BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 20 2012, pages 2584–2591
doi:10.1093/bioinformatics/bts497

Sequence analysis

Advance Access publication August 22, 2012

Systematic comparison of RNA-Seq normalization methods using measurement error models

Zhaonan Sun* and Yu Zhu*

Department of Statistics, Purdue University, 250N University Street, West Lafayette, IN 47906, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Further advancement of RNA-Seq technology and its application call for the development of effective normalization methods for RNA-Seq data. Currently, different normalization methods are compared and validated by their correlations with a certain gold standard. Gene expression measurements generated by a different technology or platform such as Real-time reverse transcription polymerase chain reaction (qRT–PCR) or Microarray are usually used as the gold standard. Although the current approach is intuitive and easy to implement, it becomes statistically inadequate when the gold standard is also subject to measurement error (ME). Furthermore, the current approach is not informative, because the correlation of a normalization method with a certain gold standard does not provide much information about the exact quality of the normalized RNA-Seq measurements.

Results: We propose to use the system of ME models based on qRT–PCR, Microarray and RNA-Seq gene expression data to compare and validate RNA-Seq normalization methods. This approach does not assume the existence of a gold standard. The performance of a normalization method can be characterized by a group of parameters of the system, which are referred to as the performance parameters, and these performance parameters can be consistently estimated. Different normalization methods can thus be compared by comparing their corresponding estimated performance parameters. We applied the proposed approach to compare five existing RNA-Seq normalization methods using the gene expression data of two RNA samples from the microArray Quality Control and Sequencing Quality Control projects and gained much insight about the pros and cons of these methods.

Contact: sunz@purdue.edu; yuzhu@purdue.edu

Received on April 10, 2012; revised on July 19, 2012; accepted on August 7, 2012

1 INTRODUCTION

RNA-Seq is a recently developed platform for transcriptome study using the next-generation sequencing technology. It offers the possibility to generate quality measurements of transcript/gene expression levels with a whole-genome coverage and single-nucleotide resolution. Compared with the other widely used technology Microarray, RNA-Seq is believed to have a lower level background signal and a wider dynamic range of detection (Agarwal *et al.*, 2010; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008). Since its advent, RNA-Seq has been used in many transcriptome studies (Cloonan *et al.*, 2008; Marioni *et al.*, 2008; Pan *et al.*, 2008).

Although RNA-Seq is promising, it is known that many factors could introduce variation and bias to RNA-Seq data, including the random primers (Hansen *et al.*, 2010), mapping method (Degner *et al.*, 2009) used in an RNA-Seq experiment as well as the molecular constitution and secondary structure of the RNA sample under study (Hansen *et al.*, 2010; Li *et al.*, 2010). Therefore, proper normalization methods are needed to denoise and normalize the RNA-Seq data before they are used in downstream analysis such as gene expression level quantification and differential expression detection.

Mortazavi *et al.* (2008) proposed to use Reads Per Kilobase of exon model per Million mapped reads (RPKM) to normalize RNA-Seq data. The RPKM method is by far the most widely used RNA-Seq normalization method. The validity of the RPKM method depends on the assumption that the reads count a gene receives follows a Poisson distribution with a constant intensity rate. However, the assumption of a constant intensity rate was found to be invalid in RNA-Seq data. Srivastava and Chen (2010) showed that a Poisson distribution with a constant rate cannot explain the non-uniformity of the reads across the same gene or exon, and Li *et al.* (2010) showed the existence of overdispersion in RNA-Seq reads data. A number of methods have been proposed to improve the RPKM method, and they can be grouped into three categories. The first category consists of methods that use more sophisticated distributions to accommodate overdispersion. For example, Srivastava and Chen (2010) proposed to use a two-parameter generalized Poisson model and developed a method called 'GPseq'; Wu *et al.* (2012) proposed to use Poisson mixture models and developed a method called 'PMseq'. The second category includes methods that adjust the expression level measurement of a gene by some identified sources of variation or bias. For example, Risso *et al.* (2011) proposed to use the GC content of a gene to correct its gene expression level measurement; in the rest of the article, we refer to their method as the 'GCR' method. Methods in the third category assume that the reads count a nucleotide receives not only depends on the expression level of the transcript, but it also depends on the 'sequencing preference' of its neighborhood. An example of this type of method is the 'mseq' method proposed by Li *et al.* (2010), which incorporates local sequencing preferences into the framework of Poisson linear models.

As new methods continue to appear, it is important to systematically compare different RNA-Seq normalization methods. To evaluate a normalization method, two aspects of the method are of particular importance. The first aspect is the concordance of

*To whom correspondence should be addressed.

the expression level measurements generated by a normalization method with the 'true' expression levels. The second aspect is whether a normalization method would produce consistent measurements when it is applied to RNA-Seq data of same biological condition generated from different experiments. In previous RNA-Seq studies, two major criteria are used for assessing a normalization method following the schema proposed by Irizarry *et al*. (2005), which are referred to as the *precision* and *accuracy* criteria. The precision of a method is defined to be the correlation between measurements from replicated experimental runs, and the accuracy is defined to be the correlation coefficient of the expression level measurements generated by the method with a gold standard. The expression measurements from a different platform are usually used as the gold standard.

These two criteria are intuitive and popularly used. The precision criterion can assess the consistency between repeated experimental runs. However, the validity of using the accuracy criterion to evaluate the performance of a RNA-Seq normalization method depends on how accurate the gold standard is. Currently, in the reported comparison studies of RNA-Seq normalization methods, polymerase chain reaction (PCR)-based technology and Microarray are two common choices of the gold standard. Microarray is known to suffer from various types of bias and variation. Although many normalization methods have been proposed for Microarray data, variations and biases in Microarray data cannot be completely eliminated. PCR-based technology is well acknowledged to be of high accuracy, but it is still subject to errors. It has been pointed out that there exist a number of factors that contribute variations to PCR data, including primer–dimer accumulation, PCR product reannealing and DNA polymerase binding to primers (Brownie *et al*., 1997). It is known in statistics that when the gold standard is subject to measurement errors (MEs), directly using the correlation coefficient with the gold standard to evaluate a normalization method can be inadequate. Unfortunately, a gold standard free of MEs may never be available due to the complexity involved in transcriptome study.

A statistically more appropriate approach to assessing different normalization methods is to directly estimate the bias and variance of the normalized RNA-Seq measurements using ME models. A system of such ME models lately was used to calibrate RNA-Seq and Microarray gene expression level measurements with Real-time reverse transcription polymerase chain reaction (qRT–PCR) for more accurate transcriptome profiling (Z.Sun *et al*., unpublished manuscript). Through the system, the biases and variances from qRT–PCR, Microarray and RNA-Seq can be estimated. In this article, we further utilize the system to facilitate the comparison of different RNA-Seq normalization methods. We applied the developed approach to compare five existing RNA-Seq normalization methods, which are RPKM, GPseq, mseq, GCR and PMseq, using gene expression data of two RNA samples from the MicroArray Quality Control (MAQC) and Sequencing Quality Control (SEQC) projects.

2 METHODS

We assume the expression levels of n genes are measured by different laboratories using qRT–PCR, Microarray and RNA-Seq. Specifically, k_1 laboratories use qRT–PCR, k_2 laboratories use Microarray and k_3 laboratories use RNA-Seq. The laboratories are assumed to be independent with each other. We also assume each laboratory generates one expression measurement for each gene. When technical replicates are available within one laboratory, the average value across replicates is used as the expression measurement of a gene.

Raw data generated by qRT–PCR and Microarray are denoised using standard platform-specific normalization methods such as the delta Ct method (qRT–PCR) and the Robust Multichip Average method (Microarray). We use $X = \{X_{jk} : j = 1, \ldots, n; k = 1, \ldots, k_1\}$ to denote the normalized qRT–PCR gene expression measurement data, where X_{jk} is the qRT–PCR measurement of gene j from the kth laboratory. Similarly, $Y = \{Y_{jk} : j = 1, \ldots, n; k = 1, \ldots, k_2\}$ denotes the normalized Microarray gene expression measurement data.

Let L be the number of RNA-Seq normalization methods we want to compare, and let $l = \mathbf{1}, \ldots, L$ be the indices for these normalization methods. The RNA-Seq raw data consist of the mapped short reads. By applying the L normalization methods to the RNA-Seq raw data, L sets of normalized RNA-Seq gene expression measurement data can be generated. We use $Z_{(l)} = \{Z_{jk,(l)} : j = 1, \ldots, n; k = 1, \ldots, k_3\}$ to denote the normalized RNA-Seq data by the lth method.

At the end, $L + 2$ sets of gene expression measurement data are generated, including the qRT–PCR data X from k_1 laboratories, the Microarray data Y from k_2 laboratories and L sets of RNA-Seq data $Z_{(1)}, \ldots, Z_{(L)}$ from k_3 laboratories.

As discussed in Section 1, qRT–PCR, Microarray and RNA-Seq are all subject to MEs. It is well known that MEs can become heteroscedastic, that is the size or variance of an ME depends on the magnitude of the targeted quantity. This is, in general, the case for the measurement data X, Y and $Z_{(l)}$ for $l = \mathbf{1}, \ldots, L$. One popularly used approach for handling heteroscedasticity in practice is to apply Box–Cox transformations. In particular, base 2 logarithmic transformation (log-2) is often used to transform original gene expression measurements in the literature on gene expression data analysis, and in the resulting data, the heteroscedasticity can be mitigated. In this article, we follow the convention to transform X, Y and $Z_{(l)}$ using log-2 and use the transformed data in analysis. For ease of representation and discussion, we still use X, Y and $Z_{(l)}$ to refer to the log-2 transformed data in the rest of the article.

For each l, the combined data of X, Y and $Z_{(l)}$ can be characterized by a system of ME models.

$$X_{jk} = \mu_j + \kappa_{1j} + \pi_{1jk}, \tag{1a}$$

$$Y_{jk} = \alpha_2 + \beta_2 \mu_j + \kappa_{2j} + \pi_{2jk}, \tag{1b}$$

$$Z_{jk,(l)} = \alpha_{3,(l)} + \beta_{3,(l)} \mu_j + \kappa_{3j,(l)} + \pi_{3jk,(l)}. \tag{1c}$$

In the aforementioned system, μ_j denotes the underlying expression level of gene j, which is treated as a fixed but unknown quantity, α_2 and β_2 are the intercept and slope in the linear relationship postulated between μ_j and the Microarray measurements, and $\alpha_{3,(l)}$ and $\beta_{3,(l)}$ are the intercept and slope in the linear relationship postulated between μ_j and the normalized RNA-Seq measurements by the lth method.

The other terms in system (1) represent MEs attributed to various sources. There exist two major types of MEs, which are errors due to platforms and errors due to laboratories. In model (1a), κ_{1j} represents the ME due to the qRT–PCR platform and π_{1jk} the ME due to laboratories in X_{jkr}. We assume κ_{1j}s are *i.i.d.* $N(0, \psi_1^2)$ and π_{1jk}s are *i.i.d.* $N(0, \varsigma_1^2)$. In model (1b), κ_{2j} represents the ME due to the Microarray platform and π_{2jk} the ME due to laboratories in Y_{jkr}. We assume κ_{2j}s are *i.i.d.* $N(0, \psi_2^2)$ and π_{2jk}s are *i.i.d.* $N(0, \varsigma_2^2)$. In model (1c), $\kappa_{3j,(l)}$ represents the ME due to the RNA-Seq platform and $\pi_{3jk,(l)}$ the ME due to laboratories in $Z_{jkr,(l)}$. We assume $\kappa_{3j,(l)}$s are *i.i.d.* $N(0, \psi^2_{3,(l)})$ and $\pi_{3jk,(l)}$s are *i.i.d.* $N(0, \varsigma^2_{3,(l)})$.

In the system, μ_js are referred to as incidental parameters, and the other parameters are referred to as structural parameters. We further divide the structural parameters into two groups as follows. The first group consists of the parameters related to the normalized RNA-Seq measurements, which are collected and denoted as $\boldsymbol{\theta}_{(l)} = \left[\alpha_{3,(l)}, \beta_{3,(l)}, \psi^2_{3,(l)}, \varsigma^2_{3,(l)}\right]^\top$. The second group consists of the structural parameters related to qRT–PCR and Microarray measurements, which are collected and denoted as $\boldsymbol{\omega} = \left[\alpha_2, \beta_2, \psi^2_1, \varsigma^2_1, \psi^2_2, \varsigma^2_2\right]^\top$. Consistent estimates of the structural parameters can be found by treating μ_js as *i.i.d.* normal random variables. The estimation method was originally discussed in Barnett (1969), and the explicit formulas of the estimates and their standard errors can be found in Z.Sun *et al.* (unpublished manuscript). The estimates of these two groups of structural parameters based on the combined data of X, Y and $Z_{(l)}$ are denoted as $\widehat{\boldsymbol{\theta}}_{(l)}$ and $\widehat{\boldsymbol{\omega}}_{(l)}$.

The vector $\boldsymbol{\theta}_{(l)}$ reflects the quality of the normalized RNA-Seq measurements by the lth method. As discussed in Section 1, the RNA-Seq raw data are known to suffer from various types of variation and bias. The purpose of normalization is to reduce variation and bias in the original data and produce more accurate and consistent measurements. Thus $\boldsymbol{\theta}_{(l)}$ represents the remaining variation and bias in the normalized RNA-Seq measurements by the lth method, and it can be used to characterize the performance of the lth normalization method. Furthermore, the L normalization methods can be compared by comparing $\boldsymbol{\theta}_{(l)}$ for $l = 1, \ldots, L$. In general, the smaller the remaining bias and variation are, the better the normalization method.

The total remaining variance of $Z_{jk,(l)}$ is $\psi^2_{3,(l)} + \varsigma^2_{3,(l)}$, where $\psi^2_{3,(l)}$ is the remaining variability due to the RNA-Seq platform and $\varsigma^2_{3,(l)}$ is the remaining variability due to the laboratories. The variation due to the RNA-Seq platform can be attributed to a variety of factors, such as the library preparation method, the base calling method and the mapping method. $\psi^2_{3,(l)}$ and $\varsigma^2_{3,(l)}$ together characterize the reproducibility of the normalized RNA-Seq measurements by the lth method. In the rest of the article, we simply refer to $\psi^2_{3,(l)}$ and $\varsigma^2_{3,(l)}$ as the RNA-Seq variance and the laboratory variance of the lth method, respectively.

Under the system of ME models, the bias of $Z_{jk,(l)}$ from μ_j is $\alpha_{3,(l)} + \left(\beta_{3,(l)} - 1\right)\mu_j$, which is a linear function of μ_j. When $\alpha_{3,(l)} = 0$ and $\beta_{3,(l)} = 1$, the normalized RNA-Seq measurements by the lth method are unbiased; otherwise, they are biased with $\alpha_{3,(l)}$ and $\beta_{3,(l)}$ representing the location bias and the scale bias of the normalized RNA-Seq measurements, respectively. The absolute value of $\alpha_{3,(l)}$ and $\beta_{3,(l)} - 1$ can be used to characterize the biases of the lth RNA-Seq normalization method.

We can compare the performances of the L normalization methods by applying pairwise comparison with the components of $\widehat{\boldsymbol{\theta}}_{(l)}$ for different l. For example, to compare the scale biases of the first and the second normalization methods, we need to test the null hypothesis $H_0 : \beta_{3,(1)} = \beta_{3,(2)}$. The test statistic is $z = \hat{\beta}_{3,(1)} - \hat{\beta}_{3,(2)}/\mathrm{se}(\hat{\beta}_{3,(1)} - \hat{\beta}_{3,(2)})$, where se denotes the standard error. z approximately follows the standard normal distribution under H_0.

To carry out the test, we need to resolve one more difficulty, which is the dependence between $\widehat{\boldsymbol{\theta}}_{(l)}$ for different l. For instance, in the aforementioned example, $\hat{\beta}_{3,(1)}$ and $\hat{\beta}_{3,(2)}$ are correlated, but the estimate of their covariance is not available. Thus $\mathrm{se}\left(\hat{\beta}_{3,(1)} - \hat{\beta}_{3,(2)}\right)$ is not available. Here, we propose to use a bootstrap procedure to overcome this difficulty. Let B be the number of bootstrap samples; in this article, we set $B = 500$. Each bootstrap sample is generated by sampling n genes *with replacement* from the n genes. Let $X_{(b)}$, $Y_{(b)}$ and $Z_{(l,b)}$ $(l = 1, \ldots, L)$ be the qRT–PCR, Microarray and RNA-Seq measurements, respectively, of the bth bootstrap sample. The corresponding estimates of $\beta_{3,(1)}$ and $\beta_{3,(2)}$ based on the bth bootstrap sample are denoted by $\hat{\beta}_{3,(1,b)}$ and $\hat{\beta}_{3,(2,b)}$. Let $\mathrm{S}_{\beta_3,1,2} = \begin{bmatrix} s_{\beta_3,(1,1)} & s_{\beta_3,(1,2)} \\ s_{\beta_3,(1,2)} & s_{\beta_3,(2,2)} \end{bmatrix}$ be the sample variance and covariance matrix of $\left\{\hat{\beta}_{3,(1,b)} : b = 1, \ldots, B\right\}$ and $\left\{\hat{\beta}_{3,(2,b)} : b = 1, \ldots, B\right\}$. Then $\mathrm{se}(\hat{\beta}_{3,(1)} - \hat{\beta}_{3,(2)})$ is estimated by $\sqrt{s_{\beta_3,(1,1)} + s_{\beta_3,(2,2)} - 2s_{\beta_3,(1,2)}}$.

Note that $\boldsymbol{\omega}$ does not depend on l, therefore the L sets of estimates $\widehat{\boldsymbol{\omega}}_{(l)}$ for $l = 1, \ldots, L$ are expected to be coherent with each other. If one set of estimates is incoherent with the others, it indicates that the corresponding normalization method may not be appropriate and needs further investigation. Therefore, it is always required to check the coherence between $\widehat{\boldsymbol{\omega}}_{(l)}$ before further analysis.

To sum up, the steps for systematically comparing the L RNA-Seq normalization methods can be described as follows:

(1) Conduct qRT–PCR, Microarray and RNA-Seq experiments using the same RNA sample to generate the raw expression data of n genes.

(2) Apply the standard normalization methods to generate the normalized qRT–PCR data X and the normalized Microarray data Y. Apply the L RNA-Seq normalization methods to the RNA-Seq raw data separately and generate L sets of normalized RNA-Seq data $Z_{(l)}$ $(l = 1, \ldots, L)$. Apply the log-2 transformation to X, Y and $Z_{(l)}$ $(l = 1, \ldots, L)$ and still denote the resulting data as X, Y and $Z_{(l)}$ $(l = 1, \ldots, L)$ for simplicity of discussion.

(3) Apply the system of ME models to each combination of X, Y and $Z_{(l)}$ and calculate the estimates of the structural parameters $\widehat{\boldsymbol{\theta}}_{(l)}$ and $\widehat{\boldsymbol{\omega}}_{(l)}$.

(4) Check the coherency of $\widehat{\boldsymbol{\omega}}_{(l)}$ for different l.

(5) Compare the L RNA-Seq normalization methods by comparing $\widehat{\boldsymbol{\theta}}_{(l)}$ component-wise and draw conclusions.

2.1 Single versus multiple laboratories

In system (1), the presence of multiple laboratories for a platform enables the separation of the ME due to the platform from the ME due to laboratories. For the purpose of comparing different RNA-Seq normalization methods, however, only RNA-Seq is required to have data generated from multiple laboratories. In the real dataset used later in this article, both Microarray and RNA-Seq data are from multiple laboratories, and qRT–PCR only has data from one laboratory ($k_1 = 1$). Such a scenario is common in transctriptome studies since qRT–PCR is not a massive parallel technology, and its cost is relatively high. When $k_1 = 1$, the ME due to qRT–PCR (κ_{1j}) and the ME due to the laboratory (π_{1j1}) are not separable, and their variances cannot be estimated separately. Instead, the combined variance $\tau^2_1 = \psi^2_1 + \varsigma^2_1$ can be estimated, and $\boldsymbol{\omega}$ needs to be modified to $\boldsymbol{\omega} = \left[\alpha_2, \beta_2, \tau^2_1, \psi^2_2, \varsigma^2_2\right]^\top$; see Z.Sun *et al.* (unpublished manuscript) for more details. This fortunately does not affect the estimation of the other structural parameters, in particular, those in $\boldsymbol{\theta}_{(l)}$, which are related to the quality of the normalized RNA-Seq measurements, and thus does not affect the comparison between different RNA-Seq normalization methods.

2.2 Correlation coefficients

Next, we show that treating qRT–PCR or Microarray as the gold standard and using correlation coefficient to assess the performance of a RNA-Seq normalization method can be insensitive and non-informative. Suppose qRT–PCR is used as the gold standard. Let $\bar{X}_{\cdot k}$ be the grand mean of all qRT–PCR measurements in the kth laboratory, and $\bar{Z}_{\cdot k',(l)}$ the grand mean of the RNA-Seq measurements in the k'th laboratory. For convenience, in the following discussion, we use qRT–PCR data from one laboratory and RNA-Seq data from one laboratory as an example. Under the system of ME models, the correlation coefficient of qRT–PCR measurements and RNA-Seq measurements normalized by the lth method is

$$r_{xz,l} = \frac{\sum_{j}\left(Z_{jk',(l)} - \bar{Z}_{.k',(l)}\right)\left(X_{jk} - \bar{X}_{.k}\right)}{\sqrt{\sum\left(\bar{Z}_{jk',(l)} - \bar{Z}_{.k',(l)}\right)^{2}\sum\left(X_{jk} - \bar{X}_{.k}\right)^{2}}} \rightarrow \frac{\Delta}{\sqrt{\left(\Delta + \frac{\psi^{2}_{3,(l)}+\varsigma^{2}_{3,(l)}}{\beta^{2}_{3,(l)}}\right)\left(\Delta + \psi^{2}_{1} + \varsigma^{2}_{1}\right)}} \tag{2}$$

where $\Delta = n^{-1}\sum_{j}\left(\mu_j - \bar{\mu}\right)^{2}$, which can be considered the variance of μ_js, and → means convergence in probability as $n \rightarrow \infty$. First, Δ exists in both the denominator and the numerator of $r_{xz,l}$. When Δ is dominant in magnitude, $r_{xz,l}$ can be insensitive to the other terms. In the cases where Δ is extremely large, $r_{xz,l}$ is close to 1, and the correlation coefficients fail to discriminate between different RNA-Seq normalization methods. Second, since Δ, ψ^{2}_{1} and ς^{2}_{1} do not depend on l, the differences between $r_{xz,l}$ for different l are determined by $\eta = \psi^{2}_{3,(l)} + \varsigma^{2}_{3,(l)}/\beta^{2}_{3,(l)}$, which is a function of the variances and biases of the lth normalization method. Therefore, the correlation coefficient-based approach for comparing different normalization methods does not provide detailed information about the variances and biases of the compared methods, and thus it is not as informative as the proposed approach.

3 DATASETS

To compare the performances of the five existing RNA-Seq normalization methods mentioned in Section 1, we used the qRT–PCR, RNA-Seq and Microarray expression data of two RNA samples from the MAQC and SEQC projects. The two RNA samples are the human brain reference RNA sample (Brain) and the human universal reference RNA sample (UHR). Their qRT–PCR, Microarray and RNA-Seq data were generated by TaqMan® Gene Expression Assays, Affymetrix U133 Plus2.0 and Illumina Genome Analyzer, respectively.

3.1 Real-time reverse transcription polymerase chain reaction

The TaqMan qRT–PCR data were generated by one laboratory. It can be downloaded at Gene Expression Omnibus (GEO) under series number GSE5350. In total, 1001 genes were measured in each of the UHR sample and Brain samples, and each gene was measured in four replicates. The genes with multiple entries in the original datasets were removed to avoid ambiguity.

3.2 RNA-Seq

The RNA-Seq data were produced from two different experiments, which were conducted in two different laboratories. Data from the first experiment (Bullard *et al.*, 2010) can be downloaded from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number SRA010153, and data from the second experiment can be downloaded under accession number SRA008403 from the same website. In the following discussion, we use the accession numbers to refer to these two datasets. Short reads from SRA010153 are 35 base-pair long, and reads from SRA008403 are 31 base-pair long. We aligned the short reads to UCSC hg19 reference genome using Bowtie 0.12.7 (Langmead *et al.*, 2009). We allowed up to two mismatches and only counted the uniquely mapped reads. The mapped reads were then assigned to genes using the Reference Sequence (RefSeq) database. The region of each gene was defined as the union of all its exon regions. To avoid confusion, we only kept the non-overlapping genes. At the end, for SRA010153, we obtained 2–4 million reads per lane, and for SRA008403, we obtained 0.7–0.8 million reads per lane. The multiple lanes in the RNA-Seq experiments were treated as the technical replicates. In SRA010153, both the UHR and Brain samples have seven replicates; In SRA008403, the Brain sample has seven replicates and the UHR sample has three replicates. The numbers of replicates for the UHR sample were different in the two laboratories. This unevenness will much complicate the subsequent analysis. Therefore, we decided to use three lanes of SRA010153 for the UHR sample. We tried different combinations of three lanes and found the analysis results were similar. Therefore, we only report the results based on the first three lanes of SRA010153 later on in the article. For each of the UHR and Brain samples, we only kept those genes that receive at least one read in each lane of the two RNA-Seq experiments.

RNA-Seq reads data were then processed using five normalization methods, RPKM, GPseq, mseq, GCR and PMseq. We use $l = 1$, 2, 3, 4 and 5 as the indices for these five methods and $Z_{(l)}$ to denote the normalized RNA-Seq data by the lth method. The R package 'GPseq' (Srivastava and Chen, 2011) was used to obtain $Z_{(2)}$. The R package 'mseq' (Li, 2011) was used to obtain $Z_{(3)}$. When using mseq, the top 100 single isoform non-overlapping genes with highest RPKM expression levels were used in training, and the neighborhood of a base-pair was defined to include 25 nucleotides to its left and 15 nucleotides to its right, and model fitting and prediction were performed in a lane-by-lane fashion. We used the regression normalization method in Risso *et al.* (2011) to obtain $Z_{(4)}$, where the GC content of each gene was calculated using the RefSeq Genome database. A R script provided by the authors of Wu *et al.* (2012) was used to obtain the PMseq normalized data $Z_{(5)}$.

3.3 Microarray

The Affymetrix Microarray data were generated by five different laboratories. The probe level data were normalized by Probe logarithmic Intensity Error (PLIER) method, and it can be downloaded from GEO under series number GSE5350. Probes were mapped to genes using their RefSeq IDs. The mean of the expression levels of the probes mapped to gene j is used as the expression level of gene j.

3.4 Standardization

For measurements from all three platforms, the log-2 transformation was applied to generate the transformed data. We use genes that overlap in all laboratories of all three platforms, and genes with extremely low and extremely high expression level (i.e. qRT–PCR expression measurements below –6 or above 4 in log-2 scale) were excluded. At the end, the UHR sample included 477 genes, and the Brain sample included 409 genes. For each l, the normalized RNA-Seq data $Z_{(l)}$ by the lth method were further standardized in a laboratory-by-laboratory manner as follows. $\tilde{Z}_{jk,(l)} = Z_{jk,(l)} - \bar{Z}_{\cdot k,(l)}/S_{Zk,(l)}\ S_X + \bar{X}_{.}$, where $\bar{X}_{.}$ and S_X denote the sample mean and sample standard deviation of all qRT–PCR measurements, $\bar{Z}_{.k,(l)}$ and $S_{Z_{k,(l)}}$ denote the sample mean and sample standard deviation of the normalized RNA-Seq measurements in the kth laboratory, and $\tilde{Z}_{jk,(l)}$ is the standardized RNA-Seq measurement. The Microarray data from each laboratory were standardized in a same way. After the standardization, all the resulting expression level measurements were of the same scale.

4 RESULTS

We applied the system of ME models to analyze the five combinations of X, Y and $Z_{(l)}$ for $l = 1, \ldots, 5$ separately to compare the five RNA-Seq normalization methods. Estimates of the structural parameters were divided into two groups, denoted by $\widehat{\boldsymbol{\theta}}_{(l)}$ and $\widehat{\boldsymbol{\omega}}_{(l)}$, according to the discussion in Section 2. In this section, we first report diagnostic results regarding the assumptions imposed on the MEs. Second, we examine the coherence of $\widehat{\boldsymbol{\omega}}_{(l)}$'s for different l's. Third, we provide detailed comparison of the five RNA-Seq normalization methods based on $\widehat{\boldsymbol{\theta}}_{(l)}$. At the end of this section, we show that the comparison between the five normalization methods based on correlation coefficients is insensitive and non-informative.

4.1 Diagnostics of model assumptions

The system of ME models assumes that the involved MEs are normal random variables with homoscedastic variances, that is, it imposes the normality and homoscedasticity assumptions on the MEs. We use residuals and diagnostic tools to check these two assumptions in the subsequent sections. After the system of ME models is fitted, two types of residuals can be calculated, which are the residuals corresponding to the MEs due to a platform (platform residuals) and the residuals corresponding to the MEs due to a laboratory (laboratory residuals). The platform residuals for qRT–PCR, Microarray and RNA-Seq can be calculated by $e_{\kappa,1j} = \bar{X}_{j.} - \hat{\mu}_j, e_{\kappa,2j} = \bar{Y}_{j.} - \hat{\alpha}_2 - \hat{\beta}_2\hat{\mu}_j$ and $e_{\kappa,3j} = \bar{Z}_{j.} - \hat{\alpha}_3 - \hat{\beta}_3\hat{\mu}_j$, respectively. The laboratory residuals in Microarray and RNA-Seq can be calculated by $e_{\pi,2jk} = Y_{jk} - \bar{Y}_j$ and $e_{\pi,3jk} = Z_{jk} - \bar{Z}_j$. To check the normality assumption, we generated the QQ plots for the residuals and did not detect significant violation of the assumption. To check the homoscedasticity assumption, we generated residual plots and constructed approximate 95% confidence intervals for Box–Cox transformations. Because the diagnostic results are similar for different normalization methods, we only present those for the normalized RNA-Seq data of the Brain sample by RPKM here as an example. The platform residual plots corresponding to qRT–PCR, Microarray and RNA-Seq are presented in Figure 1 from left to right. The plots do not demonstrate strong heteroscedastic patterns, and their corresponding 95% confidence intervals for Box–Cox transformation all contain 1, indicating that there does not exist significant violation of the homoscedasticity assumption on the platform MEs. Note that due to limited space, the confidence intervals are not reported here. The laboratory residual plots corresponding to Microarray and RNA-Seq are presented in Figure 2 from left to right. Overall, these two residual plots again do not demonstrate severe heteroscedastic patterns. The approximate 95% confidence interval of Box–Cox transformation for the laboratory residuals under RNA-Seq contains 1, suggesting that the laboratory ME due to laboratory in RNA-Seq does not violate the homoscedasticity assumption. The approximate 95% confidence interval of Box–Cox transformation for the laboratory residuals under Microarray does not include 1, though the low end of the interval is close to 1, indicating that some heteroscedasticity may exist in the laboratory ME under the Microarray platform.

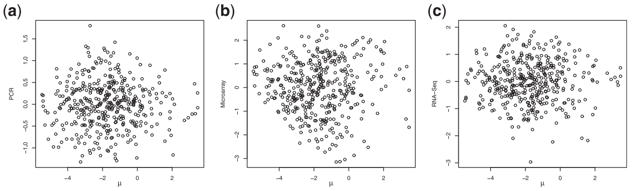

Fig. 1. Plots of platform residuals for qRT–PCR (**a**), Microarray (**b**) and RNA-Seq (**c**) based on the RPKM measurements in the Brain sample

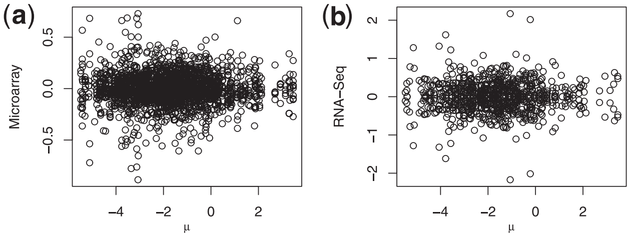

Fig. 2. Plots of laboratory residuals for Microarray (**a**) and RNA-Seq (**b**) based on the RPKM measurements in the Brain sample

Note that the system of ME models are postulated for the log-2 scale of the original expression measurements. As discussed in Section 2, log-2 transformation in general can mitigate heteroscedasticity and further lead to homoscedastic errors. The diagnostic results earlier in the text suggest that this was indeed the case for most of the MEs except those due to laboratories under Microarray. Comparing the size of the residuals due to laboratories under microarray and the sizes of other residuals, we found the former is about one-fourth of the latter, indicating that the MEs due to laboratories under Microarray are relatively small. The purpose of this study is to compare the performances of RNA-Seq normalization methods, and the homoscedasticity assumption holds well for all MEs related to RNA-Seq. These two facts lead us to believe that the minor violation of the homoscedasticity assumption does not affect the usefulness of the proposed model and approach and alter the validity of conclusions we present in the subsequent sections. More complicated models can be considered, for example, models based on other power transformations and models allowing heteroscedastic errors. We will investigate these extensions and report the results in the future.

4.2 Coherency between $\widehat{\omega}_{(l)}$

Each parameter in $\boldsymbol{\omega}$ has five estimates obtained from the five combined data. Table 1 lists $\widehat{\boldsymbol{\omega}}_{(l)}$ and their standard errors (in parentheses) for $l = 1, \ldots, 5$ for the Brain sample. For each parameter, we applied pairwise comparison with the help of the bootstrap procedure to check if its five estimates are coherent. Components of $\hat{\omega}_{(l)}$ for different l were not found to be significantly different, indicating that the five systems are coherent for the Brain sample. Similarly, we also found that elements of $\hat{\omega}_{(l)}$ in the five systems for the UHR sample were also coherent with each other.

Table 1. Estimates of structural parameters related to qRT–PCR and Microarray ($\hat{\omega}_{(l)}$) in the Brain sample

	$\hat{\alpha}_2$	$\hat{\beta}_2$	$\hat{\tau}_1^2$	$\hat{\psi}_2^2$	$\hat{\varsigma}_2^2$
RPKM	–0.2693 (0.0838)	0.8183 (0.0280)	0.6269 (0.1145)	1.4899 (0.1263)	0.0245 (0.0008)
GPseq	–0.2727 (0.0852)	0.8164 (0.0295)	0.6207 (0.1181)	1.4940 (0.1279)	0.0245 (0.0008)
Mseq	–0.2755 (0.0940)	0.8147 (0.0373)	0.6155 (0.1388)	1.4975 (0.1368)	0.0245 (0.0008)
GCR	–0.3135 (0.0853)	0.7927 (0.0299)	0.5439 (0.1213)	1.5438 (0.1303)	0.0245 (0.0008)
PMseq	–0.2571 (0.0816)	0.8254 (0.0256)	0.6490 (0.1085)	1.4750 (0.1234)	0.0245 (0.0008)

Table 2. Estimates of structural parameters related to RNA-Seq ($\hat{\theta}_{(l)}$) in the Brain sample (best estimates are highlighted)

	$\hat{\alpha}_3$	$\hat{\beta}_3$	$\hat{\psi}_3^2$	$\hat{\varsigma}_3^2$
RPKM	–0.1645 (0.0824)	0.8926 (0.0338)	0.8017 (0.1046)	0.1664 (0.0082)
GPseq	–0.1910 (0.0836)	0.8755 (0.0343)	0.8599 (0.1071)	0.1704 (0.0084)
Mseq	–0.3597 (0.0819)	0.7237 (0.0336)	0.8401 (0.1009)	0.4033 (0.0199)
GCR	–0.2222 (0.0835)	0.8562 (0.0346)	0.9174 (0.1096)	**0.1585** (0.0078)
PMseq	**–0.1053** (0.0796)	**0.9127** (0.0326)	**0.6583** (0.0978)	0.1967 (0.0097)

Table 3. Estimates of structural parameters related to RNA-Seq ($\hat{\theta}_{(l)}$) in the UHR sample (best estimates are highlighted)

	$\hat{\alpha}_3$	$\hat{\beta}_3$	$\hat{\psi}_3^2$	$\hat{\varsigma}_3^2$
RPKM	–0.0748 (0.0633)	0.9707 (0.0172)	0.7331 (0.0838)	0.0632 (0.0029)
GPseq	–0.1013 (0.0652)	0.9545 (0.0176)	0.8273 (0.0887)	0.0738 (0.0034)
Mseq	–0.4071 (0.0705)	0.6771 (0.0195)	1.1220 (0.1110)	0.6712 (0.0307)
GCR	–0.1579 (0.0641)	0.9273 (0.0177)	0.8681 (0.0889)	**0.0513** (0.0023)
PMseq	**–0.0214** (0.0603)	**0.97363** (0.0163)	**0.5689** (0.0748)	0.0911 (0.0042)

4.3 Comparison of normalization methods

Components in $\boldsymbol{\theta}_{(l)}$ are related to the quality of RNA-Seq measurements normalized by the lth method, therefore they can be used to compare different RNA-Seq normalization methods. $\widehat{\boldsymbol{\theta}}_{(l)}$ for $l = 1, \ldots, 5$ in the Brain and UHR samples are listed in Tables 2 and 3, respectively. The standard errors are also reported in parentheses in the tables.

The platform variance $\psi^2_{3,(l)}$ represents the remaining variation due to the RNA-Seq platform in the normalized RNA-Seq measurements by the lth method. In the Brain sample, in terms of the platform variance, the order of the five normalization methods from best to worst is PMseq, RPKM, mseq, GPseq and GCR. In the UHR sample, the order is PMseq, RPKM, GPseq, GCR and mseq. PMseq has the smallest platform variances, which are more than 18% and 22% lower than the platform variances of RPKM.

The laboratory variance $\varsigma^2_{3,(l)}$ represents the remaining variation due to laboratories in the normalized RNA-Seq measurements by the lth method. In both the Brain and UHR samples, in terms of the laboratory variance, the order of the five methods from best to worst is GCR, RPKM, GPseq, PMseq and mseq.

In terms of total variance, the order of the five methods from the best to the worst is the same in both the Brain and UHR samples. PMseq has the lowest total variance, which are 0.8550 in the Brain sample and 0.6600 in the UHR sample; RPKM has the second lowest total variances, which are 0.9681 in the Brain sample and 0.7963 in the UHR sample; the third is GPseq, which has a total variance of 1.0303 in the Brain sample and 0.9011 in the UHR sample; GCR is ranked the fourth, with total variance of 1.0759 in the Brain sample and 0.9194 in the UHR sample and mseq has the largest total variance in both samples, which are 1.2434 and 1.7923 in the Brain and UHR samples, respectively.

The absolute value of $\alpha_{3,(l)}$ represents the magnitude of the location bias in the normalized RNA-Seq measurements by the lth method. The order of the five methods in terms of location bias in both samples is PMseq, RPKM, GPseq, GCR and mseq. In the Brain sample, one fails to reject the null hypothesis $H_0 : \alpha_{3,(l)} = 0$ based on $\hat{\alpha}_{3,(l)}$ and its standard error for PMseq, which implies the RNA-Seq measurements normalized by PMseq are free of location bias in the Brain sample. In the UHR sample, RNA-Seq measurements normalized by PMseq, RPKM and GPseq are free of location biases.

Table 4. Pearson correlation coefficients between RNA-Seq and qRT–PCR measurements

	RPKM	GPseq	Mseq	GCR	PMseq
Brain	0.749	0.738	0.675	0.740	0.766
UHR	0.823	0.809	0.672	0.811	0.836

Table 5. Spearman correlation coefficients between RNA-Seq and qRT–PCR measurements

	RPKM	GPseq	Mseq	GCR	PMseq
Brain	0.711	0.703	0.626	0.717	0.739
UHR	0.806	0.801	0.587	0.798	0.823

The absolute value of $\beta_{3,(l)} - 1$ represents the magnitude of the scale bias in the normalized RNA-Seq measurements by the lth method. In terms of the scale bias, the order of the five methods in both the Brain and UHR samples are PMseq, RPKM, GPseq, GCR and mseq. In the UHR sample, one fails to reject the null hypothesis $H_0 : \beta_{3,(l)} = 1$ based on $\hat{\beta}_{3,(l)}$ and its standard error for PMseq and RPKM, which implies that the RNA-Seq measurements normalized by these two methods are free of scale biases in the UHR sample. In the brain sample, RNA-Seq measurement normalized by all five methods suffers from scale biases.

According to the estimated location biases and scale biases of the five methods, the PMseq is the best, followed by RPKM. The third best method is GPseq. The fourth is GCR, and mseq measurements suffer from the largest biases.

Based on the discussion earlier in the text, we can draw the conclusion about the overall performances of the five methods. PMseq has the best performance in this dataset because it is ranked the first both in terms of total variance and in terms of bias. RPKM has the second best performance because it is ranked the second both in terms of total variance and in terms of bias. The third and fourth methods are GPseq and GCR. The mseq method has the worst performance due to the fact that it is ranked the last in terms of both total variance and bias.

4.4 ME models versus correlation coefficients

Next, we show that the conventional approach for comparing different normalization methods can be less effective and less informative than our proposed approach. We treated the qRT–PCR as the gold standard and calculated two types of correlation coefficients (Pearson's and Spearman's) of the qRT–PCR measurements with the normalized RNA-Seq measurements by the five methods. The Pearson's correlation coefficients are listed in Table 4 and the Spearman's correlation coefficients are listed in Table 5.

The ranks of the correlation coefficients are used to rank-order the RNA-Seq normalization methods. According to the Spearman's correlation coefficients, the order from best to worst in the Brain sample is PMseq, GCR, RPKM, GPseq and mseq. However, on the UHR sample, the order of GCR and RPKM are switched. In each sample, the absolute values of Spearman's correlation coefficients for these two methods only have minor differences. This demonstrates the insensitivity of the correlation coefficients when used for comparing the RNA-Seq normalization methods.

According to the Pearson's correlation coefficients, in both the Brain and UHR samples, the order from best to worst is PMseq, RPKM, GCR, GPseq and mseq. However, the correlation coefficients of GPseq and GCR only have minor differences. Moreover, the two types of correlation coefficients do not provide any further information about the strengths and weaknesses of these normalization methods.

5 CONCLUSION

RNA-Seq raw data are subject to a variety of variations, biases and uncertainties, and therefore, they need to be normalized for subsequent gene expression analysis. The conventional approach to comparing different normalization methods requires a gold standard and uses the correlation coefficients between the normalized data and the gold standard. In this article, we show that such an approach can be insensitive and non-informative. We proposed to use ME models to characterize the normalized RNA-Seq gene expression data and use the estimates of the model parameters for systematic comparison of different normalization methods. In the proposed approach, no single platform is treated as the gold standard.

We applied the proposed approach to compare five existing RNA-Seq normalization methods using the gene expression data of the UHR and Brain samples from the MAQC and SEQC projects and showed that much more information about the pros and cons of the methods could be obtained than the conventional approach. We believe that the proposed method not only can help analysts in practice to choose between the existing normalization methods but also help researchers improve these methods or develop entirely new methods.

ACKNOWLEDGEMENTS

We thank the associate editor and three anonymous reviewers for their constructive comments and suggestions.

Funding: NSF-DMS-1000443.

Conflict of Interest: none declared.

REFERENCES

Agarwal,A. *et al.* (2010) Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, **11**, 383.

Barnett,V.D. (1969) Simultaneous pairwise linear structural relationships. *Biometrics*, **25**, 129–142.

Brownie,J. *et al.* (1997) The elimination of primer-dimer accumulation in PCR. Nucleic Acids Res., **25**, 3235–3241.

Bullard,J. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, **5**, 613–619.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-Sequencing data. *Bioinformatics*, **25**, 3207–3212.

Hansen,K.D. *et al.* (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res., **38**, e131.

Irizarry,R.A. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods*, **2**, 345–350.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., **10**, R25.

Li,J. (2011) *mseq: Modeling non-uniformity in short-read rates in RNA-Seq data*. R package version 1.2. http://CRAN.R-project.org/package = mseq.

Li,J. *et al*. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol., **11**, R50.

Marioni,J. *et al*. (2008) RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res., **18**, 1509–1517.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–628.

Nagalakshmi,U. *et al*. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Pan,Q. *et al*. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet., **40**, 1413–1415.

Risso,D. *et al*. (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 480.

Srivastava,S. and Chen,L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-Seq data. Nucleic Acids Res., **38**, e170.

Srivastava,S. and Chen,L. (2011) *GPseq: using the generalized Poisson distribution to model sequence read counts from high throughput sequencing experiments*. R package version 0.5. http://CRAN.R-project.org/package = GPseq.

Wu,H. *et al*. (2012) Pm-seq: using finite poisson mixture models for rna-seq data analysis and transcript expression level quantification. *Stat. Biosci. doi:10.1007/s12561-012-9070-9*.