A Powerful Approach for Identification of Differentially Transcribed mRNA Isoforms

Yuan-De Tan and Joel. R. Neilson

Department of Molecular Physiology and Biophysics and Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, 77030

Next generation sequencing is being increasingly used for transcriptome-wide analysis of differential gene expression. The primary goal in profiling expression is to identify genes or RNA isoforms differentially expressed between specific conditions. Yet, the next generation sequence-based count data are essentially different from the microarray data that are continuous type, therefore, the statistical methods developed well over the last decades cannot be applicable. For this reason, a variety of new statistical methods based on count data of transcript reads has been correspondingly developed. But currently the transcriptomic count data coming only from a few replicate libraries have high technical noise and small sample size bias, performances of these new methods are not desirable. We here developed a new statistical method specifically applicable to small sample count data called mBeta t-test for identifying differentially expressed gene or isoforms on the basis of the Beta t-test. The results obtained from simulated and real data showed that the mBeta t-test method significantly outperformed the existing statistical methods in all given scenarios. Findings of our method were validated by qRT-PCR experiments. The mBeta t-test method significantly reduced true false discoveries in differentially expressed genes or isoforms so that it had high work efficiencies in all given scenarios. In addition, the mBeta t-test method showed high stability in performance of statistical analysis and in estimation of FDR. These strongly suggests that our mBeta t-test method would offer us a creditable and reliable result of statistical analysis in practice.

Development of high-throughput sequencing technologies in recent years (Cloonan et al. 2008a; Cloonan et al. 2008b; Mortazavi et al. 2008)has massively been increasing genomic data and led sequencing cost to rapidly go down so that the sequencing technologies as platforms for studying gene expression or sub-gene expression have become more and more attractive (McCarthy et al. 2012). Current next generation sequencing (NGS) technologies such as RNA-seq (Cloonan et al. 2008a; Cloonan et al. 2008b; Mortazavi et al. 2008), Tag-seq (Morrissy et al. 2009), deepSAGE (t Hoen et al. 2008), SAGE-seq (Wu et al. 2010), and PAS-seq (Shepard et al. 2011) can generate short reads of sequence tags, that is, sequences of 35-300 bp that correspond to fragments of the original RNA. In particular, 3P-seq or PAS-seq(Shepard et al. 2011), a deep sequencing-based method for quantitative and global analysis of RNA polyadenylation has been used to study expression behavior of RNA isoforms in a variety of human and mouse cells.

To evaluate differential expression between conditions or cases, sequences need to be mapped to genome and annotated. After doing so, the sequence data can be transformed to count data at genomic level of interest. Although RNA-Seq can be used to study differential transcription of novel exons, splice-variants and isoforms-specific (Denoeud et al. 2008; Li et al. 2010; Pan et al. 2008; Wang et al. 2009) and allele-specific expression(Degner et al. 2009; Montgomery et al. 2010), our focus here is on differential expression of genes or isoforms due to alterative polyadenylation signals and cleavage sites in 3' untranslated regions (3'UTR) .

A RNA sample may be thought of as a RNA population and each RNA sequence as one individual. Sequencing a RNA sequence is a random process of sampling from a RNA population. If each individual RNA has equal chance to be selected for sequencing, then probability of sequencing a RNA sequence is proportional to the length of waiting time (Anders et al. 2010). Thus number of RNA read counts for a given genomic feature should follow Poisson distribution. The Poisson distribution implicates that only one parameter determines count variation of reads. However, since the Poisson model is just with respect to noise but does not deal with biological source of variation, that is, difference in transcription between samples, some counts are over-dispersed between samples under the Poisson Model. Accordingly, variation of read counts consists of two parts: noise(Poisson) and biological variability, that is, $\sigma^2 = \sigma^2_{\text{noise}} + \sigma^2_{\text{biological}} = \mu + a\mu^2$ where $\sigma^2_{\text{noise}} = \mu$ is variance of Poisson distribution and $\sigma^2_{\text{biological}}$ is biological variance, which is determined by biological effect *a*. This is just characteristic of negative binomial distribution (Anders et al. 2010; Robinson et al. 2008).

To identify differential transcription of RNA tags, many statistical methods have been proposed so far. At early stage, most of methods are based on Poisson distribution (Madden et al. 1997) or normal approximations (Kal et al. 1999; Man et al. 2000; Michiels et al. 1999), permutation (Zhang et al. 1997), beta distribution (Baggerly et al. 2003) or over-dispersed logistic linear distribution (Baggerly et al. 2004). As RNA count data have become more and more prevalent, newer statistical methods such as Exact test(Robinson

et al. 2008), empirical Bayesian method (Hardcastle et al. 2010), DESeq (Anders et al. 2010), generalized linear modeling (McCarthy et al. 2012), and likelihood ratio tests (Wang et al. 2010) have recently been developed.

Despite the development of technologies decreasing costs of sequencing, RNA-Seq experiments still remain expensive for many researchers so that RNA-Seq studies have to be limited to only a very small number of replicate libraries for each condition or case. The basic scientific need to assess differential transcription due to biological variation remains undiminished but the problem becomes complicated by the fact that different genes or transcripts may have different degrees of biological variation. There is therefore a need to estimate biological variation as reliably as possible from a few replicate libraries (McCarthy et al. 2012). The classical statistical methods are not applicable for such data. To address this problem, many methods such as empirical Bayesian method (baySeq) (Hardcastle et al. 2010), DESeq (Anders et al. 2010) and Exact test (Robinson et al. 2010b; Robinson et al. 2007; Robinson et al. 2008) adopt variation information across transcriptome and the GLM (McCarthy et al. 2012)uses similarity information between genes. However, none of these methods considers fudge effect of small sample size. So-called fudge effect is such an effect on which differences between two means are small but sample variances are also much smaller such that statistics are inflated (Tusher et al. 2001, Tan et al 2007). This is because sample size is so small that there is a big chance to give rise to small difference among replicates in a large-scale data (see Discussion Section for more detail). The fudge phenomenon broadly exists in high-throughput data, especially, in transcriptomic data

because there are a lot of very small counts. Therefore, to suppress such an effect can greatly improve performance of the statistical methods in identification of differentially expressed genes or isoforms. For doing so, we are required to develop novel methods or to modify the existing statistical methods.

Our development work is based on Beta t-test of Baggerly et al (2003) (Baggerly et al. 2003)because this method is not sensitive to data distributions (see Discussion Section). On the other hand, the Beta t-test approach optimizes weights for replicate libraries. The weighting and optimal strategy may be useful for excluding artificial or technical noise in count data and hence the genes or isoforms with better consistent counts in replicates libraries but having differential transcriptions between conditions would be identified with higher probabilities. The third, a very important point, is that it is t-test, a classical distance-variance test approach, that is clear and simple to understand gene differential expression but its fudge effect is also significant. For this reason, we are highly motivated to develop a novel beta t-statistic by which gene mRNAs to be tested can be separated into two different groups with least type I and type II errors.

Results

Statistical Methods

Here we follow the notations of Baggerly et al (Baggerly et al. 2003) but for the convenience of description, we use isoforms as features of study. However, our method is

130 available to all types of count data. Let X_i be count for an isoform of interest in library *i*.
131 Let p_i be true proportion of an isoform and N_i be total count in library *i*, that is, size of
132 library *i*. We suppose that the proportion p_i of a count in library *i* follows a beta
133 distribution,

134 $$p_i \sim Beta\ (\alpha, \beta), \tag{1}$$

135 but as mentioned above, the count for an isoform has binomial or negative binomial
136 distribution. In our current study, we consider the binomial distribution instead of negative
137 binomial distribution (see Discussion Section). Since $\hat{p}_i = X_i / N_i$ is an estimate of p_i, the
138 mean and variance of the estimated proportion for this isoform in library *i* are given by α,
139 β and N_i (see Supplemental Appendix A).

140

141 Considering the case of small sample size, we use weight to correct biases of expectation
142 and variance of estimated proportion p. Supposing that we have *m* replicate libraries in a
143 condition, the mean and variance of proportions in *m* replicate libraries can be linearly
144 combined by weights (Baggerly et al. 2003):

145 $$E\left(\sum_{i=1}^{m} w_i \hat{p}_i\right) = \sum_{i=1}^{m} w_i E(\hat{p}_i) = \frac{\alpha}{\alpha+\beta} \sum_{i=1}^{m} w_i = \frac{\alpha}{\alpha+\beta} \tag{2a}$$

146 $$V\left(\sum_{i=1}^{m} w_i \hat{p}_i\right) = \sum_{i=1}^{m} w_i^2 V(\hat{p}_i) = \frac{w_i^2 \alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}\left[\frac{1}{\alpha+\beta} + \frac{1}{N_i}\right] \tag{2b}$$

where the sum of weights over m replicates is constrained to be 1. Equation (2a) indicates that this combination has the correct mean. Using a partial derivative of variance of weighted proportions with respect to weights, solution for weight vectors can be analytically given by

$$w_i \propto \left[\frac{1}{\alpha + \beta} + \frac{1}{N_i} \right]^{-1}. \quad (3)$$

Equation (3) indicates that the weights are determined by the means and sizes of libraries. Here two extreme cases may occur: If $\alpha + \beta \to \infty$, then weight w_i is proportional to size of library i, N_i, meaning that distribution of p_i is degenerate so that there is no change in the true proportion going from sample to sample. If, on the other hand, $\alpha + \beta$ is very small, then the weights would be roughly the same for all libraries. The true optimum lies somewhat in between. With the weights, the proportion for an isoform count in a condition is now estimated by

$$\hat{p} = \sum_{i=1}^{m} w_i \hat{p}_i \quad (4)$$

and its variance is also estimated, in an unbiased fashion, by

$$\hat{V}^* = \frac{\sum_{i=1}^{m} (w_i \hat{p}_i)^2 - (\sum_{i=1}^{m} w_i^2)\, \hat{p}^2}{1 - (\sum_{i=1}^{m} w_i^2)}. \quad (5)$$

Since we have weights for all parameters (α, β, p, and V) in a condition, then an iterative search algorithm for optimal estimation of these parameters can be driven by weights (see Supplemental Appendix B).

Despite the estimation of variances of proportions in a condition is unbiased and optimized, those isoforms of extremely small counts would have very small and similar proportions in a few replicate libraries, which leads variances to be much smaller than differences between means so that the t-values are inflated (see Discussion Section). To avoid occurrence of this phenomenon, a modified estimator of variance is found to be

$$\hat{V} = \max\left[\hat{V}^{*}, \hat{V}^{\#}\right]. \tag{6}$$

In Baggerly et al(Baggerly et al. 2003), $\hat{V}^{\#}$ is given by

$$\hat{V}^{\#} = \frac{\frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m} N_i}\left(1 - \frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m} N_i}\right)}{\sum_{i=1}^{m} N_i}. \tag{7a}$$

From Equation(7a), we can find that when $\sum_{i=1}^{m} X_i = 0$, then $\hat{V}^{\#} = 0$. In addition, $\frac{\sum_{i=1}^{m} X_i}{(\sum_{i=1}^{m} N_i)^2}$ allows $\hat{V}^{\#}$ to be extremely small. In order to avoid the extreme small variance, we modify $\hat{V}^{\#}$ as

$$\hat{V}^{\#} = \frac{\frac{1+\sum_{i=1}^{m} X_i}{\frac{1}{m}\sum_{i=1}^{m} N_i}\left(1 - \frac{1+\sum_{i=1}^{m} X_i}{\frac{1}{m}\sum_{i=1}^{m} N_i}\right)}{\frac{1}{m}\sum_{i=1}^{m} N_i} \tag{7b}$$

178 $\hat{V}^{\#}$ in Equation (7b) is larger than that in Equation(7a). Equation (7b) shows that (1) lower

179 bound of $\hat{V}^{\#}$ is $\frac{m^2}{(\sum_{i=1}^{m} N_i)^2}\left(1-\frac{m}{\sum_{i=1}^{m} N_i}\right)$, not zero and (2) $\hat{V}^{\#} > \hat{V}^{*}$ when $\sum_{i=1}^{m} X_i$ is extremely

180 small. Thus $\hat{V} = \hat{V}^{\#}$ in case of extremely small $\hat{V}^{*}$.

181

182 By the above optimal estimation, we obtain $\hat{p}_A$ and $\hat{p}_B$, $\hat{V}_A$ and $\hat{V}_B$ in conditions A and B,

183 respectively. Using these estimates, the t-statistic similar to Z statistic suggested by Kar et

184 al (Kal et al. 1999) is found to be

185 $$t = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}_A + \hat{V}_B}} \qquad (8)$$

186 with degrees of freedom

187 $$df = \frac{\left(\hat{V}_A + \hat{V}_B\right)^2}{\frac{\hat{V}_A^2}{N_A - 1} + \frac{\hat{V}_B^2}{N_B - 1}} \qquad (9)$$

188 (Baggerly et al. 2003) where $N_A = \sum_{i=1}^{m_A} N_{Ai}$ and $N_B = \sum_{i=1}^{m_B} N_{Bi}$. With *df*, we can obtain p-

189 value for each t-statistic from the t-distribution. For count-based transcriptional data,

190 however, since replicate numbers are very small (for example, 3 replicate libraries), a

191 potential for "fudge" effect exists in a population. Although Equation (6) inserts another

192 estimator of variance as a lower bound so as to avoid occurrence of zero or extremely small

193 variance, the fudge effect still exists in Equation (8) due to small sample size. To remove

194 this effect, the t-statistic is modified as

$$t^* = \frac{\rho}{\omega}\frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}_A + \hat{V}_B}}. \quad (10)$$

Here ρ is defined as $\rho = \sqrt{\phi\varphi}$ where ψ is referred to as "polar ratio" (see Appendix C1). Equation (C1) indicates that if two data sets $X_A = \{X_{A1}, \cdots, X_{Am_A}\}$ and $X_B = \{X_{B1}, \cdots, X_{Bm_B}\}$ for an isoform do not overlap, then $\psi > 1$, otherwise, $\psi \le 1$. In statistical theory, two count sets that are definitely separated have a higher probability of showing that they come from two different populations than those that are overlap. ζ is referred to as log odds ratio (see Supplemental Appendix C2). If two data sets $X_A = \{X_{A1}, \cdots, X_{Am_A}\}$ and $X_B = \{X_{B1}, \cdots, X_{Bm_B}\}$ do not overlap and have a big gap, then $\overline{X}_A < \overline{X} < \overline{X}_B$ or $\overline{X}_A > \overline{X} > \overline{X}_B$ but $\hat{\sigma}^2 >> \hat{\sigma}^2_A$ and $\hat{\sigma}^2 >> \hat{\sigma}^2_B$, leading to $\overline{X}\hat{\sigma}^2 > \overline{X}_A\hat{\sigma}^2_A + \overline{X}_B\hat{\sigma}^2_B$ and $\zeta > 1$, otherwise, $\zeta \le 1$ where $\overline{X}$ and $\hat{\sigma}^2$ are respectively overall estimated mean and variance of counts for a given isoform; $\overline{X}_i$ and $\hat{\sigma}^2_i$ are estimated mean and variance of counts for this isoform in condition *i*, *i* = 1 for A and 2 for B. For $\psi > 1$ and $\zeta > 1$, we have $\rho > 1$. The t*-statistic is potentially preferable to the t-statistic in two aspects: (1) isoforms with small counts are not easily found to have differential expression and (2) t-values with $\rho > \omega$ are inflated but those with $\rho < \omega$ are shrunken. As a result, the fudge effect is suppressed and truly differentially expressed isoforms would be found with a high probability (very low p-value). ω is a threshold whose value is inversely proportional to sample size and determined by simulated null data based on the real data. Here we simulate null data, perform our modified Beta t-test(mBeta t-test) with setting $\rho = 1$ and $\omega = 1$ for all isoforms on the null simulated data, find false DE isoforms, calculate their ρ values, then order them from the smallest to the

largest, $\rho_1 < \rho_2 < \cdots \rho_j < \cdots < \rho_k$, and calculate quantiles. We set $q_1 = 1/k, q_2 = 2/k, \cdots,$ $q_j = j/k, \cdots q_k = 1$where k is number of false discoveries in a null simulated data. Setting $q_j \geq 0.85$, then we choose $\omega = \rho_j$ value. This means that 85% false discoveries would have $\rho \leq \omega$ and be excluded. If, however, k ≤5, then $q_j \geq 0.85$ is not informative. In this case, we take $\omega = \rho_1$. This process is done on all given simulated null datasets. We choose $\overline{\omega} = \frac{1}{S}\sum_{s=1}^{S} \omega_s$ over S simulated null datasets. The p-value for each t*-value can be obtained from t-distribution using degrees of freedom given or by performing the bootstrap (Storey et al. 2005) (see Supplemental Appendix D).

Simulation comparisons

We used the 12 scenario stimulation datasets (see Simulation in Materials and Methods) to compare 6 statistical methods for identifying isoforms differentially expressed between two given conditions. The 6 statistical methods chosen here are Beta t-test (Baggerly et al. 2003), empirical Bayesian method (Hardcastle et al. 2010), Exact test (Robinson et al. 2010b; Robinson et al. 2007; Robinson et al. 2008), GLM (McCarthy et al. 2012), DESeq (Anders et al. 2010)and our new Beta t-test (mBeta t-test). The empirical Bayesian (eBayesian) method was implemented on R package baySeq and the Exact test and GLM methods on R package edgeR(Robinson et al. 2010a). DESeq was implemented by R package DESeq(Anders et al. 2010). The Beta and mBeta t-test methods were performed in Matlab. We simulated null data to determine ω value for mBeta t-test. We set FDR cutoff = 0.05 and chose estimated FDRs smaller than but closest to this cutoff as acceptable levels

for differential expressions of isoforms because the FDR cutoff of 0.05 is widely accepted in multiple tests, especially, in genome-wide studies. We counted isoforms identified to be differentially expressed by these methods, false discoveries and calculated means and standard deviations(stdev) of the numbers of findings and true FDRs for each method chosen and then summarized them in Tables 1-3.

For small condition effect (A=100) or low artificial noise proportion (Q=10%), or low proportion (P=10%) of DE isoforms, the Beta t-test method had higher power. Since mBeta t-test is a modified beta t-test, in the case of low P or small A, mBeta t-test, eBaysian, Exact test and GLM had similar powers, while in higher P and larger A scenarios, Beta and mBeta t-test had lower powers than eBayesian, Exact test and GLM. In all 12 scenarios (Tables 1-3), DESeq always had very low powers. This is because DESeq always had extremely low true FDRs, indicating that DESeq is a very conservative method that would miss many true differentially expressed isoforms in practice. Beta t-test had much higher true FDRs than its estimated FDRs in all these scenarios, meaning that in the Beta t-test method's findings, there would be much more false discoveries than estimated, so this is not a conservative method. eBayesian showed higher powers in low artificial noise proportion (Tables 1 and 3) but it also had higher true FDRs than estimated FDRs in most cases. In high artificial noise proportion (Q=30%) scenario, eBayesian performed well (Table 2). GLM showed high powers in all 12 given scenarios but its true FDRs were much larger than estimated in 9 scenarios (Table 1-3), suggesting that this method is also not conservative. In low DE isoform proportion (P=10), low artificial noise proportion (Q=10%) or small condition

effect (A=100) scenario, Exact test performed poorly because its true FDRs were much larger than its estimated values in most cases, however, in high P(30%), high Q(30%),and large A(300) scenarios, Exact test had a good performance. Similarly to DESeq, mBeta t-test also had lower true FDRs than its estimated values in all 12 given scenarios but its powers were much higher than DESeq (Tables 1-3), showing that the mBeta t-test method is conservative and powerful.

Stability is an important property of a statistical method. To rate stabilities of these statistical methods in performance, we here used standard deviations (stdev) of finding numbers and true FDRs listed in Tables 1-3 as criterion. Small stdev means that this method has a small fluctuation and hence a high stability in identification of DE isoforms, while larger stdev indicates that it has a bigger fluctuation and hence lower stability. Thus, for each scenario, we ordered these methods by using stdev from the smallest to the largest, assigned order scores (from 1 to 6) to them and averaged their order scores over 12 scenarios. Thus, the order score can be used to measure relative stability of a method: the smaller order score, the higher stability. Table 4 summarizes the results of stability analysis. For findings, mBeta t-test had the highest stability, while GLM had the lowest. Beta t-test, eBayesian, DESeq and Exact test got similar order scores and so they had proximate stabilities. For true FDR, as expected, DESeq showed the highest stability. mBeta t-test was in the second highest rank. GLM and Beta t-test were lowest. Exact test and eBayesian showed similar stabilities.

To comprehensively rate these statistical methods, we follow Tan (Tan 2011) and define work efficiency of a statistical method as

$$w = \phi\varphi \tag{11}$$

where $\phi = \frac{N_f}{NP}$ and $\varphi = 1$ if true FDR<=0.05 and $\varphi = 0$, otherwise. Here N is number of isoforms in simulated data, P, proportion of DE ioforms given in simulation, $N_P = NP$, and N_f, number of isoforms found to be differentially expressed by a statistical method. ϕ is used to measure power (ability or probability) of a statistical method for identifying a differentially expressed isoform and φ to measure conservativeness of this method. The performance of a method must be evaluated by its power and conservativeness. If a method has a high power to find DE isoforms with no conservativeness, its findings would then be unreliable and incredible; if a method has a low power with high conservativeness, then the method would loss many findings. So such two types of statistical methods would have low work efficiencies in identification of DE isoforms.

Table 5 lists work efficiencies of the 6 large-scale statistical methods in 4 pairs of scenarios. From Table 5, one can find that eBayesian and GLM had higher work efficiencies in 3 replicate libraries than in 5 replicate libraries. This is because in 5 replicate libraries, the two methods underestimated FDR at cutoff α=0.05 (Tables 1-3) so that they lost conservativeness. Beta t-test had work efficiency of zero in all scenarios. Exact test had lower efficiencies in low P, low Q, small A and 3 replicates than in high P, high Q, large A

and 5 replicates. For the DESeq and mBeta t-test approaches, the work efficiency was greatly raised with increment of sample size. This is due to the fact that the two methods promoted its power with the same conservativeness. Similar results also can be seen in condition effects A=100 versus A=300. These two methods did not significantly respond to change in proportion of DE isoforms and in artificial noise proportion. However, in all simulated scenarios, the mBeta t-test method showed the highest work efficiencies.

In order to display FDR profiles along FDR cutoff, we here plotted averaged true FDRs against averaged estimated FDRs by these compared methods from cutoff = ~ 0 to ~ 0.21 in scenario 1. To evaluate the FDR profiles of these statistical methods, we also plotted a theoretical FDR profile (a diagonal line for true FDR against true FDR) for each method. Figure 1 shows that the estimated FDR curves of eBayesian, GLM and Exact test and Beta t-test are much below their theoretical lines, indicating that these methods, especially, Beta t-test, indeed heavily underestimated their FDRs, while DESeq too much overestimated its FDRs along the cutoff. Therefore, DESeq indeed is a too stringent and too conservative method. The mBeta t-test method overestimated FDR and hence is conservative.

Real experimental data

The simulated data are generally made in a known distribution and all factors impacting on differential expressions are well controlled, hence evaluation of these statistical methods on the simulated data is conducted in ideal and known states. Obviously, such an evaluation

has a limited significance for their application. Real data are therefore required in comparison of statistical methods. However, since everything, in particular, noise distribution in real data is unknown, a direct evaluation of statistical methods is impossibly realized by comparison between true and estimated FDRs. For this reason, we here employed an indirect way for doing so. First, we compared the performances of these statistical methods on our two real datasets: Jurkat T-cell isoform transcriptomic data and Jurkat T-cell gene transcriptomic data. The Jurkat T-cell isoform dataset contains 64428 isoforms which attribute to 14603 genes. These 64428 isoforms were annotated according to alternative poly(A) and cleavage sites within genes and studied on differential transcriptions between resting and stimulating immune states each with 3 replicate libraries by using PAS-seq (Shepard et al. 2011). We used edgeR (Robinson et al. 2010a)to normalize the Jurkat T-cell isoform and gene data. After filtering, 13409 isoforms in the isoform data and 9572 genes in the gene data were available for differential expression analyses. eBayesian had no results because either it was running too long (over at least 2 days in isoform data, infinite loop might occur) or showed NA result (in gene data). GLM obtained 4376 genes and 5039 isoforms of being differentially expressed from our gene and isoform data, respectively, at FDR cutoff of 0.05. Since ratios of findings are so high (45% in the gene data and 37% in the isoform data), we did not believe that this method could work on these real data. DESeq found 261 DE genes and 287 DE isoforms at the same FDR cutoff. The ratios of findings are so low (3% in gene data and 2% in isoform data), basically, DESeq also did not work on the two transcriptomic data. The Beta t-test, mBeta t-test and Exact test methods worked and the results obtained at FDR cutoff of 0.05 are listed in Table 6.

In the next step, we compared their findings using Venn Diagram Generator (http://simbioinf.com/mcbc/applications/genevenn/). Figure 2A shows that except that the three methods shared 554 DE genes, Exact test and mBeta t-test shared 22 DE genes in their 2019 findings while the former and Beta t-test shared merely 3 DE genes in their 2760 findings. Similar result was also found in the isoform data (Fig. 2B). In addition, if a gene was found to be differentially expressed by a method only, then it is highly possible that this DE gene would be falsely discovered. Figure 3 visualizes heat maps of 10 genes identified to be differentially expressed by mBeta t-test method (Fig. 3A), 770 by Beta t-test only (Fig. 3B), and 220 by Exact test only (Fig. 3c). Indeed, the method-specific genes do not display obvious expression difference between no stimulation (NS) and 48h poststimulation. Thus, we defined no share ratio of findings (m_i / M_i where m_i is number of method *i*-specific findings, and M_i is numbers of findings identified by method *i*) as least true false discovery rate (least true FDR is corresponding to q-value defined by Storey et (Storey et al. 2003)). Using this indirect method, we obtained the least true FDRs for the findings of the Exact test, mBeta t-test and Beta t-test methods, respectively, in the two real transcriptomic datasets (Table 6). From Table 6, one can see that Exact test and Beta t-test extremely underestimated their FDRs while mBeta t-test still overestimated FDR in its findings. These are well consistent with those found in the simulated data.

Experimental Validation

qRT-PCR experiments were carried out to valid our method's findings. Our qRT-PCR experiments were done in Jurkat T-cells at rest and 48 hours after stimulation. To make our qRT-PCR results be more representative, we randomly chose the genes that were found to be up-regulated or down-regulated at 48 hours after stimulation and not to be differentially expressed by mBeta t-test. We used relative differences between stimulation and rest and relative variation coefficient (VC) (see Materials and Methods Section) to do comparison between the RNA-seq and qRT-PCR data. The results show that genes UBL3, MST123 and KIAA0465 that were found to be up-regulated to respond to stimulation (blue columns in Fig. 4A) in RNA-seq data also displayed positive response to stimulation in qRT-PCR data (red columns in Fig. 4A). Gene CD47 negatively responded to stimulation in both datasets while gene TESK2 was not detected to have significantly difference between stimulation and rest in these two datasets. In expression direction and relative expression amount, these two datasets show cc=0.9 (Pearson correlation coefficient) (Fig. 4 A), suggesting that our transcriptomic data were well consistent with qRT-PCR data. Using relative VC, we found that gene UBL3 had small expression noise at rest and stimulation in these two datasets, while genes TESK2 had bigger expression noises in the transcriptomic data (Fig. 4 B). This is why gene UBL3 was detected to be differentially expressed but TESK2 was not though both UBL3 and TESK2 had small counts of mRNA reads in transcriptomic data.

Discussion

Although our simulated data were made in the NB distributions, Beta and mBeta t-test worked well if estimation of FDR was not considered. This is because not only the NB counts can be well approximated by binomial distributions but also the frequencies (p) of counts of mRNA reads in libraries can be approximated by beta distributions. While the eBayesian, GLM, Exact test, and DESeq approaches are merely based on the NB distribution. Therefore, for real datasets whose distributions are often unknown, mBeta t-test will perform well. For example, as seen in Result Section, eBayesian and GLM, DESeq could not work on our Jurkat T-cell isoform and gene data, while Exact test, Beta t-test and mBeta t-test worked even though their results have big differences. We also applied these methods to our leukemia transcriptomic data containing 10299 genes (data not yet published), the results show that all methods chosen can work but eBayesian has very low power (it just found 165 DE genes), while GLM, Exact test and mBeta t-test identified, respectively, 780, 733 and 711 DE genes and hence performed very similarly. These strongly suggest that eBayesian and GLM are specific to the NB distribution.

In addition, in genome-wide data, especially, in transcriptomic data, sample sizes usually are very small, for example, 3 or 2 replicate libraries in each condition due to high cost and biological resource limitation. Small samples would lead to a fudge effect (Tan 2011; Tan et al. 2011; Tan et al. 2006). For example, in count data containing more than ten thousands of isoforms, two 3-replicate small-count sets would have a larger probability of showing that they would be sampled from the same distribution not only than two 5-replicate small-count sets but also than two 3-replicate big-count sets. On the other hand, in

transcriptome-wide data, small-count data have more chance to be weakly fluctuated by noise and to form extremely small within-variances than big-count data, giving rise to inflating t-statistics. For general statistical methods, the genes or isoforms with small count data in small samples would easily be found to be differentially expressed between conditions due to inflation of statistic. To address this problem, many methods developed for identifying differentially expressed genes in microarray data introduce a constant to shrink statistics. For example, In SAM (Tusher et al. 2001), two-sample t-test is modified as S-test by adding a minimized coefficient of variation S_0 of differences between two conditions to denominator. In the regularized *t* test (Baldi et al. 2001), two-sample t-test is modified by combining gene-specific variance with global average variance. The two methods shrink all t-statistics across the whole genes detected in microarrays. So they have low powers (Tan 2011). Tan et al (Tan et al. 2006) used a conditional shrinking method to address the problem of inflating t-test. But this conditional shrinking method cannot be introduced to Beta t-test because the Beta t-test is based on differences in frequencies (proportions) of tags between conditions (Baggerly et al. 2003).

Baggerly et al (Baggerly et al. 2003) employed a weighting and iteration strategy to look for an optimal estimation of parameters beta and alpha of frequency that is assumed to follow beta distribution for a tag in a condition and furthermore developed a new t-test, we called Beta t-test. Weight and optimization is a strategy for excluding artificial or technical noise in count data. Although Baggerly et al (Baggerly et al. 2003) recognized small-count effect on t-tests and tried to avoid the problem of t-value inflation using alternative variance

given in Equation (18a), our practice demonstrated that Equation (18a) does not substantially reduce the fudge effect. For this reason, we modify the alternative variances by utilizing means of total counts over all libraries in a condition for those isoforms with very small counts. Analytically, it can be seen that the alternative variance defined in Equation (18b) is larger than that in Equation (18a). Our simulation really showed that the above small-count effect on testing for differential expressions of isoforms was mostly reduced by our modified alternative variances.

In order to eliminate small sample effect, we introduced a gene-or isoform-specific variable ρ into the Baggerly et al.s' (Baggerly et al. 2003)Beta t-test. ρ is used to measure overlap between two count sets. If two count sets more overlap, then ρ becomes smaller; if two count sets separate, then ρ >1. The larger gap between two count sets, the larger ρ. In theory, two count sets that are separated have a higher probability of showing that they came from two different populations than those that overlap. Besides, if noise within count sets is large, then ζ is small, which makes ρ become small. Thus, ρ shrinks t-values of overlapped count sets and inflates t-values of separated count sets with small noise. As seen in Tables 1-3, compared to the Beta t-test method, our mBeta t-test approach did not obviously decrease its power but significantly reduce false discovery rate so that it has higher work efficiencies. Considering sample size effect, we set a threshold ω for ρ. That is, t-values are inflated with $\rho > \omega$, or shrunken with $\rho < \omega$. As a result, almost all of small t-values are compressed into a short interval close to zero but the t-values with $\rho > \omega$ are further enlarged and a region in which truly differential expressions of isoforms and

452 expression noises are mixedly distributed becomes very narrow(Fig. 5). Since p-value only
453 depends on t-value given degree of freedom, p-values with inflating t-values become
454 smaller while those with shrinking t-values become larger. Thus a lot of false discoveries
455 are also compressed into the zero neighbors so that very few false positives would be
456 found (Fig.5). Threshold ω depends on sample size. The larger sample size, the smaller ω.
457 However, when sample size is large, ω becomes very small, ability of ρ controlling false
458 discoveries becomes very weak because the gaps between two datasets are vanished and
459 there is not fudge effect in such data.

460

461 ROC is popularly used to evaluate statistical methods (Hardcastle et al. 2010) (Robinson et
462 al. 2007) but ROC has two fatal drawbacks. First, ROC cannot evaluate the FDR estimation.
463 For multiple tests, since FDRs are unknown, they must be estimated so that one can
464 determine which features have statistical significances. Precise or conservative estimation
465 of FDRs is important for an experimental scientist or statistician to choose a statistical
466 method in practice because if a method significantly underestimates FDRs in findings, then
467 it would provide much more false findings than expected or if it much overestimates FDRs,
468 we would then loss many true findings (Tan 2011). Second, in some cases, the ROC curves
469 of the methods are tightly close to each other or overlap together, meaning that these
470 methods have similar sensitivity against specificity but it does not suggest that they have
471 the same or approximate performances because they may have different estimations of
472 FDRs. For example, we employed simulated data of scenarios 1 and 4 to makeup
473 eBayesian, Exact test, DESeq and mBeta t-test ROC curves. The results show that mBeta t-

474 test performed best at specificity<0.3 in scenario 1 (Fig. 6A) or <0.5 in scenario 4 (Fig. 6B),
475 DESeq had the slight higher sensitivity than mBeta t-test when specificity > 0.3 in scenario
476 1 (Fig. 6 A) or >0.5 in scenario 4 (Fig. 6B), eBayesian had the lowest curve, Exact test and
477 GLM had almost the same curve, performed better than eBayesian when specificity >0.4
478 (Fig. 6). However, these ROC curves did not show significant difference, in particularly, in
479 scenario 4. For this reason, our evaluation of these methods chosen was based on
480 comparison between true and estimated FDRs. The simulated data showed that the
481 eBayesian and GLM methods worked well in the ideal NB distributions, lower proportions
482 of DE isoforms, smaller condition effect and smaller number of replicate libraries but they
483 performed poorly in the case in which sample sizes were larger and proportion of DE
484 isoform was higher or condition effect was bigger. This is because in such a scenario they
485 generally had a high power to find DE isoforms but, on the other hand, their FDRs are often
486 notably underestimated. Therefore, we evaluated performance of a statistical method in
487 power (ability or probability to find DE genes or isoforms) and conservativeness of FDR
488 estimation (reliability of findings). A statistical method with high power but no
489 conservativeness of FDR estimation would offer us a lot of unreliable findings or a
490 statistical method with low power but conservativeness would miss many true DE genes or
491 isoforms. That is to say, these two types of methods would have low work efficiency to
492 perform their statistical analysis of real data.

493

494 It is required to indicate that the Exact test method performs very fast, in contrast, the
495 eBayesian method (baySeq) is very intensive in computation and took a long time (maybe

infinite loop occurred in calculation of posterior probabilities when performed on our real data). Nevertheless, the current version of baySeq is not available for our real data. The mBeta t-test method is also computationally intensive because it runs literately to look for an optimal estimation of weight and beta and alpha parameters for each isoform. However, it finishes its work in 15 minutes for more 10 thousands of isoforms if we don't use bootstrap to calculate p-values.

Materials and Methods

Cell Lines and Stimulation

Jurkat T-cell lines were obtained from the ATCC and maintained in RPMI (ATCC) with 10% fetal bovine serum supplemented with penicillin and streptomycin (Gibco). Jurkat T-cells were stimulated with plate-bound antibodies coated with a solution of 1 μg anti-CD3 (OKT3 – eBiosciences) and 5 μg anti-CD28 (CD28.2 – BD Pharmingen). Activation of T-cell was monitored via flow cytometry detection of CD69 expression (FN-50) 48 hours after stimulation (Simms et al. 1996).

High-throughput Sequencing Library Generation

Total RNA was harvested from resting and stimulated cells with Trizol reagent and processed as per manufacturer instructions. Polyadenylated RNA was isolated with the Poly(A)-Purist MAG kit as per manufacturer instructions. Libraries for high-throughput

sequencing were established as described (Shepard et al. 2011) with minor modifications. Libraries were sequenced via 50 bp paired-end sequencing on an Illumina GAIIx in genome sequencing center in Baylor College of Medicine. The listed reagents were from Life Technologies.

Annotation and pipeline analyses

Paired-end reads were first subjected to a profiler removing A and T homopolymer runs within the forward and reverse reads, respectively. Pairs in which the length of both reads was greater than 25 bp were mapped to the human genome reference (UCSC hg19) with Bowtie(Langmead et al. 2009) using default parameters. The reads were simultaneously mapped to the UCSC KnownGene database to identify putative exon-spanning reads or pairs. The union of mate pairs mapping uniquely to the genome and those mapping specifically to the KnownGene database were condensed to isoforms by 3' alignment identity, filtered for false priming, and assigned gene identity and region (e.g. CDS, 3' UTR) using UCSC KnownGene annotations. Cleavage and polyadenylation sites were defined as the median genome coordinate of all reads within a 20-base pair sliding window of an adjacent read. Defined sites were carried forward for analysis only if they were present in all libraries representing an individual cellular type or state. Intralibrary isoform representation was normalized to pseudocounts using the negative binomial method(Robinson et al. 2008).

qRT-PCR

Total RNA was respectively isolated from Jurkat T-cells at rest and 48hours after stimulation using TRIZOL(Invitrogen). RNA was digested with DNase I to remove contaminating genomic DNA. One microgram of total RNA was used to generate cDNA with the ImProm-II™ Reverse Transcription System (Promega) and real-time PCR was performed in triplicates in an Eppendorf Mastercycler. qRT-PCR was performed using a pair of primers(forward and reverse primers) designed for a selected gene. $\Delta\Delta C_T$ method was used to calculate relative quantitation of qRT-PCR product on a 7500 Fast Real-Time PCR machine with SDS software (Applied Biosystems).

Simulations

To evaluate statistical properties of various approaches, we used the NB pseudorandom generator to create RNA isoform count datasets in 12 scenarios, each repeated three times for calculations of means and standard deviations. Our simulations were conducted on our Jurkat T-cell isoform data from which we took 18290 isoforms and 3 replicated libraries in each of two conditions (resting and stimulating states). We set two levels (A=100 and 300) of condition (or treatment) effect on differential transcription of isoforms and linearly assigned the effects $\tau = UA$ to differentially expressed isoforms where U is uniform variable ($U \in (0,1]$), two levels of proportions of differentially expressed isoforms: P =10 and 30%, two levels of artificial noise proportions: Q=10 and 30% and 2 levels of sample sizes: R=3 and 5 replicate libraries. Here artificial noise (also called technique or Poisson noise) indicates that the noise does not comes from experimental system error but come from

techniques such as sequencing, mapping, annotation and pipline analysis, etc. In simulated data, isoforms with averaged count <5 were filtered, thus 18162 isoforms were available for analysis.

Software

A package for implementing mBeta t-test was written in Matlab and a program for generating simulation data was written in R. They are available for request.

Acknowledgement

We are indebted to Natee Kongchan for library preparation and Alexander Ruch and Jixin Deng for primary processing of our data sets. This work was funded by The Cancer Prevention and Research Institute of Texas (HIRP100475) and the National Cancer Institute (CA126752, CA131474).

References

Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106.

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.

Baggerly, K.A., Deng, L., Morris, J.S., and Aldaz, C.M. 2004. Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics* **5:** 144.

Baldi, P. and Long, A.D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17:** 509-519.

582 Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L.,
583 Wani, S., Bethel, G. et al. 2008a. Stem cell transcriptome profiling via massive-scale mRNA
584 sequencing. *Nat Methods* **5:** 613-619.
585 Cloonan, N. and Grimmond, S.M. 2008b. Transcriptome content and dynamics at single-nucleotide
586 resolution. *Genome Biol* **9:** 234.
587 Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. 2009. Effect of
588 read-mapping biases on detecting allele-specific expression from RNA-sequencing data.
589 *Bioinformatics* **25:** 3207-3212.
590 Denoeud, F., Aury, J.M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G.,
591 Wincker, P., Scarpelli, C. et al. 2008. Annotating genomes with massive-scale RNA sequencing.
592 *Genome Biol* **9:** R175.
593 Hardcastle, T.J. and Kelly, K.A. 2010. baySeq: empirical Bayesian methods for identifying differential
594 expression in sequence count data. *BMC Bioinformatics* **11:** 422.
595 Kal, A.J., van Zonneveld, A.J., Benes, V., van den Berg, M., Koerkamp, M.G., Albermann, K., Strack, N.,
596 Ruijter, J.M., Richter, A., Dujon, B. et al. 1999. Dynamics of gene expression revealed by
597 comparison of serial analysis of gene expression transcript profiles from yeast grown on two
598 different carbon sources. *Mol Biol Cell* **10:** 1859-1872.
599 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of
600 short DNA sequences to the human genome. *Genome Biol* **10:** R25.
601 Li, J., Jiang, H., and Wong, W.H. 2010. Modeling non-uniformity in short-read rates in RNA-Seq data.
602 *Genome Biol* **11:** R50.
603 Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H., and Beaudry, G.A. 1997. SAGE transcript profiles for
604 p53-dependent growth regulation. *Oncogene* **15:** 1079-1085.
605 Man, M.Z., Wang, X., and Wang, Y. 2000. POWER_SAGE: comparing statistical tests for SAGE
606 experiments. *Bioinformatics* **16:** 953-959.
607 McCarthy, D.J., Chen, Y., and Smyth, G.K. 2012. Differential expression analysis of multifactor RNA-Seq
608 experiments with respect to biological variation. *Nucleic Acids Res* **40:** 4288-4297.
609 Michiels, E.M., Oussoren, E., Van Groenigen, M., Pauws, E., Bossuyt, P.M., Voute, P.A., and Baas, F. 1999.
610 Genes differentially expressed in medulloblastoma and fetal brain. *Physiol Genomics* **1:** 83-91.
611 Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and
612 Dermitzakis, E.T. 2010. Transcriptome genetics using second generation sequencing in a
613 Caucasian population. *Nature* **464:** 773-777.
614 Morrissy, A.S., Morin, R.D., Delaney, A., Zeng, T., McDonald, H., Jones, S., Zhao, Y., Hirst, M., and Marra,
615 M.A. 2009. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res*
616 **19:** 1825-1835.
617 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying
618 mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621-628.
619 Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. 2008. Deep surveying of alternative splicing
620 complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413-
621 1415.
622 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. 2010a. edgeR: a Bioconductor package for differential
623 expression analysis of digital gene expression data. *Bioinformatics* **26:** 139-140.
624 Robinson, M.D. and Oshlack, A. 2010b. A scaling normalization method for differential expression
625 analysis of RNA-seq data. *Genome Biol* **11:** R25.
626 Robinson, M.D. and Smyth, G.K. 2007. Moderated statistical tests for assessing differences in tag
627 abundance. *Bioinformatics* **23:** 2881-2887.
628 Robinson, M.D. and Smyth, G.K. 2008. Small-sample estimation of negative binomial dispersion, with
629 applications to SAGE data. *Biostatistics* **9:** 321-332.

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.

Simms, P.E. and Ellis, T.M. 1996. Utility of flow cytometric detection of CD69 expression as a rapid method for determining poly- and oligoclonal lymphocyte activation. *Clin Diagn Lab Immunol* **3:** 301-304.

Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100:** 9440-9445.

Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., and Davis, R.W. 2005. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A* **102:** 12837-12842.

t Hoen, P.A., Ariyurek, Y., Thygesen, H.H., Vreugdenhil, E., Vossen, R.H., de Menezes, R.X., Boer, J.M., van Ommen, G.J., and den Dunnen, J.T. 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* **36:** e141.

Tan, Y.D. 2011. Work efficiency: A new criterion for comprehensive comparison and evaluation of statistical methods in large-scale identification of differentially expressed genes. *Genomics* **98:** 390-399.

Tan, Y.D. and Fornage, M. 2011. Effects of genetic and environmental factors and gene-environment interaction on expression variations of genes related to stroke in rat brain. *American Journal of Molecular Biology* **1:** 87-113

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.

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.

Wang, L., Feng, Z., Wang, X., and Zhang, X. 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26:** 136-138.

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57-63.

Wu, Z.J., Meyer, C.A., Choudhury, S., Shipitsin, M., Maruyama, R., Bessarabova, M., Nikolskaya, T., Sukumar, S., Schwartzman, A., Liu, J.S. et al. 2010. Gene expression profiling of human breast tissue samples using SAGE-Seq. *Genome Res* **20:** 1730-1739.

Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., and Kinzler, K.W. 1997. Gene expression profiles in normal and cancer cells. *Science* **276:** 1268-1272.

Figure legends

Figure 1 Profiles of true versus estimated FDR

Estimated curve was made by plotting estimated FDR against true FDR along cutoff of ~0 to ~ 0.21 and theoretical line is a diagonal line made by plotting true FDR against true FDR along the same cutoff. The true FDR was calculated by counting false positives in findings of a statistical method at an FDR cutoff point in a

simulated dataset and the estimated FDR was predicted by a statistical method. The true and estimated FDRs given figure 1 were averaged over three simulated datasets in scenario 1.

Figure 2 Venn diagram analyses of the Exact test, Beta t-test and mBeta t-test approaches

A: DE genes found in Jurkat T-cell gene transcriptomic data. B: DE isoforms found in Jurkat T-cell isoform transcriptomic data.

Figure 3 Heatmaps of method-specific differentially expressed genes

NS.A, NS.B, NS.C are replicate libraries A, B and C in resting state (no stimulation). 48h.A, 48h.B and 48h.C are replicate libraries A, B and C in stimulating state (48h poststimulation). Method-specific DE genes were treated as false DE genes. Heatmaps show that these genes really do not have obvious expression difference between resting and stimulating states. A: 10 DE genes found by mBeta t-test only. B: 770 DE genes found by Beta test only. C: 220 DE genes found by Exact test only.

Figure 4 Comparison between RNA-seq data and qRT-PCR data on 5 genes chosen.

In RNA-seq data genes UBL3, MST123, CD47 and KIAA0465 were found to be differentially expressed between rest and stimulation. TESK2 had no differential expression. Genes UBL3, MST123, KIAA0465 were significantly higher at stimulation than at rest, while genes CD47 and BCL11B were significantly lower at stimulation than at rest.

A: Relative expression comparisons of 5 genes between RNA-seq and qRT-PCR. In RNA-Seq data, relative expression of a gene is defined as $d_g / \bar{d}$ where $d_g = \bar{n}_{gt} - \bar{n}_{g0}$, $\bar{d}$ is averaged value, g= UBL3, MST123, CD47, KIAA0465 and TESK2, $\bar{n}$ is averaged count of reads over three replicates and t =48 hour. In qRT-PCR, the relative expression of a gene is defined as $\Delta_g - \Delta_b$ where $\Delta_g = \overline{CT}_{gt} - \overline{CT}_{g0}$, b = background gene for control, and $\overline{CT}$ is averaged CT value over three replicates and CT is log2 transformed threshold value of amplification in qRT-PCR. In our experiment, we used TBP (TATA binding protein) as background expression because it has no change in expression with time.

B: Relative expression variation coefficient comparison of 5 genes between RNA-seq and qPCR data. Relative expression variation coefficient is defined as $VC_{gt} / \overline{VC}_t$ where $VC_{gt} = \bar{n}_{gt} / s_{gt}$ and $\overline{VC}_t$ is averaged variation coefficient over all selected genes at time t of stimulation.

Figure 5 Plots of t-statistics versus log FC.

logFC=log (mean in 48h poststimulation/mean in NS) and t-values were given by two beta t-test methods from the simulated data in which 10% of isoforms randomly assigned with condition effect $\tau \leq 100$ were

differentially expressed between two conditions each with three replicate libraries. Simulation was conducted by NB pseudorandom generator based on real isoform transcriptomic data.

A The Beta t-statistics are distributed in interval between -16 and 16, while those for no differential expressed genes in Beta t-test are distributed in interval between -4.8 and 4.8. Thus a lot of false discoveries (blue dots) are distributed in the neighboring areas of t-value ≥4.8 or t-value ≤ -4.8.

B: Plots of the mBeta t-statistics versus log FC. The t-statistic interval is enlarged from below -50 to over 100, while the t-statistics for no differential expressed genes are strongly compressed into a very narrow area close to zero. Thus a lot of false discoveries of the Beta t-test method are also moved into this area so that very few false positives (blue dots) would be found.

Figure 6 ROC comparison of statistical methods

ROC curves of the eBayesian, Exact test, GLM, DESeq and mBeta t-test methods were made from simulated datasets. Sensitivity = true positive fraction (TPF) and specificity = false positive fraction(FPF). A: simulated data came from scenario 1(proportion of differentially expressed isoforms =10%, technical noise proportion= 10% and treatment effect A= 100, sample size = 3) and B: from scenario 4(proportion of differentially expressed isoforms =30%, technical noise proportion= 10% and treatment effect A= 300, sample size = 3).

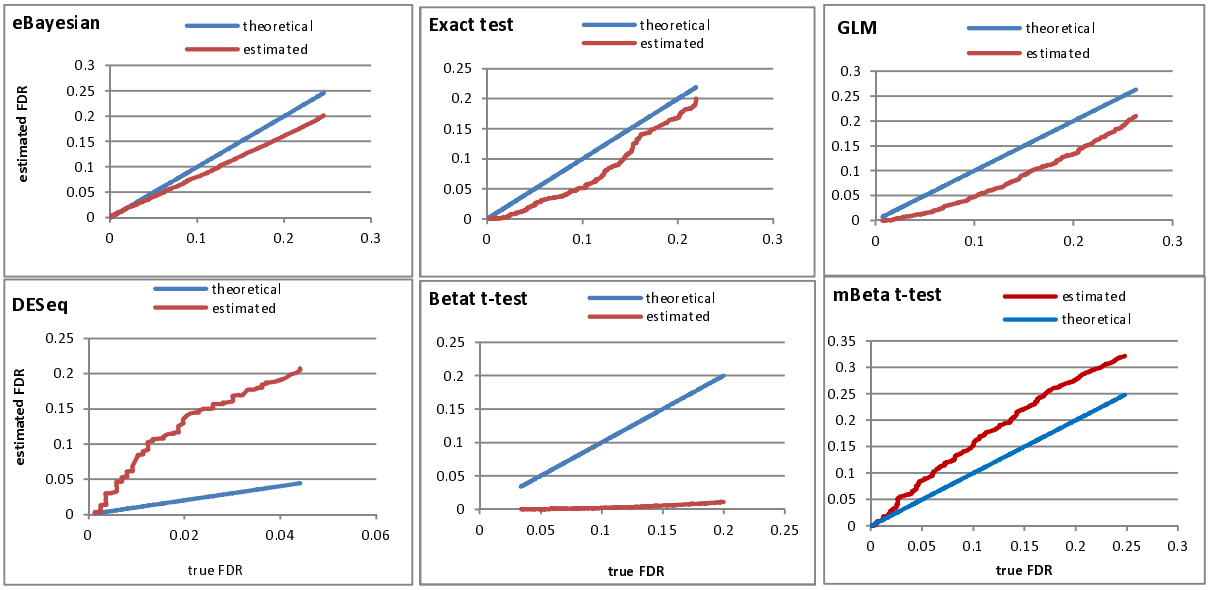


Figure 1

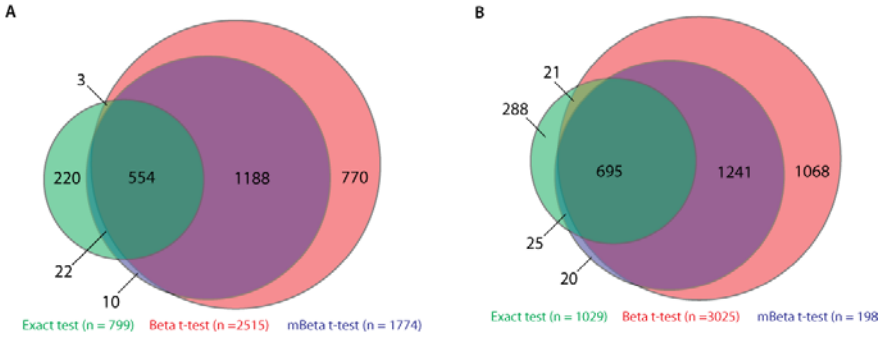
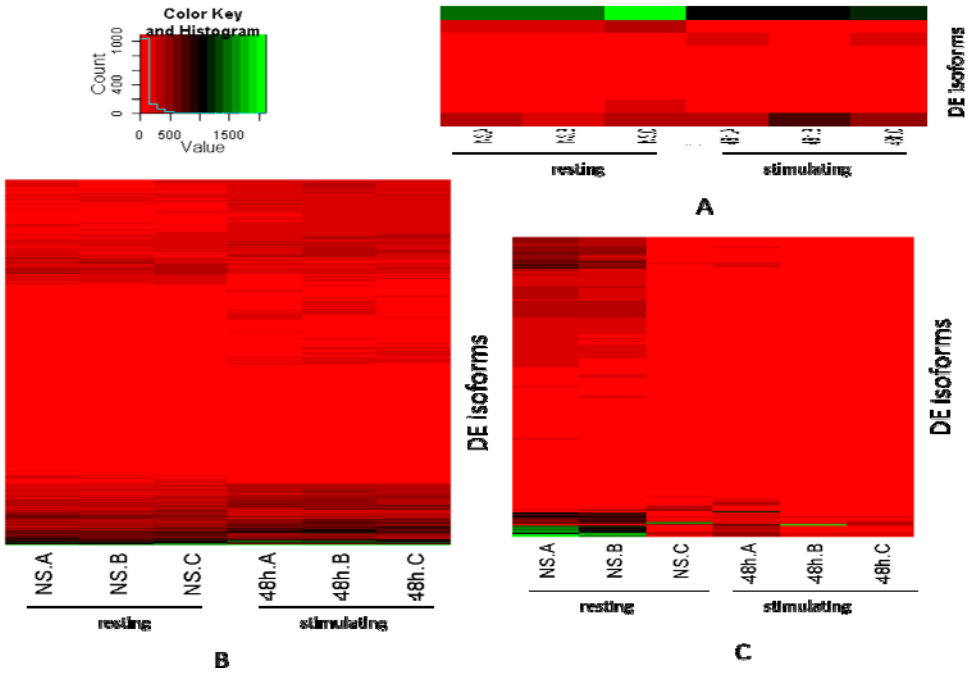


Figure 2



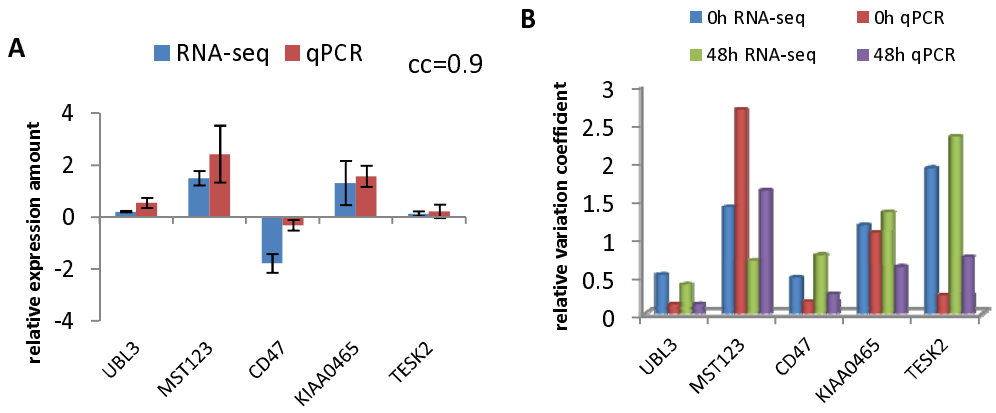
741

742

743

744 Figure 3

745



746

747 Figure 4

748

749

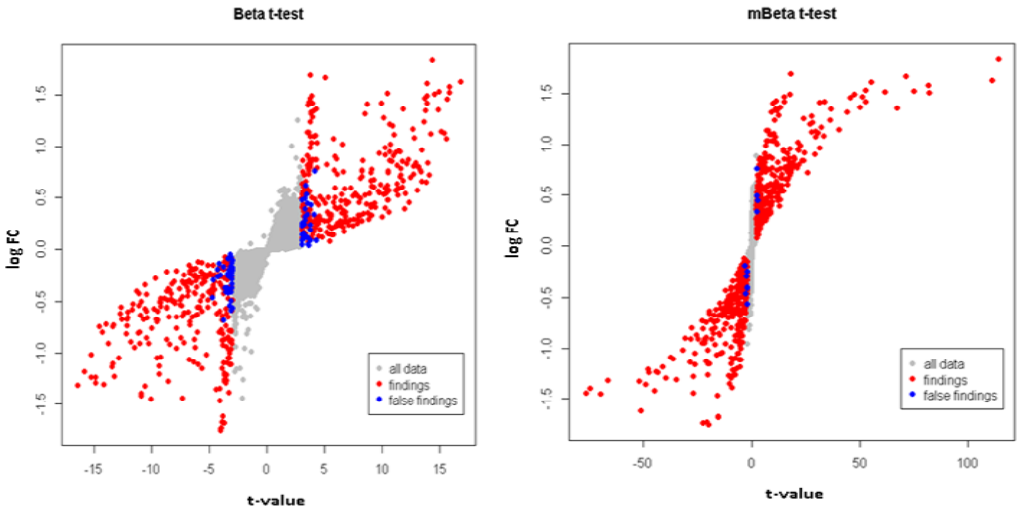


Figure 5

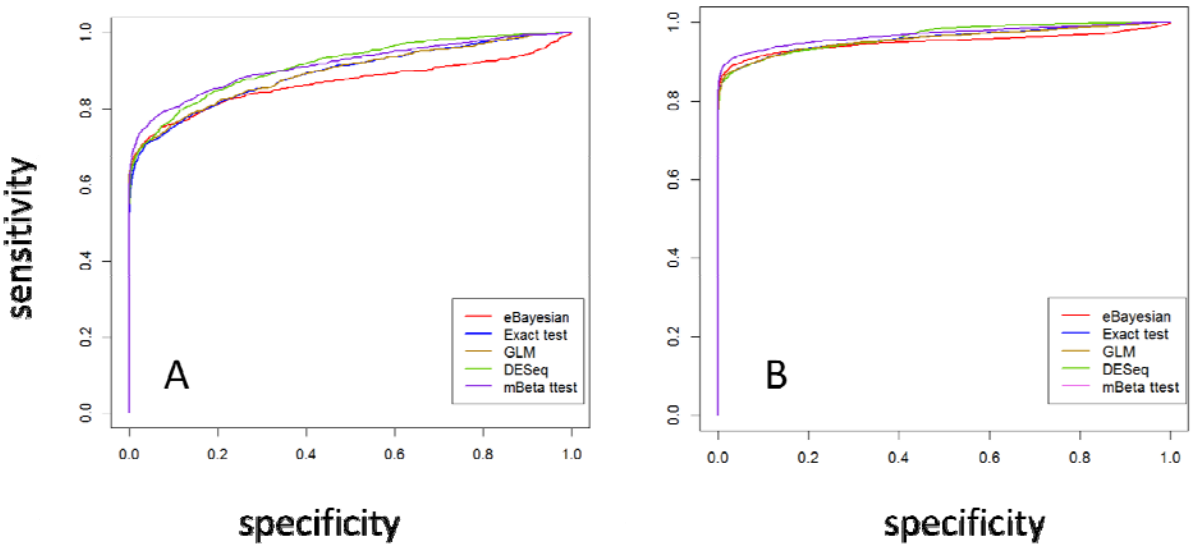


Figure 6

Table 1 Results of performing statistical methods on simulated data of 18162 isoforms and two conditions in simulation scenarios 1 - 4 where simulations were based on real transcriptomic data

scenario	Statistical Method	numbers of findings		estimated FDR	true FDR	
		mean	stdev		mean	stdev
Scenario 1	eBayesian	684.3	20.74	0.049901661	0.023415632	0.005316
	Exact test	740.3	34.70	0.049290389	0.086608166	0.009598
	GLM	753.3	36.55	0.049609439	0.096184239	0.009912
	DESeq	552.0	24.24	0.049794642	0.007797678	0.002519
	beta t-test	800. 7	4.93	0.049847667	0.144082436	0.010284
	mBeta t-test	726.0	3.46	0.048788614	0.037659939	0.008904
Scenario 2	eBaysian	990. 7	17.61	0.04970673	0.06424989	0.004605
	Exact test	1003.0	28.68	0.04971551	0.089023333	0.002429
	GLM	1015.3	31.34	0.049347345	0.097436683	0.0047
	DESeq	855.0	14.42	0.048391895	0.011703212	0.005338
	beta t-test	1000.0	25.23	0.049856733	0.155683147	0.008473
	mBeta t-test	957.3	12.50	0.048912878	0.035496106	0.007327
Scenario 3	eBayesian	2271.3	49.52	0.049921947	0.027993743	0.002909
	Exact test	2299.3	36.69	0.049936596	0.053499728	0.002031
	GLM	2316. 7	35.21	0.049880522	0.058151276	0.003686
	DESeq	1815. 7	32.02	0.049790935	0.006585122	0.00228
	beta t-test	2323.0	46.11	0.049676633	0.071477617	0.009545
	mBeta t-test	2093.0	32.90	0.049291886	0.012110596	0.002648
Scenario 4	eBayesian	3024.0	166.91	0.04991893	0.070821217	0.01752
	Exact test	2880.3	145.08	0.049928763	0.044367555	0.008163
	GLM	2896.0	150.06	0.049981628	0.048324792	0.009311
	DESeq	2517. 7	177.80	0.049904472	0.005894893	0.001385
	beta t-test	2913.0	40.92	0.049404633	0.073357708	0.007807
	mBeta t-test	2789.0	41.76	0.049425529	0.008624077	0.002894

Scenario 1: P(proportion of DE isoforms)=10%, Q(artificial noise proportion)=10%, A(condition effect) =100, R(replicate number)=3; Scenario 2: P=10%, Q=10%, A=300, R=3; Scenario 3: P=30%, Q=10%, A=100, R=3;Scenario 4: P=30%, Q=10%, A=300, R=3. $\omega = 1$ for mBeta t-test method in scenarios 1-4. FDR underestimated was marked by red color.

Table 2 Results of performing statistical methods on simulated data of 18162 isoforms and two conditions in scenarios 5 - 8 where simulations were based on real transcriptomic data

scenario	Statistical methods	number of findings		estimated FDR	true FDR	
		mean	stdev		mean	stdev
Scenario 5	eBayesian	536.0	19.31	0.049475743	0.014225	0.004023
	Exact test	582.7	40.15	0.049037986	0.053917	0.022639
	GLM	592.0	41.90	0.049785069	0.062515	0.024089
	DESeq	400.7	8.32	0.049205147	0.004114	0.003743
	Beta t-test	586.7	20.42	0.049744667	0.165699	0.010472
	mBeta t-test	608.3	37.87	0.049095016	0.045627	0.02317
Scenario 6	eBayesian	887.3	28.91	0.049840282	0.034131	0.002471
	Exact test	901.0	12.76	0.049361917	0.048886	0.005728
	GLM	911.0	13.11	0.049262212	0.054955	0.007382
	DESeq	746.3	19.08	0.047905044	0.006171	0.005225
	Beta t-test	892.0	38.22	0.0495794	0.13778	0.017347
	mBeta t-test	864.3	27.06	0.04986485	0.027286	0.005041
Scenario 7	eBayesian	1880.0	37.98	0.049924036	0.016844	0.000593
	Exact test	1905.0	63.26	0.049760675	0.033698	0.003521
	GLM	1936.3	67.33	0.04987274	0.039498	0.004252
	DESeq	1398.0	39.39	0.049872216	0.004724	0.002398
	Beta t-test	1800.7	151.30	0.0498517	0.070921	0.011422
	mBeta t-test	1887.3	74.19	0.049352367	0.018178	0.00276
Scenario 8	ebayesian	2780.3	71.62	0.049903333	0.049308	0.00472
	Exact	2653.3	97.38	0.04993212	0.033475	0.008219
	GLM	2668.3	98.19	0.049910444	0.036769	0.008319
	DESeq	2275.3	81.32	0.049506184	0.004339	0.002514
	Beta t-test	2509.7	40.82	0.0498865	0.060843	0.010993
	mBeta t test	2599.7	29.73	0.049559249	0.014071	0.004408

Scenario 5: P=10%, Q=30%, A=100, R=3; Scenario 6: P=10%, Q=30%, A=300, R=3; Scenario 7: P=30%, Q=30%, A=100, R=3;Scenario 8: P=30%, Q=30%, A=300, R=3. $\omega = 1$ for mBeta t-test method in scenarios 5-8. FDR underestimated was marked by red color.

Table 3 Results of performing statistical methods on simulated data of 18162 isoforms and two conditions in scenarios 9-12 where simulations were based on real isoform transcriptomic data

scenario	Statistical method	number of findings		estimated FDR	true FDR	
		mean	stdev		mean	stdev
Scenario 9	eBayesian	1466.7	136.99	0.049816998	0.065909	0.017935
	Exact test	1455.7	135.50	0.04942623	0.075021	0.023157
	GLM	1492.0	157.43	0.049120039	0.092843	0.032046
	DESeq	1270.7	116.81	0.04968407	0.016749	0.004976
	Beta t-test	1447.7	103.58	0.049672	0.116815	0.028366
	mBeta t-test	1457.3	114.02	0.048886388	0.03596	0.009967
Scenario 10	eBayesian	1822.3	95.61	0.049864672	0.104564	0.013078
	Exact test	1768.3	117.50	0.049594284	0.08398	0.028075
	GLM	1794. 7	131.76	0.049606781	0.094059	0.033016
	DESeq	1593.3	125.52	0.049260009	0.021667	0.012233
	Beta t-test	1772.0	67.55	0.049848	0.132159	0.025635
	mBeta t-test	1697.667	67.26	0.049301841	0.032401	0.007489
Scenario 11	eBayesian	4894. 7	289.89	0.049930905	0.081428	0.011789
	Exact test	4633.3	303.32	0.049761768	0.051046	0.013995
	GLM	4681.3	332.24	0.049960589	0.057404	0.018121
	DESeq	4206.3	317.16	0.049814792	0.013886	0.005984
	Beta t-test	4700. 7	362.29	0.0493765	0.070815	0.014574
	mBeta t-test	4438. 3	149.79	0.04952921	0.0124	0.000406
Scenario 12	ebayesian	5701.0	206.72	0.049939398	0.099753	0.01504
	Exact test	5453. 7	100.48	0.049737062	0.048492	0.006854
	GLM	5454. 7	145.40	0.049840291	0.053679	0.008784
	DESeq	4915.3	228.00	0.049430541	0.012464	0.005037
	Beta t-test	5091. 7	128.48	0.049957167	0.066915	0.013236
	mBeta t-test	5072.3	135.57	0.049633984	0.01087	0.002098

Scenario 9: P=10%, Q=10%, A=100, R=5; Scenario 10: P=10%, Q=10%, A=300, R=5; Scenario 11: P=30%, Q=10%, A=100, R=5; Scenario 4: P=30%, Q=10%, A=300, R=5. $\omega = 0.45$ for the mBeta t-test method in scenarios 9-12. FDR underestimated was marked by red color.

802

803 Table 4 Stability analysis of methods

Standard deviation of findings		Standard deviation of true FDR	
method	Averaged order score	method	Averaged order score
mBeta t-test	2.333333	DESeq	1.833333
Exact test	3.166667	mBeta t-test	2.5
Beta t-test	3.25	eBayesian	3
eBayesian	3.416667	Exact test	3.5
DESeq	3.833333	Beta t-test	5.083333
GLM	5	GLM	5.083333

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

Table 5 Averaged work efficiencies of statistical methods

	eBayesian		Exact test		GLM		DESeq		Beta t-test		mBeta t-test	
Scenario factors	mean	stdev	mean	stdev	mean	stdev	mean	stdev	mean	stdev	mean	stdev
3 replicate libraries	0.48697	0.321	0.3726	0.40689	0.27557	0.3884	0.57566	0.15142	0	0	0.69207	0.12025
5 replicate libraries	0	0	0.46259	0.53764	0	0	0.81272	0.15142	0	0	0.87067	0.07198
proportion of DE genes/isoforms=10%	0.30974	0.35327	0.13241	0.32434	0	0	0.63812	0.09414	0	0	0.75333	0.14388
proportion of DE genes/isoforms=30%	0.3281	0.36791	0.49808	0.42951	0.16917	0.31848	0.66774	0.18908	0	0	0.7625	0.14959
artificial effect proportion =10%	0.31775	0.36784	0.21165	0.42329	0.2128	0.42559	0.62857	0.19842	0	0	0.7298	0.1188
artificial effect proportion =30%	0.3281	0.36791	0.49808	0.42951	0.16917	0.31848	0.66774	0.13814	0	0	0.7625	0.14959
condition effect =100	0.3827	0.30323	0.23505	0.37554	0.09485	0.23234	0.5427	0.19842	0	0	0.66057	0.12078
condition effect =300	0.2666	0.41316	0.57015	0.44847	0.27258	0.4228	0.76666	0.1635	0	0	0.84263	0.07676

Table 6. Results of performances of 3 statistical methods on two real gene transcriptomic data

data type		Exact test	mBeta t-test[a]	Beta t-test
Gene count data	# of DE genes found	799	1774	2515
	estimated FDR	0.0499	0.0499	0.0489
	least true FDR	0.2753	0.0056	0.3062
Isoform count data	# of DE isoforms found	1029	1981	3025
	estimated FDR	0.0499	0.0498	0.0499
	least true FDR	0.2799	0.0101	0.3530

a: $\omega = 1$