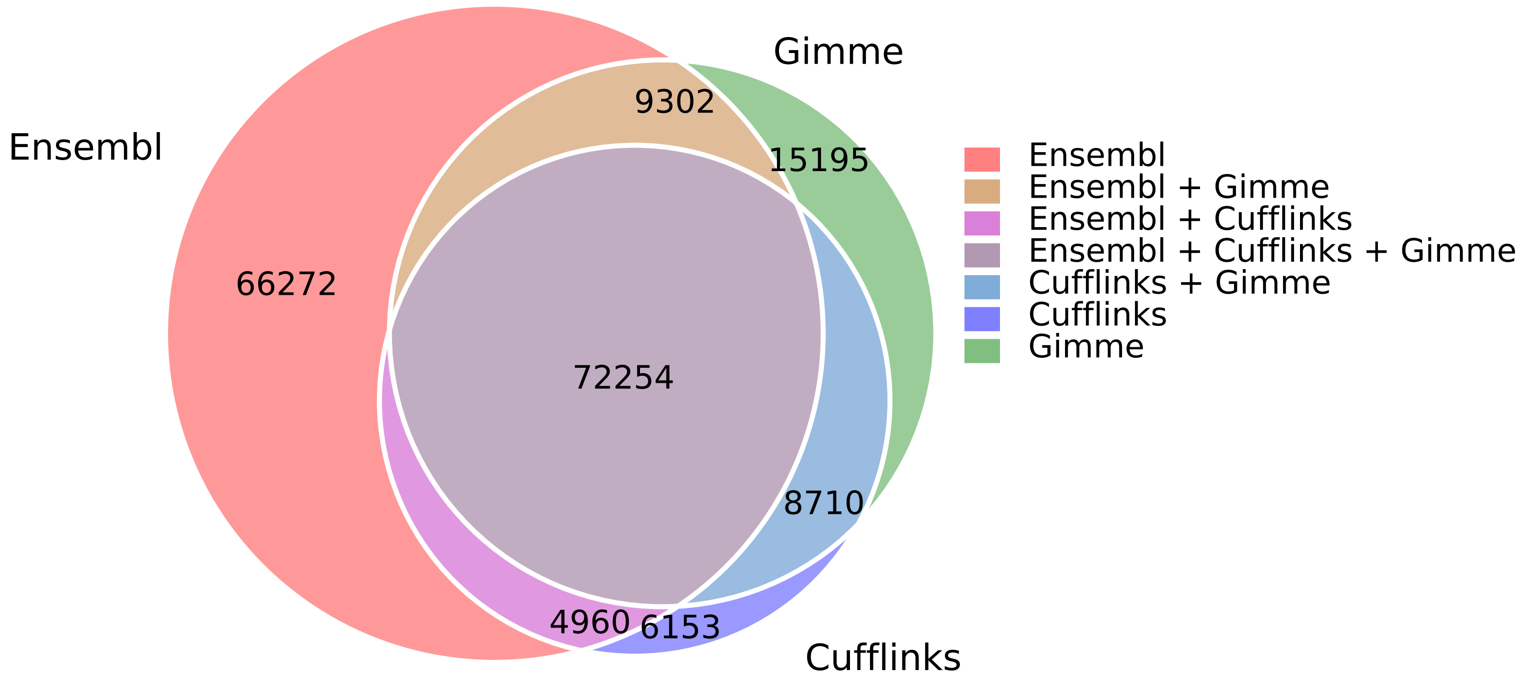

Figure 2.6: **Splice sites in chicken Ensembl gene models detected by Cufflinks and the *de novo* assembly pipeline.** Cufflinks detects many annotated isoforms that are not detected by the pipeline. The figure also shows that both methods detect a large number of unannotated splice junctions, which suggests that those junctions may be genuine.

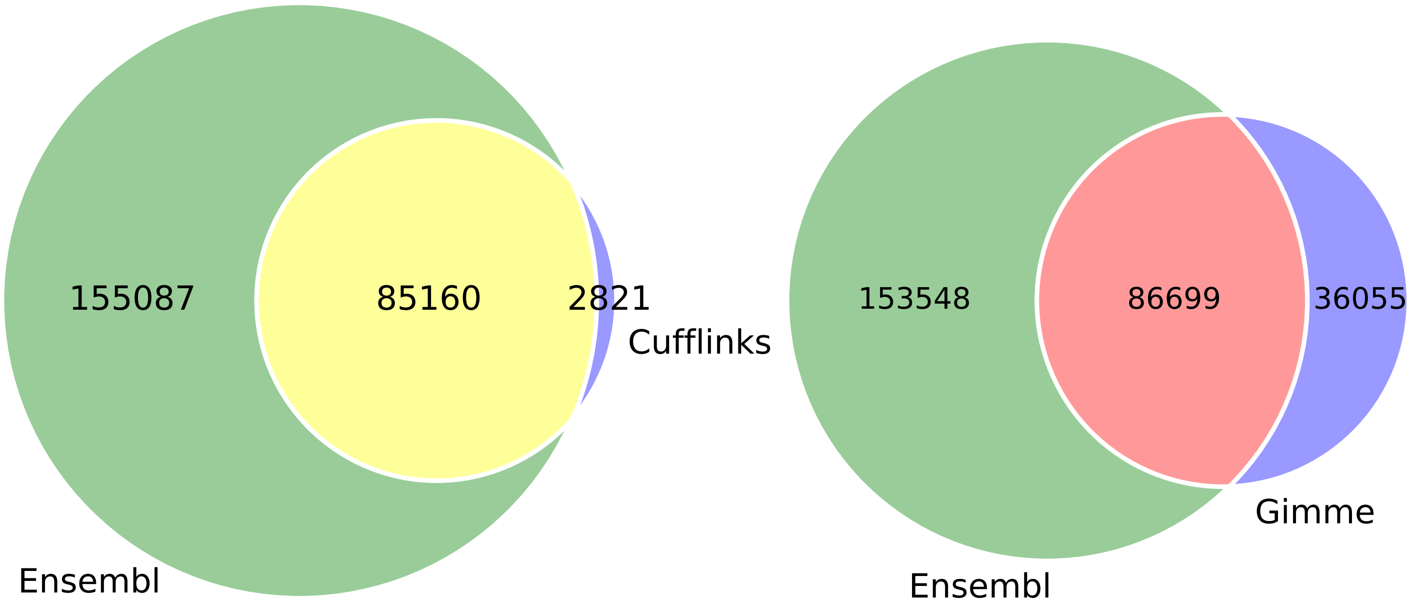

Figure 2.7: **Splice sites in mouse Ensembl gene models, Cufflinks and the *de novo* assembly pipeline.** Compared to Gimme, Cufflinks only detects a small number of splice junctions not annotated in mouse Ensembl. Cufflinks and Gimme detect a similar number of splice junctions found in mouse Ensembl.

2.2.4.5 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. 2.5. This resulted in a decreased number of total genes, suggesting that some fragmented genes were merged together to form more complete gene structures The merged gene models recover 49.3% and 56.2% of splice junctions from ESTs and Ensembl respectively (Figure 2.8), which is about 10% greater than that from corresponding unmerged gene models.

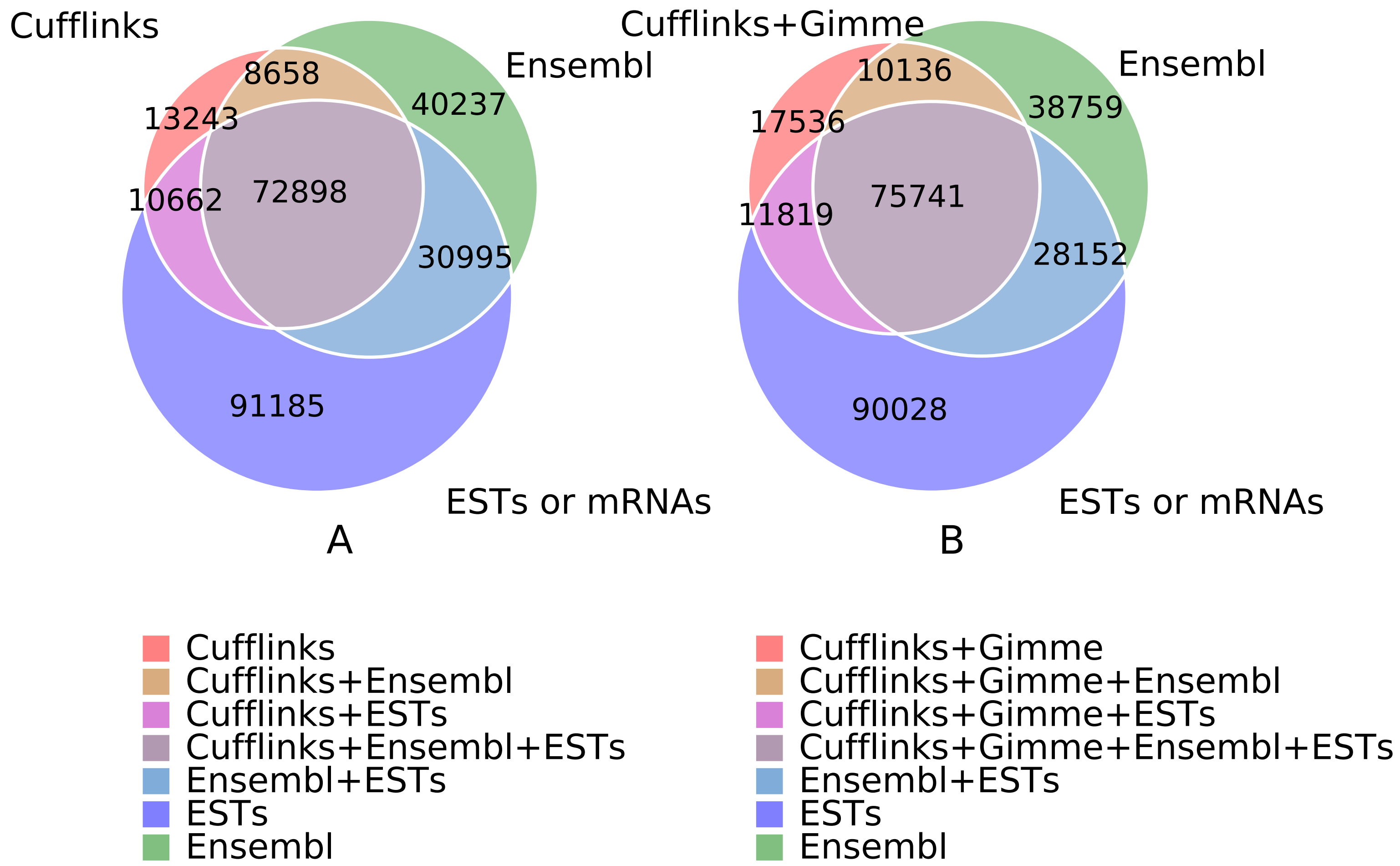

Figure 2.8: **Splice junctions found in merged models.** Merged models (B) find more splice junctions in Ensembl and ESTs + mRNAs than Cufflinks models only (A).

2.2.4.6 Validating chicken sequences by using mouse homologs

To validate our predicted isoforms, we extracted putative coding sequences from our gene models with ESTScan [18]. 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 ≥ 1.0(Figure 2.9). These matches have a bit-score/length ratio greater than 1, which indicates a good agreement between chicken and mouse proteins.

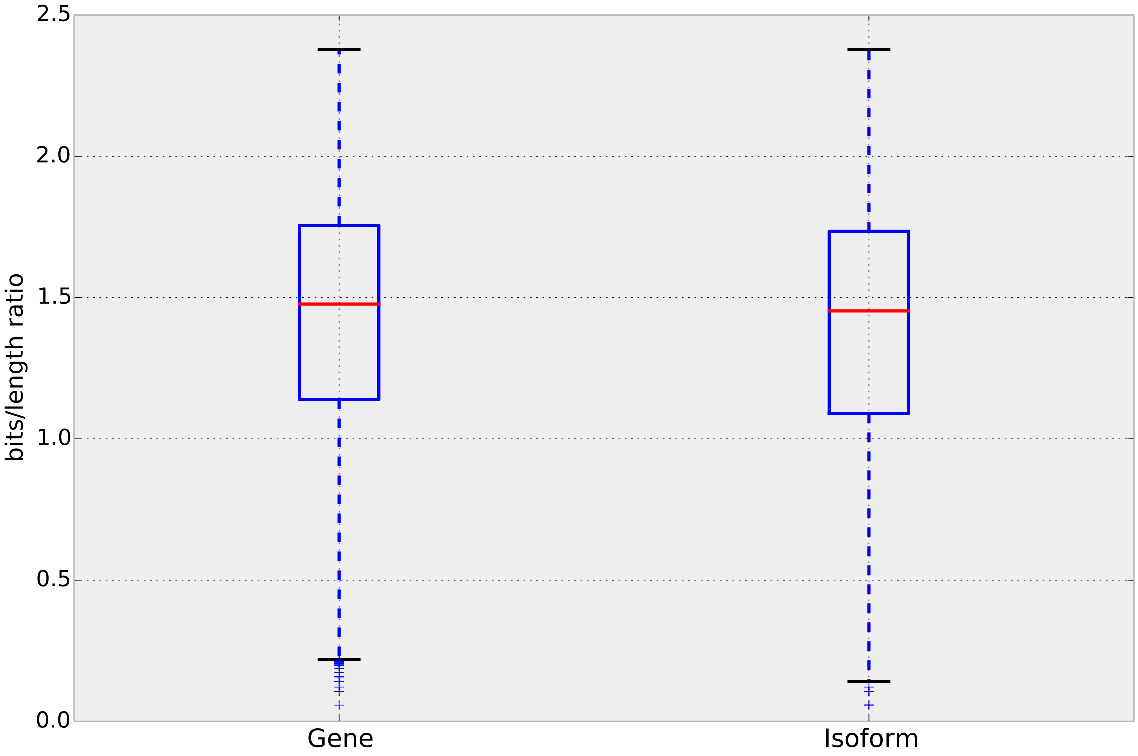

Figure 2.9: **Box plots of bit score/length ratio of isoforms and genes that match mouse proteins.**

2.3 Discussion

For organisms with a reference genome and Ensembl annotation, building gene models from RNA-Seq reads using Cufflinks with Ensembl gene models as a reference guide seems to be a preferable method for gene and isoform expression analysis. However, we have shown that Ensembl models in chicken are missing a substantial number of splice variants found in ESTs and mRNAs and Cufflinks does not recover all splice variants. A previous study by Schulz *et al.* [57] demonstrated that Cufflinks outperformed *de novo* assembly in detecting genes and isoforms in Ensembl annotations in mouse. However, we found that Cufflinks almost exclusively finds splice junctions in mouse Ensembl gene models whereas our gene models include splice junctions not in Ensembl gene models but supported by ESTs (Figure 2.7).

To build mouse Cufflinks gene models, we used Tophat default parameters claimed to be fine-tuned for mammalian RNA-Seq reads, but Cufflinks still poorly detected splice junctions not in Ensembl. This suggests that the parameters may have been set based on mammalian Ensembl models, which could limit the efficiency of finding novel splice variants in other non-mammal organisms. In contrast, *de novo* assembly does not rely on species-specific parameters. It is, therefore, recommended that a combination of Cufflinks models and transcripts from *de novo* assembly should be used to build gene models that include more splice variants.

We have also shown that the local assembly could greatly increase sensitivity of splice variant detection. However, how the local assembly affects the assembly process is not clearly understood. We speculate that reads with multiple alignments may play a significant role in enhancing the assembly of splice variants. Understanding of the mechanisms may lead to a technique that can be applied to find splice variants in organisms that lack a reference genome.

Both Cufflinks and our pipeline rely heavily on a reference genome; therefore, the quality of the genome will greatly affect 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 false splice junctions, which in turn produce a large number of splice variants as observed in some chromosomes.

2.4 Materials and Methods

2.4.1 Reads quality trimming

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.

2.4.2 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.

2.4.3 Mapping reads to the genome and gene models

Single and paired-end reads were mapped to the chicken genome by Tophat [66] release 1.3.1 using default parameters without annotations. All reads were mapped to cDNA sequences derived from gene models by Bowtie [26] 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`.

2.4.4 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 (Figure 2.10). Global and local assembly was performed using Velvet version 1.2.03 [82] 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 [57]. A poly-A tail, short transcripts and transcripts with low complexity are removed by seqclean [58] with default parameters. Redundant transcripts are removed by cd-hit-est from the CD-HIT suite [33]. 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.

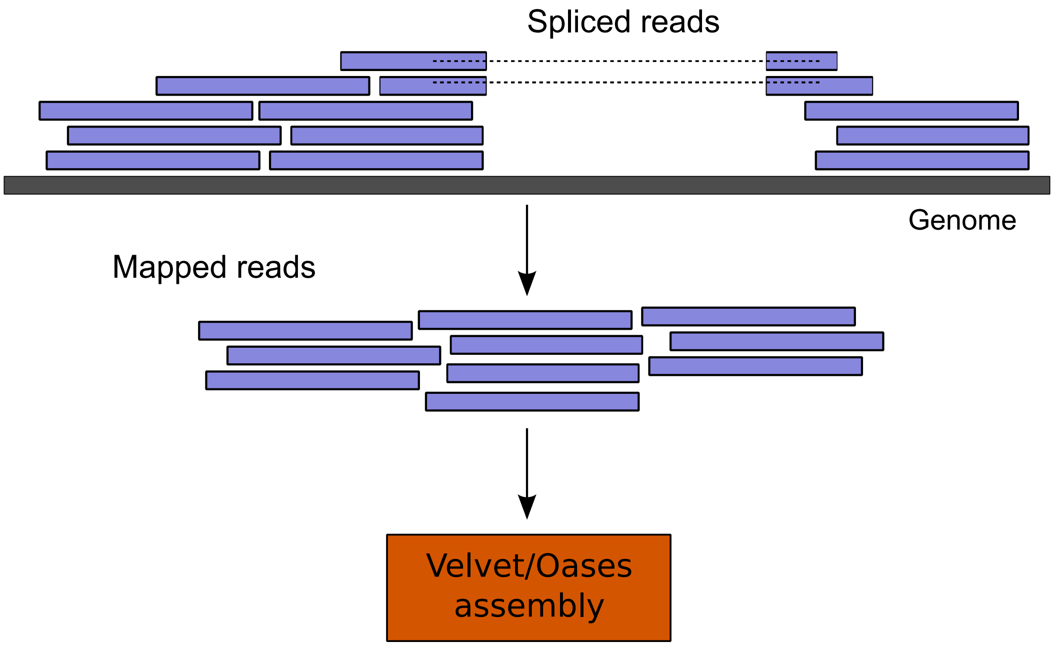

Figure 2.10: **Local Assembly Pipeline.** Reads are first mapped to a chicken genome. Then only mapped reads are assembled by Velvet and Oases. Reads mapped to each chromosome are assembled separately.

2.4.5 Gene model construction

2.4.5.1 Overall pipeline

Figure 2.11 depicts an overall gene model construction pipeline. Transcripts of all datasets from local and global assembly were mapped to the chicken genome using BLAT [24]. Alignments and gaps from BLAT outputs are considered exons and introns respectively. Optionally, data from other sources (ESTs, RefGenes, Cufflinks, Ensembl and 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). After that, transcripts that are highly similar ($> 99\%$) are clustered and removed by CD-HIT version 4.5.6 [33]. Only a representative of each cluster is kept in gene models.

2.4.5.2 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 one of whose 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 (Figure 2.2). 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`.

2.4.6 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

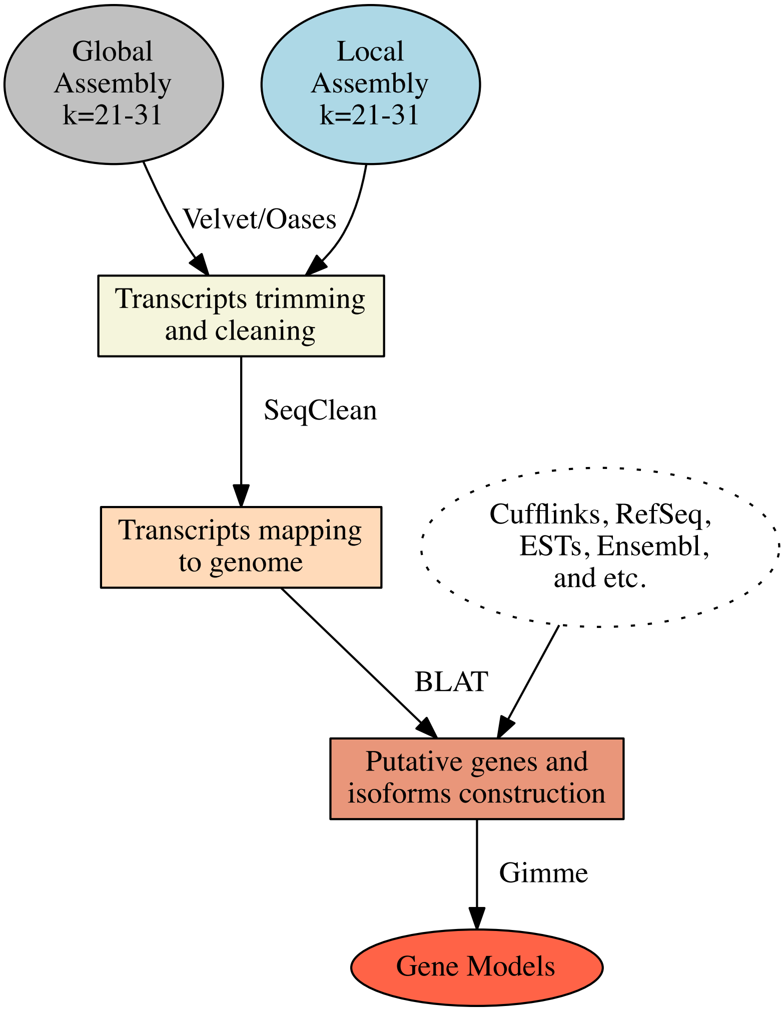
Global Assembly k=21-31
Local Assembly k=21-31
Velvet/Oases
Transcripts trimming and cleaning
SeqClean
Transcripts mapping to genome
Cufflinks, RefSeq, ESTs, Ensembl, and etc.
BLAT
Putative genes and isoforms construction
Gimme
Gene Models

Figure 2.11: **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. Gene models from Cufflinks can also be incorporated to build the gene models.

using tools from ESTScan. Only protein sequences longer than 50 bp are included in the analysis.

2.4.7 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.

2.4.8 Sequence homology analysis

Protein sequences translated from each isoform using ESTScan were searched against mouse reference proteins by BLAST 2.2.25+ [64]. 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.

2.4.9 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 [32] and Pysam [49].

2.4.10 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 [68]. All gene models are then merged together using Cuffmerge.

2.4.11 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.

2.4.12 Pipeline and Scripts

The pipeline and scripts used in this study is hosted at

```
https://github.com/likit/gimme_protocols
```

CHAPTER 3

COMPARISON OF GENE NETWORK INFERRED BY RNA-SEQ ANALYSIS

3.1 Introduction

Differential expression (DE) and pathway analysis are widely used techniques to identify candidate genes and pathways associated with phenotypes or diseases [61, 3, 31]. The advent of high-throughput technologies such as microarrays and next-generation sequencing (NGS) has led to an explosion of tools and pipelines for exploring expression data. Most of the tools for DE and pathway analyses rely on publically available gene sets such as Ensembl for quantification of gene expression and downstream pathway annotation.

Recently, RNA sequencing (RNA-Seq) by NGS technology has not only allowed biologists to study expression of annotated genes but also to discover novel genes and isoforms as well as to obtain evidence of transcribed but untranslated regions [46, 36, 43, 51]. This technology also provides an opportunity for biologists to generate more comprehensive transcriptomes for organisms by reconstructing transcript sequences from short reads using *de novo* assembly or *ab initio* methods. However, because executing these tools requires significant computational expertise, researchers may rely on existing annotation resources instead of building new gene models.

The consequences of using different gene models on gene expression analysis and pathway predictions have not been explored. We speculated that more comprehensive gene models could affect these analyses significantly, especially in organisms where the available gene annotations are relatively sparse compared to NIH model organisms. Incomplete gene models are insensitive to read mapping, which would affect predicted gene expression levels and the significance of differential expression calculations; in addition, missing splice variants and exons limit the power of differential isoform expression analysis. Missing gene models also decrease the power of GO and KEGG pathway analyses by potentially eliminating genes important for pathways.

In this article, we present comparisons of differential expression computations and KEGG pathway analyses of RNA-seq data between Ensembl gene models, *ab initio* and *de novo* constructed gene models, and merged gene models built from all of the above. We also show the effect on predictions of differentially expressed pathways when including KEGG pathway annotations from human in our analysis. We demonstrate that the gene models used and the extent of the annotation source significantly affect predictions and their statistical support. Finally, we discuss the implications for the investigation of organisms with relatively incomplete gene annotations.

3.2 Results

3.2.1 The set of differentially expressed genes varies widely by gene model set used

The first step in pathway enrichment analysis is to identify differentially expressed genes. Starting with each of four different sets of gene models, we used RSEM to estimate gene expression and then applied EBSeq to identify differentially expressed genes. For gene models, we used the public Ensembl gene models together with three sets of custom gene models. The custom-constructed gene models were, (1) constructed by *de novo* mRNAseq assembly with Velvet/Oases [82, 57], (2) a genome reference-guided gene set constructed with Cufflinks [68], and (3) a merged gene set constructed with Gimme from a combination of all three other gene sets; see Methods for details. All custom gene sets were constructed using the same Marek's Disease Virus mRNAseq data sets used for differential expression analysis.

We next examined the rate of read mapping to the gene models. The percentages of reads mapped to the Ensembl models are lowest in all samples compared to other gene model sets (Table 3.1). Fewer than 60% of reads mapped to the Ensembl models indicating that the gene models do not include all coding and non-coding regions expressed in the samples. The results from the assembly models confirm that many reads are from regions not included in the Ensembl models. Note that the number of mapped single-end reads is highest in the assembly models, even though those models have the fewest distinct genes.

Table 3.1: **Rate of reads mapped to Ensembl-matched gene models**

	Single Control	Infected	Paired-end Control	Infected	Total Genes
Ensembl	61.93%	63.59%	57.83%	59.24%	15,943
Assembly	81.40%	84.01%	64.24%	66.75%	9,002
Cufflinks	78.16%	81.28%	76.56%	77.44%	14,020
Merged	78.14%	81.48%	75.87%	77.23%	13,973
Total reads	27,618,789	29,693,654	42,632,733	30,804,398	

We compared differentially expressed gene predictions between data sets by linking each gene to an Ensembl gene via BLAST, which provided a common reference identifier. A summary of the correspondence is in Table 3.2. From *de novo* assembly, 9002 (56.46%) of the gene models matched to one of the approximately 16,000 Ensembl genes. Cufflinks and merged models matched considerably more of the Ensembl genes – 14020 (87.9%) and 13973 (87.6%), respectively.

Table 3.2: **Genes and DE genes matched Ensembl genes**

	Ensembl	Assembly	Cufflinks	Merged
All	15943	9002	14020	13973
DE	2538	2109	3433	3402

Of the genes in the different data sets with identifiable correspondence to Ensembl, RSEM predicted that between 2109 and 3433 genes were differentially expressed (DE) across the gene model sets. However, the number of DE genes shared between two different gene sets was much lower, with a maximum of 3069 out of 3428 (89.5%) DE genes shared between Gimme and Cufflinks (Figure 3.1D). Gimme incorporates the Cufflinks gene models, however, and so the best agreement on DE genes between two independent data sets is between Ensembl and Cufflinks, where 1991 genes are DE in both Ensembl (1991 of 2538 DE genes, 78.4%) and Cufflinks (1991 of 3428, 58.0%). Despite this correspondence, the number of disjoint DE genes between samples is often equal to or larger than the number of DE genes in common between any two gene sets.

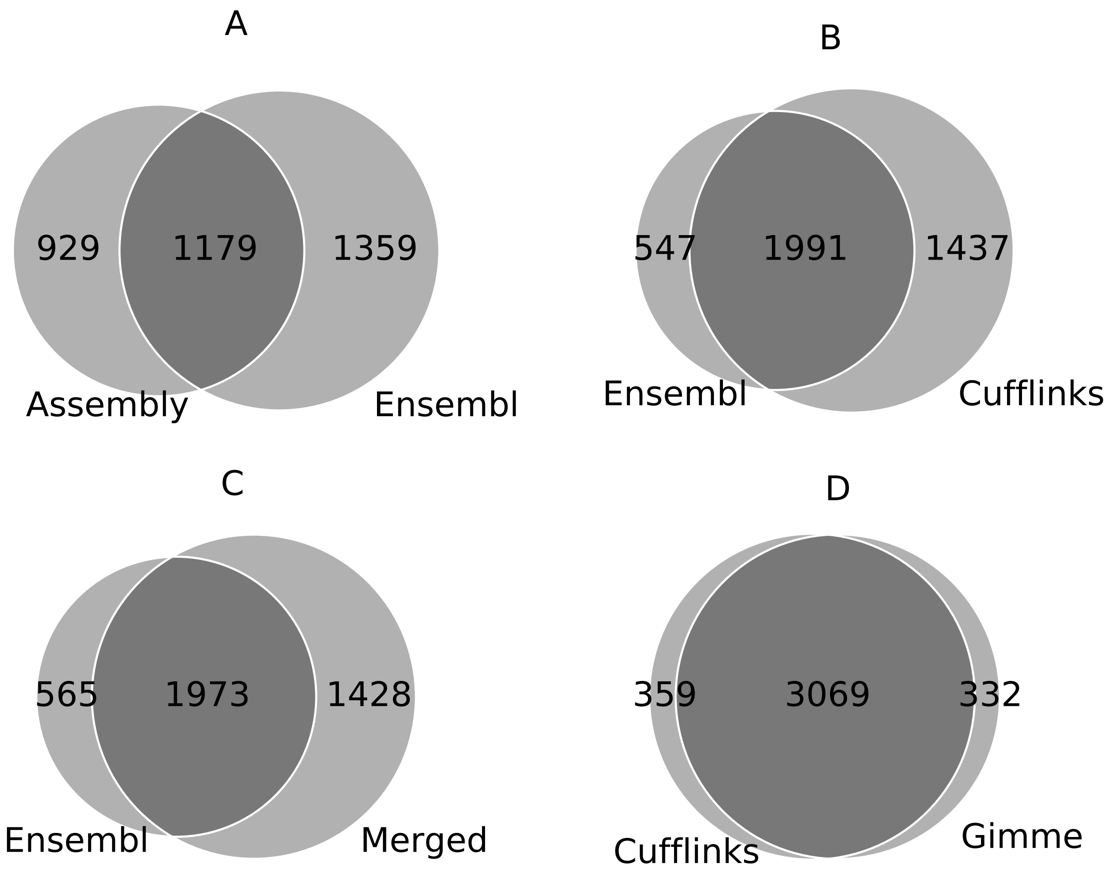

Figure 3.1: **Comparison of differentially expressed genes.** DE genes were matched to Ensembl genes via BLAST. (A-C) The number of DE genes in other gene model sets differ greatly when compared to Ensembl models. (D) Although Cufflinks models were incorporated into Gimme models, not all DE genes in Cufflinks were found in Gimme models.

3.2.2 Variation in differential expression predictions is due to variation in read mapping

To identify the factors contributing to variation in DE analysis, we examined estimated read counts of genes identified as differentially expressed in one gene model set versus another. For example, the *IFNB* gene was not DE when the Ensembl models were used, but was DE when the merged Gimme models were used; upon examination, we found that the number of reads mapping to the *IFNB* gene model was much higher in the merged models (Figure 3.2). This discrepancy in mapped reads was due to a substantial number of reads mapped to the extended regions in the Gimme model, which are not included in the Ensembl model (Figure 3.3). Another example is the

IDH3A gene, shown in Figure 3.4. This gene was expressed in both control and infected groups, but is only identified as DE when the Gimme models are used (Figure 3.5). This is because the complete 5′ UTR is included in only one of the two Ensembl isoforms and not in the RefSeq model, whereas both 3′ and 5′ UTRs are included in the Gimme model. The size of the 3′ UTR in the Gimme model is slightly larger than that of RefSeq and the size of the Gimme 5′ UTR is the same as that of Ensembl model.

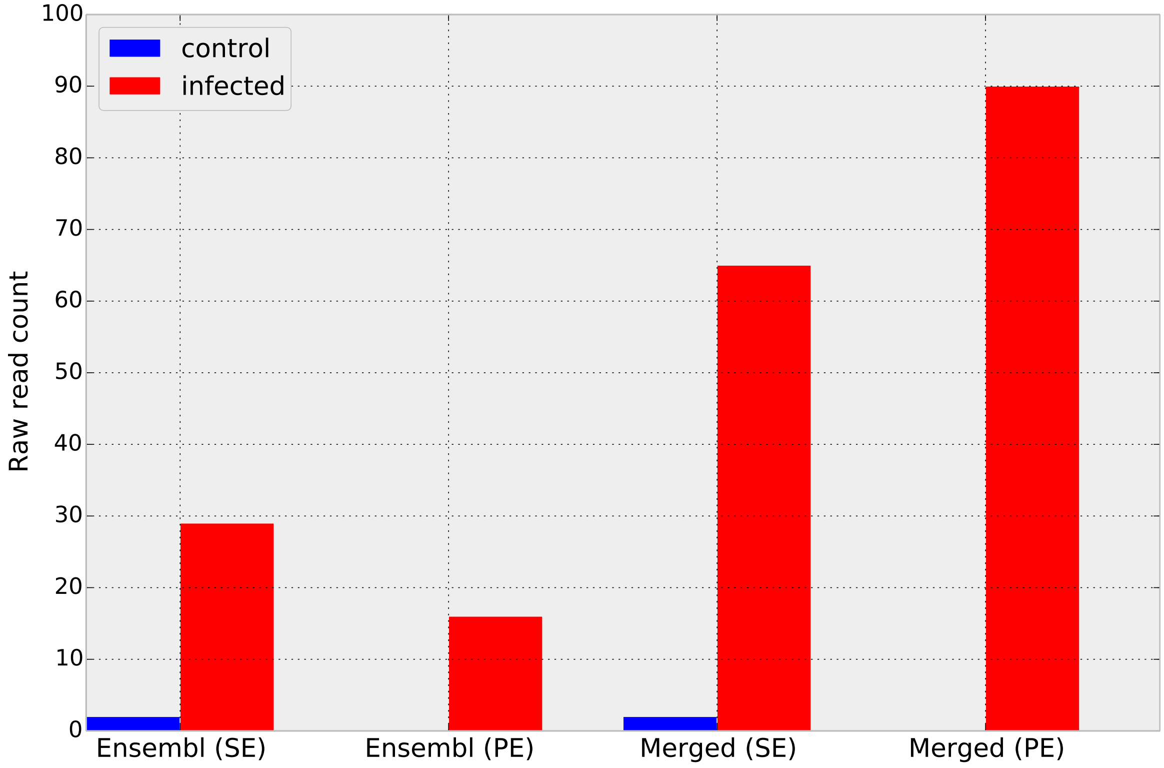

Figure 3.2: **Read counts of *IFNB* gene.** SE=single-end, PE=paired-end

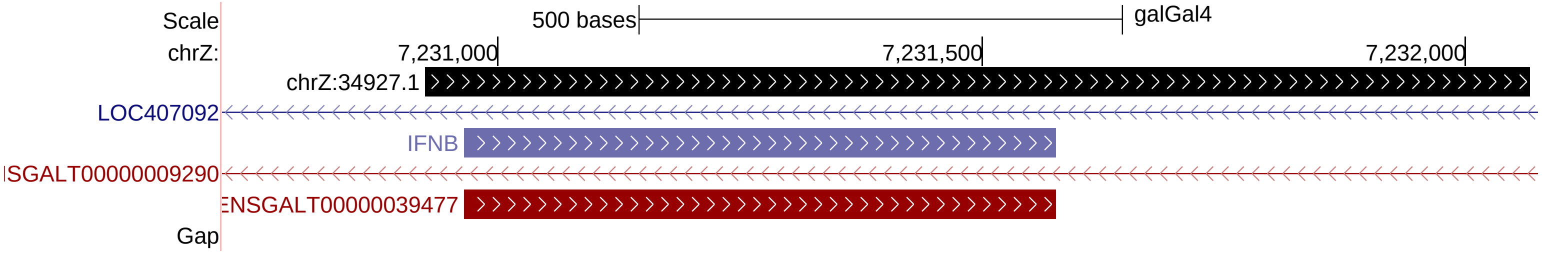

Figure 3.3: ***IFNB* gene models from Ensembl (ENSGALT00000039477, red), Gimme (chrZ:34927.1, black) and RefSeq (blue).** RefSeq and Ensembl models only include the CDS region; whereas, Gimme model include extended regions that could be unstranslated regions (UTRs). The extended regions are of notable sizes, which could account for a substantial number of mapped reads.

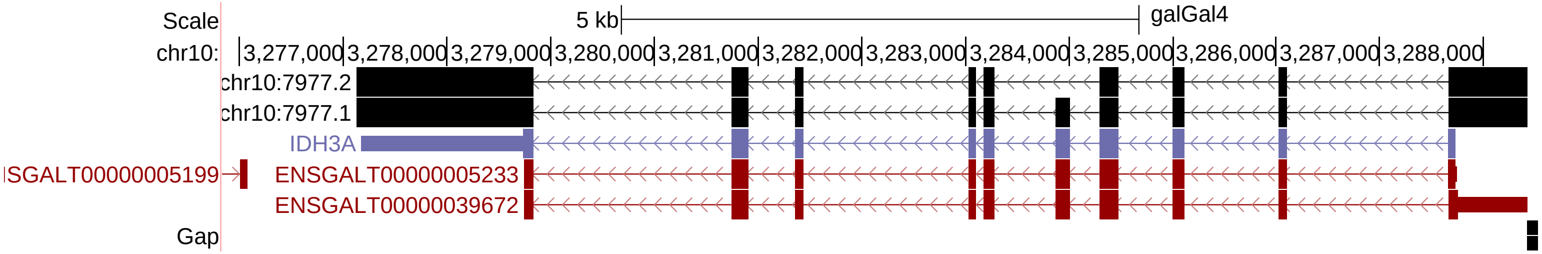

Figure 3.4: ***IDH3A* gene models from Ensembl (ENSGALT00000005233,39672, red), Gimme (chr10:7977.1, chr10:7977.2, black) and RefSeq (blue).** RefSeq model only includes CDS and 3′ UTR. Ensembl model only includes CDS and 5′ UTR in one isoform. Gimme model includes CDS and both UTRs.

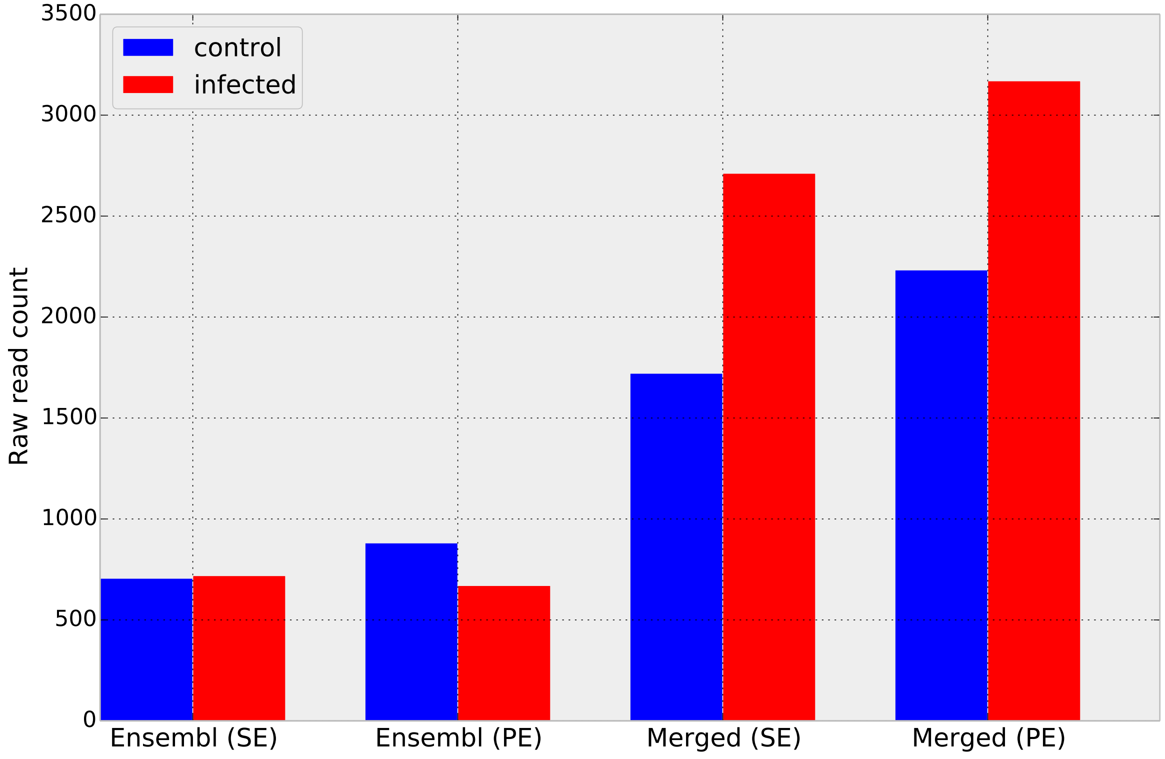

Figure 3.5: **Read counts of *IDH3A* gene.** SE=single-end, PE=paired-end

We next examined the gene size distribution across all four gene model sets used (Figure 3.6). Both the Cufflinks *ab initio* gene models and the Gimme merged gene models are significantly longer than the Ensembl and assembly-based gene models, suggesting that they are more complete.

Finally, we examined the read counts and length-normalized read counts for control and infected mRNAseq samples (Figure 3.7). In all cases, both the read counts and length-normalized read counts were significantly higher for the Gimme gene models.

To illustrate the effect of gene models variation on gene expression estimates, comparisons of effective gene sizes from all gene models and read counts from Ensembl and merged models are shown in Figure 3.6 and 3.7. Effective gene sizes of 1,563 differential-expressed gene vary greatly among gene models (Figure 3.6). The effective gene size is the average of sizes of all transcripts from the same gene. The median of effective gene sizes of Ensembl models is 2413 bp (IQR 2250 bp); whereas those of Cufflinks and merged models are much higher (3320 bp (IQR 2550 bp) and 3270 bp (IQR 2595 bp) respectively). This is expected because Cufflinks and combined models

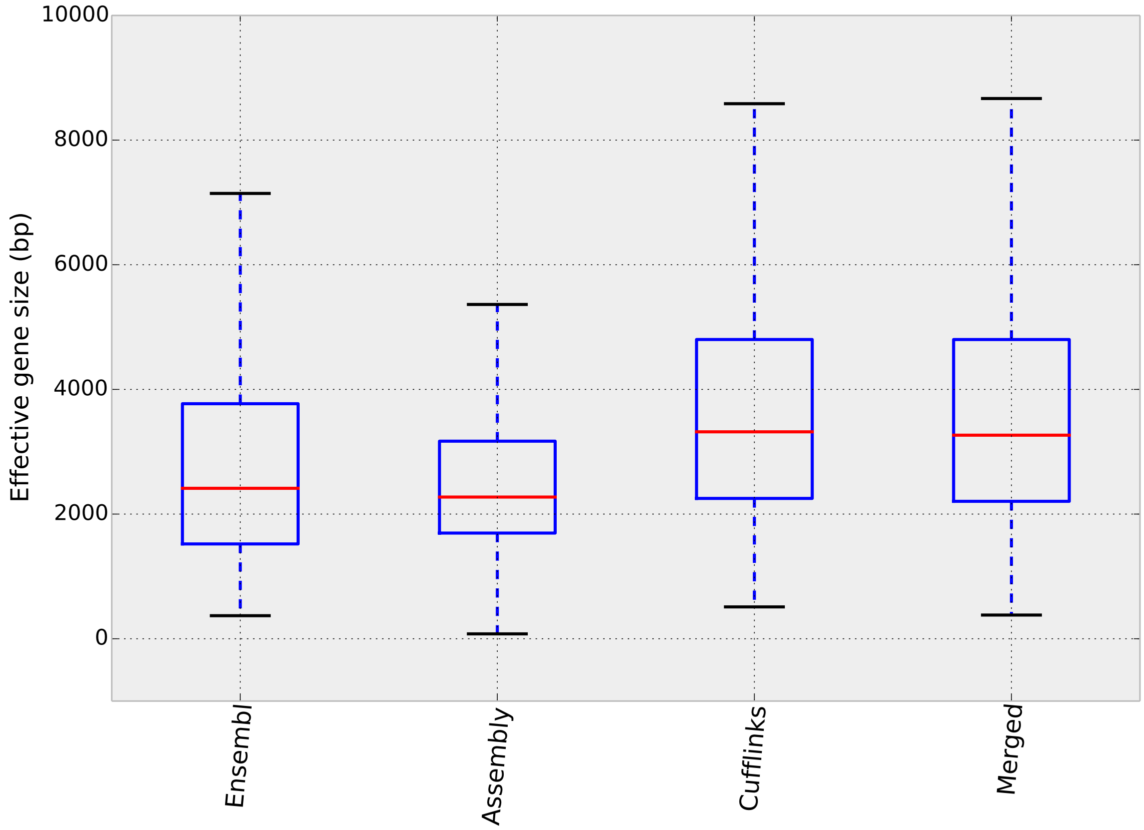

Figure 3.6: **Gene sizes distribution.** Distribution of the same set of genes from different models are plotted. Sizes of genes in Ensembl models are slightly shorter than those from combined models. This is due partly to the fact that some genes may not contain UTRs. The sizes of genes from *de novo* assembly are significantly smaller than other models indicating that transcripts are incomplete or fragmented.

are supposed to include UTRs. On the other hand, the median of the effective gene sizes of gene models from *de novo* assembly is close to that of Ensembl models; however, the IQR is much smaller (median = 2273 bp, IQR = 1474 bp) suggesting that transcripts from *de novo* assembly are incomplete compared to Ensembl models. The difference in gene sizes between Ensembl and combined models results in a substantial deviation of read counts as shown in Figure 3.7. Medians of read counts from the same condition were significantly different between Ensembl and combined models. After normalization by gene sizes, the difference were diminished suggesting that the deviation of read counts is partly due to the size of the gene models.

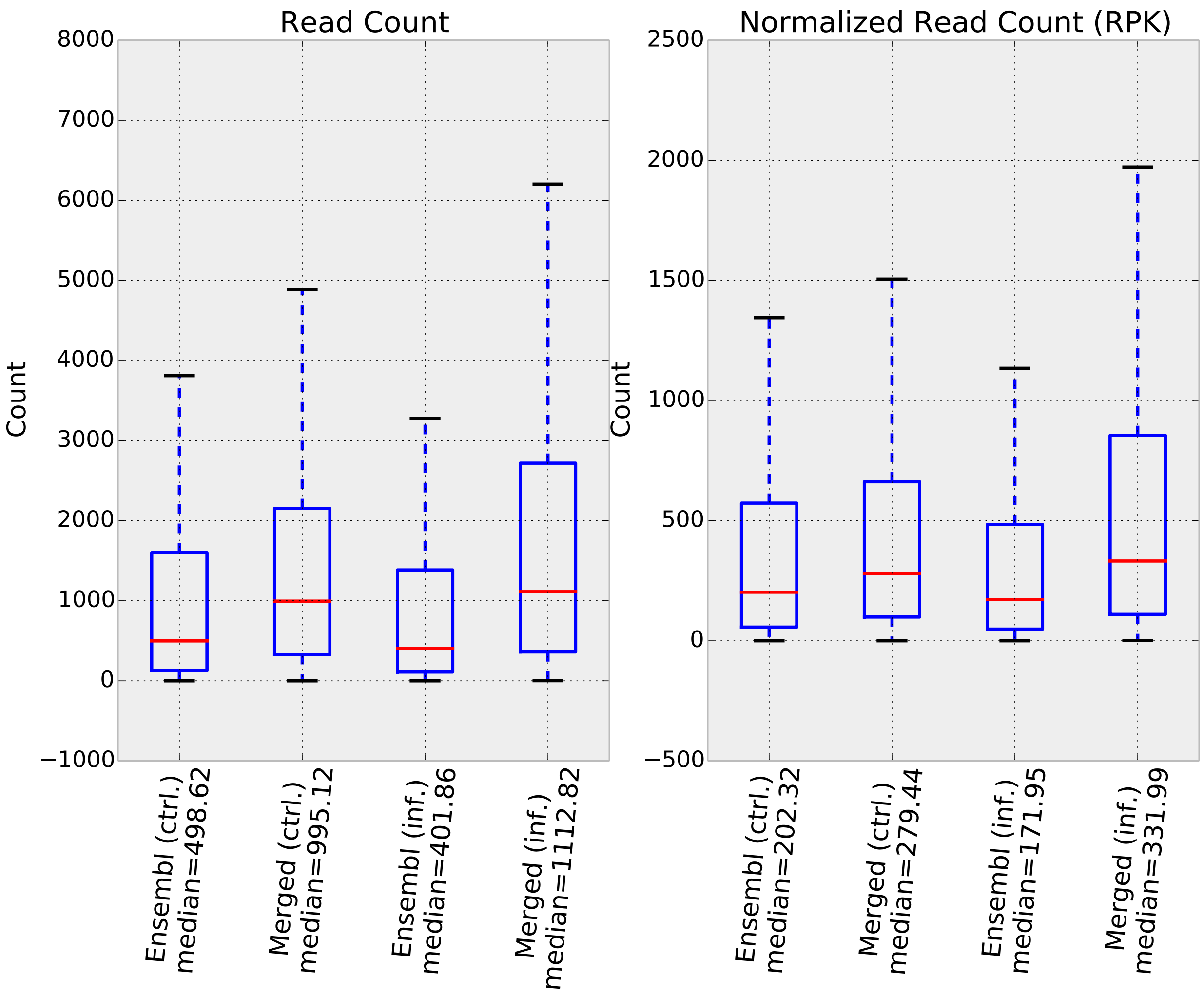

Figure 3.7: **Read counts comparisons.** In the same condition, read counts from Gimme models are significantly higher than those from Ensembl gene models. Read counts are more similar after normalization by gene lengths. Read counts were normalized by effective gene sizes using this formula: $1000 \times \frac{count}{length} = ReadPerKilobase(RPK)$.

3.2.3 A majority of differential-expressed genes are not annotated in KEGG pathway

We next used the Kyoto Encyclopedia of Gene and Genomes (KEGG) to annotate the differentially-expressed genes with putative function and identify enriched pathways using GOSeq [80]. Because there are only a small number of species-specific chicken annotations in the KEGG database, we also used homology search against human Ensembl proteins (Release 74) to transfer gene annotations from human genes to our chicken gene models.

Transferring gene annotations from human to chicken did not result in a large increase in the number of genes that were annotated, but the number of genes with pathway annotations did increase. In Figure 3.8, we show the effect of this transfer on the pathway annotations for different gene model sets. For the merged data set, ∼27% of DE gene models had species-specific KEGG annotations, while approximately 36% had human KEGG annotations.

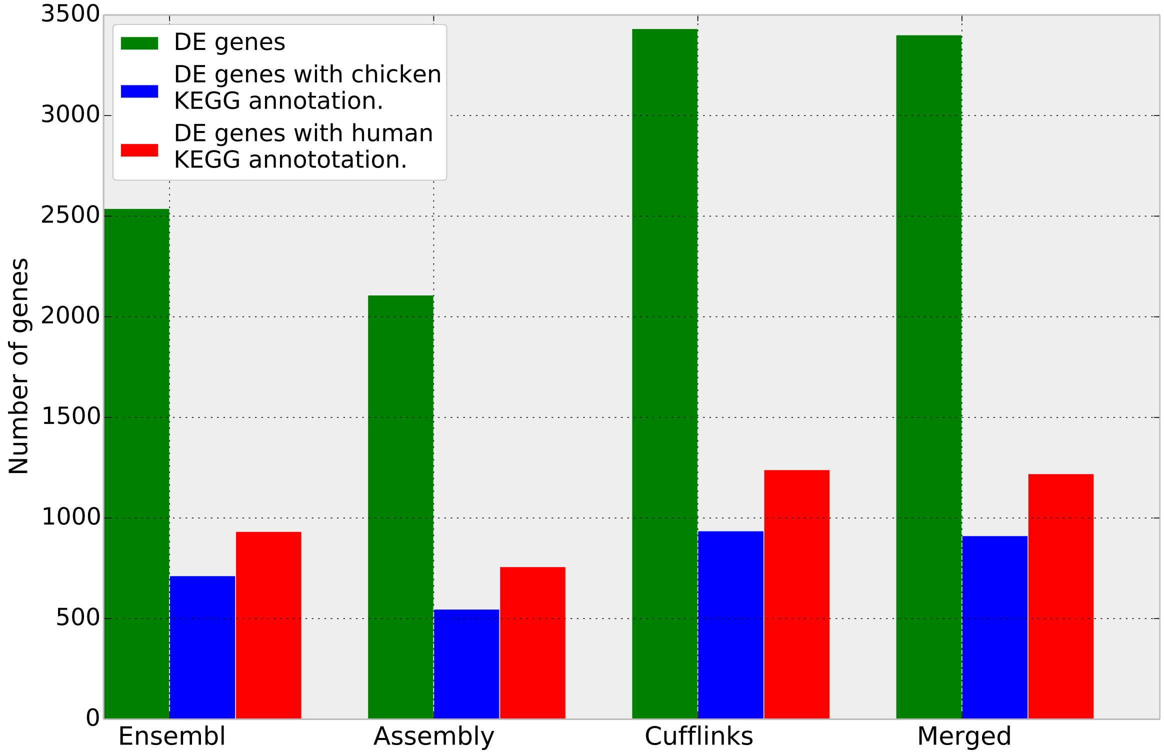

Figure 3.8: **DE genes with chicken and human KEGG annotations.**

Unsurprisingly, transferring gene annotations also led to a dramatic difference in the KEGG pathways predicted to be enriched between the two conditions. Figure 3.9 shows the enriched KEGG pathways in the merged gene models with first, only chicken annotation and second, with transferred human annotations. A total of 26 pathways were predicted to be enriched in one or both of the annotation sets, with 6 in common, 9 unique to the species-specific annotations, and 11 unique to the human annotation set. A majority of the chicken-specific annotations were immune system annotations, reflecting the extensive work done on chicken immunity [7].

3.2.4 Pathway predictions vary widely between gene model sets

To compare the pathway predictions using chicken+human annotations, we calculated the Spearman rank correlation coefficient for all possible pair-wise comparisons. The results, shown in Table 3.3, show that the correlations are very poor. The weakest correlation is between the pathway predictions for the Ensembl gene model set and the merged gene model set.

Table 3.3: **Spearman Rank Correlation**

	Assembly	Ensembl	Cufflinks	Merged
Assembly	1.0	0.39	0.45	0.50
Ensembl		1.0	0.48	0.36
Cufflinks			1.0	0.76
Merged				1.0

Five pathways are predicted to be enriched only when using the Ensembl gene models (Figure 3.10), including the antigen processing and presentation pathway, which is important for immune responses. In addition, the T cell receptor signaling pathway is predicted to be enriched in all of the gene model sets excepting only the Ensembl models.

The correlation coefficient between enriched pathways predicted from Ensembl gene models and Cufflinks gene models is rather high, which may be because the Ensembl models were integrated with RNA-Seq data to construct the Cufflinks models. The integration is done based on the criteria that incomplete transcripts or transcripts without novel introns compared to the Ensembl models are discarded and UTRs from RNA-Seq data are used to extend Ensembl UTRs [52]. Therefore, Cufflinks models are a superset of the Ensembl models and should contain all pathways from Ensembl models. However, eighteen pathways are unique to Cufflinks and seven pathways are unique to Ensembl; this may be due to a difference in the read mapping percentage.

Finally, the correlation between enriched pathways predicted using the Cufflinks-based differential expression and the Gimme-based differential expression is the highest. Although twelve enriched pathways are unique to Cufflinks, and four pathways are unique to the merged models, the correlation coefficient is high because the common pathways have similar significance scores.

Twenty nine enriched pathways are different between the merged models and Ensembl and this results in a slightly lower correlation coefficient than that between the combined models and Cufflinks. Table 3.4 shows common and unique enriched pathways from Ensembl and combined models. A majority of unique enriched pathways from Ensembl are involved in metabolism and cellular activities. In contrast, unique pathways from combined models are involved in immune response and cancer, which we would expect to be perturbed by MDV infection.

Table 3.4: **Pathways from Ensembl and merged gene models**

pathway ID	Term	Adjusted p-value
	Pathways in Ensembl only	
00980	Metabolism of xenobiotics by cytochrome P450	2.3e-06
00480	Glutathione metabolism	1.9e-05
00982	Drug metabolism - cytochrome P450	9.5e-04
04142	Lysosome	1.1e-03
04010	MAPK signaling pathway	3.0e-03
00400	Phenylalanine, tyrosine and tryptophan biosynthesis	3.4e-03
00500	Starch and sucrose metabolism	4.9e-03
00983	Drug metabolism - other enzymes	5.5e-03
00360	Phenylalanine metabolism	6.3e-03
	Common pathways	
04060	Cytokine-cytokine receptor interaction	5.8e-15
04145	Phagosome	4.7e-07
01100	Metabolic pathways	1.1e-06
04620	Toll-like receptor signaling pathway	4.8e-05
04623	Cytosolic DNA-sensing pathway	1.2e-04
04210	Apoptosis	1.8e-04
04630	Jak-STAT signaling pathway	2.1e-04
04672	Intestinal immune network for IgA production	2.4e-04
04622	RIG-I-like receptor signaling pathway	3.6e-04
04621	NOD-like receptor signaling pathway	1.4e-03
04650	Natural killer cell mediated cytotoxicity	2.0e-03
	Pathways in merged models only	
04514	Cell adhesion molecules (CAMs)	1.8e-04
04141	Protein processing in endoplasmic reticulum	10.0e-04
04510	Focal adhesion	2.3e-03
04512	ECM-receptor interaction	3.3e-03

For non-model organisms without high-coverage functional pathway annotation or gene ontology (GO), an alternative method for functional analysis is to use pathway annotations of closely related or well-studied model organisms such as mouse and human. This approach could increase sensitivity of pathway analysis; however, it will not include species-specific genes and pathways. For instance, in this study, some pathways involved in chicken immune responses including natural killer cell mediated cytotoxicity, intestinal immune network for IgA production, and focal adhesion were not enriched in human annotation with merged models (Figure 3.9).

To overcome this problem, we propose to use a custom pathway annotation built from a combination of annotations from the organism and related species. To validate our method, we built a custom annotation by assigning human annotations to chicken genes with no pathway annotations and used the custom annotation for pahtway analysis. The results (shown at the bottom graph of Figure 3.10) include some significantly enriched pathways involved in immune responses such as complement and coagulation cascades, T cells receptor pathway, and chemokines signaling pathway. Almost all enriched pathways from either Cufflinks or Ensembl models are also included. Besides the top three most significant pathways, pathways from custom annotations, Cufflinks and Ensembl have only slightly different significance scores. This indicates that custom annotation can help increase the sensitivity of pathway analysis, especially in semi-model organisms. However, results have to be interpreted with great care due to biological differences between organisms.

3.3 Discussion

3.3.1 Variation in gene model sets greatly affects expression estimates

Methods for differential gene expression analysis typically estimate gene expression from the number of reads mapped to each gene model. Reads that do not map to gene models, such as reads belonging to unannotated UTRs or exons, are generally excluded. This exclusion leads to insensitivity in estimates of expression, and can be particularly problematic when genes have long UTRs and uenven read coverage.

Here, we show that using the same parameters for mapping and differential exprssion analysis with several different gene model sets results in large differences in both read mapping and the prediction of differentially expressed (DE) genes. While some differences are expected, we found that the differences were surprisingly large: for most of the comparisons, the DE genes disjoint between the two gene model sets were larger in number than the DE genes shared between the two gene model sets.

These differences are due to differences in the number of reads mapping to the gene models. Unsurprisingly, many more reads map to the gene model data sets constructed from the mRNAseq data – between 15% and 20% (Table 3.1). This results in many more reads mapping to each gene model (Figure 3.6a) and significantly higher RPKM estimates for each mRNAseq data set (Figure 3.6b).

3.3.2 Conclusion

As the use of RNA-Seq to study semi- and non-model organisms has vastly increased recently, one objective of this study is to demonstrate the potential problem of transcriptome study in organisms that lack high quality genome, gene models and functional pathway annotation. Gene models derived from *de novo* assembly are usually the only choice for those organisms; however, the results are relatively poor compared to others. Custom gene models built from Cufflinks (integrated with Ensembl) helped increase sensitivity of pathway analysis and should be used if a reference genome is available and *de novo* assembly is not feasible. However, the method is reliant on a quality of a genome sequence and subject to read-mapping biases. Combined gene models, although appearing to have lower sensitivity compared to those from Cufflinks and Ensembl, still recover most of the pathways with high significance scores. The advantage of using combined gene models is it allows the integration of gene models from *de novo* assembly, which are not affected by read-mapping biases. Results from our study show that they significantly increase sensitivity of genes and isoforms detection. Depending on available resources and the objective of the study, we suggest that appropriate gene models should be carefully chosen to maximize the quality of the analysis. Moreover, using functional pathway annotation from a related species is always necessary for those organisms, but it will not include species-specific genes and pathways. The problem is compounded when only gene models from *de novo* assembly are available. Therefore, biologists studying non-model organisms need to be aware of these limitations when interpreting the results from transcriptome analysis.

3.4 Materials and Methods

3.4.1 Sequences and quality trimming

mRNAs were extracted from spleens of control and infected line 7 chickens (14 dpi). Sequence libraries were prepared by standard Illumina unstranded single- and paired-end protocols. Library size of the paired-end datasets is approximately 175 bp. Read lengths are 75 bp in both single- and paired-end libraries. Reads were quality trimmed by condetri 2.1 [60] with quality score cutoff of 30. The first 10 bases were removed due to a non-uniform distribution of nucleotides.

3.4.2 Custom gene models construction

Reads were mapped to chicken genome (galGal4) by TopHat 2.0.9 [66] and gene models were constructed with Ensembl models release 73 as a reference by Cufflinks 2.1.1 [68]. Velvet 1.2.03 [82] and Oases 0.2.06 [57] were used to assemble reads with hash lengths ranging from $21-31$. Transcripts from all hash lengths were then reassembled with OasesM. Combined gene models were constructed by Gimme [48], which combined Cufflinks models and alignments of *de novo* transcripts to the genome. Transcripts were mapped to the genome using BLAT [24] and for each transcript, only an alignment with the best mapping score was used.

3.4.3 Differential gene expression and gene ontology

Quality filtered reads were mapped to transcripts from all gene models without poly-A tail added by Bowtie 1.0 [27]. Estimated gene expression and differential expression analysis was performed by RSEM 1.2.7 [30] and EBseq [29] respectively. DE genes were annotated by homologous sequences in chicken and human proteins from Ensembl release 73 and 74 respectively. GOSeq 1.14.0 [80] was used to perform enrichment analysis using KEGG annotations from org.Gg.eg.db [8] and org.Hs.eg.db [9]. Human KEGG annotations were added to all DE genes to create combined KEGG pathways.

3.4.4 Pipeline and scripts

The pipeline and scripts used in this study is available at

`https://github.com/likit/RNASeq-methods-comparison`

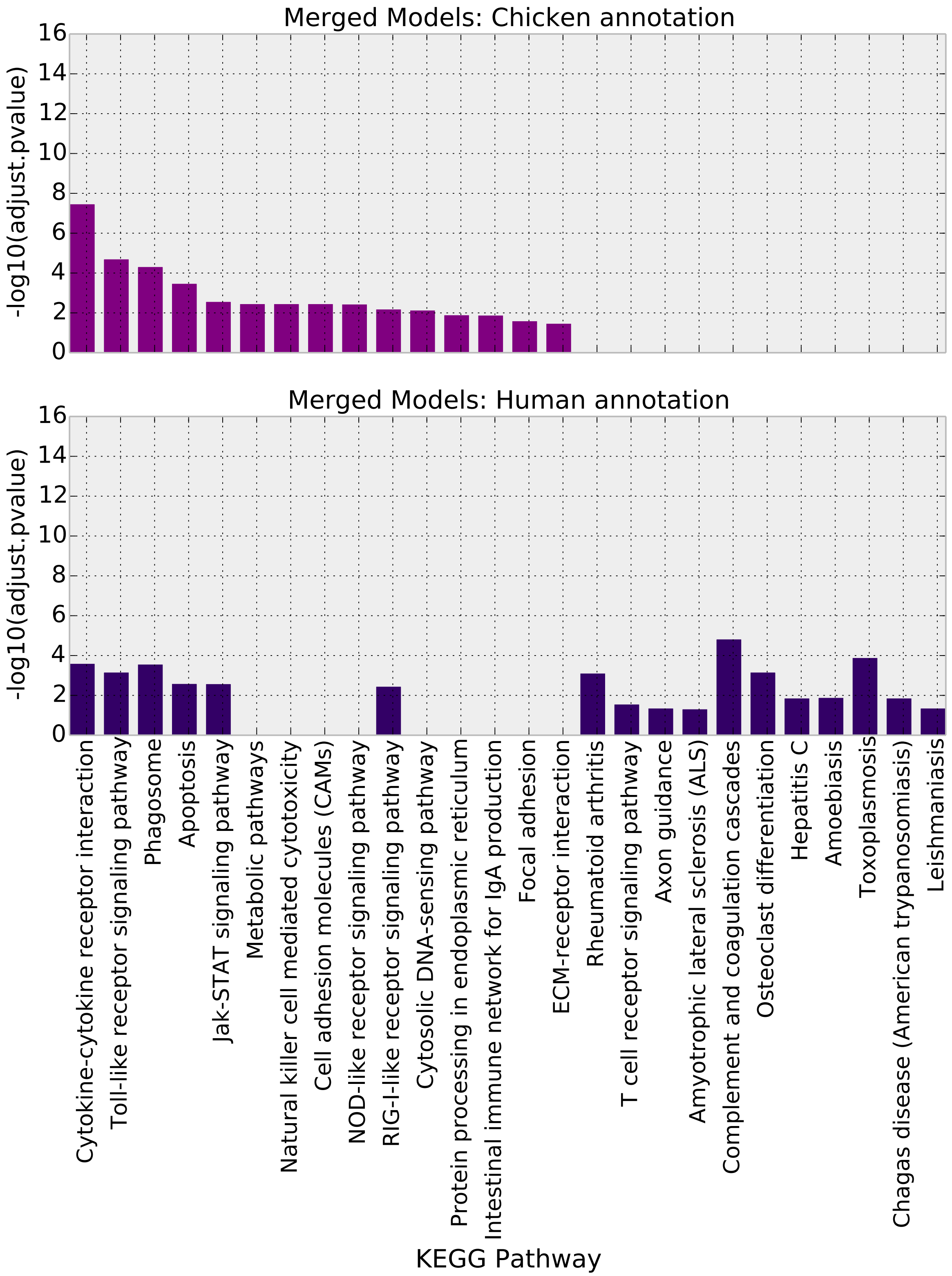

Figure 3.9: **Enriched KEGG Pathways from merged models with chicken and human annotation**

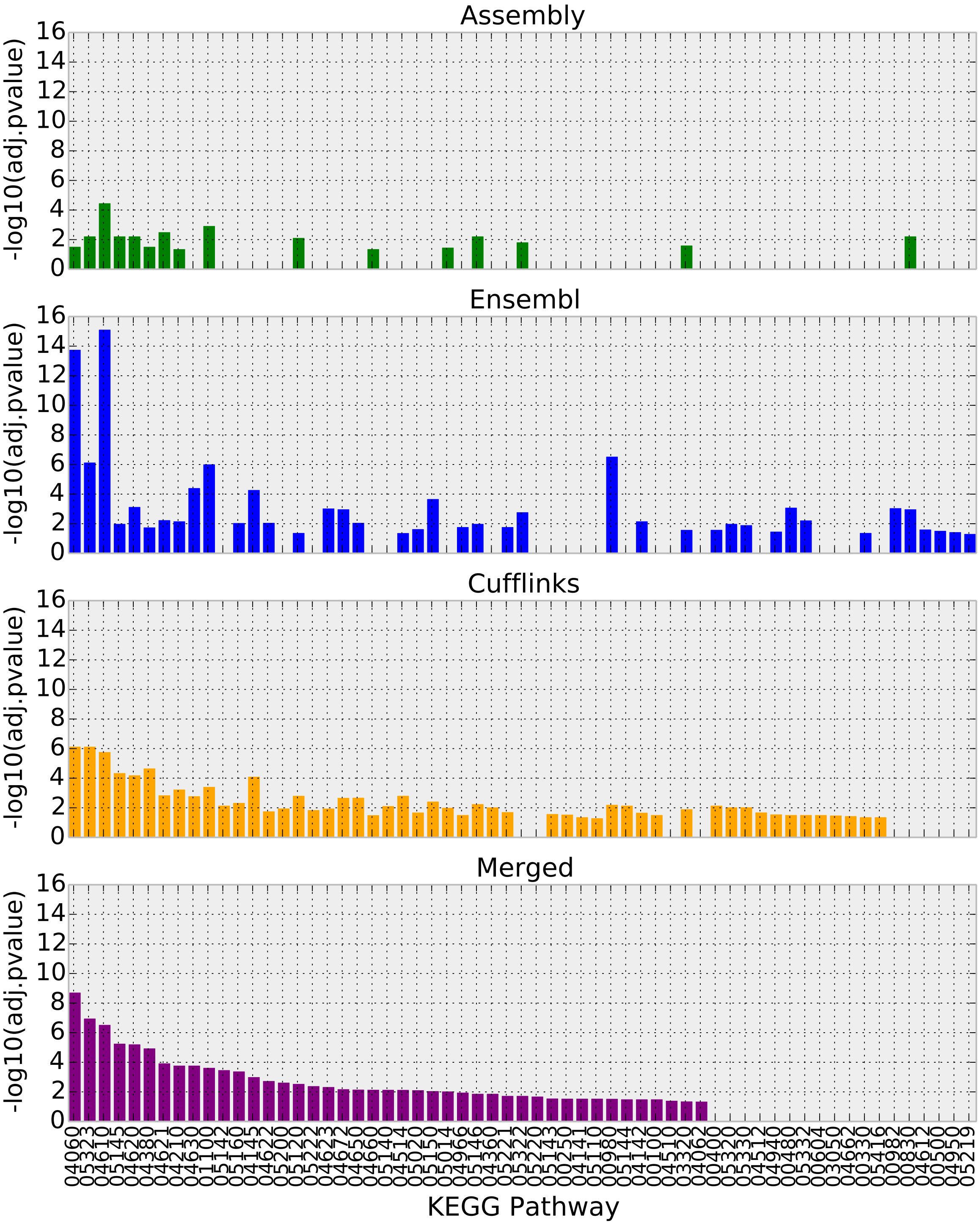

Figure 3.10: **Enriched KEGG Pathways from different gene models**

CHAPTER 4

A GENOME-WIDE SCAN FOR GENES AND ISOFORMS RESPONSIBLE FOR MD RESISTANCE

4.1 Introduction

Marek's disease (MD) is an economically significant chicken disease that affects the poultry industry worldwide with estimated annual cost of \$2 billion [41]. The disease is caused by the highly oncogenic Marek's disease virus (MDV), an alphaherpesvirus that induces T-cell lymphomas in susceptible birds. Vaccination is the primary control measure, which is effective in reducing incidence of tumor formation. However, since MD vaccines are not sterilizing, they do not prevent infection or horizontal spread of the virus. As a consequence, MDV field strains that overcome vaccinal protection have arisen repeatedly over time. Therefore, there is a need for sustainable alternative controls measures, such as improving genetic resistance.

Many studies have reported strong associations between MHC alleles and resistance or susceptibility to MD. For example, chickens with MHC allele B^{21} are resistant in contrast to chickens with the B^{19} allele, which are susceptible. ADOL lines 6 and 7, both share the same MHC B^{2} allele, yet exhibit different phenotypic responses; e.g., challenge with the JM/102W strain typically result in 0 and 100% MD incidence for lines 6 and 7, respectively. Thus, the major unanswered questions are what genetic factors, especially those that are non-MHC, contribute to susceptibility and resistance to the disease and what are the main contributing mechanisms.

In the past decades, significant efforts have been made to study variations in global gene expression between resistant and susceptible birds using microarray and RNA-Seq methods in order to identify non-MHC genes that contribute to resistance to MD [54, 39, 69, 79, 6]. However, none of the studies have investigated differential expression of alternative isoforms, which are known to play a significant role in many biological events including immune responses. In addition, studies have shown that isoform expression levels can provide better signatures for some diseases [84].

Changes of isoform expression levels are governed partly by two types of *cis*-regulatory elements: Exon Splicing Enhancer (ESE) and Exon Splicing Silencer (ESS), which are located within an exon sequence. A number of sequence motifs of ESE and ESS have been identified in human and some other organisms and can be predicted *in silico*. Mutations that disrupt or create those motifs could alter splicing patterns leading to aberrant alternative splicing. A number of disease-associated single-nucleotide polymorphsims in coding regions (SNPs) that affect ESEs and ESSs have been well characterized [4, 73]. Therefore, variations in isoform expression could lead to identification of SNPs that underlie genetic resistance to MD. In this article, we reported differential-expressed genes and isoforms that may contribute to resistance to MD as well as SNPs that can potentially affect isoform expression levels.

4.2 Results and Discussion

4.2.1 Differential expression results from our method are comparable to previous studies

To study gene and isoform expression, we incorporated Ensembl gene model release 73 with *de novo* and a reference-guided transcriptome assembly to build custom chicken gene models. The models, therefore, include both Ensembl annotated transcripts and putative genes and isoforms. The advantage of using custom gene models is it allows an investigation of unannotated genes and isoforms, which is necessary for in-depth study of gene expression.

Some DE genes that were reported by previous microarray studies were also found to be differentially expressed in this study. For example, B6.1 (Bu-1) is known to be down regulated approximately 2.3 fold in susceptible chickens with the MHC allele B^{19} at 4 days post infection (d.p.i) [54]. It was also found to be down regulated $\sim$3-fold in the susceptible line in our study. Similarly, *GMZA* reported to be upregulated across genetically different chickens (B^{19}, B^{21} alleles), and was also found to be highly upregulated here. In contrast, some genes that have been reported to be highly expressed in resistant chickens were downregulated in both lines. Those genes are *AMIGO2, MMP13* and *CLEC3B*, which were found to be downregulated more than 2-

fold [54]. Other immune genes reported to be highly expressed in susceptible chickens including *AVD, ART1, NOS2, CXCL13L2, MX1* and *SOCS1* [61] were also found to be highly upregulated in both lines. However, our results show a similar expression patterns for *IL6* and *IL18*, which were only upregulated in the susceptible line at an early stage of infection ($3-5$ d.p.i).

In contrast, *IL15* has been reported to be non differentially expressed between control and infected chickens in both lines [20]; however, here it was only upregulated in the susceptible line. Expression of *IL15* is induced by *TLR9*, which binds to non-methylated CpG residues present in the genomes of many DNA viruses, including herpes simplex virus. This cytokine auto-regulates the expression of *CD40*, which is a transmembrane receptor required for activation of macrophages by CD4 T cells. Consequently, *CD40* was only upregulated in the susceptible line (data not shown).

4.2.2 Differential gene expression indicates active immune responses to ongoing lytic infection in the susceptible line

Many genes were found to be differentially expressed (DE) between control and infected chickens in both lines. While the number of unique downregulated genes in both lines was approximately equal, the number of unique upregulated genes in the susceptible line was much greater compared to the resistant line (Figure 4.1).

Interestingly, some genes that were differentially expressed in both lines were regulated in the opposite direction (Table 4.1). Among genes downregulated in the resistant line but upregulated in the susceptible line were *LL* (lung lectin) and *SFTPA1*, which encode a calcium-dependent C-type lectin and a lung surfactant protein respectively. Both molecules are important in innate immunity [17, 25]. *LIMS1* is involved in cell differentiation and proliferation and *PPARG* is a suppressor of the *NFκB*-mediated proinflamatory response. On the other hand, nearly all genes upregulated in the resistant line but downregulated in the susceptible line are involved in cell survival such as mRNAs splicing, cell growth, and protein synthesis, except CD7 whose function is involved in T cell-B cell interaction. This difference suggests that even at this stage of infection in the resistant line, the lytic phase could be repressed. Therefore, only genes involved in cell division are upreg-

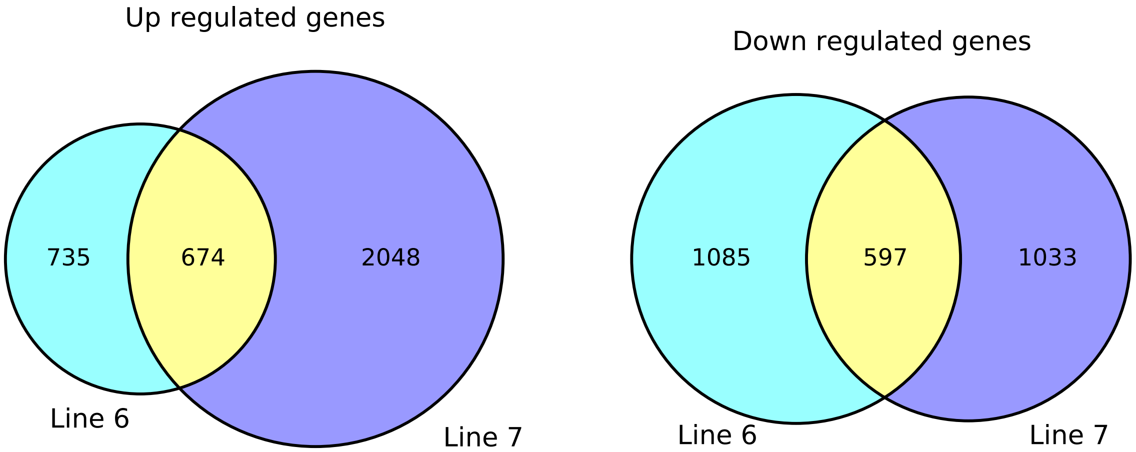

Figure 4.1: **Differential-expressed genes in response to MDV infection.** More genes are differentially expressed between control and infected chickens from line 7 than line 6.

ulated possibly to repair the initial damage due to infection in the resistant line. In comparison, the lytic phase in the susceptible line may still continue and as a result, genes involved in immune responses are still upregulated.

Table 4.1: **Genes regulated in opposite directions in response to MDV infection**

Gene	Description	log_2FC	
		Resistant	Susceptible
LL	Lung lectin	-3.36	8.71
GIF	Gastric intrinsic factor	-2.15	3.11
C14ORF1	Chromosome 14 open reading frame 1	-3.11	2.71
SFTPA1	Surfactant protein A1	-4.84	3.73
SCAF8	SR-related CTD-associated factor 8	-8.71	8.25
PPARG	Peroxisome proliferator-activated receptor frame 1	-6.99	2.06
NDUFA4	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4	1.66	-1.03
RAD17	RAD17 homolog (*S. pombe*)	1.45	-1.80
RPL39	Ribosomal protein L39	3.34	-1.63
ATP8A2	ATPase, aminophospholipid transporter, class I, type 8A, member 2	7.66	-7.80
NDUFB3	NADH dehydrogenase (ubiquinone) 1 beta	1.03	-1.34
PSMG3	Proteosome assembly chaperone 3	1.52	-1.26
MED9	Mediator complex subunit 9	8.23	-3.15
PNISR	PNN-interacting serine/arginine-rich protein	5.58	-1.81
S1PR1	Sphingosine-1-phosphatase receptor 1	1.31	-6.96
CD7	CD7 molecule	2.88	-1.38
LOC100858785	Unknown	1.26	-1.75
THOC7	THO complex 7 homolog (Drosophila)	7.01	-6.90
DNAJA2	DnaJ (Hsp40) homolog, subfamily A, member 1	2.34	-6.02

(-) down-regulated, (+) up-regulated

In addition, type I interferon (*IFN-γ* and *IFN-β*) as well as *INF-α3* were found to be highly upregulated in infected chickens in both lines (Table 4.2). However, expression of genes encoding their corresponding receptors were not different in the resistant line, but upregulated in the susceptible line. This could also reflect the ongoing immune response in the susceptible line.

Table 4.2: **Cytokine-related gene expression in response to MDV infection**

		log_2FC	
Symbol	Description	Resistant	Susceptible
IL2RG	Interleukin 2 receptor, γ	–	0.55
IL6	Interleukin 6 (interferon, β 2)	–	5.11
IL6ST	Interleukin 6 signal transducer (gp130, oncostatin M receptor)	–	1.36
IL8L1	Interleukin 8-like 1	1.90	–
IL15	Interleukin 15	–	1.06
IL18	Interleukin 18 (interferon-γ inducing factor)	1.92	4.07
IL18R1	Interleukin 18 receptor 1	1.94	1.64
IFNG	Interferon-γ	5.14	4.90
IFNB	Interferon-β	4.83	5.64
IFNA3	Interferon-α 3	4.09	5.48
IFNGR1	Interferon-γ receptor 1	–	2.04
IFNGR2	Interferon-γ receptor 2	–	0.50
IFNAR1	Interferon-α,β receptor 1	–	1.46
IFNAR2	Interferon-α,β receptor 2	–	0.58

4.2.3 Functional analysis of differential-expressed genes indicates inactive adaptive immune responses in the resistant line

To determine pathways that were perturbed during the infection, data were analyzed by GOSeq, which accounts for gene length bias unique to the RNA-Seq data [80]. Significantly perturbed pathways (FDR $<$ 0.1) from both lines that involved in immune response include the TLR signaling pathway, cytokine-cytokine receptor interaction, intestinal immune network for IgA production, and cell Jak-STAT signaling pathway (Figure 4.2). Some other pathways important in response to viral infection and only significantly enriched in the susceptible line include phagosome, apoptosis, RIG-I-like receptor signaling pathway, NOD-like receptor signaling pathway, and lysosome. For the phagosome pathway, a pathway that includes genes important for stimulation of the adaptive immune response, although MHC class I (*BF1*) was differentially expressed in both lines, other genes involved in expressing newly synthesized MHC class I were only upregulated in the susceptible line suggesting that new MHC I molecules were actively produced. Furthermore, Gene Ontology analysis of biological processes (GO:BP) (data not shown) shows that categories involved in both adaptive and innate immune responses were enriched in the susceptible line. On the other hand, only categories involved in innate immune responses were enriched in the resistant line. In addition, enrichment of the apoptosis pathway in the susceptible line suggests that the programmed cell death could be induced by the CTL response to eliminate ongoing viral infection.

At this stage of infection, our results suggest that lytic infection of MDV stimulates both innate and adaptive immune responses, which leads to activation of T cells in the susceptible line. Only activated T cells are believed to be infected by MDV, therefore, the lytic phase could facilitate the spread of the viruse by enhancing expansion of activated T cells. Due to the cell-associate nature of MDV, the viruses transfer to T cells via cell-to-cell contact between B cells and T cells during antigen presentation or B cell activation by T helper cells. Therefore, it is beneficial for the host to restrain such contact. However, it is not clear how chickens in the resistant line control the lytic infection of MDV. Two mechanisms have been speculated to contribute to MD resistance. First innate immune responses could be highly effective and could activate strong adaptive immune responses that rapidly control viral replication and force the viruses to enter into the latent phase. Second, the innate immune responses itself could be highly effective in limiting viral replication [61].

4.2.4 Genes with differential exon usage (DEU) in response to MDV infection can be divided into four groups based on their patterns of expression

The immune system is isoform-rich and many genes express different isoforms with distinctive functions in response to stimuli such as stress, chemicals and infection. Changes in expression of splice forms of immune related genes have been reported to be associated with increased sus-

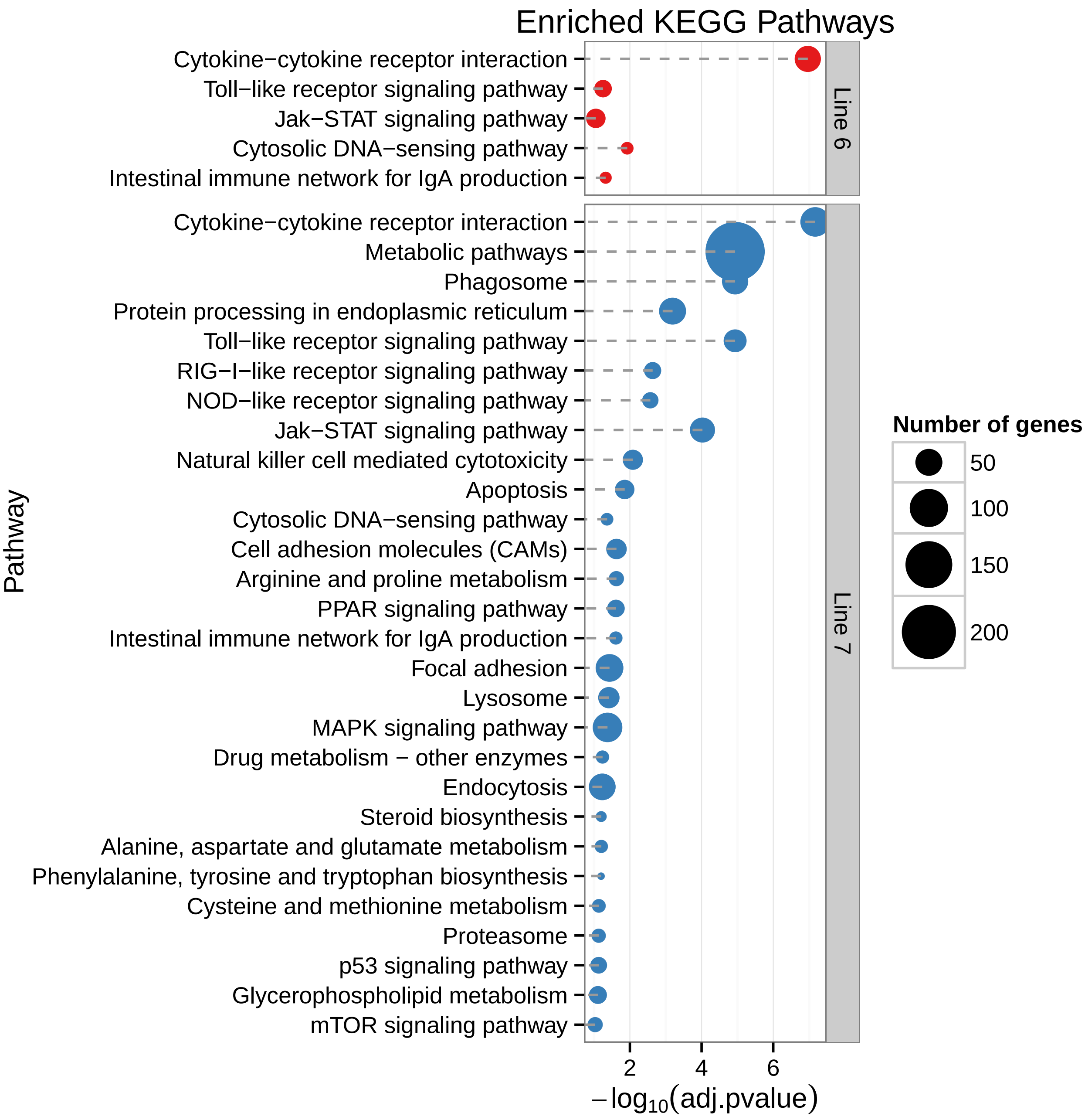

Figure 4.2: **Enriched KEGG pathways.** Significantly enriched KEGG pathways from differentially expressed genes from lines 6 and 7 by GOSeq (FDR<0.1).

ceptibility to and poor prognosis, of diseases [38]. Studying differential isoform expression could therefore shed light into inherent differences between lines that confer resistance or susceptibility to MD.

In the past decades, microarray technology has been used to study gene and isoform expression in many studies, but its sensitivity for detection of structurally similar isoforms is low, and known or predicted annotations are required to design probes [21]. Although the RNA-Seq method can provide a reliable estimate of exon expression compared to microarrays [44] and is not constrained to the same limitations, studying isoform expressions using RNA-Seq is still not straightforward because of the short read lengths. Reads from current Illumina technology are generally not long enough to span across all exons in an isoform. In most cases, only exons in close proximity are covered by the same read, which makes it difficult to accurately predict a full structure of the isoform. In addition, some genes are fused due to overlapping untranslated regions (UTRs), which can also result in erroneous predicted isoform structures.

Due to those issues, it is not feasible to accurately estimate expression of isoforms, especially when gene annotation is constructed from *de novo* assembly [65]. To avoid these issues, we chose to study exon expression instead of isoform expression. Using MISO with the exon-centric method, only reads spanning across a few exons are used and only exons involved in a splicing event are examined. The expression of exon inclusion is calculated as Percent Spliced In (Psi or Ψ), which can be used to infer the portion of transcripts that include the exon in each sample [22]. In this study, we investigated the three most common alternative splicing events in vertebrates, which are skipped exons (SE), an alternative $3'$ (A3SS) and $5'$ (A5SS) splice site. Lists of DEU genes from the resistant line that show difference in Ψ greater than 0.20 when compared to the susceptible line in infected chickens are shown in Tables 4.3, 4.4 and 4.5. Genes can be categorized roughly into four groups based on the pattern of Ψ across control and infected birds in both lines.

Group I (Table 4.3) includes genes with Ψs that were up- or down-regulated in infected chickens in the resistant line only. This group includes *BCL11B* (B-cell CLL/lymphoma 11B zinc finger proteins), a B-cell lymphoma associated C2H2-type zinc finger protein encoding gene, which functions as a tumor-suppressor for T-cell lymphoma in human. According to homologous alignments on the UCSC genome browser, a splice form with the skipped exon is similar to mouse *BCL11B isoform b*. The skipped exon was expressed 30% in the infected chickens from the resistant line;

whereas it was rarely expressed (4-7%) in the control resistant line and both groups in the susceptible line. The skipped exon was not found to encode any known protein domain, however, it is in the middle of two adjacent C2H2-type finger protein domains. *GEMIN6* plays a role in the assembly of spliceosomal snRNP in cytoplasm. *SRSF6* (SR splicing factor 6) encodes a nuclear protein that belongs to the splicing factor protein family.

In group II (Table 4.4), Ψ values were relatively stable in control and infected chickens within line, but not between lines. Genes that could play an important role in immune responses are *RAC3*, *HCK*, and *ITGB2*. *RAC3* (Ras-related C3 botulinum toxin This gene encodes small GTPases, belonging to the Ras family, that regulate a wide variety of cellular events including cell growth, cytoskeletal reorganization, and the activation of protein kinases. The role of small GTPases in immune responses is discussed further below. *HCK* transmits signals from cell surface receptors such as *FCGR1A, FCGR2A, IL2, IL6, IL18*, and integrins (*ITGB1, ITGB2*). *ITGB2* (CD18) encodes subunit β_2 integrin of *LFA-1* and *CR3* receptors. *LFA-1* plays an important role in adhesion of lymphocytes with other cells. *CR3* binds to a vast array of ligands and molecules including complement C3bi, microbial proteins, ICAM-1 and -2, ECM proteins, and coagulation proteins. It plays a significant role in neutrophil and monocyte activation including phagocytosis, adhesion and migration. The role of *ITGB2* in immune responses is discussed further in the next section. *DYNLL2, SEPT11* and *PFN2* are also involved in cell rearrangement and cytokinesis. In particular, *DYNLL2* is a dynein protein that have been demonstrated to regulate T cell activation by driving T cell receptor microclusters (TCR-MCs) toward the center of an immune synapse [16].

Group III (Table 4.5) includes genes that exhibit differential isoform expression only in response to the infection in susceptible line. A number of genes in this group encode proteins that are parts of spliceosome: *SRSF3*, *HNRNPDL*, *SFSWAP*, *THOC1*, *RNPC3* and *SRSF5*. *PODXL* encodes PODX-like proteins that function in an integrin-dependent manner as both pro-adhesive and anti-adhesive molecules. This protein is involved in cell-to-cell contact, cell trafficking, and cancer progression [42, 62].

The last group (Group IV, Table 4.5) only has three genes: *GOSR1*, *SRSF6* and *ENSGAL00000026498*.

The Ψ value differences of these genes were greater than 0.20 the cutoff between control and infected chickens in the resistant and susceptible lines and were significantly different between infected chickens in the resistant and susceptible lines. *GOSR2* encodes a trafficking membrane protein important for transporting proteins from the *cis*- to the *trans*-golgi network and *SRSF6* (serine/arginine-rich splicing factor 6) encodes a protein involved in mRNA splicing.

Table 4.3: **DEU between the resistant line and the susceptible line in infected birds, group I**

			Resistant (Ψ)		Susceptible (Ψ)	
Type	Ensembl	Symbol	Un	Inf	Un	Inf
SE	ENSGALG00000011127	BCL11B	0.07	**0.30**	0.06	0.04
SE	ENSGALG00000013137	INO80C	0.15	**0.35**	0.95	0.86
A5SS	ENSGALG00000013821	GEMIN6	0.84	**0.61**	0.81	0.85
A5SS	ENSGALG00000009824	C7H2ORF77	0.49	**0.26**	0.68	0.62
A5SS	ENSGALG00000002144	THRAP3	0.31	**0.51**	0.28	0.18
A3SS	ENSGALG00000020987	ZDHHC7	0.42	**0.23**	0.57	0.55
A3SS	ENSGALG00000005685	KSR1	0.77	**0.44**	0.72	0.65
A3SS	ENSGALG00000027665	SYNGR1	0.46	**0.23**	0.68	0.60
A3SS	ENSG00000163875*	MEAF6	0.28	**0.57**	0.40	0.29

*Human homologs, Un=uninfected, Inf=infected, SE=skipped exon, A5SS=5′ splice site, A3SS=3′ splice site. Bold face indicates that there is a SNP between the lines 6 and 7 within an alternative exon.

Table 4.4: **DEU between the resistant line and the susceptible line in infected birds, group II**

			Resistant (Ψ)		Susceptible (Ψ)	
Type	Ensembl	Symbol	Un	Inf	Un	Inf
SE	ENSGALG00000005522	DYNLL2	**0.01**	**0.02**	0.20	0.25
SE	ENSGALG00000004971	URM1	**0.07**	**0.03**	0.18	0.23
SE	ENSGALG00000015709	TACC3	**0.87**	**0.93**	0.77	0.72
SE	ENSGALG00000014642	LOC374195	**0.60**	**0.57**	0.70	0.80
SE	ENSGALG00000011682	CNOT4	**0.57**	**0.62**	0.40	0.41
SE	ENSGALG00000007511	ITGB2	**0.17**	**0.22**	0.02	0.01
SE	ENSGALG00000006522	HCK	**0.47**	**0.59**	0.99	0.97
SE	ENSGALG00000000904	C11H16ORF57	**0.92**	**0.98**	0.84	0.78
A5SS	ENSGALG00000010836	AHR	**0.97**	**0.99**	0.63	0.59
A5SS	ENSGALG00000011488	CMTM7	**0.55**	**0.67**	0.37	0.41
A3SS	ENSGALG00000008939	FUBP1	**0.42**	**0.26**	0.59	0.54
A3SS	ENSGALG00000008507	THOC2	**0.52**	**0.53**	0.69	0.78
A3SS	ENSGALG00000002859	RAC3	**0.69**	**0.84**	0.67	0.61
A3SS	ENSGALG00000012050	TNRC6B	**0.57**	**0.39**	0.94	0.93
A3SS	ENSGALG00000010410	PFN2	**0.71**	**0.78**	0.53	0.50
A3SS	ENSGALG00000027908	LOC422528	**0.28**	**0.39**	0.13	0.09
A3SS	ENSGALG00000011476	SEPT11	**0.78**	**0.86**	0.60	0.58

*Human homologs, Un=uninfected, Inf=infected, SE=skipped exon, A5SS=5′ splice site, A3SS=3′ splice site. Bold face indicates that there is a SNP between the lines 6 and 7 within an alternative exon.

Table 4.5: **DEU between the resistant line and the susceptible line in infected birds, group III and IV**

			Resistant (Ψ)		Susceptible (Ψ)	
Type	Ensembl	Symbol	Un	Inf	Un	Inf
SE	ENSGALG00000003861	HERC4	0.31	0.37	0.45	**0.06**
SE	ENSGALG00000009029	TSPAN12	0.08	0.15	0.20	**0.47**
SE	ENSGALG00000009520	MARCH1	0.42	0.54	0.66	**0.34**
SE	ENSGALG00000008320	EDEM1	0.95	0.99	0.90	**0.72**
SE	ENSGALG00000000533	SRSF3	0.36	0.38	0.30	**0.16**
SE	ENSGALG00000023199	HNRPDL	0.39	0.40	0.30	**0.18**
SE	ENSGALG00000006157	DDX26B	0.67	0.60	0.57	**0.84**
SE	ENSGALG00000001745	PSTPIP2	0.07	0.05	0.11	**0.26**
SE	ENSG00000175029*	CTBP2	0.38	0.38	0.23	**0.12**
A5SS	ENSGALG00000008038	SF3B1	0.41	0.57	0.55	**0.31**
A5SS	ENSGALG00000002487	SFSWAP	0.58	0.73	0.55	**0.41**
A3SS	ENSGALG00000005162	RNPC3	0.55	0.36	0.42	**0.67**
A3SS	ENSGALG00000000720	LOC419563	0.88	0.94	0.85	**0.70**
A3SS	ENSGALG00000009421	SRSF5	0.55	0.72	0.54	**0.39**
A3SS	ENSGALG00000014915	THOC1	0.33	0.48	0.33	**0.23**
A3SS	ENSGALG00000000189	YTHDC2	0.44	0.59	0.42	**0.32**
SE	ENSGALG00000001107	GOSR2	0.73	**0.92**	0.38	**0.59**
SE	ENSG00000124193*	SRSF6	0.43	**0.71**	0.54	**0.34**
A3SS	ENSGALG00000026498	Unknown	0.12	**0.70**	0.10	**0.34**

*Human homologs, Un=uninfected, Inf=infected, SE=skipped exon, A5SS=5′ splice site, A3SS=3′ splice site. Bold face indicates that there is a SNP between the lines 6 and 7 within an alternative exon.

4.2.5 Roles of *LFA-1* and actin cytoskeleton in T cells activation

By grouping genes based on patterns of Ψs, we found that many genes in group II: *ITGB2, PFN2, DYNLL2, SEPT11*, and *RAC3*, are involved in cytokinesis or cell synapse, which are important for T cell activation. As described above, *ITGB2* encodes the β-subunit of integrins including LFA-1, which is exclusively expressed in lymphocytes and plays a major role in lymphoproliferation, antigen presentation, T cell activation, and cytotoxicity. Integrins are special kind of receptors that transmit signals bidirectionally across the cell membrane. They are heterodimeric composed of an α (large) and a β (small) subunit [74]. The β_2 (CD18) subunit encoded by *ITGB2* is expressed on lymphocytes and antigen presenting cells (APCs) as a component of *LFA-1* and *CR3* receptors. LFA-1 binds to its ligand ICAM-1 to help form a synapse that brings APCs and T cells together to initiate antigen presentation leading to T cell activation [12].

Absence of LFA-1 leads to impaired functions of lymphocytes in proliferation and tumor rejection [55, 56]. Mutations in the *ITGB2* gene have been associated with type 1 leukocyte adhesion deficiency (LAD-1), an autosomal-recessive inherited disease found in a few families. The disease is characterized by impairment of lymphocytes in adherent-dependent functions, lack of accumulation to the site of infection and recurrent bacterial and fungal infection [63]. In addition, the response of lymphocytes to mitogens is decreased in patients with LAD [63]. The decrease in responsiveness to mitogens has been shown to correlate with resistance to MD by Lee and Bacon [28], who illustrated that resistant birds (MD resistant lines 6 and N) were less responsive to phytohemagglutinin (PHA) than MD susceptible birds (line 7 and P).

The actin cytoskeleton is very important in T cell activation because it enhances the activity of LFA-1 by increasing its avidity and recruiting signaling molecules necessary for downstream signaling [12, 70]. Cytoskeleton proteins binding to cytoplasmic domain of LFA-1 are thought to play an important role in driving LFA-1 to aggregate on the cell surface, resulting in increased avidity. Aggregation of LFA-1 has been demonstrated to be essential for lymphocytes to bind to the ligand [71]. Interestingly, *RAC3* and *PFN2*, which are involved in the actin cytoskeleton pathway (Figure 4.3), also expressed different ratios of alternative splice forms between lines . These gene products are also found in three other pathways that are involved in immune responses (Table 4.6). It could be speculated that pre-mRNA splicing of these genes is co-regulated by splicing regulators or some genetic factors.

Table 4.6: **Pathways containing *RAC3, ITGB2* and *PFN2***

Pathway ID	Description	Gene
hsa04810	Regulation of actin cytoskeleton	*RAC3, ITGB2, PFN2*
hsa04015	RAP1 signaling pathway	*RAC3, ITGB2, PFN2*
hsa04650	Natural killer cells cytotoxicity	*RAC3, ITGB2*
hsa05416	Viral myocarditis	*RAC3, ITGB2*

4.2.6 Prediction of functional domains of splice forms of genes in the actin cytoskeleton pathway

To predict the function of the alternative splice forms of genes in the actin cytoskeleton pathway, transcript sequences were translated to protein sequences by ESTscan [19]. Protein sequences were then searched for annotated protein domains using the standalone version of InterPro Scan [50]. Besides *ITGB2*, other genes have alternative exons located in coding regions that could potentially affect functional protein domains in some ways. The exon with an alternative 3′ splice site of *RAC3* encodes part of a protein domain identified as a small GTPase of the Ras subfamily (ProSiteProfiles:PS5142 and SMART:SM00173). Rac3 is highly homologous to Rac1 and has been reported to possess the ability to promote membrane ruffling, transformation, activation of c-Jun transcriptional activity and co-activation of NFκB [76]. Activated Rac also regulates production of superoxide in neutrophils and macrophages.

The alternative exon of *PFN2* seems to disrupt the coding sequence that encodes the profilin domain (Pfam:PF00235). The profilin domain is essential for almost all organisms and its functions include regulating actin polymerization, controlling complex networks of molecular interaction and transmitting signals from small-GTPase pathways. It also binds to Rac effector molecules and a number of other ligands [77].

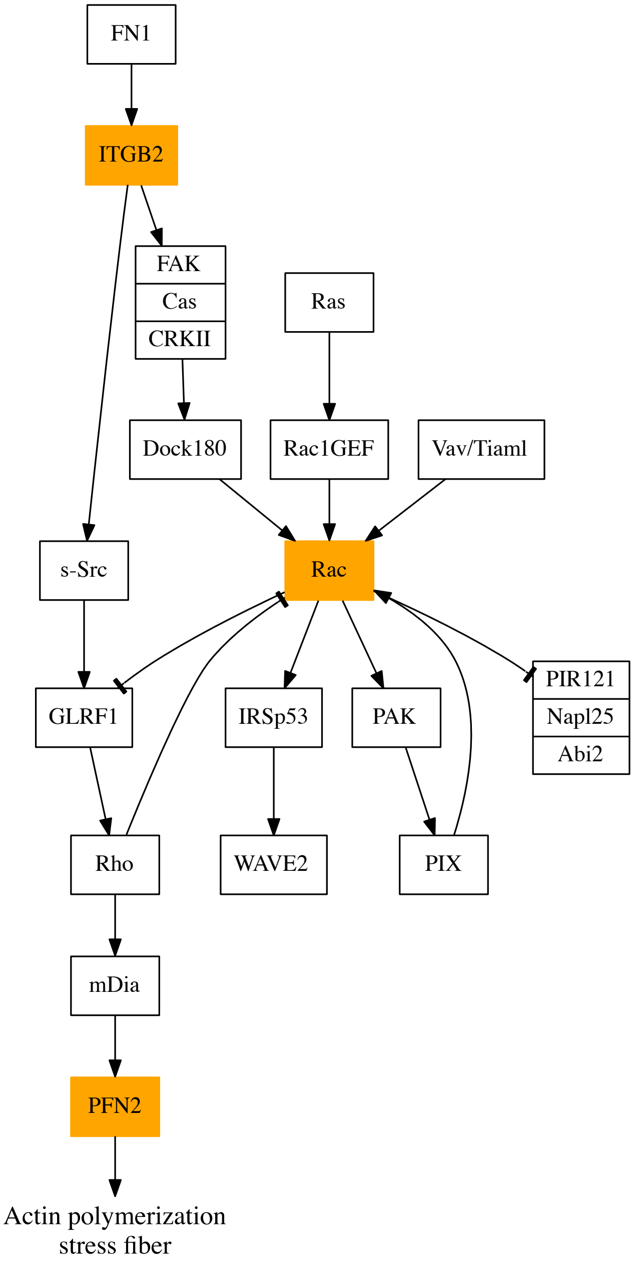

Figure 4.3: **Human regulation of actin cytosekeleton pathway.** An exon of *ITGB2, RAC* and *PFN2* is not differentially expressed between the control and the infected groups, but between resistant and susceptible chickens. These three genes indirectly interact with one another in the pathway. Only genes that interact with these genes are shown in this figure.

Even though the exact mechanism is not known, lack of responsiveness of T cells to stimuli appears to benefit resistant birds because in these birds MDV can not induce T cells to proliferate and cause them to undergo neoplastic transformation. It has also been suggested that the mechanism that controls both lymphocyte proliferation induced by MDV and lymphocyte proliferation induced by the immune response is the same [45]. Therefore, it may be useful to consider a link between the deficiency of lymphocytes in the resistant line to the alternative splice form of *ITGB2* that is only expressed in the resistant line. Although the exon included in the alternative splice form is non-coding, it could serve important functions in translation or posttranscriptional regulation.

4.2.7 Prediction of *cis*-regulatory elements in alternative splicing exons of genes in group II

Among all groups, alternative splicing of genes in the group II is most likely to be regulated by genetic factors because the ratios of isoform expression in this group were relatively stable within line, but were significantly different between lines. Investigation of nucleotide differences within exons of both lines could reveal a possible role of SNPs in regulating alternative splicing in this group. We obtained a sequence of alternative exons from the resistant line and used Human Splicing Finder (HSF) to determine whether SNPs from the susceptible line could alter predicted ESEs or ESSs. Results from some genes involved in cytokinesis are discussed in this section. Exonic SNPs from the resistant and susceptible lines are listed in Table 4.7.

Table 4.7: **SNP between resistant and susceptible lines found in an exon of ITGB2, PFN2 and DYNNL2 (Group II)**

Gene	Chromosome	Position	Reference	Resistant	Susceptible	Strand
ITGB2	7	7183696	C	*T*	.	-
PFN2	9	23221934	- -	.	*AA*	+
DYNLL2	19	8694149	G	.	*A*	-

For *ITGB2*, a SNP (T) at position 26 of the cDNA from the resistant line, which corresponds to position 7,183,696 on chromosome 7 is located in a predicted binding site for SC35, which is an exon enhancer. Although the exon is not expressed in the susceptible line, we found that there is no polymorphisms between lines according to SNP data from genome resequencing. Therefore, this SNP may not account for exclusion of the exon in the susceptible line (Figure 4.4). Exon sequences of *PFN2* from the lines 6 and 7 differ at position 23,221,934 on chromosome 9. A small insertion of two AA nucleotides is predicted to create a new binding site for Tra2-β splicing regulator, which serves as a stabilizer of an enhancer complex [35] (Table 4.8). From exon expression data, Ψ of an exon with alternative splicing increases from 0.20-0.30 to about 0.50 in the susceptible line. Tra2 could possibly increase inclusion of the exon with alternative $3\prime$ splice site via ESE-dependent $3\prime$ splice site activation (Figure 4.5). Even though the linear distance between Tra2-β binding site and the alternative $3\prime$ splice site is greater than 1kb, Tra2 could possibly get close to the $3\prime$ splice site in the secondary structure of the mRNA.

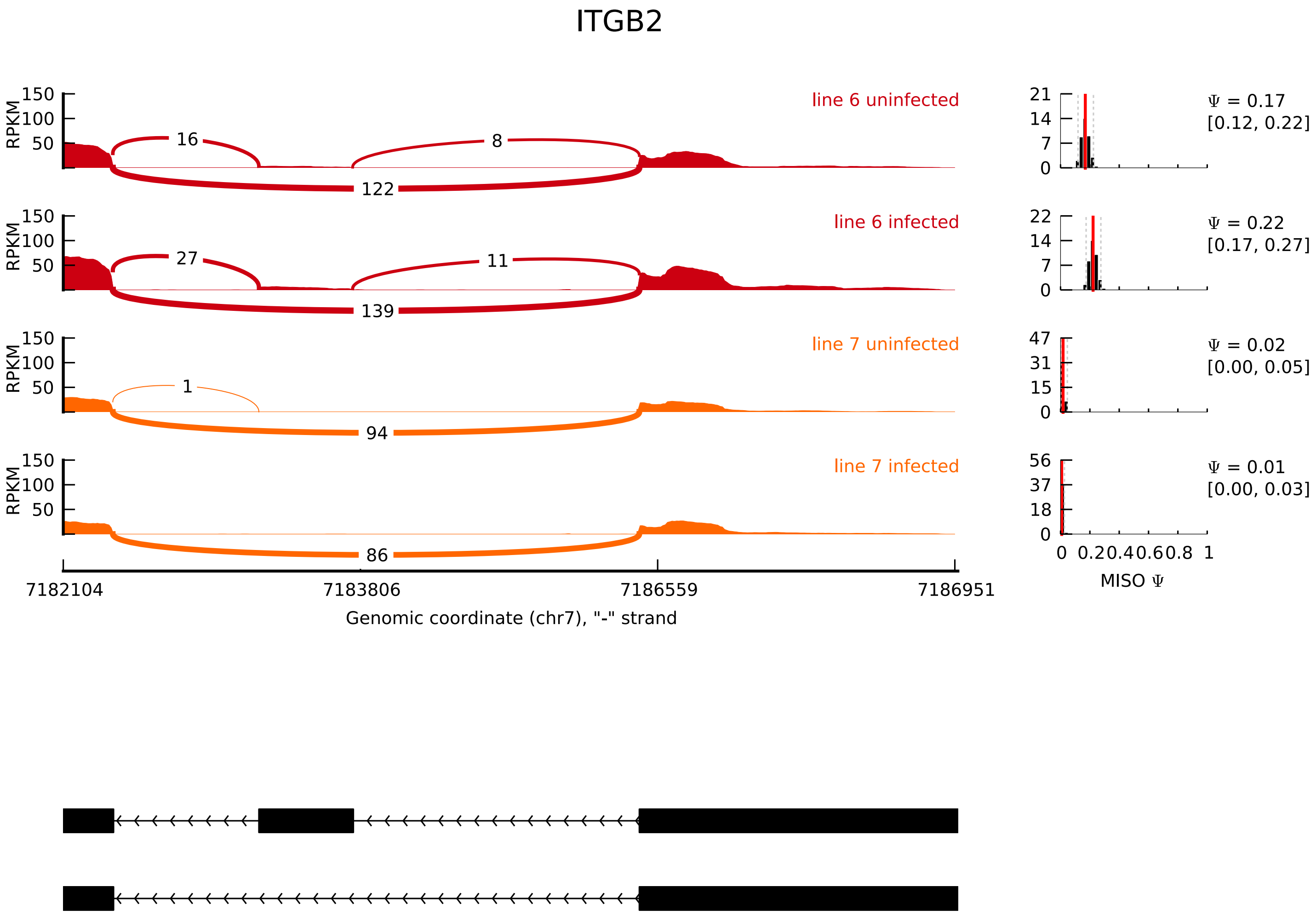

Figure 4.4: **ITGB2 exon expression.** The cassette exon is predicted to be skipped in line 7 but expressed ∼20% in line 6. The histogram depicts read coverage and the curve lines depict spliced reads. The distribution of Ψ values are plotted on the right side with confidence intervals in the square brackets. At the bottom of the plot, black think horizontal lines depict exons and thin lines with arrow heads depict introns.

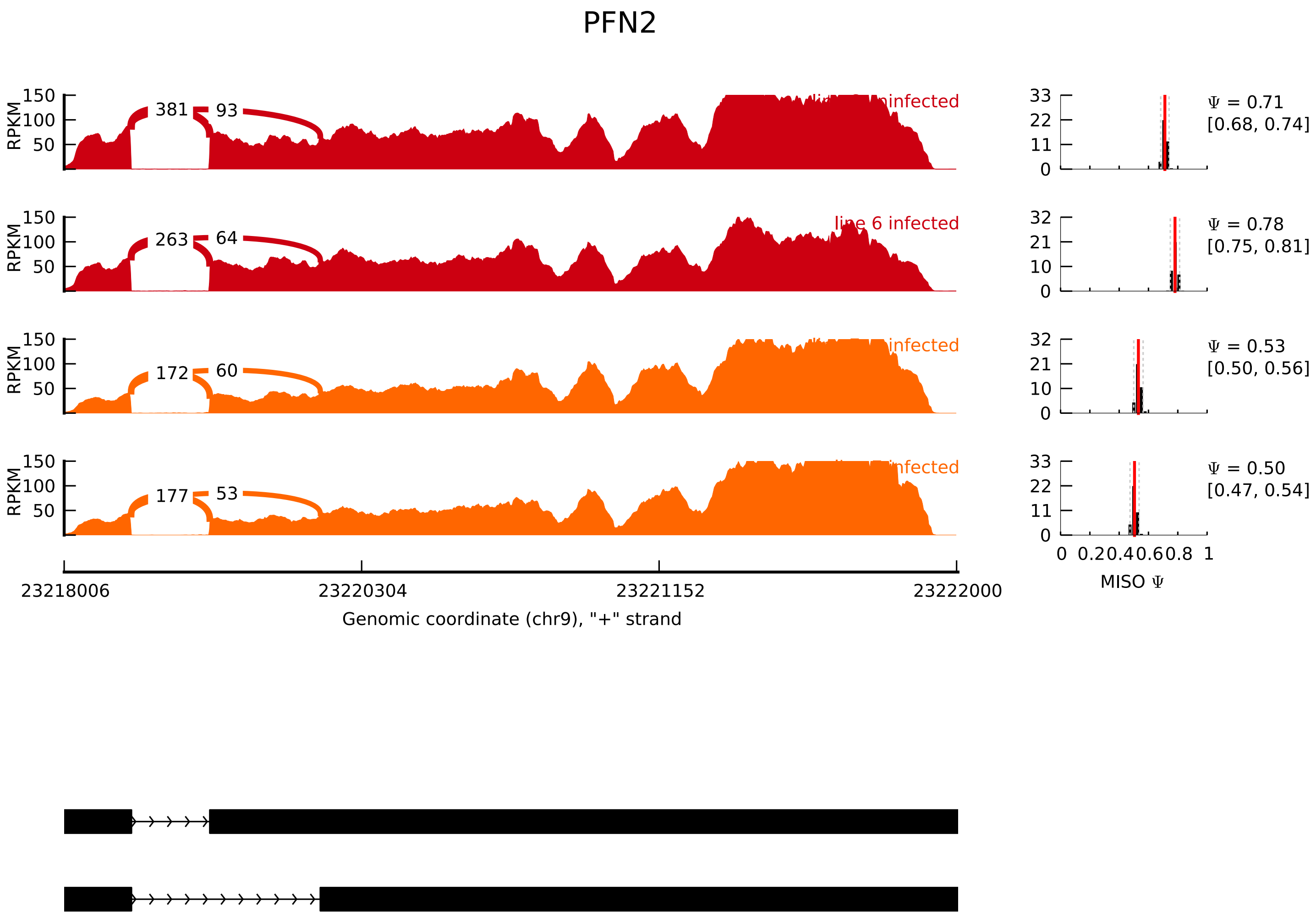

Figure 4.5: **PFN2 exon expression.** The long exon is predicted to be more expressed in line 6 (control and infected group) than in line 7. The histogram depicts read coverage and the curve lines depict spliced reads. The distribution of Ψ values are plotted on the right side with confidence intervals in the square brackets. At the bottom of the plot, black think horizontal lines depict exons and thin lines with arrow heads depict introns.

Replacement of an A with a G nucleotide in the skipped exon of *DYNLL2* from line 6 is predicted to slightly alter the binding site of several ESEs as well as to create a new binding site for 9G8 (Table 4.8). This exon is upregulated in the susceptible line compared to the resistant line, therefore, the presence of the new binding site for 9G8 exon enhancer helps support the expression results. In addition, the G nucleotide in this position matches the reference nucleotide, therefore, we could expect this exon to be expressed in other datasets. According to EST tags on the UCSC genome browser, the exon has been found and sequenced from chicken eyes (15d post-hatched, EST sequence:DR424100).

Table 4.8: **esults from human splicing finder**

Gene	cDNA Position	Linked SR protein	Type	Reference Motif	Mutant Motif	Variation
DYNLL2	2	SF2/ASF (IgM-BRCA1)	ESE[1]	CTCCGGG (86.38)	CTCCGAG (72.69)	-15.85%
	2	SF2/ASF, SF2/ASF (IgM-BRCA1)	ESE[1]	CTCCGGG (79.91)	CTCCGAG (72.69)	-9.03%
	4	SF2/ASF (IgM-BRCA1)	ESE[1]	CCGGGGT (73.00)	CCGAGGT (86.23)	18.12%
	4	SF2/ASF (IgM-BRCA1), SF2/ASF	ESE[1]	CCGGGGT (73.00)	CCGAGGT (82.94)	13.61%
	6	9G8	ESE[2]		GAGGTG (60.67)	New site
	6	hnRNP A1	ESS[4]		GAGGTG (74.05)	New site
PFN2	2068	Tra2-β	ESE[1]	AAAAT (81.02)	AAAAa	+16.19%
	2069	Tra2-β	ESE[1]		AAAaa (94.14)	New site
	2070	Tra2-β	ESE[1]		AAAaaT (81.02)	New site
	2066		ESS[3]		ACAAAAaa (38.13)	New site
	2067		ESS[3]		CAAAAaaT (28.85)	New site

[1] ESE Finder matrices for SRp40, SC35, SF2/ASF and SRp55 proteins. [2] ESE motifs from HSF. [3] Predicted PESS Octamers from Zhang & Chasin. [4] hnRNP motif.

There are too many SNPs in the exon of *SEPT11*, making it unfeasible to predict which SNP might regulate the exon expression. Therefore, we do not discuss these two genes in this section. However, results from HSF analysis of these two genes and other genes in this group are provided in the supplementary materials. Experimental validation of exonic SNPs provided by this study could shed some light on underlying polymorphisms that contribute to resistance to MD.

4.3 Conclusion

Custom gene models built from combination of gene models from *de novo* assembly, reference-based assembly, and Ensembl have allowed us to identify genes and isoforms that might play an important role in resistance to MD. Results from gene expression analysis indicated that adaptive immune responses were more highly activated during lytic infection in the susceptible line, than in the resistant line. Because only activated T cells are thought to be infected by MDV, we speculate that enhancement of adaptive immune responses could help spread the viruses by recruiting and activating more T cells. In contrast, the delay or reduction of adaptive immune responses could benefit the host by limiting infection of activated T cells.

To elucidate the molecular mechanism of MD resistance, we investigated differential isoforms expression between lines and identified a number of genes that could be responsible for difference in immune responses. Notably, this incudes several genes involved in actin cytoskeleton structure and cytokinesis, which are important for the functions of lymphocytes and immune cells but have not been of great interest in the field of MD research. Even though we mainly discuss the possible role of *ITGB2* in MD resistance, other genes cannot be precluded and should be a candidate for further investigation and experimental studies. Moreover, the full mechanism of MD resistance is highly complex and more data from different stages of infection as well as greater sequencing depth will be required to identify all genes and isoforms involved. To enhance the study of unannotated gene and isoform expression, our approach of constructing gene models from RNA-Seq should be iteratively used to extend the Ensembl data to construct more complete gene models.

4.4 Materials and methods

4.4.1 Sequences and quality trimming

mRNAs were extracted from spleens of control and infected chickens lines 6 and 7 (4 d.p.i). Sequence libraries were prepared by standard Illumina unstranded single- and paired-end protocols. Library size of the paired-end datasets is approximately 175 bp. Read lengths are 75 bp in both single- and paired-end libraries. Reads were quality trimmed by Condetri 2.1 [60] with quality score cutoff of 30. The first 10 bases were removed due to non-uniform distribution of nucleotides.

4.4.2 Gene model construction

Due to lack of complete gene models for chickens, we employed two methods to construct gene models from RNA-Seq reads. First, short reads were assembled using Velvet/1.2.03 [82] and Oases/0.2.06 [57] to obtain long transcripts. Assembly was done with hash lengths range from 21 to 31 for both local and global assembly (described in Gimme paper). Poly-A tails were trimmed and low complexity transcripts were removed by Seqclean [58]. All transcripts were then aligned to the chicken reference genome (Galgal4, with unplaced contigs and random chromosomes removed) with BLAT [24]. Second, reads were aligned to the reference genome using Tophat/2.0.9 [66] and Ensembl gene model release 73 was used to guide reference-based assembly by Cufflinks2 [68]. Alignments from BLAT and models from Cufflinks were then combined to construct gene models by Gimme (manuscript in preparation).

4.4.3 Differential gene expression analysis and Gene Ontology

To identify DE genes, reads were mapped to transcripts by RSEM v.1.2.7 [30], which is also used to estimate gene expression and identify DE genes. Data from single- and paired-end datasets from the same line were treated as biological replicates. To identify enriched pathways and ontology terms, a list of DE genes was analysed by GOSeq v.1.10.0 based on chicken KEGG annotations. P-values were corrected by Benjamini-Hochberg multiple testing correction. Genes, pathways and

GO terms with corrected P-value < 0.1 were considered significant. Pathview [37] was used to create a KEGG pathway diagram with colors representing relative level of gene expressions.

4.4.4 Differential exon usage analysis

Gene models were converted to alternative splicing models using a Python script. In order to increase sensitivity, read counts from single- and paired-end samples were combined and treated as single-end reads for splicing event analysis with MISO/0.4.9 [22]. Splicing events with Bayes factor > 10 and $\Delta\Psi > 0.20$ were considered significant. Read coverages and Ψ distributions were plotted using Sashimi plot [23].

4.4.5 Variant calling and *in silico* splicing analysis

Variants were called using mpileup command from SAMTools/0.1.18 [32] and BCFTools [2]. Only variants with quality score ≥ 20 were used for mutation analyses. Exon enhancers and suppressors were predicted using the Human Splicing Finder web portal [11]. Human default parameter settings were used in all analyses.

4.4.6 Protein domains search

Transcripts were translated to protein sequences using ESTScan 3.0.3 [19] with chicken HMM matrices built from chicken cDNAs and RefSeq sequences. Protein sequences were searched against InterPro database using InterProScan/5.44.0 [50].

4.4.7 Pipeline and scripts

The pipeline and scripts used in this study are available at

`https://github.com/likit/mdv_rnaseq_paper`.

CHAPTER 5

CONCLUSIONS

5.1 Summary

The goal of the projects discussed in this dissertation is to identify genes and isoforms that contribute to MD resistance using RNA-Seq data. However, available chicken gene models are not suitable for expression analysis because they are not complete. For example, some of them do not include full UTRs and some exons expressed. We have demonstrated that different gene models yield different results in differential expression analysis as well as biological pathway analysis. Therefore, it is important to improve on existing gene models using RNA-Seq data.

We have developed a pipeline that can combine both *de novo* assembly and reference-based assembly to construct gene models that we have shown to recover more splice variants in the dataset than either method alone. We have also described the local assembly technique that is more memory efficient than global assembly (conventional *de novo* assembly) and can recover some unique splice variants not found by conventional assembly.

We used the merged gene models from Cufflinks and *de novo* assembly described above to study gene and isoform expression from RNA-Seq data collected from spleens of resistant and susceptible inbred chicken lines (lines 6 and 7 respectively) at 4 dpi – lytic phase of infection. We found that more genes were differentially expressed in the susceptible line than the resistant line and those DE genes were enriched in genes active in both the innate and the adaptive immune system. In contrast, only pathways involved in the innate immune system were enriched in DE genes within the resistant line. The results indicate that the adaptive immune responses were enhanced in the susceptible line at this stage of infection. The adaptive immune responses involve recruiting and activating T cells, which results in more target T cells for MDV because only activated T cells are thought to be infected by MDV. More infected T cells could lead to a larger number of T cells that transform into T cells lymphoma in the susceptible line.

To investigate further, we used our merged gene models to study isoform expression between susceptible and resistant lines. Splice variants are highly important in the immune system and many genes have been shown to play a role in increasing risk and prognosis of diseases. Several mutations at binding motifs of splicing factors that alter splicing patterns and are associated with diseases have also been characterized. However, reads from RNA-Seq are too short to span across the entire transcripts; therefore, it is difficult to estimate isoform expression accurately. Moreover, we cannot be certain that our gene models include all expressed isoforms due to the method used to build gene models from short reads. For these reasons, we decided to use an exon centric approach, which only compares exon expression. Exon expression can be estimated more accurately because reads can span across exon junctions of an exon of interest and its adjacent exons.

Using this method, we identified many genes that have differential exon expression between susceptible and resistant lines. Intriguingly, we found that some genes have different patterns of exon expression between lines and they were involved in the actin cytoskeleton pathway and some other pathways involved in immune responses. The actin cytoskeleton pathway is important in cell-to-cell contact, cytokinesis, phagocytosis, and antigen presentation which are all important in adaptive immune responses. Based on the results, we hypothesized that splice variants may play a role in limiting or controlling the elicitation of the adaptive immune response in the resistant line, which in turn limiting the spread of the virus to activated T cells. Results from splicing finder prediction also showed that exonic SNPs between susceptible and resistant lines may regulate splicing patterns by disrupting or creating new binding sites for splicing factors. Results from our study, after further validation, could be used to facilitate the marker-assisted selective breeding to develop chickens with genetic resistant to MD.

5.2 Future work

Results from our study have shown that isoforms could play a significant roles in MD resistance. Therefore, to better understand a mechanism that controls MD resistance, especially at an isoform-level, we need to obtain more sequencing data to identify more isoforms involved in immune responses to MDV infection. Ideally, we need to obtain data from multiple time points and with longer read lengths.

However, obtaining more data is relatively inexpensive in regard to time and labor spent on analyzing data. To enhance sensitivity of gene and isoform detection, gene models need to be updated using new data to ensure that the models include all genes and isoforms expressed in the dataset. To our knowledge, there is no available integrated system for building gene models from multiple methods and existing gene models, and identifying alternative exon usage and exonic SNPs, as well as predicting protein domains from exons of interest as described in this dissertation. The whole process is complicated and only practical for adept bioinformaticians. To accommodate biologists, we plan to develop an open protocol for isoform analysis based on the pipeline used in this dissertation from the beginning to the end. Every step in the protocol will be thoroughly documented and the protocol will be available online and can be updated by users. Once established, the protocol can be used by biologists to iteratively update gene models and perform gene and isoform expression analysis from RNA-Seq data from chickens as well as other organisms with low-quality gene annotation.

The local assembly described in this dissertation can recover unique splice variants not found in the global assembly. However, the method relies on read alignment to a reference genome. Therefore, it is dependent on the quality of the genome as well as efficiency of the short read aligner used. In addition, for most non-model organisms, a high-quality reference genome is not available; thus, this method cannot be employed. Even though we have successfully used the method to recover many splice variants, the inherent mechanism of local assembly has not been understood. An understanding of the mechanism could lead to an algorithm that improves sensitivity of splice variants detection from *de novo* assembly and can be used in organisms without a reference genome.

APPENDIX

The appendix contains a glossary of some technical terms used within this disseration.

Alternative Splicing: A mechanism that produces a variety of mRNA variants from a single gene by removing introns and joining exons in different manners. Alternative splicing is tightly controlled and can be stimulated by either internal or external stimuli.

***de novo* Assembly**: A method used to generate a longer transcript or a contig by concatenating overlapping short reads without the use of a reference genome or transcriptome as a guide.

Differential Gene Expression: A statistically significant difference in expression of a gene between samples.

Differential Exon Usage: A statistically significant difference in expression of an exon between samples.

Global Assembly: See *de novo* Assembly.

Indel: Insertion or deletion of nucleotides or amino acids.

K-mer: A subsequence (of length k) from a read obtained from DNA or mRNA sequencing, where k is an integer. An example of k=4 or 4-mers is ACGT.

KEGG Pathway Annotation: A collection of manually curated pathway maps representing connections and interactions of genes and molecules involved in biological pathways, diseases, drugs, and chemical substances.

Local Assembly: A technique used to increase sensitivity of splice variant detection and facilitate assembly. In contrast to global assembly, only reads mapped to a genome are assembled.

Paired-End Reads: A pair of short reads with a specific insert size (a distance between mates) generated by next-generation sequencing technology. Two mates are generally sequenced from opposite directions.

Pipeline (computing): A set of commands and tools used for computational analyses that are usually organized in a specific order or series, where the output of one tool/command is the input of next one.

Polymorphism: A variation of nucleotides or amino acids between two or more DNA, mRNA and protein sequences.

Read Mapping: Alignment of reads to a genome or a reference set of transcripts.

Short Reads: A short stretch of nucleotides (50-400bp) produced by next-generation sequencing technology.

Single-End Reads: A short reads produced by next-generation sequencing technology, usually sequenced from one side of a DNA or an mRNA.

Transcriptome: Total mRNA molecules expressed from tissues or cells.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] L. D. BACON. Influence of the major histocompatability complex on disease resistance and productivity. *Poultry Science*, 66(5):802–811, 1987.

[2] BCFTools. http://samtools.github.io/bcftools/.

[3] J. Beane, J. Vick, F. Schembri, C. Anderlind, A. Gower, J. Campbell, L. Luo, X. H. Zhang, J. Xiao, Y. O. Alekseyev, et al. Characterizing the impact of smoking and lung cancer on the airway transcriptome using rna-seq. *Cancer prevention research*, 4(6):803–817, 2011.

[4] B. J. Blencowe. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in biochemical sciences*, 25(3):106–110, 2000.

[5] W. Briles, R. W. Briles, W. McGibbon, H. Stone, et al. Identification of b alloalleles associated with resistance to marek's disease. In *Resistance and immunity to Marek's disease. A seminar in the EEC Joint Programme of the'Mechanisms of resistance to Marek's disease', held in the Reichstag, West Berlin, 31 October-2 November 1978.*, pages 395–416. Commission of the European Communities., 1980.

[6] N. Bumstead. Genomic mapping of resistance to marek's disease. *Avian Pathology*, 27(S1):S78–S81, 1998.

[7] D. W. Burt. Chicken genome: current status and future opportunities. *Genome research*, 15(12):1692–1698, 2005.

[8] M. Carlson. *org.Gg.eg.db: Genome wide annotation for Chicken.* R package version 2.14.0.

[9] M. Carlson. *org.Hs.eg.db: Genome wide annotation for Human.* R package version 2.14.0.

[10] B. Collard, M. Jahufer, J. Brouwer, and E. Pang. An introduction to markers, quantitative trait loci (qtl) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica*, 142(1-2):169–196, 2005.

[11] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud. Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*, 37(9):e67–e67, 2009.

[12] M. L. Dustin and J. A. Cooper. The immunological synapse and the actin cytoskeleton: molecular hardware for t cell signaling. *Nature immunology*, 1(1):23–29, 2000.

[13] I. M. Gimeno. Marek's disease vaccines: a solution for today but a worry for tomorrow? *Vaccine*, 26:C31–C41, 2008.

[14] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, July 2011.

[15] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5):503–510, May 2010.

[16] A. Hashimoto-Tane, T. Yokosuka, K. Sakata-Sogawa, M. Sakuma, C. Ishihara, M. Tokunaga, and T. Saito. Dynein-driven transport of t cell receptor microclusters regulates immune synapse formation and t cell activation. *Immunity*, 34(6):919–931, 2011.

[17] A. Hogenkamp, N. Isohadouten, S. S. Reemers, R. A. Romijn, W. Hemrika, M. R. White, B. Tefsen, L. Vervelde, M. van Eijk, E. J. Veldhuizen, et al. Chicken lung lectin is a functional c-type lectin and inhibits haemagglutination by influenza a virus. *Veterinary microbiology*, 130(1):37–46, 2008.

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

[19] C. Iseli, C. V. Jongeneel, and P. Bucher. Estscan: a program for detecting, evaluating, and reconstructing potential coding regions in est sequences. In *ISMB*, volume 99, pages 138–148, 1999.

[20] P. Kaiser, G. Underwood, and F. Davison. Differential cytokine responses following marek's disease virus infection of chickens differing in resistance to marek's disease. *Journal of virology*, 77(1):762–768, 2003.

[21] M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic acids research*, 28(22):4552–4557, 2000.

[22] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Publishing Group*, 7(12):1009–1015, Nov. 2010.

[23] Y. Katz, E. T. Wang, J. Silterra, S. Schwartz, B. Wong, J. P. Mesirov, E. M. Airoldi, and C. B. Burge. Sashimi plots: Quantitative visualization of RNA sequencing read alignments. *Audio and Electroacoustics Newsletter, IEEE*, pages –, June 2013.

[24] W. Kent. BLAT—The BLAST-Like Alignment Tool. *Genome research*, 2002.

[25] P. S. Kingma and J. A. Whitsett. In defense of the lung: surfactant protein a and surfactant protein d. *Current opinion in pharmacology*, 6(3):277–283, 2006.

[26] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[27] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[28] L. F. LEE and L. D. BACON. Ontogeny and line differences in the mitogenic response of chicken lymphocytes. *Poultry science*, 62(4):579–584, 1983.

[29] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

[30] B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[31] C. Li, Y. Zhang, R. Wang, J. Lu, S. Nandi, S. Mohanty, J. Terhune, Z. Liu, and E. Peatman. Rna-seq analysis of mucosal immune responses reveals signatures of intestinal barrier disruption and pathogen entry following *Edwardsiella ictaluri* infection in channel catfish *Ictalurus punctatus*. *Fish & shellfish immunology*, 32(5):816–827, 2012.

[32] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

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

[34] H.-C. Liu, H.-J. Kung, J. E. Fulton, R. W. Morgan, and H. H. Cheng. Growth hormone interacts with the marek's disease virus sorf2 protein and is associated with disease resistance in chicken. *Proceedings of the National Academy of Sciences*, 98(16):9203–9208, 2001.

[35] A. J. Lopez. Alternative splicing of pre-mrna: developmental consequences and mechanisms of regulation. *Annual review of genetics*, 32(1):279–305, 1998.

[36] T. Lu, G. Lu, D. Fan, C. Zhu, W. Li, Q. Zhao, Q. Feng, Y. Zhao, Y. Guo, W. Li, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by rna-seq. *Genome research*, 20(9):1238–1249, 2010.

[37] W. Luo and C. Brouwer. Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

[38] K. W. Lynch. Consequences of regulated pre-mrna splicing in the immune system. *Nature Reviews Immunology*, 4(12):931–940, 2004.

[39] R. W. Morgan, L. Sofer, A. S. Anderson, E. L. Bernberg, J. Cui, and J. Burnside. Induction of host gene expression following infection of chicken embryo fibroblasts with oncogenic marek's disease virus. *Journal of virology*, 75(1):533–539, 2001.

[40] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, 2004.

[41] C. Morrow and F. Fehler. Marek's disease: a worldwide problem. *Marek's disease, An Evolving Problem*, pages 49–61, 2004.

[42] J. S. Nielsen and K. M. McNagny. The role of podocalyxin in health and disease. *Journal of the American Society of Nephrology*, 20(8):1669–1676, 2009.

[43] T. D. Otto, D. Wilinski, S. Assefa, T. M. Keane, L. R. Sarry, U. Böhme, J. Lemieux, B. Barrell, A. Pain, M. Berriman, et al. New insights into the blood-stage transcriptome of plasmodium falciparum using rna-seq. *Molecular microbiology*, 76(1):12–24, 2010.

[44] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 2008.

[45] F. Pazderka, B. M. Longenecker, G. R. Law, H. A. Stone, and R. F. Ruth. Histocompatibility of chicken populations selected for resistance to marek's disease. *Immunogenetics*, 2(1):93–100, 1975.

[46] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[47] J. K. Pickrell, A. A. Pai, Y. Gilad, and J. K. Pritchard. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet*, 6(12):e1001236, 2010.

[48] L. Preeyanon. Gimme: A lightweight reference-guided transcripts assembler: http://github.com/likit/gimme.

[49] Pysam. A python module for manipulating samfiles: http://github.com/pysam-developers/pysam.

[50] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic acids research*, 33(suppl 2):W116–W120, 2005.

[51] N. Raghavachari, J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, J. Christopher, P. J. Munson, and G. J. Kato. A systematic comparison and evaluation of high density exon arrays and rna-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC medical genomics*, 5(1):28, 2012.

[52] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[53] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of RNA-seq data. *Nature methods*, 7(11):909–912, Oct. 2010.

[54] A. Sarson, P. Parvizi, D. Lepp, M. Quinton, and S. Sharif. Transcriptional analysis of host responses to marek's disease virus infection in genetically resistant and susceptible chickens. *Animal genetics*, 39(3):232–240, 2008.

[55] K. Scharffetter-Kochanek, H. Lu, K. Norman, N. Van Nood, F. Munoz, S. Grabbe, M. McArthur, I. Lorenzo, S. Kaplan, K. Ley, et al. Spontaneous skin ulceration and defective t cell function in cd18 null mice. *The Journal of experimental medicine*, 188(1):119–131, 1998.

[56] R. Schmits, T. Kündig, D. M. Baker, G. Shumaker, J. Simard, G. Duncan, A. Wakeham, A. Shahinian, A. Van Der Heiden, M. F. Bachmann, et al. Lfa-1-deficient mice show normal ctl responses to virus but fail to reject immunogenic tumor. *The Journal of experimental medicine*, 183(4):1415–1426, 1996.

[57] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, Feb. 2012.

[58] 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: http://compbio.dfci.harvard.edu/tgi/software/.

[59] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, June 2009.

[60] L. Smeds and A. Künstner. Condetri-a content dependent read trimmer for illumina data. *PloS one*, 6(10):e26314, 2011.

[61] J. Smith, J.-R. Sadeyen, I. R. Paton, P. M. Hocking, N. Salmon, M. Fife, V. Nair, D. W. Burt, and P. Kaiser. Systems analysis of immune responses in marek's disease virus-infected chickens identifies a gene involved in susceptibility and highlights a possible novel pathogenicity mechanism. *Journal of virology*, 85(21):11146–11158, 2011.

[62] A. Somasiri, J. S. Nielsen, N. Makretsov, M. L. McCoy, L. Prentice, C. B. Gilks, S. K. Chia, K. A. Gelmon, D. B. Kershaw, D. G. Huntsman, et al. Overexpression of the anti-adhesin podocalyxin is an independent predictor of breast cancer progression. *Cancer research*, 64(15):5068–5073, 2004.

[63] T. A. Springer, M. L. Dustin, T. K. Kishimoto, and S. D. Marlin. The lymphocyte function associated lfa-1, cd2, and lfa-3 molecules: cell adhesion receptors of the immune system. *Annual review of immunology*, 5(1):223–252, 1987.

[64] T. Tatusova. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences - Tatusova - 2006 - FEMS Microbiology Letters - Wiley Online Library. *FEMS microbiology letters*, 1999.

[65] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2013.

[66] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111, Apr. 2009.

[67] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012.

[68] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010.

[69] R. L. Vallejo, L. D. Bacon, H.-C. Liu, R. L. Witter, M. A. Groenen, J. Hillel, and H. H. Cheng. Genetic mapping of quantitative trait loci affecting susceptibility to marek's disease virus induced tumors in f2 intercross chickens. *Genetics*, 148(1):349–360, 1998.

[70] Y. van Kooyk and C. G. Figdor. Avidity regulation of integrins: the driving force in leukocyte adhesion. *Current opinion in cell biology*, 12(5):542–547, 2000.

[71] Y. Van Kooyk, P. Weder, K. Heije, and C. G. Figdor. Extracellular ca2+ modulates leukocyte function-associated antigen-1 cell surface distribution on t lymphocytes and consequently affects cell adhesion. *The Journal of cell biology*, 124(6):1061–1070, 1994.

[72] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov. 2008.

[73] G.-S. Wang and T. A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, 2007.

[74] H. Wang, D. Lim, and C. E. Rudd. Immunopathologies linked to integrin signalling. In *Seminars in immunopathology*, volume 32, pages 173–182. Springer, 2010.

[75] L. Wang, X. Wang, X. Wang, Y. Liang, and X. Zhang. Observations on novel splice junctions from RNA sequencing data. *Biochemical and biophysical research communications*, 409(2):299–303, June 2011.

[76] S. Werbajh, I. Nojek, R. Lanz, and M. A. Costas. Rac-3 is a nf-κb coactivator. *FEBS letters*, 485(2):195–199, 2000.

[77] W. Witke. The role of profilin complexes in cell motility and other cellular processes. *Trends in cell biology*, 14(8):461–469, 2004.

[78] R. Witter. Control strategies for marek's disease: a perspective for the future. *Poultry science*, 77(8):1197–1203, 1998.

[79] N. Yonash, L. Bacon, R. Witter, and H. Cheng. High resolution mapping and identification of new quantitative trait loci (qtl) affecting susceptibility to marek's disease. *Animal genetics*, 30(2):126–135, 1999.

[80] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Method gene ontology analysis for rna-seq: accounting for selection bias. *PMC free article][PubMed]*, 2010.

[81] N. D. Young. A cautiously optimistic vision for marker-assisted breeding. *Molecular breeding*, 5(6):505–510, 1999.

[82] D. Zerbino. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 2008.

[83] D. R. Zerbino, G. K. McEwen, E. H. Margulies, and E. Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PloS one*, 4(12):e8407, Jan. 2009.

[84] Z. Zhang, S. Pal, Y. Bi, J. Tchou, and R. V. Davuluri. Isoform level expression profiles provide better cancer signatures than gene level expression profiles. *Genome medicine*, 5(4):33, 2013.