Gene 518 (2013) 164–170

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Identifying differentially spliced genes from two groups of RNA-seq samples

Weichen Wang [a,b,1], Zhiyi Qin [a,1], Zhixing Feng [a], Xi Wang [a,c], Xuegong Zhang [a,d,*]

[a] MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China
[b] Department of Operational Research and Financial Engineering, Princeton University, NJ 08544, USA
[c] School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan NSW 2308, Australia
[d] School of Medicine, Tsinghua University, Beijing 100084, China

ARTICLE INFO

*Article history:*
Accepted 27 November 2012
Available online 8 December 2012

*Keywords:*
Alternative splicing
RNA-Seq
Negative binomial model
Differential splicing
Exon-centric method

ABSTRACT

Recent study revealed that most human genes have alternative splicing and can produce multiple isoforms of transcripts. Differences in the relative abundance of the isoforms of a gene can have significant biological consequences. Identifying genes that are differentially spliced between two groups of RNA-sequencing samples is an important basic task in the study of transcriptomes with next-generation sequencing technology.

We use the negative binomial (NB) distribution to model sequencing reads on exons, and propose a NB-statistic to detect differentially spliced genes between two groups of samples by comparing read counts on all exons. The method opens a new exon-based approach instead of isoform-based approach for the task. It does not require information about isoform composition, nor need the estimation of isoform expression. Experiments on simulated data and real RNA-seq data of human kidney and liver samples illustrated the method's good performance and applicability. It can also detect previously unknown alternative splicing events, and highlight exons that are most likely differentially spliced between the compared samples.

We developed an NB-statistic method that can detect differentially spliced genes between two groups of samples without using a prior knowledge on the annotation of alternative splicing. It does not need to infer isoform structure or to estimate isoform expression. It is a useful method designed for comparing two groups of RNA-seq samples. Besides identifying differentially spliced genes, the method can highlight on the exons that contribute the most to the differential splicing. We developed a software tool called DSGseq for the presented method available at http://bioinfo.au.tsinghua.edu.cn/software/DSGseq.

## 1. Introduction

The rapid development of next-generation-sequencing has made it possible to read nucleotide compositions of DNA or RNA molecules at high throughput and high accuracy. It provides the new technique called RNA sequencing or RNA-seq to study transcriptomes by sequencing RNA molecules instead of hybridizing them to predesigned probes as in the microarray technology. RNA-seq generates tens of millions of short reads with lengths of 25 to over 100 nt randomly sampled from RNA molecules in the sample, providing digital measurements on abundances of all transcripts. An important advantage of RNA-seq is that it can reveal previously un-annotated transcripts and provide information on fine structures of transcripts, especially for alternative splicing genes. Recent RNA-seq studies have updated people's understanding on alternative splicing and revealed that most human genes may have alternative splicing (Pan et al., 2008; Wang et al., 2008).

Comparing transcriptomes of two groups of samples and identifying differentially expressed (DE) genes are basic and important tasks. These have been widely studied with microarrays. The digital nature of RNA-seq brings new characteristics in the gene expression measurement. New methods have been proposed for identifying differentially expressed genes from RNA-seq data, such as DEGseq (Wang et al., 2010), DESeq (Anders and Huber, 2010), baySeq (Hardcastle and Kelly, 2010) and Cufflinks (Trapnell et al., 2010). These methods focused on the overall expression of a gene regardless of its alternative isoforms. As many genes can be alternatively spliced, and different relative abundances of splicing isoforms can have crucial impact on phenotypes (Pan et al., 2008; Wang et al., 2003, 2008), it is in many scenarios more important to study the expression pattern or the relative proportion of multiple isoforms of a gene besides its overall abundance.

In contrast to the rapid development of methods for studying DE genes, there are fewer methods for detecting differentially spliced

*Abbreviations:* DSG, differentially spliced gene; NB, negative binomial; IPV, isoform proportion vector; MSDR, mean sequencing depth ratio; RPKM, reads per kilobase per million reads.

* Corresponding author at: FIT 1-107, Tsinghua University, Beijing, 100084, China. Tel.: +86 10 62794919; fax: +86 10 62773552.

*E-mail addresses:* weichenw@princeton.edu (W. Wang), qzy06@mails.tsinghua.edu.cn (Z. Qin), zhangxg@tsinghua.edu.cn (X. Zhang).

[1] These authors contributed equally to this work.

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.045

(DS) genes, i.e., genes that have different proportions of isoform expression between compared samples. Some RNA-seq data analysis tools like Cufflinks (Trapnell et al., 2010) have included features for detecting differential splicing (called Cuffdiff), but most of them are based on estimating the expression of all isoforms first and then comparing proportions of the isoforms. The inference of isoforms and estimation of isoform expression are still current topics under extensive research (Jiang and Wong, 2009; Richard et al., 2010; Roberts et al., 2011; Wu et al., 2011). Such inferences and estimations may introduce extra inaccuracy or uncertainty for studying DS genes. Actually, the accurate estimation of isoform expression is more demanding than detecting differences in their expression, and therefore it may not be a good strategy to detect differences in splicing based on the inference and estimation of isoform expression. In Stegle et al. (2010), the authors discussed the question and proposed to directly use exons not shared by isoforms to detect different abundance of isoforms. They proposed a non-parametric test for detecting differential read coverage based on a measurement defined in the Reproducing Kernel Hilbert Space. It is a promising framework but depends on known annotation of isoform structures, and the selection of kernels is still a topic of future study (Stegle et al., 2010). Singh et al. (2011) developed a method for detecting differential transcription by constructing Aligned Cumulative Transcript Graphs (ACT-Graph) from mapped exon reads and junction reads and comparing graphs between samples with a flow difference metric (FDM). Simulation experiments showed superior performance over other methods. However, the construction of ACT-Graphs requires information of junction reads. When the sequencing depth is not very high or the read length is short, the proportion of junction reads is small, and therefore many splicing sites may not be covered by junction reads. Besides, junction read mapping is still a challenging task in both the mapping accuracy and computation cost (Wang et al., 2011). These factors can be limitations for such graph-based methods. On the other hand, all existing methods were designed for comparing two individual samples. Due to the intrinsic diversity between biological samples, it's more important to compare two groups of samples to find real signals that discriminate the two groups. Several methods have included the option of comparing samples with replicates. However, this was typically done by the combination of multiple pair-wise comparisons of single samples. The underlying tests were not designed for comparing two groups.

In this paper, we study the question of identifying genes that are differentially spliced between two groups of samples, either of technical replicates or biological replicates. We use "differential splicing" or DS to refer to that the relative abundances of a gene's multiple isoforms are different between samples of the compared groups. As discussed above, for the purpose of inferring differential splicing, we do not necessarily need to estimate the expression of isoforms as differential splicing can be reflected from read distributions across the exons composing the isoforms. In this way, we also don't need to know isoform structures or even the existence of multiple isoforms. We use the negative binomial (NB) distribution to model read-counts on all exons of a gene. It considers over-dispersion in read-count distribution and borrows information across samples to get better estimation of the signal of each exon. An NB-statistic is proposed to assess differential splicing. It takes into account both the between-group variation and within-group variation, similar to Baggerly et al. (2003). The proposed method can not only identify differentially spliced genes, but also detect exons that show biggest differences, i.e., exons that are most likely to be differentially spliced between the compared samples. Previously unknown alternative splicing events can be detected in this way. We also consider the possible position-associated bias in read distribution in many RNA-seq datasets and develop a bias-correction option. The comparison with the popular software Cuffdiff on both simulated and real datasets illustrated the good performance and applicability of the proposed method. The proposed software DSGseq can be accessed at http://bioinfo.au.tsinghua.edu.cn/software/DSGseq.

## 2. Methods

### 2.1. The model

Let $g$ denote a gene to be studied and $K^{(g)}$ be the set of its possible isoforms, and the number of isoforms is $r^{(g)}$. Note that $K^{(g)}$ or $r^{(g)}$ needs not to be known in our method. Assume that gene $g$ has $m^{(g)}$ non-redundant exons, which are segments that cover all exon regions of the gene and don't overlap. This is a mathematical definition of exon, which can be an exon in the biological definition, and can also be a segment of a biological exon. For example, in the case of alternative 5′- or 3′-end events, we define the alternative part of the exon as a non-redundant exon and the remaining part as another non-redundant exon. Fig. 1 shows an example of the mathematical definition of exons in a gene with two isoforms. Long exons can be split into several non-redundant exons if necessary. When considering junction reads covering parts of two exons, we can take the junction part as an extra non-redundant exon. In the remaining part of the paper, we use the term "exon" as defined in this way. We can define a gene structure matrix $A=(a_{fj})$ of dimension $r^{(g)}$ by $m^{(g)}$ to indicate the composition of all isoforms, where $a_{fj}=1$ means that isoform $f$ contains exon $j$, and $a_{fj}=0$ means that exon $j$ is not included in isoform $f$. Again, this annotation is for the purpose of derivation only. The proposed method does not require knowing this matrix.

Let $k_f^{(g)}$ be the copy number of RNA transcripts in the form of isoform $f$, and $l_j^{(g)}$ be the length of exon $j$. We consider the possible sequencing preference of each exon and model it as a multiplier vector $\{\beta_j\}_{j=1}^{m}$. In the following description, we consider only one gene, so we omit the superscript ($g$) in all annotations.

The RNA sequencing procedure can be viewed as a random sampling procedure from the multiple copies of RNA transcripts in the cells. The probability that a read falls in the region of exon $j$ can be written as

$$p_j = \frac{\beta_j l_j \sum_{f\in K} a_{fj} k_f}{\sum_{i=1}^{m} \sum_{f\in K} a_{fi} \beta_i l_i k_f} = \frac{1}{\Xi} \beta_j l_j \sum_{f\in K} a_{fj} k_f. \tag{1}$$

The matrix form for all $m$ exons is $\mathbf{p} = \frac{1}{\Xi} BLA^T k$ where $\mathbf{p}$ and $\mathbf{k}$ are column vectors of $p_j$s and $k_f$s. $\mathbf{B}$ and $\mathbf{L}$ are diagonal matrices with diagonal elements $\{\beta_j\}_{j=1}^{m}$ and $\{l_j\}_{j=1}^{m}$, respectively, and $\Xi$ is a normalization factor. This relationship implies that there is a one-to-one correspondence between $\mathbf{p}$ and $\mathbf{k}$ if identifiability is assumed, i.e., the matrix $\mathbf{A}$ is of full row rank. This assumption requires that there is no isoform that can be composed by the combination of other isoforms. With this assumption, the information of isoform expression is fully reflected in reads at all exons. Checking the current annotation of the human genome, we found that most genes can satisfy this assumption. Therefore, we can detect differences in isoform proportion from the information at all exons without having to estimate the expression of all isoforms. The method we propose is to detect differential splicing between two groups of samples based on the estimated expression probability vector of all exons of the gene.

Let $n$ be the number of samples in the study and $M_i$ is the total number of reads (read-count) received at the studied gene. Let $Y_{ij}$ be the read-count of sample $i$ at exon $j$, $i=1,2,\cdots,n$. Differences in $M_i$ reflect the differential total expression and sequencing coverage of the gene between the compared samples. For studying differential splicing, we are comparing the read count vector $\mathbf{Y}_i=\{Y_{ij}\}$ conditional on $M_i$. That is, we focus on the proportion of isoform expression instead of the overall expression of the whole gene. $Y_{ij}/M_i$ can be taken as an estimate of the $p_j$ in sample $i$. Similar to some methods for studying differential expression of genes (Anders and Huber, 2010), we model the read count by the negative binomial (NB) model to consider over-dispersion in the data:

$$Y_{ij} \sim NB\left(\mu_{ij}, \phi\right), \mu_{ij} = M_i p_j, \quad i = 1, \cdots, n, j = 1, \cdots, m \tag{2}$$