Annual Report

Likit Preeyanon

Research Summary

Introduction

I proposed to study alternative splicing in chicken line 6 and 7 that contribute to resistance and susceptibility to Marek's disease. The hypothesis is *"Alternative isoforms and SNPs that alter a splicing pattern contribute to genetic resistance or susceptibility to MD"*. However, chicken gene annotations are not of good quality and not complete. Furthermore, chicken genome sequence, even the latest version still contains many errors from misassembly. Therefore, methods developed for RNA-Seq analysis that require a high quality genome sequence and gene annotations cannot be used directly for this study.

To overcome these obstacles, I have been developing a computational pipeline to construct gene models from RNA-Seq data. The pipeline includes two assembly methods as well as in-house software that I have developed. The pipeline is shown in Fig.1.

Because of a poor quality of chicken genome sequence, a lot of work has been done to minimize the false positive results. Based on observation, false positive splicing predictions are caused by genome as well as mRNA missassembly. Therefore, each step in the pipeline as well as the software is tuned to reduce errors and also increase sensitivity. Many methods have been used to evaluate the quality of the gene models built from the pipeline as discussed further.

Gene Annotation Pipeline

The first step of the pipeline is to obtain transcript sequence from short-read assembly. The assembly is done by Velvet[**?**] and Oases[**?**] assembler for both global and local assembly methods. The difference between global and local assembly method is that in local aseembly, only reads mapped to chickgen genome are assembled and reads mapped to each chromosome are assembled separately. In contrast, all reads are assembled at once in global assembly. I found that the combination of both methods help improve sensitivity of splice junction detection. Moreover, hash lengths[1] ranging

[1] a number of overlapped nucleotides required by Velvet assembler

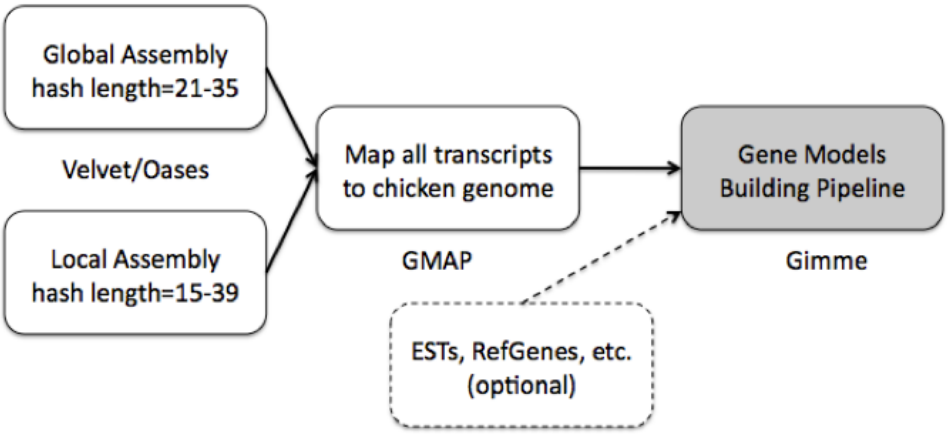

Figure 1: **Gene model construction pipeline.** Transcripts are obtained from two assembly methods – global and local assembly. Transcripts are aligned to a chicken genome by GMAP. Gimme then constructs gene models based on alignments of transcripts.

from 21-31 and 15-39 are used for global and local assembly to enhance sensitivity. Transcripts are then mapped to chicken genome using GMAP[**?**].

The next step is to feed alignments from both method to in-house program, called *"Gimme"*, to construct the gene models. The program will first filter out transcripts that do not meet the criteria. As a default, transcripts containing more than one very small exon (shorter than 40bp) are excluded. Very small exons tend to be a result of misalignment. Table 1 shows a comparison of the number of transcripts excluded from the program from different datasets. Transcripts from global and local assembly are from single-end datasets from chicken line 6 and 7, pre- and post infection. As shown in Table 1, a majority of full-length transcripts such as chicken reference mRNAs are not excluded from the program. On the other hand, fragmented sequences such as expressed sequence tags (ESTs) are greatly affected by the criteria. The results show that transcripts from global and local assembly are more complete than ESTs and the criteria does not filter out high quality alignments.

Table 1: **Alignment Summary**

Dataset	Total	Included	Excluded
Chicken mRNA	27,965	25,136	2,829 (10%)
Chicken EST	559,062	289,105	269,957 (48%)
Global assembly	1,044,023	883,286	160,737 (15%)
Local assembly	2,070,547	1,654,657	415,890 (20%)

The program constructs gene models using information from alignments. Basically, transcripts sharing common introns are grouped together. Intron and exon graphs are built from exon junctions. Then the program traverses through all possible paths to construct all possible isoforms of each gene. Incomplete exons at the end of transcript are not included in the gene model. (Exon 3a and 3b in Fig.2)

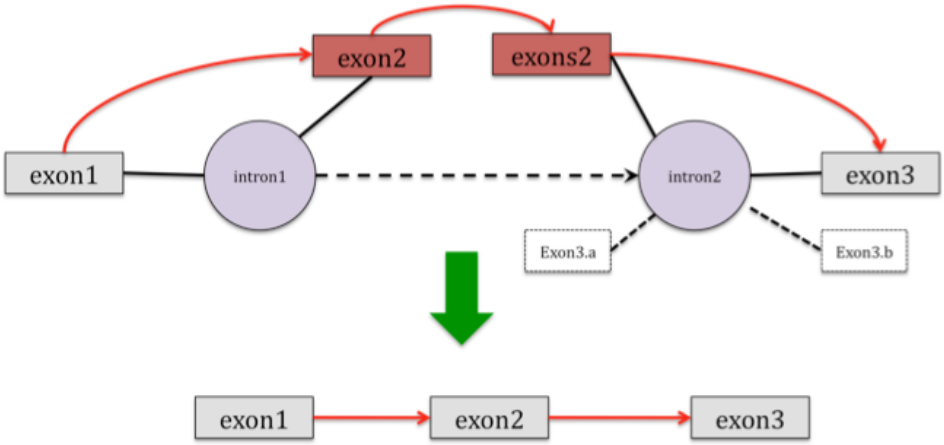

Figure 2: **Intron and exon graphs.** Introns sharing at least one exon are grouped together. Then an exon graph is made using exons as nodes.

A large number of isoforms is obtained from the program; however, not all of them are actually expressed. Furthermore, we do not need to know all isoforms that are expressed in chicken to identify alternative splicings in downstream analysis. Therefore, a set of all possible transcripts are reduced to only a set of minimal transcripts that contains all splice junctions and untranslated regions (UTRs). Then, transcripts that are more than 95% similar are grouped together using ESTScan[**?**]. Only a representative of each group is kept. Table 2 shows the number of genes, maximum and minimum isoforms from each dataset after removing similar sequences. This step helps remove false exon junctions from errors in genome sequence.

Gene Models Validation

The simplest method to validate the gene models is to map sequencing reads back to the transcripts derived from gene models. A large number of reads is expected to map to transcripts if the gene models are high quality. In addition, results from mapping will also be used to evaluate splice junctions found by the pipeline. The results show that approximately 65-74% of reads are mapped to the gene models, indicating that the gene models are good.

Table 2: **Number of putative genes and isoforms**

Method	Gene	Isoform			
		Maximum	Minimum	Multiexon*	Two-exon*
Global	18,017	27,780	24,403	17,428	6,975 (26%)
Local	17,126	22,264	21,002	15,342	5,660 (26%)
Global + Local	18,256	29,592	25,405	18,430	6,975 (27%)
Cufflinks	27,886	33,850	19,038	16,320	2,718 (14%)

**Minumum set*

Those unmapped reads can be reads from poly-A tail, sequencing error as well as reads from missing parts of the genome.

Table 3: **Single-end reads mapped to gene models**

Dataset	Mapped	Unmapped
Line 6 uninfected	21,262,978 (70.64%)	8,839,234 (29.36%)
Line 6 infected	18,655,749 (64.85%)	10,110,835 (35.15%)
Line 7 uninfected	18,796,057 (68.06%)	8,822,732 (31.94%)
Line 7 infected	21,409,427 (72.10%)	8,284,227 (27.90%)

To evaluate splice junctions detected by the pipeline, the number of reads mapped across each splice junction is obtained. Splice junctions with more than four mapped reads are more likely to be genuine junctions. As shown in Fig.2, more than 80% of splice junctions are crossed by more than four reads. This suggests that a majority of splice junctions are genuine.

To investigate further, splice junctions detected by the pipeline are compared with another dataset and method. Of total 112,089 junctions detected by the pipeline, about 61% (68,905) are supported by ESTs and 83% are supported by Cufflinks[**?**]. On the other hand, the pipeline detects about 81% (93,756) of junctions detected by Cufflinks. The results suggest that the pipeline can detect a majority of exon junctions detected by another method in addition to novel junctions.

Based on an assumption that most isoforms are translated to functional protein sequences, a large number of translatable transcripts are expected from gene models. To evaluate this, transcripts are translated to protein sequences using ESTScan, which is designed to work with ESTs and can deal with indel and frameshift well. Of total 22,884 minimal transcripts, 90% (25,405) transcripts are translatable (longer than 50aa). To investigate further, all protein sequences are BLASTed[**?**] against mouse protein sequences. Bitscore to length ratios are then calculated for each sequence and plotted in Fig.3. A majority of isoforms have a ratio greater than 1.0, which indicates a good alignment. This result help confirm that the protein

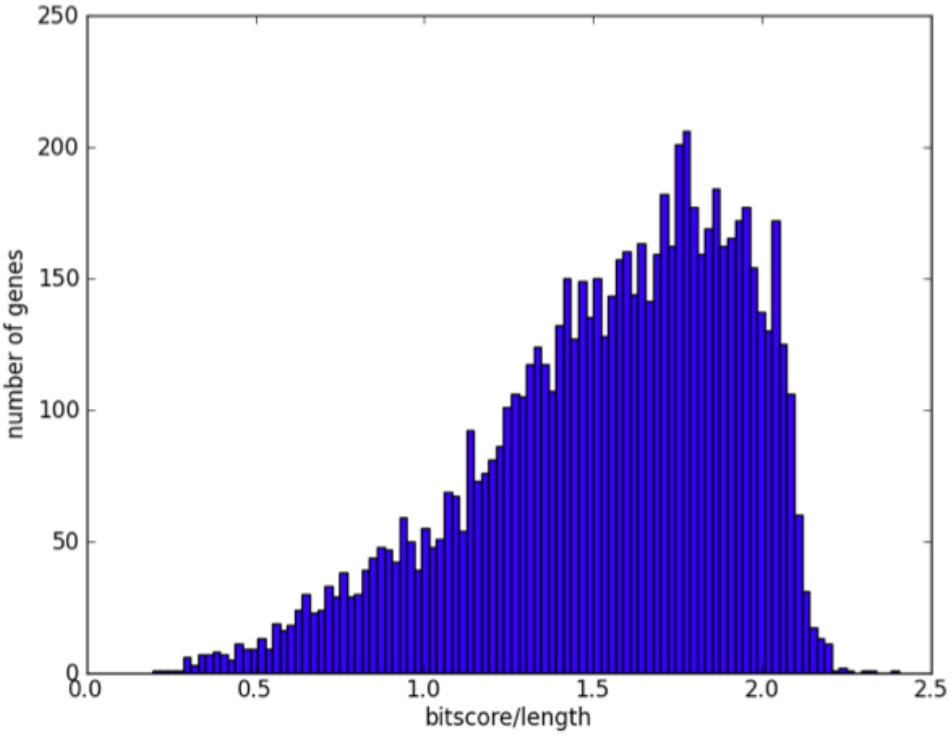

Figure 3: **Histogram of bit score/length ratio of isoforms that match mouse proteins.**

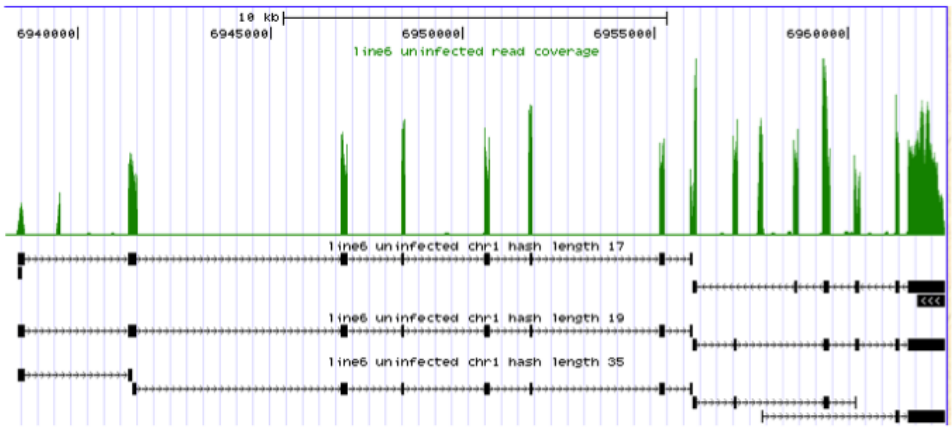

Figure 4: **Reads mapped across exon junctions.**

sequences are translatable and might be functional.

Future work

Gene models obtained from the pipeline will be used to search for differentially expressed genes between line 6 and 7 as well as pre- and post infection using a software such as DEGSeq[**?**]. Differentially expressed exons, which indicate alternative splicing, will also be identified using DEXSeq[unpublished]. Then, a role of genes found from both analyses will be predicted from pathway analysis and gene ontology. To understand how alternative splicing contributes to resistance or susceptibility of MD, protein structure of differentially expressed isoforms will be predicted using SMART[**?**]. Alteration of functional protein domains will suggest a potential gain or loss of function.

Prospective Publication

First Author Paper

Title: RNASeq Assembly Detects Many Alternative Splicing

This paper reports the pipeline for building gene models from RNA-Seq data using assembly method. It also includes validations of gene models and comparison of the pipeline to other existing softwares. It also demonstrates that the pipeline detects many novel/unannotated genes and isoforms.

The results in this paper have been presented as a poster titled: *"Alternative Splicing Detection Using RNA-Seq data: An Assembly Approach"* at the following conferences:

Plant and Animal Genome (PAG) XX, January 13-18, 2012 at San Diego, CA.

International Society for Computational Biology (ISCB), December 8-10, 2011 at Aspen, CO.

Great Lake Bioinformatics Conference (GLBIO), May 2-4, 2011, Athens, OH.

The manuscript is now in preparation. It is more than 80% complete and is expected be submitted by the end of July 2012.

Title: Roles of Transcriptome Variation to Resistance of Marek's disease

This paper reports results from RNAseq analysis including differential gene expresseion, alternative splicing as well as allele specific expression between chicken line 6 and 7. The paper also reports candidate genes that contribute to resistance/susceptibility to MD. Predicted function of isoforms/genes from protein sequences as well as pathway analysis are included in this paper. Furthermore, causative single nucleotide polymorphisms are also identified.

Data required for the analysis is now available. Analysing data and writing should take approximately 5-6 months. I plan to submit the paper

by January 2012.

Title: Digital Normalization with RNA-Seq Assembly

This paper shows comprehensive comparisons of isoforms detected by raw data and digital normalized data. RNA sequences from chicken and lamprey will be used to illustrate the effeciency of the methods.

Preliminary results for this paper should be done within 2 months. Depending on the prelimanary result, a manuscript may be submitted by January 2013.

Others

Title: Roles of Meq Protein in Marek's disease

I have been involved in analysing Chip-seq and microarray data for this project in Hans Cheng's laboratory. The paper is planned to be submitted by the end of this year.

Graduation Timeline

- July 2012 First paper submitted.
- January 2013 Second and third paper submitted.
- May 2013 Dissertation defense.
- June 2013 Disssertation submitted.