Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data

Jun Li[1] Robert Tibshirani[2]

Abstract

We discuss the identification of features that are associated with an outcome in RNA-Sequencing (RNA-Seq) and other sequencing-based comparative genomic experiments. RNA-Seq data takes the form of counts, so models based on the normal distribution are generally unsuitable. The problem is especially challenging because different sequencing experiments may generate quite different total numbers of reads, or "sequencing depths". Existing methods for this problem are based on Poisson or negative-binomial models: they are useful but can be heavily influenced by "outliers" in the data. We introduce a simple, non-parametric method with resampling to account for the different sequencing depths. The new method is more robust than parametric methods. It can be applied to data with quantitative, survival, two-class, or multiple-class outcomes. We compare our proposed method to Poisson and negative-binomial based methods in simulated and real data sets, and find that our method discovers more consistent patterns than competing methods.

Keywords

Nonparametric; Differential expression; RNA-Seq; FDR; Resampling.

[1]Department of Statistics, Stanford University

[2]Departments of Health Research & Policy, and Statistics, Stanford University

Corresponding author: Jun Li, Department of Statistics, Stanford University, Stanford, CA 94305, USA. Email: junli07@stanford.edu

1 Introduction

Biological conditions and disease statuses are known to be largely characterized by differences in gene expression levels (see for instance [8, 26, 10, 6]). In the past decade, microarrays have been the primary choice for genome-wide gene expression analysis. Each array measures the expression levels of all genes from one sample, and using multiple arrays, expression levels in different samples are captured. Genes that are differentially expressed among samples can then be identified by statistical algorithms (see for example [9, 13, 19, 31]).

In the recent years, RNA-Seq has become a very competitive alternative to the microarrays (see for instance [17, 18, 25, 33, 34]). Figure 1 illustrates the process of using RNA-Seq for comparative experiments. In each experiment, mRNA are amplified, shattered, and reverse transcribed into cDNA. These short pieces of cDNA are sequenced, giving a list of short sequences called *reads*. These reads are then mapped to the reference genome using an appropriate algorithm, telling us which region each read comes from. Finally, for a set of regions of interest on the genome, such as genes, exons, or junctions, we count the number of reads mapped unambiguously to each of them, and use this count as a measure of expression of the region. This measure is a nonnegative integer, in contrast to the continuous value obtained from a microarray. In comparative experiments, RNA-Seq measurements are done for multiple samples. While the expressions of each sample are summarized by a vector of counts, the expressions of all experiments are finally put together and form a matrix, as shown in Figure 1.

Suppose that we have data from n RNA-Seq experiments, and each of them produces counts for p regions of interest. Statistically, we treat each experiment as a "sample", and each region of interest as a "feature". The data we have is an $n \times p$ matrix $\mathbf{N}$, whose element N_{ij} is the number of reads mapped to Feature j in Experiment i, $1 \leq i \leq n$, $1 \leq j \leq p$. It is important to note that the expectation of N_{ij} depends not only on the expression of Feature j, but also on the length of the read list (that is, the total number of reads) generated by Experiment i (See Figure 1). For example, if Experiment 1 and 2 use the identical biological sample (so every feature is equally expressed in the two experiments), but Experiment 1 has one million reads in total and Experiment 2 has two million reads in total, then it is likely that $N_{2j} \simeq 2N_{1j}$, for any j. Thus, counts from

different experiments are not directly comparable before being "scaled" or "normalized" properly. As a (relative) measure of the length of the list of reads, *sequencing depth* is introduced. Suppose Feature j is non-differentially expressed in Experiment $1, \ldots, n$. Using Experiment 1 as the base level whose sequencing depth is set to one, the sequencing depth of Experiment i is the ratio of expected values $\mathbf{E}(N_{ij})/\mathbf{E}(N_{1j})$, $1 \leq i \leq n$.

In comparative experiments, the goal is to correlate gene expression with an "outcome" for that sample. This outcome can be (1) two-class, such as diseased versus healthy; (2) multiple-class, like subtype A versus subtype B versus subtype C of a disease; (3) quantitative, like a continuous value measuring the virus concentration in a patient's blood; or (4) survival, like the survival time of a patient. The task of comparative experiments is to identify features that are overexpressed/underexpressed in samples in one/several classes, samples with larger quantitative outcomes, or samples that survive longer. If such overexpression or underexpression is found in a feature, we say this feature is differentially expressed.

Many methods have been developed to identify differentially expressed features from RNA-Seq data. These methods are parametric: they assume that (maybe after some simple transformation) each N_{ij} is drawn from a particular distribution, such as Gaussian ([4, 12]), Poisson ([16, 7, 32, 11, 14]), negative binomial ([23, 24, 22, 11, 1]), et al ([2, 15]). It is reported that data from technical replicates can often be well characterized by Poisson distribution ([16]), while data from biological replicates have much larger variance and negative binomial models seem to be more appropriate ([21]). More precisely, in each class, $N_{ij} \sim \text{Poisson}(d_i \cdot \nu_j)$ in technical replicates, where d_i denotes the sequencing depth of Experiment i, and ν_j denotes the expression of Feature j. In biological replicates, as samples in the same class may still have different expressions, it is better to assume $N_{ij} \sim \text{Poisson}(d_i \cdot \nu_{ij})$. However, this model contains too many parameters, so people use $N_{ij} \sim$ negative binomial$(d_i \cdot \nu_j)$ instead. Most parametric methods (e.g. [22, 1, 14]) estimate the sequencing depth d_i first, and then include it as a known term in their model. Then parametric test statistics are employed to test differential expression. As many of these methods utilize the most powerful test statistic, they are very efficient when the distributional assumption holds, even when the sample size is small. However, there is no guarantee that the real data can be well characterized by the assumed distribution.

If this distribution is a poor approximation, the results from the parametric method will not be reliable.

Here we examine the sequencing data from [35], which we denote as the "Witten data". This data set contains 58 samples, evenly assigned to two classes, and 714 features (miRNAs). 186 of the features are discarded from the analysis as they have no more than 0.5 reads averaged across samples. We apply the popular program *edgeR* ([21]), which assumes negative binomial distribution, to this data. When we plot the counts of the 20 most significant features reported by *edgeR*, we see that many of them are unlikely to follow a negative binomial distributions: one count dominates the class declared as overexpressed (we call it *leading class*), while all the other counts are very small or even zero. Figure 2 shows three of these genes. They are the 7th, 10th and 11th most significant features detected by *edgeR*, and the largest count contains 99%, 88%, and 84% of all reads in the leading class. Here, and in all similar plots in this paper, counts from different experiments are scaled by the sequencing depths. Among the 528 features, 192 of them have more than 50% of reads concentrated on only one count. Provided each class contains 29 samples, 50% is exceedingly large. It is well known that the negative binomial distribution often has its largest mass not far from the mean, so it is very unlikely that the counts follow a negative binomial distribution. If we still treat the distribution of counts as negative binomial, these large counts should be "outliers". There are possible reasons for outliers. A gene may be very highly expressed in one individual but not others. In this case, this high expression is a characteristic of this individual, and not related to the outcome. Mapping errors may also produce outliers. In Section 3, we will use simulation to show that parametric methods are very sensitive to the presence of outliers, where they generally fail to give a reasonable estimate of the false discovery rate.

Nonparametric methods are a way to finesse the difficulty of modeling counts. Without relying on underlying distributional assumption, they can give reliable results on a vast variety of data sets. In this paper, we describe a simple nonparametric method to measure the significance of features from RNA-Seq data. We also implement the usual permutation plug-in method to estimate the false discovery rate (FDR). Under various simulation scheme, we show that this statistic is competitive with the best parametric-based statistic under moderate sample size when the assumed parametric model holds.

When the assumed model does not hold, our statistic is able to select significant features much more efficiently than parametric methods. Also, in contrast to parametric methods, our method gives a reliable estimate of the FDR. On several real data sets, our method is able to find features that are expressed consistently higher in one class, and these are more likely to be biologically meaningful.

Moreover, the use of current parametric methods is limited in the outcome types that they can handle. Except for *PoissonSeq* ([14]), to our knowledge, existing methods can only be used for data with two-class outcomes. *PoissonSeq* can also be used for data with quantitative outcomes and multiple-class outcomes, but not survival outcomes. Because of the complexity of parametric methods, it is often difficult to extend them to other types of outcomes. In contrast, our nonparametric method can be used for all of the types of outcomes mentioned above. Further, the resampling strategy that we developed (Section 2.2) eliminates the difference between sequencing depths of experiments, making it easy to generalize our method to other possible types of outcomes.

The rest of this paper is organized as follows. In Section 2, we propose a nonparametric statistic for data with a two-class outcome and the associated resampling strategy, as well as a permutation plug-in method to estimate the false discovery rate (FDR). In Section 3, we study the performance of our nonparametric method on simulated data sets, and compare it with three available methods, *edgeR*, *PoissonSeq*, and *DESeq*. In Section 4, we apply our method as well as *edgeR*, *PoissonSeq*, and *DESeq* on three real RNA-Seq data sets, and compare the list of features that are called as differentially expressed by different methods. In Section 5, we extend our nonparametric statistic to other types of outcomes, and show their performance on simulated data sets. Section 6 contains the discussion.

2 A nonparametric method for two-class data

2.1 Wilcoxon statistic

For Feature j, suppose that we have counts $N_{1j}, \ldots, N_{nj}$ from either Class 1 or Class 2. Suppose Class k contains n_k samples, $k = 1, 2$ and $n_1 + n_2 = n$. Let $C_k = \{i : \text{Sample } i \text{ is from Class } k\}$, $k = 1, 2$. If the sequencing depths of all n experiments are

the same, then $N_{i_1 j} > N_{i_2 j}$ indicates the expression of Feature j is higher in Experiment i_1 than i_2. Let $R_{ij}(N)$ be the rank of N_{ij} in $N_{1j}, \ldots, N_{nj}$. Then the two-sample Wilcoxon statistic (also called the "Mann-Whitney statistic") is

$$T_j = \sum_{t \in C_1} R_{tj}(N) - \frac{n_1(n+1)}{2} \tag{2.1}$$

Here we assume that there are no ties between $N_{1j}, \ldots, N_{nj}$. The constant term is set as $-n_1(n+1)/2$ instead of $-n_1(n_1+1)/2$ (the usual definition) to make $\mathbf{E}T_j = 0$ when Feature j is not differentially expressed. A larger absolute value of T_j is stronger evidence of differential expression of Feature j, and positive/negative T_j indicates Feature j is over-expressed/under-expressed in Class 1. The Wilcoxon statistic (2.1) only depends on the ranks, and it is nonparametric.

2.2 Resampling strategy

Statistic (2.1) makes sense only if the sequencing depths of samples are the same. Otherwise, $N_{1j}, \ldots, N_{nj}$ are not comparable. Unfortunately, the sequencing depths of different samples are often very different in real data sets.

One idea to solve this problem might be to simply scale each count N_{ij} by the sequencing depth for Sample i. However we have found that this works poorly, as it does not produce counts with the appropriate amount of variation. So we use a resampling strategy instead.

Suppose the sequencing depths of the experiments are $d_1, \ldots, d_n$. Here we assume they are known. Denote $d_{\min} = \min_{i=1,\ldots,n} d_i$, and $i_{\min} = \arg\min_{i=1,\ldots,n} d_i$. That is, the $i_{\min}$th experiment has the smallest sequencing depth $d_{\min}$. Recall that N_{ij} is the number of reads mapped to Feature i in Experiment j (See Figure 2). We keep the whole list of reads generated by Experiment $i_{\min}$ unchanged, and shorten lists generated by other experiments so that they also have sequencing depth $d_{\min}$. To do this, we randomly select each read with probability $d_{\min}/d_i$, and discard it with probability $1 - d_{\min}/d_i$. After selection, the number of reads mapped to Feature j in Experiment i is

$$N'_{ij} \sim \text{binomial}(N_{ij}, d_{\min}/d_i). \tag{2.2}$$

We call this sampling method "down sampling".

It can be viewed in another way. If $N_{ij} \sim \text{Poisson}(d_i \cdot \nu_{ij})$, where ν_{ij} is the expression of Feature i in Experiment j, and we generate N'_{ij} by (2.2), then $N'_{ij} \sim \text{Poisson}(d_{\min} \cdot \nu_{ij})$ exactly. In this case, Experiment i after down sampling has the expected sequencing depth $d_{\min}$.

The Wilcoxon statistic for the down sampled data can be defined accordingly as

$$T'_j = \sum_{t \in C_1} R_{tj}(N') - \frac{n_1(n+1)}{2}, \tag{2.3}$$

where $R_{ij}(N')$ is the rank of N'_{ij} in $N'_{1j}, \ldots, N'_{nj}$.

In our experience, this down sampling method works well, but can be inefficient when $d_{\min}$ is small, as too many reads are discarded. In this case, we instead resample each experiment to a sequencing depth that is the geometric mean of the sequencing depths for all experiments. More specifically, we let $\bar{d} = (\Pi_{i=1}^n d_i)^{1/n}$, and resample using

$$N'_{ij} \sim \text{Poisson}(\frac{\bar{d}}{d_i} N_{ij}). \tag{2.4}$$

We call this sampling method "Poisson sampling". It is worth noting that even if N_{ij} follows a Poisson distribution, N'_{ij} does not. It has the expected value $\bar{d} \cdot \nu_{ij}$, but the variance is inflated by a factor of $\bar{d}/d_i + 1$. However, it turns out that this inflation does not significantly harm the performance of the method (See section 3). Generally, it is impossible to generate $\text{Poisson}(\bar{d} \cdot \nu_{ij})$ from $\text{Poisson}(d_i \cdot \nu_{ij})$ for any unknown ν_{ij} if $\bar{d} > d_i$. (Proof given by Persi Diaconis, not shown; personal communication.) Comparing down sampling and Poisson sampling on simulation data under various scheme (results not shown), we find they give very similar results when $d_{\max}/d_{\min} < 10$, and Poisson sampling is significantly better than down sampling otherwise. Hence, we use Poisson sampling hereafter. In the appendix, we give a simple theoretical analysis for the two methods, based on Pitman efficiency.

In Equation (2.3), we assumed no ties between $N'_{1j}, \ldots, N'_{nj}$. However, since they are all integers, ties may occur. To break ties, we add a small random number to each count, i.e., $N'_{ij} \leftarrow N'_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim$ i.i.d. $\text{Uniform}(0, 0.1)$, $1 \le i \le n$, $1 \le j \le p$.

In the above, we assume the sequencing depths $d_1, \ldots, d_n$ are known. In practice,

they can be accurately estimated by several methods such as *TMM* ([22]), *DESeq* ([1]), quantile normalization ([7]), and the one proposed in [14]. The last one is used in our paper: it is a simple estimate, based on the mean read count over those features that seem to be null in the data set.

2.3 Multiple resampling

The above resampling strategy makes the Wilcoxon statistic applicable, but it has two drawbacks. First, only parts of the data are used: many reads are discarded during the resampling procedure. Second, resampling, as well as adding small numbers for breaking the ties, brings randomness to the results, which might be substantial for features with small counts. These two drawbacks may finally lower the power of our nonparametric method. To minimize these limitations, we repeat the resampling S times ($S > 1$) and take the average. That is, if the rank of N'_{ij} in $N'_{1j}, \ldots, N'_{nj}$ in Resampling s is $R_{tj}(N'^s)$, we use the statistic

$$T_j^\star(\text{two-class}) = \frac{1}{S}\sum_{s=1}^{S}\left(\sum_{t\in C_1} R_{tj}(N'^s) - \frac{n_1(n+1)}{2}\right). \tag{2.5}$$

This multiple resampling strategy actually increases the power of Wilcoxon statistic defined by (2.3) by reducing its variance. In simulation data, we find that $S = 20$ is large enough to give a stable value of $T_j^\star$ and gain sufficient power.

2.4 Estimating the false discovery rate

Given $T_1^\star, \ldots, T_p^\star$, we often set a cutoff, say C, and call features with $|T_j^\star| > C$ as significant. It is important to know the accuracy of our findings. When p is large, the preferred measure of accuracy is the false discovery rate ([3]), FDR, the expected proportion of false positives in the set of features called significant. Several ways have been proposed to estimate FDR. When p-values can be easily calculated by the (exact or asymptotic) distribution of the statistic, one can use methods by Benjamini and Hochberg ([3]). When the distribution of the statistic is unknown, a popular choice is the permutation plug-in estimate ([31, 27, 30, 29, 28]), which uses permutations to generate the null distribution of the statistic.

In our cases, the distribution of Wilcoxon statistics defined by Equation (2.3) are known, but $T_j^\star$ defined by (2.5) are using the average of S Wilcoxon statistics, and its distribution is no longer known. Hence, we use the permutation plug-in method to estimate FDR in the following steps (See [30] for a detailed description).

1. Compute $T_1^\star, \ldots, T_p^\star$ based on the data.

2. Permute the n outcome values B times. In the bth permutation, compute statistics $T_1^{\star b}, \ldots, T_p^{\star b}$ based on the permuted data.

3. For a range of values of the cutpoint C, compute $\hat{V} = \frac{1}{B}\sum_{j=1}^{p}\sum_{b=1}^{B} I_{(|T_j^{\star b}|>C)}$, and $\hat{R} = \sum_{j=1}^{p} I_{(|T_j^\star|>C)}$.

4. Estimate the FDR at the cutpoint C by $\hat{\text{FDR}}_C = \hat{\pi}_0 \hat{V}/\hat{R}$.

In Step 4 above, $\hat{\pi}_0$ is an estimate of π_0, the true proportion of null features in the population. The estimation is typically made by comparing the numbers of observed and permutation statistics that fall in the non-significant range of values. We use the usual estimate $\hat{\pi}_0 = 2\sum_{j=1}^{p} I_{(|T_j^\star|\le q)}/p$, where q the median of all permuted values $|T_j^{\star b}|$, $j = 1, \ldots, p$, $b = 1, \ldots B$.

In this paper, we call our nonparametric method *SAMseq*.

3 A simulation study

Methods based on the Poisson distribution assume that

$$N_{ij} \sim \text{Poisson}(\mu_{ij}), \tag{3.1}$$

and

$$\log \mu_{ij} = \log d_i + \log \nu_j + \gamma_j I_{(j\in C_2)}. \tag{3.2}$$

Here d_i is the sequencing depth of Experiment i, ν_j captures the expression level of Feature j in the first class, and γ_j is the differential expression. Current negative binomial-distribution based methods assume

$$N_{ij} \sim \text{negative binomial}(\mu_{ij}), \tag{3.3}$$

with μ_{ij} also specified by Equation 3.2.

Li et al. ([14]) proposed a way to simulate d_i, ν_j and γ_j in Model (3.2) to mimic real data, and we employ it here. Briefly it is as follows: (1) d_i are simulated so that the total number of reads are similar to real RNA-Seq experiments and the maximum sequencing depth is about 7 times of the minimum, (2) ν_j are simulated so that the profile of gene expression levels are similar to a real RNA-Seq data set ([16]), (3) γ_j are simulated so that the average fold change for the significant features is about 2.7. $p = 20,000$ features are simulated, which is roughly the number of genes in the human genome. For the negative binomial distributed data, a constant dispersion parameter 0.25 is used. Different from Li et al. ([14]) , 30% instead of 10% of the features are set to be differential expressed so that the differences between different methods are more clearly shown. We simulate 12 samples in each of the two classes.

We next compare the performance of our method with other methods. Li et al. ([14]) did a detailed comparison of many methods on simulated data sets, including (1) their own method, *PoissonSeq*, (2) *SAM* (Significance Analysis of Microarrays, [31]) applied to the square root of normalized data, (3) The Poisson distribution based method proposed by [16], which is implemented in an R package *DEGSeq* ([1]), and (4) *edgeR* ([21]), a negative binomial based method with sequencing depths estimated by *TMM* ([22]). Of all these methods, only *edgeR* and *PoissonSeq* can give reasonable estimates of FDRs in both Poisson and negative binomial cases. So here we focus our comparison on these two methods, as well as a newly developed one called *DESeq* ([1]).

Here we give a brief introduction to *edgeR*, *DESeq*, and *PoissonSeq*. Both *edgeR* and *DESeq* assume a negative binomial distribution for the data. They estimate the dispersion parameter first, calculate the values of an (approximately) exact statistic, which are then converted to p-values by their known distribution. Finally, FDRs are estimated by Benjamini and Hochberg ([3]). The main difference between *edgeR* and *DESeq* is their different models for the dispersion parameter. (See [21] and [1] for details.) On the other hand, *PoissonSeq* always assumes Poisson distribution for the data; overdispersed data are transformed to Poisson using a simple order transformation. Score statistics are calculated, and the FDRs are obtained by a modified version of the permutation plug-in method.

3.1 Data without outliers

We first simulate Poisson and negative binomial data with no outliers. Figure 3(A, B) gives the plots of the true (solid lines) and estimated (broken lines) FDRs of the four methods. All FDR curves are the mean of 20 simulations. In the case of Poisson data (Figure 3(A)), our method and *PoissonSeq* give significantly lower true FDRs than *edgeR* and *DESeq*. While *edgeR* and *DESeq* greatly under-estimate FDRs, both our method and *PoissonSeq* slightly over-estimate FDRs. In the case of negative binomial data (Figure 3(B)), *PoissonSeq* gives slightly smaller true FDRs, and *DESeq* gives slightly higher true FDRs. While *DESeq* under-estimates FDRs a bit, the other three methods give accurate estimates of FDRs.

These simulations show that with moderate sample size, our non-parametric method gives competitive results to popular parametric methods, when the latter's distributional assumption holds.

3.2 Data with outliers

Parametric methods can work well when their assumption holds. However, real data sets often deviate from their assumed model. In particular, real data sets often contain outliers, as we have shown in the Introduction. Here we simulate data with such outliers. We still generate μ_{ij} according to (3.2), but then, we let $\mu_{ij} \leftarrow 10\mu_{ij}$ with probability 0.01. That is, 1% of the counts are outliers.

Results are shown in Figure 3(C, D). In both the Poisson and negative binomial cases, the true FDRs of *SAMseq* are only slightly higher than those in Figure 3(A, B), where no outliers present in the data. Also, the estimates of FDRs are still very accurate. So certain amount of outliers barely hurt the performance of our nonparametric method. However, the performance of the three parametric methods is a different story. Their true FDRs become unacceptably high, and further, they greatly underestimate the FDRs. Underestimating FDRs in real applications is often very dangerous. We see that parametric methods can completely fail when the underlying distribution does not hold strictly.

3.3 Data with small sample size

In the above, we simulated data sets with a moderate sample size (12 samples in each class). For even smaller data sets, we may worry about the performance of our nonparametric method, since in that case (1) nonparametric methods are often inefficient compared to parametric methods when the distributional assumption holds, and (2) the number of possible permutations is too small to generate an accurate null distribution. Here we simulate data with only 5 samples in each class, either Poisson distributed with outliers or negative binomial distributed with outliers. Except for the sample size, the other parameters are simulated by the same way as in Section 3.2.

The plots of true and estimated FDRs are shown in Figure 3(E, F). We find that the true FDRs of *SAMseq* are still much smaller than *edgeR PoissonSeq* and *DESeq*, although it is unable to differentiate among the most significant features (top ∼2,000 in Figure 3(E), and top ∼600 in Figure 3(F)), as there are too few possible values of the Wilcoxon statistic. *SAMseq* overestimates FDRs in both panels, mainly because of the overestimation of π_0. Fortunately, this overestimation should still be acceptable. The three parametric methods again fail to give reasonable estimates of FDRs.

4 Performance on real data sets

4.1 Description of the data sets

We compare *PoissonSeq*, *edgeR*, *DESeq*, and *SAMseq* on three real sequencing data sets: an RNA-Seq data set from [16], a Tag-Seq data set from [12], and an miRNA-Seq data set from [35]. For short, we call them Marioni data, t'Hoen data, and Witten data, according to the names of the first author. These three data sets use similar next-generation sequencing techniques to generate reads, although reads are mapped to different regions: genes (RNAs), tags, and miRNAs, respectively.

Marioni data contains five technical replicates in each class. The original file contains 32,000 genes, but many of them have no more than 5 reads totally. These genes are removed, leaving 18,228 for analysis. The counts in this data set are considered to be Poisson distributed with few outliers ([16]).

t'Hoen data contains four biological replicates in each class. This data set is sig-

nificantly overdispersed with outliers. The original file contains 844,316 tags (features). Filtering features with no more than 5 reads totally, 111,809 features are left for analysis.

As we introduced in the Introduction, Witten data contains 29 biological replicates in each class, and 528 features after filtering those with too small counts. This data set is also largely overdispersed with outliers.

4.2 Estimated FDRs

We apply *PoissonSeq*, *edgeR*, *DESeq* and *SAMseq* on the three datasets, and Figure 4 shows the plots of the estimated FDR curves. On Marioni data, the four curves have very similar shape, though *edgeR* and *DESeq* are unable to estimate the proportion of null genes and their largest FDRs are always 1. Also, we list the 10,000 most significant genes by each method, and count how many genes also appear in the list by other methods. We find that the overlap is $\sim$90% of each pair. These observations agree with the conclusion by [16] that this data set contains little noise, and also show that our non-parametric method performs competitively with parametric methods on Poisson distributed data with few outliers.

On the other hand, the four methods perform quite differently on t'Hoen data and Witten data, both of which are heavily overdispersed and with outliers. Note that this tells us nothing about the true FDR curves, as the estimates might be quite far from the true values (see Figure 3). We also calculate the percentage of common genes in the top calls (top 2,000 on t'Hoen data and top 150 on Witten data) by different methods. The numbers are quite low, especially between nonparametric and parametric methods ($\sim$25% on t'Hoen data and $\sim$53% on Witten data).

4.3 Different features detected by different methods

To figure out why nonparametric and parametric methods give such different results, we take a closer look at the most significant features found by them. We have checked the Witten data in the Introduction, and now we consider t'Hoen data. We find that the top features detected by the parametric methods (*edgeR*, *PoissonSeq*, and *DESeq*), and by *SAMseq* show quite different patterns. Features found by parametric methods tend to have one or two extremely large values in one class, which might be "outliers". If

we delete them, the feature can become insignificant. In contrast, top features found by *SAMseq* often have similar counts in each class, and counts in one class are *consistently* larger than the other class–for this dataset, this means that all four counts in one class are larger than any count in the other class.

Many features are detected only by *SAMseq* or only by parametric methods. In Figure 5, we show several such examples. The top left panel shows a feature from t'Hoen data detected only by *SAMseq*. Class 2 has consistently larger counts but the difference is not large enough to be detected by parametric methods. The bottom left panel shows a feature from Witten data detected only by *SAMseq*. Most values in Class 2 are lower than in Class 1, so *SAMseq* deems this feature as differentially expressed, but there are three large values in Class 2, making the mean values of the two classes almost the same, and parametric methods report that this feature is not differentially expressed. The top right panel shows a feature from t'Hoen data detected only by parametric methods. Only one sample has a non-zero count. The bottom right panel shows a feature from Witten data detected only by parametric methods. The second class have a very large count, making its mean much larger than that of the first class. However, if we delete this large count, the mean of the second class will conversely be smaller than the first class.

It seems that parametric methods, especially *edgeR* and *DESeq*, favor features with outliers. We divide features into groups according to the proportion of reads concentrating on the largest count in the leading class, and count the proportion of features that are called significant in each group. We plot their relation in Figure 6. It is clear that when the proportion is very high, *edgeR* and *DESeq* are more likely to select it as significant. On the contrary, *SAMseq* tends to detect features that have consistent expression in the leading class. For example, on Witten data, 80% of features whose proportion of reads in the largest count is larger than 80% are detected as very significant by *edgeR*, 73% are detected by *DESeq*, and 47% are detected by *PoissonSeq*; these three are much higher than their own average. On the contrary, only 13% of such features are detected to be differentially expressed by *SAMseq*.

5 Application to other types of outcomes

In the above, we discussed a nonparametric method for data with a two-class outcome. The extension to other types of outcomes is straightforward, and we give details next.

5.1 Multiple-class

Suppose there are K classes, and Class k contains n_k samples, $k = 1, \ldots, K$ and $\sum_{k=1}^{K} n_k = n$. Let $C_k = \{i : \text{ Sample } i \text{ is from Class } k\}$. We use the Kruskal-Wallis statistic (the Wilcoxon statistic for multiple classes), and the multiple resampling version is

$$T_j^\star(\text{multiple-class}) = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{12}{n(n+1)}\sum_{k=1}^{K}\frac{(\sum_{t\in C_k} R_{tj}(N'^s))^2}{n_k} - 3(n+1)\right). \tag{5.1}$$

This statistic is unsigned.

5.2 Quantitative

Here each outcome y_i is a real number, $i = 1, \ldots, n$. The statistic we use is Spearman's rank correlation coefficient, which is the (Pearson's) correlation between $R_{1j}(N'), \ldots, R_{nj}(N')$ and $R_1(y), \ldots, R_n(y)$, the ranks of $y_1, \ldots, y_n$. By using the rank of y_i, we only assumes a monotone relation between the mean of the count and the outcome, rather than a linear relation. Accordingly, the multiple sampling version is defined as

$$T_j^\star(\text{quantitative}) = \frac{1}{S}\sum_{s=1}^{S}\text{corr}(\{R_i^s(y)\}, \{R_i(y)\}). \tag{5.2}$$

5.3 Survival

Here each outcome is a pair (t_i, δ_i), where t_i is the survival time (may have ties), and δ_i is an indicator of whether the failure is observed ($\delta_i = 1$) or censored ($\delta_i = 0$), $i = 1, \ldots, n$. For Feature j, we use $R_{1j}(N'), \ldots, R_{nj}(N')$ as the single predictor in a Cox proportional hazards model. Possible ties in the survival times are handled by Breslow's method ([5]). Cox model uses the partial likelihood, which involves only the ranks of the survival times, making the model semiparametric. We use its score statistic, and the multiple resampling

version is

$$T_j^\star(\text{survival}) = \frac{1}{S}\sum_{s=1}^{S} \frac{\sum_{i=1}^n \delta_i (R_{ij}(N'^s) - B_{ij}^s/A_i)}{\sqrt{\sum_{i=1}^n \delta_i (A_i C_{ij}^s - (B_{ij}^s)^2)/(A_i^s)^2}}, \tag{5.3}$$

where $A_i = \sum_{k=1}^n I_{t_k \geq t_i}$, $B_{ij}^s = \sum_{k=1}^n R_{kj}(N'^s) I_{t_k \geq t_i}$, and $C_{ij}^s = \sum_{k=1}^n R_{kj}^2(N'^s) I_{t_k \geq t_i}$.

5.4 A simulation study

As in Section 3.2, we simulate data following a Poisson distribution with outliers and a negative binomial distribution with outliers. In all cases below, we simulate 20,000 features and 24 samples.

The simulation of data with 4-class outcomes and quantitative outcomes are the same as that in Li et al. ([14]). In the 4-class case, we simulate the mean according to $\log \mu_{ij} = \log d_i + \log \nu_j + \sum_{k=2}^4 \gamma_{jk} I_{(j \in C_k)}$, where the distribution of d_i and ν_j are the same as the two-class case, and $\gamma_{j2}, \gamma_{j3}, \gamma_{j4} \sim$ i.i.d. $N(0,1)$. We simulate 6 samples in each class.

In the quantitative case, we simulate the mean as $\log \mu_{ij} = \log d_i + \log \nu_j + \gamma_j y_j$. The distribution of d_i and ν_j are the same as the two-class case, $y_j \sim \text{Uniform}(-1, 1)$, and $\gamma_j \sim N(0,1)$.

For survival data, to simulate correlated survival time and counts, we assume that there are latent variables $y_1, \ldots, y_n$, and simulate d_i, ν_j, y_j and μ_{ij} the same as the above. Then we let the true survival time $t_i^{\text{surv}} = 1 + y_i + \text{Uniform}(0, 0.2)$, the censoring time $t_i^{\text{cens}} = \text{Uniform}(1, 2)$, and so $t_i = \min(t_i^{\text{surv}}, t_i^{\text{cens}})$, $\delta_i = I_{(t_i^{\text{cens}} \geq t_i^{\text{surv}})}$. To get some ties, we finally let $t_i \leftarrow \lfloor 20 t_i \rfloor$, where $\lfloor\rfloor$ means the integer part.

Among the four methods, only *SAMseq* and *PoissonSeq* can be used on data with multiple-class outcomes and quantitative outcomes. We find that the results (not shown) are quite like the case of two-class outcomes (Figure 3): when no outliers present, both methods estimate FDRs accurately, but when there are outliers, only *SAMseq* gives accurate estimates.

SAMseq is the only method that is applicable to data with survival outcomes. The FDR curves are shown in Figure 7. We find that *SAMseq* gives accurate estimates of FDRs.

6 Discussion

In this paper, we developed a nonparametric method that can be used on data with two-class, multiple-class, quantitative, and survival outcomes. The major strength of our method is its robustness. According to our simulation studies, the presence of outliers does not significantly hurt its performance. It overestimates the FDR in some cases, but not by very much. What is important, it never seems to significantly underestimate the FDR in our different simulation settings. In contrast, although parametric methods sometimes show better true FDRs, their estimates of FDRs can be far too low if the distributional assumption does not hold.

We argue that accurately estimating FDRs can be as important as, if not more important than, having a low actual FDR. For real data sets, the true FDRs are unknown and computational methods can only give estimated FDRs. For example, if on a data set, method 1 gives a list of 100 significant genes, and correctly estimates the FDR to be 10%. Method 2 gives a list of 200 significant genes, but underestimates the FDR from 10% to 1%. In this case, although Method 2 is more powerful in detecting significant genes, using the whole list of 200 significant genes and believing that there are only about 2 false positives could be dangerous.

Another important advantage of our nonparametric method is its simplicity. Previous methods often consider both the experimental effect (that is, different sequencing depths for the samples) and the feature effect (that is, whether different values of outcomes influence the feature expression, which is the effect we want to test) at the same time. These two effects are confounded, making the statistical model and test statistic relatively complicated. Our resampling strategy removes the experimental effects first, which simplifies the problem and makes all our test statistics one-dimensional.

The simplicity of our method makes it easy to adapt to different settings. The field of significance testing for one-dimensional problems is well developed and hence it is usually easy to find an appropriate test for a given setting. We have applied our method to data with two-class, multiple-class, quantitative and survival outcomes. We believe it will be easy to extend it to other types of outcomes and to other experimental designs.

The most significant features that we find have different "patterns" from those found by parametric methods. In the real data set we analyzed, we find that *edgeR*, *PoissonSeq*,

and *DESeq* favor features with outliers, since one outlier is sufficient to make the mean of one class much larger than the other class and completely change the value of the parametric statistic. On the other hand, one outlier changes only a little of our nonparametric statistic. The most significant features detected by our method are those whose counts are *consistently* higher in one class. Based on our discussions with biologists, features with consistent patterns are often more valuable and trustworthy.

As with other nonparametric methods, the main limitation of our method is its relatively low power for data with small sample size. In Section 3.3, we show the results for data with 5 samples in each class. When the sample size is even smaller and the underlying distribution is Poisson or negative binomial with no outliers, the true FDR of our method will be much higher than parametric methods like *PoissonSeq*, *edgeR*, and *DESeq*, although the estimated FDRs are still accurate, no matter whether outliers present. In this case, parametric methods should be preferred if the possible outliers can be carefully handled.

Our nonparametric method is computationally fast. On a Windows 7 laptop with a 2.40 GHz processor and 2 GB memory, estimating the FDR curve for 20,000 features and 12 samples on the basis of 100 permutations takes ∼23 seconds for two-class data, ∼42 seconds for 4-class data, ∼15 seconds for quantitative data, and ∼70 seconds for survival data. *PoissonSeq* takes ∼15 seconds for two-class, multiple-class, and quantitative outcomes. *DESeq* takes 10∼80 seconds for two-class data. *EdgeR* takes ∼3 minutes for two-class data. All our functions are written in `R` and they will be made freely available as an extension package for the R statistical environment ([20]).

In this paper, we have discussed the identification of differentially expressed genes in RNA-Seq data. In recent years, DNA-Seq, ChIP-Seq, 3SEQ, and other approaches related to RNA-Seq have risen in popularity. Our proposed methods should be applicable to data generated by these related technologies.

Acknowledgments

We are grateful to Persi Diaconis, Zhongyang Zhang, Jonathan Taylor, Trevor Hastie, and Hui Jiang for fruitful discussions. We also thank an anonymous reviewer for insightful comments.

The second author was supported by National Science Foundation [Grant DMS-9971405]; and National Institutes of Health [Contract N01-HV-28183].

Conflict of Interest Statement

None Declared.

References

[1] Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biology **11**, R106 (2010)

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

[3] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. **85**, 289–300 (1995)

[4] Bloom, J.S., Khan, Z., Kruglyak, L., Singh, M., Caudy, A.A.: Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics **10**, 221 (2009)

[5] Breslow, N.E.: Analysis of survival data under the proportional hazards model. International Statistical Review **43**, 45–57 (1975)

[6] Brown, P., Botstein, D.: Exploring the new world of the genome with dna microarrays. Nature genetics **21**, 33–37 (1999)

[7] Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S.: Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC Bioinformatics **11**, 94 (2010)

[8] DeRisi, J., Iyer, V., Brown, P.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**, 680–6 (1997)

[9] Dudoit, S., Yang, Y., Callow, M., Speed, T.: Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments (2000). Unpublished, available at http://www.stat.berkeley.edu/users/sandrine

[10] Eisen, M., Brown, P.: Dna arrays for analysis of gene expression. Methods in Enzymology **303**, 179–205 (1999)

[11] Hardcastle, T.J., Kelly, K.A.: bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics **11**, 422 (2010)

[12] 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., den Dunnen, J.T.: Deep sequencing-based expression analysis shows major advances in robustness,

resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res **36**(21), e141 (2008)

[13] Kerr, M., Martin, G., Churchill, G.: Analysis of variance for gene expression microarray data. Journal of Computational Biology **7**, 819–837 (2000)

[14] Li, J., Witten, D.M., Johnstone, I., Tibshirani, R.: Normalization, testing, and false discovery rate estimation for rna-sequencing data. submitted (2010)

[15] Lu, J., Tomfohr, J.K., Kepler, T.B.: Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. BMC Bioinformatics **6**, 165 (2005)

[16] Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y.: Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res **18**(9), 1509–17 (2008)

[17] Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods **5**, 621–628 (2008)

[18] Nagalakshmi, U., Wong, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M.: The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **302**, 1344–1349 (2008)

[19] Newton, M., Kendziorski, C., Richmond, C., Blatter, F., Tsui, K.: On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology **8**, 37–52 (2001)

[20] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2011). URL http://www.R-project.org. ISBN 3-900051-07-0

[21] Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–40 (2010)

[22] Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol **11**(3), R25 (2010)

[23] Robinson, M.D., Smyth, G.K.: Moderated statistical tests for assessing differences in tag abundance. Bioinformatics **23**(21), 2881–7 (2007)

[24] Robinson, M.D., Smyth, G.K.: Small-sample estimation of negative binomial dispersion, with applications to sage data. Biostatistics **9**(2), 321–32 (2008)

[25] Shendure, J.: The beginning of the end for microarrays? Nat Methods **5**(7), 585–7 (2008)

[26] Spellman, P.T., Sherlock, G., Iyer, V.R., Zhang, M., Anders, K., Eisen, M.B., Brown, P.O., Botstein D.and Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces by microarray hybridization. Mol. Cell. Biol. **9(12)**, 3273–975 (1998)

[27] Storey, J.: A direct approach to false discovery rates. Journal of the Royal Statistical Society B. **64**, 479–498 (2002)

[28] Storey, J.: The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics **31**, 2013–2025 (2003)

[29] Storey, J., Tibshirani, R.: Sam thresholding and false discovery rates for detecting differential gene expression in dna microarrays. In: The Analysis of Gene Expression Data: Methods and Software. Edited by G Parmigiani, ES Garrett, RA Irizarry and SL Zeger. Springer (2002)

[30] Storey, J., Tibshirani, R.: Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences **100**, 9440–5 (2003)

[31] Tusher, V., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. Proc. Natl. Acad. Sci. USA. **98**, 5116–5121 (2001)

[32] Wang, L., Feng, Z., Wang, X., Zhang, X.: Degseq: an r package for identifying differentially expressed genes from rna-seq data. Bioinformatics **26**(1), 136–8 (2010)

[33] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10**(1), 57–63 (2009)

[34] Wilhelm, B., Landry, J.: RNA-Seq - quantitative measurement of expression through massively parallel RNA-sequencing. Methods **48**, 249–257 (2009)

[35] Witten, D., Tibshirani, R., Gu, S.G., Fire, A., Lui, W.O.: Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biol **8**, 58 (2010)

Appendix

Pitman Efficiency is a measure for how well a test statistic performs on data. We calculate the Pitman efficiency for Wilcoxon statistic on resampled data (Equation 2.3). The following situation is considered. (We will show later that our resampled data can be approximated by this situation.)

$$X_i \sim N(b(\mu+\delta), c_1(\mu+\delta)), \quad Y_j \sim N(b\mu, c_2\mu), \tag{6.1}$$

where $b, c_1, c_2 > 0$ are known constants, μ and δ are unknown, $i, j = 1, \ldots, m$. We want to test $H_0 : \delta = 0$ v.s. $H_1 : \delta \neq 0$ by Wicoxon statistic $T = \sum_{i=1}^{m} R_i - m(m+1)/2$, where R_i is the rank of X_i among all Xs and Ys. The Pitman efficiency (PE) is defined as $\mathrm{PE} = (\frac{\mathrm{d}\mathbf{E}T}{\mathrm{d}\delta}|_{\delta=0})/\sqrt{\mathrm{var}T|_{\delta=0}}$. It is not hard to show that

$$\mathrm{PE} = \sqrt{\frac{6m^2}{\pi(2m+1)}} \cdot \frac{b}{\sqrt{(c_1+c_2)\mu}} \cdot \left[1 + 6 \cdot \frac{m-1}{m+0.5} \cdot \left(f(\sqrt{\frac{c_2}{c_1}}) + f(\sqrt{\frac{c_1}{c_2}})\right)\right]^{-1/2}, \tag{6.2}$$

where $f(x) = P(U > xV, U > xW) - \frac{1}{3}$, with $U, V, W \sim$ i.i.d. $N(0,1)$. When $x > 0$, $f(x) \in (-\frac{1}{6}, \frac{1}{6})$, and when $x = 1$, $f(x) = 0$, but there is no closed form generally; we estimate it by simulation.

Now we show that under a much simplified scheme and use Gaussian approximation, our resampled data can be expressed as (6.1). Suppose each class contains the same number of samples, $m = n/2$. Let every sample in each class has the same sequencing depth. Suppose $N_{ij} \sim \text{Poisson}(d_{\max} \cdot \nu_j)$ if $i \in C_1$, and $N_{ij} \sim \text{Poisson}(d_{\min} \cdot \nu)$ if $i \in C_2$. Here ν_j is the expression of Feature j (assuming to be the same in all samples), $d_{\max}$ and $d_{\min}$ are the sequencing depths, with $d_{\max} \geq d_{\min}$. To study which resampling method is preferred for different values of $d_{\max}/d_{\min}$, we reparametrize using $\mu_j = \bar{d} \cdot \nu_j = \sqrt{d_{\max} d_{\min}}$ and $D = d_{\max}/d_{\min}$, then the Poisson means of the two classes are $D^{1/2}\mu_j$ and $D^{-1/2}\mu_j$.

If down sampling is applied, $N'_{ij} \sim \text{Poisson}(D^{-1/2}\mu_j)$, $i = 1, \ldots, n$. So N'_{ij} in either class has mean $D^{-1/2}\mu_j$ and variance $D^{-1/2}\mu_j$. If Poisson sampling is applied, then for $i \in C_1$, $\mathbf{E}N'_{ij} = \mathbf{E}_{N_{ij}}(\mathbf{E}_{N'_{ij}|N_{ij}} N'_{ij}) = \mathbf{E}_{N_{ij}}(\frac{\bar{d}}{d_{\max}} N_{ij}) = \bar{d}\nu_j = \mu_j$, and $\text{var}N'_{ij} = \text{var}_{N_{ij}}(\mathbf{E}_{N'_{ij}|N_{ij}} N'_{ij}) + \mathbf{E}_{N_{ij}}(\text{var}_{N'_{ij}|N_{ij}} N'_{ij}) = \text{var}_{N_{ij}}(\frac{\bar{d}}{d_{\max}} N_{ij}) + \mathbf{E}_{N_{ij}}(\frac{\bar{d}}{d_{\max}} N_{ij}) = (\frac{\bar{d}}{d_{\max}} + 1) \cdot \bar{d}\nu_j = (1 + D^{-1/2})\mu_j$. Similarly, we get that for $i \in C_2$, $\mathbf{E}N'_{ij} = \mu_j$ and $\text{var}N'_{ij} = (1 + D^{1/2})\mu_j$.

Thus, if we approximate N'_{ij} by a Gaussian distribution with mean $\mathbf{E}N'_{ij}$ and variance $\text{var}N'_{ij}$, then (1) N'_{ij} generated by down sampling can be written in the form of (6.1) with $b = c_1 = c_2 = D^{-1/2}$ and $\mu = \mu_j$, (2) N'_{ij} generated by Poisson sampling can be written in the form of (6.1) with $b = 1$, $c_1 = 1 + D^{-1/2}$, $c_2 = 1 + D^{1/2}$ and $\mu = \mu_j$. Plugging them in to 6.2, we get the relative efficiency of Poisson sampling to down sampling

$$\frac{\text{PE}_{\text{Poisson sampling}}}{\text{PE}_{\text{down sampling}}} = \frac{\sqrt{2}}{1 + D^{-1/2}} \cdot \left[1 + 6 \cdot \frac{m-1}{m+0.5} \cdot (f(D^{-1/4}) + f(D^{1/4}))\right]^{-1/2}.$$

By simulation, we find that $f(D^{-1/4}) + f(D^{1/4})) \in (-0.01, 0.07)$ for any $D > 1$. So the second term is roughly 1 for any value of D and sample size m. The relative efficiency is monotone increasing as D increases, and crosses 1 when $D \simeq 6$. This indicates that under this simplified situation, Poisson sampling is less efficient than down sampling when $D < 6$, and more efficient for larger values of D.

The above analysis of Pitman efficiency is only on a simplified case. It is not clear to what extent these results will hold on real data, given that Gaussian distribution can be a poor approximation for count data, and each experiment has a different sequencing depth on real data. It is also not clear to what extent the results will hold for Wilcoxon statistic under multiple resampling (Equation 2.5).

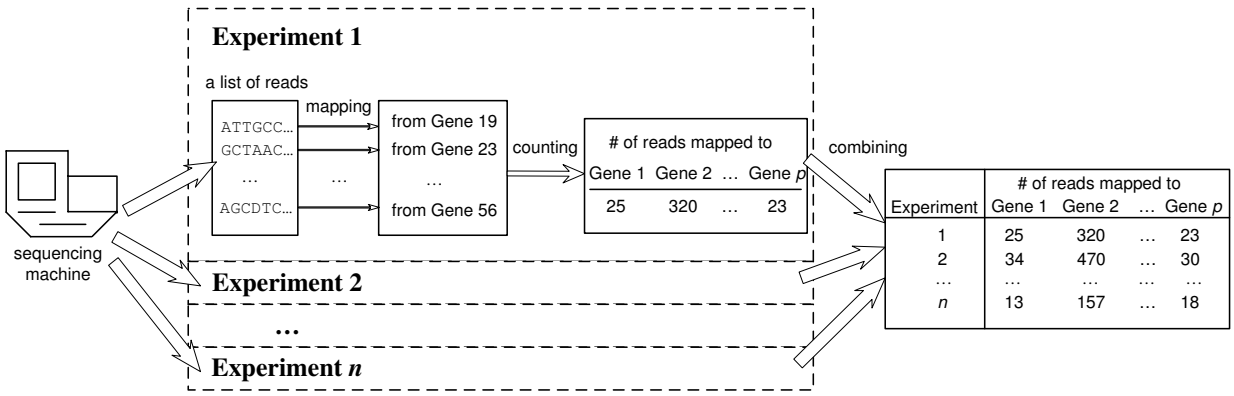

Figure 1: Using RNA-Seq technique for comparative experiments. The sequencing machine generates a list of reads from each sample. Each read in the list is mapped (matched) to a region (here we use a gene) on the genome. Then we sum up the number of reads mapped to each gene, giving a count as a measure of its expression. Each RNA-Seq experiment results in a vector of counts, with length p equal to the number of genes. Combining results from n experiments, the final data can be summarized as an $n \times p$ matrix.

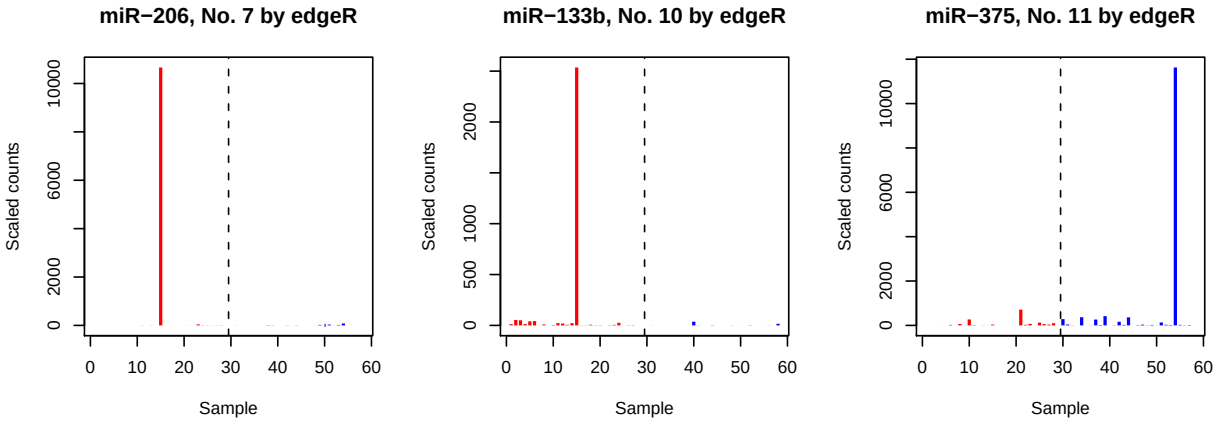

Figure 2: Counts from some miRNAs found to be very significant by *edgeR* do not seem to follow negative binomial distributions. Each panel shows the counts from one miRNA in the Witten data ([35]). These miRNAs are the 7th, 10th and 11th most significant features detected by *edgeR*. The heights of vertical bars show the scaled counts from the samples. The first 29 bars, colored red, are samples from the one class, and the other 29 bars, colored blue, are from the other. The black broken line is also drawn to separate the two classes. In each panel, we see that one count has much larger values than all the other counts.

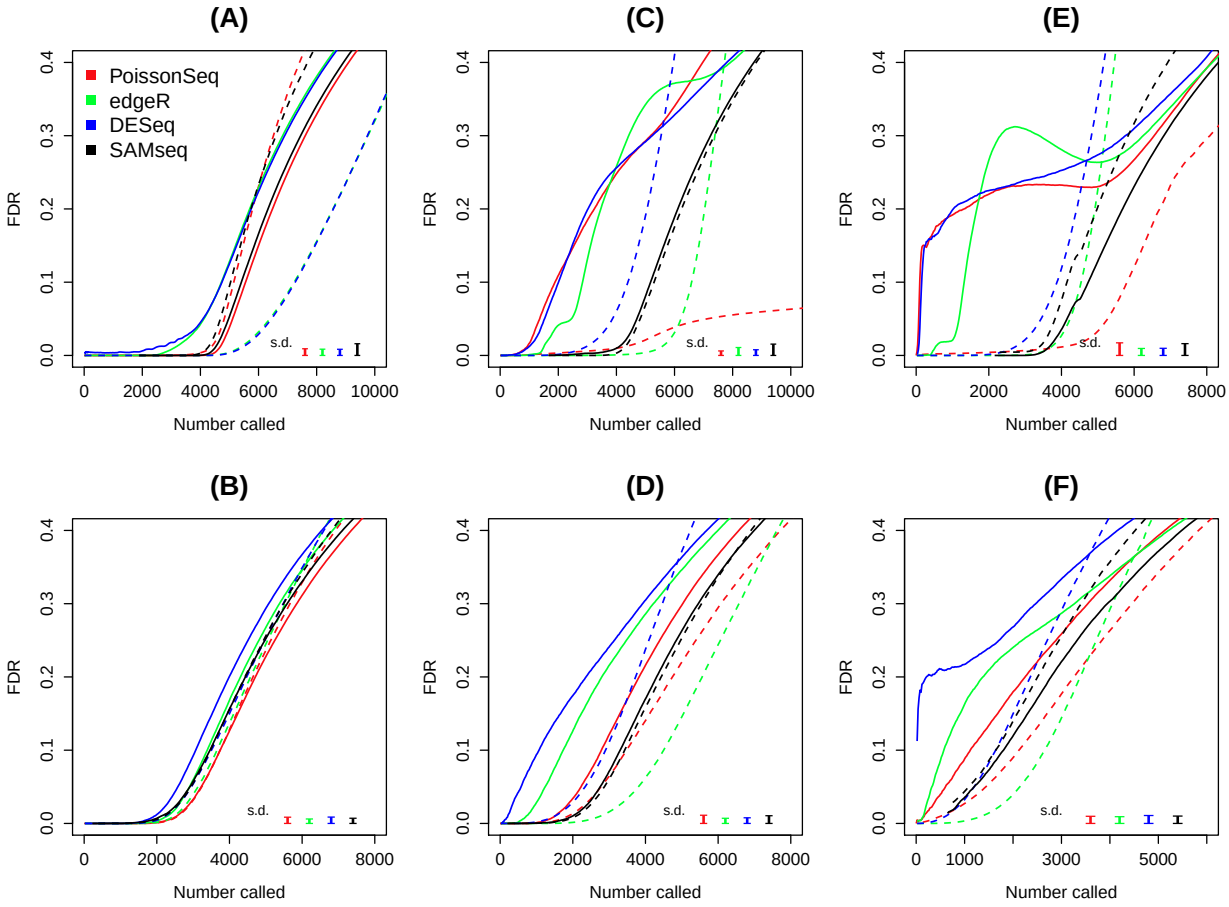

Figure 3: FDR curves for simulated data. (A) Poission distributed data, 12 samples in each class; (B) negative binomial distributed data, 12 samples in each class; (C) Poission distributed data with outliers, 12 samples in each class; (D) negative binomial distributed data with outliers, 12 samples in each class; (E) Poission distributed data with outliers, 5 samples in each class; (F) negative binomial distributed data with outliers, 5 samples in each class. The solid curves show the true FDRs; the broken curves are the estimates. These are results (averaged over 20 simulations) on same simulation data sets using different methods: *PoissonSeq*, *edgeR*, *DESeq*, and *SAMseq*. Some black lines do not go through origin point since a proportion of the most significant features have the same value of the statistic (2.5) and need to be called at the same time. The average standard errors of the estimates are shown as vertical bars on the bottom right.

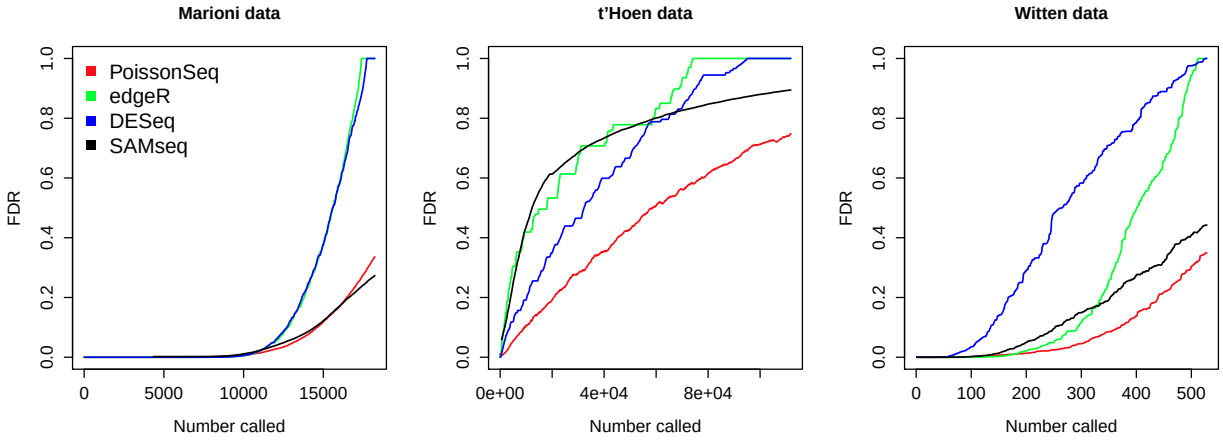

Figure 4: Estimated FDR curves for three real data sets: Marioni data, t'Hoen data, and Witten data (from left to right). In real data sets, the true FDR curves are unknown and so we cannot tell which method is performing better.

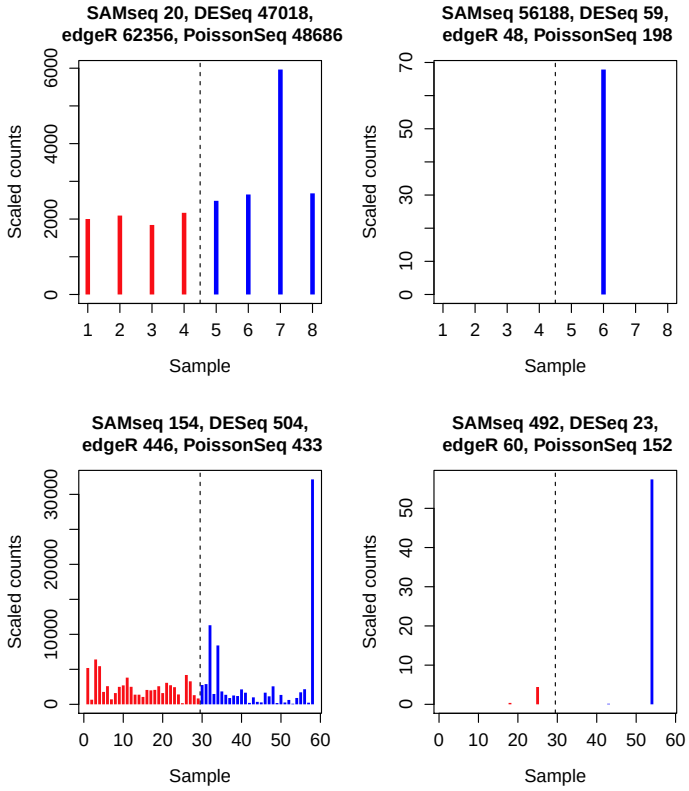

Figure 5: Examples for features deemed to be differentially expressed only by *SAMseq* on t'Hoen data (top left) and on Witten data (bottom left), or only by parametric methods (*PoissonSeq*, *edgeR*, and *DESeq*) on t'Hoen data (up right) and Witten data (bottom right). The title of each subfigure shows the ranks of significance by different methods, like the first one is the 20th most significant feature by *SAMseq*, 47018th by *DESeq*, 62356th by *edgeR*, and 48686th by *PoissonSeq*. In each panel, bars with different colors are samples from different classes. The black broken line is used to separate the two classes. The length of each bar is the scaled count of a sample.

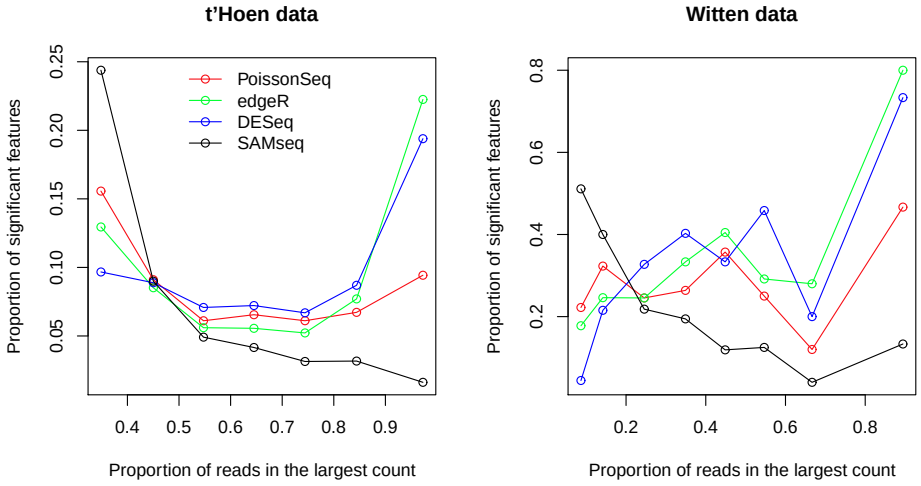

Figure 6: Relations between proportions of reads in the largest count and proportions of significant features in two real data sets: t'Hoen data (left panel), and Witten data (right panel). Features are divided into groups according to what proportion of reads concentrate on the largest count of that class. An proportion near 1 means one count is much larger than all other counts, that is, it has an outlier. We then count in each group what proportion of features are among the list of most significant features (top 10,000 for t'Hoen data, and top 150 for Witten data). We see that features with outliers are more easily been called significant by parametric methods, especially *edgeR* and *DESeq*, while our nonparametric method favors features with similar counts in different samples.

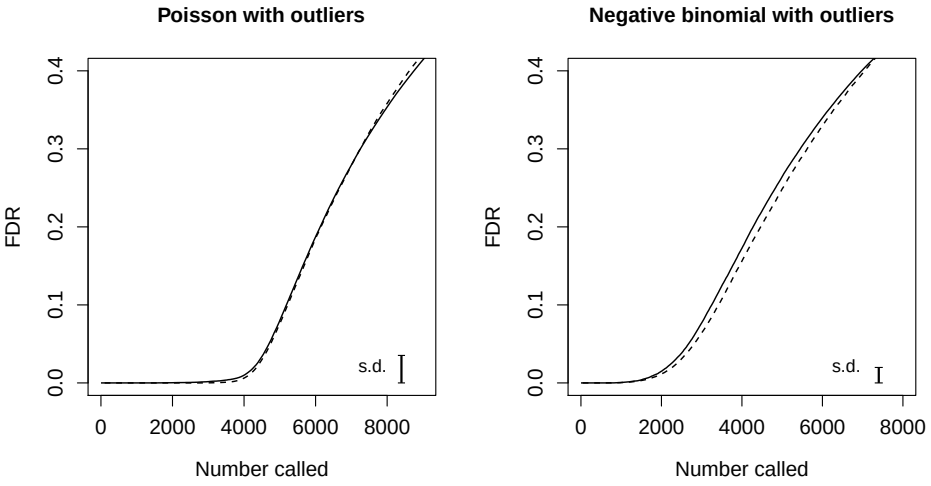

Figure 7: FDR curves for simulated data with survival outcomes (averaged over 20 simulations). The left panel is Poisson distributed data with outliers, and the right panel is negative binomial distributed data with outliers. The solid curves show the true FDRs; the broken curves are the estimates. The average standard errors of the estimates are shown as vertical bars on the bottom right.