論文の内容の要旨

Abstract of Dissertation

論文題目 **A global approach for identification and quantification of splicing variants using RNA-Seq data**

（RNAシーケンシングデータからのスプライシング変異産物の検出と定量に関する包括的アプローチ）

氏　名　桑　飞

Name of Author SANG FEI

Abstract

Alternative Splicing (AS) plays a major role in generating the diversity of transcriptome and proteome, and affects development and disease by producing structurally and functionally distinct mRNA and protein variants. Transcripts from approximately 95% of multi-exon human genes are spliced in more than one way, and in most cases the resulting transcripts are variably expressed between different cell and tissue types. The second generation sequencing technology is now opening unprecedented routes to address the analysis of entire transcriptome. However, accurate quantification of the abundances of transcripts still remains a big challenge. We have developed a global approach that allows the identification and quantification of splicing variants derived solely from the read count and the base count in RNA-Seq data. It is based on an explicit statistical model and enables the identification of splicing variants using the known gene annotation or the assembled transcript structure information, as well as the relative quantification. To test our method, we generated and analyzed 120 million paired 75-bp artificial mouse RNA-Seq reads from about 18,000 various transcripts.

Analysis Pipeline

Our analysis can be summarized into 4 main phases: Mapping Phase, Counting Phase, Identifying Phase and Quantifying Phase. In Mapping Phase, the align tools, eg. Tophat, are applied to map the short RNA-Seq reads to a genome reference to generate ordinary reads and splice reads; In Counting Phase, two different calculations are used to count the number of mapped reads and the number of mapped bases for a given location respectively. Moreover, in order to simply all the alternative splicing events into exon skipping, a special subexon matrix was generated. If an exon appears in two or more isoforms at different lengths, we divide this exon into several different parts and treat each part as a subexon. Then according to this subexon matrix the structure of each transcript is identified and described, which means if this transcript contains a subexon, the value of this subexon in the matrix can be set to 1, otherwise to 0. The count calculation is preformed depending on the subexon matrix. During the counting phase, the base count is calculated separately for ordinary reads and splice reads in order to maintain the valuable splicing information which is of much importance to estimate isoform expression level; In Identifying Phase, two strict assumptions help us filter noisy incorrectly expressed transcripts caused by mapping errors or incomplete express transcripts due to low read coverage. One is “If one exon was not expressed, the whole transcript was not expressed.” The other one is “If one junction between two exons was not expressed, the whole transcript was not expressed.” Perhaps these two assumptions are too strict to succeed capturing the low expressed transcripts; In Quantifying Phase, we developed a novel global approach, a probabilistic model that uses information of both normal

reads and splice reads in RNA-Seq data to perform more accurate estimations of transcript expressions.

Results

Analysis with a known gene reference

FluxSimulator totally selected 15,333 genes and 18,143 isoforms for mouse 20M dataset, 15,475 genes and 18,330 isoforms for mouse 40M dataset, 15,524 genes and 18,411 isoforms for mosue 60M dataset, 13,531 genes and 18,425 isoforms for human 60M dataset. The total number of mouse expressed transcripts selected by FluxSimulator did not rise along with the increasing sequenced reads, which indicated that the expression of each transcript grew up. Totally we identified 7,279 genes and 8,090 isoforms in mouse 20M dataset, 8,014 genes and 8,928 isoforms in mouse 40M dataset, 8,399 genes and 9,452 isoforms in mouse 60M dataset respectively, as well as 5,308 genes and 6,535 isoforms in human 60M dataset. Cufflinks identified 11,602 genes and 12,448 isoforms in mouse 20M dataset, 12,401 genes and 13,317 isoforms in mouse 40M dataset, 12,660 genes and 13,579 isoforms in mouse 60M dataset respectively, as well as 10,496 genes and 11,946 isoforms in human 60M dataset.

		Mouse		**Human**
Dataset	**20M**	**40M**	**60M**	**60M**
Simulated Genes	15,333	15,475	15,524	13,531
Simulated Transcripts	18,143	18,330	18,411	18,425
Cufflinks Genes	11,602	12,401	12,660	10,496
Cufflinks Transcripts	12,448	13,317	13,579	11,946
Identified Genes	7,279	8,014	8,399	5,308
Identified Transcripts	8,090	8,928	9,452	6,535

Table 1. Summary of simulated results and estimated results

The comparison results of mouse 60M dataset reported that most cufflinks results were underestimated the expression level of isoforms, with only 325 accurate predictions under the error rate below 10%. In addition, 6,176 non-expressed transcripts were incorrectly predicted by Cufflinks giving estimated expression values, while 1,276 false positive results existed in Cufflinks results. Our results showed higher correlation value than cufflinks, which means the results were more similar with the true expression levels. 9,316 non-expressed transcripts were incorrectly predicted by our method. Although more false negative estimations existed in our results due to the strict assumptions in the Identifying Phase, only 406 false positive estimations are included in our results. Furthermore 4,515 out of 9,049 transcripts were properly estimated under the error rate below 10%, 49.8% accuracy, much better than Cufflinks. However Cufflinks captured more transcripts with a few mapped reads than our method. Genes with multiple isoforms were specially derived from the mouse 60M results to access the expression ratio estimations, 8,126 isoforms obtained from Cufflinks results, 4,699 isoforms from our results. Cufflinks had 4,125 (about 51%) well-estimated results with error rate below 10% while 4,150 (about 88%) by our method. Besides, our expression level estimations were also more accurate with 0.814 correlation value than Cufflinks estimations with 0.320 correlation value which contained some well-estimated expression percentage but with inaccurate expression levels. For example, Rasgrf1 had one long isoform and one short isoform. Both our method and Cufflinks gave the accurate expression percentage estimation, however the expression levels that Cufflinks produced were much lower than our estimations and true expression levels.

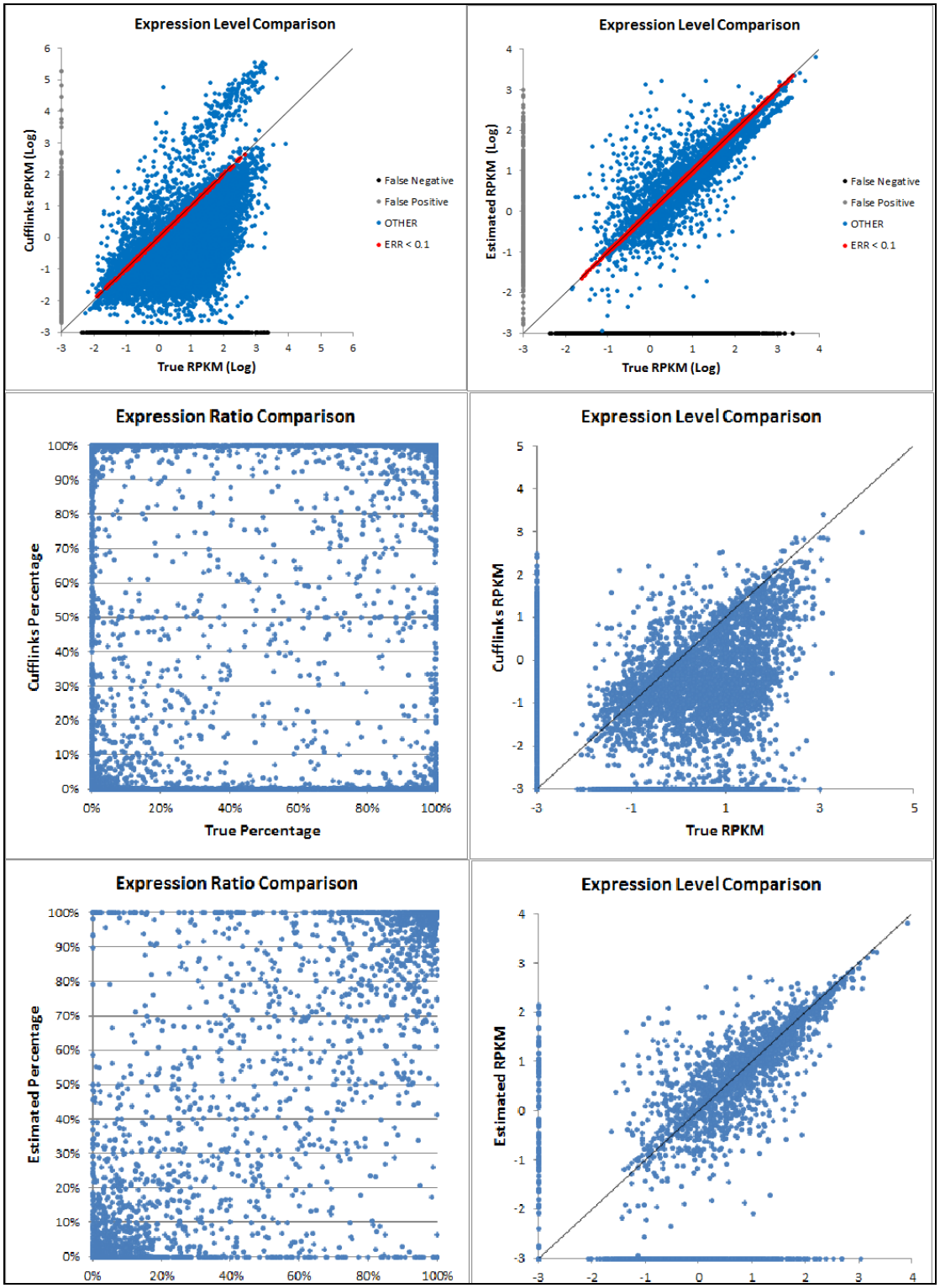

Figure 1. Expression level comparison between estimated values calculated by Cufflinks and our method and true values calculated by simulated expressions when using a known gene reference. Epression level raito between estimated values calculated by Cufflinks and our method and true values calculated by simulated expressions when using a known gene reference.

Analysis with a de novo assembled reference

In order to compare with Cufflinks results, we also used Trinity, a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data, to help assemble short RNA-Seq reads. First, Trinity was applied to perform assembly; Then, those assembled transcripts were mapping to the genome reference by Blat; At last, the identification and quantification were analyzed by our global approach. Here only 60M dataset were analyzed. In the first assembly step, Trinity generated more transcripts (23,781) than the simulated transcripts (18,793), indicating the results of Trinity included much more false positive data than Cufflinks which reconstructed 12,058 transcripts. After filtering the transcripts which were not included in the gene annotation reference, Cufflinks only kept 2,139 transcripts while Trinity kept 7,204. Their distributions of expression levels were almost similar with the true expressions. Still most cufflinks results were underestimated the expression level of isoforms, with only 132 accurate predictions under the error rate below 10%. 5,231 false negative estimations and 51 false positive estimations remained in the Cufflinks results. Our results showed 3,594 transcripts were properly estimated under the error rate below 10%. 294 false negative estimations and 58 false positive estimations remained in the results. Although Trinity generated much more false positive noises, it captured many isoforms that cufflinks failed to detect, about 5,000 transcripts in mouse 60M dataset. Genes with multiple isoforms were specially derived from the 60M results to access the expression ratio estimations, 485 isoforms obtained from Cufflinks results, 645 isoforms from our results. Cufflinks had 303 (about 62.4%) well-estimated results while 416 (about 64.5%) by our method. The accuracies of percentage estimations were almost the same, however, our method provided more accurate expression level estimations than Cufflinks.

Conclusion and Discussion

Through the comparison between Cufflinks estimations and true expression levels, it was clear that Cufflinks underestimated the expressions of most transcripts. Two possible factors may lead to the underestimation of Cufflinks. One is the mapping problem, which Cufflinks estimation depends on. In this study, we use Tophat as our main mapping tool that was developed by the same group of Cufflinks. Since Cufflinks requires the mapped fragments to estimate the transcript abundance, the case of only one fragment read mapped to the genome reference cannot be used in Cufflinks abundance estimation. As we discuss above, only 36% of total simulated reads in mouse 60M dataset were perfect fragment alignments, which may cause the underestimation of Cufflinks. However, when we only use these perfect fragment alignments to estimate transcript abundance by Cufflinks, the results were still underestimated although the estimations turned better. The other reason that caused the underestimation may be the Poisson Model applied by Cufflinks. However, the authors of Cufflinks did not discuss about how they calculated the fragment count, therefore, we cannot find the real reason of underestimation. In our study, we encouraged to apply the base count instead of the read count. Read Count overestimates the counts of short exon, because the read length is longer than the exon size. Because the subexon matrix was applied in our statistical model, the frequency of short exons became much higher. Base Count played an essential role in our estimation of transcript expression levels. Our results illuminate that our global approach was more accurate than Cufflinks wide-used currently. Our results provide a probabilistic model for RNA-seq analysis, offer more accurate isoform expression estimation and develop a pipeline for studies of gene and isoform expression.