

# RAX2: a genome-wide detection method of condition-associated transcription variation

Yuan-De Tan<sup>1</sup>, Jixin Deng<sup>1</sup> and Joel R Neilson<sup>1,2,\*</sup>

<sup>1</sup>Department of Molecular Physiology and Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA and <sup>2</sup>Dan L. Duncan Cancer Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received January 20, 2015; Revised April 4, 2015; Accepted April 17, 2015

## ABSTRACT

Most mammalian genes have mRNA variants due to alternative promoter usage, alternative splicing, and alternative cleavage and polyadenylation. Expression of alternative RNA isoforms has been found to be associated with tumorigenesis, proliferation and differentiation. Detection of condition-associated transcription variation requires association methods. Traditional association methods such as Pearson chi-square test and Fisher Exact test are single test methods and do not work on count data with replicates. Although the Cochran Mantel Haenszel (CMH) approach can handle replicated count data, our simulations showed that multiple CMH tests still had very low power. To identify condition-associated variation of transcription, we here proposed a ranking analysis of chi-squares (RAX2) for large-scale association analysis. RAX2 is a non-parametric method and has accurate and conservative estimation of FDR profile. Simulations demonstrated that RAX2 performs well in finding condition-associated transcription variants. We applied RAX2 to primary T-cell transcriptomic data and identified 1610 (16.3%) tags associated in transcription with immune stimulation at FDR < 0.05. Most of these tags also had differential expression. Analysis of two and three tags within genes revealed that under immune stimulation short RNA isoforms were preferably used.

## INTRODUCTION

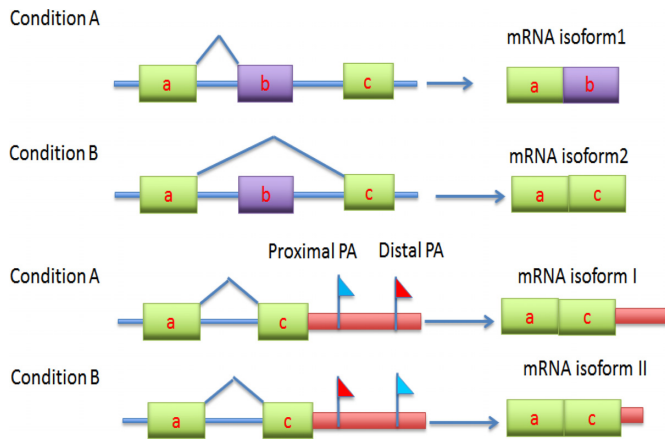
It has recently been revealed that alternative splicing and alternative cleavage and polyadenylation (ACP) is not only a universal post-transcription processing step in eukaryotic gene expression but also a versatile mechanism for post-transcriptional regulation of genes (1–3). After transcription, a pre-mRNA is capped at the 5' end, spliced and

cleaved in the 3'-untranslated region (3' UTR) to yield a new open end that allows to add a polyadenylation (poly(A)) tail (3,4). A poly(A) tail at 3' end may protect the mRNA from unregulated degradation, trigger export of the mRNA to cytoplasm and assist recognition by translation machinery (3–5). A poly(A) signal at which the pre-mRNA is cleaved and referred to as poly(A) site is recognized and activated by a batch of protein factors called polyadenylation factors (3,4,6,7). Alternative splice and poly(A) sites significantly increase the complexity of transcriptomes and proteomes because they lead to multiple isoforms or variants of an mRNA. Using next-generation sequencing (NGS) techniques, it has been observed that over 50% of transcription units in the mammalian genome are characterized by ACP (8–13). As alternative splice and poly(A) sites can uncover most transcription variation at subgene level, to investigate association of transcription variation with tissue types or cell state change has become very important for exploring etiologic mechanism of disease of interest (8,14–20). But almost all of the current studies are based on differential analysis. This may be because the current large-scale statistical methods such as DESeq (21), baySeq (22), edgeR exact test (23,24), edgeR GLM (25), DEXSeq (26), Cuffdiff (27,28), DiffSplice (29), SplicingCompass (30) are differential analysis methods. Of these methods, the former four are used to identify differential transcription between conditions and the latter four focus on finding differential splicing between conditions. However, unlike gene differential expression, splice switch or usage switch of alternative poly(A) sites occurs on the same transcription unit and are associated with a change in condition (Figure 1). As shown in Figure 1, switch usage or switch splicing is involved in two conditions and two poly(A) sites or two splice sites:

	condition 1	condition 2
site A	high expression	low expression
site B	low expression	high expression

High-expression switches from site A in condition 1 to site B in condition 2. This is a case that high expression of sites is associated with change in condition. If one just looks at the difference in expression of site A or site B be-

\*To whom correspondence should be addressed. Tel: +713 798 8851; Fax: +713 798 3475; neilson@bcm.edu



**Figure 1.** Demonstration for association of usage of alternative splice and poly(A) sites with conditions: splicing switch and usage switch of poly(A) sites due to change in condition. Under normal condition, the gene transcription product is isoform 1 due to splicing event between exons a and b, but in stress or stimulation, for example, the gene product is isoform 2 by switching splicing from exon b to exon c (top two figures). This is called a splicing switch. Likewise, under normal condition, cells use distal poly(A) site and the transcription produces mRNA isoform I but when cells are stressed or stimulated, poly(A) site usage switches from distal site to proximal site and produces isoform II (bottom two figures). This is called poly(A) site usage switch.

tween two conditions, then this is another case that one concerns differential expression of single sites between two conditions. In the first case only an association method can be used to test if sites A and B are associated with conditions 1 and 2 in expression but for the second case one can use differential methods to test for differential expression of each site between two conditions. So the use of differential analysis methods to identify usage and splicing switches from site to site on the same RNA unit is not ideal. To our best knowledge, no existing large-scale association analysis methodologies have been proposed. The traditional association methods such as Pearson chi-square test and Fisher exact test are single-test methods and work on the count data without replicates, they hence could not directly be applied to high-throughput transcriptomic count data of RNA reads with replicate libraries. The Cochran Mantel Haenszel (CMH) chi-square test approach may handle the data with replication, but it is a single test method, not a large-scale association test approach. This is why we need to develop a large-scale association statistical method for identifying tissue-associated or condition-associated isoform transcriptions. This method is named ranking analysis of chi-squares (RAX2). RAX2, which extends the traditional chi-square test to replicated and high-throughput transcriptomic count data of RNA reads derived from NGS, is based on comparison between a set of ranked chi-square statistics for association effects and a set of ranked null chi-square values across a set of given thresholds and estimation of a false discovery rate (FDR) profile.

## MATERIALS AND METHODS

### Cell lines and stimulation

Human primary CD4 T cells were obtained from buffy coats derived from the Gulf Coast Regional Blood Bank via the EasySep Negative Selection Kit (Stem Cell Technologies) as per manufacturer's instructions. Purity (>90%) was assessed via flow cytometry for CD3 and CD4 (UCHL1 and RPA-T4, respectively). Cells were maintained in RPMI (ATCC) with 10% fetal bovine serum supplemented with 10 mM HEPES pH 7.4, nonessential amino acids, 2 mM L-glutamate, 50  $\mu$ M  $\beta$ -mercaptoethanol, 100 units/ml penicillin and 100  $\mu$ g/ml streptomycin (all from Gibco/Life Technologies). T-cells were stimulated with plate-bound antibodies (1  $\mu$ g/ml anti-CD3 (OKT3 – eBiosciences), 5  $\mu$ g/ml anti-CD28 (CD28.2 – BD Pharmingen). Activation of T-lymphocytes was verified via flow cytometric detection of CD69 expression (FN-50) 16 h after stimulation, and cells were harvested at 48 h.

### High-throughput sequencing library generation

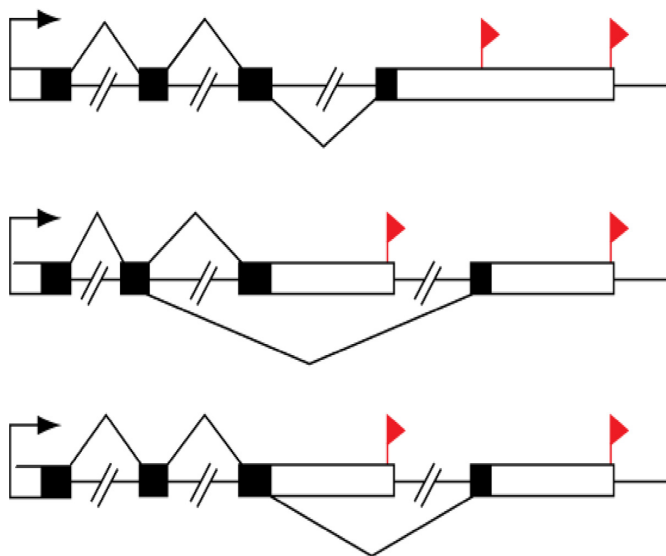
Total RNA was harvested from resting and stimulated cells with Trizol reagent (Life Technologies) as per manufacturer instructions. Polyadenylated RNA was isolated with the Poly(A)-Purist MAG (Ambion/Life Technologies) kit as per manufacturer instructions. High-throughput sequencing libraries were generated essentially as described (2), with the exception that 'barcoded' linkers were used to facilitate multiplexing. Libraries were sequenced via 50 bp paired-end sequencing on an Illumina GAIIX.

### Library processing and mapping

Paired end reads were mapped to the hg18 build of the human genome using the paired-end mapping module of BWA (31), default alignment stringency and requiring that each read be mapped in a proper pair. To rescue reads crossing splicing junctions, nonmapping reads were remapped to the UCSC KnownGene reference and projected back to the hg18. Individual reads were condensed into tags based on their 3' coordinate using a sliding window of 20 nucleotides and using the frequency-weighted median 3' coordinate as the tag identifier. Tags were then filtered by using a progressive filtering strategy to assess adenosine and guanine composition in the five, ten and fifteen bases followed the tag-mapping site. Tags were assigned to individual transcription units based on UCSC KnownGene annotations. For each transcription unit, the aggregate tags mapped to the unit were ranked based on frequency. Tags were extracted from the highest to the lowest frequency until the extracted tags represented greater than 90% of the aggregate frequency for the gene. The remaining tags were discarded. Libraries were normalized using a negative binomial model within the DE-seq package (21). For 'gene' analysis, the summed frequency of all tags mapped to each transcription unit was considered as a single entity.

### RAX2 packages

The methods described in this paper, including estimations of null chi-square distribution and FDR, are implemented

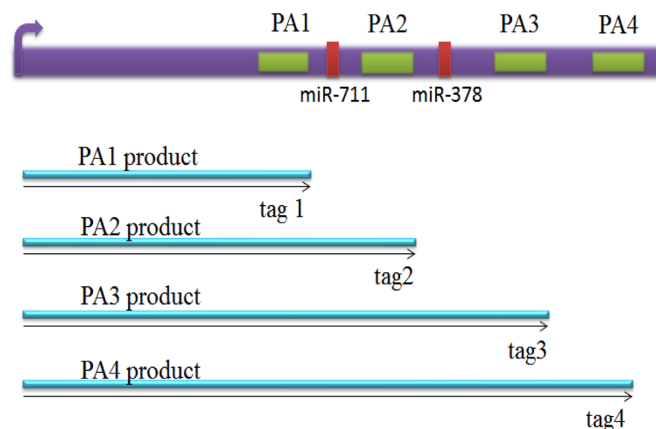


**Figure 2.** Transcription unit structures and different 3' UTR isoforms. A majority of mammalian transcription units are characterized by alternative cleavage and polyadenylation (ACP). The majority of these contain multiple cleavage and polyadenylation sites in their terminal exon (top), impacting untranslated region identity without changing the coding sequence. Transcription units may also be characterized by mutually exclusive terminal exon structure (middle) or composite terminal exon structure (bottom). In the latter two cases, ACP is coupled to mRNA splicing, and both the coding sequence and the untranslated region are impacted. Black box: exon, white box: untranslated region (UTR), red flag: poly(A) site and -// -: intron.

in software package RAX2. RAX2 was written in R. The current version of RAX2 is designed to analyze count data in two distinct states and the results are output in MS DOS csv format. Performance of RAX2 and the other statistical analysis process are given in Supplementary User Guidance for RAX2. The RAX2 package can be found in Supplementary Package or Bioconductor.

**Models**

For convenience, our discussion about RAX2 is based on ACP sites within genes. ACP can occur concomitantly with or independently of alternative splicing (Figure 2). ACP independent of alternative splicing, that is, all cleavage and polyadenylation (poly(A)) sites are in the terminal exon of the transcription unit, results in a transcription unit with a tandem untranslated region (tandem UTR). Both alternative splicing of terminal exons and tandem UTR usage are visible for directed 3' end sequencing methodologies. Since either process has the potential to alter 3' UTR identity, we do not differentiate between them in this analysis. For a given transcription unit, transcript variants derived from the first poly(A) site (poly(A) site 1) are assumed to have a transcript length from the transcriptional start position to poly(A) site 1. Similarly, transcript variants with poly(A) sites 2 and 3 are also assumed to be derived from the transcriptional start position to poly(A) sites 2 and 3, respectively. Therefore, within such a transcription unit, the transcript isoforms are one-to-one corresponding to poly(A) sites. For the sake of convenience, we refer to the transcript



**Figure 3.** A model for multiple poly(A) sites or tags within genes. mRNA of Hsp70.3 is a typical model for multiple poly(A) sites (green boxes) and microRNA sites (red boxes). It has a common start point (promoter  $\curvearrowright$ ), four poly(A) sites (PA1, PA2, PA3 and PA4) in 3' UTR coding for several mRNA isoforms that are different in their 3' end. Alternative polyadenylation changes the length of the 3' UTR and it also can change microRNA binding sites in 3' UTR. While microRNAs tend to repress translation and promote degradation of the mRNAs they bind to. Since transcription products are one-to-one corresponding to poly(A) sites, we define the transcription products as tags.

variant at poly(A) site 1 as tag1. Similarly, the transcript variants at poly(A) sites 2, 3 and 4 are also defined as tags 2, 3 and 4, respectively (Figure 3). In general, within gene  $g$  ( $g = 1, \dots, G$ ),  $Z_g$  alternative poly(A) sites ( $Z_g$  is number of poly(A) sites within gene  $g$ ) correspondingly have  $Z_g$  alternative tags. For example, gene AARS2 on chromosome 6 has four poly(A) sites ( $Z_g = 4$ ) and hence its mRNA transcription unit (called gene) has four isoforms or tags (see columns gene name and pos (position) in Supplementary Table S1). Supplementary Table S1 lists such count data of reads of tags within genes where columns 1–7 list information of tags or poly(A) sites within genes: tagid, geneid, name, DNA strand, position on transcription unit and annotation; columns 8–10 list count data of cells in normal and stimulated states. The data were normalized so that they look decimal. Here, we let  $n_{gzjv}$  be count of tag  $z$  ( $z = 1, \dots, Z_g \geq 2$ ) within gene  $g$  ( $g = 1, 2, \dots, G$ ) in cell state  $j$  in replicate  $v$  ( $v = 1, \dots, r$ ). Our concern here is if change in count of a tag  $z$  at a special poly(A) site *within* a gene of interest is associated with change in condition (cell state) in the context of a discontinuous process. For the convenience, we focus on two states: treatment state (TS) and normal state (NS). We set  $j = 1$  for NS and  $j = 2$  for TS. To find if the transcription variation of an individual tag is associated with change in cell state, we need to compare the count of tag  $z$  to that of all the other tags (denoted by  $\bar{z}$ ) within gene  $g$ . Therefore, a tag or poly(A) site has two states:  $z$  and  $\bar{z}$ . We set  $i = 1$  for  $z$  and  $i = 2$  for  $\bar{z}$ . Thus count  $n_{gzjv}$  in the raw data table can be converted into two-by-two table count  $n_{gzijv}$  where  $n_{gzijv}$  is count of RNA reads of tag  $z$  within gene  $g$  in tag state  $i$  in condition  $j$  in replicate  $v$  ( $v = 1, 2, \dots, r$ ) and there are  $Z_g r$  two-by-two table datasets ( $n_{gz11v}, n_{gz12v}, n_{gz21v}, n_{gz22v}$ ) for gene  $g$ . Since our interest is in regard to tag  $z$  (or poly(A) site  $z$ ) within gene  $g$  instead of gene  $g$  it-



self, totally we have  $S = \sum_{g=1}^G Z_g$  tags or poly(A) sites in all genes of study. To implement ranking analysis of all tags, we combine subscripts  $g$  and  $z$  into  $s$  where  $s = 1, \dots, S$ . Thus, each tag  $s$  has two states ( $s$  and  $\bar{s}$ ). We still set  $i = 1$  for tag state  $s$  and  $i = 2$  for tag state  $\bar{s}$ . The two-by-two table count dataset ( $n_{gz11v}, n_{gz12v}, n_{gz21v}, n_{gz22v}$ ) is then rewritten as ( $n_{s11v}, n_{s12v}, n_{s21v}, n_{s22v}$ ).

Since there are replicate observations for each tag, each count consists of two components: the association effect and random noise. Thus, a model for each count in such a two-by-two table is

$$n_{sijv} = n_{sij} + e_{sijv} \quad (1)$$

where  $n_{sij}$  is count of RNA reads of tag  $s$  due to association effect between state  $i$  of poly(A) site  $s$  and cell state  $j$  and  $e_{sijv}$  is count of RNA reads due to a special noise between state  $i$  of poly(A) site  $s$  and cell state  $j$  in replicate observation  $v$ . Since  $n_{sijv}$  follows Poisson distribution (21,24) or negative binomial distribution (21,24,25) or binomial distribution (32),  $e_{sijv}$  may also follow one of these distributions. But we here do not concern distribution of  $e_{sijv}$  because we do not need to separate it from  $n_{sijv}$ . Let  $f_{sij}$  and  $f_{sijv}$  be frequencies of counts  $n_{sij}$  and  $e_{sijv}$ , respectively. As  $n_{sij}$  is count of RNA reads of tag  $s$  in state  $i$  and in cell state  $j$ ,  $n_{sij} = n_s f_{sij}$  and  $e_{sijv} = n_s f_{sijv}$  where  $n_s$  is total count of RNA reads of tag  $s$  within  $g$  across two cell states and two tag states over all replicate libraries. Thus, model (1) may be rewritten as

$$n_{sijv} = n_s f_{sij} + n_s f_{sijv}. \quad (2)$$

### Chi-square statistics

For poly(A) site  $s$ , the mean count ( $\bar{n}_{sij}$ ) of RNA reads over  $r$  replicates consists of association effect  $n_{sij}$  between site state  $i$  and cell state  $j$  and mean noise ( $\bar{e}_{sij}$ ):

$$\begin{aligned} \bar{n}_{sij} &= \frac{1}{r} \sum_{v=1}^r n_{sijv} = \\ n_s f_{sij} + \frac{n_s}{r} \sum_{v=1}^r f_{sijv} &= n_s f_{sij} + n_s \bar{f}_{sij} = n_{sij} + \bar{e}_{sij} \end{aligned} \quad (3)$$

where  $f_{sij}$  is the frequency of association effect  $n_{sij}$  between state  $i$  of tag  $s$  and cell state  $j$ ,  $f_{sijv}$ , frequency of a special noise in replicate observation  $v$  and  $\frac{1}{r} \sum_{v=1}^r f_{sijv} = \bar{f}_{sij}$ . For poly(A) site  $s$ ,  $\bar{n}_{sij}$  is expected as

$$E(\bar{n}_{sij}) = E(n_{sij}) + E(\bar{e}_{sij}). \quad (4)$$

(See Supplementary Note S1 for detail), that is, expectation of mean count ( $\bar{n}_{sij}$ ) of RNA reads of tag  $s$  in tag state  $i$  and cell state  $j$  equals expectation of association effect ( $n_{sij}$ ) plus expectation of mean noise ( $\bar{e}_{sij}$ ). Thus, Pearson chi-square statistics for mean count  $\bar{n}_{sij}$  that is estimate of association effect  $n_{sij}$  is

$$\chi_s^2(\bar{n}) = \chi_s^2(n + \bar{e}) \quad (5)$$

(See Supplementary Note S2 for derivation). From Equation (5), we can see that  $\chi_s^2(\bar{n}) \neq \chi_s^2(n)$  unless the mean of noises is zero ( $\bar{e} = 0$ ), meaning that the Pearson chi-square

for  $\bar{n}_{sij}$  is not an unbiased estimate of the chi-square statistic for  $n_{sij}$ . For this reason, we cannot use  $\chi_s^2(\bar{n})$  as standard chi-square statistic, that is,  $P$ -value for  $\chi_s^2(\bar{n})$  obtained from a chi-square distribution has a big bias. In addition,  $\chi_s^2(\bar{n}) = \chi_s^2(\bar{e})$  if, for a tag  $s$ , there is no association effect  $n_{sij}$  between tag states and cell states, say,  $\bar{n}_{sij} = \bar{e}_{sij}$ .

### Null chi-squares

Similarly to using within-group variance to estimate null variance, we also employ within-group chi-square to estimate null chi-square. To this end, we need to construct another contingent table in which we have site states  $s$  and  $\bar{s}$  within gene  $g$ , and  $r$  replicates in cell state  $j$ . Set  $i = 1$  for  $s$  and  $i = 2$  for  $\bar{s}$ . A set of two-by-two tables (see Supplementary Note S3) is constructed with pairs of replicates and states of tag  $s$  in cell state  $j$ . Thus the null chi-square is easily estimated by  $\chi_s^2(e)$  (see Supplementary Note S3 for derivation).

### Ranking analysis of chi-squares

Since  $\chi_s^2(\bar{n})$  does not have one-to-one correspondence to  $\chi_s^2(e)$ , to compare them, we need to separately rank tags by amounts of  $\chi_s^2(\bar{n})$  and  $\chi_s^2(e)$  from the smallest to the largest. Let  $\chi_{s^*}^2(\bar{n})$  and  $\chi_{s^*}^2(e)$  be two Pearson chi-square values at position  $s^*$  in the ranking linear space(\*), the chi-square values are smallest at  $s^* = 1$  and largest at  $s^* = S^*$ . Then we can compare  $\chi_{s^*}^2(\bar{n})$  to  $\chi_{s^*}^2(e)$  at position  $s^*$  in a one-to-one fashion. Given threshold  $\Delta$ , expression change of tag  $s^*$  is declared to be associated with change in cell state if and only if

$$\chi_{s^*}^2(\bar{n}) - \chi_{s^*}^2(e) > \Delta. \quad (6)$$

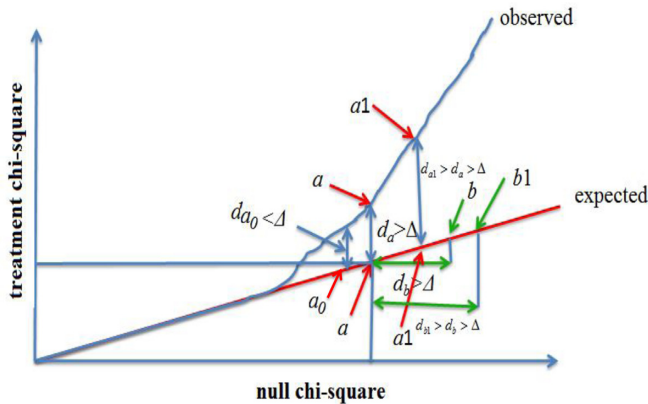
### FDR estimation

In large-scale statistical analysis, we do not need to be concerned with so-called type I error but we must consider how to control false discovery rate (FDR) because in, for example, 10 000 hypotheses to be tested, at least 500 hypotheses would commit type I error if  $\alpha = 0.05$  is chosen as significance level (33). Obviously 500 hypotheses rejected by chance are not acceptable. This is why we here want to control FDR rather than type I error. A reasonable control of FDR involves a reliable estimation of FDR. Given threshold  $\Delta$ , the number ( $N_\Delta$ ) of tags whose expression changes are declared to be uniquely associated with change in cell state may consist of the number of truly positive tags,  $T_\Delta$ , and number of falsely positive tags,  $F_\Delta$ :

$$N_\Delta = T_\Delta + F_\Delta. \quad (7)$$

Therefore, FDR is expected as  $FDR_\Delta = E(F_\Delta/N_\Delta)$  at threshold  $\Delta$ . As  $F_\Delta$  is unknown in experimental datasets, FDR must be estimated.

The existing methods for estimation of FDR are not suitable to our current chi-square statistics because, as seen above, chi-square statistics are remarkably biased against standard chi-square statistics such that  $P$ -values obtained from chi-square distribution are also biased. For this reason, we here propose a novel method. The principle of the



**Figure 4.** A demonstration plot of ranked treatment chi-square versus null chi-square. The expected linear plot (red line) is given by null hypotheses that the treatment chi-square is equal to null chi-square at each chi-square point. The observed linear plot is given by ranked observed treatment chi-squares versus ranked estimated null chi-squares. Given a threshold  $\Delta$ , all treatment chi-squares with  $d_s \geq d_a$  are declared to be significant or interested where  $d_a = \chi_a^2(\bar{n}) - \chi_a^2(e) \geq \Delta$  and  $s = a + 1, \dots, G$ . All null chi-squares with  $d_t \geq d_b$  are defined as potential false positives with probability given in Equation (9) where  $d_b = \chi_b^2(e) - \chi_a^2(e) \geq \Delta$  and  $t = b + 1, \dots, G$ .

method is shown in Figure 4. As seen in Figure 4, ordered and observed chi-square  $\chi_{s^*}^2(\bar{n})$  more and more deviates from ordered and expected chi-square  $\chi_{s^*}^2(e)$  as chi-square value increases.

Given threshold  $\Delta$ , we find that  $d_a = \chi_a^2(\bar{n}) - \chi_a^2(e) > \Delta$  at position  $a$  in the ranking linear space (see Figure 4) is the smallest distance among all  $s^*$ . We define  $a$  as  $a_\Delta$  and  $d_a$  as  $d_{a_\Delta}$ . All  $d_{s^*} = \chi_{s^*}^2(\bar{n}) - \chi_{s^*}^2(e) > \Delta$  would be larger than or equal to  $d_{a_\Delta}$  where  $s^* = a_\Delta, a_\Delta + 1, \dots, S$ , that is, all tags at  $s^* \geq a_\Delta$  would be declared as positive tags. So, the number of positives declared by Equation (6) is obtained by

$$N_\Delta = S - a_\Delta + 1. \quad (8)$$

In order to avoid confusion,  $s^*$  in the ordered null chi-square space is replaced with  $t^*$ . Likewise, for a given threshold  $\Delta$ , at  $t^* = b_\Delta$ , we also find that  $d_{b_\Delta} = \chi_{b_\Delta}^2(e) - \chi_{a_\Delta}^2(e) > \Delta$  is smallest among all  $t^*$  (see Figure 4). All  $\chi_{t^*}^2(e) - \chi_{a_\Delta}^2(e) > \Delta$  would be larger than or equal to  $d_{b_\Delta}$ . In addition, within the ranking space of  $\chi_{t^*}^2(e)$ , null tag at  $t^* = b_\Delta + j$  has smaller chance to be chosen as a false positive than at  $t^* = b_\Delta + k$  where  $j < k = 1, \dots, S - b_\Delta$  and the ranking space of  $\chi_{t^*}^2(e)$  varies with sample size. In other words, the probability that a null tag at position  $t^*$  is chosen as false positive is determined by replicate number  $r$ , chi-square space, tag space  $S^*$ , null chi-square values and  $t^*$ . The logic relationship is that position  $t^*$  in the ranking linear space is positively related to the probability that the null tag at position  $t^*$  is chosen as a false-positive tag and null chi-square value is also positively related to the probability that the null chi-square is chosen as false-positive chi-square. As position  $t^*$  is fixed in a given tag space, the ratio  $(t^*/S^*)$  would not alone fit to changeable false-positive probability from practical data; while  $\chi_{t^*}^2(e)$  is variable and depends on a practical dataset and its values at some nearby positions may not vary or very approximate, so ratio  $(\chi_{t^*}^2(e)/\chi_{S^*}^2(e))$  is also

not alone appropriately used as the false-positive probability. To obtain an accurate estimate of false-positive probability we need to combine them. A good and simple combination way is geometric mean. We use geometric mean  $\sqrt{(\chi_{t^*}^2(e)/\chi_{S^*}^2(e))(t^*/S^*)} = [(\chi_{t^*}^2(e)t^*) / (\chi_{S^*}^2(e)S^*)]^{0.5}$  to estimate the probability that a null tag at position  $t^*$  is chosen as false positive. Additionally, in a single hypothesis test the  $P$ -value of statistic is negatively related to sample size, similarly, the probability that the null chi-square at position  $t^*$  is chosen as false-positive chi-square is also negatively related to sample size because increment of sample size would let space of  $\chi_{t^*}^2(e)$  be shrunken. On the other hand,  $\chi_{S^*}^2(e) = \max(\chi_s^2(e))$  is positively related to the probability. We therefore replace 0.5 with  $x = (2r/\chi_{S^*}^2(e))^2$ . Thus, probability that a tag at position  $t^*$  is chosen as a false-positive tag is given by

$$p(t^*|\chi_{t^*}^2(e)) = \left( \frac{\chi_{t^*}^2(e)t^*}{\chi_{S^*}^2(e)S^*} \right)^x. \quad (9)$$

The number of false positives in  $N_\Delta$  findings is estimated by

$$F_\Delta = \sum_{t^*=b_\Delta}^{S^*} p(t^*|\chi_{t^*}^2(e)). \quad (10)$$

When  $x \rightarrow 0$ ,  $p(t^*|\chi_{t^*}^2(e)) \rightarrow 1$  such that  $F_\Delta \rightarrow S - b_\Delta + 1$ . This is an extreme case in which FDR is very conservatively estimated (see Supplementary Figure S1A1, A2 and A3). In general, larger sample sizes would have a smaller  $\chi_{S^*}^2(e)$ , leading to a larger  $x$  value that causes a smaller  $p(t^*|\chi_{t^*}^2(e))$ . This is agreeable with the fact that larger sample sizes would have a smaller FDR. With  $F_\Delta$ , FDR for  $N_\Delta$  findings declared by chi-square tests at threshold  $\Delta$  is estimated as

$$FDR_\Delta = \frac{F_\Delta}{N_\Delta}. \quad (11)$$

## RESULTS

### Simulation evaluation of RAX2

We use the following steps to generate the null data of  $S$  tags with  $r$  replicates based on the real data in a group:

Step1: Calculate four variances and four means from two-by-two tables over  $r$  replicates for each tag within a gene.

Step2: Choose randomly one variance and one mean with equal probability for each tag within a gene.

Step 3: Generate a two-by-two random null or baseline count data with  $r$  replicates using negative binomial pseudorandom generator in R environment with the mean as  $\mu$  and the variance as dispersion (size).

Step 4: Repeat steps 1–3 until  $g = G$ .

We then adopted uniform and normal pseudorandom generators to generate count data with association effects:

$$n_{gzij} = U_{gzij} \frac{N_{gzi} N_{gzj}}{N_{gz}} \quad (12)$$

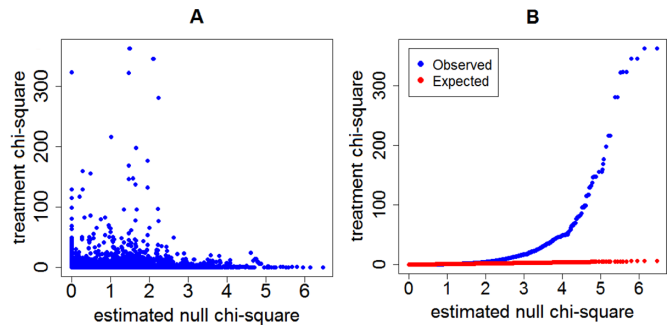
where  $N_{gzi}$  and  $N_{gzj}$  are normal variables with  $\sigma = 50$  and mean = 100 (these values are arbitrary because associate

effect does not depend on individual values) within gene  $g$ ,  $N_{gz} = \sum_{i=1}^2 N_{gzi} + \sum_{j=1}^2 N_{gzj}$ , and  $U_{gzij} = U_{gzi} U_{gzj}$  where  $U_{gzi}$  and  $U_{gzj}$  are uniform variables,  $0 < U_{gzi} \leq 1$  and  $0 < U_{gzj} \leq 1$ ,  $i = 1$  and  $2$ ,  $j = 1$  and  $2$ ,  $z = 1, \dots, Z_g$ . The association effect  $n_{gzij}$  is randomly assigned to 10, 20 and 30% of null tags (the null data above). Using these simulated datasets, we compared the estimated FDR given by a statistical method to its true FDR across a set of given thresholds ( $\Delta = U\chi_{S^*}^2(e)$ ,  $0 \leq U \leq 1$ ).

Supplementary Figure S1 summarizes these results in the simulated scenarios where 10% (A1 and B1), 20%, (A2 and B2) and 30% (A3 and B3) of tags have association effects on transcription with cell states and  $r = 6$ ,  $x = 0$  and  $x = (2r/\chi_{S^*}^2(e))^2$  where  $\chi_{S^*}^2(e) = \max(\chi_s^2(e))$ . With  $x = 0$  (Supplementary Figure S1A1–3), FDR was significantly overestimated in all three given scenarios such that many tags with true association effects would be missed at FDR  $< 0.05$  level. In contrast, with  $x = (2r/\chi_{S^*}^2(e))^2$  (Supplementary Figure S1B1–3), the estimated FDR curve is slightly higher than the true one in each of these three scenarios, finding tags with a true association effect at a given threshold with higher power. Supplementary Figure S1C1–3 is obtained with  $x = (2r/\chi_{S^*}^2(e))^2$  from the simulated data with  $r = 3$  and 10, 20 and 30% of tags having association effect, respectively. The fact that these FDR curves are very similar to those in Supplementary Figure S1B1, B2 and B3 indicates that our estimate of FDR profile with  $x = (2r/\chi_{S^*}^2(e))^2$  is absolutely conservative but close to its true one across a set of given thresholds in these scenarios.

### Comparison with Pearson chi-square, Fisher exact test and CMH chi-square test approaches

In order to show strong advantages of our method, we applied Fisher exact test, Pearson chi-square test and CMH chi-square test approaches to the simulated data and used Benjamini-Hochberg (BH) multiple-testing procedure (34,35) and q-value (36,37) to estimate FDR. Since the Fisher exact test and Pearson chi-square methods do not work on the replicated data, we utilized mean counts over  $r$  replicate libraries to construct a two-by-two table for each tag. These association test methods were performed in R packages (fisher.test, chsq.test and CMH.test), the BH procedure was conducted in R package p.adjust with method = BH. Package qvalue (36,37) was downloaded from Bioconductor and its qvalue.gui() was used to calculate q-values. The  $P$ -values were obtained from application of these three approaches to our simulated data of 15391 tags with three replicates among which 10% of tags were given with association effects. The three sets of 15391  $P$ -values for Fisher exact tests, Pearson chi-square tests and CMH chi-square tests were listed in Supplementary Table S2 and adjusted by BH-procedure and q-value method. Supplementary Figure S2 displays the results of plotting estimated against true FDR values across all cutoff points. One can find from Supplementary Figure S2 that FDRs are completely overestimated by BH-procedure and q-value approach. The same case was also found in the simulated data with five replicates (data not shown). As indicated above, Pearson chi-square based on mean counts over replicates is not unbiased and



**Figure 5.** Scatter and linear plots of treatment chi-squares versus estimated null chi-squares. (A) Scatter plot of treatment chi-squares against estimated null chi-squares derived from primary experimental data. Over 80% of treatment chi-squares fall in the null chi-square distribution. (B) Linear plot of ranked treatment chi-squares against ranked estimated null chi-squares. The observed linear dots (blue line) were significantly deviated from the expected linear dots (red line) across the estimated null chi-square distribution.

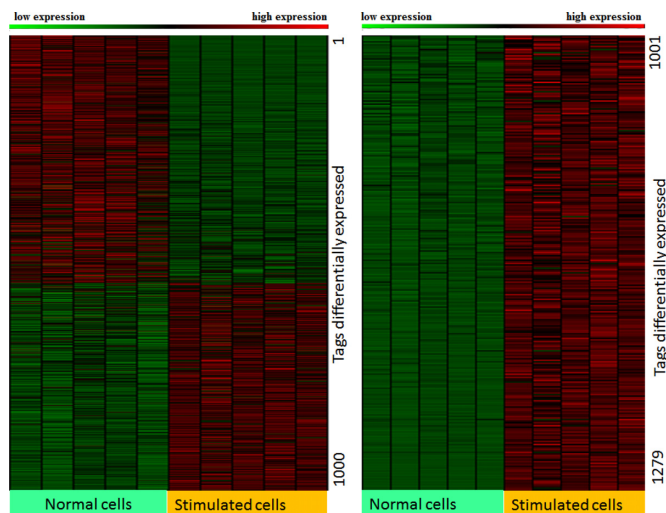
the  $P$ -value for such chi-square obtained from a chi-square distribution is greatly biased against its true value (See Supplementary Table S2). So BH and q-value procedures are not available to adjust such biased  $P$ -value profiles.

### RAX2 analysis of the primary T-cell data

We then assessed the performance of RAX2 on 3p-seq datasets derived from resting and stimulated primary human CD4<sup>+</sup> T cells (Supplementary Table S1). After normalizing and filtering, the data contain 16247 tags scattered in 10160 transcription units or genes. Transcription units with a single tag were filtered and the remainderd 9899 tags scattered in 3812 genes were available for RAX2 analysis. We performed classical chi-square and the BH-procedure on our primary T-cell transcriptomic data. As expected in Supplementary Figures S2, only 747 tags were found to be associated with cell states at FDR  $< 0.05$  (see Supplementary Table S4). We therefore applied RAX2 to the real data. The estimated null chi-square value falls in a range of 0 to 6.7 (Figure 5A). The treatment chi-square falls in interval between 0 and 363, which is larger than that in the estimated null chi-square distribution. The observed linear dots are deviated from the expected linear dots when the null chi-square is larger than 2.5 (Figure 5B).

The results obtained by applying RAX2 to our real transcriptomic data are illustrated in Supplementary Tables S3–S5. At FDR  $< 0.05$ , 1610 (16.3%) tags were found to be associated in transcription with antigen receptor stimulation (Supplementary Table S3). Among 1610 tags, 1279 have fold change  $> 1.4$  or  $< 0.715$ . Heatmap shows that these 1279 associated tags also displayed strong differential expression between rest and stimulation (Figure 6). We chose genes with two tags identified at FDR  $< 0.05$  (see Supplementary Table S6), and plotted logarithms of the ratio of transcription relative frequencies of proximal tags in stimulation to those in normal condition against those of distal tags and observed four patterns of transcription variation: forward and backward ‘switch’ changes; positively and negatively accordant changes (Figure 7A) where forward switch is defined when high expression is switched from proximal





**Figure 6.** Heatmap of tags differentially expressed between two conditions detected by RAX2. The heatmap was made with the data of 1279 tags with fold change (average of tag expressions in stimulated cells/average of tag expressions in normal cells or average of tag expressions in normal cells/average of tag expression in stimulated cells) > 1.4 selected from 1610 associated tags detected by RAX2.

site to distal site and backward switch when high expression is switched from distal site back to proximal site. We call such an association expression positively accordant expression when expression of two tags is highly raised by stimulation or negatively accordant expression when both come down due to stimulation. In our data, genes with forward switch were many fewer than genes with backward switch (148 versus 186), implicating that under stimulation, usages of many distal poly(A) sites switched to proximal poly(A) sites. In addition, positively accordant genes were many more than negatively accordant genes (186 versus 85) (Figure 7B). This means that stimulation significantly promoted transcription of tags. Figure 7C shows that proximal tags positively associated with stimulation are significantly more than distal tags positively associated with stimulation (372 versus 334). This result indicates that proximal poly(A) sites were preferably used under stimulation. This is consistent with the previous studies (18–20). Sandberg *et al.* (17) found that tandem UTR length is highly negatively correlated to proliferation ratio. That is, proliferative cells more use proximal poly(A) sites. Mayr and Bartel (16) also observed that cancer cells preferably used short 3' UTR while tissue cells more used long 3'UTR because cancer cells have high proliferation ratio but tissue cells are highly differentiated. CD3/CD28 costimulation triggered a series of physiological activations and expansion (growth) of T-cells. Figure 8 displays examples for forward and backward switches, positively and negatively accordant changes. Figure 8A is backward switch between two tags within gene PHF6 (PHD finger protein 6) potentially playing a role in transcription regulation. Stimulation made high transcription level switched from distal poly(A) site backward to proximal poly(A) site. In gene SRP68 (Signal Recognition Particle of 68 kDa, a ribonucleoprotein complex), the transcription pattern of two tags (Figure 8B) is completely in-

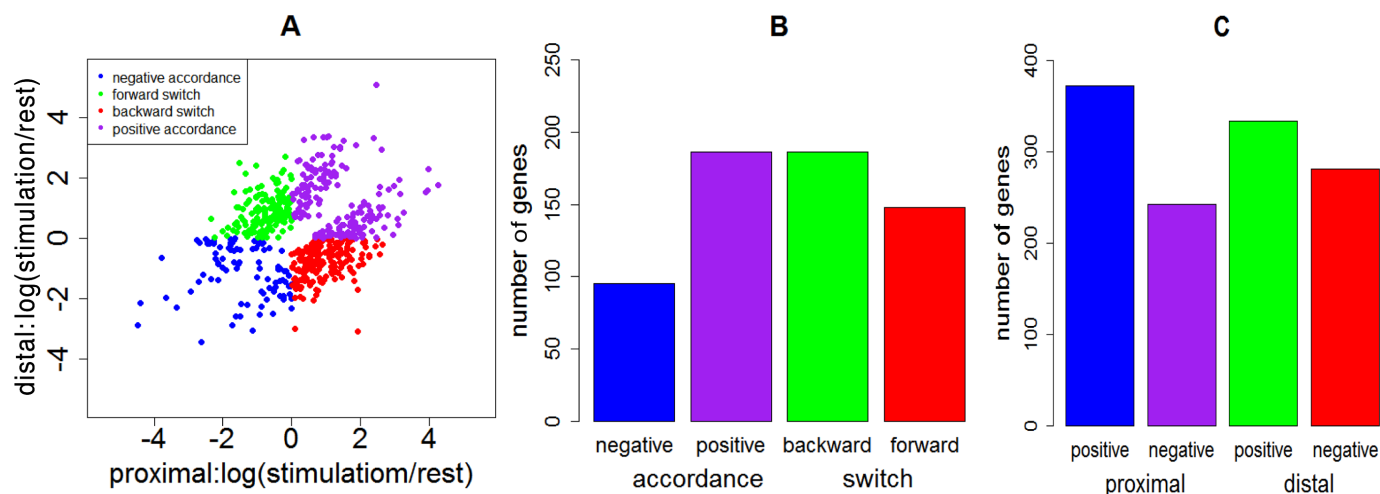
verse with that in gene PHF6, called forward switch. Figures 8C and 8D show negatively and positively accordant transcription of proximal and distal tags in genes ARL4C (ADP-Ribosylation Factor-Like 4C) and IL24 (interleukin 24). Stimulation triggered high transcription of two tags in gene IL24 but strongly suppressed transcription of proximal and distal tags in gene ARL4C.

In our result, 162 genes were found to have three poly(A) sites whose usages were associated with stimulation (Supplementary Table S8). Supplementary Figure S3 displays three scatter plots of these genes with proximal tags versus middle tags (Supplementary Figure S3A1), proximal tags versus distal tags (Supplementary Figure S3A2) and middle tags versus distal tags (Supplementary Figure S3A3). In plot of proximal versus middle tags (Supplementary Figure S3A1), accordantly changed tags were many more than switched tags (47 versus 9 in Supplementary Figure S3B1). Among accordantly changed tags, proximal and middle tags positively responding to stimulation are more than those negatively responding to stimulation (Supplementary Figure S3C1). This result suggests that stimulation indeed raised transcription of tags but did not significantly trigger transcription switch between proximal and middle poly(A) sites. However, Supplementary Figures S3A2 and S3A3 definitely show that stimulation remarkably increased transcription switch from distal poly(A) sites backward to middle poly(A) site (Supplementary Figure S3B2) or backward to proximal poly(A) sites (Supplementary Figure S3B3) while positively and negatively accordant transcriptions between proximal and distal poly(A) sites and between middle and distal poly(A) sites became very weak. In genes with proximal and middle poly(A) sites, positive tags were more than negative tags but in genes with distal poly(A) sites, negative tags were more than positive tags (Supplementary Figure S3C2–3). These results strongly indicate that, as seen in genes with two tags, in genes with three tags, CD3/28 costimulation resulted in more usage of shorter tags.

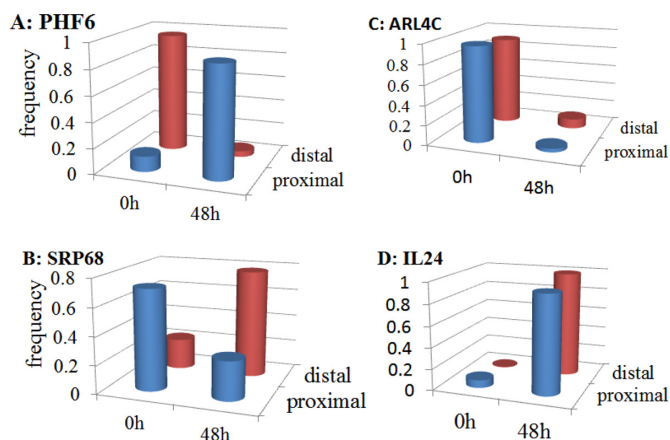
## DISCUSSION

Transcriptional profiling via NGS technologies has underscored the complexity of transcript isoform variation in mammalian cells. The extent of this variation, whether the identity and function of a protein product (e.g. alternative splicing) or the visibility of the gene product is altered due to the posttranscriptional regulatory machinery (e.g. ACP), has been broadly appreciated. It is of great interest to leverage newer technologies to globally assess the impact of these processes on human disease, particularly since several individual examples in which dysregulation of either process contributes to human disease exist (5).

The identification of relative expression changes of individual mRNA isoforms derived from a single transcription unit can in theory be performed by considering each of these isoforms from the transcription unit as a single entity and applying statistical approaches DESeq (21), baySeq (22), edgeR exact test (23,24), edgeR GLM (25), DEXSeq (26), Cuffdiff (27,28), DiffSplice (29) and SplicingCompass (30) to determine whether a given isoform is differentially expressed between two different cell states. However, these approaches do not directly assess whether isoforms within



**Figure 7.** Two-way scatter plot for four association patterns of transcription of two tags within genes between two cell states. A two-way scatter plot displays distributions of scatter dots in four phases: Phase II for forward switch, phase IV for backward switch, phase I for positive accordance and phase III for negative accordance. (A) Plot of proximal tags versus distal tags where coordinates  $x$  and  $y$  are differences between ratios of counts of tags in stimulation and those in rest state. The ratio = sum of transcription counts of a tag over all replicates in a cell state / sum of transcription counts of this tag over all replicates and all cell states. (B) Numbers of genes with four transcription patterns of proximal and distal tags. (C) Numbers of genes with proximal tags positively and negatively responses to stimulation and numbers of genes with distal tags positively and negatively response to stimulation



**Figure 8.** Examples for association patterns of transcriptional representation of two tags within genes between two cell states. Gene products shown here contain two tags defined by usage of distal and proximal poly(A) sites. Association between relative expression of two tags within genes and the cell states, detected by RAX2, shows backward switch (A), forward switch (B), negative accordant changes (C) and positive accordant changes (D). y axis corresponds to ratio of count sum of a tag over all replicate libraries in a cell state (0 h or 48 h post stimulation) to the sum of the tag over all replicate libraries across all cell states.

genes are associated with the conditions in transcription. Valid statistical approaches for doing so are Pearson chi-square test and Fisher exact test, but these methodologies cannot currently use information of variation in replicate libraries. As seen in Supplementary Appendix B, the Pearson chi-square of mean count over all given replicates is biased because mean noise cannot be excluded. CMH chi-square test approach can be applied to repeat two-by-two table data or stratified count data and indeed its power is significantly higher than Pearson chi-square tests and Fisher Exact tests (see Supplementary Table S2), so it will be promised to ex-

tend our RAX2 to CMH and to develop a multiple CHM test approach.

Therefore, we need to develop a novel chi-square test methodology. Within the context of NGS datasets where individual transcription units have a potential to be characterized by the production of multiple mRNA isoforms, it is not clear what the relationship is between replication and the  $P$ -values of chi-square statistics. To avoid this puzzle issue, a nonparametric approach is preferably chosen. This approach requires that an observed Pearson chi-square profile be compared to the null Pearson chi-square profile given a threshold. Since null chi-square profile does not exist, it must be estimated. In our methodology, we adopt a manner similar to estimation of null variance from replicate within-cell data to estimate the null chi-square profile. To choose appropriate threshold, we need to estimate FDR cutoff within which the findings declared by comparison between observed and null chi-square profiles are believed to be reliable.

While there are several alternative approaches for estimation of FDR, for example, the Benjamini-Hochberg procedure (34,35), the Benjamini-Liu procedure (38) and the Pounds-Cheng procedure (39), Tan-Xu multiprocedures (40), these FDR estimators are based on  $P$ -value profile and are not suitable to our nonparametric method. The permutation-based estimator developed by Tusher *et al.* (33) has been shown both in theory and in simulation to be handicapped by bias in small sample sizes (41,42). Storey and Tibshirani (37) proposed a  $q$ -value as a new FDR estimator. Similarly, the  $q$ -value approach is also based on  $P$ -values (36,37). The other ranking analysis methods (41–43) are designed to be applicable to continuous data sets or cannot otherwise be used to this context. To accurately estimate FDR for the chi-square findings, we here proposed a novel approach. Our approach is based on a simple principle: given a threshold, tags were declared to be associ-



ated with conditions by comparing treatment ranking chi-square profile to null ranking chi-square profile and number of false tags associated with conditions was determined by comparing null chi-square at low ranking position to that at higher ranking position. For a given threshold  $\Delta$ , a tag at  $t^* = b_{\Delta} + k$  in a linear space of the null chi-squares has growth of probability to be declared as a false positive with increment of  $k$ . As seen in Figure 4, if a treated chi-square distribution falls into the null chi-square distribution, then we cannot find  $a_{\Delta}$  value except for  $a_{\Delta} = 1$  across a set of given thresholds, and hence we also cannot find  $b_{\Delta}$  value except for  $b_{\Delta} = S$ . Thus,  $N_{\Delta} = 0, F_{\Delta} = 0$ , and FDR cannot be defined in such a case. Supplementary Figure S1 shows that estimated FDR is larger than its true FDR when threshold is small but it tends to be very close to its true value as threshold increases. This property guarantees that estimation of FDR is conservative and reliable at any threshold level.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We are indebted to Natee Kongchan and Arindam Chaudhury for library preparation and Alexander Ruch for pipeline construction. We thank two reviewers for their constructive suggestions that allowed us to greatly improve the paper.

## FUNDING

This work was funded by research grants to J.R.N. from the Cancer Prevention and Research Institute of Texas [HIRP100475] and the National Cancer Institute [CA131474], as well as a Career Development Award to J.R.N. from CA126752.

Conflict of interest statement. None declared.

## REFERENCES

- Colgan,D.F. and Manley,J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev.*, **11**, 2755–2766.
- Shepard,P.J., Choi,E.A., Lu,J., Flanagan,L.A., Hertel,K.J. and Shi,Y. (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, **17**, 761–772.
- Zhao,J., Hyman,L. and Moore,C. (1999) Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.*, **63**, 405–445.
- Shen,Y., Venu,R.C., Nobuta,K., Wu,X., Notibala,V., Demirci,C., Meyers,B.C., Wang,G.L., Ji,G. and Li,Q.Q. (2011) Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing. *Genome Res.*, **21**, 1478–1486.
- Dankwardt,S., Hentze,M.W. and Kulozik,A.E. (2008) 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.*, **27**, 482–498.
- Hunt,A.G., Xu,R., Addepalli,B., Rao,S., Forbes,K.P., Meeks,L.R., Xing,D., Mo,M., Zhao,H., Bandyopadhyay,A. et al. (2008) Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC Genomics*, **9**, 220.
- Shi,Y., Di Giammartino,D.C., Taylor,D., Sarkeshik,A., Rice,W.J., Yates,J.R. 3rd, Frank,J. and Manley,J.L. (2009) Molecular architecture of the human pre-mRNA 3' processing complex. *Mol. Cell*, **33**, 365–376.
- Ji,Z., Lee,J.Y., Pan,Z., Jiang,B. and Tian,B. (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 7028–7033.
- Lee,J.Y., Yeh,I., Park,J.Y. and Tian,B. (2007) PolyA.DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, **35**, D165–D168.
- Meyers,B.C., Vu,T.H., Tej,S.S., Ghazal,H., Matvienko,M., Agrawal,V., Ning,J. and Haudenschild,C.D. (2004) Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nat. Biotechnol.*, **22**, 1006–1011.
- Shen,Y., Ji,G., Haas,B.J., Wu,X., Zheng,J., Reese,G.J. and Li,Q.Q. (2008) Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res.*, **36**, 3150–3161.
- Tian,B., Pan,Z. and Lee,J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.*, **17**, 156–165.
- Wu,X., Liu,M., Downie,B., Liang,C., Ji,G., Li,Q.Q. and Hunt,A.G. (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 12533–12538.
- Flavell,S.W., Kim,T.K., Gray,J.M., Harmin,D.A., Hemberg,M., Hong,E.J., Markenscoff-Papadimitriou,E., Bear,D.M. and Greenberg,M.E. (2008) Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron*, **60**, 1022–1038.
- MacDonald,C.C. and McMahon,K.W. (2011) Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond. *Wiley Interdiscip. Rev. RNA*, **1**, 494–501.
- Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3' UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
- Sandberg,R., Neilson,J.R., Sarma,A., Sharp,P.A. and Burge,C.B. (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, **320**, 1643–1647.
- Wang,E.T., Sandberg,R., Luo,S., Khrebukova,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.
- Winter,J., Kunath,M., Roepcke,S., Krause,S., Schneider,R. and Schweiger,S. (2007) Alternative polyadenylation signals and promoters act in concert to control tissue-specific expression of the Opitz Syndrome gene MID1. *BMC Mol. Biol.*, **8**, 105.
- Yu,M., Sha,H., Gao,Y., Zeng,H., Zhu,M. and Gao,X. (2006) Alternative 3' UTR polyadenylation of Bzw1 transcripts display differential translation efficiency and tissue-specific expression. *Biochem. Biophys. Res. Commun.*, **345**, 479–485.
- Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
- Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
- Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.
- McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
- Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.
- Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
- Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript

- assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
29. Hu, Y., Huang, Y., Du, Y., Orellana, C.F., Singh, D., Johnson, A.R., Monroy, A., Kuan, P.F., Hammond, S.M., Makowski, L. *et al.* (2012) DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.*, **41**, e39.
  30. Aschoff, M., Hotz-Wagenblatt, A., Glatting, K.H., Fischer, M., Eils, R. and Konig, R. (2013) SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics*, **29**, 1141–1148.
  31. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
  32. Baggerly, K.A., Deng, L., Morris, J.S. and Aldaz, C.M. (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**, 1477–1483.
  33. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 5116–5121.
  34. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)* **57**, 289–300.
  35. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
  36. Dabney, A.R. and Storey, J.D. (2006) A reanalysis of a published Affymetrix GeneChip control dataset. *Genome Biol.*, **7**, 401.
  37. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 9440–9445.
  38. Benjamini, Y. and Liu, W. (1999) A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Stat. Plan. Inference*, **82**, 163–170.
  39. Pounds, S. and Cheng, C. (2006) Robust estimation of the false discovery rate. *Bioinformatics*, **22**, 1979–1987.
  40. Tan, Y.D. and Xu, H. (2014) A general method for accurate estimation of false discovery rates in identification of differentially expressed genes. *Bioinformatics*, **30**, 2018–2025.
  41. Tan, Y.D., Fornage, M. and Fu, Y.X. (2006) Ranking analysis of microarray data: a powerful method for identifying differentially expressed genes. *Genomics*, **88**, 846–854.
  42. Tan, Y.D., Fornage, M. and Xu, H. (2008) Ranking analysis of F-statistics for microarray data. *BMC Bioinformatics*, **9**, 142.
  43. Tan, Y.D. (2011) Ranking analysis of correlation coefficients in gene expressions. *Genomics*, **97**, 58–68.