Lecture 6

Lecturer: Jingyi Jessica Li *Scribe: Wanlu Liu*

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 mulitple 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 nosie limits the capacity of EST-based analysis for detecting differential alternative splicing[4]. As the development of high-throughput RNA sequencing techonology (RNA-seq), it has become feasible to conduct genome-wide quantitative analyses of RNA alternative aplicing [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 (ψ) 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

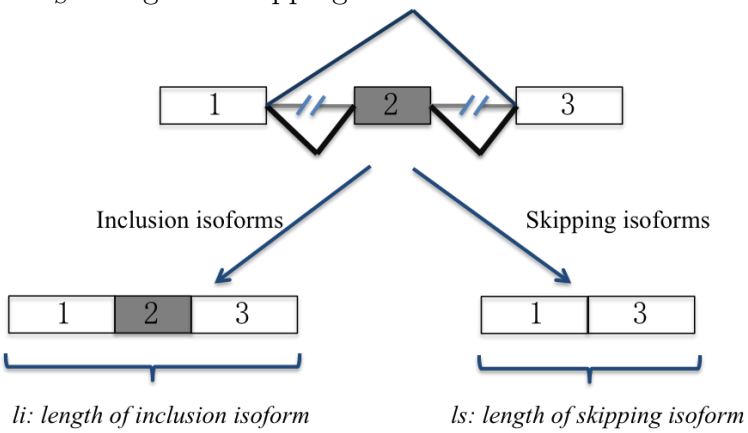


Figure1. Illustration of alternative splicing and notations. The pre-mRNA transcripts could be spliced into inclusion isforms and skipping isoforms. We denote the l_I as the length of inclusion isoform while l_S as the length of skipping isoform.

2.2 Likelihood for N_I

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

2.3 Example

Consider we have two RNA-seq data with different exon inclusion level ψ_1 and ψ_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:

$$-2log\left(\frac{max(\psi_1,\psi_2)L_o(\psi_1,\psi_2)}{max(\psi_1,\psi_2)L(\psi_1,\psi_2)}\right) \sim \chi_1^2$$

while,

$L_o(\psi_1, \psi_2)$ is constraint likelihood under $|\psi_1 - \psi_2| \leq c$

$L(\psi_1, \psi_2)$ is unconstraint likelihood
For example, for the gene RLEN, it has different exon inclusion level in brain(89

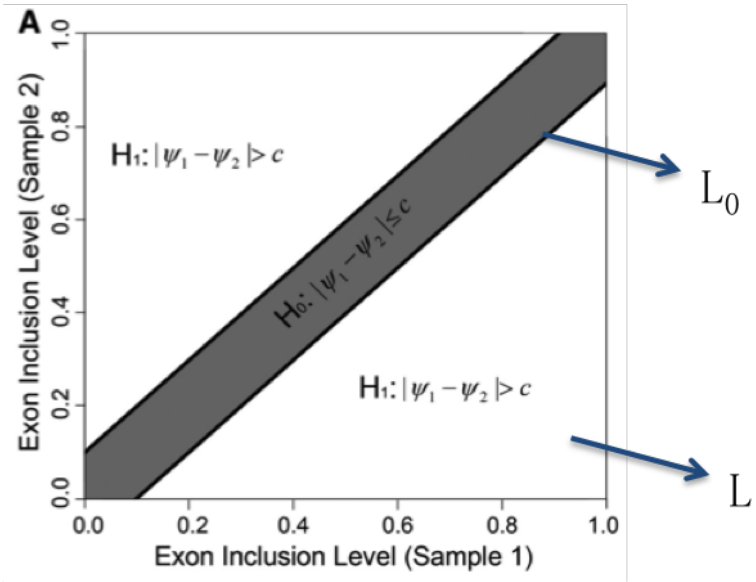
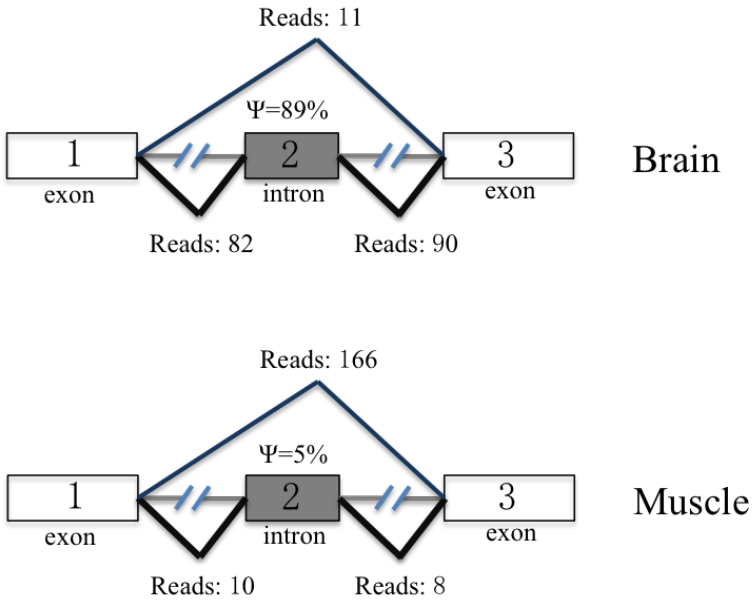


Figure2. Different alternative spliced form of gene RELN have different inclusion level in brain and muscle.

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



References

[1] Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5), 345-355.

[2] Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *TRENDS in Genetics*, 17(2), 100-107.

[3] Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, 30, 3754-3766.

[4] Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics*, 5, 72.

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

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

[7] Griffith,M., Griffith,O.L., Mwenifumbo,J., Goya,R., Morrissy,A.S., Morin,R.D., Corbett,R., Tang,M.J., Hou,Y.C., Pugh,T.J. et al. (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods*, 7, 843-847.

[8] Lalonde,E., Ha,K.C., Wang,Z., Bemmo,A., Kleinman,C.L., Kwan,T., Pastinen,T. and Majewski,J. (2011) RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.*, 21, 545-554.

[9] Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, 31, 5635-5643.

[10] Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation.*Nat. Methods*, 7, 1009-1015.

[11] Xing,Y., Yu,T., Wu,Y.N., Roy,M., Kim,J. and Lee,C. (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.*, 34, 3150-3160.

[12] Shen,S., Park,J.W., Huang,J., Dittmar,K.A., Lu, Z., Zhou,Q., Carstens, R.P. and Xing, Yi. (2012) MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.*, 40(8): e61-e61.