Identification of Differential Alternative Splicing Events with an Adjusted Beta-Distribution Model

Kan Liu[1], Qian Du[1], Guodong Ren[2], Bin Yu[1], Chi Zhang[1]
1. School of Biological Sciences
University of Nebraska
Lincoln, Nebraska, USA
2. State Key Laboratory of Genetic Engineering Institute of Plant Biology,
School of Life Sciences,
Fudan University,
Shanghai, China

***Abstract*— High-throughput next generation sequencing of cDNA, i.e. RNA-Seq, presents an unprecedented resource for characterizing the alternative splicing (AS) in complex eukaryotic transcriptomes. Accumulating evidences indicate that AS is developmentally regulated, but the precise responses of AS event to development is not well understood. Here, we describe a new method, based on an adjusted beta-distribution model, for detection of differential AS patterns from RNA-Seq data comparisons. Applying our method to two datasets of RNA-Seq for zika infection in human cells and pollen tissue in *Arabidopsis thaliana*, we identified 1,871 differentially AS events for 1,394 protein-coding genes in human and 496 differentially AS events for 358 protein-coding genes in Arabidopsis, respectively. The results included known AS events reported before as well as novel events, which demonstrate that the biological replicates are important in the effective identification using β-distribution. With a high accurate rate, our new method in differential AS identification will facilitate future investigation on transcriptomic annotation.**

Keywords— alternative splicing (AS), next generation sequencing, RNA-seq

I. INTRODUCTION

Alternative splicing is a critical mechanism for expanding regulatory and functional diversity from a limited number of gene templates, and is particularly complex in higher plants such as *Arabidopsis thaliana* [1, 2]. According to conservative estimates based on annotated gene models, about 40% of multi-exon genes in higher plants can produce multiple transcript isoforms through the regulation of alternative splicing [3]. In human, over 90% of multi-exon genes are alternatively spliced [4]. Yet, it was difficult to estimate the prevalence and the biological significance of AS in these complex model eukaryotic genomes before the high-throughput sequencing technology became available. Due to the high cost of sequencing process, earlier RNA-Seq studies limited to small sequencing depth and without biological replicates of each sample. With the decrease of sequencing cost of RNA-Seq, it is common to use the replicated samples with higher sequencing coverage for transcriptome analysis to reduce noise and variance unwanted. There is no doubt that biological or the sequencing performance variance of single sample will detriment the identification of alternative splicing of high-throughput sequencing studies. Therefore, there is a tremendous need for new and robust analytic tools to detect differential AS with various types of replicate high-throughput datasets.

A few analytic tools to identify the differential alternative splicing transcripts are available from replicated RNA-Seq data, such as DEXSeq[5], DSGseq[6], SplicingCompass[7], and rMATS[8] from read-count based models, which are to quantify transcript with single isoforms, while other tools, such as Cufflinks[9] and DiffSplice[10], use isoform resolution models instead. For example, rMATs uses a statistical model incorporating the information of alternative splicing in splice junction and exon reads to estimate the isoforms proportion with the consideration of paired replicated samples. Counting units in rMATs can be full or truncated exonic regions for the statistical filtration. For genes with replicates, rMATs, however, showed a biased lower estimation of false-negative ratio of AS events when the read coverage of AS events is low.

RNA-Seq experiment has been widely used bcause it can produce millions of reads in a single experiment and, hence, detect novel transcripts and facilitates gene prediction [11]. Robust and acurate AS detection methods are in great demand for RNA-seq data analysis. For this aim, we developed a new method for identification of different AS events, which can identify differential AS events with an adjusted β-distribution model based on replicate RNA-seq data. We applied our method to two test RNA-seq datasets and showed a robust performance on differential AS detection in both human and plant. Compared with various types of evidence, the results obtained by our method are accurate.

II. METHOD AND MATERIALS

A. Pipeline of identifying differential alternative splicing events

All short reads underwent quality controlled with FastQC (v0.11.5) and trimmed with Trimommatic (0.32) [12]. Trimmed high quality reads were then aligned onto the corresponding reference genomes using STAR [13] with mismatches up to three. The count of reads for different AS events, including alternative 3' and 5' splice sites (A3SS,

978-1-5090-4767-3/17/$31.00 ©2017 IEEE

A5SS), skipped exons (SE), mutually exclusive exons (MXE), and retained introns (RI), were obtained by HTSeq-count (0.6.1p1) with the exon and junction information from gene annotations. For each AS event, two types of reads, inclusion reads and skipping reads, are counted seperatively. Raw read counts of AS events for all replicates were grouped in inclusion read group and skipping read group, and each group were normalized using quantile normalization method with R package "preprocessCore". Raw input table containing the detailed list of inclusion and skipping read count in plain text format is required for the process, and the minimum coverage of 5 was required. Signicantly differentially AS patterns were identified by the adjusted β-distribution model by considering both the overall gene expression level and variation in alternative splicing levels among replicates. The detail of this model for test is described in the following section. Finally, the significantly differential AS events were reported with p-value less than 0.001 and the fold change of inclusion/skipping portion between two groups large than 1.5 or less than 0.5.

B. *Adjustied β-distribution model*

To test the difference of an AS event between two groups, such as control and treatment groups, the proportion of inclusion reads was assumed to follow β-distribution, and the probability density function is

$$f(p) = \frac{1}{Beta(\phi\mu, \phi(1-\mu))} p^{\phi\mu-1}(1-p)^{\phi(1-\mu)-1},$$

where p denotes the proportion of the number of inclusion reads in all reads aligned for this given AS event in a group, μ represents the expectation of p, and ϕ is a scale parameter. For a given AS event, the proportion p from i^{th} group can be fitted with a logistic model,

$$\log(\frac{p_i}{1-p_i}) = \tau_i,$$

where $p_i \sim Beta(\mu_i, \phi_i)$ and τ_i represents fixed effect from i^{th} group. Compared with the commonly used binomial distribution, the scale parameter ϕ_i in beta distribution could be used to account for extra dispersion in the data. In addition, in our model, the scale parameter ϕ_i for an AS event was adjusted with the read count values by fitting a normal distribution, *i.e.* $\phi_i = f(\phi'_i, N(R))$, where R is the total number of reads for the given AS, and ϕ'_i is estimated by the variance of dataset. Therefore, for example, high read counts would lead to smaller variance of the scale parameters. This adjustment could be biologically meaningful; more read counts were always regarded to be more reliable. However, for extremely low read count of some AS events, this model would not fit, and hence, the Fisher's exact test would be used instead. Finally, q values were calculated based on the p values to solve the problem caused by multiple tests.

C. *GO enrichment analysis*

Gene Ontology (GO) enrichment analysis was performed using Gorilla [14], and lists of enriched GO terms were generated with an adjusted p-value <0.05 for both species. The GO term association information for Human and Arabidopsis was obtained from the GO database.

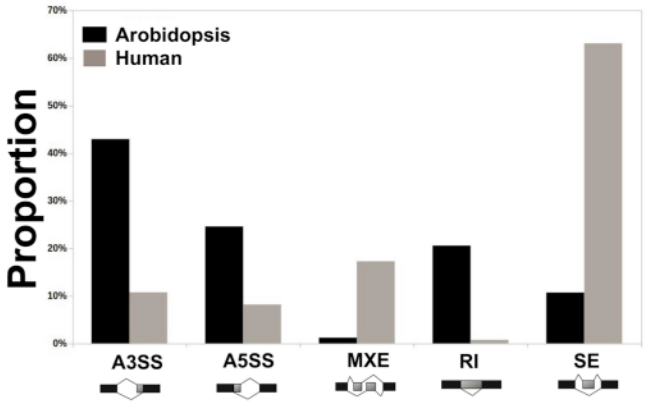
D. *Datasets for test*

Two published RNA-seq dataset were used to test our method. (1) The RNA-seq dataset for pollen and leaf tissues of *Arabidopsis thaliana* were obtained from SRA (SRR847501-SRR847504) [15]. Each library was sequenced for 75 cycles by the single-end type, and the total number of reads is about 40 million per library. (2) The RNA-seq dataset for human cells with zika infection were obtained from SRA (SRR3194428-SRR3194431) [16]. Each library was sequenced on an Illumina HiSeq2500 for 75 cycles by the single-end type.

III. RESULTS

Using the new method, we conducted genome-wide identification of differential alternative splicing events for zika infection in human and pollen/leaf tissues in Arabidopsis. According to the read mapping results, about 85% and 91% of all these genes were expressed with a minimum read coverage larger than 5. There was a high concordance between replicates in the read count number called in RNA-Seq of both datasets, illustrating the high sensitivity of RNA-Seq in terms of qualitative expression values for AS study.

Fig. 1. The proportion of each type of AS events in Arabidopsis and Human zika infection cases, respectively.



To identify the most abundantly expressed gene in zika dataset, we ranked the genes according to their normalized expression values. A total of 1,871 significant differential AS events were identified between zika infection and control. Out of 1,871 AS events, there 1,180 intron retention (SE), which dominate the total differential AS events. Please see Fig. 1 for the distribution of five types of AS events.

For the Arabidopsis case, a total of 496 significant differential AS events were identified between pollen and leaf tissues. In Fig.1, one can see that 43% of these differential AS events is A3SS, and it agrees with the discovery of Loraine et al. [3]. In addition, most of these A3SS events are relevant to plant development, such as AT4G14385, which was reported by Loraine et al. [3]. Many research focused on intron-retention isoforms from ancestor genes coupled with RNA to DNA reverse and recruitment into the genome. Based on our result, there are much more RI related genes proved this

evolutionary significance.

We compared the predicted differential AS events from rMATs and our method and there are many overlaps in all five AS types. Besides, our method also reported some AS events with large fluctuation of coverage across the spliced region, which are usually underestimated by other models. For example, in the zika case, 11 RI were reported by rMATs and a total of 14 by our method. The overlap of predicted AS events between two methods are usually the stable cases. We also compared the AS events exclusively discovered by our method with the Ref. [16], and found that 21% of these AS events were reported.

We further conducted a GO enrichment analysis on genes that have at least one differential AS events. For Arabidopsis, many relevant GO terms for biological proceses, such as single-organism cellular process and phosphate-containing compound metabolic process, were enriched by genes with differential AS events. Many genes are enriched in cellular component terms, including categories related to cell wall, pollen tube and cell projection as well as function terms, such as protein phosphorylation and kinase activity [3]. For zika infection case, enriched GO terms are shown in Table 1. These GO terms showed that the discovered differential AS and their parent genes are relevant to the response to zika infection.

In our predicted differential AS events, many of them are reported being relevant to biological processes. For example, the Arabidopsis gene, AT3G04620, has a differential RI event identified by our method. It encodes an alba DNA/RNA-binding protein and is invovled in pollen sperm cell differentiation [17]. The gene product of AT1G64980, discovered to have SE event, is a nucleotide-diphospho-sugar transferases superfamily protein (CDI), and works for pollen tube growth [18]. The gene AT3G06330, encoding a putative RING-type ubiquitin ligase, was identited to has a different A5SS event by our algorithm. This AS event also was confirmed by experimental work; the 5' donor site was preferred in pollen, while the upstream donor site of preferred in leaves [19].

TABLE I. ENRICHED GO-TERM FOR THE ZIKA INFECTION CASE

Term	Description	Pvalue	AS[a]
GO:0043170	macromolecule metabolic process	$7.05x10^{-10}$	37,40,46,227,4
GO:1902582	single-organism intracellular transport	$3.24x0^{-6}$	5,7,6,33,0
GO:0009058	biosynthetic process	$1.14x10^{-5}$	29,25,28,144,3
GO:0000398	mRNA splicing, via spliceosome	$4.04x10^{-5}$	2,2,3,23,1
GO:0006396	RNA processing	$2.83x10^{-7}$	7,9,7,47,1

[a] Numbers of genes with a specific type of AS, in order of A5SS, A3SS, MXE, SE, RI.

For the human zika infection case, the human *SNRPN* gene, encoding a serine/arginine-rich (SR) RNA-binding protein, was found to have a different SE event by our method. This gene was thought to play a role in immume response for virual infection. According to our method, genes of serine/arginine-rich splicing factor 3 and 11 were differentially spliced in control and zika patients. Both genes showed significant regulation on the splicing process of multi-exon transcripts. One more example is that the human *CD46* gene, encoding a membrane cofactor protein, is relevant to human immune system and has a SE event in zika patients.

IV. DISCUSSION

The significant AS events identified by our method can be combined with other analysis to help understand the biological systems and facilitate genome annotation. We applied our method with RNA-Seq datasets from two Arabidopsis tissues, and based on TAIR10 gene sequences and models. Comparing the detected differential AS events and the corresponding gene expression levels, we found that a majority of genes with differential AS events do not have differential expression levels in pollen development. This result demonstrated that AS functions as an complementary mechemism for gene regulation for plant development. For the human case, more altervative splicing variants were discovered in zika infection patients than controls. Our analysis discovered 1,871 differential AS events of 1,394 genes in zika infection samples, and only 875 of these genes were differentially expressed at the gene level. This agrees with the discovery that AS plays an important role in regulation of host genes during early virus infection [20]. An thorough comparison and assessment of the performance is challenging because it is hard to get a benchmark of AE events.

ACKNOWLEDGMENT

We thank Dr. Shangang Jia and Weilong Yang for technical assistance and method discussion. This study is supported by NE Soybean Board funds.

REFERENCES

[1] V. Costa, C. Angelini, I. De Feis, and A. Ciccodicola, "Uncovering the complexity of transcriptomes with RNA-Seq," *J Biomed Biotechnol,* vol. 2010, p. 853916, 2010.

[2] M. A. Campbell, B. J. Haas, J. P. Hamilton, S. M. Mount, and C. R. Buell, "Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis," *BMC Genomics,* vol. 7, p. 327, Dec 28 2006.

[3] A. E. Loraine, S. McCormick, A. Estrada, K. Patel, and P. Qin, "RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing," *Plant Physiol,* vol. 162, pp. 1092-109, Jun 2013.

[4] A. E. Loraine, G. A. Helt, M. S. Cline, and M. A. Siani-Rose, "Protein-based analysis of alternative splicing in the human genome," *Proc IEEE Comput Soc Bioinform Conf,* vol. 1, pp. 118-24, 2002.

[5] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome Res,* vol. 22, pp. 2008-17, Oct 2012.

[6] W. Wang, Z. Qin, Z. Feng, X. Wang, and X. Zhang, "Identifying differentially spliced genes from two groups of RNA-seq samples," *Gene,* vol. 518, pp. 164-70, Apr 10 2013.

[7] M. Aschoff, A. Hotz-Wagenblatt, K. H. Glatting, M. Fischer, R. Eils, and R. Konig, "SplicingCompass:

differential splicing detection using RNA-seq data," *Bioinformatics,* vol. 29, pp. 1141-8, May 01 2013.

[8] S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu*, et al.*, "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data," *Proc Natl Acad Sci U S A,* vol. 111, pp. E5593-601, Dec 23 2014.

[9] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren*, et al.*, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat Biotechnol,* vol. 28, pp. 511-5, May 2010.

[10] Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson*, et al.*, "DiffSplice: the genome-wide detection of differential splicing events with RNA-seq," *Nucleic Acids Res,* vol. 41, p. e39, Jan 2013.

[11] A. C. English, K. S. Patel, and A. E. Loraine, "Prevalence of alternative splicing choices in Arabidopsis thaliana," *BMC Plant Biol,* vol. 10, p. 102, Jun 04 2010.

[12] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics,* vol. 30, pp. 2114-20, Aug 01 2014.

[13] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha*, et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics,* vol. 29, pp. 15-21, Jan 01 2013.

[14] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics,* vol. 10, p. 48, Feb 03 2009.

[15] A. D. Estrada, N. H. Freese, I. C. Blakley, and A. E. Loraine, "Analysis of pollen-specific alternative splicing in Arabidopsis thaliana via semi-quantitative PCR," *PeerJ,* vol. 3, p. e919, 2015.

[16] H. Tang, C. Hammack, S. C. Ogden, Z. Wen, X. Qian, Y. Li*, et al.*, "Zika Virus Infects Human Cortical Neural Progenitors and Attenuates Their Growth," *Cell Stem Cell,* vol. 18, pp. 587-90, May 05 2016.

[17] M. Borg, L. Brownfield, H. Khatab, A. Sidorova, M. Lingaya, and D. Twell, "The R2R3 MYB transcription factor DUO1 activates a male germline-specific regulon essential for sperm cell differentiation in Arabidopsis," *Plant Cell,* vol. 23, pp. 534-49, Feb 2011.

[18] H. M. Li, H. Chen, Z. N. Yang, and J. M. Gong, "Cdi gene is required for pollen germination and tube growth in Arabidopsis," *FEBS Lett,* vol. 586, pp. 1027-31, Apr 05 2012.

[19] Y. Wang, W. Z. Zhang, L. F. Song, J. J. Zou, Z. Su, and W. H. Wu, "Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis," *Plant Physiol,* vol. 148, pp. 1201-11, Nov 2008.

[20] E. Petersen, M. E. Wilson, S. Touch, B. McCloskey, P. Mwaba, M. Bates*, et al.*, "Rapid Spread of Zika Virus in The Americas--Implications for Public Health Preparedness for Mass Gatherings at the 2016 Brazil Olympic Games," *Int J Infect Dis,* vol. 44, pp. 11-5, Mar 2016.