


Gene expression

MIXnorm: normalizing RNA-seq data from formalin-fixed paraffin-embedded samples

Shen Yin^{1,2}, Xinlei Wang ^{1,*}, Gaoxiang Jia¹ and Yang Xie²

¹Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA and ²Department of Population and Data Sciences, Quantitative Biomedical Research Center, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on September 23, 2019; revised on February 21, 2020; editorial decision on February 26, 2020; accepted on February 28, 2020

Abstract

Motivation: Recent studies have shown that RNA-sequencing (RNA-seq) can be used to measure mRNA of sufficient quality extracted from formalin-fixed paraffin-embedded (FFPE) tissues to provide whole-genome transcriptome analysis. However, little attention has been given to the normalization of FFPE RNA-seq data, a key step that adjusts for unwanted biological and technical effects that can bias the signal of interest. Existing methods, developed based on fresh-frozen or similar-type samples, may cause suboptimal performance.

Results: We proposed a new normalization method, labeled MIXnorm, for FFPE RNA-seq data. MIXnorm relies on a two-component mixture model, which models non-expressed genes by zero-inflated Poisson distributions and models expressed genes by truncated normal distributions. To obtain maximum likelihood estimates, we developed a nested EM algorithm, in which closed-form updates are available in each iteration. By eliminating the need for numerical optimization in the M-step, the algorithm is easy to implement and computationally efficient. We evaluated MIXnorm through simulations and cancer studies. MIXnorm makes a significant improvement over commonly used methods for RNA-seq expression data.

Availability and implementation: R code available at <https://github.com/S-YIN/MIXnorm>.

Contact: swang@smu.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Human tissue biospecimens are of two primary types, fresh-frozen (FF) and formalin-fixed paraffin-embedded (FFPE) tissues. As fresh tissues deteriorate rapidly at room temperature, FF samples must be frozen instantly after collection and then stored in freezers. FF tissues are well suited for molecular analysis using gene expression measurements as freezing preserves RNA well. However, they are expensive to store and transport, and difficult to collect for large-scale studies. By contrast, FFPE samples can be stored at room temperature and kept for a long time. Due to the ease of handling and inexpensive storage, numerous FFPE tissue samples have been deposited into tissue banks and pathology laboratories around the world, and are readily available (Perlmutter *et al.*, 2004; Reis *et al.*, 2011; Ripoli *et al.*, 2016). The ubiquity of FFPE tissue specimens has made them an invaluable resource in biomedical research, with great potential for predictive and prognostic biomarker discovery.

However, the quality of RNA extracted from FFPE tissues is a concern due to chemical modifications and continued degradation over time. The process of using formalin to fix and paraffin embedding to preserve tissues for an extended period of time is designed to

well preserve cellular proteins rather than preserving RNA. Consequently, assays using microarray or quantitative polymerase chain reaction (qPCR) often have limited reproducibility and sensitivity when measuring gene expression from such samples. In order to exploit the vast collection of FFPE samples, substantial effort has been devoted to development and/or validation of advanced technologies that can reliably probe their gene expression levels. For high-throughput profiling, RNA-sequencing (RNA-seq), which uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample, is in common use. Recent studies have shown that for a wide variety of human tumor tissues (e.g. bladder, colon, prostate and renal carcinoma), RNA-seq can be used to measure mRNA of sufficient quality extracted from FFPE tissues to provide biologically relevant transcriptome analysis (Graw *et al.*, 2015; Grenier *et al.*, 2017). Meanwhile, recent FFPE RNA-Seq solutions, such as Illumina total RNA-Seq, enable researchers to produce high-quality results from degraded samples. As a result, a drastically increasing number of studies have used RNA-seq on FFPE specimens (e.g. Lin *et al.*, 2014; Morton *et al.*, 2014).

A critical step when analyzing RNA-seq data is normalization. Normalization removes systematic biases that affect measured gene

expression levels (e.g. variability in experimental conditions, sample collection and preparation and machine parameters, etc.), while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription. A number of normalization methods for RNA-seq data have been developed (e.g. Dillies *et al.*, 2013). A common approach is to normalize the measured expression using (estimated) scaling factors. The most straightforward normalization method, Reads Per Million (RPM) (Mortazavi *et al.*, 2008), estimates the scaling factor by dividing the total read count of a sample by 1 000 000. The normalized data are the read counts divided by the scaling factors. The upper quartile (UQ) (Bullard *et al.*, 2010) method estimates the scaling factor by the upper quartile of the read counts within each sample. DESeq (Anders and Huber, 2010) works under the assumption that only a small subset of genes are differentially expressed (DE). First, for each gene, the ratio of its read count over its geometric mean across all samples is calculated. Then, the scaling factor is estimated by the median ratio within each sample. Thus, it is also referred to as median normalization. Trimmed Mean of M-values (TMM) (Robinson and Oshlack, 2010) is also based on the assumption that most of the genes are not DE, where one sample is chosen as the reference sample and the others as test samples. The log ratio of the read count between each test sample and the reference is computed for each gene. Then for each test sample, TMM estimates the scaling factor by the weighted mean of log ratios after exclusion of the genes with extreme average expression or with largest log ratios. PoissonSeq (PS) (Li *et al.*, 2012) models RNA-seq data by a Poisson log-linear model. The normalization is done implicitly by including the scaling factor as a term in the model.

Though a number of normalization methods are available for RNA-seq data, none has been specifically designed for FFPE samples, of which a prominent feature is sparsity (i.e. excessive zero or small counts), caused by RNA degradation in such samples. The quantile-based methods become problematic due to excess zeros that cause ranking ties. For DESeq, the geometric mean is only well defined for genes with at least one read count in every sample. The zero inflation is also a concern for methods that implicitly use scaling factors, such as PS since they all rely on Poisson or negative binomial (NB) distributions for modeling count data.

To illustrate characteristics of RNA-seq data from FFPE samples, we begin by presenting an exploratory analysis in Section 2.1 using a real data example. In Section 2.2, we propose a novel normalization method, called MIXnorm, based on a two-component mixture model for log read counts, to capture the sparsity as well as major mean and variance structures underlying the data. Due to whole-genome sequencing, the number of parameters involved is often very large. We develop an efficient nested expectation-maximization (EM) algorithm to fit the proposed mixture model, where parameters are updated via closed-form solutions iteratively. Section 3 briefly summarizes simulation studies and expounds two real data applications to compare the performance of the proposed MIXnorm to five commonly used RNA-seq normalization methods, including UQ, DESeq, RPM, PS and TMM. Section 4 concludes the article with a brief discussion. Technical details, performance evaluation via simulation and additional analysis results are available through online [Supplementary Material](#).

2 Materials and methods

2.1 An exploratory analysis

As mentioned in the introduction, a striking feature of FFPE RNA-seq data is the sparsity, which can be observed in multiple datasets from independent studies. An example is provided here using paired FF and FFPE samples from a published study, RNA-seq validation of the Complexity INdex in SARComas (CINSARC) prognostic signature (Lesluyes *et al.*, 2016). Prognosis of metastatic outcomes in soft tissue sarcomas is important because of its high recurring rate (up to 50% of recurrence). CINSARC, a gene signature that consists of 67 genes, has been identified as a valuable prognostic factor in sarcomas. This signature was originally identified on FF samples

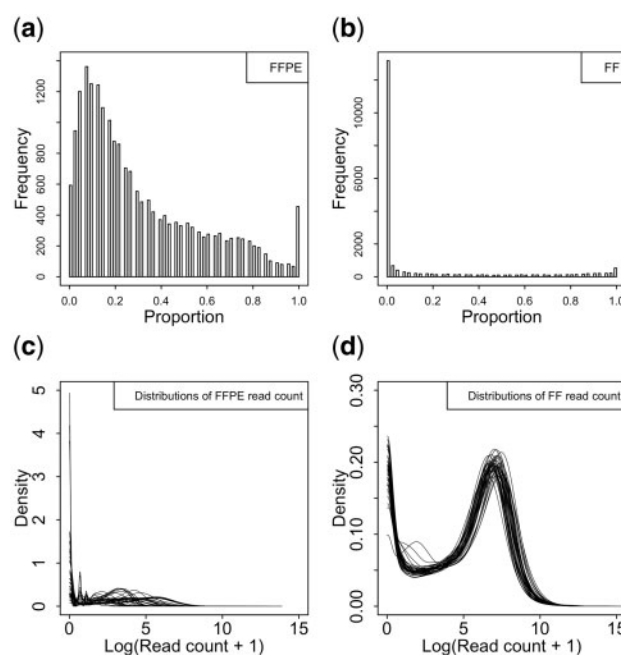


Fig. 1. An exploratory analysis of RNA-seq data in Lesluyes *et al.* (2016). a and b) The histogram of zero-count proportion among 41 FFPE/FF samples (represented by the horizontal axis) based on a total of 20 242 genes. c and d) Empirical densities of log read counts for the 41 FFPE/FF samples. Each curve in (c) and (d) represents the density for one sample across all the 20 242 genes

assayed by the microarray platform. The study goal of Lesluyes *et al.* (2016) was to evaluate the prognostic performance of CINSARC on both FF and FFPE samples. Thus, the resulting dataset contains gene expression levels for 20 242 protein-coding genes, measured by whole-genome NGS on paired FF and FFPE samples from 41 patients, though their primary interest lied on the CINSARC gene signature.

We first transformed the raw read counts in this dataset into the natural logarithm scale. In order to deal with zero counts, we define the log count $L \equiv \log(C + 1)$, where C is the raw count. Figure 1a shows that among a total of 20 242 genes, there is a significant portion of genes with more than 50% zero counts in FFPE samples while Figure 1b shows that over 65% genes, represented by the left-most bar, do not have any zero count in FF samples. Further, Figure 1c and d shows that for each sample, regardless of sample types, the commonly used Poisson or NB distributions for count data are far from being adequate to capture the bimodal density of gene expression (with one spike at zero). Two other interesting observations from Figure 1c and d are: (i) the locations of the distributions of 41 FFPE samples vary much more than those of FF samples, indicating great heterogeneity in RNA degradation levels among the FFPE tissues and (ii) densities from different FF samples show highly similar variability while those from FFPE samples do not (the spread of the curves varies tremendously).

The above findings indicate that existing normalization methods for RNA-seq data, all developed based on FF or like samples, are ill-suited for FFPE samples as they cannot cope with the highly complex features of such data. We proceed to develop a robust yet powerful method, MIXnorm, based on a two-component mixture model to capture the distinct bimodality as well as major mean and variance structures underlying the data. The first component is to model non-expressed genes, whose read counts should be zero or relatively small due to non-specific binding. These genes include biologically zero-expression genes that may exist, or those with low expression but cannot be expressed due to various experimental limitations (e.g. drop-outs), or those that should be expressed but cannot because of high-level mRNA degradation. For the non-expressed genes, we use a zero-inflated Poisson (ZIP) distribution to capture the spike at zero for each sample, of which the Poisson

mean reflects the background noise level. The second component is to model expressed genes, and we use a truncated normal (TN) distribution for log gene read counts of each sample to approximate the roughly bell-shaped curve centered at the second mode.

2.2 The MIXnorm method

2.2.1 The statistical model for FFPE data

Let C_{ij} denotes the raw count of gene j from sample i and $L_{ij} \equiv \log(C_{ij} + 1)$ is the natural logarithm transformed count for $i = 1, \dots, I, j = 1, \dots, J$. We define a latent binary variable D_j : $D_j = 0$ indicates gene j is non-expressed in this study, meaning that observed non-zero counts of gene j are due to background noise; $D_j = 1$ indicates gene j is expressed, with mean expression > 0 . The following mixture model is proposed for FFPE data:

$$C_{ij} \sim \text{ZIP}(\pi_j, \delta_i), \text{ if } D_j = 0, \quad (1)$$

$$\begin{aligned} L_{ij} &\sim \text{TN}(\mu_i, \sigma_i^2, 0, +\infty), \text{ if } D_j = 1, \\ D_j &\sim \text{Bernoulli}(\phi), \end{aligned} \quad (2)$$

where $0 \leq \pi_j, \phi \leq 1, \delta_i, \sigma_i > 0$ for $i = 1, \dots, I, j = 1, \dots, J$. Here, $\text{ZIP}(\pi_j, \delta_i)$ stands for a ZIP distribution, with probability π_j being zero and probability $1 - \pi_j$ being from a Poisson distribution with mean δ_i ; $\text{TN}(\mu_i, \sigma_i^2, 0, +\infty)$ stands for a normal distribution with mean μ_i and variance σ_i^2 , left truncated at zero as $L_{ij} > 0$; and ϕ is the proportion of expressed genes in the study. Figure 1a clearly shows the zero-count proportion varies across different genes, and so π_j is assumed to be gene-specific instead of being constant. The δ_i reflects sample-specific background noise and should be relatively small. Figure 1c shows that the location and spread of L_{ij} both vary a lot from sample to sample, meaning that the sample-specific mean μ_i and variance σ_i^2 are necessary for FFPE data. We note that L_{ij} is a discrete random variable with support $\{0, \log(1), \log(2), \dots\}$, but in (2), a continuous distribution is used to approximate the discrete distribution of L_{ij} .

Let $\Theta = (\pi, \delta, \mu, \sigma, \phi)$ denotes the collection of all the parameters in the mixture model, where $\pi = (\pi_1, \dots, \pi_J)$, $\delta = (\delta_1, \dots, \delta_I)$, $\mu = (\mu_1, \dots, \mu_I)$ and $\sigma = (\sigma_1, \dots, \sigma_I)$. The (incomplete) likelihood function is

$$\begin{aligned} L(\Theta|C) &= \prod_{j=1}^J p(C_j|\Theta) \\ &= \prod_{j=1}^J [p(C_j|D_j = 1, \mu, \sigma)p(D_j = 1|\phi) \\ &\quad + p(C_j|D_j = 0, \pi_j, \delta)p(D_j = 0|\phi)] \\ &= \prod_{j=1}^J \left[\prod_{i=1}^I p(C_{ij}|D_j = 1, \mu_i, \sigma_i) \cdot \phi \right. \\ &\quad \left. + \prod_{i=1}^I p(C_{ij}|D_j = 0, \pi_j, \delta_i) \cdot (1 - \phi) \right], \end{aligned}$$

where $p(C_{ij}|D_j = 0, \pi_j, \delta_i)$ is the probability mass function (PMF) of C_{ij} of non-expressed genes, i.e. the ZIP distribution in (1); $p(C_{ij}|D_j = 1, \mu_i, \sigma_i)$ is the PMF of C_{ij} for expressed genes, which will be approximated by a probability density function (PDF) with $\log(C_{ij} + 1)$ following the TN distribution on $[0, +\infty)$ in (2). See Web Appendix A in the Supplementary Material for a detailed justification about the validity of using the PDF to approximate the PMF.

2.2.2 Model fitting via an EM algorithm

A common method for estimating parameters of a model with a latent variable structure is to employ an EM algorithm (Dempster et al., 1977) to obtain their maximum likelihood estimates (MLEs). The complete-data log-likelihood with the latent variables D is given by

$$\begin{aligned} \ell(\Theta|C, D) &= \sum_{j=1}^J \log p(C_j, D_j|\Theta) \\ &= \sum_{j=1}^J D_j \cdot \{\log(\phi) + \log[p(C_j|D_j = 1, \mu, \sigma)]\} \\ &\quad + \sum_{j=1}^J (1 - D_j) \cdot \{\log(1 - \phi) + \log[p(C_j|D_j = 0, \pi_j, \delta)]\}. \end{aligned} \quad (3)$$

Let $\Theta^{(t)} = (\pi^{(t)}, \delta^{(t)}, \mu^{(t)}, \sigma^{(t)}, \phi^{(t)})$ be the parameter estimates at the t th iteration. The distribution of D given the observed data C and the current parameter estimates $\Theta^{(t)}$ is

$$p(D|C, \Theta^{(t)}) = \prod_{j=1}^J \frac{p(C_j, D_j|\Theta^{(t)})}{p(C_j|\Theta^{(t)})} = \prod_{j=1}^J \left(w_j^{(t)} \right)^{D_j} \left(1 - w_j^{(t)} \right)^{1-D_j},$$

where

$$w_j^{(t)} = \frac{\phi^{(t)} p(C_j|D_j = 1, \mu^{(t)}, \sigma^{(t)})}{\phi^{(t)} p(C_j|D_j = 1, \mu^{(t)}, \sigma^{(t)}) + (1 - \phi^{(t)}) p(C_j|D_j = 0, \pi_j^{(t)}, \delta^{(t)})}. \quad (4)$$

Each iteration of an EM algorithm consists of two steps, the expectation (E) step and the maximization (M) step. The E-step calculates the expected complete-data log-likelihood given C and $\Theta^{(t)}$, where the expectation is taken over the latent variables D . Since $\ell(\Theta|C, D)$ in (3) is linear in D_j , and $E(D_j|C, \Theta^{(t)}) = w_j^{(t)}$, we have

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) &= E_{D|C, \Theta^{(t)}} \ell(\Theta|C, D) \\ &= \sum_{j=1}^J (1 - w_j^{(t)}) [\log(1 - \phi) + \log p(C_j|D_j = 0, \pi_j, \delta)] \\ &\quad + \sum_{j=1}^J w_j^{(t)} [\log(\phi) + \log p(C_j|D_j = 1, \mu, \sigma)]. \end{aligned} \quad (5)$$

In essence, the E-step calculates the conditional expectation of D given C and $\Theta^{(t)}$. The M-step updates the parameter estimates by maximizing the expected log-likelihood (5). Note that (5) can be maximized with respect to $\phi, (\mu, \sigma)$ and (π, δ) separately. The updated parameter estimates in the $(t + 1)$ th iteration are given by

$$\begin{aligned} \phi^{(t+1)} &= \frac{\sum_{j=1}^J w_j^{(t)}}{J}, \\ (\mu_i^{(t+1)}, \sigma_i^{(t+1)}) &= \underset{\mu_i, \sigma_i}{\operatorname{argmax}} \sum_{j=1}^J \log \text{TN}(L_{ij}|\mu_i, \sigma_i, 0, \infty) \cdot w_j^{(t)}, \\ &\quad i = 1, \dots, I, \end{aligned} \quad (6)$$

$$(\pi^{(t+1)}, \delta^{(t+1)}) = \underset{\pi, \delta}{\operatorname{argmax}} \sum_{j=1}^J \sum_{i=1}^I \left[\log \text{ZIP}(C_{ij}|\pi_j, \delta_i) \cdot (1 - w_j^{(t)}) \right], \quad (7)$$

where the maximization in (7) has constraints $\pi_j \in [0, 1]$ and $\delta_i > 0$; $\text{TN}(\cdot|\cdot)$ stands for the PDF of the TN distribution, $\text{ZIP}(\cdot|\cdot)$ stands for the PMF of the ZIP distribution, both with distributional parameters specified after ‘—’. The update for $\phi^{(t+1)}$ has a closed form. Other parameters can be updated by a Newton–Raphson type method numerically within each iteration t .

The PMF $\text{ZIP}(C_{ij}|\pi_j, \delta_i)$ in (7) cannot be factored into functions of π_j and δ_i . Therefore, the update of (π, δ) involves multi-dimensional optimization, which can be computationally intensive when $I + J$ is large, as is typical for high-throughput profiling, such as RNA-seq. Another drawback of the above algorithm is numerical instability due to the use of the Newton–Raphson method for an approximate solution in the M-step. Dempster et al. (1977) proved that for an EM-type algorithm, the (incomplete) likelihood in every iteration never decreases as t increases. Thus, the incomplete likelihood is typically used to monitor the convergence of the algorithm.

However, this monotone convergence property does not necessarily hold if the E- or M-step is not computed exactly. In such situations, the incomplete log-likelihood may fluctuate around a fixed point for a long time. Due to this instability, when applying the above EM algorithm to real data, we observed that it would not converge, especially when a small tolerance value is selected to terminate the iterative process.

2.2.3 Review of nested EM algorithms

van Dyk (2000) described how nesting two or more EM algorithms could take advantage of closed-form conditional expectations and lead to algorithms with both ease of implementation and computing efficiency (i.e. fast and stable convergence). Assume the missing data can be split into two (or more) sets $Y_{\text{mis } 1}$ and $Y_{\text{mis } 2}$ such that the complete data can be expressed by $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis } 1}, Y_{\text{mis } 2})$, where $Y_{\text{mis } 1}$ and $Y_{\text{mis } 2}$ can be introduced under a data augmentation scheme to aid the computation. Let Θ denotes the vector of all parameters involved, and \mathcal{H} is the parameter space. Define the nested conditional expectation of log-likelihood by

$$\bar{Q}(\Theta|\Theta_1, \Theta_2) = E\{E[l(\Theta|Y_{\text{obs}}, Y_{\text{mis } 1}, Y_{\text{mis } 2})|Y_{\text{obs}}, Y_{\text{mis } 1}, \Theta_1]|Y_{\text{obs}}, \Theta_2\}, \quad (8)$$

where Θ_1 and Θ_2 denote different realizations of Θ , and $\bar{Q}(\Theta|\Theta_1, \Theta_2)$ is a function on $\mathcal{H} \times \mathcal{H} \times \mathcal{H}$. The outer expectation in (8) is taken with respect to $Y_{\text{mis } 1}$ whereas the nested inner expectation is taken with respect to $Y_{\text{mis } 2}$. According to van Dyk (2000), the t th iteration of a nested EM algorithm repeats the following cycle K times.

Cycle k for $k = 1, \dots, K$:

E-step: compute

$$\bar{Q}^{\sim}(\Theta|\Theta^{(t+\frac{k-1}{K})}, \Theta^{(t)}) = E\left\{aE\left[l(\Theta|Y_{\text{com}})|Y_{\text{obs}}, Y_{\text{mis } 1}, \Theta^{(t+\frac{k-1}{K})}\right]|Y_{\text{obs}}, \Theta^{(t)}\right\},$$

M-step: update the parameter estimates by

$$\Theta^{(t+\frac{k}{K})} = \arg \max_{\Theta} \bar{Q}^{\sim}(\Theta|\Theta^{(t+\frac{k-1}{K})}, \Theta^{(t)}).$$

Upon completion of the K th cycle, set $\Theta^{(t+1)} = \Theta^{(t+\frac{K}{K})}$. That is, run K cycles of the inner EM algorithm for each iteration of the outer EM.

When the missing data structure is complex, direct calculation of $E[l(\Theta|Y_{\text{com}})|Y_{\text{obs}}, \Theta^{(t)}]$ is usually difficult. Moreover, we may not be able to directly sample from $p(Y_{\text{mis } 1}, Y_{\text{mis } 2}|Y_{\text{obs}}, \Theta)$, and thus a Monte-Carlo EM algorithm is not feasible as well. A nested EM algorithm takes advantages of subdividing the missing data so that $p(Y_{\text{mis } 1}|Y_{\text{obs}}, \Theta)$ and $p(Y_{\text{mis } 2}|Y_{\text{obs}}, Y_{\text{mis } 1}, \Theta)$ are both known distributions or easy to sample directly. Theoretical properties of nested EM algorithms have been well studied. Theorem 1 in van Dyk (2000) guarantees that, like EM algorithms, nested EM algorithms enjoy the monotone convergence property, and so the incomplete-data likelihood $p(Y_{\text{obs}}|\Theta)$ can be used to detect convergence.

2.2.4 Model fitting via a nested EM algorithm

Below, we introduce additional latent variables so that a nested EM-type algorithm can be constructed to improve computational efficiency. Based on Lambert (1992), a ZIP distribution can be thought of as a mixture of two states, the perfect zero state and the Poisson state. Suppose, we knew which zeros came from the perfect zero state and which came from the Poisson state. That is, for a non-expressed gene j , we define $Z_{ij} = 1$ when C_{ij} is from the perfect zero state and $Z_{ij} = 0$ when C_{ij} is from the Poisson state, for $i = 1, \dots, I$. Obviously, $Z_{ij}|D_j = 0 \sim \text{Bernoulli}(\pi_j)$. Further, we augment the TN data by (hypothesized) missing observations, which borrows ideas from Tanner and Wong (1987) and McLachlan and Jones (1988). That is, the augmented data follow a normal distribution so that the posterior distributions of the parameters or their functions are

straightforward to calculate. For sample i , apart from the observed J genes, there are T_i unobserved genes with $D_i = 1$ and their log count $L_{ij} < 0$, $j = J+1, \dots, J+T_i$, such that $L_{ij} \sim N(\mu_i, \sigma_i)$, for $j = 1, \dots, J+T_i$. Here, the number of observations T_i falling in $(-\infty, 0)$ is also latent. Note that, we now have a quite complex latent variable structure. However, by nesting inner EM algorithms inside an outer EM, we do not need the actual realizations of the unobservable random variables T_i and L_{ij} for $j = J+1, \dots, J+T_i$. To iteratively update the parameter estimates, only the conditional expectations of the corresponding sufficient statistics are required.

A nested EM algorithm is invoked by treating $Y_{\text{com}} = (C, D, Z, T, L_t)$ as the complete data, where $T = (T_1, \dots, T_I)$ and L_t is an array with elements L_{ij} for $i = 1, \dots, I$ and $j = J+1, \dots, J+T_i$. The complete-data log-likelihood is then given by

$$\begin{aligned} \ell(\Theta|C, D, Z, T, L_t) = & \sum_{i=1}^I \sum_{j=1}^{J+T_i} \{D_{ij}[\log \phi + \log N(L_{ij}|\mu_i, \sigma_i)] \\ & - \log(C_{ij}+1) + (1-D_{ij})[\log(1-\phi)] \\ & + Z_{ij} \log \pi_j + (1-Z_{ij}) \log(1-\pi_j)] \\ & + (1-D_{ij})(1-Z_{ij})(C_{ij} \log \delta_i - \delta_i - \log C_{ij}!)\} \\ & + \sum_{i=1}^I \sum_{j=J+1}^{J+T_i} [\log N(L_{ij}|\mu_i, \sigma_i) - \log(C_{ij}+1)]. \end{aligned} \quad (9)$$

Let $Y_{\text{obs}} = C$ be the observed data. $Y_{\text{mis } 1}$ denotes D and $Y_{\text{mis } 2}$ denotes the rest of the unobserved data (Z, T, L_t) . Following the notation used in Dempster et al. (1977) denote $Y_{\text{mis } 1}^{(t)} = E(Y_{\text{mis } 1}|Y_{\text{obs}}, \Theta^{(t)})$. It is clear from (9) that $E(\ell(\Theta|Y_{\text{com}})|Y_{\text{obs}}, Y_{\text{mis } 1}, \Theta^{(t+\frac{k-1}{K})})$ is linear in $Y_{\text{mis } 1}$. Therefore, the outer E-step can be simplified by computing $Y_{\text{mis } 1}^{(t)}$ only once per iteration and then run K inner EM cycles with $(Y_{\text{obs}}, Y_{\text{mis } 1}^{(t)})$ treated as observed data. Specifically, the outer E-step calculates $w_j^{(t)} = E(D_j|C, \Theta^{(t)})$, the conditional expectation of D . Then, the inner EM treats $(C, w^{(t)})$ as observed data, where $w^{(t)} = (w_1^{(t)}, \dots, w_J^{(t)})$. Since Z and L_t are independent, we are essentially nesting two inner EM algorithms here. The inner E-step involving Z can be simplified to calculate the conditional expectation of Z_{ij} given $(C, w^{(t)}, \Theta^{(t+\frac{k-1}{K})})$ by noting that the complete-data log-likelihood (9) is linear in Z_{ij} and $\sum_{i=1}^I Z_{ij}$ is the complete-data sufficient statistic for π_j . The inner E-step involving L_t and T calculates the expected values of the sufficient statistics $s_i = \sum_{j=1}^{J+T_i} D_j L_{ij}$ and $S_i = \sum_{j=1}^{J+T_i} D_j L_{ij}^2$ for the normal distribution parameters (μ_i, σ_i) conditioning on the observed data, $w^{(t)}$ and $\Theta^{(t+\frac{k-1}{K})}$. For detailed steps of our nested EM algorithm, see Web Appendix B in the Supplementary Material.

Compared to (6) and (7), the nested EM algorithm greatly simplifies the process of updating $(\pi, \delta, \mu, \sigma)$ by providing closed-form formulas and so avoids the need for high-dimension optimization as well as the issue of numerical instability.

Finally, we need to determine the number of cycles K in each EM iteration. Note that, the purpose of the inner EM cycles is not to reach convergence, but rather to move quickly toward the mode of the incomplete-data log-likelihood with a small computational cost. Because EM algorithms usually make a significant progress in the first few iterations, van Dyk (2000) suggested to fix K at some small value. We choose $K = 5$ in our implementation.

2.2.5 Normalizing gene expression and identifying expressed genes

Once the mixture model is fitted and the MLE $\hat{\Theta}$ is obtained from the nested EM algorithm, the normalized expression N_{ij} of gene j from sample i can be obtained by

$$N_{ij} = E(D_j|C_j, \hat{\Theta}) \times \left\{ L_{ij} - \left[\hat{\mu}_i + \frac{\psi\left(-\frac{\hat{\mu}_i}{\hat{\sigma}_i}\right)}{\Phi\left(-\frac{\hat{\mu}_i}{\hat{\sigma}_i}\right)} \hat{\sigma}_i \right] \right\},$$

where $E(D_j|C_j, \hat{\Theta})$ is calculated by (4) from the last E-step, which estimates the probability of gene j being expressed, and the term in the braces is the estimated expression for an expressed gene after removing the sample-specific effect. Clearly, the normalized expression is in the log scale. It is easy to use MIXnorm for detecting expressed genes. Gene j is identified as expressed if $w_j^{(t)} > c_w$ at convergence, where $c_w \in [0, 1]$ is a cut-off value. As shown in [Supplementary Table S1](#) in [Web Appendix C2](#), the choice of c_w seems not to have a noticeable impact on the classification performance of MIXnorm. In fact, $w_j^{(t)}$ in (4) is determined by the ratio of $p(C_j|D_j = 1, \mu^{(t)}, \sigma^{(t)})$ and $p(C_j|D_j = 0, \pi_j^{(t)}, \delta^{(t)})$, which are the likelihoods of the data modeled by TN and ZIP distributions, respectively. These two likelihoods are usually separate well. Thus, it is not surprising for us to observe that in our simulations, $w_j^{(t)}$ was either close to zero or close to one when MIXnorm converges, and so different threshold values in a quite wide range would not affect the detection performance much. We mention that MIXnorm is directly applicable to FF or like samples. This is because FF samples may be viewed as a reduced case of FFPE samples (i.e. little degradation in FF samples compared to severe and diverse degradation in FFPE samples). However, it is inappropriate to apply existing methods to FFPE data as they do not have the capacity to deal with the more complex data structure, as mentioned in the introduction.

3 Results

3.1 Simulations

Simulation studies were conducted to compare MIXnorm with five methods commonly used for normalizing RNA-seq data, including UQ, PS, DESeq, RPM and TMM. Here, we used a data-generating model that is modified from the proposed mixture model, in order to better mimic real situations. In our six simulation studies, we examined the impact of the proportion of expressed genes on the normalization performance in study I, the impacts of the sample-specific effects in study II, the impacts of the gene-specific effects in study III, the sensitivity to violations of model assumptions in study IV, the performance of directly and separately applying MIXnorm when DE genes exist across different conditions in study V and the relationship between the sample size (number of genes) and computing time of MIXnorm in study VI. For details about the data-generating models, process and simulation settings, see [Supplementary Web Appendix C1](#). All the results are reported and discussed in [Supplementary Web Appendix C2](#). We find that MIXnorm consistently outperforms the existing methods in nearly all the settings and is more robust to changes of sample-specific or gene-specific effects as well as violations of model assumptions. We also find that the computing time of MIXnorm has almost a perfect positive linear relationship with the sample size and number of genes, respectively. When DE genes exist, we recommend applying MIXnorm to normalize data from different groups separately instead of applying it to pooled data.

3.2 Real data applications

3.2.1 Soft tissue sarcomas data

The soft tissue sarcomas dataset was used for our exploratory analysis in Section 2.1, which contains expression levels for 20 242 protein-coding genes from paired FF and FFPE samples of 41 patients measured by RNA-seq. Note that, the availability of paired FF samples would enable us to quantitatively assess and compare the performance of different RNA-seq normalization methods. Since the true (normalized) gene expression is unknown, it is generally difficult to compare the performance on real data. Nevertheless, such paired FF data, after normalization to remove technical effects, can be used as a surrogate of the truth. This is because FF tissues are known to maintain RNA very well (much lower degradation of

RNA and no methylene crosslink between RNA and proteins) and thus are considered as a gold standard for most molecular assays ([Solassol et al., 2011](#)). To be specific, the gene-wise Pearson correlations between normalized FFPE and FF data (in the log scale) were computed and compared among the six different methods (MIXnorm, DESeq, RPM, TMM, PS and UQ). The correlations between original FFPE and FF data (without using any normalization method, also in the log scale) were computed to provide a baseline. We used the same approach as described in [Supplementary Web Appendix C2](#) to deal with genes that have zero SD when computing Pearson correlations.

Since the soft tissue sarcomas data were collected primarily for the analysis of the CINSARC gene signature, we evaluated the performance on all the 20 242 genes as well as the 67 genes in the gene signature. [Table 1](#) summarizes gene-wise correlations for the CINSARC gene signature in the left panel, and gene-wise correlations for all the genes in the middle panel, where genes in the CINSARC signature show considerably higher correlations than the population of the protein-coding genes for the methods MIXnorm, DESeq and RPM. Among all the methods, MIXnorm results in the highest quartiles. DESeq is the second best in this real data application, which is also one of the recommended normalization methods for high-throughput RNA-seq data ([Dillies et al., 2013](#)). The most straightforward normalization method RPM gives better result compared to PS and TMM for genes in the CINSARC signature. UQ failed to normalize the data. After removing genes with zero raw read counts across all samples, there are still several FFPE samples with more than 75% zero counts, which makes the scaling factors of UQ equal zero. Note that, for DESeq, genes with at least one zero read count were removed before calculating the scaling factors, which removed 97% of genes in the FFPE RNA-seq data.

[Figure 2](#) plots the normalized FFPE and FF expression levels in the log scale for all 67 genes in the CINSARC signature, where the left panel shows scatterplots for MIXnorm, TMM and DESeq, and the right panel shows scatterplots for RPM, PS and the original data. We observe that all these genes were identified as expressed genes by MIXnorm, as one may expect. For all methods except MIXnorm, there are genes whose normalized FF expression is high but normalized FFPE expression is low or almost zero, resulting in an obvious horizontal line at $y = 0$. This suggests that the existing methods were not able to handle genes with zero or low expression well in FFPE samples. The Pearson correlation coefficients between normalized FF and FFPE expression levels reported in [Figure 2](#) also indicate that MIXnorm has the best overall performance for this gene signature.

3.2.2 Clear cell renal cell carcinoma data

Our second application uses the clear cell renal cell carcinoma (ccRCC) dataset from [Eikrem et al. \(2016\)](#), of which RNA-seq data from FFPE samples were used to simulate synthetic data in Section 3.1.

ccRCC is the most common subtype of renal cell carcinoma, and is resistant to conventional chemotherapy and radiotherapy. Therefore, it is only curable by early surgical tumor removal when a surgery is able to eradicate the disease. Reversal of cancer gene expression is predictive of therapeutic potential. Much effort has been made to develop molecular signatures of disease progression for ccRCC. Among many, [Eikrem et al. \(2016\)](#) aimed to validate RNA-seq outcomes from FFPE biopsies with paired RNAlater stored samples for ccRCC patients. The data include 16 adult patients from Haukeland University Hospital. Four core biopsies were obtained from each patient, including two with ccRCC and two from adjacent normal tissues. The two pairs of ccRCC and normal tissues were then stored in FFPE and RNAlater, respectively. The RNA-seq data obtained from these tissues contain genes annotated by Ensembl. We converted the Ensembl ID to the HGNC symbol by Biomart and kept the protein-coding genes only. The processed dataset contains 18 458 protein-coding genes and 32 paired FFPE and RNAlater samples.

The proposed MIXnorm model in Section 2.2.1 is quite general, we believe. In practice, however, model assumptions (mainly, zero

Table 1. Data applications

Method	Soft tissue sarcomas						ccRCC		
	CINSARC gene signature			20 242 protein-coding genes			18 458 protein-coding genes		
	First Qu.	Median	Third Qu.	First Qu.	Median	Third Qu.	First Qu.	Median	Third Qu.
MIXnorm	0.333	0.455	0.517	0.098	0.235	0.384	0.304	0.524	0.789
DEseq	0.165	0.260	0.354	0.019	0.160	0.298	0.203	0.418	0.609
RPM	0.146	0.243	0.350	0.010	0.156	0.297	0.204	0.422	0.612
TMM	0.010	0.098	0.161	0.021	0.159	0.291	0.110	0.267	0.463
PS	−0.126	0.002	0.154	−0.374	−0.148	0.036	0.071	0.285	0.491
UQ	—	—	—	—	—	—	0.187	0.407	0.610
Original	0.020	0.107	0.181	0.011	0.146	0.277	0.142	0.299	0.485

Note: The left and middle panels show gene-wise correlations between normalized FFPE and FF expression for soft tissue sarcomas data; the right panel shows gene-wise correlations between normalized FFPE and RNAlater expression for ccRCC data. The UQ method failed to work for soft tissue sarcomas data due to excess zeros. The highest quartiles are highlighted in bold.

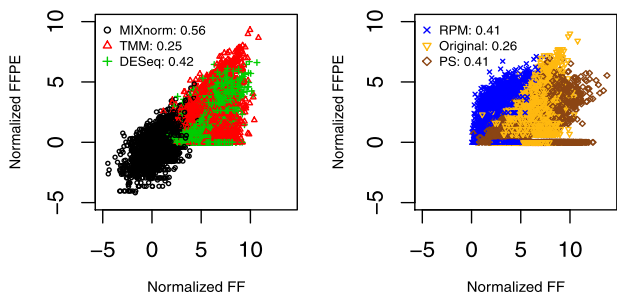


Fig. 2. Soft tissue sarcomas data example: the normalized FFPE versus FF expressions in the log scale from all 41 samples for all 67 genes in the CINSARC gene signature. The left panel shows scatterplots for MIXnorm, TMM and DESeq, and the right panel shows scatterplots for RPM, PS and the original data (without any normalization). Pearson correlation coefficients are reported for each method in the legend

inflation and truncated normality) may not roughly hold. Thus, before applying MIXnorm to RNA-seq data, we recommend that users conduct an explanatory analysis as we did for the soft tissue sarcomas data using log transformed read counts (i.e. Fig. 1), and look for clear bimodality with the first spike occurring at zero and approximately Gaussian curves around the second mode for most samples. The empirical densities of log read counts for the 32 paired FFPE and RNAlater samples, shown in Supplementary Figure S7, suggest the suitability of the proposed MIXnorm for the ccRCC data. We also suggest conducting a confirmatory analysis after applying MIXnorm, by visually examining Q–Q plots or conducting distributional tests, to check whether the assumption of truncated normality is adequate for expressed genes in most of the samples. Both Q–Q plots (Supplementary Fig. S8) and *P*-values from Kullback–Leibler tests (Supplementary Fig. S9) suggest that there was no gross departure from the assumed TN distributions for ccRCC FFPE data. For detail, see Web Appendix D of Supplementary Material.

RNAlater is an aqueous, non-toxic tissue storage reagent that rapidly permeates tissues to stabilize and protect cellular RNA in unfrozen specimens. It is considered to be comparable to the FF procedure. Therefore, the normalized RNAlater data were used as a surrogate of the gold standard in this application. The 18 458 gene-wise Pearson correlations between normalized FFPE and RNAlater data in the log scale were computed to evaluate the performance. As suggested in Section 3.1, we performed MIXnorm separately on the tumor and normal tissues. The gene-wise correlations were then calculated from all 32 paired samples. The quartiles and median of the correlations are summarized in the right panel of Table 1. Compared to the original data, the MIXnorm, DEseq, RPM and UQ normalized data improve the gene-wise Pearson correlations. Clearly, MIXnorm performs the best among all methods. DEseq,

Table 2. ccRCC data example: summary of differential expression analysis based on different normalization methods

	FFPE DE genes	RNAlater DE genes	Common DE genes	Common top 20 DE
MIXnorm	1488	1482	1036	13
DEseq	1014	951	680	7
RPM	999	926	676	9
TMM	1073	1067	632	7
PS	1001	1300	652	8
UQ	1002	943	679	8
Original	1041	1096	646	9

Note: The second column is the number of DE genes identified from the FFPE data; the third column is the number of DE genes identified from the RNAlater data; the fourth column is the number of common genes between the two sets of DE genes; and the last column is the number of common genes among the two sets of top 20 DE genes from FFPE and RNAlater.

RPM and UQ have similar quartiles in this application. We note that, the ccRCC FFPE data have better quality compared to the soft tissue sarcomas FFPE data. In fact, DEseq only needs to remove 32% of genes that have zero raw read counts. UQ needs to remove 5% of genes with zero raw read counts across all samples. Obviously, the performance of the quantile-based methods heavily depends on the data quality. Further, in real applications with FFPE samples, none of the existing normalization methods is robust while MIXnorm seems to be superior. After all, only MIXnorm is specifically designed for FFPE RNA-seq data.

As requested by one of the reviewers, we provide additional results in Supplementary Table S2 to investigate the impact of removing genes with low expression on the performance of different normalization methods. We find that MIXnorm gives similar results regardless of removal of such genes or not, and maintains its top performance. For detail, see Supplementary Web Appendix D.

This paired design allows us to conduct differential expression analysis between ccRCC and normal conditions using both FFPE and RNAlater samples, and to access the validity of using FFPE samples for such analysis. We identified DE genes [Benjamini–Hochberg adjusted *P*-value <0.05 from paired *t*-tests and absolute log2 fold change (FC) >2] from each of the two tissue sources based on the different normalization methods and report results in Table 2. We find that MIXnorm gives the highest number of common DE genes from the two sources. Furthermore, among the two sets of top 20 DE genes identified from RNAlater and FFPE samples, MIXnorm gives the highest number (13) of common genes while the other methods gives 9 or less. Table 3 summarizes the FFPE and RNAlater log2 FCs of 13 shared genes identified by MIXnorm, of which Spearman correlation is 0.88.

Table 3. ccRCC data example: the 13 shared genes among the two sets of top 20 DE genes from FFPE and RNAlater, ordered by the absolute value of the RNAlater log2 FC

	CA9	SLC6A3	NDUFA4L2	UMOD	GP2	CLCNKA	CDCA2	TNFAIP6	SLC4A11	KNG1	SLC12A1	AQP2	NELL1
RNAlater log2 FC	8.04	7.22	6.39	-6.15	-5.51	-5.28	5.23	5.17	-5.08	-5.02	-4.95	-4.92	-4.77
FFPE log2 FC	5.66	6.31	4.89	-5.62	-4.96	-5.69	5.05	5.45	-5.22	-5.03	-4.89	-4.89	-5.02

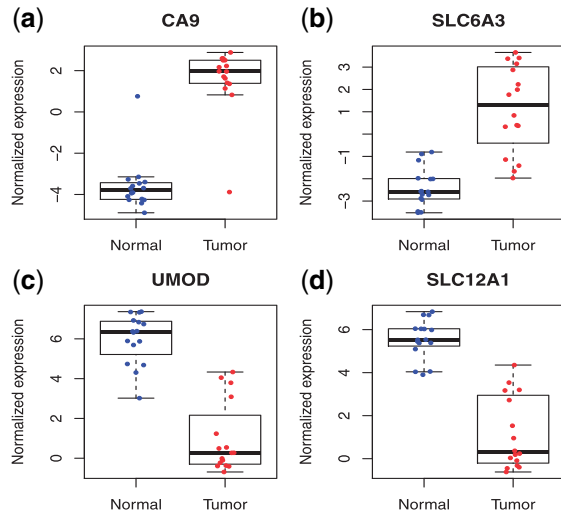
**Fig. 3.** ccRCC data example: normalized expressions levels of CA9 (a), SLC6A3 (b), UMOD (c) and SLC12A1 (d) from FFPE samples

Table 3 confirms strong over-expression of SLC6A3 and CA9 and under-expression of UMOD and SLC12A1 in ccRCC tissues, previously identified by immunohistochemistry studies (Eikrem *et al.*, 2016; Schrödter *et al.*, 2016; Wozniak *et al.*, 2013). The normalized expression levels of the four genes from FFPE samples are plotted in Figure 3, which clearly show the up- and down-regulation of these genes. It is interesting to note that the most up-regulated gene SLC6A3 identified by FFPE data is associated with the process of producing dopamine transporter (DAT). The importance of expression changes of DAT has been widely studied in Parkinson's syndrome and attention-deficit/hyperactivity disorder (Nutt *et al.*, 2004; Schrödter *et al.*, 2016). Recently, Hansson *et al.* (2017) studied FF samples from The Cancer Genome Atlas database and identified the DAT SLC6A3 as a specific biomarker for ccRCC. Our application demonstrates that SLC6A3 expression measured from FFPE samples may also serve as a highly specific biomarker for ccRCC. Tostain *et al.* (2010) presented a comprehensive study on the carbonic anhydrase 9 (CA9) as a marker for diagnosis, prognosis and treatment in ccRCC. It has been shown that CA9 mRNA expression measured by reverse transcription polymerase chain reaction and CA9 antigen detected by ELISA are promising molecular markers for diagnosis and prognosis of ccRCC (Tostain *et al.*, 2010). Our analysis further suggests that CA9 expression measured from FFPE RNA-seq may also serve as a molecular marker for ccRCC. It is worth noting that among the common top 20 DE genes, all normalization methods except MIXnorm failed to identify SLC12A1. SLC12A1 is a protein-coding gene that encodes kidney specific sodium-potassium-chloride cotransporter and is known to be associated with Bartter syndrome and Antenatal Bartter Syndrome. Schrödter *et al.* (2016) found that SLC12A1 expression was decreased in FF ccRCC tissues. Our analysis finds that after MIXnorm normalization, FFPE tissues are also able to detect down-regulation of SLC12A1.

4 Discussion

In recent years, many studies have been conducted to evaluate the feasibility of using FFPE specimens with RNA-seq, the dominant

high-throughput technology in gene expression profiling. These studies have collectively provided overwhelming evidence of reliable expression profiles obtained from FFPE specimens. However, none of the existing methods was developed for normalizing FFPE RNA-Seq data, a critical step in data analysis. Motivated by real data from FFPE tissues, we developed a two-component mixture model, which intends to capture major characteristics of the FFPE RNA-seq data accurately. Due to the resulting complex likelihood function, direct maximization can be unrealistic and time-consuming. By designing a nested EM-type algorithm that is easy to implement and computationally efficient, we greatly reduced the difficulty of finding the MLE.

We have shown that MIXnorm maintains top performance across various simulation settings and in two real data applications, compared to five existing RNA-seq normalization methods. The advantage of MIXnorm becomes more significant when the proportion of expressed genes becomes small. This may be due to the fact that MIXnorm is able to identify expressed genes from non-expressed genes accurately, and then models the two groups separately by ZIP and TN distributions. Besides the improvement in performance, MIXnorm has two other merits: (i) it handles genes with high-proportion zeros rigorously while existing methods typically require removal of such genes beforehand and (ii) it can output a parameter that represents the proportion of expressed genes, which can serve as an overall quality score for an RNA-seq experiment using FFPE tissues.

In MIXnorm, we employed ZIP instead of zero-inflated NB distributions to model non-expressed genes. This is mainly because after sorting out expressed genes, over-dispersion would not be a major issue. Also, NB and Poisson models often give similar parameter estimates, and NB fitting leads to larger standard error (SE) estimates than Poisson fitting. However, for the purpose of normalization, the SE estimates would not affect the results. Thus, ZIP was used also for simplicity.

We mention that FF data show simpler patterns than FFPE data, which can be modeled by simplifying the FFPE model proposed in Section 2.2.1, i.e. setting $\sigma^2 \equiv \sigma^2$ in (2), as Figure 1d shows a constant variance of L_{ij} across samples. In our soft tissue sarcomas data example, the estimated SDs of the TN distributions are much more consistent for the FF samples (coefficient of variation $CV = 0.05$) than those for the FFPE samples ($CV = 0.40$). That is why one can apply MIXnorm directly to FF or like samples, as discussed in Section 2.2.5. However, for computational efficiency, we can further simplify the nested EM algorithm to accommodate a common variance σ^2 , to be able to run faster for FF data.

Single-cell RNA-sequencing (scRNA-seq) has become widely used for transcriptome analysis in many biological studies. Like FFPE RNA-Seq data, scRNA-seq data have the sparsity feature. However, we do not recommend that MIXnorm be applied to such data blindly. scRNA-seq experiments aim to capture the heterogeneity among individual cells, where different cell types or transient states may make a gene expressed only in some cell subpopulations (Vallejos *et al.*, 2017). As discussed in the introduction, commonly used normalization methods for bulk RNA-seq data (e.g. DEseq, TMM, PS, UQ, etc.) are typically based on scaling factors, which assume that most of the genes are not differently expressed across different samples in a study (Anders and Huber, 2010; Robinson and Oshlack, 2010). Obviously, this key assumption is not valid for scRNA-seq data. We note that MIXnorm is essentially a scaling factor-based method, too. The scaling factor for each sample is estimated by the mean of the sample-specific TN distribution. In particular, MIXnorm assumes that a gene is either expressed or not

across all samples in a study, which is invalid for scRNA-seq data. Thus, we believe that MIXnorm is not suitable for normalizing scRNA-seq data.

Funding

This work was supported by the National Institutes of Health [grant R15GM131390 to X.W.].

Conflict of Interest: none declared.

References

- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
- Bullard, J. et al. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, **11**, 94.
- Dempster, A.P. et al. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.*, **39**, 1–38.
- Dillies, M.-A. et al.; on behalf of The French StatOmique Consortium. (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.
- Eikrem, O. et al. (2016) Transcriptome sequencing (RNAseq) enables utilization of formalin-fixed, paraffin-embedded biopsies with clear cell renal cell carcinoma for exploration of disease biology and biomarker development. *PLoS One*, **11**, e0149743.
- Graw, S. et al. (2015) Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci. Rep.*, **5**, 12335.
- Grenier, J.K. et al. (2017) RNA-seq transcriptome analysis of formalin fixed, paraffin-embedded canine meningioma. *PLoS One*, **12**, e0187150.
- Hansson, J. et al. (2017) Overexpression of functional SLC6A3 in clear cell renal cell carcinoma. *Clin. Cancer Res.*, **23**, 2105–2115.
- Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.
- Lesluyes, T. et al. (2016) RNA sequencing validation of the Complexity INdex in SARComas prognostic signature. *Eur. J. Cancer*, **57**, 104–111.
- Li, J. et al. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538.
- Lin, X.S. et al. (2014) Differentiating progressive from nonprogressive T1 bladder cancer by gene expression profiling: applying RNA-sequencing analysis on archived specimens. *Urol. Oncol.*, **32**, 327–336.
- McLachlan, G.J. and Jones, P.N. (1988) Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, **44**, 571–578.
- Mortazavi, A. et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.
- Morton, M.L. et al. (2014) Identification of mRNAs and lincRNAs associated with lung cancer progression using next-generation RNA sequencing from laser micro-dissected archival FFPE tissue specimens. *Lung Cancer*, **85**, 31–39.
- Nutt, J.G. et al. (2004) The dopamine transporter: importance in Parkinson's disease. *Ann. Neurol.*, **55**, 766–773.
- Perlmutter, M.A. et al. (2004) Comparison of snap freezing versus ethanol fixation for gene expression profiling of tissue specimens. *J. Mol. Diagn.*, **6**, 371–377.
- Reis, P.P. et al. (2011) mRNA transcript quantification in archival samples using multiplexed, color-coded probes. *BMC Biotechnol.*, **11**, 46.
- Ripoli, F.L. et al. (2016) A comparison of fresh frozen vs. formalin-fixed, paraffin-embedded specimens of canine mammary tumors via branched-DNA assay. *Int. J. Mol. Sci.*, **17**, E724.
- Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
- Schrödter, S. et al. (2016) Identification of the dopamine transporter SLC6A3 as a biomarker for patients with renal cell carcinoma. *Mol. Cancer*, **15**, 10.
- Solassol, J. et al. (2011) KRAS mutation detection in paired frozen and formalin-fixed paraffin-embedded (FFPE) colorectal cancer tissues. *Int. J. Mol. Sci.*, **12**, 3191–3204.
- Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, **82**, 528–540.
- Tostain, J. et al. (2010) Carbonic anhydrase 9 in clear cell renal cell carcinoma: a marker for diagnosis, prognosis and treatment. *Eur. J. Cancer*, **46**, 3141–3148.
- Vallejos, C.A. et al. (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, **14**, 565–571.
- van Dyk, D.A. (2000) Nesting EM algorithms for computational efficiency. *Stat. Sin.*, **10**, 203–225.
- Wozniak, M.B. et al. (2013) Integrative genome-wide gene expression profiling of clear cell renal cell carcinoma in Czech Republic and in the United States. *PLoS One*, **8**, e57886.