frontiers in Genetics

METHODS
published: 24 March 2021
doi: 10.3389/fgene.2021.650795

SMIXnorm: Fast and Accurate RNA-Seq Data Normalization for Formalin-Fixed Paraffin-Embedded Samples

Shen Yin[1,2], *Xiaowei Zhan*[1], *Bo Yao*[1], *Guanghua Xiao*[1], *Xinlei Wang*[2]* *and Yang Xie*[1]*

[1] *Department of Population and Data Sciences, Quantitative Biomedical Research Center, The University of Texas Southwestern Medical Center, Dallas, TX, United States,* [2] *Department of Statistical Science, Southern Methodist University, Dallas, TX, United States*

OPEN ACCESS

Edited by:
Chaeyoung Lee,
Soongsil University, South Korea

Reviewed by:
Arpit Tandon,
Sciome LLC, United States
Bruce Alex Merrick,
National Institute of Environmental Health Sciences (NIEHS), United States

***Correspondence:**
Xinlei Wang
swang@smu.edu
Yang Xie
Yang.Xie@UTSouthwestern.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *08 January 2021*
Accepted: *01 March 2021*
Published: *24 March 2021*

Citation:
Yin S, Zhan X, Yao B, Xiao G, Wang X and Xie Y (2021) SMIXnorm: Fast and Accurate RNA-Seq Data Normalization for Formalin-Fixed Paraffin-Embedded Samples. Front. Genet. 12:650795. doi: 10.3389/fgene.2021.650795

RNA-sequencing (RNA-seq) provides a comprehensive quantification of transcriptomic activities in biological samples. Formalin-Fixed Paraffin-Embedded (FFPE) samples are collected as part of routine clinical procedure, and are the most widely available biological sample format in medical research and patient care. Normalization is an essential step in RNA-seq data analysis. A number of normalization methods, though developed for RNA-seq data from fresh frozen (FF) samples, can be used with FFPE samples as well. The only extant normalization method specifically designed for FFPE RNA-seq data, MIXnorm, which has been shown to outperform the normalization methods, but at the cost of a complex mixture model and a high computational burden. It is therefore important to adapt MIXnorm for simplicity and computational efficiency while maintaining superior performance. Furthermore, it is critical to develop an integrated tool that performs commonly used normalization methods for both FF and FFPE RNA-seq data. We developed a new normalization method for FFPE RNA-seq data, named SMIXnorm, based on a simplified two-component mixture model compared to MIXnorm to facilitate computation. The expression levels of expressed genes are modeled by normal distributions without truncation, and those of non-expressed genes are modeled by zero-inflated Poisson distributions. The maximum likelihood estimates of the model parameters are obtained by a nested Expectation-Maximization algorithm with a less complicated latent variable structure, and closed-form updates are available within each iteration. Real data applications and simulation studies show that SMIXnorm greatly reduces computing time compared to MIXnorm, without sacrificing the performance. More importantly, we developed a web-based tool, *RNA-seq Normalization (RSeqNorm)*, that offers a simple workflow to compute normalized RNA-seq data for both FFPE and FF samples. It includes SMIXnorm and MIXnorm for FFPE RNA-seq data, together with five commonly used normalization methods for FF RNA-seq data. Users can easily upload a raw RNA-seq count matrix and select one of the seven normalization methods to

produce a downloadable normalized expression matrix for any downstream analysis. The R package is available at https://github.com/S-YIN/RSEQNORM. The web-based tool, *RSeqNorm* is available at http://lce.biohpc.swmed.edu/rseqnorm with no restriction to use or redistribute.

Keywords: RNA-sequencing, normalization, FFPE, formalin-fixed paraffin-embedded samples, archived samples, statistical methods

1. INTRODUCTION

The application of next-generation sequencing (NGS) on measuring transcript abundance is widely known as RNA-seq. RNA-seq works by sequencing a library of cDNA fragments in a high-throughput manner in order to provide a comprehensive quantification of transcriptomic activities in biological samples (Quinn et al., 2018). In practice, normalization is an important step in RNA-seq data analysis since raw counts are often not directly comparable between samples (Dillies et al., 2013). Normalization brings out the biologically relevant information in gene expression by removing the systematic noise that arises from various experimental reasons (such as batch effect, lane effect, sequencing bias, etc.). Recent studies have shown that the raw sequencing data without normalization could cause invalid inference from many conventional statistical analyses and measurements (Quinn et al., 2018).

Fresh-frozen (FF) tissue biospecimens are considered the gold standard in molecular analysis using gene expression, as freezing preserves RNA well. A number of normalization methods have been well-studied on FF bulk RNA-seq data (Dillies et al., 2013). Most existing methods, including Reads Per Million (RPM), Upper-Quartile (UQ), DESeq, Trimmed Mean of *M*-values (TMM), etc., are based on scaling factor estimation, where the normalized expression is obtained by dividing the raw count by an estimate of the sample-specific scaling factor. For example, RPM (Mortazavi et al., 2008) estimates the scaling factor for each sample by the total number of reads divided by $1,000,000$. Similarly, UQ (Bullard et al., 2010) estimates the scaling factor by the upper quartile of counts across all genes for each sample. DESeq (Anders and Huber, 2010) normalization first calculates the geometric means of the counts for all genes as an average reference library. Then the ratio of the count in each sample to that in the reference library is computed. The scaling factor for each sample is estimated as the median of this ratio across all genes. TMM (Robinson and Oshlack, 2010) normalization selects one sample as a reference, and the *M*-values are calculated as the log ratios of the read count between each test sample and the reference for all genes. Then for each test sample, the scaling factor is estimated by the weighted mean of *M*-values after removing the genes with extreme average expression or *M*-values. On the other hand, normalization can be done implicitly by accounting for the size factor as a term in the RNA-seq data model. PoissonSeq (PS) (Li et al., 2012) models RNA-seq data by a Poisson log-linear model, where a set of sample-specific parameters is included as offset parameters in the linear predictor to account for different sequencing depths.

Recently, there has been increasing interest in performing transcriptome profiling on Formalin-Fixed Paraffin-Embedded (FFPE) tissues, as they are widely available from routine diagnostic sample preparation (Morton et al., 2014). Being able to successfully measure mRNA abundance from FFPE samples could greatly facilitate biomarker discoveries and genomic studies of clinical samples (Graw et al., 2015; Grenier et al., 2017). The major challenge of adapting FFPE biospecimens in molecular analysis is that the chemical process is designed to preserve cellular proteins rather than preserving RNA. As a result, RNA from FFPE tissues is usually degraded, which could limit gene expression analysis. Studies have shown that the fixation process, storage time, specimen size and conditions play important roles in the RNA quality from FFPE samples (Von Ahlfen et al., 2007). Although such samples may suffer from the chemical modifications and continued degradation over time, recent studies have shown that RNA-seq can measure the RNA expression from FFPE samples in sufficient quality (Li et al., 2014). Due to chemical modifications and the RNA degradation, the RNA quality and abundance extracted from the FFPE samples vary a lot. Therefore, the normalization step is even more important for RNA-seq data measured from FFPE samples than those measured from FF samples. Though the forementioned existing methods are applicable to FFPE RNA-seq data, none of them were specifically designed and validated on FFPE samples. The mRNA expression measured from FFPE can have lower quality and higher sparsity (i.e., many zero counts occur), compared with FF samples. Consequently, the zero-inflation makes the assumption of model-based FF RNA-seq normalization invalid. For most scaling factor-based RNA-seq normalization, practitioners need to discard genes with many zeros beforehand, which may be a significant portion of the data when applied to FFPE samples. To the best of our knowledge, MIXnorm (Yin et al., 2020) is the only method that is designed to normalize FFPE RNA-seq data but that is applicable to FF RNA-seq data as well. MIXnorm addresses the zero inflation by separately modeling the expressed genes and non-expressed genes in a mixture model. Yin et al. (2020) considered normalized expression from paired FF (or like) samples as a surrogate of the truth and showed quantitatively that MIXnorm consistently outperforms commonly used FF RNA-seq normalization methods when applied on FFPE samples by comparing the gene-wise correlations. In addition, it has been shown that after MIXnorm normalization, RNA-seq data from FFPE tissues enable us to detect important up- or down-regulated genes. However, MIXnorm relies on a complex latent variable structure for model fitting, which causes a heavy computational burden for large data sets. We show in this paper that the

statistical model of MIXnorm can be properly simplified to still capture the main characteristics of the FFPE RNA-seq data. We propose a simplified version of MIXnorm, labeled SMIXnorm, for FFPE RNA-seq data normalization. The fitting of SMIXnorm requires a less complicated latent variable structure. We show through simulation studies and real data applications that SMIXnorm retains almost the same performance as MIXnorm, while greatly reducing the computing time.

More importantly, there is a lack of platforms that integrate existing methods and produce normalized data by different methods. Evans et al. (2018) mentioned that the selection of normalization methods played an important role in downstream analysis due to the different assumptions those methods made. Furthermore, it is important to raise the awareness of the separate normalization methods for FFPE samples, as more and more applications involve RNA-seq data from such samples. We developed a web portal, *RNA-seq Normalization (RSeqNorm)* (http://lce.biohpc.swmed.edu/rseqnorm/), to conduct normalization for both FF and FFPE RNA-seq data. It offers seven normalization methods, with accompanying diagnostic plots for users to visually examine the RNA-seq data quality. Based on this platform, we compared different normalization methods using both comprehensive simulation studies and real data applications. These results, together with the *RSeqNorm* web portal, will facilitate users to select the best normalization method for their application.

The paper is structured as follows. In section 2.1, we present the SMIXnorm method. The statistical model of SMIXnorm is simplified from that of MIXnorm. An efficient nested Expectation-Maximization (EM) algorithm is designed for model fitting. We further justify the simplifications by comparing SMIXnorm to MIXnorm from a technical point of view. An introduction of the web portal *RSeqNorm* is given in section 2.2. Section 3 reports simulation studies and real data analyses. Finally, a brief concluding discussion is made in section 4.

2. MATERIALS AND METHODS

2.1. The SMIXnorm Method

2.1.1. The Simplified Statistical Model

Assume the RNA-seq count data from FFPE samples can be summarized by a matrix $\boldsymbol{C}_{I\times J}$, where C_{ij} is the number of reads in sample i for gene j. We adopt a similar latent variable framework as in MIXnorm to address the zero inflation of such data. That is, the binary latent variable $D_j = 1$ indicates gene j is expressed and $D_j = 0$ indicates gene j is not expressed in the study for $j = 1, ..., J$. Then we model the count data as a mixture of zero-inflated Poisson (ZIP) and normal distributions,

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

$$L_{ij} \sim \text{N}\,(\mu_i, \sigma_i^2), \text{if } D_j = 1, \tag{2}$$

$$D_j \sim \text{Ber}(\phi), \tag{3}$$

where $L_{ij} = \log(C_{ij} + 1)$ denotes the log transformed count, $0 \le \pi_j, \phi \le 1$, $\delta, \mu_i, \sigma_i \ge 0$ for $i = 1, ..., I$ and $j = 1, ...J$. The model assumes an unobserved variable D_j which follows a Bernoulli distribution with parameter ϕ. Genes with $D_j = 0$ are considered not expressed. These include low-expression genes that have abundance below detection limit, or biologically non-expressed genes that are absent from the biological sample of interest, or genes that should have been expressed but suffer from high-level mRNA degradation. The observed counts from non-expressed genes are due to background noise and are modeled by a zero-inflated Poisson distribution with gene-specific probability of extra zeros π_j and a common expected Poisson count δ. Genes with $D_j = 1$ are expressed genes and we model the log counts of those genes by a normal distribution with sample-specific location and scale parameters. Compared to the model of MIXnorm, we use the normal distribution in (2) instead of a truncated normal distribution and a common Poisson mean δ rather than the sample-specific mean δ_i. As will be discussed in section 2.1.3, these would greatly facilitate the computation while not hurting much the performance. Note that the normal distribution assigns positive densities to negative log counts L_{ij}, which never occur in real data. However, it is reasonable to assume that the negative values only take a negligible portion of the density in modeling expressed genes as the mean (log) counts of such genes are usually well above zero. Further, SMIXnorm is directly applicable to FF or like samples based on the same argument in Yin et al. (2020) that FF samples may be considered a reduced case of FFPE samples.

2.1.2. Model Fitting

Let $\Theta = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \delta, \phi)$ denote the set of all parameters in the simplified model. The observed data likelihood as a function of Θ is defined as follow:

$$\begin{aligned} L(\Theta|\boldsymbol{C}) &= \prod_{j=1}^{J} p(\boldsymbol{C}_j|\Theta) \\ &= \prod_{j=1}^{J} [p(\boldsymbol{C}_j|D_j = 1, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(D_j = 1|\phi) \\ &\quad + p(\boldsymbol{C}_j|D_j = 0, \pi_j, \delta) p(D_j = 0|\phi)] \\ &= \prod_{j=1}^{J} [\prod_{i=1}^{I} p(C_{ij}|D_j = 1, \mu_i, \sigma_i) \cdot \phi \\ &\quad + \prod_{i=1}^{I} p(C_{ij}|D_j = 0, \pi_j, \delta) \cdot (1 - \phi)], \end{aligned} \tag{4}$$

where $p(C_{ij}|D_j = 1, \mu_i, \sigma_i)$ is the probability density function (pdf) of C_{ij} for expressed genes such that $\log(C_{ij} + 1)$ follows the normal distribution in (2), and $p(C_{ij}|D_j = 0, \pi_j, \delta)$ is the zero-inflated Poisson probability mass function (pmf) for non-expressed genes as described in (1). Note that a continuous distribution is used to approximately model the discrete random variable $\log(C_{ij} + 1)$. See Yin et al. (2020) for a detailed justification for the approximation.

A common approach to obtain the maximum likelihood estimate (MLE) of Θ is to treat $(\boldsymbol{C}, \boldsymbol{D})$ as the complete data and update the parameter estimates iteratively by an EM algorithm.

However, direct implementation of the EM algorithm requires Newton-Raphson type maximization to update the parameters within the ZIP component, which may cause the EM algorithm fail to converge due to numerical instability. By a similar approach in Yin et al. (2020), we update the parameter estimates from the ZIP component by nesting another EM algorithm and avoid the Newton-Raphson type optimization. Specifically, we construct another latent variable Z_{ij} for genes not expressed so that the ZIP distribution can be treated as a mixture of two states, the perfect zero state and the Poisson state. Assume $Z_{ij} = 1$ when C_{ij} is from the perfect zero state and $Z_{ij} = 0$ when C_{ij} is from the Poisson state, satisfying $Z_{ij}|D_j = 0 \sim \text{Ber}(\pi_j)$. Let $\boldsymbol{Y}_{\text{com}} = (\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{Z})$ denotes the complete data. The complete-data log-likelihood with latent variables $\boldsymbol{D}$ and $\mathbf{Z}$ is given by

$$\begin{aligned}\ell(\Theta|\mathbf{C}, \boldsymbol{D}, \mathbf{Z}) = \sum_{j=1}^{J}\sum_{i=1}^{I}\Big\{ & D_j\left[\log\phi + \log \mathrm{N}\left(L_{ij}|\mu_i, \sigma_i\right) - \log(C_{ij}+1)\right] \\ & + (1-D_j)\left[\log(1-\phi) + Z_{ij}\log\pi_j \right. \\ & \left. + \left(1-Z_{ij}\right)\log\left(1-\pi_j\right)\right] \\ & + \left(1-D_j\right)\left(1-Z_{ij}\right)\left[C_{ij}\log\delta - \delta - \log C_{ij}!\right]\Big\}. \quad (5)\end{aligned}$$

The outer EM treats $\boldsymbol{C}$ as observed data and $\boldsymbol{D}$ as missing data. Let $\Theta^{(t)} = (\boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\pi}^{(t)}, \delta^{(t)}, \phi^{(t)})$ denote the set of current parameter estimates after t iterations of the algorithm. The distribution of $\boldsymbol{D}$ given the observed data and $\Theta^{(t)}$ is

$$p(\boldsymbol{D}|\boldsymbol{C}, \Theta^{(t)}) = \prod_{j=1}^{J}\frac{p(\boldsymbol{C}_j, D_j|\Theta^{(t)})}{p(\boldsymbol{C}_j|\Theta^{(t)})} = \prod_{j=1}^{J}\left(w_j^{(t)}\right)^{D_j}\left(1-w_j^{(t)}\right)^{1-D_j}, \quad (6)$$

where

$$w_j^{(t)} = \frac{\phi^{(t)}p(\boldsymbol{C}_j|D_j=1, \boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)})}{\phi^{(t)}p(\boldsymbol{C}_j|D_j=1, \boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)}) + (1-\phi^{(t)})p(\boldsymbol{C}_j|D_j=0, \pi_j^{(t)}, \delta^{(t)})}. \quad (7)$$

Note that in (5), the complete-data log-likelihood is linear in $\boldsymbol{D}$. Therefore, the outer E step, which calculates the conditional expectation of the complete-data log-likelihood with respect to the missing data $\boldsymbol{D}$, is reduced to computing $w_j^{(t)} = \mathbf{E}(D_j|\boldsymbol{C}, \Theta^{(t)})$ only once per iteration. Within each iteration of the outer EM, the inner EM repeats K cycles and treats $\boldsymbol{Z}$ as missing data and $(\boldsymbol{C}, \boldsymbol{w}^{(t)})$ as observed data, where $\boldsymbol{w}^{(t)} = (w_1^{(t)}, ..., w_J^{(t)})$. By the similar argument and that $\sum_{i=1}^{I} Z_{ij}$ is the complete-data sufficient statistic for π_j, the kth cycle of the inner E step involving $\boldsymbol{Z}$ is reduced to calculating the conditional expectation of Z_{ij} given $(\boldsymbol{C}, \boldsymbol{w}^{(t)}, \Theta^{(t+\frac{k-1}{K})})$.

Details of the nested EM algorithm are summarized in section 1 in **Supplementary Material**. The convergence can be detected using the change of the observed-data log-likelihood $\ell(\Theta|\boldsymbol{C})$ in two consecutive iterations which is trivial to obtain via (4). We set the number of inner EM cycles $K = 5$ within each iteration in our implementation as in Yin et al. (2020).

Similar to MIXnorm, SMIXnorm normalization relies on the accurate classification of the genes that are expressed or not. This is indicated by the latent variable D_j. We estimate D_j by its conditional expectation given the observed data at the last E-step. Assume that the majority of genes are not differentially expressed across samples (Robinson and Oshlack, 2010). Then the means of the normal distributions μ_i's can be treated as the sample-specific noises. The normalized expression for sample i and gene j can be given by

$$N_{ij} = \mathbf{E}\left(D_j|\boldsymbol{C}_j, \hat{\Theta}\right) \times (L_{ij} - \hat{\mu}_i), \quad (8)$$

where $\hat{\Theta}$ denotes the MLE of Θ. In our numerical experiments, when gene j is not expressed, it is often the case that the conditional expectation $\mathbf{E}\left(D_j|\boldsymbol{C}_j, \hat{\Theta}\right) \approx 0$, which makes $N_{ij} \approx 0$; when gene j is expressed, $\mathbf{E}\left(D_j|\boldsymbol{C}_j, \hat{\Theta}\right)$ is often close to 1. Thus, we may simply output zero for genes with $\hat{D}_j = 0$, and $L_{ij} - \hat{\mu}_i$ for genes with $\hat{D}_j = 1$ in the actual implementation. The proposed method normalizes the data by subtracting the estimated sample-specific noise from the log count. Therefore, the normalized data are in the log scale.

2.1.3. SMIXnorm vs. MIXnorm

There are two major simplifications when comparing SMIXnorm to MIXnorm. First, SMIXnorm assumes a common Poisson mean δ for the non-expressed genes, where MIXnorm allows the sample-specific Poisson means δ_i. Note that δ appears in the normalization step (8) through the conditional expectation of D_j that can be computed by Equation (7). We observe in practice that after the nested EM algorithm converges, the conditional expectation $\mathbf{E}(D_j|\boldsymbol{C}, \Theta^{(t)})$ [i.e., $w_i^{(t)}$ in Equation 7] is not sensitive to the choice of a common or sample specific Poisson mean. With the parameters set to their MLE, this conditional expectation mainly depends on the ratio between $p(\boldsymbol{C}_j|D_j = 0, \hat{\pi}_j, \hat{\delta})$ (or $p(\boldsymbol{C}_j|D_j = 0, \hat{\pi}_j, \hat{\delta}_i)$ in MIXnorm) and $p(\boldsymbol{C}_j|D_j = 1, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}})$. The Poisson mean essentially reflects the background noise level and is supposed to be a small positive number. Thus, the former distribution, regardless of using $\hat{\delta}$ or $\hat{\delta}_i$, puts most of its probability mass near 0, whereas the latter distribution in practice has negligible density around small values near 0. Consequently, the conditional expectation in Equation (7) is close to either 1 or 0 depending on whether the gene is expressed or not and reducing δ_i to δ has little effect in the normalization step.

We further simplify the model by ignoring the truncation on the normal distributions. Compared to MIXnorm, the sample-specific noise among expressed genes is captured by the mean of a normal distribution in SMIXnorm instead of the mean of a truncated normal distribution. However, we argue that these two estimates are asymptotically identical. Assume we model a set of continuous positive real numbers $\boldsymbol{X} = (X_1, ..., X_J)$ by a truncated normal distribution $\text{TN}(\theta, \tau^2, 0, +\infty)$, where θ and τ are the mean and standard deviation of the corresponding normal distribution before truncation. To estimate the mean

of the truncated normal distribution, which is a function of θ and τ, say $m(\theta, \tau)$, one common approach is to first obtain the maximum likelihood estimates $(\hat{\theta}, \hat{\tau})$, then the maximum likelihood estimate of the mean is $m(\hat{\theta}, \hat{\tau})$ based on the invariance property of MLE. Another approach for a point estimate is the method of moments. The method of moments estimate of $m(\theta, \tau)$ is simply the sample mean $\bar{X}$, which is essentially the MLE of μ if we ignore the truncation and model the data with $N(\mu, \sigma^2)$. Under the large sample situation, as is often the case in RNA-seq data involving many genes from whole-genome experiments, both $m(\hat{\theta}, \hat{\tau})$ and $\bar{X}$ are asymptotically consistent estimators for $m(\theta, \tau)$. The use of a normal distribution without truncation is appealing since it has closed-form parameter updates within the nested EM algorithm, while a truncated normal distribution requires additional latent variable structures and data augmentation to obtain closed-form parameter updates (Yin et al., 2020). Overall, SMIXnorm produces similar normalized expression compared to MIXnorm whereas greatly simplifies the model fitting process, as will be confirmed in section 3 by numerical evidence.

2.2. *RSeqNorm* Web Portal

The *RSeqNorm* web portal (http://lce.biohpc.swmed.edu/rseqnorm) provides a set of analysis routines for normalization of RNA-seq data from either FF or FFPE samples. The workflow is illustrated in **Figure 1**. Users provide raw sequence read count data (e.g., from RNA-seq experiments) in the form of an integer-valued matrix $\boldsymbol{C}$. The web portal can compute the normalized expression as well as diagnostic information for download. We implement SMIXnorm, MIXnorm and five commonly used normalization methods, including Reads Per Million (RPM), Upper-Quartile (UQ), DESeq, Trimmed Mean of M-values (TMM) and PoissonSeq (PS). Though developed for FFPE data, SMIXnorm and MIXnorm are directly applicable to FF data normalization. For FF data normalization, there seems to have no unanimously best normalization method. We suggest using the methods offered by *RSeqNorm* and evaluating the normalization performance using prior information or known biological knowledge. For example, users may conduct a differential expression (DE) test following the normalization step and select the method that detects more genes that are known to be differentially expressed in the literature.

RSeqNorm accepts the raw read count matrix in a comma-separated values (CSV) file. The (i, j) element of the count matrix records how many reads have been assigned to gene j in sample i. An example input file is downloadable from the *RSeqNorm* website. Detailed file requirements are shown in **Figure 2**. SMIXnorm and MIXnorm require additional input arguments including the maximum number of iterations [range $(10, 50)$, default value 15] and the convergence threshold [recommend range $(1e-5, 1)$, default value 0.01] for the nested EM algorithm. We note that the observed-data log-likelihood as a function of all parameters may have a large curvature near the MLE. However, the SMIXnorm and MIXnorm normalized expression values are not sensitive to small variations of the parameter estimates. Therefore, the convergence criterion here is defined as the maximum absolute change in the parameter updates between the previous and current iterations, instead of using the change in the observed-data log-likelihood. The algorithm stops when the absolute change is smaller than the predetermined threshold value or the maximum number of iterations is reached. We mentioned in section 2.1.3 that the conditional probability of being expressed is often close to either 0 or 1. As a consequence, the normalized expression level for a non-expressed gene is usually a trivial number ($<$ 10e−10 in our implementation to real data). Therefore, an approximate set of normalized expression is also available for SMIXnorm and MIXnorm, which normalizes genes with $\hat{D}_j = 0$ directly to 0 across all samples, where $\hat{D}_j = I(w_j^{(t)} > c_w)$ at convergence, $I(\cdot)$ is the indicator function and c_w is set to 0.5 in *RSeqNorm*. In practice, we observe that the choice of c_w is not sensitive in identifying expressed genes. The conditional probability of being expressed $(w_j^{(t)})$ and the proportion of expressed genes identified $(\hat{\phi})$ are returned in the R package *RSEQNORM*, which is downloadable from our website. Note that $\hat{\phi}$ may reflect the overall data quality for an RNA-seq experiment (i.e., a value close to 1 indicates high quality).

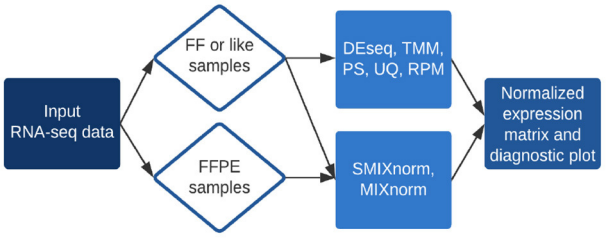


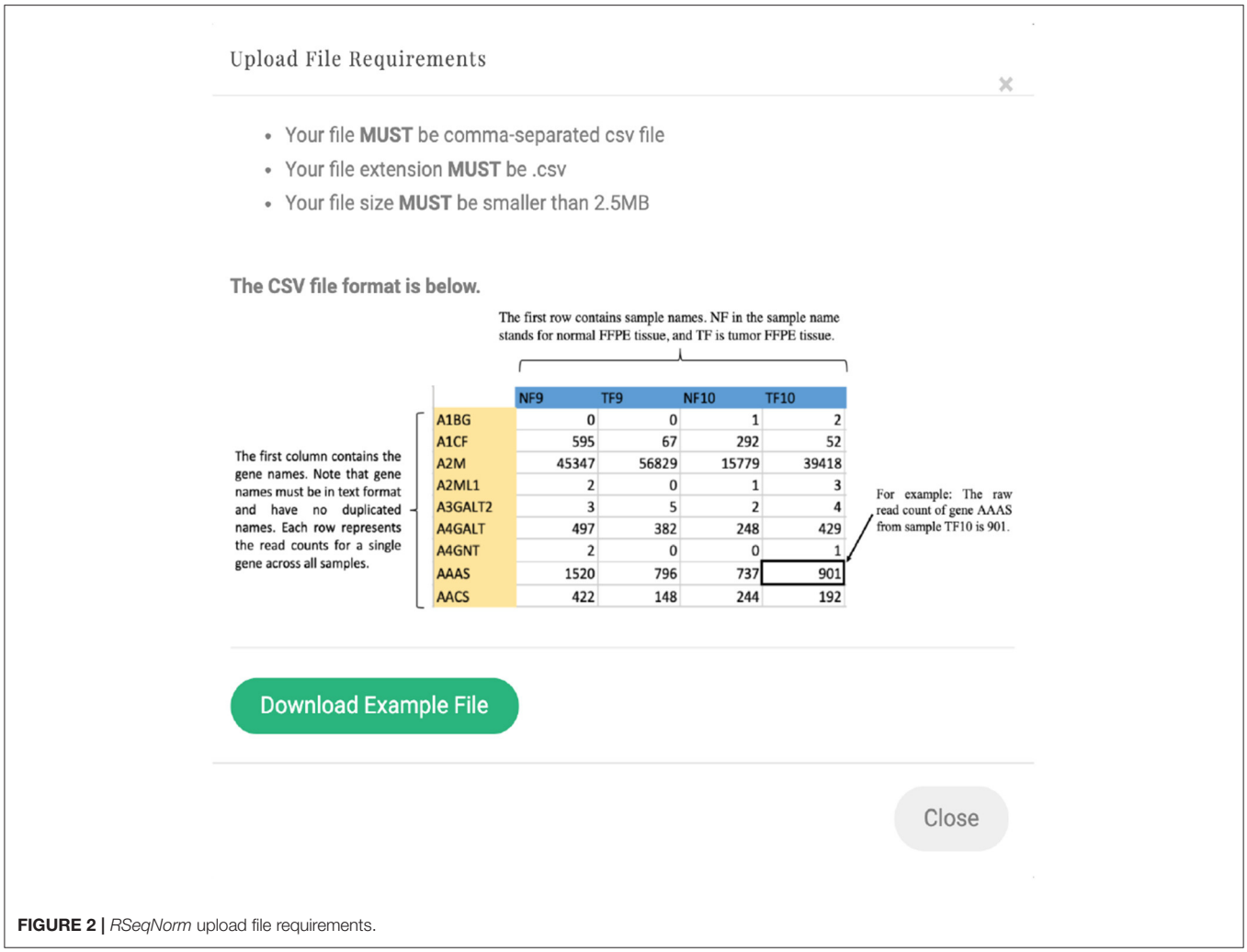
FIGURE 1 | Summary of *RSeqNorm* web-portal process.

RSeqNorm returns the normalized expression in a CSV file with the same dimension as the user input file. User may leave the web portal after successfully submitting the job and an email notification will be sent to the user with a download link when the normalization is finished. A histogram of zero-count proportions is also returned as shown in **Figure 3**, where SMIXnorm is used on the example data provided by *RSeqNorm*. The histogram shows the distribution of zero-count proportions among all samples (represented by the horizontal axis) over all input genes. High frequencies near 0 indicate that most biologically expressed genes are actually expressed in all samples in the experiment and so the data are of high quality. Ideally, one would expect that a gene is either expressed among all samples or not expressed in any of the samples. In this ideal case, the histogram shows frequencies only at 0 and 1 on the horizontal axis.

3. RESULTS

3.1. Simulation

We conducted simulation study to show that SMIXnorm greatly reduces computing time and maintains performance comparable to MIXnorm that is better than all FF RNA-seq normalization methods. Our simulation study evaluates the computing time of SMIXnorm and MIXnorm with different sample sizes. Note that the SMIXnorm model cannot be used to simulate the data



	NF9	TF9	NF10	TF10
A1BG	0	0	1	2
A1CF	595	67	292	52
A2M	45347	56829	15779	39418
A2ML1	2	0	1	3
A3GALT2	3	5	2	4
A4GALT	497	382	248	429
A4GNT	2	0	0	1
AAAS	1520	796	737	901
AACS	422	148	244	192

FIGURE 2 | *RSeqNorm* upload file requirements.

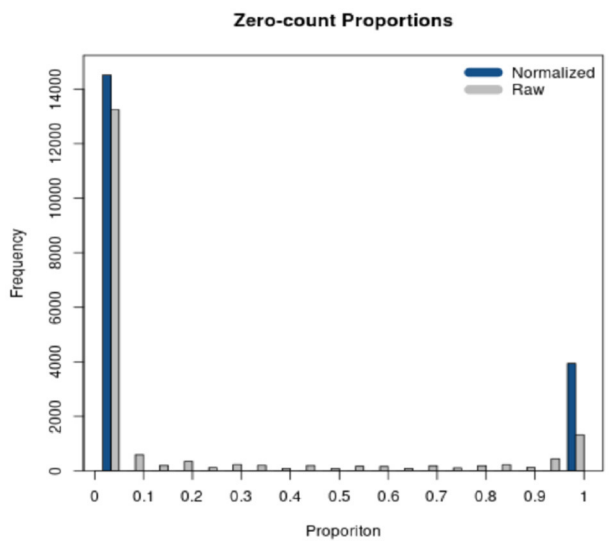


FIGURE 3 | Diagnostic plot returned by *RSeqNorm* using SMIXnorm.

since it permits negative log transformed counts that do not exist in practice. Here, a modified MIXnorm model was used to generate synthetic data sets (under the same settings as in Simulation VI in Yin et al., 2020). The model parameters were set to their MLEs estimated from a public RNA-seq dataset for FFPE soft tissue sarcomas samples (Lesluyes et al., 2016), which contains expression levels for 20, 242 protein-coding genes from 41 patients.

The sample size was set to a multiple of 41 from 41 to 2, 050. The average computing time for SMIXnorm and MIXnorm over 50 replicates is plotted against sample size in **Figure 4**. Both SMIXnorm and MIXnorm show a linear relationship between the average computing time and the sample size with Pearson correlations greater than 0.99. The linear regression fit of the SMIXnorm computing time against sample size results in a slope estimate of 0.051, compared to 0.297 from MIXnorm. To evaluate the performance of SMIXnorm and MIXnorm together with five commonly used FF RNA-seq normalization methods (PS, UQ, DESeq, RPM, and TMM), we calculated the gene-wise Pearson correlations for the 20, 242 genes between the normalized and true expression for each of the 50 simulated

data set under the setting of 41 samples. The results are reported in **Supplementary Figure 1**, showing that SMIXnorm and MIXnorm have almost identical performance and they consistently beat the other methods.

3.2. Real Data Analysis

When dealing with real data where true expression is unknown, researchers often consider frozen sections as the gold standard for most molecular assays. As mentioned in the introduction, the frozen process maintains RNA well, compared to the FFPE process. Therefore, we used the paired FF RNA-seq expression after normalization as a surrogate to the true gene expression in our three data examples, in order to evaluate the performance of different normalization methods.

3.2.1. Colorectal Cancer Data

The colorectal cancer data (Omolo et al., 2016) contains 54 selected FFPE tumor specimens from a larger multi-center cohort with paired FF samples. Gene expression levels were measured by whole genome RNA-seq (RNA-Acc) assay and Affymetrix GeneChip (Affy) platform on the FFPE samples and FF samples, respectively.

The activation of RAS signaling pathway is frequent in human cancer. Recent studies have shown that RAS mutations account for approximately 40% of colorectal cancers and lung cancers (Omolo et al., 2016). A number of RAS pathway activation gene expression signatures have been identified using multiple types of cancer cell lines and human FF samples. Omolo et al. (2016) evaluated an 18-gene RAS pathway signature on FFPE samples in five technology platforms. We focus on the Illumina whole genome RNA-seq of FFPE samples and the gold standard FF samples measured by Affymetrix GeneChip. The Affy_FF samples were normalized using the RMA method (Irizarry et al., 2003; Omolo et al., 2016).

To assess the performance of the translation of the gene signature from FF to FFPE samples, we considered the same metric as in Omolo et al. (2016). The RAS pathway activation score is defined as the mean normalized expression levels of the 18 RAS genes for each sample. We calculated the Spearman correlation using the 54 pairs of FF and FFPE RAS pathway activation scores for each normalization method and summarized the results in **Table 1**. The raw data give the lowest correlation as expected. SMIXnorm and MIXnorm give almost the same results and the highest correlations. The p-value based on a two-sided permutation test for the hypothesis $H_0 : \rho = 0$ is reported in the parenthesis. SMIXnorm, MIXnorm, PS, DESeq, RPM and UQ report significant correlations at the significance level of 0.05. MIXnorm takes about 42 s to normalize the colorectal cancer data, while SMIXnorm only takes about 7.5 s.

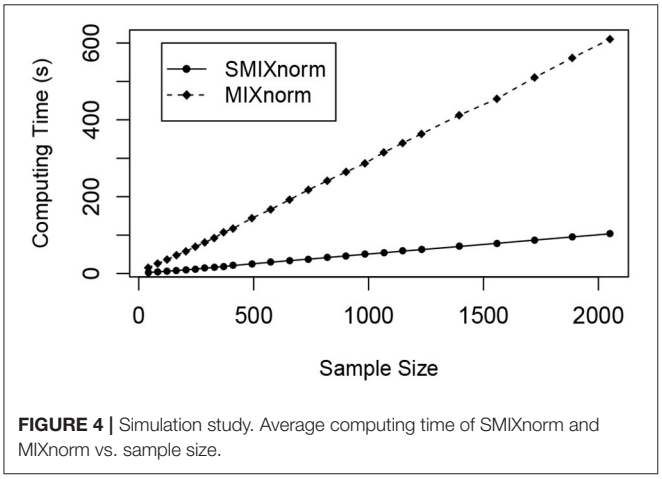


FIGURE 4 | Simulation study. Average computing time of SMIXnorm and MIXnorm vs. sample size.

3.2.2. Soft Tissue Sarcomas Data

The soft tissue sarcomas data (Lesluyes et al., 2016) measure the expression levels of 20,242 protein-coding genes from 41 patients with paired FF and FFPE samples. To evaluate the normalization performance, the gene-wise Pearson correlations between normalized FFPE and FF expression levels (in the log scale) were computed and compared among the seven different methods (SMIXnorm, MIXnorm, DESeq, RPM, TMM, PS, and UQ). A gene expression signature, Complexity INdex in SARComas (CINSARC), which contains 67 genes, has been identified as an important prognostic factor in sarcomas using fresh frozen samples. Lesluyes et al. (2016) showed that CINSARC remains a potential prognostic factor using the FFPE RNA-seq data. Therefore, we evaluated the performance of different normalization methods on all the 20,242 genes as well as the 67 genes in the CINSARC gene signature.

Figure 5 shows the gene-wise Pearson correlations for all the 20,242 genes using violin plots. The dot and line in each violin plot represent the mean and standard deviation of the gene-wise Pearson correlations. The gene-wise correlations from the 67 genes in the CINSARC signature are summarized in **Table 2** using the first, second, and third quartiles. Note that UQ failed to normalize the data as the scaling factor estimates equal 0 for some samples due to the excess zero counts. The genes in the gene signature have higher correlations than those calculated from the population of all protein-coding genes for all the normalization methods. The shapes of the violin plots suggest that SMIXnorm and MIXnorm have almost identical results and SMIXnorm gives the highest mean correlation while PS gives the lowest among all the methods. The three commonly used FF RNA-seq normalization methods, TMM, DESeq, and RPM, give similar results on all protein-coding genes and there is no clear improvement compared to the original correlations

TABLE 1 | Correlations ρ between normalized FF and FFPE RAS pathway activation scores.

	SMIXnorm	MIXnorm	PS	DESeq
ρ	0.343 (0.012)	0.343 (0.011)	0.324 (0.017)	0.298 (0.029)
	RPM	UQ	TMM	Raw
ρ	0.286 (0.036)	0.270 (0.049)	0.153 (0.269)	0.125 (0.366)

p-values in the parenthesis are based on a two-sided permutation test for the hypothesis $H_0 : \rho = 0$.

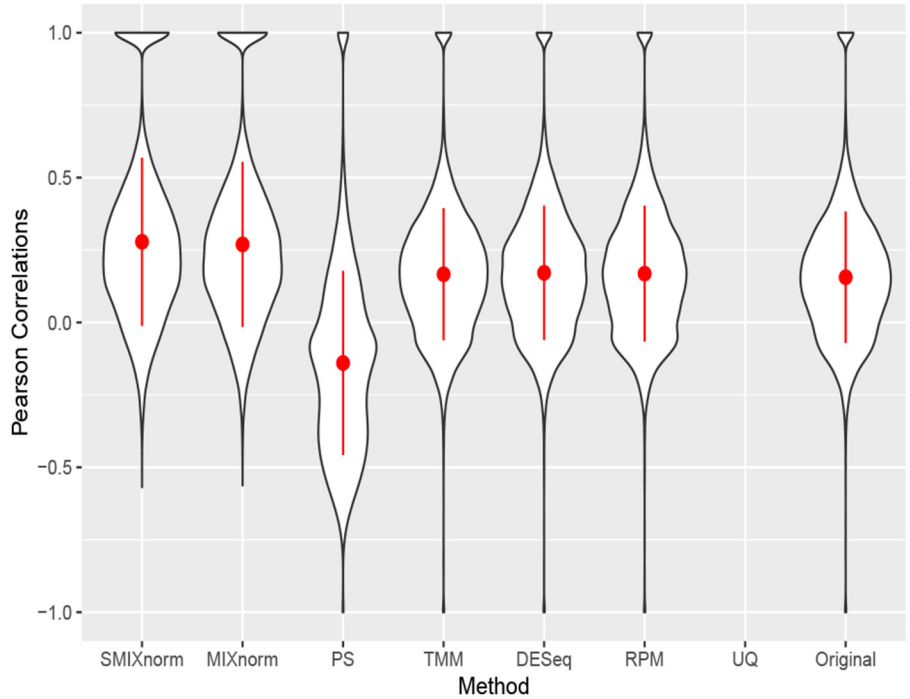


FIGURE 5 | Gene-wise correlations between normalized FFPE and FF expression for soft tissue sarcomas data on all 20, 242 protein coding genes. The UQ method failed to normalize the data due to excess zero counts.

TABLE 2 | Gene-wise correlations between normalized FFPE and FF expression for soft tissue sarcomas data on the CINSARC gene signature.

	First Qu.	Median	Third Qu.
SMIXnorm	0.344	0.465	0.529
MIXnorm	0.333	0.455	0.517
DESeq	0.165	0.260	0.354
RPM	0.146	0.243	0.350
TMM	0.010	0.098	0.161
PS	−0.126	0.002	0.154
UQ	–	–	–
Original	0.020	0.107	0.181

The UQ method failed to normalize the data due to excess zero counts.

calculated without any normalization. However, DESeq and RPM show much higher correlations on the CINSARC gene signature compared to TMM. Overall, SMIXnorm and MIXnorm show similar results and are consistently better than the other methods. We further note that in this example, a poor choice of normalization method (e.g., UQ, TMM, or PS) may yield results worse than those from original unnormalized data. MIXnorm takes about 5 min to normalize the soft tissue sarcomas data and we note that the algorithm reaches the default maximum number of iterations before convergence. On the other hand, SMIXnorm converges in 6 iterations and takes about 10 s to normalize the data.

3.2.3. Clear Cell Renal Cell Carcinoma Data

The clear cell renal cell carcinoma (ccRCC) data are available in the repository Gene Expression Omnibus from a published study (Eikrem et al., 2016). ccRCC is the most common and aggressive histological type among the primary renal neoplasms. Metastasis is a major cause of ccRCC patient death due to the resistance to standard chemotherapy and radiotherapy. Hence, much effort has been made to unravel the underlying molecular mechanisms of ccRCC, for example, by applying gene expression analysis to develop molecular signatures of disease progression, which plays an important role in assessing the carcinogenesis and development of disease as well as guiding clinical decisions. The ccRCC RNA-seq data contain 32 pairs of FFPE and RNAlater samples with 18, 458 protein-coding genes converted from 64, 253 Ensembl annotated genes. Following Yin et al. (2020), the RNAlater samples were considered the gold standard in this study and gene-wise Pearson correlations were computed to compare the performance of the seven normalization methods. The results are shown via violin plots in **Figure 6**. Again, we observe that SMIXnorm and MIXnorm give almost the same results and consistently perform the best among all methods. It is interesting to note that TMM, which performs the best among existing FF normalization methods in soft tissue sarcomas data on all protein-coding genes, gives worse results than DESeq, RPM, and UQ in this application. In fact, TMM and PS show no advantage compared to the original correlations without any normalization. MIXnorm takes about 10.5 s to normalize the ccRCC FFPE data, while SMIXnorm only takes about 3.3 s.

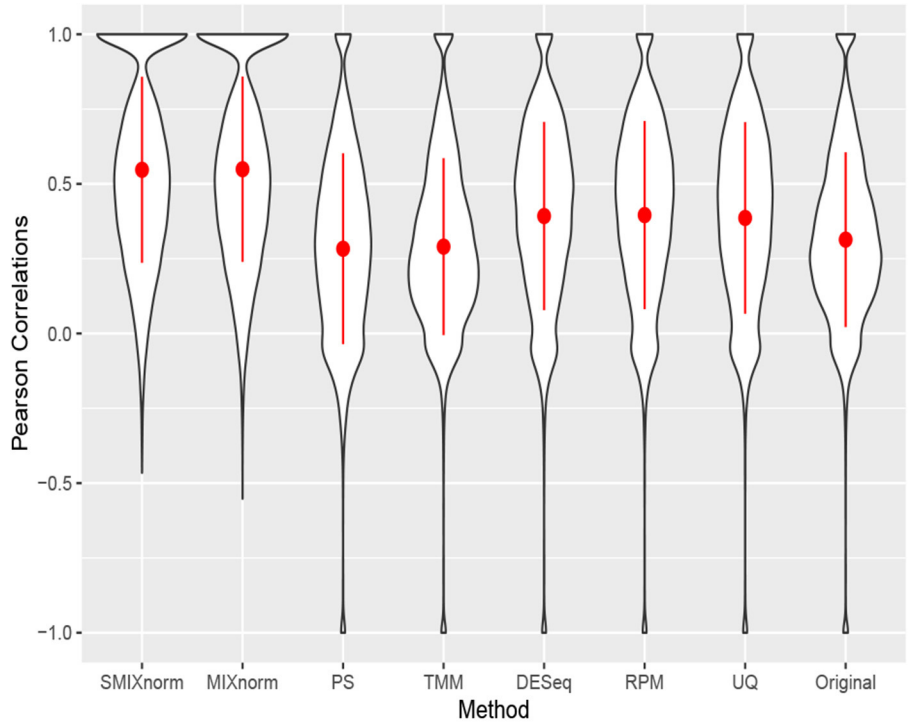


FIGURE 6 | Gene-wise correlations between normalized FFPE and RNAlater for ccRCC data on 18, 458 protein coding genes.

The paired design of the ccRCC study allows us to conduct differential expression (DE) analysis between ccRCC and normal tissues. Specifically, two ccRCC and two adjacent normal tissues were obtained from each of the 16 patient. The two pairs were then stored in FFPE and RNAlater, respectively. We identified DE genes [Benjamini–Hochberg adjusted $P < 0.05$ from paired t-tests and absolute log 2 fold change (FC) $>$ 2] from FF and FFPE samples based on normalized expression levels and summarize results in **Table 3**. We find that SMIXnorm and MIXnorm give similar numbers of common DE genes from the two sources, and are much higher than the numbers from the other methods. Among the common DE genes identified from MIXnorm, SMIXnorm is able to identify 999 out of the 1, 036 genes (i.e., 96.4%). Furthermore, among the two sets of top 20 DE genes identified from RNAlater and FFPE samples, SMIXnorm and MIXnorm share the same common genes and show almost identical log 2 FC (**Table 4**). Our analysis shows that using either SMIXnorm or MIXnorm for normalization, we are able to detect a set of up- or down-regulated genes from FFPE RNA-seq data that is similar to that from FF or like RNA-seq data.

TABLE 3 | Summary of differential expression analysis based on different normalization methods from the ccRCC data.

	FFPE DE genes	RNAlater DE genes	Common DE genes	Common top 20 DE
SMIXnorm	1,490	1,486	1,023	13
MIXnorm	1,488	1,482	1,036	13
DESeq	1,014	951	680	7
RPM	999	926	676	9
TMM	1,073	1,067	632	7
PS	1,001	1,300	652	8
UQ	1,002	943	679	8
Original	1,041	1,096	646	9

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; the last column is the number of common genes among the two sets of top 20 DE genes from FFPE and RNAlater.

4. DISCUSSIONS

We have developed an efficient normalization method, named SMIXnorm, for FFPE RNA-seq data normalization. Modified from the MIXnorm statistical model, we use a similar two-component mixture model to separately model the expressed and non-expressed genes. The simplifications of the statistical model avoid the complex likelihood function and the need of a complicated latent variable structure to invoke the nested EM algorithm. We have shown through real data applications and simulation studies that SMIXnorm greatly reduces the complexity of the likelihood function and the computing time without sacrificing the performance. Though other FF normalization methods take only about 1 s to normalize each dataset in our three data applications, their performance is not comparable to that of SMIXnorm and MIXnorm for FFPE

TABLE 4 | Shared DE genes among the two sets of top 20 DE genes from FFPE and RNAlater samples in the ccRCC data, ordered by the absolute value of the SMIXnorm normalized RNAlater log2 FC.

	SMIXnorm RNAlater log2 FC	MIXnorm RNAlater log2 FC	SMIXnorm FFPE log2 FC	MIXnorm FFPE log2 FC
CA9	8.02	8.04	5.66	5.66
SLC6A3	7.20	7.22	6.30	6.31
NDUFA4L2	6.38	6.39	4.88	4.89
UMOD	−6.17	−6.15	−5.63	−5.62
GP2	−5.53	−5.51	−4.96	−4.96
CLCNKA	−5.29	−5.28	−5.70	−5.69
CDCA2	5.21	5.23	5.04	5.05
TNFAIP6	5.16	5.17	5.45	5.45
SLC4A11	−5.10	−5.08	−5.23	−5.22
KNG1	−5.04	−5.02	−5.04	−5.03
SLC12A1	−4.96	−4.95	−4.90	−4.89
AQP2	−4.94	−4.92	−4.90	−4.89
NELL1	−4.79	−4.77	−5.03	−5.02

samples. Some FF normalization methods perform even worse than the use of original data without any normalization.

We mentioned in the soft tissue sarcomas application that MIXnorm failed to converge at the default maximum number of iterations and tolerance level. Due to the severely different RNA degradation levels among the 41 soft tissue sarcomas FFPE samples, 10 FFPE samples have more than 10,000 zero counts and 9 FFPE samples have less than 3,000 zero counts. Consequently, MIXnorm gives negative estimated sample-specific location parameters of the truncated normal distributions for those samples with a higher proportion of zero counts, which blurs the distinction between expressed and non-expressed genes. SMIXnorm models the expressed genes by normal distributions without truncation, which naturally constrains the location parameters to be non-negative in all the EM iterations as sample means are used for estimation. Thus, SMIXnorm seems to be more robust and converges faster than MIXnorm.

Recently, single-cell RNA sequencing (scRNA-seq) becomes an important technology in molecular analysis. While bulk RNA-seq measures the expression in the population level across cells, scRNA-seq allows for the cell level resolution and therefore, reveals heterogeneity of cell subpopulations. Similar to FFPE RNA-seq data, a prominent feature of scRNA-seq data is the sparsity. The high proportion of zero count arises for both biological reasons and technical reasons (Lun et al., 2016; Vallejos et al., 2017). The most commonly used scRNA-seq normalization method is SCnorm (Bacher et al., 2017), which uses quantile regression to group genes by estimated count-depth relationships, and then estimates different scaling factors within each group via a second quantile regression. Other popular scRNA-seq normalization methods, including BASiCS (Vallejos et al., 2015), SAMstrt (Katayama et al., 2013), and GRM (Vallejos et al., 2017), rely on spike-ins. Therefore, most scRNA-seq normalization methods are not directly applicable to the FFPE RNA-seq data that typically do not have spike-ins.

With the rapid adaption of FFPE samples in RNA-seq analysis, it is important for users to realize that different normalization procedures should be used for FFPE vs. FF data. We offer *RSeqNorm*, a comprehensive and user-friendly normalization toolkit for RNA-seq data. To the best of our knowledge, *RSeqNorm* is the only available web-based tool that integrates different normalization methods for both FFPE and FF RNA-seq data. It includes seven normalization methods, among which five are commonly used (RPM, UQ, DESeq, TMM, and PS) for FF or like RNA-seq data. Though MIXnorm and SMIXnorm are specifically designed for FFPE RNA-seq normalization, they can be applied to FF data directly. However, it is generally inappropriate to implement other existing methods on FFPE data. The input for all methods is a read count matrix at the gene level. The output is an expression matrix after normalization with the same dimension as the input data. The R package, RSEQNORM, which implements SMIXnorm and MIXnorm is downloadable from the website (http://lce.biohpc.swmed.edu/rseqnorm).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SY and XW proposed and implemented the SMIXnorm method. SY, XZ, XW, and GX performed data analysis and wrote the manuscript. XZ, GX, and YX provided resources and helpful discussions. SY, XZ, and BY developed the web portal *RSeqNorm*. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been partially supported by the NIH grant [R15GM131390, R35 GM136375, P50CA70907, P30CA142543, 1R01GM115473, and 1R01CA172211], and the Cancer Prevention and Research Institute of Texas [RP190107 and RP180805].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.650795/full#supplementary-material

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14, 584–586. doi: 10.1038/nmeth.4263

Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14, 671–683. doi: 10.1093/bib/bbs046

Eikrem, O., Beisland, C., Hjelle, K., Flatberg, A., Scherer, A., Landolt, L., 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. doi: 10.1371/journal.pone.0149743

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* 19, 776–792. doi: 10.1093/bib/bbx008

Graw, S., Meier, R., Minn, K., Bloomer, C., Godwin, A. K., Fridley, B., et al. (2015). Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci. Rep.* 5:12335. doi: 10.1038/srep12335

Grenier, J. K., Foureman, P. A., Sloma, E. A., and Miller, A. D. (2017). RNA-seq transcriptome analysis of formalin fixed, paraffin-embedded canine meningioma. *PLoS ONE* 12:e0187150. doi: 10.1371/journal.pone.0187150

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Katayama, S., Töhönen, V., Linnarsson, S., and Kere, J. (2013). Samstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 29, 2943–2945. doi: 10.1093/bioinformatics/btt511

Lesluyes, T., Pérot, G., Largeau, M. R., Brulard, C., Lagarde, P., Dapremont, V., et al. (2016). RNA sequencing validation of the complexity index in sarcomas prognostic signature. *Eur. J. Cancer* 57, 104–111. doi: 10.1016/j.ejca.2015.12.027

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13, 523–538. doi: 10.1093/biostatistics/kxr031

Li, P., Conley, A., Zhang, H., and Kim, H. L. (2014). Whole-transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq. *BMC Genomics* 15:1087. doi: 10.1186/1471-2164-15-1087

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Morton, M. L., Bai, X., Merry, C. R., Linden, P. A., Khalil, A. M., Leidner, R. S., 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. doi: 10.1016/j.lungcan.2014.03.020

Omolo, B., Yang, M., Lo, F. Y., Schell, M. J., Austin, S., Howard, K., et al. (2016). Adaptation of a RAS pathway activation signature from FF to FFPE tissues in colorectal cancer. *BMC Med. Genomics* 9:65. doi: 10.1186/s12920-016-0225-2

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34, 2870–2878. doi: 10.1093/bioinformatics/bty175

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). Basics: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* 11:e1004333. doi: 10.1371/journal.pcbi.1004333

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14:565. doi: 10.1038/nmeth.4292

Von Ahlfen, S., Missel, A., Bendrat, K., and Schlumpberger, M. (2007). Determinants of RNA quality from FFPE samples. *PLoS ONE* 2:e1261. doi: 10.1371/journal.pone.0001261

Yin, S., Wang, X., Jia, G., and Xie, Y. (2020). MIXnorm: normalizing RNA-seq data from formalin-fixed paraffin-embedded samples. *Bioinformatics* 36,3401–3408. doi: 10.1093/bioinformatics/btaa153

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.