1 Introduction

Alternative splicing plays crucial roles in development and disease and is very important in regulating gene function in higher eukaryotes [1]. The importance of alternative splicing is remarkably highlighted by its ability of generating multiple mRNA and protein isoform from a single gene [2]. Xu et al. first applied expressed sequence tags (ESTs) to detect the tissue-specific exons [3]. However, low throughput and high noise limits the capacity of EST-based analysis for detecting differential alternative splicing [4]. As the development of high-throughput RNA sequencing technology (RNA-seq), it has become feasible to conduct genome-wide quantitative analyses of RNA alternative splicing [5][6]. By comparing the RNA-seq data from two biological conditions, exons with changes in exon inclusion levels could be identified. In previous study, different approaches such as Fisher exact test [7][8] and Bayesian statistics [9][10][11] have been applied to estimate the statistical significance of the differential alternative splicing events.

In this paper [12], in order to test flexible hypothesis of differential alternative splicing patterns on RNA-seq, Shen et al. have developed MATS (multivariate analysis of transcript splicing) based on a Bayesian statistical framework. MATS uses a multivariate uniform prior to model the between sample correlation in exon splicing patterns, and a Markov chain Monte Carlo (MCMC) method coupled with a simulation-based adaptive sampling procedure to calculate the P-value and false discovery rate (FDR) of differential alternative splicing. MATS has several advantages compared to previous methods of detecting differential alternative splicing. First of all, MATS provides the flexibility for using user-defined pattern to identify differential alternative splicing events. Also, the multivariate uniform prior implemented in MATS is more general and better captures the genome-wid similarity in exon splicing patterns between biological conditions. Finally, Markov chain Monte Carlo (MCMC) method coupled with a simulation-based adaptive sampling procedure employed by MATS is applicable to almost any type of null hypotheses of interest.

2 MATS, multivariate analysis of transcript splicing

2.1 Notations

Define the exon inclusion level ($\psi$) of an alternatively spliced exon as the percentage of 'exon inclusion' transcripts among all such 'exon inclusion' transcripts plus 'exon skipping'
transcripts.

Define $N_I, N_S, l_I, l_S$ as (Fig. 1):

- $N_I$: Number of reads inclusion
- $N_S$: Number of reads skipping
- $l_I$: length of inclusion isoform
- $l_S$: length of skipping isoform

Figure 1. Illustration of alternative splicing and notations. The pre-mRNA transcripts could be spliced into inclusion isoforms and skipping isoforms. We denote the $l_I$ as the length of inclusion isoform while $l_S$ as the length of skipping isoform.

### 2.2 Likelihood for $N_I$

$$N_I|\psi \sim Binomial(N_I + N_S, \frac{l_I \psi}{l_I \psi + l_S (1 - \psi)})$$

### 2.3 Example

Consider we have two RNA-seq data with different exon inclusion level $\psi_1$ and $\psi_2$. $c$ represents the user-defined threshold for splicing change. The null and alternative hypotheses are:

- $H_0 : |\psi_1 - \psi_2| \leq c$
- $H_1 : |\psi_1 - \psi_2| > c$

Then, the test statistics are:

$$-2\log\left(\frac{\max(\psi_1, \psi_2) L_0(\psi_1, \psi_2)}{\max(\psi_1, \psi_2) L(\psi_1, \psi_2)}\right) \sim \chi^2_1$$

while,

$L_0(\psi_1, \psi_2)$ is constraint likelihood under $|\psi_1 - \psi_2| \leq c$
$L(ψ_1, ψ_2)$ is unconstraint likelihood

For example, for the gene RLEN, it has different exon inclusion level in brain.

Figure 2. Different alternative spliced form of gene RELN have different inclusion level in brain and muscle.

Based on MATS, we can get the $H_0$, $H_1$ and unconstrained $L$ and constrained $L_0$ as shown in Figure 3.
References


