博士論文

A comprehensive evaluation of methods for differential expression analysis on multi-group RNA-seq count data

（RNA-seqの多群間比較用カウントデータに基づく発現変動解析手法の評価）

湯 敏

Contents

List of Figures

List of Tables

List of Additional files

Chapter 1 Introduction

1.1 Microarray

Since the discovery, in 1952, that DNA is genetic material, its exploration has never stopped. Nowadays, it is well known that biodiversity derives from genetic diversity. Differences in the information encoded in the genome shapes the thousands and thousands of species, which have completely different phonotypes, behavior characteristics, and metabolism. In the past century, one of the main interests in the genetic community has been the comparative analysis of genetic structure and DNA sequences in closely related species with the hope of revealing the fundamental cause of completely different species with very similar genomes. In the past decades, scientists have elucidated the cellular functions of many genes from experiments on model organisms (i.e., mouse, fruit fly, and thale cress) [1]. The standard biological technologies are gene knockout (KO) and transgenesis. The biggest advantage of these technologies is that the experimental results can be easily observed from the difference between the mutant and normal individuals. However, these technologies usually involve long periods of work, high workload, huge financial budget, and limited data yield. As a result, they have been limitedly used in some big laboratories. Moreover, there is little possibility to simultaneously analyze several interesting gene loci at the tissue or cellular level.

When microarray technology arrived in 1983, it became possible for the first time to analyze the mechanisms of each gene in the actual sense [2]. In 1995, Schena and Shalon developed a high-capacity system using microarrays to monitor the expression of several genes in parallel. In that system, a microarray is prepared by high-speed robotic printing of complementary DNA sequences on glass slides after chemical and heat treatments. Fluorescent probes are prepared from two mRNA sources by a single round of reverse transcription in the presence of fluorescein- or lissamine-labeled nucleotide analogs. Then, the two probes are mixed in equal proportions, hybridized to the microarray, and separately scanned for fluorescein and lissamine emission after the independent excitation of the two fluorophores. The two kinds of gene-specific color scanning

densities can determine the difference in the gene's expression in two individual spots [3].

After steady improvement over the past 30 years, microarrays now allow the monitoring of expression levels of thousands to tens of thousands of targets simultaneously in a single sample [4, 5]. Depending on the experimental needs, the targets can be oligonucleotide, cDNA, protein, SNP, even BAC, and others. With the establishment of companies such as Affymetrix, Agilent, Applied Microarrays, Arrayit, and Illumina, microarrays have become widely used as the technology of choice for high-throughput transcript profiling and have built an excellent public reputation. As a result, numerous microarray platforms have been invented and used, and concerns about their reliability and consistency have been raised. Several comparison studies have been published with contradictory results. Some have reported agreement in conclusions across different platforms while others have not [6-15]. Irizarry *et al*. [16] also demonstrated the existence of relative large differences in data obtained across different labs using the same platforms. To corroborate the reliability of the technology, the US Food and Drug Administration (FDA) and National Institutes of Health (NIH) launched the first MicroArray Quality Control (MAQC) project in 2006. The MAQC Consortium demonstrated intraplatform consistency for various microarray-based and alternative technologies across several test sites as well as a high level of interplatform concordance in terms of genes identified as differentially expressed and confirmed that microarrays can reliably identify differentially expressed genes (DEGs) between sample classes or populations [17]. Moreover, the MAQC microarray data set was positively validated by quantitative gene expression technologies, which support the use of microarray platforms for the quantitative characterization of gene expression [18]. Therefore, it is still mainly used on a large scale around the world and many studies using microarray data have been published.

However, there are several inherent limitations to this good technology. The issues of hybridization, cross-hybridization, dye-based detection, and design constraints that preclude or seriously limit the detection of splice patterns and unknown previously unmapped genes

make microarrays difficult to use in standard array designs to provide full sequence and transcriptome comprehensiveness [19-21]. Its advantages and disadvantages are summarized in Table 1.

Although the expression level of each gene is estimated as the logarithm of fluorescence density in the microarray screenshot, it is also possible, by means of transcriptome profiling, to detect differential expression (DE) of specific genes between different samples using this early high-throughput methodology. There were many challenges facing the analysis of microarray data initially. For instance, because of the high cost, only few replicates could be made. In addition, DE was determined by simplistic statistics, such as fold change. Advanced statistical testing procedures, such as those based on modified *t*-tests, have been used after the variability over replicates became apparent [22].

Table 1 - Advantages and disadvantages of Microarray

Advantages	Disadvantages
Robust, reliable method, proven over decades of use	Need prior sequence knowledge
High throughput method	Can not detect structural variations
Streamlined handing - can be easily automated	Can not detect isoforms
Straightforward data analysis	Hybridization and sample labeling biases
Short turn around time	Not an absolute quantitation method
Low cost	

1.2 NGS and RNA-seq

Nucleic acid sequencing is a methodology for determining the exact order of nucleotides present in a given DNA or RNA molecule and can unfold the complex information in the genome. As we all know, Edward Sanger developed the first-generation sequencing in 1975 [23]. Although this technology had been adopted for almost 20 years, the requirement for the construction of clone libraries has limited its widespread use. It was mainly utilized to complete international or national vast projects (e.g., the Human Genome Project in 2003). The purpose of these projects was to construct full-length gene structures of model organisms. Because the technology is very expensive and time-consuming, these projects had to be conducted by the collaboration of several countries and completed by many workers. However, the demand for cheaper and faster sequencing methods has greatly increased. In 1998, the speed of sequencing made rapid progress as a result of a big revolution in the industry. The new technology has been called high-throughput sequencing (HTS) or next-generation sequencing (NGS). NGS performs massively parallel sequencing, during which millions of fragments of DNA from a single sample are sequenced in a single run. NGS allows an entire genome to be sequenced in a matter of days and at a small fraction of the cost of the above projects.

Based on its strong capability of sequencing large amounts of DNA fragments simultaneously, NGS has been used in a range of quantitative assays. In particular, it became possible to sequence cDNA reversed from RNA of cells or tissues, which is a well-known process termed RNA sequencing (RNA-seq). To be exact, the term RNA-seq refers to a wet-lab experimental procedure that generates short DNA sequence reads derived from the entire set of RNA molecules present at a featured biological stage [24-33]. In recent years, RNA-seq has become one of the important and classic applications of NGS technology. In the preliminary era of RNA sequencing, 454 technology of the Roche Corporation, Solexa technology of Illumina Corporation, and SOLiD technology of ABI Corporation were the most representative NGS platforms, with new

platforms continuously being launched (Figure 1). In recent years, a wide range of novel applications have been added to RNA-seq, including genome-guided or *de novo* assembly of transcripts, the discovery of new fusion genes in cancer, transcript identification, and the quantification of alternative splicing in tissues, populations, and diseases [24, 25, 33-40]. With continuing technical improvements and decreasing costs, it has become a more and more popular choice for transcriptome studies, and many large projects have been completed, such as the Encyclopedia of DNA Elements, The Cancer Genome Atlas, and projects of the International Cancer Genome Consortium [35, 37, 39]. As in phase I of the MAQC I project, which tested the intra- and interplatform and across-site agreement for gene expression microarrays, the FDA launched the Sequencing Quality Control (SEQC/MAQC-III) project, in which the performance of RNA-seq across laboratory sites is assessed while many types of sequencing platforms and data analysis pipelines are tested. As a result, the measurements of relative expressions have been demonstrated to be accurate and reproducible across sites and platforms [41]. Collaborating in the MAQC projects, the Association of Biomolecular Resource Facilities NGS (ABRF-NGS) study in 2014 provided a broad guideline for cross-platform standardization, evaluation, and improvement of RNA-seq [42].

As a matter of fact, RNA-seq is one type of frequently used quantitative assay with the capability of high-throughput sequencing of DNA fragments. In some sense, the large amounts of DNA fragments can reflect a biological system's repertoire of RNA molecules. In a common RNA-seq process, the first step is the extraction of RNA transcripts from tissues or cells. The extraction contains all species of RNA transcripts, including messenger RNA (mRNA), non-coding RNA, and small RNA. However, unlike small RNAs [microRNA (miRNA), Piwi-interacting RNA (piRNA), short interfering RNA (siRNA), and many others], which can be directly sequenced after adapter ligation, larger RNA (i.e., mRNA) molecules must be fragmented into smaller pieces (200–500 bp) to be compatible with the most deep-sequencing technologies. Because only

mRNA has much biological variation, while the others are conserved, we simply refer to mRNA as all RNA transcripts for convenience. After fragmentation, the mRNA fragments are reversely transcribed to cDNA, and a cDNA library is constructed. After adapter ligation to each cDNA, the library is fast-sequenced on a NGS platform. The process is summarized in Figure 2.

In contrast to the analog-style signals obtained from fluorescent-dye-based microarrays, the NGS platforms produce discrete digital counts of the number of detected transcripts, which are called “read counts.” Typically, these read counts are assigned to a class based on the common mapped region (i.e., gene, exon, or genomic loci) of a genome or a reference transcriptome (Figure 3). This alignment process is called “read mapping” and is achieved by computational tools, such as Tophat [43], Bowtie [44], BWA [45, 46], and HTSeq [47]. In RNA-seq, the number of reads in a class is an important summary statistic, and these read counts have been found as a good approximation to be linearly related to the abundance of the target transcript [28]. Collectively, here we refer to the class as a *gene*, although a class may also refer to, for example, a transcription factor binding site. Therefore, if the RNA (particularly mRNA) of a sample is thoroughly extracted, the abundance of the mapped reads or number of read counts can approximately reflect the real expression level of the corresponding gene to a very high accuracy. As a result of the above wet RNA-seq experiment, researchers can start the downstream analysis with a so-called “count matrix” or “count data,” where each row indicates the gene, each column indicates the sample, and each cell indicates the number of count reads mapped to the gene in the sample [28]. Computer tools, such as BEDTools [48], featureCounts [49], or Cufflinks [50], can be used to produce the so-called count data. In general, the analysis procedure involves the following three steps: normalization of raw count data, statistical modeling of gene expression, and test for DE. In general, the second and third steps are wrapped into one method. Therefore, the analysis procedure is recognized by most people as the following two steps: data normalization and DEG identification.

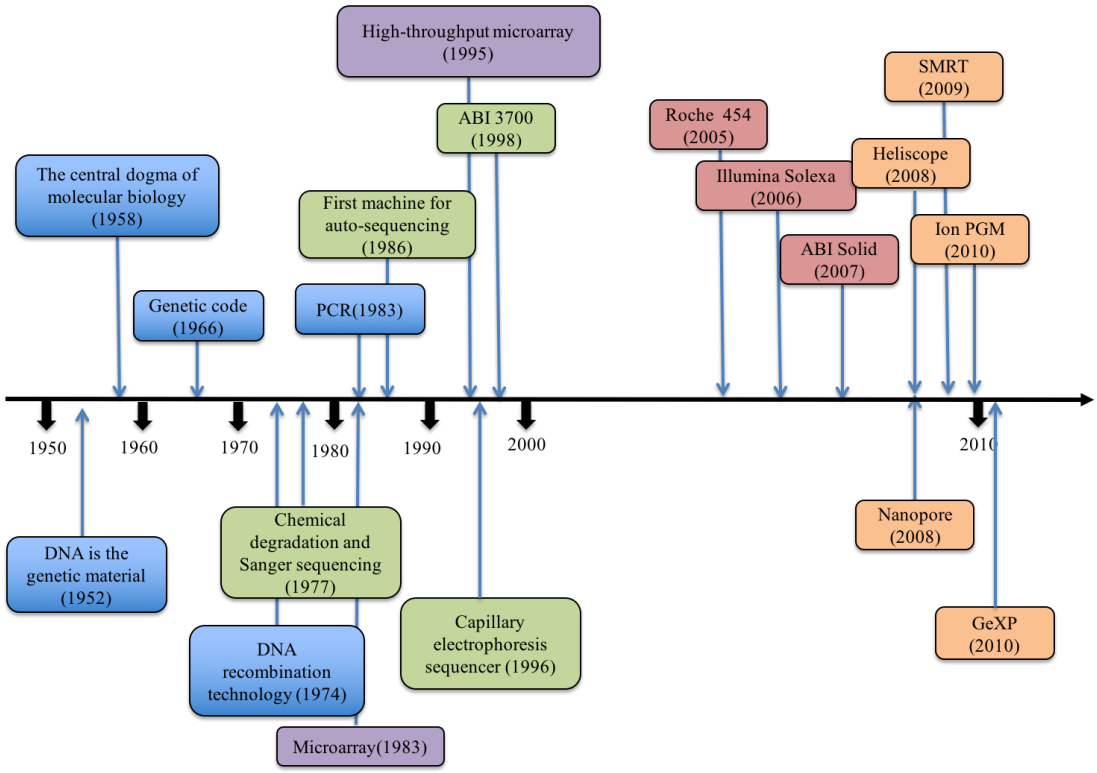


Figure 1 - Modern history of sequencing development in genetics

The lawngreen boxes show the achievements of Sanger sequencing. The deeppink boxes show the representative platforms of NGS. The yellow boxes show the representative platforms of third generation sequencing. The purple boxes show the early development of microarray.

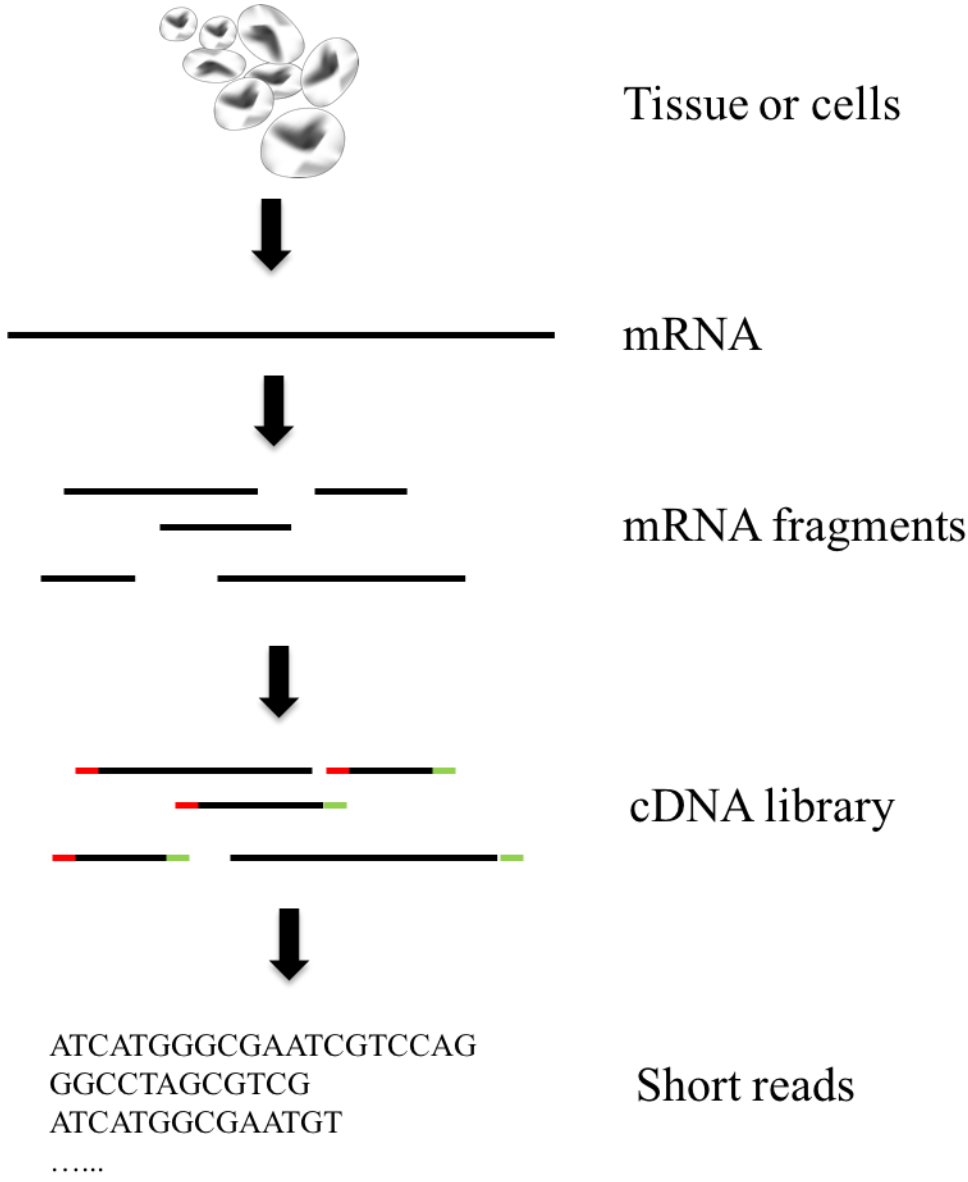


Figure 2 - A typical wet RNA-seq experiment

Briefly, the extracted long RNAs from cells or tissues are first converted into a library of cDNA fragments. Adaptors are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using NGS technology.

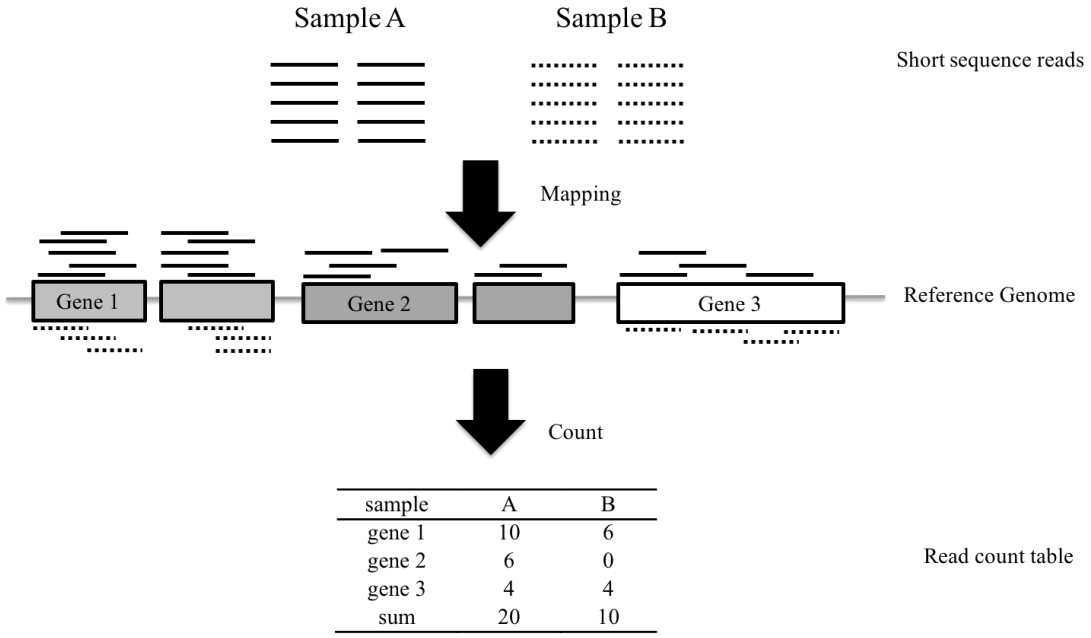


Figure 3 - An illustration for read mapping in RNA-seq data

The resulting sequence reads are aligned to the reference genome. Raw count data will be summarized in a table, where each row indicates the gene (or transcript). Each column indicates the sample (or library), and each cell indicates the number of reads that are mapped to the gene in the sample.

Compared with microarrays, RNA-seq offers several key advantages. First, unlike hybridization-based microarrays, RNA-seq is not limited to detecting transcripts corresponding to existing genomic sequences. A second advantage of RNA-seq relative to microarrays is that it has a very low background signal as DNA sequences can be unambiguously mapped to unique regions of the genome. Further, it does not have an upper limit for quantification, which correlates with the number of sequences obtained. Third, RNA-seq has a large dynamic range of expression levels over which the vast majority of gene expressions can be detected. Finally, without cloning or amplification steps, RNA-seq requires less RNA sample [51, 52]. With the decreasing cost in recent years, more and more researchers prefer RNA-seq as the first option to perform a transcriptome study. As the most cutting-edge technology for bioinformation studies, more and more sequence centers with multiple NGS machines are being set up worldwide (Figure 4). Most of the centers are concentrated in the USA and Western Europe. In Asia, China and Japan take the lead in the RNA-seq field.

However, RNA-seq is a technology under active development, and there remain several challenges. First, although the price of a single run continues to be low, it is still a big financial impediment in case many replicates need to sequenced. Second, gene expression spans several orders of magnitude, with some genes represented by only a few reads. Third, reads originate from mature mRNA (exons only) as well as from incompletely spliced precursor RNA (containing intronic sequences), making it difficult to identify the mature transcripts. And last but not the least, reads are short and genes can have many isoforms as a result of alternative splicing events, which makes it challenging to determine which isoform produced each read. The detailed advantages and disadvantages of RNA-seq are summarized in Table 2.

Because the final aims of microarrays and RNA-seq are the same, and microarrays will still be used on a large scale for a long time in the future, it is required to check the two technologies for consistency. Bottomly *et al.* [53] compared RNA-seq (Illumina GA IIx) with two microarray platforms (Illumina MouseRef-8 v2.0 and Affymetrix MOE 430 2.0) to detect

striatal gene expression between B6 and D2 inbred mouse strains. The overlaps show great concordance (Figure 2 of [53]). Marioni *et al*. [30] estimated gene expression differences between the liver and kidney RNA samples using multiple biological replicates and compared the results from the two different platforms (RNA-seq and microarray). They demonstrated that the RNA-seq platform can detect 81% of DEGs from the microarray platform, and the Spearman correlation coefficient of fold change ratios between them was 0.73, which is higher than that across different microarray platforms in the MAQC I project [17, 30].

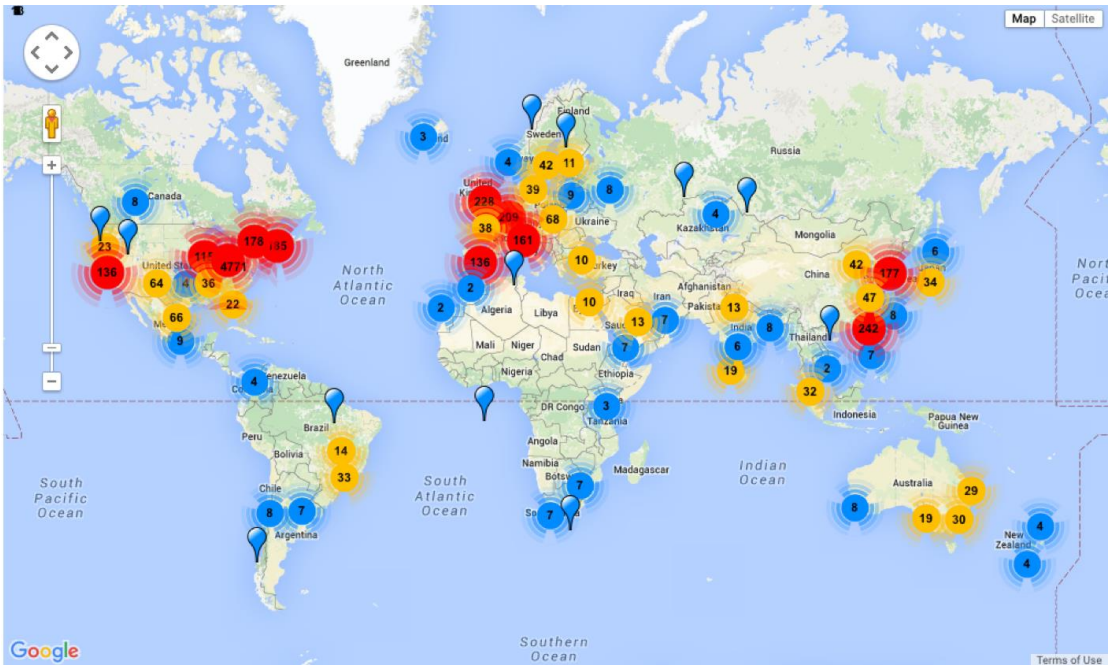


Figure 4 - A world map for the distribution of high-throughput sequencers

The numbers in the cycles indicates the number of sequencers. The platforms include 455, Hiseq, HiSeq X Ten, Illumina GA2, Ion Torrent, MiSeq, MinION, NextSeq, PacBio, Polonator, Proton, and SOLiD. The figure was a screen shot from http://omicsmaps.com/ on Nov.11, 2015.

Table 2 - Advantages and disadvantages of RNA-seq

Advantage	Disadvantage
Provides a comprehensive view of the transcriptome	Technology is new to most researchers
Not dependent on any prior sequence knowledge	Data storage is more challenging
Increasing dynamic range and turntable sensitivity	Analysis is more complex, no standard protocol
Can detect structural variation	Expensive
A truly digital solution (absolute abundance)	Specialized computing infrastructure and personnel are required

1.3 Normalization of read counts

One of the final analysis goals of microarrays and RNA-seq is to identify DE in different samples. Prior to this, however, the variability among the samples should be normalized because it has been demonstrated many times that the normalization procedure has great impact on the inference of DEGs [53-55]. In microarray data analysis, the comparisons of expression levels can be made more reliable by normalizing for systematic biases, such as dye effect and hybridization artifacts. Although these inherent technical biases of microarrays do not exist in RNA-seq experiments, two main sources of systematic variability have been reported in addition to the ones derived from different platforms or sites. The first is within-sample gene-specific effects, such as GC-content biases and gene length. The former shows strong sample-specific effects on RNA-seq; the latter represents a trend that longer genes obtain more reads in the read mapping process [56]. To remove these technical variations, several methods have been proposed [54, 57]. However, in DE analysis, where the genes are individually tested for expression differences between samples, such within-sample biases are usually ignored because they probably contribute equally to all samples. The second variation is between-sample distribution differences in read counts, such as differences in total counts (i.e., sequencing depth or library size). As mentioned above, the gene-specific reads approximate the expression level of the gene. In other words, the sum of the reads in a sample indicates the full expression level of all genes. Therefore, normalization by library size is particularly important for DEG detection in different samples because different samples generally have different library sizes. Most normalization methods are developed to address this issue.

The most straightforward way to adjust for the variation of library sizes is to simply rescale or resample the read counts of all samples to be equal. However, such a normalization neglects the fact that read counts inherently represent the relative expression level of genes. Consider, for example, a situation in which two library sizes are the same but the reads in one sample are evenly distributed while the reads in the other sample

are not, resulting in more falsely called DEGs [53]. To account for this issue and make the counts comparable across samples, inter-sample normalization is achieved by scaling the raw read counts in each sample by a single sample-specific factor reflecting its library size, which is also called the *normalization factor*. In recent years, numerous methods have been proposed for calculating these scaling factors.

To begin with, the famous proposal by Mortazavi *et al.* [28] is to divide the number of reads C_i from a specific gene i with a L_i gene length simply by the total number of reads ($N = \sum_i C_i$) in each library. This normalization procedure is named reads per kilobase of exon model per million mapped reads ($RPKM_i$), which can be viewed as the normalized read count of gene i:

$$RPKM_i = \frac{raw\ read\ counts\ of\ gene\ i\ (C_i)}{mapped\ reads\ (N) \times\ length\ of\ gene\ i\ (L_i)}$$

A variation is fragments per kilobase of exon model per million mapped fragments ($FPKM$) [40]. $FPKM$ corrects differences in both library size and gene length by normalizing the number of reads from a specific gene by both its length and the total number of mapped reads in the sample.

$$FPKM = \frac{total\ gene\ fragments}{mapped\ reads\ (millions)\ \times\ gene\ length\ (KB)}$$

The difference between $RPKM$ and $FPKM$ is the counting object. The former involves mapped reads, whereas the latter involves mapped fragments. Hence, the formulas are slightly different. When raw data originate from paired-end sequencing, fragments are sequenced from both ends, providing two reads for each fragment. However, in the scheme of these two approaches, the proportional representation of each gene is not independent from the expression levels of all other genes. Often large proportions of the reads are mapped onto a small fraction of highly expressed genes, for which small expression changes will skew the counts of lowly expressed genes. As a result, erroneous DE from lowly expressed

genes is easily inferred. Therefore, *RPKM* and *FPKM* mitigate but do not completely eliminate the bias derived from gene length [58].

Per-sample total read count (TC) is a variant of *RPKM*. Of a fixed count for all samples, TC scales each sample to the average total count per sample. More specifically, gene counts are divided by the library size associated with their sample and multiplied by the mean total count from the whole count data.

A quartile of the per-sample count distribution (e.g., upper quartile, UQ) is similar to TC in principle. The difference is that UQ scales the expression level at the 75^{th} percentile of each sample to the average of all samples [59].

Median (Med) is also similar to TC in principle. The difference is that gene counts are divided by the median counts associated with their sample [59].

The normalization factors computed from the above methods are permanent across all samples. Alternatively, the following two methods turn out robust summaries including sample-specific normalization factors by relating each sample to a pre-defined reference sample.

The trimmed means of M values (TMM) normalization method was proposed by Robinson and Oshlack [60] and originally implemented in the edgeR Bioconductor package [61]. TMM has been implemented in many other packages. Its assumption is that most genes are non-DEGs. In executing this method, one sample is considered the reference sample and the others are the test samples. The TMM factor is computed as the weighted mean of log ratios between the reference and test samples, after the exclusion of the most expressed genes, not expressed genes, and genes with the biggest log ratios. According to the above hypothesis, this TMM should be close to 1. If not, an estimate of the correction factor will be provided by the value to adjust the library size [60].

The DESeq normalization method is implemented in the DESeq Bioconductor package [62] and is based on the hypothesis that most of the genes are not DEGs, but non-DEGs. It computes a scaling factor for a given sample by computing the median of the ratio, for each gene, of its read count over its geometric mean across all samples. Because most

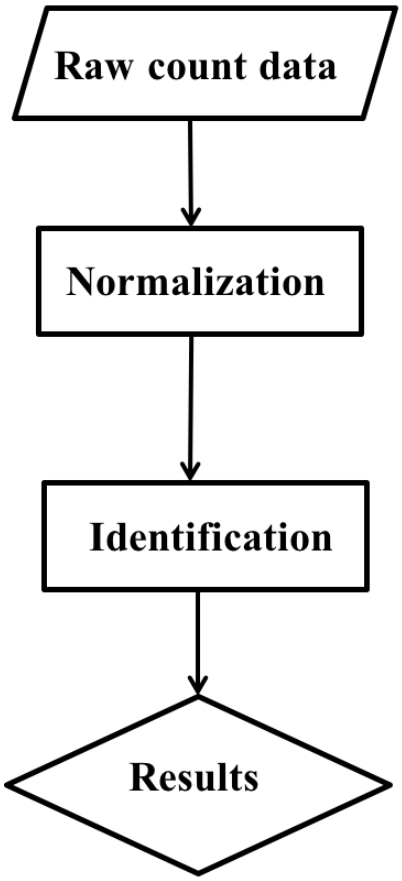
genes are assumed to be non-DEGs, they should have similar read counts across samples, which results in a ratio of 1. The median ratio for each sample can be used to generate a correction factor that should be applied to all read counts for this sample to fulfill the hypothesis.

In addition, there are other methods to calculate statistics similar to normalization factors.

Quantile (Q) normalization was initially proposed for microarray data. Now, it can also be used to deal with RNA-seq count data. Q sorts the counts from each sample and sets the values to be equal to the quantile mean from all samples to make the counts across all samples have the same distribution. It can be implemented in the limma Bioconductor package by calling the "normalize.quantiles" function. Recently, a new normalization function termed voom, designed specifically for RNA-seq data, was added to the limma package. It performs a locally weighted scatterplot smoothing (LOWESS) regression to estimate the mean–variance relationship and transforms the read counts to the appropriate log form for linear modeling [60, 63].

The PoissonSeq Bioconductor package [64] defines a gene set that is least differentiated between two conditions using a goodness-of-fit estimate, which is then used to compute library normalization factors.

Although several systematical comparison studies have been reported, this important step of RNA-seq analysis is still not resolved and completely investigated because of unknown nuisance technical effects [53, 55, 65]. In particular, more complex experiments are usually accompanied more strongly by these unknown effects. In a recent study, Risso *et al*. [29] described a new normalization strategy named remove unwanted variant using pilot data from the SEQC project, which can remove the unwanted variation as much as possible from RNA-seq data.



sample	A	B
gene 1	10	6
gene 2	6	0
gene 3	4	4
sum	20	10

sample	A	B
gene 1	10	12
gene 2	6	0
gene 3	4	8
sum	20	20

sample	*p* value	FDR
gene 1	0.004	0.03
gene 2	0.0031	1
gene 3	0.815	0.1
gene 4	0.0003	0.0001
….	…	…

Figure 5 - From count data to DEG identification in RNA-seq data analysis

There are two main steps in the RNA-seq data analysis: normalization and identification. The purpose of normalization is to reduce the systematic variability. After the normalization step, the expression levels from an individual gene between different samples can be compared. Usually, every gene will be distributed a *p* value or posterior probability, which indicates the significance of DE. A user-defined threshold will be set to cut the genes into DEGs and non-DEGs groups.

1.4 Statistical modeling of read counts

Over the past decades, various statistical methods have been developed for analyze expression profiling data generated by microarrays. Up to now, we can note that the data from microarray and RNA-seq are inherently different. As mentioned earlier, microarray data is “analog” since expression levels are represented as continuous hybridization signal intensities. In contrast, RNA-Seq data is “digital”, representing expression levels as discrete counts. This inherent difference leads to the difference in the parametric statistical methods that since they often depend on the assumptions of the random mechanism that generates the data. For example, the normal distribution is a common distribution for statistical comparisons involving continuous data. It is generally assumed that the log intensities (or expression levels) in a microarray experiment are approximately normally distributed. However, this kind of distribution cannot be directly applied to model the read counts in an RNA-seq experiment without first examining the underlying distributions.

In contrast, there are several kinds of count-based distribution suitable for modeling discrete read counts. Since the short sequence reads were independent samples from a population with given, fixed fractions of genes, the read counts would follow a multinomial distribution, which can be approximated by the Poisson distribution [30, 53]. An essential property of Poisson distribution is that the mean (μ) equals to variance (v). Therefore, read counts across technical replicates derived from a single source fit well to a Poisson distribution. However, most of the time, the variance of gene expression on biological replicates (BRs) derived from different individuals is larger than its mean expression values. In other words, the assumption of Poisson distribution is too restrictive as it predicts smaller variations than that is seen in the data with BRs.

To address this so-called over-dispersion problem, for data with BRs derived from different individuals, the read counts well fit to an over-

dispersed Poisson distribution such as negative binomial distribution (NB) [61, 62, 66], beta-binomial (BB) model [67, 68], Poisson-Tweedie model [69] and so on. In particular, the Poisson-Tweedie model well captures the biological variation (especially for zero-inflation and heavy tail behavior, for details see [69]) when many BRs are available. As an increase in sample size (i.e., the number of BRs) precedes an increase in sequencing depth (i.e., the number of sequenced reads) [70-72], a more complex model such as Poisson-Tweedie may be the first choice for count data with many BRs. However, as many replicates are still difficult to take due to sequencing cost and the small amount of the target RNA sample, RNA-seq data with few BRs have mainly been stored. As a result, the methods based on the NB model have been widely used as a common choice for DE analysis of RNA-seq data with few BRs [61, 62, 73, 74].

In fact, the NB distribution describes a failure distribution (y) of one event, whose incidence (p) is permanent in a Bernoulli's experiment. In that experiment, the event comes out r times. The conclusive parameter (Y) indicating the probability of all failures will be calculated by the following formula.

$$f(y;p,r) = P(Y = y) = \frac{\Gamma(y)}{\Gamma(r)\Gamma(y-r+1)} p^r (1-p)^{y-r}$$

However, to read counts, the expected value and dispersion is very important. Assuming the expected value is μ and the dispersion value is ϕ ($\phi > 0$), the p and r will be replaced like in the following equations.

$$p = \frac{1}{1+\mu\phi}$$

$$r = \phi^{-1}$$

After the replacement, the distribution parameters (μ and ϕ) will change the above the formula.

$$f(y;\mu,\phi) = P(Y = y) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)} \left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}} \left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y}$$

Then the expected value and variance will be described like in the following formulas.

$$E(Y) = \mu$$

$$V(Y) = \mu + \phi\mu^2$$

In other words, when NB distribution is used to model the read counts, the NB distribution has parameters, which are uniquely determined by mean μ and variance v. Their relation is defined as $v = \mu + \phi\mu^2$, where ϕ is the dispersion factor. This conversion was first introduced by Robinson *et al.* [61] and was expanded in some other studies.

1.5 DEG identification

As mentioned above, the relative expression level of genes can be quantified by the read counts using RNA-seq technology. Based on its high throughput, RNA-seq can capture the expression of almost all genes in a specific sample. The set of gene-wise counts makes up the expression profile for the sample, in which the expression level of unknown genes can also be seen. After the normalization of raw read counts, it becomes possible to find the causes of different physiological characteristics between samples by testing DE. As a result, it becomes easier for researchers to analyze the differences between samples at a higher resolution, particularly to identify potential DEGs. Most of the popular normalization methods are model-based, as described in Section 1.3. Usually, the test is performed after the estimation of the parameters for the appropriate statistical model.

Take the detection of DEGs from two samples as an example. The raw data are normalized with various normalization methods. After that, or at the same time, the parameters of the statistical model fitted for this data are estimated before testing DE. As a result, a *p* value will be calculated for each gene. A low *p* value suggests that the possibility that the expression levels of a specific gene across samples are occasionally the same is very low. Conversely, the possibility that the gene is differentially expressed in the samples is high. To address the multiplicity problem of *p* values for multiple genes, the false discovery rate (FDR) controlling approach proposed by Benjamini and Hochberg is adopted. The genes are subsequently listed by ascending order of their FDRs. Given a threshold, the genes whose FDRs are smaller than the threshold will be identified as DEGs, while the others are identified as non-DEGs. In other words, in the new list, DEGs are top-ranked while non-DEGs are bottom-ranked and their order results from the significance of DE.

A variation of Fisher's exact test is used in both the edgeR and DESeq packages. It is adopted for NB distribution, and the returned *p* values are calculated from the derived probabilities [61, 62].

Limma uses a moderated *t*-statistic to compute *p* values in which both the standard error and the degrees of freedom are modified. The standard

error is moderated across genes with a shrinkage factor, which effectively borrows information from all genes to improve the inference on any single gene. The degrees of freedom are also adjusted by a term that represents the a priori number of degrees of freedom for the model [75].

The baySeq approach estimates two models (DEG and non-DEG) for every gene. The posterior likelihood of the model for DEG, given the observed data, is used [66].

In the PoissonSeq method, the test for DE is simply a test for the significance of the g_i term (i.e., the correlation of gene i expression with the two conditions), which is evaluated by score statistics. By simulation experiments, it was shown that these score statistics follow a chi-squared distribution, which is used to derive p values for DE [64].

The test statistics $T = E[\log(x)]/\mathrm{Var}[\log(x)]$ is employed by Cuffdiff, where x is the ratio of the normalized counts between two conditions and this ratio approximately follows a normal distribution; therefore, a t-test is used to calculate the p value for DE.

All methods use standard approaches for multiple hypothesis correction (i.e., Benjamin–Hochberg) except for PoissonSeq, which implements a novel estimation of FDR for count data that is based on permutation.

1.6 The purpose of this study

Since 2005, the RNA-seq technology has been more and more popular in many biological fields. However, lacking of effective and robust analysis methods limits the researchers to obtain reliable bioinformation. Although there are many well-established methods available for DE analysis in microarray, they cannot be immediately transferable to the analysis of RNA-seq data since the radical difference between the two kinds of digital data. Therefore, in the past decade, the bioinformatics community is continuously launching methods for analysis of RNA-seq count data (Table 3). Meanwhile, several methods for microarray data analysis have been adapted to RNA-seq count data and it was demonstrated that the adapted methods perform comparably to the methods designed for RNA-seq [75]. In order to compare these methods comprehensively, several evaluation studies have also been reported [71, 76]. However, these evaluations are limited the two-group comparisons (i.e., two cellular conditions or phenotypes). On the other hand, with the ongoing update, more and more methods start to have the capability of dealing with multi-group (>2) or multi-factored RNA-seq experiments where multiple biological conditions and different sequencing protocols are included [61, 62]. Therefore, accumulations of comparative studies for multi-group data are also desired.

In this study, we elaborate the exact usages of the state-of-the-art methods, which have the capability of dealing with multi-group RNA-seq count data. We also compare 12 pipelines available in nine R packages for detecting DE from multi-group RNA-seq count data, focusing on three-group data with or without replicates. We evaluate those pipelines on the basis of both simulation data and real count data [77].

Table 3 - Methods for calling DEGs in RNA-seq data analysis

Method	Proposed Year	Statistical Model	Multiple Group	Programming Language	Reference
DESeq	2010	NB	Yes	R	[62]
edgeR	2010	NB	Yes	R	[61]
DEGseq	2010	NB	No	R	[78]
baySeq	2010	NB	Yes	R	[66]
GPseq	2010	Poisson	No	R	[79]
ASC	2010	NB	No	R	[80]
NOIseq	2011	NULL	No	R	[81]
TSPM	2011	Poisson	No	R	[82]
NBPSeq	2011	NB	No	R	[83]
PoissonSeq	2012	Poisson	Yes	R	[64]
BitSeq	2012	NB	No	C/C++	[84]
QuasiSeq	2012	NB	No	R	[85]
GFOLD	2012	Poisson	No	C/C++	[86]
TCC	2013	NB	Yes	R	[87]
Cuffdiff2	2013	NB	No	C/C++	[88]
SAMseq	2013	NULL	Yes	R	[89]
EBSeq	2013	NB	Yes	R	[90]
DSS	2013	NB	No	R	[91]
ShrinkSeq	2013	NB	No	R	[92]
NPEBseq	2013	NB	No	R	[93]
DESeq2	2014	NB	Yes	R	[94]
voom	2014	NB	Yes	R	[75]
BADGE	2014	NB	No	Matlab	[95]
edgeR_robust	2014	NB	Yes	R	[96]

This table lists 24 methods for DE detection (ascending order by the publication), most of which can be implemented by installing the R packages from Bioconductor or CRAN website. Null in statistical model column means the methods are non-parametrical.

Chapter 2 Analysis methods for RNA-seq data analysis

2.1 DE analysis methods investigated in the present study

In the above chapter, we have illustrated the two steps for analyzing two-group RNA-seq count data in detail: one step for normalizing raw count data and the other step for identifying DEGs. In our previous studies [87, 97], we regarded the two steps as *X* for data normalization and *Y* for DEG identification. Therefore, we refer *X*-*Y* pipeline to describe the analysis procedure of RNA-seq count data. In most of the public R packages in Bioconductor and CRAN websites, each of the R packages has its own methods for the elements of *X*-*Y* pipeline.

Here we take the most frequently used edgeR as examples to roughly illustrate our analysis design. The edgeR manipulates the raw RNA-seq count data as input. It first calculates normalization factors (or size factors) for individual sample as *X*, then construct the model (i.e., estimate the parameters on the model in which the calculated normalization factors are used to re-scale the raw counts), and calculate *p* values (i.e., perform the statistical test using the model) as *Y*. Previous studies have demonstrated that *X* has more impact than *Y* on the ranked gene list [97-99], the normalization method TMM (Trimmed Means of M values) and the identification method adapted Fisher's exact Test implemented in edgeR for two-group data comparison generally give satisfactory results [55]. When comparing multi-group data, the default normalization method is also TMM while the default DEG identification method is the likelihood ratio test based on generalized linear models (GLM) whose error structure follows the negative binomial distribution. One of the models corresponds to alternative hypothesis and the other corresponds to null hypothesis. In this study, we termed the default pipelines *X*-*Y* for edgeR as "edgeR - edgeR (or *E*-*E*)" (Figure 6). Accordingly, the default procedures in DESeq and DESeq2 can also be described as *D*-*D* and *S*-*S*.

In section 1.3, we discussed that a more accurate normalization factor can estimate the gene expression more precisely. For this purpose, we previously designed a robust multi-step normalization procedure called TbT [97]. According to the above suggested theory, TbT consists of three steps: *X* using TMM (step 1), *Y* using an empirical Bayesian method implemented in the baySeq package [66] (step 2), and *X* using TMM after elimination of the estimated DEGs (step 3) comprising the TMM-baySeq-TMM normalization pipeline. The key concept is to alleviate the negative effect of potential DEGs before calculating the normalization factor in step 3. As

mentioned previously [97], the DEG elimination strategy (called DEGES) can be repeated until the calculated normalization factors converge. The iterative TbT can be described as a TMM-(baySeq-TMM)$_n$ procedure. Accordingly, a generalized pipeline with the multi-step normalization can be described as *X*-(*Y*-*X*)$_n$-*Y* in which the *X*-(*Y*-*X*)$_n$ with $n >= 2$ corresponds to the iterative DEGES-based normalization (Figure 7).

Our TCC package [87] implements the proposed pipeline *X*-(*Y*-*X*)$_n$-*Y*. Recommendations are made from an extensive simulation analysis: (1) edgeR-(edgeR-edgeR)$_3$-edgeR on two-group RNA-seq data with few replicates and (2) DESeq-(DESeq-DESeq)$_3$-DESeq on two-group data without replicates [12]. However, similar to many other studies [70-73, 76, 99, 100], the performance evaluations were limited to a two-group comparison. While many R packages as well as TCC can perform DE analysis on more complex experimental designs [66, 67, 74, 94, 96, 101, 102], there have been few evaluation studies on RNA-seq data three-group data.

To investigate the performance of DE pipelines for a multi-group comparison, a total of 12 pipelines available in nine packages were mainly evaluated in this study: TCC (ver. 1.7.15) [87], edgeR (ver. 3.8.5) [61], DESeq (ver. 1.18.0) [62], DESeq2 (ver. 1.6.3) [94], voom [38] in limma (ver. 3.22.1) [75], SAMseq [89] in samr (ver. 2.0), PoissonSeq (ver. 1.1.2) [64], baySeq (ver. 2.0.50) [66], and EBSeq (ver. 1.6.0) [90].

The initial aim of current study was to evaluate 12 pipelines available in nine R packages when analyzing multi-group RNA-seq data. In particular, our primary interest is to investigate the effectiveness of the DEGES-based pipeline in TCC under such more complex designs. We report pipelines suitable for multi-group comparison. Note that TCC can perform several combinations for the DE pipeline *X*-(*Y*-*X*)$_n$-*Y* with $n = 3$ as recommended [87]. We sometimes refer to this DEGES-based pipeline as *XYX-Y* with the fixed number of *n* for short. We basically confine individual methods (*X* and *Y*) in each pipeline to functions provided by the same packages (i.e., edgeR or DESeq2) for simplicity. That is, the edgeR-related pipeline is "edgeR-(edgeR-edgeR)$_3$-edgeR", where *X* = TMM and *Y* = DEG identification method, implemented in edgeR. Although we previously termed this pipeline "iDEGES/edgeR-edgeR" [87], here we abbreviate it to *EEE-E* for convenience. Similarly, the "DESeq-(DESeq-DESeq)$_3$-DESeq" pipeline can be shortened to *DDD-D*. This is because (1) users can select, for example, different DEG identification methods *Y* for steps 2 and 4 and (2) we will discuss some possible combinations such as *DED-S* for the "DESeq-(edgeR-

DESeq)$_3$-DESeq2" pipeline. In this sense, the DEGES-based pipeline can also be denoted as *X*-(*Y*-*X*)$_n$-*Z* or *XYX*-*Z*. For convenience, we summarized the information of all pipelines in Table 3.

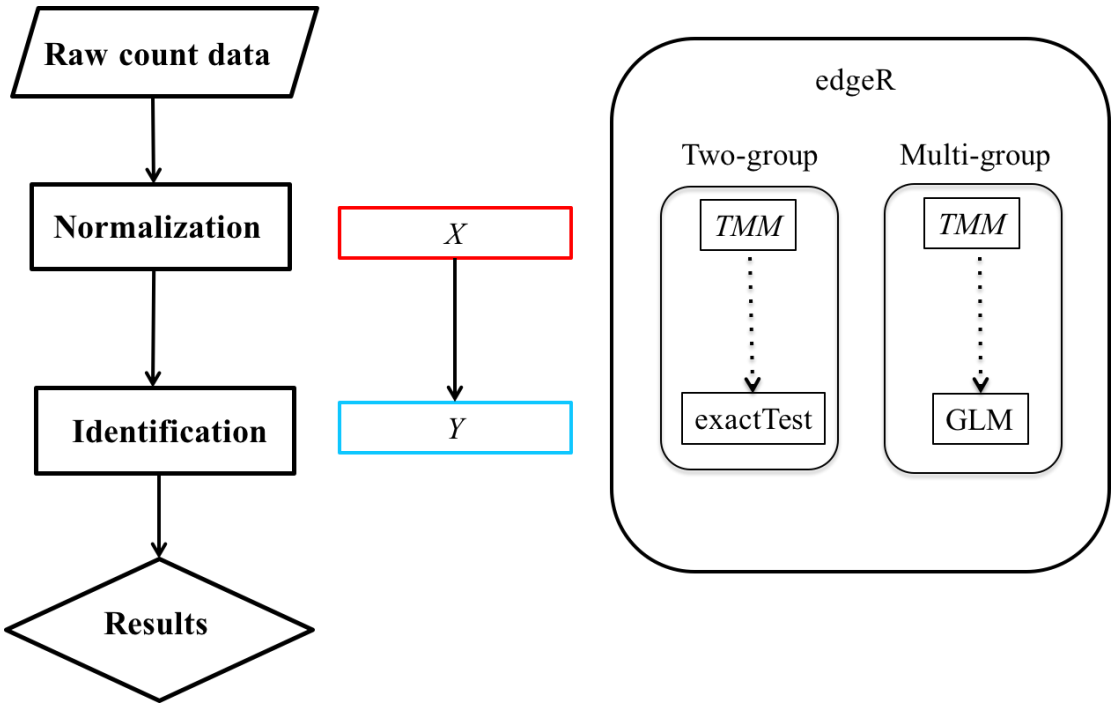


Figure 6 - Traditional two step procedure for RNA-seq data analysis

In this present study, we refer to the two steps as *X* for data normalization and *Y* for DEG identification. Taking edgeR package as an example, TMM is for *X*, an adapted Fisher's exact test for *Y* in two-group comparison and a statistical test in GLM model for *Y* in multi-group comparison.

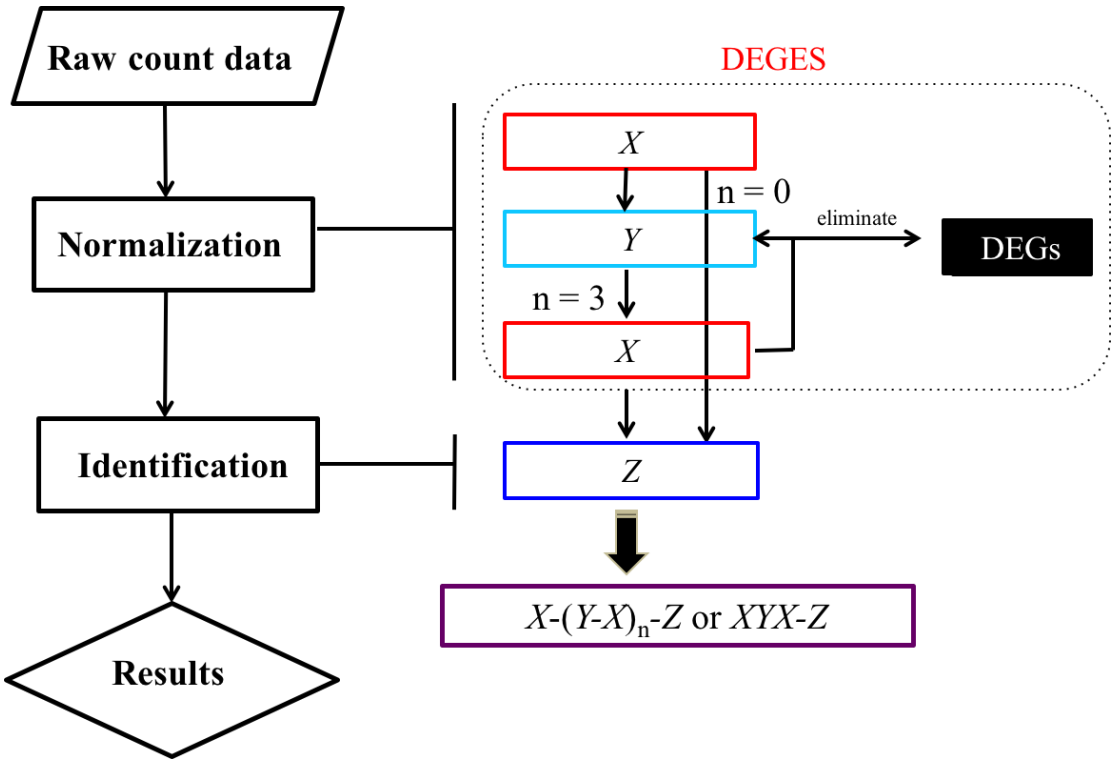


Figure 7 - DE analysis pipeline with DEGES-based normalization method

A generalized pipeline with the multi-step normalization can be described as X-(Y-X)$_n$-Y in which the X-(Y-X)$_n$ with n >= 2 corresponds to the iterative DEGES-based normalization.

Table 4 - Information about all of the pipelines involved in this study

Pipeline	Abbreviation	Package	Version	DEGES -based	Bayesian -based
voom	voom	limma	v3.22.1	-	-
SAMseq	SAMseq	samr	v2.0	-	-
PoissonSeq	PoissonSeq	PoissonSeq	v1.1.2	-	-
baySeq	baySeq	baySeq	v2.0.50	-	+
EBSeq	EBSeq	EBSeq	v1.6.0	-	+
edgeR-edgeR	*E-E* (edgeR)	TCC	v1.7.15	-	-
edgeR_robust	edgeR_robust	edgeR	v3.8.5	-	-
DESeq-DESeq	*D-D* (DESeq)	TCC	v1.7.15	-	-
DESeq2-DESeq2	*S-S* (DESeq2)	TCC	v1.7.15	-	-
edgeR-(edgeR-edgeR)$_3$-edgeR	*EEE-E* (TCC)	TCC	v1.7.15	+	-
DESeq-(DESeq-DESeq)$_3$-DESeq	*DDD-D* (TCC)	TCC	v1.7.15	+	-
DESeq2-(DESeq2-DESeq2)$_3$-DESe2	*SSS-S* (TCC)	TCC	v1.7.15	+	-
edgeR-DESeq	*E-D*	TCC	v1.7.15	-	-
DESeq-edgeR	*D-E*	TCC	v1.7.15	-	-
edgeR-DESeq2	*E-S*	TCC	v1.7.15	-	-
DESeq-DESeq2	*D-S*	TCC	v1.7.15	-	-
DESeq-(edgeR-DESeq)$_3$-edgeR	*DED-E*	TCC	v1.7.15	+	-
edgeR-(DESeq-edgeR)$_3$-edgeR	*EDE-E*	TCC	v1.7.15	+	-
DESeq-(DESeq-DESeq)$_3$-edgeR	*DDD-E*	TCC	v1.7.15	+	-
edgeR-(edgeR-edgeR)$_3$-DESeq	*EEE-D*	TCC	v1.7.15	+	-
edgeR-(DESeq-edgeR)$_3$-DESeq	*EDE-D*	TCC	v1.7.15	+	-
DESeq-(edgeR-DESeq)$_3$-DESeq	*DED-D*	TCC	v1.7.15	+	-
edgeR-(edgeR-edgeR)$_3$-DESeq2	*EEE-S*	TCC	v1.7.15	+	-
DESeq-(edgeR-DESeq)$_3$-DESeq2	*DED-S*	TCC	v1.7.15	+	-
edgeR-(DESeq-edgeR)$_3$-DESeq2	*EDE-S*	TCC	v1.7.15	+	-
DESeq-(DESeq-DESeq)$_3$-DESeq2	*DDD-S*	TCC	v1.7.15	+	-

Information for all pipelines involved in this thesis. The pipelines with "-" symbol in abbreviation are constructed in TCC package. Note that the pipelines in edgeR, DESeq and DESeq2 packages are originally inherited in TCC package. Normalization and identification methods of the above 12 pipelines in Table 4a are from individual package while the bottom 14 pipelines in Table 4b are from two different packages.

2.2 DE analysis using individual packages

As shown in Table 3, all the DEGES-based pipelines *X*-(*Y*-*X*)$_n$-*Z* or *XYX*-*Z* were performed using the TCC package. This kind of pipeline includes *EEE-E*, *DED-E*, *EDE-E*, *DDD-E*, *EEE-D*, *DED-D*, *EDE-D*, *DDD-D*, *SSS-S*, *EEE-S*, *DED-S*, *EDE-S*, and *DDD-S*. Four other pipelines (*D-E*, *E-D*, *E-S*, and *D-S*) were also performed using this package, since hey were the hybrid ones originally implemented in different packages and only can be performed via TCC package. These DEGES-based and non-DEGES-based pipelines were performed using two functions ("calcNormFactors" and "estimateDE") in the package. The genes were ranked in ascending order of the *p* values. The *p* value adjustment for the multiple-testing problem was performed using the "p.adjust" function with *method="BH"* option (Benjamin-Hochberg FDR calculation).

Two pipelines, *E-E* (the same as the default edgeR procedure) and edgeR_robust, were performed using the edgeR package. The *E-E* pipeline for analyzing count data with replicates was performed using the following functions: "DGEList", "calcNormFactors", "estimateGLMCommonDisp" with default options, "estimateGLMTrendedDisp", "estimateGLMTagwiseDisp", "glmFit", and "glmLRT." When analyzing count data without replicates, the "estimateGLMCommonDisp" function with three options (*method="deviance"*, *robust=TRUE*, and *subset=NULL*) was used and two functions ("estimateGLMTrendedDisp" and "estimateGLMTagwiseDisp") were not used, as suggested. The edgeR_robust method was performed using the following functions: "DGEList", "calcNormFactors", "estimateGLMRobustDisp", "glmFit", and "glmLRT." The gene ranking and *p* value adjustment procedure were performed in the same way as described above.

The pipeline *D-D* in the DESeq package was performed using the following functions: "newCountDataSet", "estimateSizeFactors", "estimateDispersions" with default options for analyzing data with replicates, and "fitNbinomGLMs." For analyzing data without replicates, the "estimateDispersions" function with following options was used as

suggested: *method="blind"* and *sharingMode="fit-only."* The genes were ranked in ascending order of the *p* values. The *p* value adjustment for the multiple-testing problem was performed using the "p.adjust" function with *method="BH"* option (Benjamin-Hochberg FDR calculation).

The pipeline *S-S* in the DESeq2 package was performed using the following functions: "DESeqDataSetFromMatrix", "estimateSizeFactors", "estimateDispersions", and "nbinomLRT." The genes were ranked in ascending order of the *p* values. Since this package provides adjusted *p* values, the number of DEGs satisfying the 5% FDR threshold was obtained using the values.

The pipeline voom in the limma package was performed using the following functions: "DGEList", "calcNormFactors" in edgeR, "voom", "lmFit", "eBayes", and "topTable". The gene ranking was performed using the resultant *p* values. Since this package provides adjusted *p* values, the number of DEGs satisfying the 5% FDR threshold was obtained using the values.

The pipeline SAMseq in the samr package was performed using the "SAMseq" function with following options: *nperms*=100, *nresamp*=20, *resp.type*="*Multiclass*", and *fdr.output*=1.0. Since this package only provides adjusted *p* values, the gene ranking was performed using the adjusted *p* values.

The pipeline PoissonSeq was performed the "PS.Main" function with *npermu*=500 option. The gene ranking was performed using the resultant *p* values. Since this package provides adjusted *p* values, the number of DEGs satisfying the 5% FDR threshold was obtained using the values.

The pipeline baySeq was performed using the following functions: "new", "getLibsizes" with *estimationType="edgeR"* option, "getPriors.NB" with *samplesize*=5000 and *estimation*="*QL*" options, "getLikelihoods" with *pET*="*BIC*" option, and "topCounts." Since this package only provides adjusted *p* values, the gene ranking was performed using the values. The *ordering* information in the output of the "topCounts" function was used for classifying the expression patterns of genes.

The pipeline EBSeq was performed using the following functions: "GetPatterns", "MedianNorm", "EBMultiTest" with three options

(*maxround*=5, *Qtrm*=1.0, and *QtrmCut*=-1), and "GetMultiPP." There are five expression patterns to consider when comparing three-group data. The "EBMultiTest" function was performed with the consideration of all the five possible patterns. The posterior probability obtained from the "non-DEG" pattern was used as a surrogate estimate for the adjusted *p* values. The gene ranking was performed using the values. The *MAP* information in the output of the "GetMultiPP" function was used for classifying the expression patterns of genes. Most of the above mentioned options were tested in a small scale in order to obtain the optimum paramters or the optimized option combination in one package. For example, we changed the maxit parameter of edgeR_robust from 5 to 12 and 24. Or we test different combinations with three options of DESeq package, “method”, “sharingMode”, and “fitType” in “estimateDispersions” function. Table 5 summarizes the functions and options interrogated in this study.

Table 5 - Average AUC values for simulation data with various options

Pipeline	***E-E* (edgeR)**		
Simulation condition	P_{DEG} = 5%, (0.5, 0.4, 0.1) for (PG1, PG2, PG3), and Nrep = 3		
calcNormFactors	glmLRT		AUC
method="TMM"	***test = "chisq"***		**91.47**
method="RLE"	***test = "chisq"***		91.46
method="upperquartile"	***test = "chisq"***		91.40
method="none"	***test = "chisq"***		91.19
method="TMM"	*test = "F"*		91.47
method="RLE"	*test = "F"*		91.46
method="upperquartile"	*test = "F"*		91.40
method="none"	*test = "F"*		91.19
Pipeline	***D-D* (DESeq)**		
Simulation condition	P_{DEG} = 5%, (0.5, 0.4, 0.1) for (PG1, PG2, PG3), and Nrep = 3		
	estimateDispersions		AUC
method="pooled"	***sharingMode = "maximum"***	***fitType = "parametric"***	90.60
method="pooled-CR"	***sharingMode = "maximum"***	***fitType = "parametric"***	90.43
method="blind"	***sharingMode = "maximum"***	***fitType = "parametric"***	90.47
method="pooled"	*sharingMode = "fit-only"*	***fitType = "parametric"***	90.95
method="pooled-CR"	*sharingMode = "fit-only"*	***fitType = "parametric"***	90.78
method="blind"	*sharingMode = "fit-only"*	***fitType = "parametric"***	91.46
method="pooled"	*sharingMode = "gene-est-only"*	***fitType = "parametric"***	86.79
method="pooled-CR"	*sharingMode = "gene-est-only"*	***fitType = "parametric"***	87.18
method="blind"	*sharingMode = "gene-est-only"*	***fitType = "parametric"***	85.57
method="pooled"	***sharingMode = "maximum"***	*fitType = "local"*	90.78
method="pooled-CR"	***sharingMode = "maximum"***	*fitType = "local"*	87.11
method="blind"	***sharingMode = "maximum"***	*fitType = "local"*	90.24
method="pooled"	*sharingMode = "fit-only"*	*fitType = "local"*	91.40
method="pooled-CR"	*sharingMode = "fit-only"*	*fitType = "local"*	85.56
method="blind"	*sharingMode = "fit-only"*	*fitType = "local"*	**91.69**
method="pooled"	*sharingMode = "gene-est-only"*	*fitType = "local"*	86.79
method="pooled-CR"	*sharingMode = "gene-est-only"*	*fitType = "local"*	87.18
method="blind"	*sharingMode = "gene-est-only"*	*fitType = "local"*	85.57
Pipeline	***S-S* (DESeq2)**		
Simulation condition	P_{DEG} = 25%, (0.5, 0.4, 0.1) for (PG1, PG2, PG3), and Nrep = 1		
estimateSizeFactors	estimateDispersions		
type="ratio"	***fitType="parametric"***		**82.01**
type="iterate"	***fitType="parametric"***		81.91
type="ratio"	*fitType="local"*		81.53
type="iterate"	*fitType="local"*		81.31
type="ratio"	*fitType="mean"*		76.02
type="iterate"	*fitType="mean"*		75.84

Average AUC values of 100 trails are shown. The suggested (or default) options and the highest AUC values are in bold.

2.3 ROC curve and AUC value

As mentioned above, in terms of hypothesis tests in DEG identification step, each gene will be predicted as DEG or non-DEG according to a specific threshold. In a simulation trial, all of the genes are preset to DEGs or non-DEGs, which can be regarded as real DEGs and real non-DEGs. If a real DEG is predicted as a DEG., we call the prediction is True Positive (TP); on the contrary, if the gene is predicted as a non-DEG, we call the prediction is False Negative (FN). Similarly, if a real non-DEG is predicted as a DEG, the prediction is False Positive (FP); the gene is predicted as non-DEG, the prediction is True Negative (TN). The true positive rate (TPR) and the true positive rate (TPR) can be calculated by the two following formulas.

$$TPR = \frac{TP}{FN + TP}$$

$$FPR = \frac{FP}{FP + TN}$$

The TPR is also known as sensitivity corresponding to Type I error while the FPR is also known as fall-out and can be calculated as (1 - specificity). In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting TPR against FPR at various threshold settings. The area under the ROC curve (i.e., AUC) values is used for evaluating individual combinations based on sensitivity and specificity simultaneously. A good combination should therefore have a high AUC value, which indicates high sensitivity and specificity. In this study, all of the AUC values are expressed in three or four digital percentage.

In practice, gene list ranked in accordance with the level of DE are pre-required for calculating AUC values. According to the level (i.e., *p* value × (−1)), a list of scores will be generated and the genes are ranked by the scores. Take the table of Figure 8 as an example, the larger ranking number indicates bigger probability of DE. A variable is set to decrease

progressively from 10000 to 1 with the interval of 1. In a trial, the genes with larger ranking number than the specific value of the variable is predicted as DEG, the others are predicted as non-DEGs. Then TPR and FPR can be calculated and used as a couple of coordinates for plotting a dot. Accordingly, there will be more and more dots in the plotting board with the decreasing of the variable. As a result, 10000 dots are linearly bound up to a line named ROC curve. The area under the curve is calculated as AUC value. Therefore, if the scores are distributed to the DEGs and non-DEGs randomly, the AUC value of the line connecting the start point (0, 0) and end pint (1, 1) is approximately 0.5 (green line in Figure 8). In case of DEGs with high scores and non-DEGs with low scores, the AUC value will reach 1 (red line in figure 8). In the opposite case, the AUC value will decline to 0.

To the wet lab biologists, in the process of biological data analysis, they just focus on the top significant differentially expressed genes (e.g., 10000 * 1% = 100). In this situation, the variable will be changed from 10000 to 9900. The computed AUC value from this is called partial AUC value, versus the full AUC values introduced above. In this study, we use “pAUC” function in pROC CRAN package to calculate partial AUC values.

Sample	Real DE	Predictions		
		Ranking (black line)	Ranking (read line)	Ranking (green line)
gene1	DEG	10000	9000	10
gene2	DEG	9999	7000	6000
…	…	…		
gene9999	non-DEG	2	50	500
gene10000	non-DEG	1	200	1000

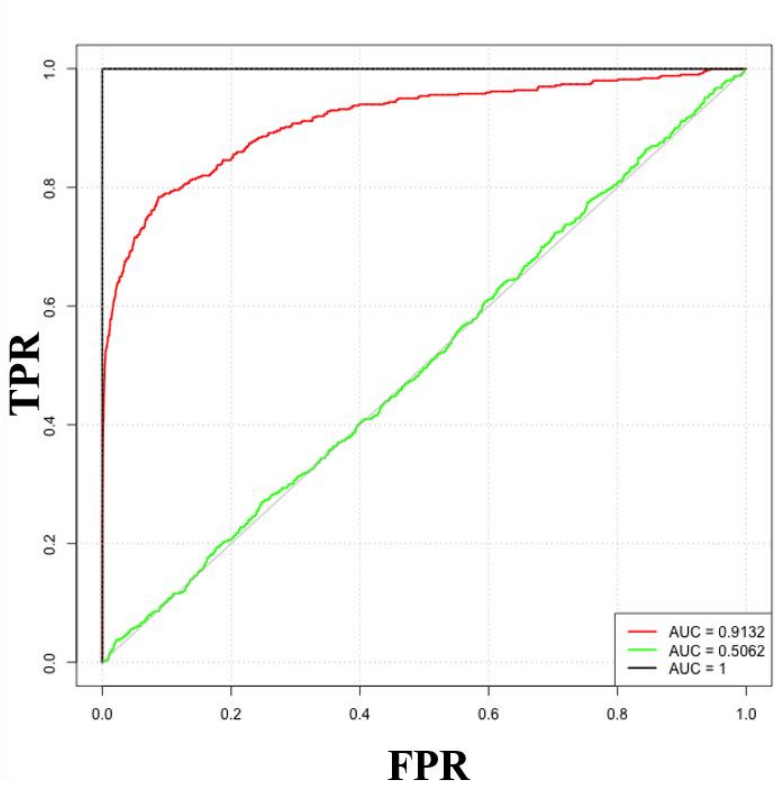


Figure 8 - ROC curve plotting and its characteristics

ROC curve is one kind of approach for evaluating the DEG detecting power by comparing the DEG information from predicted results and fact. It is plotted by true positive rate (TPR) against false positive rate (FPR). If the prediction results are generated randomly, the curve (green line) is close to the grey diagonal line. If most of the predicated DEGs hit, the curve (red line) approaches the left-up corner. The area between the curve and x-axis can be computed and the resulting value is quantified as AUC value (area under the curve). In sum, the AUC value indicates the performance of one prediction trial. The larger the AUC value is, the better the prediction is.

2.4 Computer environment

All analyses were performed using R (ver. 3.2.0 pre-release) and Bioconductor [103]. The following is the displayed information about the computer environment for this study after entering the "sessionInfo()" command in the R user interface.

```
> sessionInfo()
R version 3.2.0 pre-release (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: CentOS release 6.2 (Final)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
 [1] splines   stats4    parallel  stats     graphics  grDevices utils
 [8] datasets  methods   base

other attached packages:
 [1] EBSeq_1.6.0             gplots_2.17.0           blockmodeling_0.1.8
 [4] PoissonSeq_1.1.2        combinat_0.0-8          samr_2.0
 [7] matrixStats_0.14.2      impute_1.42.0           TCC_1.7.15
[10] ROC_1.44.0              baySeq_2.0.50           perm_1.0-0.0
[13] abind_1.4-3             edgeR_3.8.5             limma_3.22.1
[16] DESeq2_1.6.3            RcppArmadillo_0.5.300.4 Rcpp_0.12.0
[19] GenomicRanges_1.20.5    GenomeInfoDb_1.4.1      IRanges_2.2.7
[22] S4Vectors_0.6.3         DESeq_1.18.0            lattice_0.20-33
[25] locfit_1.5-9.1          Biobase_2.28.0          BiocGenerics_0.14.0

loaded via a namespace (and not attached):
 [1] gtools_3.5.0         digest_0.6.8         plyr_1.8.3
 [4] futile.options_1.0.0 acepack_1.3-3.3      RSQLite_1.0.0
 [7] ggplot2_1.0.1        gdata_2.17.0         annotate_1.46.1
[10] rpart_4.1-10         proto_0.3-10         BiocParallel_1.2.20
[13] geneplotter_1.46.0   stringr_1.0.0        foreign_0.8-65
[16] munsell_0.4.2        nnet_7.3-10          gridExtra_2.0.0
[19] Hmisc_3.16-0         XML_3.98-1.3         bitops_1.0-6
[22] MASS_7.3-43          grid_3.2.0           xtable_1.7-4
[25] gtable_0.1.2         DBI_0.3.1            magrittr_1.5
[28] scales_0.2.5         KernSmooth_2.23-15   stringi_0.5-5
[31] XVector_0.8.0        reshape2_1.4.1       genefilter_1.50.0
[34] latticeExtra_0.6-26  futile.logger_1.4.1  Formula_1.2-1
[37] lambda.r_1.1.7       RColorBrewer_1.1-2   tools_3.2.0
```

[40] survival_2.38-3 AnnotationDbi_1.30.1 colorspace_1.2-6
[43] cluster_2.0.3 caTools_1.17.1

Chapter 3 Simulation study

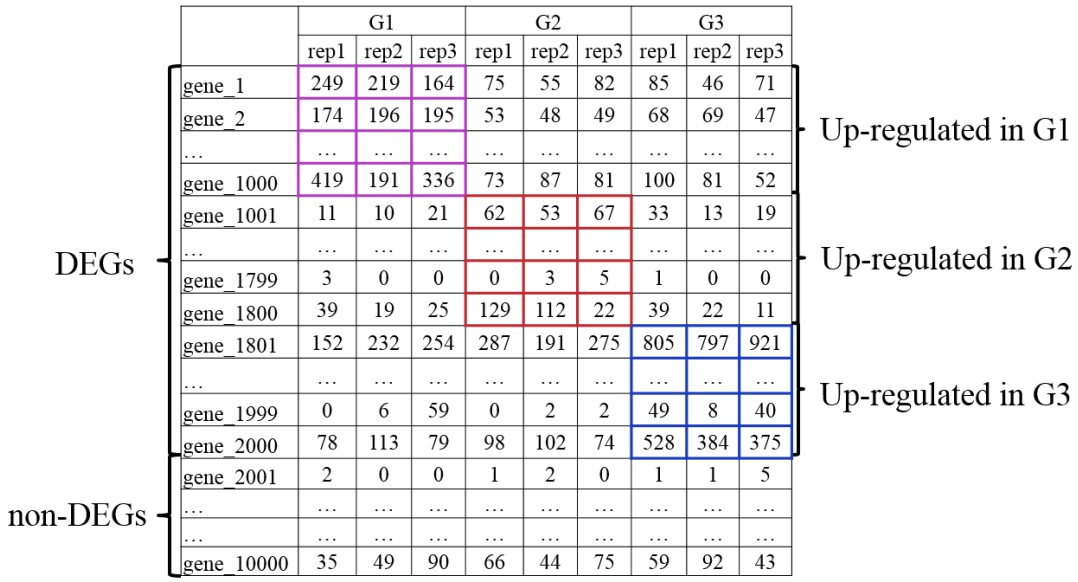
3.1 Generation of simulation data

There are several approaches developed for generating simulation data. However, most of them are completed by several codes, which are deeply hidden in the vignettes of these approaches. To the best of our knowledge, only the compcodeR package proposed by Soneson, C [104] and the TCC package proposed by Sun *et al*. [87] can generate simulation count data with one function very easily. compcodeR provides an interface for users to several popular methods for DE analysis of RNA-seq data and contains functionality for comparing the analysis results from several methods [104]. In particular, it can introduce outliers into the simulation data. However, it is restricted to generate two-group data and an extension will be needed for multi-group data simulation. On the other hand, TCC can generate the multi-group simulation data and can meet our requirements. So, in our study, all of the count data are come out from the TCC data simulation framework [105].

In TCC, the three-group simulation data were produced using the "simulateReadCounts" function. The variance (v) of the negative binomial (NB) distribution can generally be modeled as $v = \mu + \phi\mu^2$. The empirical distribution of read counts for producing the mean (μ) and dispersion (ϕ) parameters of the NB model was obtained from *Arabidopsis* data (three BRs for both the treated and non-treated samples) in [51]. The output of the "simulateReadCounts" function is stored in the TCC class object with information about the simulation conditions and is therefore ready-to-analyze.

Following our previous study [87, 97], we here demonstrate the performance of these pipelines mainly based on the same evaluation metric and simulation framework. In the simulation study, we use the AUC value as a main measure for comparison, which evaluates both sensitivity and specificity of the pipelines simultaneously [76, 104, 106-110]. To perform the multi-group comparison as simply as possible, here we firstly focus on the three-group data (i.e., G1 vs. G2 vs. G3) with equal numbers of BRs (i.e., 1, 3, 6, and 9 replicates per group; Nrep = 1, 3, 6, and 9). The gene ranking was performed on the basis of an ANOVA-like *p*

value or the derivatives, where a low *p* value for a gene indicates a high degree of DE in at least one of the groups compared. The simulation conditions are as follows: the total number of genes is 10,000 (Ngene = 10000), 5 or 25% of the genes are DEGs (P_{DEG} = 5 or 25%), the levels of DE are four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}) are (1/3, 1/3, 1/3), (0.5, 0.3, 0.2), (0.5, 0.4, 0.1), (0.6, 0.2, 0.2), (0.6, 0.3, 0.1), (0.7, 0.2, 0.1), and (0.8, 0.1, 0.1). Figure 9 shows an exact sample for three-group count data. Among the 10000 genes, there are 1000, 800 and 200 up-regulated genes distributed in group 1 (G1), group 2 (G2), and group 3 (G3) separately. The expression levels of up-regulated genes are obviously higher in the designated group than in the other two groups.



	G1			G2			G3		
	rep1	rep2	rep3	rep1	rep2	rep3	rep1	rep2	rep3
gene_1	249	219	164	75	55	82	85	46	71
gene_2	174	196	195	53	48	49	68	69	47
…	…	…	…	…	…	…	…	…	…
gene_1000	419	191	336	73	87	81	100	81	52
gene_1001	11	10	21	62	53	67	33	13	19
…	…	…	…	…	…	…	…	…	…
gene_1799	3	0	0	0	3	5	1	0	0
gene_1800	39	19	25	129	112	22	39	22	11
gene_1801	152	232	254	287	191	275	805	797	921
…	…	…	…	…	…	…	…	…	…
gene_1999	0	6	59	0	2	2	49	8	40
gene_2000	78	113	79	98	102	74	528	384	375
gene_2001	2	0	0	1	2	0	1	1	5
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
gene_10000	35	49	90	66	44	75	59	92	43

Figure 9 - Three-group simulation data with equal number of replicates

The simulation condition is as follows: the total number of genes is 10,000 (Ngene = 10000), the number of replicates is 3 (Nrep = 3), 20% of the genes are DEGs (P_{DEG} = 20%), the level of DE is four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}) are (0.5, 0.4, 0.1) which means that there are 1000, 800 and 200 up-regulated genes in G1, G2 and G3 separately.

3.2 Results from simulation data with replicates

We first assessed the performances of a total of 12 pipelines: three pipelines in TCC (i.e., *EEE-E*, *DDD-D*, and *SSS-S*), edgeR, edgeR_robust, DESeq, DESeq2, voom, SAMseq, PoissonSeq, baySeq, and EBSeq. Table 6 lists the average AUC values of 100 trials between the ranked gene lists and the truth for various simulation conditions with Nrep = 3. Overall, the AUC values for the *EEE-E* pipeline were the highest and similar across the seven different proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}). The edgeR (i.e., the pipeline *E-E*) performed the second best overall. *EEE-E* and edgeR performed comparably under the unbiased proportion of DEGs in individual groups (1/3, 1/3, 1/3). This is quite reasonable because the *EEE-E* can be viewed as an iterative edgeR pipeline: their theoretical performances are the same under the unbiased condition [12]. Similar to the relationship between *EEE-E* and edgeR, the *DDD-D* (or *SSS-S*) can be viewed as an iterative DESeq (or DESeq2) pipeline. As expected, *DDD-D* (or *SSS-S*) consistently outperformed DESeq (or DESeq2) in all simulation conditions except for the unbiased situations.

We observed similar AUC values across the seven different proportions of DEGs for individual pipelines at P_{DEG} = 5% (Table 6a). When a higher amount of DEGs was introduced (i.e., P_{DEG} = 25%; Table 6b), the performances generally worsened in accordance with the increased degrees of biases (i.e., from left to right in Table 6). For example, the AUC values for voom under the unbiased (1/3, 1/3, 1/3) and most biased (0.8, 0.1, 0.1) proportions decreased from 87.08% to 84.56%. We observed relatively poor performances for EBSeq and voom. This is consistent with a previous simulation study on two-group data with a low number of BRs (Nrep = 2) [28]. A possible explanation of these results is that EBSeq was originally developed to detect differential isoforms (not DEGs) [41] and the large body of methodology in voom is for microarray data (not RNA-seq count data) [38]. Our current evaluation focuses on the gene-level RNA-seq count data and does not address the problem of such a detailed resolution of the analysis. SAMseq and PoissonSeq performed

stably across different proportions. This is probably because both methods are non-parametric ones that do not assume any particular distribution for the data and that are generally robust against such biased situation (Table 3). These methods, however, performed poorly overall.

Table 6 - Average AUC values for three-group simulation data with replicates

P_{G1}	33%	50%	50%	60%	60%	70%	80%
P_{G2}	33%	30%	40%	20%	30%	20%	10%
P_{G3}	33%	20%	10%	20%	10%	10%	10%
(a) P_{DEG} = 5%							
EEE-E (TCC)	91.57	**91.50**	**91.50**	**91.43**	**91.42**	**91.45**	**91.46**
DDD-D (TCC)	90.70	90.62	90.64	90.54	90.55	90.59	90.62
SSS-S (TCC)	88.34	88.33	88.30	88.24	88.23	88.21	88.30
E-E (edgeR)	**91.58**	91.48	91.47	91.38	91.37	91.38	91.34
edgeR_robust	90.95	90.86	90.85	90.75	90.74	90.74	90.73
D-D (DESeq)	90.71	90.60	90.60	90.50	90.49	90.50	90.48
S-S (DESeq2)	88.34	88.31	88.26	88.19	88.17	88.11	88.14
voom	87.16	87.01	86.99	86.88	86.91	86.88	86.86
SAMseq	85.04	84.97	84.93	84.83	84.88	84.88	84.91
PoissonSeq	87.31	87.25	87.25	87.19	87.17	87.22	87.23
baySeq	90.24	90.21	90.21	90.22	90.17	90.13	90.07
EBSeq	85.77	85.85	85.78	85.81	85.73	85.71	85.77
(b) P_{DEG} = 25%							
EEE-E (TCC)	91.47	**91.46**	**91.45**	**91.45**	**91.43**	**91.42**	**91.37**
DDD-D (TCC)	90.77	90.73	90.72	90.70	90.68	90.65	90.57
SSS-S (TCC)	88.13	88.11	88.13	88.14	88.12	88.09	88.06
E-E (edgeR)	**91.47**	91.30	91.18	91.06	90.98	90.62	89.97
edgeR_robust	90.89	90.69	90.57	90.43	90.34	89.97	89.27
D-D (DESeq)	90.77	90.54	90.37	90.25	90.15	89.73	89.04
S-S (DESeq2)	88.12	87.83	87.62	87.49	87.36	86.79	85.92
voom	87.08	86.71	86.52	86.29	86.18	85.60	84.56
SAMseq	84.95	84.82	84.82	84.77	84.75	84.72	84.63
PoissonSeq	87.22	87.18	87.14	87.13	87.11	87.06	86.97
baySeq	90.34	90.13	90.07	89.92	89.83	89.52	88.86
EBSeq	85.82	85.61	85.49	85.34	85.30	84.74	84.02

Average AUC values (%) of 100 trials for each simulation condition are shown: (a) P_{DEG} = 5% and (b) P_{DEG} = 25%. Simulation data contain a total of 10,000 genes: P_{DEG} % of genes is for DEGs, P_{G1} % of P_{DEG} in G1 is higher than in the other groups, and each group has three BRs (Nrep = 3). Seven conditions are shown in total. The highest AUC value for each condition is in bold.

The AUC values of all pipelines become larger when the number of BRs per group increases from 3 to 9 (Additional files 1 and Additional file 2), which is consistent to several previous studies. It should be noted that the relative performances for EBSeq tend to better as the number of replicates per group increases. In particular, EBSeq consistently outperformed the others when Nrep = 9 and P_{DEG} = 5% (Additional file 2), suggesting that the DEGES-based pipeline based on EBSeq could produce a more accurate ranked gene list. However, as previously discussed for the DEGES-based pipeline based on baySeq [66], bayesian methods (EBSeq and baySeq) generally require huge computation time (Additional file 3). Accordingly, the implementation of DEGES for EBSeq might be unrealistic.

As shown in Table 3, TCC can perform various combinations for the DEGES-based DE pipeline *X*-(*Y*-*X*)$_n$-*Z* or *XYX*-*Z*, where *Y* and *Z* are the DEG identification methods and *X* is the normalization method. We investigated the effect of the individual methods (used as *X*, *Y*, and *Z*) by analyzing a total of 12 pipelines (eight DEGES-based pipelines and four non-DEGES-based pipelines). Table 7 shows the average AUC values for these pipelines. Note that the values in Table 6 and Table 7 are comparable and that those for four pipelines (*EEE-E*, *DDD-D*, *E-E*, and *D-D*; gray colored in Table 7) are provided in both tables. It is clear that the choice of *Z* has more impact than the choice of *Y* on the gene ranking accuracy and that the use of the DEG identification method provided in edgeR in both *Y* and *Z* can be recommended. In comparison with the two normalization methods in *X* in the eight DEGES-based pipelines, the method in DESeq (denoted as "*D*") gave slightly higher AUC values than the TMM normalization method in edgeR (denoted as "*E*"). However, the superiority of DESeq in *X* was not observed when four non-DEGES-based pipelines *X*-*Z* were compared, where edgeR (i.e., the TMM normalization method) outperformed DESeq. In any case, the different choices in *X* have less impact than the choices in *Y* and *Z*.

It is supersized that the best pipeline was *DED-E*, followed by *EEE-E* and *DDD-E* (Table 7b). The *DED-E* and *DDD-E* pipelines consist of

methods provided by different packages. For example, *DED-E* (DESeq-(edgeR-DESeq)$_3$-edgeR) pipeline, consists of the normalization method in DESeq as *X* and the DEG identification method in edgeR as *Y* and *Z*. These results suggest that in some cases, the suitable choices of the best pipeline may slightly improve the sensitivity and specificity of DE results. We should note that the current simulation data are generated by the "simulateReadCounts" function in TCC. This is simply because, to the best of our knowledge, TCC only provides the R function that can generate multi-group simulation count data. TCC simulates all counts using NB distributions, suggesting that this simulation framework advantageously acts on the classical R packages such as edgeR and DESeq. This is probably the main reason for inferior performances of two recently published packages (edgeR_robust and DESeq2; Table 6); those are the advanced versions for edgeR and DESeq, respectively, and are robust against count outliers such as abnormally high counts (for details, see [94, 96]).

Table 7 - Effect of different choices for the possible pipelines in TCC

P_{G1}	33%	50%	50%	60%	60%	70%	80%
P_{G2}	33%	30%	40%	20%	30%	20%	10%
P_{G3}	33%	20%	10%	20%	10%	10%	10%
(a) P_{DEG} = 5%							
EEE-E	91.57	91.50	91.50	**91.43**	91.42	91.45	91.46
DED-E	91.57	**91.50**	**91.50**	91.43	**91.42**	**91.46**	**91.47**
EDE-E	91.57	91.50	91.50	91.43	91.42	91.45	91.46
DDD-E	91.57	91.50	91.50	91.43	91.42	91.45	91.46
EEE-D	90.70	90.62	90.64	90.54	90.55	90.58	90.62
DED-D	90.71	90.62	90.64	90.54	90.55	90.59	90.62
EDE-D	90.70	90.62	90.64	90.54	90.55	90.58	90.62
DDD-D	90.70	90.62	90.64	90.54	90.55	90.59	90.62
E-E (edgeR)	91.58	91.48	91.47	91.38	91.37	91.38	91.34
D-E	**91.58**	91.48	91.46	91.38	91.36	91.36	91.32
E-D	90.70	90.61	90.61	90.50	90.50	90.51	90.50
D-D (DESeq)	90.71	90.60	90.60	90.50	90.49	90.50	90.48
(b) P_{DEG} = 25%							
EEE-E	91.47	91.46	91.45	91.45	91.43	91.42	91.37
DED-E	91.47	**91.46**	**91.47**	**91.47**	**91.45**	**91.45**	**91.43**
EDE-E	91.47	91.43	91.41	91.40	91.36	91.30	91.19
DDD-E	91.47	91.44	91.43	91.42	91.39	91.36	91.29
EEE-D	90.77	90.74	90.74	90.73	90.71	90.71	90.65
DED-D	90.77	90.74	90.76	90.75	90.73	90.74	90.71
EDE-D	90.77	90.71	90.70	90.68	90.64	90.60	90.47
DDD-D	90.77	90.73	90.72	90.70	90.68	90.65	90.57
E-E (edgeR)	91.47	91.30	91.18	91.06	90.98	90.62	89.97
D-E	**91.48**	91.25	91.08	90.96	90.86	90.44	89.75
E-D	90.77	90.59	90.48	90.35	90.26	89.92	89.25
D-D (DESeq)	90.77	90.54	90.37	90.25	90.15	89.73	89.04

Legends are basically the same as in Table 6. Results of a total of 12 pipelines are shown. The AUC values for four pipelines (*EEE-E*, *DDD-D*, *E-E*, and *D-D*) colored in gray are also shown in Table 6. The *DED-E* pipeline outperforms the others overall.

3.3 Results from simulation data without replicates

Several studies [71, 76, 111] and our above results (Table 6, Additional files 1 and 2) have demonstrated biological replicate size can greatly improve the accuracy of parameter estimation in the statistical parametric model. However, unlike (multi-group) count data with replicates, there are a few packages that can manipulate count data without replicates, such as TCC, edgeR, DESeq, DESeq2 and so on. We here evaluated a total of 20 pipelines (13 DEGES-based pipelines and seven non-DEGES-based pipelines). Table 8 shows the results for simulation data without replicates under P_{DEG} = 25%. When three original non-DEGES-based pipelines *X*-*Z* are compared, *S*-*S* (i.e., DESeq2) performed the best, followed by *D*-*D* and *E*-*E*. This is completely different from Table 6. When 13 DEGES-based pipelines *XYX*-*Z* are compared, the choice of Z for the DEGES-based pipeline has more impact than the choice of *Y* on the gene ranking accuracy (similar to Table 7) and using the DEG identification method provided in DESeq2 (i.e., *S*) can be recommended as *Z*. This result may possibly be explained by the removal of outliers that do not fit the distributional assumptions of the model [89]: DESeq2 [94] implements a functionality of outlier detection and the removal on the basis of Cook's distance [112]. In the situation of count data without replicates, DEGs tend to be flagged as outliers: Cook's distances for DEGs are generally greater than those for non-DEGs. Therefore, the negative effect of 25% DEGs introduced in this simulation framework could be weakened.

In addition to the model construction only with non-outliers in the *Z* step, the DEGES-based normalization in the *XYX* step also slightly but reliably improves ranked gene lists. That is, the AUC values for *SSS*-*S* are higher than those for *S*-*S* (i.e., DESeq2) because the former pipeline is by virtue of that kind of multi-step normalization strategy originally proposed by Kadota *et al*. [97]. However, as also discussed in the TCC paper [87], DESeq and DESeq2 generally estimate FDR more conservatively than the others [74]. Indeed, we observed that the numbers of potential DEGs satisfying 10% FDR in step 2 (i.e., the *Y* step) in the *SSS*-*S* pipeline were nearly zero (i.e., the estimated P_{DEG} values were 0%)

in all simulations, although the actual P_{DEG} values were 25%. This is reasonable because any attempt to work without replicates will lead to very limited reliability [87]. TCC employs a predefined floor P_{DEG} value (= 5%) to obtain certain differences between the DEGES-based approach *SSS-S* and non-DEGES-based approach *S-S*: at least 5% of the top-ranked genes were not used when the normalization factors were calculated at step 3 in *XYX*. As an estimated P_{DEG} value of *x*% tends to work better when simulation data with the same P_{DEG} value is analyzed, the accurate estimation is the next important task.

For simulation results, some people may argue about the area under the ROC curve calculated for the AUC values. Since in reality they care more about the early behavior (left part) of the ROC curve, they think that calculating the whole area under the ROC curve (full AUC) might not be informative and calculating part of the area by controlling (1 – specificity) < 0.1 (partial AUC) is a better choice. Note that, we took both full and partial AUC values into account in this study because the full AUC values have been used widely as an important metric as well as partial AUC values [113] and we added the suggested information in Additional file 4. Similar to the results with full AUC values, we observed the overall superiority for the *EEE-E* pipeline provided in TCC. In conclusion, choosing full or partial AUC values does not have much impact on the results in the current evaluation study.

Table 8 - Average AUC values for three-group simulation data without replicates

P_{G1}	33%	50%	50%	60%	60%	70%	80%
P_{G2}	33%	30%	40%	20%	30%	20%	10%
P_{G3}	33%	20%	10%	20%	10%	10%	10%
EEE-E	77.15	76.88	76.78	76.63	76.88	76.15	75.48
DED-E	77.15	76.86	76.73	76.59	76.86	76.08	75.41
EDE-E	77.15	76.88	76.79	76.64	76.88	76.19	75.57
DDD-E	77.15	76.87	76.75	76.61	76.87	76.13	75.50
EEE-D	81.51	81.14	81.28	80.93	81.14	80.51	79.97
DED-D	81.52	81.14	81.25	80.90	81.14	80.45	79.90
EDE-D	81.49	81.14	81.28	80.94	81.14	80.55	80.05
DDD-D	81.51	81.15	81.26	80.91	81.15	80.49	79.98
E-E (edgeR)	77.15	76.87	76.76	76.60	76.87	76.10	75.36
D-E	77.15	76.86	76.71	76.57	76.86	76.04	75.35
E-D	81.49	81.13	81.27	80.91	81.13	80.46	79.86
D-D (DESeq)	81.53	81.12	81.23	80.88	81.12	80.41	79.84
SSS-S	82.46	82.18	82.08	81.98	82.18	81.52	80.97
EEE-S	**82.46**	82.18	82.08	81.98	82.18	81.50	80.89
DED-S	82.46	82.17	82.04	81.95	82.17	81.43	80.81
EDE-S	82.46	**82.18**	**82.09**	**82.00**	**82.18**	**81.54**	**80.97**
DDD-S	82.46	82.17	82.06	81.97	82.17	81.48	80.90
S-S (DESeq2)	82.46	82.16	82.01	81.92	82.16	81.38	80.73
E-S	82.46	82.17	82.07	81.96	82.17	81.45	80.76
D-S	82.46	82.16	82.02	81.93	82.16	81.39	80.74

Legends are basically the same as in Table 6. Results of a total of 20 pipelines under P_{DEG} = 25% are shown. The *EDE-S* pipeline outperforms the others overall.

3.4 Results from simulation data with other multiple groups

In order to verify the consistent of results from count data with different groups, we also investigate the performance of DE pipelines for four-group comparison. The simulation conditions are as follows: the total number of genes is 10,000 (Ngene = 10000), 5 or 25% of the genes are DEGs (P_{DEG} = 5 or 25%), the levels of DE are four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}, P_{G4}) are (0.25, 0.25, 0.25, 0.25), (0.4, 0.4, 0.1, 0.1), (0.5, 0.3, 0.1, 0.1), (0.5, 0.2, 0.2, 0.1), (0.6, 0.2, 0.1, 0.1), (0.7, 0.1, 0.1, 0.1). Since SAMseq cannot handle several simulation count data because of an unclear error, Table 9 lists the average AUC values of 100 trails for 11 pipelines. Again, the AUC values for the *EEE-E* pipeline were the highest and similar across the six different proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}, P_{G4}). The edgeR (i.e., *E-E*) performed the second best. The other results are almost the same as the results of Table 5.

In addition, we also produced five-group simulation data under the following conditions: the total number of genes is 10,000 (Ngene = 10000), 5 or 25% of the genes are DEGs (P_{DEG} = 5 or 25%), the levels of DE are four-fold in individual groups, and the proportions of DEGs up-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}, P_{G4}) are (0.25, 0.25, 0.25, 0.25), (0.4, 0.4, 0.1, 0.1), (0.5, 0.3, 0.1, 0.1), (0.5, 0.2, 0.2, 0.1), (0.6, 0.2, 0.1, 0.1), and (0.7, 0.1, 0.1, 0.1).

Table 9 - Average AUC values for four-group simulation data with replicates

P_{G1}	25%	40%	50%	50%	60%	70%
P_{G2}	25%	40%	20%	30%	20%	10%
P_{G3}	25%	10%	20%	10%	10%	10%
P_{G4}	25%	10%	10%	10%	10%	10%
(a) P_{DEG} = 5%						
EEE-E(TCC)	**91.46**	**91.42**	**91.38**	**91.38**	**91.39**	**91.49**
DDD-D(TCC)	90.51	90.39	90.39	90.39	90.38	90.49
SSS-S(TCC)	88.87	88.74	88.66	88.66	88.73	88.82
E-E(edgeR)	91.46	91.39	91.36	91.36	91.33	91.41
edgeR_robust	90.88	90.82	90.79	90.79	90.74	90.78
D-D(DESeq)	90.51	90.36	90.35	90.35	90.32	90.40
S-S(DESeq2)	88.87	88.71	88.62	88.62	88.66	88.71
voom	86.95	86.74	86.65	86.65	86.67	86.73
PoissonSeq	86.40	86.40	86.36	86.36	86.27	86.41
baySeq	90.18	90.00	90.01	90.01	90.13	90.14
EBSeq	84.93	84.92	85.06	85.06	85.10	84.90
(b) P_{DEG} = 25%						
EEE-E(TCC)	91.61	**91.61**	**91.57**	**91.61**	**91.62**	**91.57**
DDD-D(TCC)	90.69	90.69	90.63	90.66	90.69	90.60
SSS-S(TCC)	88.89	88.89	88.87	88.91	88.84	88.81
E-E(edgeR)	**91.61**	91.61	91.33	91.31	91.11	90.70
edgeR_robust	90.98	90.98	90.65	90.67	90.47	90.03
D-D(DESeq)	90.69	90.69	90.37	90.34	90.18	89.74
S-S(DESeq2)	88.89	88.89	88.46	88.41	88.02	87.47
voom	86.93	86.93	86.42	86.45	86.03	85.37
PoissonSeq	86.54	86.54	86.39	86.43	86.37	86.25
baySeq	90.41	90.41	90.06	90.06	89.88	89.48
EBSeq	85.27	85.27	84.93	84.76	84.74	84.37

Average AUC values (%) of 100 trials for each simulation condition are shown: (a) P_{DEG} = 5% and (b) P_{DEG} = 25%. Simulation data contain a total of 10,000 genes: P_{DEG} % of genes is for DEGs, P_{G1} % of P_{DEG} in G1 is higher than in the other groups, and each group has three BRs (Nrep = 3). Six conditions are shown in total. The highest AUC value for each condition is in bold.

Table 10 - Average AUC values for five-group simulation data with replicates

P_{G1}	20%	30%	40%	50%	60%
P_{G2}	20%	30%	20%	20%	10%
P_{G3}	20%	20%	20%	10%	10%
P_{G4}	20%	10%	10%	10%	10%
P_{G5}	20%	10%	10%	10%	10%
(a) P_{DEG} = 5%					
EEE-E(TCC)	**91.48**	**91.39**	**91.59**	**91.45**	**91.53**
DDD-D(TCC)	90.23	90.20	90.34	90.20	90.31
SSS-S(TCC)	88.73	88.63	88.82	88.62	88.76
E-E(edgeR)	91.47	91.38	91.57	91.41	91.47
edgeR_robust	90.84	90.75	90.97	90.82	90.80
D-D(DESeq)	90.23	90.19	90.33	90.17	90.26
S-S(DESeq2)	88.75	88.62	88.81	88.57	88.69
voom	86.83	86.65	86.82	86.69	86.57
PoissonSeq	85.71	85.78	85.90	85.80	85.87
baySeq	90.07	89.79	89.96	89.82	89.95
EBSeq	86.64	86.45	86.68	86.50	86.70
(b) P_{DEG} = 25%					
EEE-E(TCC)	**91.55**	**91.61**	**91.52**	**91.57**	**91.56**
DDD-D(TCC)	90.46	90.47	90.40	90.39	90.41
SSS-S(TCC)	88.75	88.79	88.73	88.78	88.73
E-E(edgeR)	91.55	91.51	91.38	91.27	91.03
edgeR_robust	90.91	90.90	90.75	90.59	90.36
D-D(DESeq)	90.46	90.38	90.27	90.12	89.92
S-S(DESeq2)	88.75	88.65	88.50	88.32	87.92
voom	86.82	86.71	86.50	86.20	85.74
PoissonSeq	85.78	85.86	85.76	85.71	85.66
baySeq	90.29	90.10	90.07	89.92	89.64
EBSeq	86.38	86.48	86.20	86.00	85.82

Average AUC values (%) of 100 trials for each simulation condition are shown: (a) P_{DEG} = 5% and (b) P_{DEG} = 25%. Simulation data contain a total of 10,000 genes: P_{DEG} % of genes is for DEGs, P_{G1} % of P_{DEG} in G1 is higher than in the other groups, and each group has three BRs (Nrep = 3). Five conditions are shown in total. The highest AUC value for each condition is in bold.

Chapter 4 Real data study

4.1 Real data with replicates

In addition to the simulation analysis, we also analyzed a real RNA-seq count dataset on three species: humans (HS), chimpanzees (PT), and rhesus macaques (RM) [50]. The original count dataset ("suppTable1.xls") can be downloaded from the supplementary website of [39]. Briefly speaking about the study, Blekhman et al. studied expression levels of liver samples from three males (M1, M2, and M3) and three females (F1, F2, and F3) from each species, giving a total of six different individuals (i.e., six BRs) for each species. Since they performed duplicate experiments for each individual (i.e., two technical replicates), the publicly available raw count matrix consists of 20,689 genes × 36 samples (= 3 species × 2 sexes × 3 BRs × 2 technical replicates). To correctly estimate the biological variation and to make the assumed structure of input data, we summed and collapsed the count data of technical replicates, giving a reduced number of columns in the count matrix (i.e., 18 samples; 3 species × 2 sexes × 3 BRs). The sex (i.e., males or females) was ignored in the three-group comparison of this dataset (i.e., 18 samples; 3 species × 6 BRs). The relationship of sample names between the original and current study can be seen in Figure 10.

We here compared a total of 12 pipelines in light of the overall similarity of ranked gene lists, the number of shared DEGs satisfying an FDR threshold, and so on. To compare these pipelines as simply as possible, we regarded this dataset as single-factor experimental design of three species where each has six BRs (i.e., HS_rep1-6 vs. PT_rep1-6 vs. RM_rep1-6).

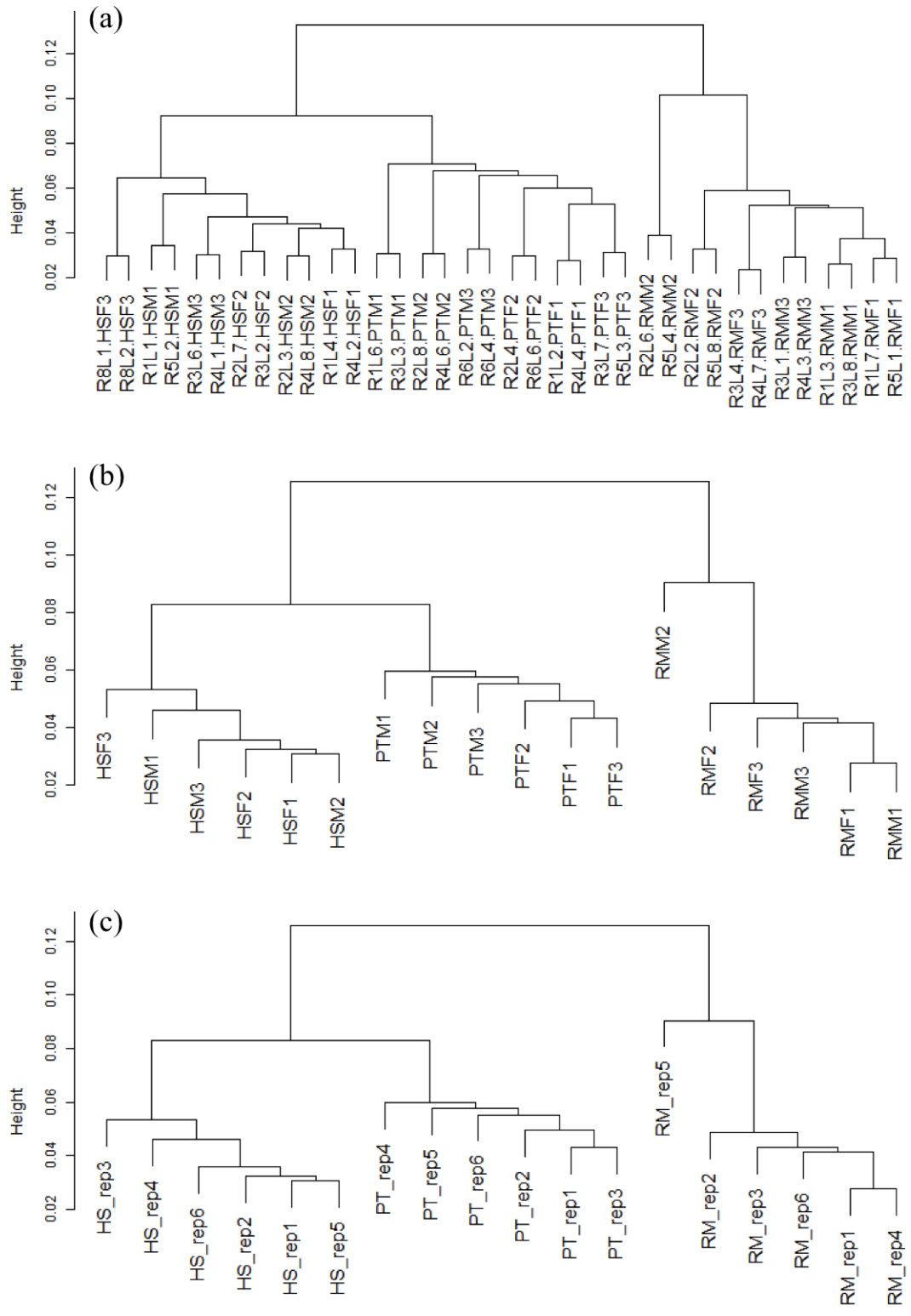


Figure 10 - Dendrogram of average-linkage hierarchical clustering for the Blekhman's count data

Results of sample clustering are shown: (a) a raw count dataset consisting of 36 samples, (b) a collapsed data consisting of 18 samples, and (c) the same data as (b) but with different sample labels. The clustering was performed using the "clusterSample" function with default options provided in TCC.

4.2 Data analysis

Figure 11 shows the dendrogram of average-linkage clustering for the 12 ranked gene lists. Seven pipelines located in the center (from *SSS-S* to *D-D*) show similar ranked gene lists. This is mainly because the seven pipelines from four packages (TCC, edgeR, DESeq, and DESeq2) commonly employ a GLM framework. Indeed, the minimum value of Spearman's correlation coefficients (*r*) among the seven pipelines was 0.9240. It is also noteworthy that ranked gene lists produced from TCC's iterative strategies and the corresponding original non-iterative strategies are particularly similar. For example, the *r* between *EEE-E* from TCC and *E-E* from edgeR was 0.9999, implying that this data is not extremely biased in light of the proportions of DEGs up- and/or down-regulated in individual groups (P_{G1}, P_{G2}, P_{G3}). That is, the proportions of DEGs in this data (P_{G1}, P_{G2}, P_{G3}) are rather closer to (1/3, 1/3, 1/3) than, for example, (0.8, 0.1, 0.1) or (0.0, 0.9, 0.1).

Note that the dendrogram shown in Figure 11 does not necessarily indicate the superiority of the seven GLM-based pipelines over the others such as EBSeq and baySeq. For example, EBSeq employs an empirical Bayesian framework that returns the posterior probabilities for each of the five possible expression patterns (or models) to each gene. We here used the posterior probability obtained from the "non-DEG" pattern as a surrogate estimate for the adjusted *p* values and ranked genes in ascending order of the values. Probably, this is the main reason for EBSeq having lower similarity than the others. We also confirmed this trend with some simulation data. As in Additional file 2, EBSeq showed the highest average AUC values in the simulation condition: P_{DEG} = 5%, (0.5, 0.4, 0.1) for (P_{G1}, P_{G2}, P_{G3}), and Nrep = 9. A typical dendrogram of 12 ranked gene lists obtained from this simulation condition is given in Additional file 9. In this trial, while EBSeq and baySeq formed one of the two major clusters, those AUC values were not at the top two: the ranks for EBSeq and baySeq were the 1st and 6th, respectively. These results indicate that the low similarities of ranked gene lists between Bayesian pipelines (such as EBSeq and baySeq) and the GLM-based pipelines do not matter.

Figure 12 shows the numbers of DEGs obtained from individual pipelines. We found that different pipelines could produce considerably different numbers of DEGs. Indeed, the numbers widely ranged from 3,832 (18.5% of all genes; DESeq) to 9,453 (45.7%; SAMseq). This trend is consistent with that in a previous comparative study [76]. We also compared the overlaps between all pairs of pipelines (Additional file 5). As expected from Figure 11, we observed similar numbers of DEGs between the three DEGES-based pipelines (*EEE-E*, *DDD-D*, and *SSS-S*) and the corresponding non-DEGES-based ones (*E-E*, *D-D*, and *S-S*) (Additional file 10). The Jaccard coefficients, defined as "intersection / union" for two sets of DEGs, for the three pairs (*EEE-E* vs. *E-E*, *DDD-D* vs. *D-D*, and *SSS-S* vs. *S-S*) were top-ranked among a total of 66 possible pairs (Additional file 6). For example, both *EEE-E* in TCC and *E-E* in edgeR reported the same numbers of DEGs (= 7,247). Of these, 7,208 DEGs (99.46%) were common, and the Jaccard coefficient was 7,208 / 7,286 = 0.9893. The overall number of common genes across the twelve sets of DEGs was 2,376 genes. Since individual sets were identified under the 5% FDR threshold, 95% of the 2,376 common DEGs can statistically be regarded as *confident*.

We next classified the expression patterns of the DEGs obtained from the 12 pipelines (Table 11). We here assigned individual DEGs to one of the ten possible patterns defined in baySeq [66]; this package returns one of these patterns to each gene. The *background* information for this data is shown in the "all_genes" row in Table 11. The "common" row indicates the percentages of individual expression patterns for the 2,376 common DEGs. The remaining rows (from *EEE-E* to EBSeq) show the distributions for each of the pipelines. It is reasonable that no DEGs identified by individual pipelines are assigned as a flat expression pattern (i.e., G1=G2=G3) for the HS (G1) vs. PT (G2) vs. RM (G3) comparison. We found that most DEGs were assigned preferably to one of four patterns (G1>G2>G3, G2>G1>G3, G3>G1>G2, and G3>G2>G1) and unpreferably to one of two patterns (G1>G3>G2 and G2>G3>G1). That is, up- (or down-) regulation in G1 for DEGs tends to coincide with G2 more than G3. This can also be seen in the results from sample clustering for raw count

data (Figure 10), implying that we can roughly predict the DE results such as those shown in Table 11 from the overall similarities of samples on raw count data.

When comparing the distributions of patterns for DEGs between pipelines, we saw high similarities overall. If anything, baySeq showed a distribution relatively different from the others in light of the higher percentages for three patterns (G1>G2=G3, G2>G1=G3, and G3>G1=G2). This kind of classification can also be performed using EBSeq [90]. EBSeq defines a total of five possible patterns when comparing three groups: Pattern 1 for non-DEG (i.e., G1=G2=G3), Pattern 2 for DE in G3 (G1=G2<G3 and G1=G2>G3), Pattern 3 for DE in G2 (G2>G1=G3 and G2<G1=G3), Pattern4 for DE in G1 (G1>G2=G3 and G1<G2=G3), and Pattern 5 for DE among all groups. Similar to baySeq, EBSeq also returns one of these patterns to each gene. The results of classification based on EBSeq are given in Additional file 7. Similar to the results from baySeq (Table 11), we observed that nearly half the DEGs were assigned to Pattern 2, where the expression patterns between G1 and G2 tend to be more similar than for G3. We also observed that the distribution for baySeq is relatively different from the others, e.g., lower percentages in Patterns 3 and 4 and a higher percentage in Pattern 5.

We next assessed the reproducibility of ranked gene lists. Remember that the real dataset we analyzed here consists of three groups, each of which has six BRs (we denote this dataset as "rep1-6"). In addition to the original three-group comparison with six replicates (i.e., HS_rep1-6 vs. PT_rep1-6 vs. RM_rep1-6), we also performed three three-group comparisons by dividing the original dataset into three; individual subsets consist of two BRs for each group. For example, the first subset (say "rep1-2") consists of a total of six samples for comparing HS_rep1-2, PT_rep1-2, and RM_rep1-2. Likewise, the third subset ("rep5-6") is for comparing "HS_rep5-6 vs. PT_rep5-6 vs. RM_rep5-6." After performing the DE analysis for the three subsets (i.e., rep1-2, rep3-4, and rep5-6), we obtained three ranked gene lists for these subsets. Accordingly, there are a total of four ranked gene lists (rep1-2, rep3-4, rep5-6, and rep1-6) for each pipeline. We evaluated the reproducibility of ranked gene lists (i) for

each subset to the original dataset (i.e., rep1-6 vs. rep1-2, rep1-6 vs. rep3-4, and rep1-6 vs. rep5-6) and (ii) among the three subsets (i.e., rep1-2 vs. rep3-4 vs. rep5-6).

Figure 13 shows the numbers of common genes between the compared sets of top-ranked genes for individual pipelines: (a) for the top 100 and (b) for the top 1,000. For example, there were 66 common genes when comparing the two sets (rep1-6 and rep5-6) of the 100 top-ranked genes obtained from the *EEE-E* pipeline (see the leftmost blue bar in Figure 13a). As shown in Table 6 and Additional files 1 and 2, the more BRs we use, and the more accurate the ranked gene lists we can obtain. Accordingly, the evaluation based on the reproducibility of ranked gene lists is analogous to a performance comparison when the available count data has only two BRs. Overall, we see high reproducibility for three edgeR-related pipelines (*EEE-E*, *E-E*, and edgeR_robust) and low reproducibility for two pipelines (SAMseq and EBSeq). This trend is consistent with the simulation results shown in Table 6 (i.e., three-group data with three BRs) and previous simulation results for two-group data with two BRs [76]. Although PoissonSeq showed the highest reproducibility when the 1,000 top-ranked genes were evaluated (Figure 13b), the performance seems unstable, especially on < 200 top-ranked genes. This is mainly due to low reproducibility of the ranked gene list for rep1-2 to the list for rep1-6. Although we saw a plausible outlying sample (RMM2 or RM_rep5) in the dendrogram of sample clustering for the raw count data, it would not be related to the dissimilarity of ranked gene lists between rep1-2 and rep1-6. The percentages of overlapping/common genes (POGs) for any numbers of top-ranked genes are given in Additional files 11, 12 and 13.

For biologists, we also list the top 20 genes with the biggest significance for DE from the analysis results of 12 pipelines, which can be validated by RT-PCR in a small scale (Additional file 8).

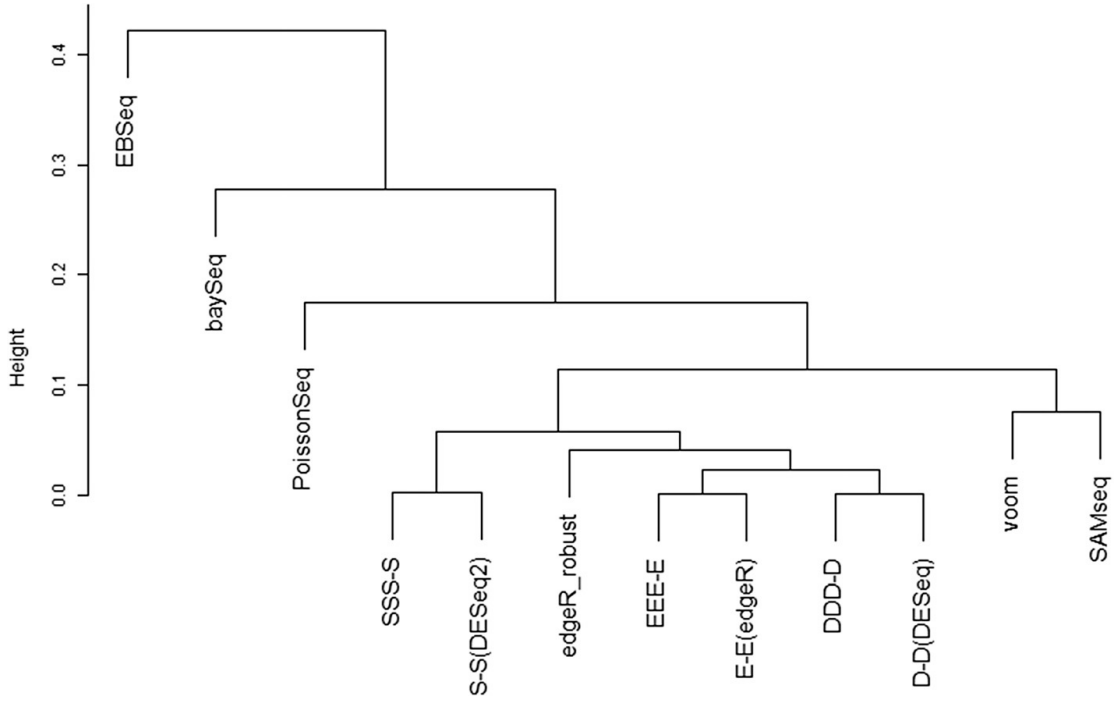


Figure 11 - Overall similarity of 12 ranked gene lists applied for Blekhman's count data

The dendrogram of average-linkage clustering is shown. Spearman's rank correlation coefficient (*r*) is used as a similarity metric; left-hand scale represents (1 – *r*).

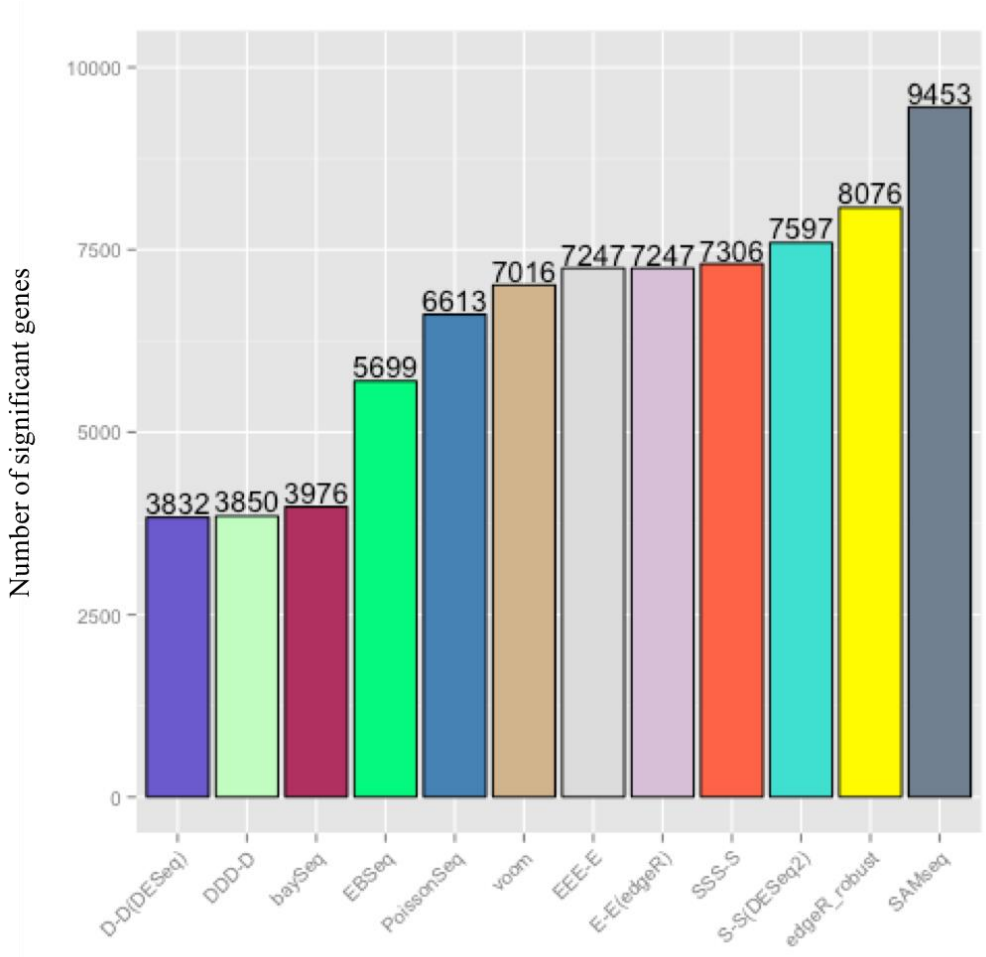


Figure 12 - Number of genes found to be significantly DE among the three species in the Blekhman's count data

Table 11 - Classification of expression patterns for DEGs

	G1=G2=G3	G1>G2=G3	G1>G2>G3	G1>G3>G2	G2>G1=G3	G2>G1>G3	G2>G3>G1	G3>G1=G2	G3>G1>G2	G3>G2>G1	Total
all_genes	13.5	2.2	15.1	8.7	2.3	15.9	9.4	2.9	15.1	14.8	20689
common	0.0	0.1	23.2	5.8	0.2	26.4	5.7	0.7	18.6	19.2	2376
EEE-E	0.0	0.6	20.7	7.4	0.7	21.9	8.1	1.6	19.9	19.2	7247
DDD-D	0.0	0.4	25.0	7.3	0.6	25.0	6.0	1.4	17.3	17.1	3850
SSS-S	0.0	0.2	19.3	7.1	0.3	21.7	9.4	0.9	19.9	21.2	7295
E-E (edgeR)	0.0	0.6	20.4	7.3	0.7	22.1	8.3	1.6	19.7	19.3	7247
edgeR_robust	0.0	0.3	20.6	8.4	0.5	22.0	8.8	1.2	19.1	18.9	8076
D-D (DESeq)	0.0	0.4	24.3	7.2	0.6	24.2	6.0	1.4	17.8	18.1	3832
S-S (DESeq2)	0.0	0.2	20.4	8.0	0.3	21.8	8.9	0.8	19.7	19.9	7585
voom	0.0	0.7	21.3	7.7	0.7	22.5	8.2	1.3	18.7	19.0	7016
SAMseq	0.0	0.2	20.9	9.7	0.3	21.8	9.2	0.8	18.9	18.3	9453
PoissonSeq	0.0	0.0	19.5	8.9	0.1	22.2	9.4	0.3	20.3	19.3	6613
baySeq	0.0	0.8	21.0	5.5	1.3	23.7	6.3	2.8	19.0	19.6	3975
EBSeq	0.0	0.0	21.0	7.0	0.1	23.7	7.1	0.3	20.8	19.9	5699

Percentages of genes assigned to each of the ten possible patterns defined as baySeq. Numbers in the "Total" column indicate the numbers of genes. For example, 13.5% of 20,689 genes in baySeq are assigned as "G1=G2=G3".

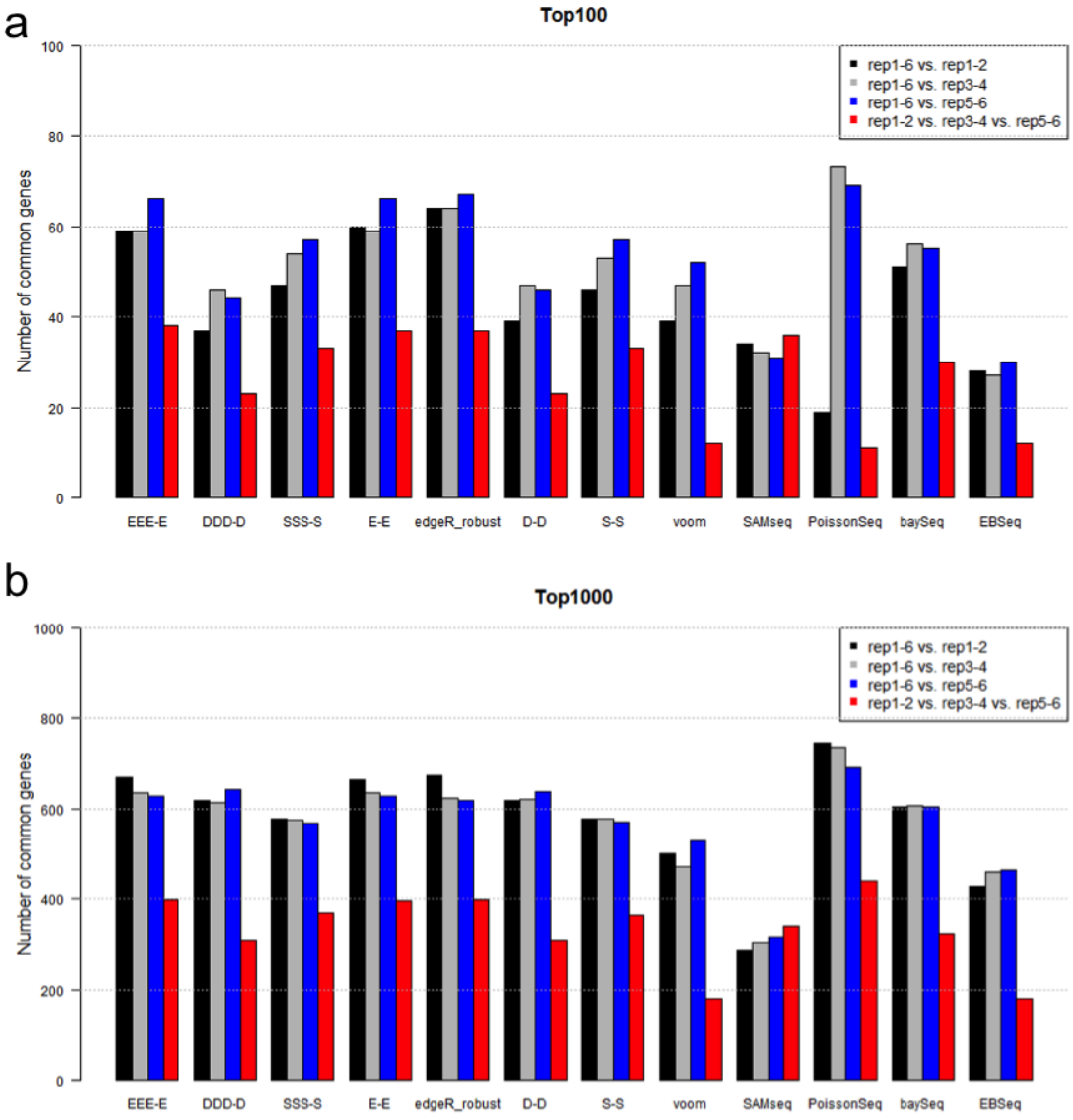


Figure 13 - Reproducibility between ranked gene lists

Numbers of common genes between top-ranked genes for individual pipelines are shown: (a) results for 100 top-ranked gene lists and (b) results for 1000 top-ranked gene lists. Bars in black (rep1-6 vs. rep1-2), gray (rep1-6 vs. rep3-4), and blue (rep1-6 vs. rep5-6) in Figure 13a indicate the numbers of common genes between the two sets of 100 top-ranked genes obtained from the individual pipelines. For example, the gray bar (rep1-6 vs. rep3-4) for *DDD-D* in Figure 13a indicates that there were 46 common genes when the 100 top-ranked genes from the dataset rep1-6 are compared with the 100 top-ranked genes from the dataset rep3-4. Analogously, bars in red (rep1-2 vs. rep3-4 vs. rep5-6) in Figure 14b indicate the numbers of common genes between the three sets of 1000 top-ranked genes for the three datasets (rep1-2, rep3-4, and rep5-6). For example, the red bar for *EEE-E* in Figure 13b indicates that there were 397 common genes (39.7 % of overlapping genes) when the three sets of gene lists (each of which contains 1000 top-ranked genes) obtained from the pipeline *EEE-E* for the three datasets were compared.

Chapter 5 Conclusion

In this study, we evaluated 12 pipelines for DE analysis of multi-group RNA-seq count data. Arguably, this experimental design has been performed well in practice. To our knowledge, this evaluation is the first comprehensive study on multi-group count data. Our main findings can be summarized as follows:

First, the idea of DEGES implemented in TCC can be applied to multi-group data. We confirmed that the AUC values for the three DEGES-based pipelines (*EEE-E*, *DDD-D*, and *SSS-S*) were higher overall than the corresponding non-DEGES-based pipelines, *E-E* (edgeR), *D-D* (DESeq), and *S-S* (DESeq2), respectively (Table 6).

Second, the choice of DEG identification method *Z* in the DEGES-based pipeline *XYX-Z* is critical for obtaining good DE results. For *Z* in the pipeline *XYX-Z*, using *E* (the DEG identification method provided in edgeR; Table 6) and *S* (provided in DESeq2; Table 8) when analyzing three-group data with and without replicates, respectively, gave higher AUC values than the others.

Third, to analyze three-group data with replicates, we obtained the results that either *DED-E* or *EEE-E* outperforms the others. We recommend these pipelines for this case (Tables 6 and 7). Both pipelines can easily be performed by using the TCC package. While *DED-E* showed the highest AUC values under the interrogated pipelines and simulation conditions, the difference between *DED-E* and the second best pipeline *EEE-E* can practically be negligible. Since *EEE-E* is the natural extension of a DEGES-based pipeline for edgeR, using *EEE-E* would be the best practice. However, note that two Bayesian pipelines (baySeq and EBSeq) perform comparably to or better than the GLM-based pipelines (edgeR, DESeq, and DESeq2) when a number of replicates are available (Additional file 1 and Additional file 2). In particular, EBSeq consistently outperformed *EEE-E* under some simulation conditions (Nrep = 9 and P_{DEG} = 5%; Additional file 2), suggesting that the DEGES-based pipeline based on EBSeq could produce a more accurate ranked gene list. Although these Bayesian pipelines tend to come at the cost of a huge computation time, their implementation and evaluation are the next important tasks.

Fourth, to analyze three-group data without replicates, we obtained the results that eight *EDE-S* or *SSS-S* outperforms the others (Table 8). Both pipelines can easily be performed by using the TCC package. While *EDE-S* showed the highest AUC values under the interrogated pipelines and simulation conditions, the difference between *EDE-S* and the second best pipeline *SSS-S* can practically be negligible. Since *SSS-S* is the natural extension of a DEGES-based pipeline for DESeq2, using *SSS-S* would be the best practice. Note that our previous recommendation for analyzing two-group data without replicates was to use *DDD-D* and that this conclusion was obtained only by evaluating a total of eight competing pipelines (*D-D*, *DDD-D*, *EDE-D*, *EbE-D*, *D-b*, *DDD-b*, *EDE-b*, and *EbE-b*, where "*b*" denotes baySeq). We expect the DESeq2-related pipelines (i.e., *EDE-S* and *SSS-S*) would be recommended for analyzing two-group data without replicates as an updated guideline. The comprehensive evaluation should, of course, be performed as a next task.

Fifth, the results of DE analysis (including existence or non-existence of DEGs) can roughly be estimated by the hierarchical dendrogram of sample clustering for the raw count data (Table 11; Figure 10; Additional files 5-7). The dendrogram of sample clustering shows some useful information about the DE results. The real count data we used here has 18.5% - 45.7% of DEGs at the 5% FDR threshold (Additional file 5). In our experience, such results (i.e., existence of large numbers of DEGs) have frequently been obtained when individual groups (G1, G2, and G3) form distinct sub-clusters where each sub-cluster consists only of members in each group (Figure 10). In other words, if members in each sub-cluster originate from plural groups, no or few DEGs would be obtained as the DE result for such indistinct data. Of course, it is critical to employ appropriate choices for distance metric and filtering of low count data for obtaining a robust dendrogram. While we employed the default options ("1 – Spearman correlation coefficient" as a distance and the use of *unique* expression patterns as an objective filtering) in the clustering function ("clusterSample") provided in TCC, further evaluation should also be performed.

Chapter 6 Future prospect

In modern biological study, a variety of "omics" have come forth, transcriptomics has become the indispensable way for researchers to understand vast quantities of bioinformation. Researchers have developed a number of methodologies in order to study one specific transcriptome profiling deeply and comprehensively. Methods based on hybridization (i.e., microarray) and NGS (i.e., RNA-seq) are two main methodologies. Depending on the price advantage and the ongoing technical improvement, microarrays will be still the mainstream products of the transcriptomics market in the short future. With the decreasing price and clear advantages over microarray, RNA-seq has been becoming the technology of choice and will replace microarray totally sooner or later in the market.

Based on its high accuracy and data yield, RNA-seq is an easier way to interpret the functional elements of the genome and reveal the molecular constituents of the cells and tissues. Therefore, it is now being adopted for clinic use (i.e., clinical diagnostic, personalized medicine) [36], especially for the studies relative to cancer and other disease [114]. For example, RNA-seq can facilitate the development of cancer database for gene mutations. Based on the DE information of mutant gene, the associated biomolecule or/and molecular complexes can be easily detected. Then the specific medicines for these targets can be designed. As a result, the medicine screening process for cancers will be more effective. Moreover, because of the existing of tissue-specificity and person-specificity, RNA-seq has become an important means in the project of “person medicine” or “accurate medicine”.

To date, NGS requires orders of magnitude more starting materials than those are found in an individual cell. However, in some cases such as circulating tumor cells (CTCs), stem cells and other rare populations, sufficient material cannot be extracted for downstream analysis on such a scale. Moreover, handing such small quantities mean that sample loss, degradation and contamination can have a pronounced effect on sequence robust and quality. Moreover, the materials in RNA-seq are based on a large population of cells, in which the relative proportions of differentially expressed transcripts in a transcriptome show highly variable [115]. Heavy amplification in large collaborative projects also

propagates errors and biases, which can lead to uneven coverage, unknown nuisance technical effects and inaccurate quantification. Therefore, there is much extensive space for NGS amplification, such as single-cell sequencing-based technologies. With the rapid progress in sequencing technologies, single-cell sequencing has been preliminary applied for some studies in which the capability of traditional NGS is weak.

On the other hand, in the process of RNA-seq data analysis, although a number of methods have been proposed, there are no well-approved methods for data normalization or DEG identification in the RNA-seq community. Further, most of the methods were developed for single-factor or two-group experiments. Since there will be more and more multi-group or multi-factored experiments, many efforts for the complexity studies like our work should be done.

Additional files

Additional file 1 - Average AUC values for simulation data with 6 BRs

P_{G1}	33%	50%	50%	60%	60%	70%	80%
P_{G2}	33%	30%	40%	20%	30%	20%	10%
P_{G3}	33%	20%	10%	20%	10%	10%	10%
P_{DEG} = 5%							
EEE-E (TCC)	**94.94**	**95.00**	**94.91**	**95.01**	**95.00**	**94.98**	**94.93**
DDD-D (TCC)	94.65	94.68	94.63	94.73	94.72	94.69	94.65
SSS-S (TCC)	93.35	93.42	93.31	93.43	93.46	93.40	93.36
E-E (edgeR)	94.94	94.98	94.88	94.98	94.95	94.90	94.81
edgeR_robust	94.18	94.21	94.14	94.21	94.23	94.16	94.13
D-D (DESeq)	94.65	94.66	94.59	94.69	94.67	94.60	94.52
S-S (DESeq2)	93.35	93.39	93.26	93.37	93.39	93.27	93.16
voom	91.36	91.36	91.29	91.34	91.36	91.25	91.15
SAMseq	90.66	90.75	90.67	90.75	90.75	90.73	90.75
PoissonSeq	91.73	91.75	91.71	91.80	91.76	91.73	91.71
baySeq	94.40	94.43	94.41	94.45	94.44	94.32	94.34
EBSeq	93.91	93.90	93.97	94.03	94.03	94.04	94.00
P_{DEG} = 25%							
EEE-E (TCC)	94.94	**94.96**	**94.94**	**94.95**	**94.93**	**94.91**	**94.88**
DDD-D (TCC)	94.75	94.79	94.77	94.77	94.74	94.74	94.72
SSS-S (TCC)	93.31	93.34	93.30	93.32	93.28	93.29	93.25
E-E (edgeR)	**94.94**	94.78	94.63	94.51	94.41	93.97	93.21
edgeR_robust	94.22	94.05	93.92	93.76	93.65	93.21	92.47
D-D (DESeq)	94.75	94.58	94.38	94.26	94.14	93.70	92.98
S-S (DESeq2)	93.31	92.94	92.57	92.38	92.18	91.44	90.24
voom	91.33	90.98	90.68	90.41	90.21	89.39	88.08
SAMseq	90.67	90.63	90.57	90.55	90.49	90.42	90.33
PoissonSeq	91.67	91.67	91.64	91.61	91.58	91.53	91.41
baySeq	94.56	94.43	94.29	94.20	94.11	93.80	93.22
EBSeq	93.90	93.73	93.61	93.40	93.32	92.74	91.84

Results are shown for a total of 12 pipelines for three-group simulation data, where each group has six (Nrep = 6) BRs. Legends are the same as in Table 6.

Additional file 2 - Average AUC values for simulation data with 9 BRs

P_{G1}	33%	50%	50%	60%	60%	70%	80%
P_{G2}	33%	30%	40%	20%	30%	20%	10%
P_{G3}	33%	20%	10%	20%	10%	10%	10%
P_{DEG} = 5%							
EEE-E (TCC)	96.65	96.62	96.67	96.73	96.63	96.69	96.68
DDD-D (TCC)	96.49	96.48	96.52	96.56	96.47	96.54	96.52
SSS-S (TCC)	95.46	95.48	95.45	95.55	95.43	95.51	95.50
E-E (edgeR)	96.65	96.61	96.64	96.69	96.59	96.62	96.57
edgeR_robust	95.79	95.81	95.79	95.86	95.77	95.81	95.79
D-D (DESeq)	96.49	96.46	96.49	96.52	96.43	96.46	96.40
S-S (DESeq2)	95.46	95.45	95.40	95.49	95.36	95.38	95.32
voom	93.71	93.70	93.67	93.74	93.63	93.60	93.51
SAMseq	93.20	93.26	93.25	93.34	93.26	93.30	93.28
PoissonSeq	93.92	93.92	93.96	94.01	93.92	93.96	93.94
baySeq	96.33	96.33	96.29	96.36	96.28	96.29	96.22
EBSeq	**96.83**	**96.89**	**96.90**	**96.97**	**96.93**	**96.96**	**96.91**
P_{DEG} = 25%							
EEE-E (TCC)	96.72	**96.73**	**96.74**	**96.72**	**96.74**	**96.72**	**96.70**
DDD-D (TCC)	96.64	96.65	96.66	96.63	96.65	96.64	96.63
SSS-S (TCC)	95.48	95.46	95.50	95.46	95.48	95.48	95.44
E-E (edgeR)	96.72	96.54	96.41	96.27	96.19	95.73	94.88
edgeR_robust	95.91	95.70	95.55	95.43	95.32	94.86	94.00
D-D (DESeq)	96.64	96.44	96.28	96.14	96.07	95.64	94.91
S-S (DESeq2)	95.48	95.03	94.70	94.46	94.30	93.49	92.17
voom	93.79	93.35	93.05	92.76	92.56	91.61	90.00
SAMseq	93.30	93.22	93.21	93.23	93.17	93.11	93.00
PoissonSeq	93.95	93.94	93.92	93.89	93.89	93.81	93.70
baySeq	96.46	96.33	96.25	96.16	96.11	95.82	95.18
EBSeq	**96.86**	96.70	96.61	96.38	96.38	95.85	94.85

Results are shown for a total of 12 pipelines for three-group simulation data, where each group has nine (Nrep = 9) BRs. Legends are the same as in Table 6.

Additional file 3 - Average computation times (in seconds) of 20 trials.

	Nrep = 3		Nrep = 6		Nrep = 9	
P_{G1}	33%	80%	33%	80%	33%	80%
P_{G2}	33%	10%	33%	10%	33%	10%
P_{G3}	33%	10%	33%	10%	33%	10%
P_{DEG} = 5%						
EEE-E (TCC)	18.3	17.7	29.8	29.6	41.6	41.3
DDD-D (TCC)	223.4	219	212.5	212.7	218.9	217.2
SSS-S (TCC)	29.7	29.2	42.2	42.2	59.2	58.4
E-E (edgeR)	4.5	4.4	7.2	7.2	10.2	10
edgeR_robust	22.4	22	35.2	35.1	48.6	48
D-D (DESeq)	55.3	54.5	52.9	52.9	54.3	53.9
S-S (DESeq2)	7.3	7.1	10.3	10.3	14.4	14.4
voom	1.6	1.6	1.7	1.7	2	2
SAMseq	14.3	13.7	19.7	19.6	24.9	24.9
PoissonSeq	5.5	5.4	5.1	5.1	5.3	5.3
baySeq	1008.6	988.5	1554.2	1557.2	2128.6	2125.5
EBSeq	456.3	455.8	797.2	853.6	1110.5	1219.4

Average computational times of 20 trials cost by 12 pipelines are shown. The simulations are under six different settings.

Additional file 4 - Average partial AUC values of 20 trails with (1 – specificity) < 0.1

	Nrep = 3		Nrep = 6		Nrep = 9	
P_{G1}	33%	80%	33%	80%	33%	80%
P_{G2}	33%	10%	33%	10%	33%	10%
P_{G3}	33%	10%	33%	10%	33%	10%
P_{DEG} = 5%						
EEE-E (TCC)	**6.94**	**6.93**	**8.11**	**8.12**	8.67	8.65
DDD-D (TCC)	6.62	6.61	7.84	7.87	8.41	8.43
SSS-S (TCC)	6.72	6.69	7.87	7.89	8.44	8.43
E-E (edgeR)	6.94	6.89	8.11	8.09	8.67	8.62
edgeR_robust	6.82	6.79	7.91	7.93	8.44	8.41
D-D (DESeq)	6.62	6.57	7.84	7.84	8.41	8.4
S-S (DESeq2)	6.73	6.65	7.87	7.84	8.44	8.38
voom	6.27	6.21	7.38	7.4	7.95	7.88
SAMseq	5.82	5.83	7.31	7.36	7.88	7.89
PoissonSeq	5.6	5.6	6.6	6.6	7.16	7.16
baySeq	6.53	6.47	7.84	7.83	8.44	8.39
EBSeq	6.12	6.09	7.92	7.94	**8.68**	**8.71**

Results are shown for a total of 12 pipelines for three-group simulation data under three different BRs (i.e., 3, 6, 9) with PDEG = 5%. Legends are the same as in Table 6.

Additional file 5 - Comparison of DEGs obtained from individual pipelines for the Blekhman's count data

	EEE-E	*DDD-D*	*SSS-S*	*E-E* (edgeR)	edgeR _robust	*D-D* (DESeq)	*S-S* (DESeq2)	voom	SAMseq	PoissonSeq	baySeq	EBSeq
EEE-E	**7247**	3843	6639	7208	6968	3825	6785	6360	6886	5046	3934	5255
DDD-D	3843	**3850**	3815	3843	3849	3778	3828	3776	3840	2980	3121	3330
SSS-S	6639	3815	**7295**	6645	6886	3801	7221	6469	7162	5153	3746	5424
E-E (edgeR)	7208	3843	6645	**7247**	6959	3825	6782	6356	6880	5036	3934	5247
edgeR _robust	6968	3849	6886	6959	**8076**	3831	7111	6784	7737	5366	3864	5490
D-D (DESeq)	3825	3778	3801	3825	3831	**3832**	3810	3759	3822	2968	3123	3311
S-S (DESeq2)	6785	3828	7221	6782	7111	3810	**7585**	6591	7425	5298	3748	5521
voom	6360	3776	6469	6356	6784	3759	6591	**7016**	6887	4809	3768	5271
SAMseq	6886	3840	7162	6880	7737	3822	7425	6887	**9453**	5897	3815	5583
PoissonSeq	5046	2980	5153	5036	5366	2968	5298	4809	5897	**6613**	3064	4374
baySeq	3934	3121	3746	3934	3864	3123	3748	3768	3815	3064	**3975**	3213
EBSeq	5255	3330	5424	5247	5490	3311	5521	5271	5583	4374	3213	**5699**

Numbers of DEGs satisfying the 5% FDR threshold and the overlaps between all pairs of pipelines are shown. The presentation method is the same as in table 1 in [76]: the numbers on the diagonal are highlighted in bold.

Additional file 6 – Jaccard coefficients from the comparison of DEGs obtained from individual pipelines for the Blekhman's count data

	EEE-E	*DDD-D*	*SSS-S*	*E-E* (edgeR)	edgeR _robust	*D-D* (DESeq)	*S-S* (DESeq2)	voom	SAMseq	PoissonSeq	baySeq	EBSeq
EEE-E	**1**	0.5298	0.84	0.9893	0.834	0.5273	0.8432	0.805	0.7017	0.5725	0.54	0.683
DDD-D	0.53	**1**	0.521	0.5298	0.4765	0.9677	0.5032	0.533	0.4058	0.3982	0.664	0.536
SSS-S	0.84	0.5205	**1**	0.8415	0.8115	0.5188	0.9428	0.825	0.7471	0.5886	0.498	0.717
E-E (edgeR)	0.989	0.5298	0.842	**1**	0.832	0.5273	0.8425	0.804	0.7006	0.5707	0.54	0.682
edgeR _robust	0.834	0.4765	0.812	0.832	**1**	0.4743	0.8317	0.817	0.7901	0.5756	0.472	0.663
D-D (DESeq)	0.527	0.9677	0.519	0.5273	0.4743	**1**	0.5009	0.53	0.4039	0.397	0.667	0.532
S-S (DESeq2)	0.843	0.5032	0.943	0.8425	0.8317	0.5009	**1**	0.823	0.7724	0.5953	0.48	0.711
voom	0.805	0.5326	0.825	0.8038	0.8166	0.5303	0.8228	**1**	0.7187	0.5452	0.522	0.708
SAMseq	0.702	0.4058	0.747	0.7006	0.7901	0.4039	0.7724	0.719	**1**	0.5799	0.397	0.583
PoissonSeq	0.573	0.3982	0.589	0.5707	0.5756	0.397	0.5953	0.545	0.5799	**1**	0.407	0.551
baySeq	0.54	0.6635	0.498	0.5398	0.472	0.6667	0.4798	0.522	0.3969	0.4072	**1**	0.497
EBSeq	0.683	0.5355	0.717	0.6815	0.6626	0.5323	0.7112	0.708	0.5834	0.551	0.497	**1**

Numbers of DEGs satisfying the 5% FDR threshold and the overlaps between all pairs of pipelines are shown. The presentation method is the same as in table 1 in [76]: the numbers on the diagonal are highlighted in bold.

Additional file 7 - Classification of expression patterns for DEGs (based on EBSeq)

	Pattern1	Pattern2	Pattern3	Pattern4	Pattern5	Total
all_genes	46.17%	36.73%	7.30%	6.42%	3.38%	20689
common	0.00%	54.04%	11.53%	10.86%	23.57%	2376
EEE-E	15.23%	46.90%	14.88%	13.37%	9.62%	7247
DDD-D	9.84%	49.17%	12.65%	12.55%	15.79%	3850
SSS-S	11.21%	48.54%	16.08%	14.74%	9.43%	7295
E-E(edgeR)	15.33%	46.87%	14.94%	13.23%	9.62%	7247
edgeR_robust	15.96%	46.24%	15.39%	13.82%	8.59%	8076
D-D(DESeq)	9.97%	49.16%	12.55%	12.37%	15.94%	3832
S-S(DESeq2)	11.52%	48.27%	16.22%	14.92%	9.07%	7585
voom	13.61%	47.25%	15.34%	14.07%	9.73%	7016
SAMseq	19.72%	44.47%	14.96%	13.50%	7.35%	9453
PoissonSeq	18.10%	45.61%	14.06%	12.19%	10.04%	6613
baySeq	16.08%	46.84%	10.92%	9.84%	16.33%	3975
EBSeq	0.00%	56.97%	16.42%	14.34%	12.27%	5699

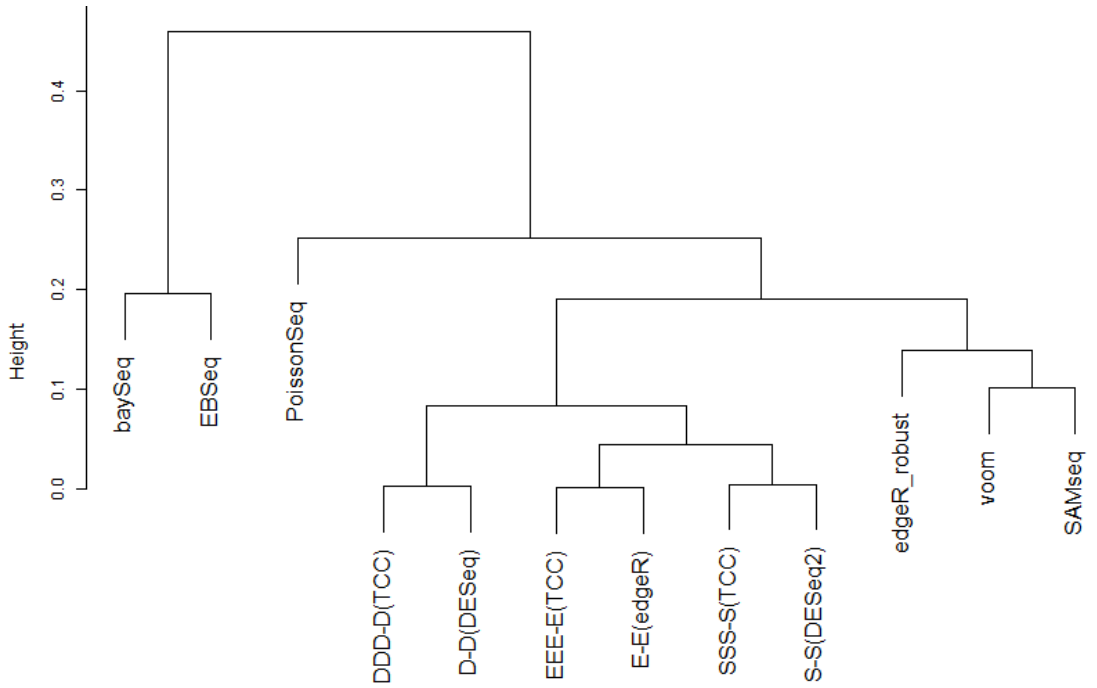
EBSeq defines a total of five possible patterns (Patterns 1 ~ 5). DEGs (satisfying 5% FDR threshold) identified by individual pipelines were assigned to one of the five possible patterns.

Additional file 8 - The top 20 DEGs detected by the 12 pipelines

***E-E*(edgeR)**	***EEE-E***	***D-D* (DESeq)**	***DDD-D***	***S-S*(DESeq2)**	***SSS-S***
ENSG00000000003	ENSG00000000003	ENSG00000000971	ENSG00000000971	ENSG00000000003	ENSG00000000003
ENSG00000000457	ENSG00000000457	ENSG00000001461	ENSG00000001461	ENSG00000000457	ENSG00000000457
ENSG00000000460	ENSG00000000460	ENSG00000001561	ENSG00000001561	ENSG00000000460	ENSG00000000460
ENSG00000000971	ENSG00000000971	ENSG00000001617	ENSG00000001617	ENSG00000000971	ENSG00000000971
ENSG00000001084	ENSG00000001084	ENSG00000001630	ENSG00000001630	ENSG00000001036	ENSG00000001084
ENSG00000001461	ENSG00000001461	ENSG00000002330	ENSG00000002330	ENSG00000001084	ENSG00000001167
ENSG00000001561	ENSG00000001561	ENSG00000002549	ENSG00000002549	ENSG00000001167	ENSG00000001461
ENSG00000001617	ENSG00000001617	ENSG00000002587	ENSG00000002587	ENSG00000001461	ENSG00000001561
ENSG00000001629	ENSG00000001629	ENSG00000002726	ENSG00000002726	ENSG00000001561	ENSG00000001617
ENSG00000001630	ENSG00000001630	ENSG00000002745	ENSG00000002745	ENSG00000001617	ENSG00000001629
ENSG00000001631	ENSG00000001631	ENSG00000002933	ENSG00000002933	ENSG00000001629	ENSG00000001630
ENSG00000002330	ENSG00000002330	ENSG00000003989	ENSG00000003989	ENSG00000001630	ENSG00000001631
ENSG00000002549	ENSG00000002549	ENSG00000004139	ENSG00000004139	ENSG00000001631	ENSG00000002330
ENSG00000002586	ENSG00000002586	ENSG00000004534	ENSG00000004534	ENSG00000002330	ENSG00000002549
ENSG00000002587	ENSG00000002587	ENSG00000004779	ENSG00000004779	ENSG00000002549	ENSG00000002586
ENSG00000002726	ENSG00000002726	ENSG00000004799	ENSG00000004799	ENSG00000002586	ENSG00000002587
ENSG00000002745	ENSG00000002745	ENSG00000005020	ENSG00000005020	ENSG00000002587	ENSG00000002726
ENSG00000002919	ENSG00000002919	ENSG00000005102	ENSG00000005102	ENSG00000002726	ENSG00000002745
ENSG00000002933	ENSG00000002933	ENSG00000005108	ENSG00000005108	ENSG00000002745	ENSG00000002919
ENSG00000003056	ENSG00000003056	ENSG00000005379	ENSG00000005379	ENSG00000002919	ENSG00000002933
edgeR_robust	**voom**	**SAMseq**	**PoissonSeq**	**baySeq**	**EBSeq**
ENSG00000000457	ENSG00000000457	ENSG00000000003	ENSG00000000003	ENSG00000001461	ENSG00000000460
ENSG00000000460	ENSG00000000460	ENSG00000000457	ENSG00000000457	ENSG00000001561	ENSG00000000971
ENSG00000000971	ENSG00000000971	ENSG00000000460	ENSG00000000938	ENSG00000001617	ENSG00000001461
ENSG00000001036	ENSG00000001084	ENSG00000000971	ENSG00000000971	ENSG00000002330	ENSG00000001561
ENSG00000001084	ENSG00000001461	ENSG00000001036	ENSG00000001036	ENSG00000002549	ENSG00000001617
ENSG00000001167	ENSG00000001561	ENSG00000001084	ENSG00000001084	ENSG00000002726	ENSG00000001629
ENSG00000001461	ENSG00000001617	ENSG00000001167	ENSG00000001461	ENSG00000002745	ENSG00000001630
ENSG00000001561	ENSG00000001626	ENSG00000001461	ENSG00000001561	ENSG00000002933	ENSG00000001631
ENSG00000001617	ENSG00000001629	ENSG00000001561	ENSG00000001617	ENSG00000003056	ENSG00000002330
ENSG00000001626	ENSG00000001630	ENSG00000001617	ENSG00000001626	ENSG00000003400	ENSG00000002549
ENSG00000001629	ENSG00000001631	ENSG00000001626	ENSG00000001629	ENSG00000003509	ENSG00000002586
ENSG00000001630	ENSG00000002330	ENSG00000001629	ENSG00000001630	ENSG00000004059	ENSG00000002587
ENSG00000001631	ENSG00000002549	ENSG00000001630	ENSG00000001631	ENSG00000004139	ENSG00000002726
ENSG00000002330	ENSG00000002586	ENSG00000001631	ENSG00000002330	ENSG00000004468	ENSG00000002919
ENSG00000002549	ENSG00000002587	ENSG00000002330	ENSG00000002549	ENSG00000004534	ENSG00000002933
ENSG00000002586	ENSG00000002726	ENSG00000002549	ENSG00000002586	ENSG00000004766	ENSG00000003056
ENSG00000002587	ENSG00000002745	ENSG00000002586	ENSG00000002726	ENSG00000004779	ENSG00000003400
ENSG00000002726	ENSG00000002919	ENSG00000002587	ENSG00000002919	ENSG00000005020	ENSG00000004059
ENSG00000002745	ENSG00000002933	ENSG00000002726	ENSG00000002933	ENSG00000005102	ENSG00000004139
ENSG00000002919	ENSG00000003056	ENSG00000002745	ENSG00000003056	ENSG00000005379	ENSG00000004534

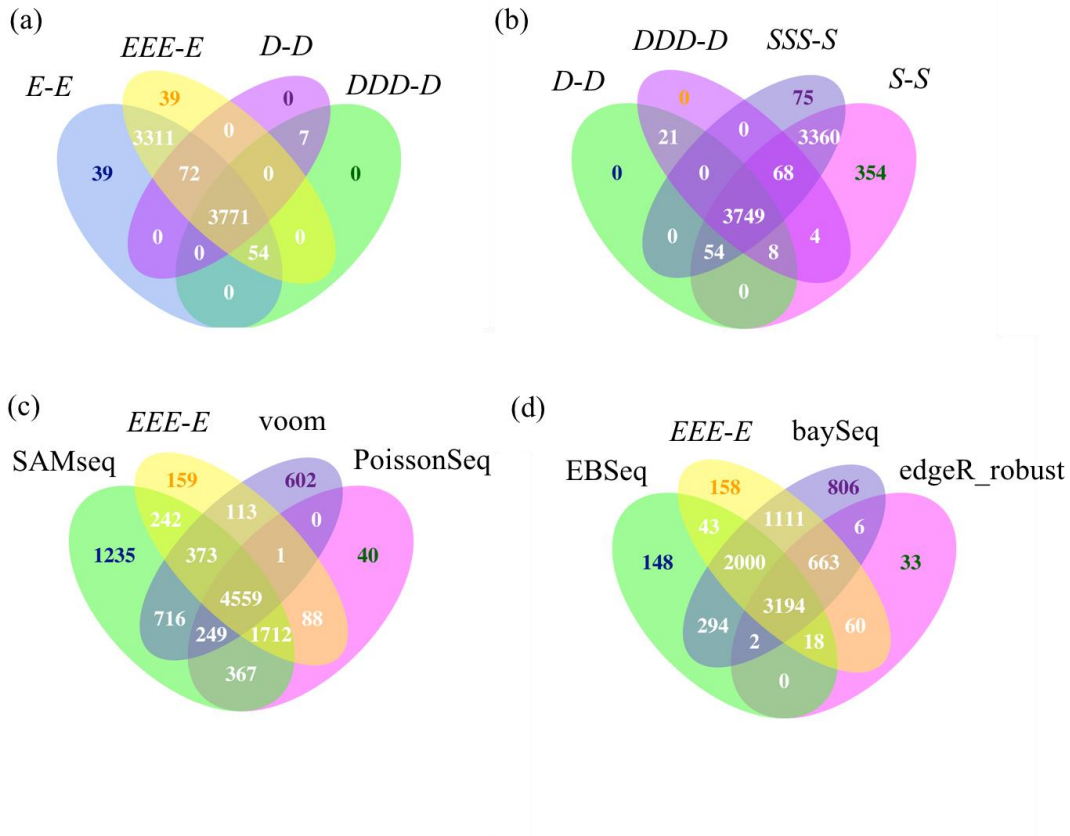
There are 38 kinds of DEGs in the top 20 genes of 12 short gene lists. The genes at the 10 top of the ranking according to the appearance frequency are in bold.

Additional file 9 - Dendrogram of average-linkage hierarchical clustering for 12 ranked gene lists



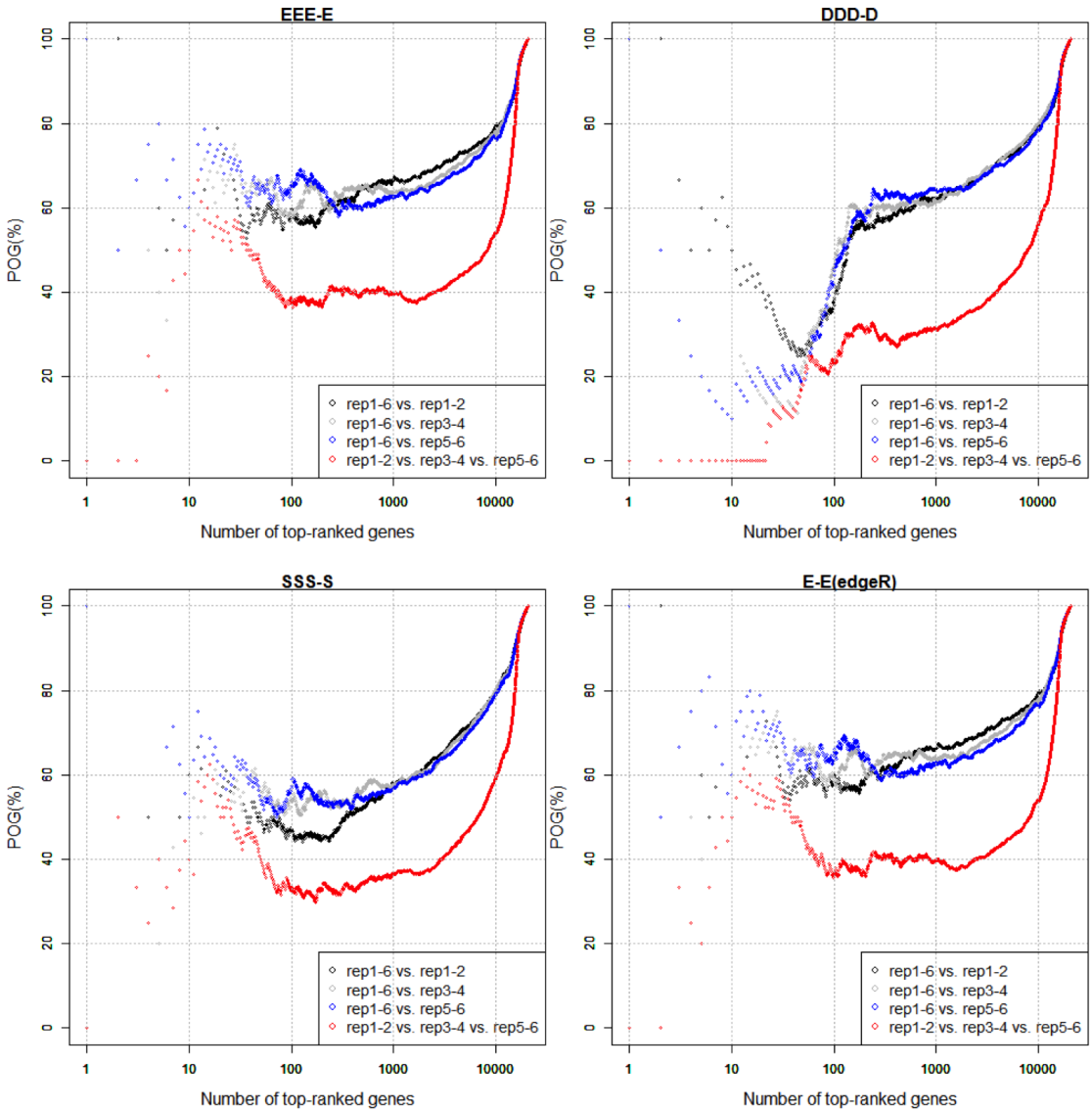
Twelve ranked gene lists used for constructing the dendrogram were obtained from the analysis of the simulation data under the following conditions: P_{DEG} = 5 %, (0.5, 0.4, 0.1) for (PG1, PG2, PG3), and Nrep = 9. The clustering was performed using the "clusterSample" function with distances defined as (1 – Spearman's rank correlation coefficient). EBSeq showed the highest AUC values (= 96.83 %) in this simulation trial, followed by *EEE-E* (96.45 %), *E-E* (96.42 %), *DDD-D* (96.35 %), *D-D* (96.31 %), baySeq (96.21 %), edgeR_robust (95.13 %), S-S (94.54 %), SSS-S (94.43 %), PoissonSeq (94.07 %), voom (92.70 %), and SAMseq (92.23 %).

Additional file 10 - Overlaps among the four sets of DEGs among the three species



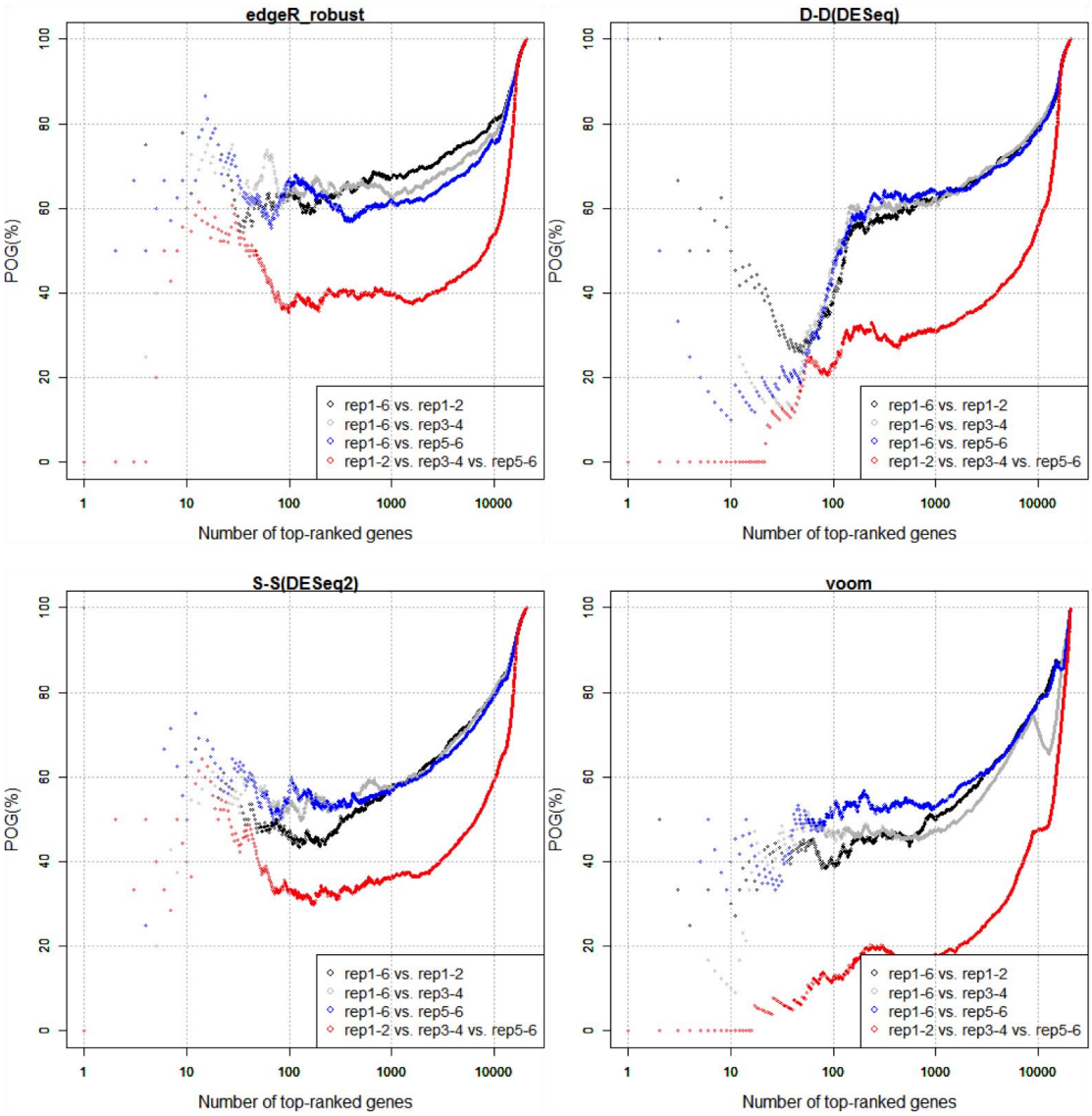
Venn diagram shows the overlaps among four pipelines. In (a) and (b), the three pairs (*EEE-E* vs. *E-E*, *DDD-D* vs. *D-D*, and *SSS-S* vs. *S-S*) show great consistent (also can be seen in Additional 6). Note that, although the number of DEGs in the first pair is same, these are still 39 genes are not in each other's DEG list. In addition, comparing pipelines *EEE-E* (DEGES-based edgeR) and edgeR_robust (the advanced version of edgeR), the vast majority of DEGs (99.59%) in the latter are included by the former. The *SSS-S* (7295) is more conservative than *S-S* (7585). Most of the DEGs (96.33%) in EBSeq are included by baySeq.

Additional file 11 - Percentages of Overlapping Genes (POGs) among ranked gene lists for *EEE-E*, *DDD-D*, *SSS-S*, and *E-E* (edgeR)



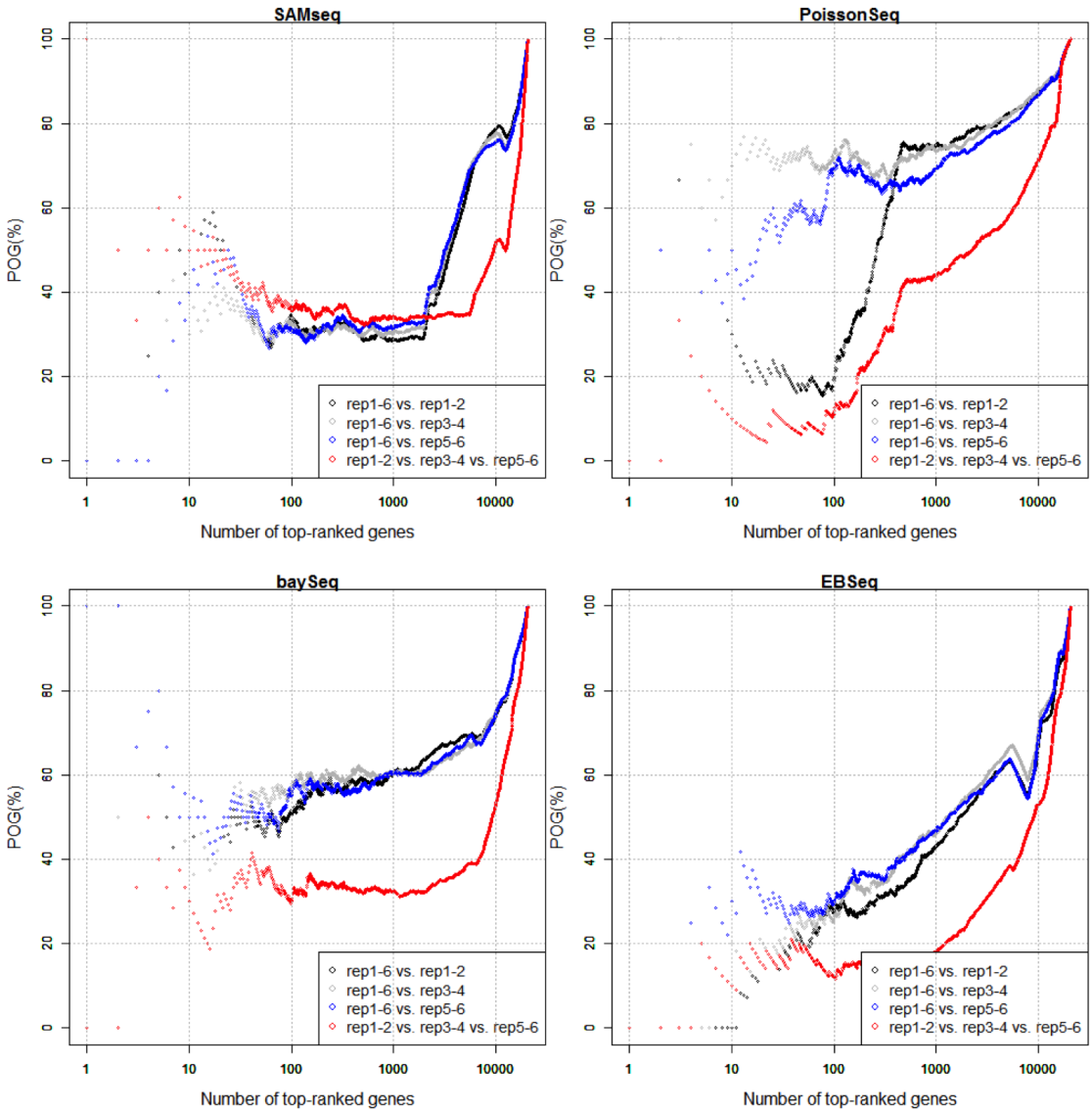
POG values for any numbers of top-ranked genes for four individual pipelines are shown. Legends are basically the same as in Figure 13.

Additional file 12 - Percentages of Overlapping Genes (POGs) among ranked gene lists for edgeR_robust, *D-D* (DESeq), *S-S* (DESeq2) and voom



POG values for any numbers of top-ranked genes for four individual pipelines are shown. Legends are basically the same as in Figure 13.

Additional file 13 - Percentages of Overlapping Genes (POGs) among ranked gene lists for SAMseq, PoissonSeq, baySeq and EBSeq



POG values for any numbers of top-ranked genes for four individual pipelines are shown. Legends are basically the same as in Figure 13.

Index

AUC: the area under the curve

BB: beta-binomial (distribution or model)

BR: biological replicate

DE: differential expression

DEG: differentially expressed genes

DEGES: DEG elimination strategy

FDR: false discovery rate

GLM: generalized linear model

NB: negative-binomial (distribution or model)

POG: percentages of overlapping genes

HS: *Homo sapiens*

PT: *Pan troglodytes*

RM: *Rhesus macaques*

TMM: trimmed mean of M values (method)

TbT: the TMM-baySeq-TMM pipeline

TCC: Tag Count Comparison

RPKM: reads per kilobase of exon model per million mapped reads

FPKM: fragments per kilobase of exon model per million mapped fragments

Acknowledgements

My deepest gratitude goes first and foremost to Professor Shimizu Kentaro[1], my supervisor, for his constant encouragement and guidance. He has walked me through all the stages of the writing of this thesis. Without his consistent and illuminating instruction, this thesis could not have reached its present form.

Second, I would like to express my heartfelt gratitude to Professor Kadota Koji[2], who led me into the world of RNA-seq. I am also greatly indebted to the professors and teachers at the Department of biotechnology: Professor Nakamura Shugo[1], Professor Terada Tohru, who have instructed and helped me a lot in the past three years.

Third, I would like to thank for the grant from the Chinese Scholarship Council (CSC), which gave me the opportunity of studying in Japan without worrying about the economic problem.

Last my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. I also owe my sincere gratitude to my friends who gave me their help and time in listening to me and helping me work out my problems during the difficult course of the thesis.

[1] Bioinformation Engineering Lab, Graduate School of Agricultural and Life Sciences, Faculty of Agriculture, The University of Tokyo.
[2] Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo.

References

[1] Fields S, Johnston M. **Cell biology. Whither Model Organism Research?** *Science* 2005, **307**(5717):1885-6.

[2] Chang TW. **Binding of cells to matrixes of distinct antibodies coated on solid surface**. *J Immunol Methods* 1983, **65**(1–2):217-23.

[3] Schena M, Shalon D, Davis RW, Brown PO. **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-70.

[4] Brown PO, Botstein D. **Exploring the new world of the genome with DNA microarrays**. *Nat Genet* 1999, **21**:33-7.

[5] Lander ES. **Array of hope**. *Nat Genet* 1999, **21**(1 Suppl):3-4.

[6] Barczak A, Rodriguez MW, Hanspers K, *et al.* **Spotted long oligonucleotide arrays for human gene expression analysis**. *Genome Res* 2003, **13**(7):1775-85.

[7] Carter MG, Hamatani T, Sharov AA, *et al.* **In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling**. *Genome Res* 2003, **13**(5):1011-21.

[8] Hughes TR, Mao M, Jones AR, *et al.* **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer**. *Nat Biotechnol* 2001, **19**(4):342-7.

[9] Kane MD, Jatkoe TA, Stumpf CR, *et al.* **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays**. *Nucleic Acids Res* 2000, **28**(22):4552-7.

[10] Kothapalli R, Yoder SJ, Mane S, *et al.* **Microarray results: how accurate are they?** *BMC Bioinformatics* 2002, **3**:22.

[11] Kuo WP, Jenssen TK, Butte AJ, *et al.* **Analysis of matched mRNA measurements from two different microarray technologies**. *Bioinformatics* 2002, **18**(3):405-12.

[12] Li J, Pankratz M, Johnson JA. **Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays**. *Toxicol Sci* 2002, **69**(2):383-90.

[13] Tan PK, Downey TJ, Spitznagel EL, *et al.* **Evaluation of gene expression measurements from commercial microarray platforms**. *Nucleic Acids Res* 2003, **31**(19):5676-84.

[14] Wang HY, Malek RL, Kwitek AE, *et al.* **Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays**. *Genome Biol* 2003, **4**(1):R5.

[15] Yuen T, Wurmbach E, Pfeffer RL, *et al.* **Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays**. *Nucleic Acids Res* 2002, **30**(10):e48.

[16] Irizarry RA, Warren D, Spencer F, *et al.* **Multiple-laboratory comparison of microarray platforms**. *Nat Methods* 2005, **2**(5):345-50.

[17] MAQC Consortium, Shi L, Reid LH, *et al.* **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nat Biotechnol* 2006, **24**(9):1151-61.

[18] Canales RD, Luo Y, Willey JC, *et al.* **Evaluation of DNA microarray results with quantitative gene expression platforms**. *Nat Biotechnol* 2006, **24**(9):1115-22.

[19] Casneuf T, Van de Peer Y, Huber W. **In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation**. *BMC Bioinformatics* 2007, **8**:461.

[20] Eklund AC, Turner LR, Chen P, *et al.* **Replacing cRNA targets with cDNA reduces microarray cross-hybridization**. *Nat Biotechnol* 2006, **24**(9):1071-3.

[21] Okoniewski MJ, Miller CJ. **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations**. *BMC Bioinformatics* 2006, **7**:276.

[22] Smyth GK. **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.

[23] Sanger F, Nicklen S, Coulson AR. **DNA Sequencing with Chain-Terminating Inhibitors**. **1977**. *Biotechnology* 1992, **24**:104-8.

[24] Yassour M, Kapian T, Fraser HB, *et al.* **Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing**. *Proc Natl Acad Sci U S A* 2009, **106**(9):3264-9.

[25] Wang ET, Sandberg R, Luo SJ, *et al.* **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-6.

[26] Sultan M, Schulz MH, Richard H, *et al.* **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. *Science* 2008, **321**(5891):956-60.

[27] Lister R, O'Malley RC, Tonti-Filippini J, *et al.* **Highly integrated single-base resolution maps of the epigenome in Arabidopsis**. *Cell* 2008, **133**(3):523-36.

[28] Mortazavi A, Williams BA, Mccue K, *et al.* **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**(7):621-8.
[29] Risso D, Ngai J, Speed TP, *et al.* **Normalization of RNA-seq data using factor analysis of control genes or samples**. *Nat Biotechnol* 2014, **32**(9):896-902.
[30] Marioni JC, Mason CE, Mane SM, *et al.* **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res* 2008, **18**(9):1509-17.
[31] Cloonan N, Forrest AR, Kolle G, *et al.* **Stem cell transcriptome profiling via massive-scale mRNA sequencing**. *Nat Methods* 2008, **5**(7):613-9.
[32] Nagalakshmi U, Wang Z, Waern K, *et al.* **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science* 2008, **320**(5881):1344-9.
[33] Maher CA, Kumar-Sinha C, Cao X, *et al.* **Transcriptome sequencing to detect gene fusions in cancer**. *Nature* 2009, **458**(7234):97-101.
[34] Guttman M, Garber M, Levin JZ, *et al.* **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nat Biotechnol* 2010, **28**(5):503-10.
[35] Denoeud F, Aury JM, Da Silva C, *et al.* **Annotating genomes with massive-scale RNA sequencing**. *Genome Biol* 2008, **9**(12):R175.
[36] Berger MF, Levin JZ, Vijayendran K, *et al.* **Integrative analysis of the melanoma transcriptome**. *Genome Res* 2010, **20**(4):413-27.
[37] Wilhelm BT, Briau M, Austin P, *et al.* **RNA-seq analysis of 2 closely related leukemia clones that differ in their self-renewal capacity**. *Blood* 2011, **117**(2):e27-38.
[38] Mortazavi A, Schwarz EM, Williams B, *et al.* **Scaffolding a Caenorhabditis nematode genome with RNA-seq**. *Genome Res* 2010, **20**(12):1740-7.
[39] Blekhman R, Marioni JC, Zumbo P, *et al.* **Sex-specific and lineage-specific alternative splicing in primates**. *Genome Res* 2010, **20**(2):180-9.
[40] Trapnell C, Williams BA, Pertea G, *et al.* **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-5.
[41] SEQC/MAQC-III. **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium**. *Nat Biotechnol* 2014, **32**(9):903-14.
[42] Li S, Tighe SW, Nicolet CM, *et al.* **Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study**. *Nat Biotechnol* 2014, **32**(9):915-25.
[43] Trapnell C, Pachter L, Salzberg SL. **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-11.
[44] Langmead B, Salzberg SL. **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-9.
[45] Li H, Durbin R. **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-60.
[46] Li H, Durbin R. **Fast and accurate long-read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-95.
[47] Anders S, Pyl PT, Huber W. **HTSeq–A Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2015, **31**(2):166-9.
[48] Quinlan AR, Hall IM. **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-2.
[49] Liao Y, Smyth GK, Shi W. **featureCounts: an efficient general-purpose read summarization program**. *Bioinformatics* 2014, **30**(7):923-30.
[50] Trapnell C, Roberts A, Goff L, *et al.* **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-78.
[51] Cumbie JS, Kimbrel JA, Di Y, *et al.* **GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences**. *PloS one* 2011, **6**(10): e25279.
[52] Holt RA, Jones SJ. **The new paradigm of flow cell sequencing**. *Genome Res* 2008, **18**(6):839-46.
[53] Bullard JH, Purdom E, Hansen KD, *et al.* **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinformatics* 2010, **11**:94.
[54] Risso D, Schwartz K, Sherlock G, *et al.* **GC-content normalization for RNA-Seq data**. *BMC Bioinformatics* 2011, **12**:480.
[55] Dillies MA, Rau A, Aubert J, *et al.* **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Brief Bioinform* 2013, **14**(6):671-83.

[56] Oshlack A, Wakefield MJ. **Transcript length bias in RNA-seq data confounds systems biology**. *Biol Direct* 2009, **4**:14.
[57] Hansen KD, Irizarry RA, Wu ZJ. **Removing technical variability in RNA-seq data using conditional quantile normalization**. *Biostatistics* 2012, **13**(2):204-16.
[58] Young MD, Wakefield MJ, Smyth GK, *et al.* **Gene ontology analysis for RNA-seq: accounting for selection bias**. *Genome Biol* 2010, **11**(2):R14.
[59] Bolstad BM, Irizarry RA, Astrand M, *et al.* **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-93.
[60] Robinson MD, Oshlack A. **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biol* 2010, **11**(3):R25.
[61] Robinson MD, McCarthy DJ, Smyth GK. **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-40.
[62] Anders S, Huber W. **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**(10):R106.
[63] Yang YH, Thorne NP. **Normalization for two-color cDNA microarray data**. *Lecture Notes-Monograph Series* 2003, 403-18.
[64] Li J, Witten DM, Johnstone IM, *et al.* **Normalization, testing, and false discovery rate estimation for RNA-sequencing data**. *Biostatistics* 2012, **13**(3):523-38.
[65] Sun Z, Zhu Y. **Systematic comparison of RNA-Seq normalization methods using measurement error models**. *Bioinformatics* 2012, **28**(20):2584-91.
[66] Hardcastle TJ, Kelly KA. **baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics* 2010, **11**:422.
[67] Hardcastle TJ, Kelly KA. **Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution**. *BMC Bioinformatics* 2013, **14**:135.
[68] Zhou YH, Xia K, Wright FA. **A powerful and flexible approach to the analysis of RNA sequence count data**. *Bioinformatics* 2011, **27**(19):2672-8.
[69] Esnaola M, Puig P, Gonzalez D, *et al.* **A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments**. *BMC Bioinformatics* 2013, **14**:254.
[70] Al Seesi S, Tiagueu YT, Zelikovsky A, *et al.* **Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates**. *BMC Genomics* 2014, **15** Suppl 8:S2.
[71] Rapaport F, Khanin R, Liang Y, *et al.* **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data**. *Genome Biol* 2013, **14**(9):R95.
[72] Liu Y, Zhou J, White KP. **RNA-seq differential expression studies: more sequence or more replication?** *Bioinformatics* 2014, **30**(3):301-4.
[73] Zhang ZH, Jhaveri DJ, Marshall VM, *et al.* **A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data**. *PloS One* 2014, **9**(8):e103207.
[74] Ching T, Huang SJ, Garmire LX. **Power analysis and sample size estimation for RNA-Seq differential expression**. *RNA* 2014, **20**(11):1684-96.
[75] Law CW, Chen Y, Shi W, *et al.* **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome Biol* 2014, **15**(2):R29.
[76] Soneson C, Delorenzi M. **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC Bioinformatics* 2013, **14**:91.
[77] Tang M, Sun J, Shimizu K, *et al.* **Evaluation of methods for differential expression analysis on multi-group RNA-seq count data**. *BMC Bioinformatics* 2015, **16**:361.
[78] Wang L, Feng Z, Wang X, *et al.* **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data**. *Bioinformatics* 2010, **26**(1):136-8.
[79] Srivastava S, Chen L. **A two-parameter generalized Poisson model to improve the analysis of RNA-seq data**. *Nucleic Acids Res* 2010, **38**(17):e170.
[80] Wu Z, Jenkins BD, Rynearson TA, *et al.* **Empirical bayes analysis of sequencing-based transcriptional profiling without replicates**. *BMC Bioinformatics* 2010, **11**:564.
[81] Tarazona S, Garcia-Alcalde F, Dopazo J, *et al.* **Differential expression in RNA-seq: A matter of depth**. *Genome Res* 2011, **21**(12):2213-23.
[82] Auer PL, Doerge RW. **A Two-Stage Poisson Model for Testing RNA-Seq Data**. *Stat Appl Genet Mol Biol* 2011, **10**(1): Article26.
[83] Di YM, Schafer DW, Cumbie JS, *et al.* **The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq**. *Stat Appl Genet Mol Biol* 2011, **10**(1): Article24.
[84] Glaus P, Honkela A, Rattray M. **Identifying differentially expressed transcripts from RNA-seq data with biological variation**. *Bioinformatics* 2012, **28**(13):1721-8.

[85] Lund SP, Nettleton D, McCarthy DJ, *et al.* **Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates**. *Stat Appl Genet Mol Biol* 2012, **11**(5).
[86] Feng J, Meyer CA, Wang Q, *et al.* **GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data**. *Bioinformatics* 2012, **28**(21):2782-8.
[87] Sun J, Nishiyama T, Shimizu K, *et al.* **TCC: an R package for comparing tag count data with robust normalization strategies**. *BMC Bioinformatics* 2013, **14**:219.
[88] Trapnell C, Hendrickson DG, Sauvageau M, *et al.* **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nat Biotechnol* 2013, **31**(1):46-53.
[89] Li J, Tibshirani R. **Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data**. *Stat Methods Med Res* 2013, **22**(5):519-36.
[90] Leng N, Dawson JA, Thomson JA, *et al.* **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments**. *Bioinformatics* 2013, **29**(8):1035-43.
[91] Wu H, Wang C, Wu Z. **A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data**. *Biostatistics* 2013, **14**(2):232-43.
[92] Van De Wiel MA, Leday GG, Pardo L, *et al.* **Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors**. *Biostatistics* 2012, **14**(1):113-28.
[93] Bi Y, Davuluri RV. **NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data**. *BMC Bioinformatics* 2013, **14**:262.
[94] Love MI, Huber W, Anders S. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
[95] Gu J, Wang X, Halakivi-Clarke L, *et al.* **BADGE: A novel Bayesian model for accurate abundance quantification and differential analysis of RNA-Seq data**. *BMC Bioinformatics* 2014, **15** Suppl 9:S6.
[96] Zhou X, Lindsay H, Robinson MD. **Robustly detecting differential expression in RNA sequencing data using observation weights**. *Nucleic Acids Res* 2014, **42**(11):e91.
[97] Kadota K, Nishiyama T, Shimizu K. **A normalization strategy for comparing tag count data**. *Algorithms Mol Biol* 2012, **7**(1):5.
[98] Luo H, Li J, Chia BK, *et al.* **The importance of study design for detecting differentially abundant features in high-throughput experiments**. *Genome Biol* 2014, **15**(12):527.
[99] Garmire LX, Subramaniam S. **Evaluation of normalization methods in mammalian microRNA-Seq data**. *RNA* 2012, **18**(6):1279-88.
[100] Seyednasrollah F, Laiho A, Elo LL. **Comparison of software packages for detecting differential expression in RNA-seq studies**. *Brief Bioinform* 2015, **16**(1):59-70.
[101] An J, Kim K, Chae H, *et al.* **DEGPACK: A web package using a non-parametric and information theoretic algorithm to identify differentially expressed genes in multiclass RNA-seq samples**. *Methods* 2014, **69**(3):306-14.
[102] 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.
[103] Gentleman RC, Carey VJ, Bates DM, *et al.* **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.
[104] Soneson C. **compcodeR-an R package for benchmarking differential expression methods for RNA-seq data**. *Bioinformatics* 2014, **30**(17):2517-18.
[105] Benidt S, Nettleton D. **SimSeq: a nonparametric approach to simulation of RNA-sequence datasets**. *Bioinformatics* 2015, **31**(13):2131-40.
[106] Dembele D, Kastner P. **Fold change rank ordering statistics: a new method for detecting differentially expressed genes**. *BMC Bioinformatics* 2014, **15**:14.
[107] Farztdinov V, McDyer F. **Distributional fold change test - a statistical approach for detecting differential expression in microarray experiments**. *Algorithms Mol Biol* 2012, **7**(1):29.
[108] Kadota K, Shimizu K. **Evaluating methods for ranking differentially expressed genes applied to microArray quality control data**. *BMC Bioinformatics* 2011, **12**:227.
[109] Kadota K, Nakai Y, Shimizu K. **Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity**. *Algorithms Mol Biol* 2009, **4**:7.
[110] Kadota K, Nakai Y, Shimizu K. **A weighted average difference method for detecting differentially expressed genes from microarray data**. *Algorithms Mol Biol* 2008, **3**:8.
[111] Khang TF, Lau CY. **Getting the most out of RNA-seq data analysis**. *PeerJ* 2015, **3**:e1360.
[112] Cook RD. **Detection of Influential Observation in Linear-Regression**. *Technometrics* 1977, **19**:15-18.

[113] Hochreiter S, Clevert DA, Obermayer K. **A new summarization method for affymetrix probe level data**. *Bioinformatics* 2006, **22**(8):943-9.
[114] Barrett CL, Schwab RB, Jung H, *et al.* **Transcriptome sequencing of tumor subpopulations reveals a spectrum of therapeutic options for squamous cell lung cancer**. *PLoS One* 2013, **8**(3):e58714.
[115] Sanchez A, Golding I. **Genetic Determinants and Cellular Constraints in Noisy Gene Expression**. *Science* 2013, **342**(6163):1188-93.