REVIEWS

Statistical modeling for differential transcriptome analysis using RNA-Seq technology

Jeffrey Charles Miecznikowski[1, 2], Song Liu[2], Xing Ren[1]

1. Department of Biostatistics, SUNY University at Buffalo, Buffalo, NY, USA. 2. Department of Biostatistics, Roswell Park Cancer Institute, Buffalo, NY, USA

Correspondence: Jeffrey Charles Miecznikowski. Address: Department of Biostatistics, Kimball Tower 723, 3435 Main St, Buffalo, NY, 14214 USA. Telephone: 716-881-8953. Fax: 716-829-2200. Email: jcm38@buffalo.edu

Received: June 14, 2012 **Accepted:** August 6, 2012 **Published:** October 1, 2012
DOI: 10.5430/jst.v2n5p33 **URL:** http://dx.doi.org/10.5430/jst.v2n5p33

Abstract

RNA-Seq is a recently developed technology for transcriptome profiling. Numerous advantages of RNA-Seq suggest that it will be the platform of choice for genome-wide expression studies. RNA-Seq generates large volumes of data which require statistical methods for data processing and accurate inference. This article reviews the RNA-Seq technologies followed by a detailed discussion of current statistical methods for normalization and differential expression analysis.

Key words

RNA-Seq, Normalization, Biological counts, Differential expression

1 Transcriptome profiling

Over 99.9% of genome sequences are the same in all humans [1, 2], yet individuals show great distinction from each other. In a multicellular organism nearly all cells contain the same genome, but they develop into different tissues. A major source for many of these variations is the different gene expression patterns [3]. In control of gene expression, the transcriptome is the complete set of ribonucleic acid (RNA) transcripts, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and other non-coding RNA in a given cell type. It is the connection between genes and phenotype. The constitution of the transcriptome for an organism varies at different developmental stages or under different physiological conditions. As a quantitatively cataloged transcriptome provides us information on the underlying genetic mechanisms, the transcriptome is studied in relation to many diseases, *e.g.*, cancers. One of the main goals in a cancer transcriptome study is quantifying the changes in expression levels of all the transcripts in tumor cells. In this manuscript, we discuss the cutting edge methods for quantifying the transcriptome and how the resulting data is used to determine significantly differentially expressed transcripts.

2 Microarrays

Microarrays have been the primary technology for quantitative transcriptome analysis since the mid-1990s [4], and they have discovered many results in cancer research [5-8]. Although expression studies by microarrays have been very successful in the last decade, there are at least three intrinsic limitations to this hybridization-based technology. First, the background noise in microarrays is large due to cross-hybridization of closely related genes. Second, the microarray signal often reaches a limit of detection or saturation, therefore microarrays have a limited dynamic range (a few hundredfold) [9]. Third, microarray analysis requires prior knowledge on the genome sequence and thus is not suitable for non-model organisms. In part these limitations lead researchers to develop sequence based technologies.

3 Sequence based approaches

Sequence-based approaches were developed initially from Sanger sequencing [10, 11]. The original protocol was low throughput and not quantitative. Many technical improvements since the initial days have led to numerous accomplishments, including the Human Genome Project [12]. Quantitative methods were developed based on the use of tagged sequences, including SAGE (serial analysis of gene expression) [13], CAGE (cap analysis of gene expression) [14] and MPSS (massively parallel signature sequencing) [15]. These methods do not require prior sequence annotation and can directly determine cDNA sequences. However, only a portion of the transcriptome can be analyzed. The low coverage of these technologies showed a need for sequencing with higher throughput.

4 RNA-Seq

With the development of high throughput deoxyribonucleic acid (DNA) sequencing technologies, the next generation sequencing technologies (NGS) allow researchers to read huge volumes of sequences quickly. RNA sequencing (RNA-Seq) is a revolutionary tool for transcriptome analysis based on NGS. From an application’s standpoint, RNA-Seq has been applied to a number of transcriptome expression analyses in cancer research, see Table 1. This recent popularity in cancer research has resulted in several high-throughput sequencing platforms commercially available for RNA-Seq, including Illumina/Solexa, Roche/454, Applied Biosystems/SOLiD, Pacific Biosciences’ RS, Helicos Biosciences’ Heliscope and Ion Torrent’s Proton [16]. These platforms support massively parallel sequencing and therefore have improved efficiency. For example, the latest Illumina HiSeq 2500 system can output 120 gigabases (Gb) in a rapid run of 27 hours and up to 600Gb in a high output run of 11 days. The following subsections provide details on sample preparation and a summary of the advantages and challenges for RNA-Seq data.

Table 1. A list of recent applications of RNA-Seq in cancer research

Cancer type	Platform	Reference
granulosa-cell tumor	Illumina GAII	[32]
chronic myelogenous leukemia and prostate cancer	Illumina GA and 454 FLX	[33]
brain cancer	Illumina GA	[34]
prostate cancer	Illumina GAII	[29]
oral squamous cell carcinomas	SOLiD	[35]
prostate cancer	Illumina GA	[36]
devil facial tumor	454	[28]

4.1 Preparation, sequencing and alignment

In general, to prepare a complementary DNA (cDNA) library for RNA sequencing, mRNAs are extracted from tissues and randomly sheared into short strands. These fragmented mRNAs are reverse transcribed to cDNAs using random primers. Then adapters are added and ligated to cDNAs on one or both ends for sequencing purposes. The ligated cDNAs usually undergo electrophoresis so that cDNAs with certain length are selected and then followed by polymerase chain reaction (PCR) amplification to obtain a cDNA library.

The sequences of the ligated ends on cDNAs are read by a high-throughput sequencing instrument. Sequencing is done in a massively parallel fashion and a huge amount (many millions to billions) of short reads is obtained. The raw data file contains the sequence of all the reads and their corresponding quality scores indicating the confidence of the read at every base. Using alignment software, these reads are mapped to a reference genome or assembled de novo without the reference. Alignment takes into account factors likes reads with exon junctions or polyA ends, reads that can be matched to multiple locations, and single nucleotide polymorphisms (SNPs). There are many alignment programs available, such as MAQ [17], Bowtie [18], BWA [19], ELAND (by Illumina), SOAP2 [20], SHRiMP [21] and many more. In these alignment programs, certain mismatches are allowed for polymorphisms and sequencing errors. During alignment, raw measurements of transcripts are summarized from the mapped reads. Although it is beyond the scope of this review, summarization is an important step as there are many different ways to map the reads and summarize the counts. After summarization, the output data contains the transcript IDs and the corresponding number of reads.

4.2 Advantages and challenges

RNA-Seq has three major advantages over hybridization-based technologies such as microarrays. First, the assembly of short reads in RNA-Seq does not rely on known genome sequences which make it applicable to organisms with unknown sequences. Second, RNA-Seq has a very low background noise, a very large dynamic range (over 10000-fold), and is highly precise and reproducible [9, 22, 23]. For these reasons, expression levels determined by RNA-Seq are seen as more accurate than microarray expression levels [9]. Third, since RNA- Seq provides detailed sequence information of the transcriptome, it can be used to detect allele specific expressions [24, 25], alternate splicing [23, 26, 27], gene fusions [28-30], and novel promoters [31].

Cost is currently the major disadvantage of RNA-Seq. The instrument and labeling kits are very expensive compared to microarray chips and images. However, RNA-Seq is expected to replace microarrays in various applications as the cost decreases. RNA-Seq also brings new challenges in expression analysis. Statistical methods developed for microarrays cannot be directly applied to RNA-Seq data due to the intrinsic differences between the technologies such as sequencing depth. In RNA-Seq, sequencing depth measures how many times a sample is sequenced on average [37]. The sequencing depth reflects the total number of reads from a sample. A RNA-Seq sample is often sequenced in several parallel lanes and lanes within the same run using the same RNA sample often have different depths [22]. Since the observed number of reads is proportional to depth, summarized counts of transcripts should to be normalized with regard to depth before statistical analysis of differential expression (DE). Another issue is the different type of measurements between gene expression microarrays and RNA-Seq technology. In microarrays, fluorescence intensity is a surrogate of transcript level [4]. Thus most methods for microarrays use continuous distributions (*e.g.* log-normal) to model microarray data [37]. In RNA-Seq, however, raw measurements of expression are given by the numbers of reads, which are non-negative integers. The methods based on continuous distribution assumptions are not appropriate for RNA-Seq, especially for low expression genes. Several R packages have been developed for statistical testing for DE using RNA-Seq data [38]. Those include edgeR [39], DESeq [40], DEGSeq [41], baySeq [42], BBSeq [43], TSMP [44], NBPSeq [45] and PoissonSeq [46]. Additionally, databases like SEQC (SEquencing Quality Control) have been established to assess the performance of the NGS technologies. SEQC, also known as MAQC-III (the third phase of the MAQC project), is a follow up from the MAQC and MAQC-II projects [47, 48]. It aims at assessing the technical reproducibility of NGS technologies such as RNA-Seq by generating benchmark datasets with known reference samples. The following sections detail the methods used to model, normalize and determine differential expression in RNA-Seq technologies.

5 Statistical models for RNA-Seq data

In general for RNA-Seq technology, the input data for statistical analysis is a matrix $\mathrm{Y} = [y_{ij}]$ where y_{ij} denotes the number of reads of transcript i in sample j. Let n_j be the total number of reads in sample j, *i.e.* the sum of column j, $n_j = \sum_{i=1}^{n} y_{ij}$. For a robust statistical test for DE genes, a distribution must be specified for the number of reads y_{ij}.

5.1 Binomial and poisson models

Most statistical models for DE using RNA-Seq data are derived from the assumption that y_{ij} follows a binomial distribution; $y_{ij} \sim \mathrm{Bin}(n_j, p_i)$, where n_j is the total number of reads in sample j and $p_i$is the true proportion (unknown) of transcript i in the sample relative to all other transcripts. Because of the large number of genes in the human genome, p_i is very small for any given transcript i. In this situation, the binomial model can be approximated by the Poisson model; $y_{ij} \sim \mathrm{Poisson}(\mu_{ij})$, where the mean parameter $\mu_{ij} \cong n_j p_i$ is the expected value of the number of reads of transcript i in sample j.

The Poisson model may look appropriate for modeling y_{ij}, but it does not take into account the technical sources that might cause the number of reads to vary. A certain degree of overdispersion is observed in technical replicates in Marioni *et al.* [22] (see Figure 1). Furthermore, the Poisson model cannot explain the overdispersion observed in biological replicates [49]. By the Poisson model, the mean and variance of read counts are equal. If the sample variance is plotted against sample mean for all the transcripts, a linear relationship is expected under the Poisson model. From Figure 1 and the mean-variance figure in [43], one can see that transcripts with high expression levels tend to have variance greater than the mean. The degree of dispersion also increases with mean expression level. To account for overdispersion more flexible models have been introduced assuming μ_{ij} or p_i to be a random variable instead of a constant [39, 40, 42, 43, 45].

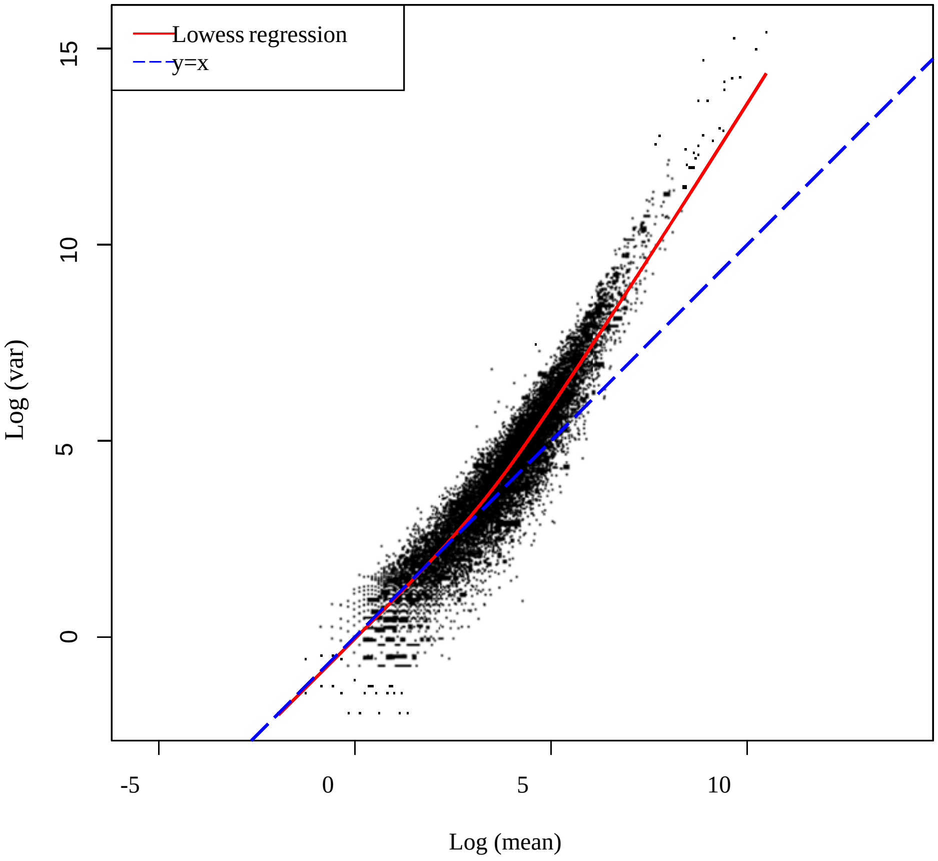

Figure 1. A kidney sample from Marioni *et al.* was sequenced in seven lanes on Illumina Genome Analyzer [22]. The scatterplot shows the sample variance versus sample mean of the seven lanes on the log scale for the read count of each transcript. Overdispersion is evidenced by the deviation from the scatterplot to the line y = x which corresponds to the Poisson model where the mean is equal to the variance.

5.2 Negative binomial

The negative binomial model; $y_{ij} \sim \mathrm{NB}(\mu_{ij}, \varphi_i)$, with the mean parameter μ_{ij} and the dispersion parameter φ_i is an extension of the Poisson model where the variance is larger than the mean. It is more appropriate to model count data with overdispersion. The negative binomial model is also known as the gamma-Poisson model, since it can be derived from the Poisson model by assuming the Poisson mean λ_{ij} follows a Gamma distribution, $\lambda_{ij} \sim \Gamma(\mu_{ij}, \mu_{ij}\varphi_i)$, where μ_{ij} is the Gamma mean and $\mu_{ij}\varphi_i$ is the variance. The ratio of variance to mean in the negative binomial model is $\varphi_i + 1$, thus the Poisson model is a special case of negative binomial when $\varphi_i = 0$. The negative binomial model is the most common model in DE testing for overdispersed RNA-Seq data and is implemented in the R packages edgeR [39], DESeq [40], NBPSeq [45] and baySeq [42].

5.3 Beta binomial

The beta binomial model is an extension from the binomial model. Similar to the negative binomial, it can also explain the overdispersion in biological replicates by allowing the relative abundance of the *i*th transcript p_i to vary according to a beta binomial distribution. It can be derived from the binomial model by assuming p_i follows a beta distribution, $p_i \sim Beta(\alpha_{1i}, \alpha_{2i})$. Thus in this framework, the conditional probability of read count y_{ij} given total count n_j is beta binomial, $y_{ij}|n_j \sim Beta(\alpha_{1i}, \alpha_{2i})$. As the total count of reads is large, the beta binomial behaves similarly to the negative binomial model. The beta binomial model is implemented in the R package BBSeq [43].

5.4 Power transformation

Although the negative binomial model is most commonly used, it does not have maximum likelihood solutions with closed forms and the estimation of the dispersion parameter φ_i requires biological replicates which may not be available for experiments with quantitative outcomes, *e.g.*, survival. Li *et al.* take a different approach by seeking a power transformation of the over-dispersed data [46]. They first select a conservative set *S* of transcripts, where all the transcripts in *S* are believed to be not differentially expressed. Define $O = \sum_{i \in S}\{GOF_i - (\mathrm{m} - 1)\}$, where GOF_i is the goodness of fit test statistic for transcript *i* based on the Poisson model, and $m - 1$ is the expected value for $\mathrm{GOF_i}$ based on m samples. (The selection of *S* and goodness-of-fit statistic will be discussed in the next section.) Thus *O* can be interpreted as the overall dispersion over the Poisson model; $\mathrm{O} \approx 0$ if the transformed data follows the Poisson distribution. The goal is to obtain θ so that $\mathrm{O} \approx 0$ for the power transformed data $y_{ij} \longrightarrow y_{ij}^{\theta}$. In fact, the requirement that θ is the same for all transcripts is too restrictive, so the data is divided into multiple groups with θ estimated for each group. Their power transformation approach is implemented in the R package PoissonSeq [46]. In addition to modeling the read count data, researchers must normalize the RNA-Seq data to account for systematic variation.

6 Normalization of RNA-Seq data: a matter of depth

In most statistical packages for DE analysis, the parameter of interest is the true expression level q_i for transcript *i* in the sample. For each transcript *i*, the goal is often to test whether q_i is different in treatment group 1 versus 2, $H_0: q_{i1} = q_{i2}$ vs. $H_1: q_{i1} \neq q_{i2}$. We can specify a relationship *f* between q_i and the mean parameter μ_{ij} in the model such that $\mu_{ij} = f(d_j, q_i)$, where d_j is the sequencing depth for sample *j* and μ_{ij} is modeled as in previous sections. The sequencing depth d_j measures the number of sequenced transcripts in sample *j*. A simple relationship is $\mu_{ij} = d_j q_i$ as in [40], or $\mu_{ij} = d_j e^{q_i}$ as in [50].

The raw measure of expression for transcript *i* in sample *j* is the number of reads y_{ij}. However, y_{ij} is biased towards longer transcripts because they generate more short reads given the same expression level [51]. However, length is not as critical since the effect of length is negated when testing the same transcript. Previous studies have shown that normalization with regard to sequencing depth is an essential step for accurate comparison of mean transcript expression level across samples [39, 40, 50, 52].

It should be pointed out that the choice of normalization is not independent of the model specified for DE testing. For example, quantile normalization produces non-integer counts, making count-based models such as Poisson or negative binomial distributions inappropriate. Instead, using estimated depth $\hat{d}_j$ as a scaling factor for each sample is preferred such that the integer counts are preserved.

An intuitive method for normalization uses the total count of reads in each sample as scaling factor. Another common method uses RPKM (reads per kilobase per million mapped) as the scaling factor to adjust for both total count and transcript length [23]. These methods, however, can lead to biased results because they are overly simplistic in certain situations. As an example, the total count can be heavily affected by a small proportion of highly expressed genes [46, 50]. For example, suppose there are two samples with 100 transcripts. In the control sample, all 100 transcripts have 10 reads. In the treatment sample, 90 transcripts remain the same with 10 reads each, while the other 10 are up-regulated and have 50 reads. The treatment sample has a total count 1.4 times as many as the control if both are sequenced at the same depth. If scaled by the total count, 90% of the transcripts will appear to be down-regulated in treatment group although they are not differentially expressed. For this reason, the following more robust methods have been proposed to estimate sequencing depth.

6.1 Trimmed mean of *M* values

Robinson and Oshlack propose a trimmed mean of *M* values (TMM) as a scaling normalization approach [49]. Under the assumption that the majority of genes are not DE, the TMM method equates the overall expression levels of genes between samples. The relationship function f they specify is $\mu_{ij} = \frac{q_{ij}}{s_j} n_j$, where the mean parameter μ_{ij} is the expected value of y_{ij}, q_{ij} is the true expression level of transcript i in sample, n_j is the total number of reads in sample *j*, and size factor $s_j = \sum_i q_{ij}$ is the expected number of total transcripts in sample *j*. Since s_j is unknown, they proposed to estimate the relative ratio of two samples *j* and *j'* by $f_{jj'} = s_j / s_{j'}$ via a trimmed mean of log ratios. Define the log-fold change as $M_i = log_2 \frac{y_{ij}/n_j}{y_{ij'}/n_{j'}}$ and the absolute expression level for absolute expression level $A_i = \frac{1}{2} log_2 (\frac{y_{ij}}{n_j} \frac{y_{ij'}}{n_{j'}})$ for $n_j \neq 0$. The *M* value can be explained as the difference of proportion for observed counts on $\log_2$ scale, while the *A* value is the average. Assuming most of the genes are not differentially expressed, $f_{jj'}$ can be estimated from the weighted trimmed mean of *M* values. The value of *f* is a scaling factor for normalization.

6.2 Median ratio of counts

Anders and Huber propose a similar approach to TMM in [40]. If a transcript is not differentially expressed in sample *j* and *j'*, the ratio of expected counts $EY_{ij}/EY_{ij'}$ is equal to the ratio of depths $d/d_{j'}$. They propose using the median ratio of observed counts to estimate the relative depth [40]. They use the estimator,

$$\hat{d}_j = median_i \frac{y_{ij}}{(\prod_{j=1}^{m} y_{ij})^{1/m}} \tag{1}$$

where the denominator is the geometric mean for transcript i in all samples. Thus the scaling factor is computed as the median of the ratio between sample j and the overall mean.

6.3 Upper-quartile normalization

Inspired by the normalization procedure of microarray data, Bullard *et al*. propose to match all the samples to a reference by using the quantiles of the distribution [50]. A simple way to match the quantiles is to scale the counts by the median. However, due to the frequent existence of zero and low count transcripts, Bullard *et al*. propose to only use the upper-quartile of counts for non-zero transcripts. The normalized counts are then rounded to integers to preserve the count nature of the data. They compare the upper- quartile approach to two other normalization approaches, (1) “RPKM” and (2) counts of “housekeeping” transcripts using qRT-PCR as the golden standard. They found their approach yields better concordance with qRT-PCR and that it significantly reduces the bias, thus improving sensitivity.

6.4 Goodness of fit approach

Li *et al*. propose a goodness of fit approach adapted from the total count estimation [46]. Without losing generality, let $\sum_{j=1}^{m} d_j = 1$, where d_j is the relative depth for sample j and m is the number of samples. Total count normalization gives an estimator $\hat{d}_j = \frac{\sum_{i\in S} y_{ij}}{\sum_{i\in S} y_{i.}}$, where $y_{i.} = \sum_{i\in S} y_{ij}$ and S is the full set of genes. A better approach is to use a set S containing only non-differentially expressed genes. Thus they employ a Poisson goodness-of-fit test statistic $GOF_j = \sum_{j=1}^{m} \frac{\left(y_{ij}-\hat{d}_j y_{i.}\right)^2}{\hat{d}_j y_{i.}}$ to obtain an optimized set S. Genes with GOF values in the $(\epsilon, 1-\epsilon)$ quantile are chosen into set S. A conservative value for ϵ is 0.25. Updated S gives an updated $\hat{d}_j$, which in turn is used to update S. This recursive algorithm converges quickly to give the final estimates for d_j and S. After modeling the read counts and normalizing the data, it is common to test for differential expression of transcripts between two conditions.

7 Testing for differential expression

Commonly we are interested in testing whether the expression level of a transcript is the same between treatments. For a comparison of two treatments A and B, the hypothesis is $H_0: q_{iA} = q_{iB}$ vs. $H_1: q_{iA} \neq q_{iB}$ where q_{iA} and q_{iB}denote the amount of transcript i in treatments A and B, respectively. Given the specified model and scaling factor estimated, the standard procedures may employ a Wald test, score test [46], or the likelihood ratio test [22]. Due to the cost of sequencing, however, usually only a small number of samples are available. This raises the question about the appropriateness of procedures based on large sample approximations [53].

Robinson and Smyth developed an exact test for small sample estimation in the negative binomial model [53]. It was originally applied to serial analysis of gene expression (SAGE) data [13]. Similar to Fisher’s exact test, they replace the hypergeometric probabilities with negative binomial probabilities. Anders and Huber follow the same strategy [40]. Specifically, for a comparison between samples in treatment A and treatment B, they define $k_{iT} = \sum_{j\in T} y_{ij}$, $(T = A, B)$, where T denotes the set of sample indices in Treatment T. Thus k_{iT} is the sum of transcript i in treatment T and $k_{iS} = k_{iA} + k_{iB}$ is the overall sum. Given the negative binomial model, the probability of the $k_{iA} = a$ and $k_{iB} = b$, denoted $p(a, b)$ can be calculated for any values a and b. Then the two sided p-value for the exact test is the probability of observing treatment sums more extreme than the observed combination of k_{iA} and k_{iB}, conditional on the overall sum k_{iS}. In other words, the p-value for transcript i is given by the following,

$$p_i = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b)\le p(k_{iA},k_{iB})}} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)} \tag{2}$$

where the denominator is the probability of observing the overall sum k_{iS}, and the numerator is the sum of probabilities less than or equal to $p(k_{iA}, k_{iB})$ given the overall sum k_{iS}.

8 Multiple testing

Commonly, a separate test is performed on the null hypothesis for each transcript, and a p-value is computed for each test. Although a p-value < 0.05 or 0.01 is usually considered significant for a single test, this decision rule presents a problem in genome-wide multiple testing, where tens of thousands of tests are performed simultaneously. Suppose in a comparison between biological replicates, none of 10,000 transcripts are differentially expressed. If the decision rule calls p-values less than 0.01 significant, on average 100 transcripts will be incorrectly identified as DE. In general we want to discover as many differentially expressed transcripts as possible while keeping the false discovery rate (FDR) [54] relatively low. Conventional statistical adjustments for multiple testing, such as the Bonferroni correction, aim to control the family wise error rate over the whole family of transcripts. These procedures lack sufficient power and are too conservative for transcriptome-wide studies. There is a consensus that FDR estimation procedures are a good alternative approach [55]. Benjamini and Hochberg coined the term FDR and provide a procedure for its control [54]. Several other procedures have been developed to estimate FDR [56-59], and many of them have been widely applied to microarray data. Li *et al*. showed that the standard plug-in permutation greatly overestimates the true FDR [46]. Due to the mean-variance dependency in RNA-Seq models, the test statistic has very different permutation distributions for null and non-null transcripts. These approaches need to be modified in order to be used for RNA-Seq data. Li *et al*. propose an adapted approach by excluding non-null genes from the permutation distribution as implemented in PoissonSeq [46].

9 Future work

A precise catalogue of all transcripts across diverse cell types provides us insight about gene functions and pathways. RNA-Seq technology is a powerful tool to quantify the transcriptome in the tissues under different physiological conditions. It has broadened our view of the expression studies in transcriptome analysis. In this review, we have outlined the major steps in RNA-Seq technology and the statistical analysis of the RNA-Seq data. Although this review mainly discussed two group comparisons, some statistical packages reviewed in this manuscript are capable of handling more complex designs including experiments with quantitative outcomes or multiple treatment conditions [60]. On the other hand, duplicate samples are required to estimate the overdispersion in some models, but they are not generally available for quantitative outcomes [46]. This could restrict the application of some of the methods.

As previously discussed, RNA-Seq data processing starts with alignment of short reads. The alignment method affects the summarized counts of reads and there are many algorithms for sequence alignment and reads summarization. These methods vary in short reads mapping and transcriptome reconstruction, thus they have different impact on the obtained matrix of summarized reads. Other studies indicate that the RNA-Seq platforms also produce bias in generating reads [61, 62]. So far, there has been little research on the choice of these methods and its impact on the DE analysis [63].

To analyze RNA-Seq data more accurately, other experimental aspects should be taken into account. Unlike the microarray studies, a technical feature of RNA-Seq is the bias caused by the sequence of transcript. A major aspect is the length bias, in which longer transcripts have more reads than short ones at the same expression level. Meanwhile, the so-called “GC-content” (percentage of G and C bases) bias is observed in several studies [64], in which transcript fragments of high GC-content are preferentially detected in the sequencing process. Although the sequence effects such as length and GC-content are negated when testing for the same transcript, these biases result in greater statistical power in DE analysis for transcripts with longer sequences and higher GC content. This can significantly affect the results of multiple testing and the downstream analyses, such as Gene Ontology (GO) for enrichment among a set of DE genes [65]. As the understanding of the bias sources grows, more robust statistical models will be needed to account for these sources of technical variation.

Both RNA-Seq and microarrays are effective tools for transcriptome profiling and they have shown similar performance in some studies [37, 66-68]. Although RNA-Seq results are believed to be highly reproducible, there are discrepancies between the expression levels measured by RNA-Seq and microarrays [22]. For example, a large range of low expression genes in microarrays are not detectable by RNA-Seq [37]. This demonstrates a need for validation datasets and appropriate validation criteria. Real-time polymerase chain reaction (RT-PCR) has been the choice to assess the accuracy of microarray and RNA-Seq technologies, but the small scale of RT-PCR may restrict its application to validation in genome-wide studies.

In microarray expression studies involving multiple testing, false discovery rate (FDR) has been the choice for genome-wide error control, yet there is limited work on how to control FDR in RNA-Seq data analysis. The methods previously developed for microarrays are not appropriate for RNA-Seq. It is been shown that FDR may not be controlled well by the traditional Benjamini-Hochberg procedure and the rate of errors is underestimated [69]. Current procedures for FDR estimation need to be re-evaluated for RNA-Seq data with regard to the difference in DE and non-DE genes, sample exchangeability, and gene independency [46]. Ultimately, as the cost of RNA-Seq continues to decrease, more flexible statistical frameworks are needed to handle complex RNA-Seq experiments.

Conflict of interests

The author(s) declare that they have no competing interests.

References

[1] Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, et al.. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998; 280(5366): 1077–1082. PMid:9582121 http://dx.doi.org/10.1126/science.280.5366.1077

[2] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al.. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409(6822): 928–933. PMid:11237013 http://dx.doi.org/10.1038/35057149

[3] Harrison PW, Wright AE, Mank JE. The evolution of gene expression and the transcriptome-phenotype relationship. Semin Cell Dev Biol. 2012; 23(2): 222–9. http://dx.doi.org/10.1016/j.semcdb.2011.12.004

[4] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270(5235): 467–470. PMid:7569999 http://dx.doi.org/10.1126/science.270.5235.467

[5] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al.. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769): 503–511. PMid:10676951 http://dx.doi.org/10.1038/35000501

[6] Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP. Gene expression during the life cycle of drosophila melanogaster. Science. 2002; 297(5590): 2270–2275. PMid:12351791 http://dx.doi.org/10.1126/science.1072152

[7] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al.. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439): 531–537. PMid:10521349 http://dx.doi.org/10.1126/science.286.5439.531

[8] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998; 9(12): 3273–3297. PMid:9843569

[9] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1): 57–63. PMid:19015660 http://dx.doi.org/10.1038/nrg2484

[10] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975; 94(3): 441–448. http://dx.doi.org/10.1016/0022-2836(75)90213-2

[11] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977; 74(12): 5463–5467. http://dx.doi.org/10.1073/pnas.74.12.5463

[12] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al.. The sequence of the human genome. Science. 2001; 291(5507): 1304–1351. PMid:11181995 http://dx.doi.org/10.1126/science.1058040

[13] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995; 270(5235): 484–487. PMid:7570003 http://dx.doi.org/10.1126/science.270.5235.484

[14] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al.. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA. 2003; 100(26): 15 776–15 781.

[15] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al.. Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays. Nat Biotechnol. 2000; 18(6): 630–634. PMid:10835600 http://dx.doi.org/10.1038/76469

[16] Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011; 11(5): 759–769. PMid:21592312 http://dx.doi.org/10.1111/j.1755-0998.2011.03024.x

[17] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18(11): 1851–8. PMid:18714091 http://dx.doi.org/10.1101/gr.078212.108

[18] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3): R25. PMid:19261174 http://dx.doi.org/10.1186/gb-2009-10-3-r25

[19] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009; 25(14): 1754–1760. PMid:19451168 http://dx.doi.org/10.1093/bioinformatics/btp324

[20] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. Soap2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25(15): 1966–1967. PMid:19497933 http://dx.doi.org/10.1093/bioinformatics/btp336

[21] Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. Shrimp: accurate mapping of short color-space reads. PLoS Comput Biol. 2009; 5(5): e1000 386.

[22] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-Seq: an assessment of technical repro- ducibility and comparison with gene expression arrays. Genome Res. 2008; 18(9): 1509–1517. PMid:18550803 http://dx.doi.org/10.1101/gr.079558.108

[23] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5(7): 621–628. PMid:18516045 http://dx.doi.org/10.1038/nmeth.1226

[24] Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M. Computational analysis of whole-genome differential allelic expression data in human. PLoS Comput Biol. 2010; 6(7): e1000 849.

[25] Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science. 2010; 329(5992): 643–648. PMid:20616232 http://dx.doi.org/10.1126/science.1190830

[26] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008; 40(12): 1413–1415. PMid:18978789 http://dx.doi.org/10.1038/ng.259

[27] Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al.. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321(5891): 956–960. PMid:18599741 http://dx.doi.org/10.1126/science.1160342

[28] Murchison EP, Tovar C, Hsu A, Bender HS, Kheradpour P, Rebbeck CA, Obendorf D, Conlan C, Bahlo M, Blizzard CA, et al.. The tasmanian devil transcriptome reveals schwann cell origins of a clonally transmissible cancer. Science. 2010; 327(5961): 84–87. PMid:20044575 http://dx.doi.org/10.1126/science.1180616

[29] Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, et al.. Discovery of non-ets gene fusions in human prostate cancer using next-generation rna sequencing. Genome Res. 2011; 21(1): 56–67. PMid:21036922 http://dx.doi.org/10.1101/gr.110684.110

[30] Li Y, Chien J, Smith DI, Ma J. Fusionhunter: identifying fusion transcripts in cancer using paired-end RNA-Seq. Bioinformatics. 2011; 27(12): 1708–1710. PMid:21546395 http://dx.doi.org/10.1093/bioinformatics/btr265

[31] Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PLoS One. 2008; 3(12): e3839. PMid:19052635 http://dx.doi.org/10.1371/journal.pone.0003839

[32] Shah SP, Kobel M, Senz J, Morin RD, Clarke BA, Wiegand KC, Leung G, Zayed A, Mehl E, Kalloger SE, et al.. Mutation of fox l2 in granulosa-cell tumors of the ovary. N Engl J Med. 2009; 360(26): 2719–2729. PMid:19516027 http://dx.doi.org/10.1056/NEJMoa0902542

[33] Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009; 458(7234): 97–101. PMid:19136943 http://dx.doi.org/10.1038/nature07638

[34] Markovets AA, Herman D. Analysis of cancer metabolism with high-throughput technologies. BMC Bioinformatics. 2011; 12 Suppl 10: S8. PMid:22166000 http://dx.doi.org/10.1186/1471-2105-12-S10-S8

[35] Tuch BB, Laborde RR, Xu X, Gu J, Chung CB, Monighetti CK, Stanley SJ, Olsen KD, Kasperbauer JL, Moore EJ, et al.. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. PLoS One. 2010; 5(2): e9317. PMid:20174472 http://dx.doi.org/10.1371/journal.pone.0009317

[36] Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci USA. 2008; 105(51): 20 179–20 184.

[37] Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011; 9: 34. PMid:21627854 http://dx.doi.org/10.1186/1741-7007-9-34

[38] R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria 2008. Available from: http://www.R-project.org, ISBN 3-900051-07-0.

[39] Robinson MD, McCarthy DJ, Smyth GK. Edger: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1): 139–140. PMid:19910308 http://dx.doi.org/10.1093/bioinformatics/btp616

[40] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10): R106. PMid:20979621 http://dx.doi.org/10.1186/gb-2010-11-10-r106

[41] Wang L, Feng Z, Wang X, and Zhang X. Degseq: an R package for identifying differentially expressed genes from RNA-Seq data. Bioinformatics. 2010; 26(1): 136–138. PMid:19855105 http://dx.doi.org/10.1093/bioinformatics/btp612

[42] Hardcastle TJ, Kelly KA. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11: 422. PMid:20698981 http://dx.doi.org/10.1186/1471-2105-11-422

[43] Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics. 2011; 27(19): 2672–2678. PMid:21810900 http://dx.doi.org/10.1093/bioinformatics/btr449

[44] Auer PL, Doerge RW. Statistical design and analysis of rna sequencing data. Genetics. 2010; 185(2): 405–416. PMid:20439781 http://dx.doi.org/10.1534/genetics.110.114983

[45] Di Y, Schafer DW, Cumbie JS, Chang JH. The nbp negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol. 2011; 10(1). http://dx.doi.org/10.2202/1544-6115.1637

[46] Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estima- tion for RNA-Sequencing data. Biostatistics. 2011; 13(3): 523-38.

[47] Shi L, Reid L, Jones W, Shippy R, Warrington J, Baker S, Collins P, De Longueville F, Kawasaki E, Lee K, et al.. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. Nature biotechnology. 2006; 24(9): 1151–1161. PMid:16964229 http://dx.doi.org/10.1038/nbt1239

[48] Shi L, Campbell G, Jones W, Campagne F, Wen Z, Walker S, Su Z, Chu T, Goodsaid F, Pusztai L, et al.. The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models. Nature biotechnology. 2010; 28(8): 827. PMid:20676074 http://dx.doi.org/10.1038/nbt.1665

[49] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioin- formatics. 2007; 23(21): 2881–2887. PMid:17881408 http://dx.doi.org/10.1093/bioinformatics/btm453

[50] Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11: 94. PMid:20167110 http://dx.doi.org/10.1186/1471-2105-11-94

[51] Oshlack A, Wakefield MJ. Transcript length bias in RNA-Seq data confounds systems biology. Biol Direct. 2009; 4: 14. PMid:19371405 http://dx.doi.org/10.1186/1745-6150-4-14

[52] Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-Sequencing differential expression analysis with myrna. Genome Biol. 2010; 11(8): R83. PMid:20701754 http://dx.doi.org/10.1186/gb-2010-11-8-r83

[53] Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to sage data. Biostatistics. 2008; 9(2): 321–332. PMid:17728317 http://dx.doi.org/10.1093/biostatistics/kxm030

[54] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995; 85: 289–300.

[55] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 2006; 7(1): 55–65. PMid:16369572 http://dx.doi.org/10.1038/nrg1749

[56] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA. 2001; 98(9): 5116–5121. PMid:11309499 http://dx.doi.org/10.1073/pnas.091062498

[57] Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(3): 479–498. http://dx.doi.org/10.1111/1467-9868.00346

[58] Storey JD. The positive false discovery rate: A bayesian interpretation and the q-value. Annals of Statistics. 2003; 31: 2013–2035. http://dx.doi.org/10.1214/aos/1074290335

[59] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003; 100(16): 9440–9445. PMid:12883005 http://dx.doi.org/10.1073/pnas.1530509100

[60] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40(10): 4288-97. PMid:22287627 http://dx.doi.org/10.1093/nar/gks042

[61] Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol. 2010; 11(5): R50. PMid:20459815 http://dx.doi.org/10.1186/gb-2010-11-5-r50

[62] Hansen KD, Brenner SE, Dudoit S. Biases in illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010; 38(12): e131. PMid:20395217 http://dx.doi.org/10.1093/nar/gkq224

[63] Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010; 11: 220. PMid:21176179 http://dx.doi.org/10.1186/gb-2010-11-12-220

[64] Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-Seq data using conditional quantile normalization. Biostatistics. 2012; 13(2): 204-16. PMid:22285995 http://dx.doi.org/10.1093/biostatistics/kxr054

[65] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000; 25(1): 25–29. PMid:10802651 http://dx.doi.org/10.1038/75556

[66] Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics. 2010; 11: 383. PMid:20565764 http://dx.doi.org/10.1186/1471-2164-11-383

[67] Liu S, Lin L, Jiang P, Wang D, Xing Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res. 2011; 39(2): 578–588. PMid:20864445 http://dx.doi.org/10.1093/nar/gkq817

[68] Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics. 2009; 10: 221. PMid:19435513 http://dx.doi.org/10.1186/1471-2164-10-221

[69] Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-Seq data. Am J Bot. 2012; 99(2): 248–256. PMid:22268221 http://dx.doi.org/10.3732/ajb.1100340