Comparison of Normalization Methods for Analysis of TempO-Seq Targeted RNA Sequencing Data

Pierre R. Bushel1,2,*, Stephen S. Ferguson3, Sreenivasa C. Ramaiyahgar1, Richard S. Paules3 and Scott S. Auerbach3

1 Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences of National Institutes of Health, Durham, NC, United States, 2 Massive Genome Informatics Group, National Institute of Environmental Health Sciences National Institutes of Health, Durham, NC, United States, 3 Biomolecular Screening Branch, National Institute of Environmental Health Sciences of National Institutes of Health, Durham, NC, United States

Analysis of bulk RNA sequencing (RNA-Seq) data is a valuable tool to understand transcription at the genome scale. Targeted sequencing of RNA has emerged as a practical means of assessing the majority of the transcriptomic space with less reliance on large resources for consumables and bioinformatics. TempO-Seq is a templated, multiplexed RNA-Seq platform that interrogates a panel of sentinel genes representative of genome-wide transcription. Nuances of the technology require proper preprocessing of the data. Various methods have been proposed and compared for normalizing bulk RNA-Seq data, but there has been little to no investigation of how the methods perform on TempO-Seq data. We simulated count data into two groups (treated vs. untreated) at seven-fold change (FC) levels (including no change) using control samples from human HepaRG cells run on TempO-Seq and normalized the data using seven normalization methods. Upper Quartile (UQ) performed the best with regard to maintaining FC levels as detected by a limma contrast between treated vs. untreated groups. For all FC levels, specificity of the UQ normalization was greater than 0.84 and sensitivity greater than 0.90 except for the no change and +1.5 levels. Furthermore, K-means clustering of the simulated genes normalized by UQ agreed the most with the FC assignments [adjusted Rand index (ARI) = 0.67]. Despite having an assumption of the majority of genes being unchanged, the DESeq2 scaling factors normalization method performed reasonably well as did simple normalization procedures counts per million (CPM) and total counts (TCs). These results suggest that for two class comparisons of TempO-Seq data, UQ, CPM, TC, or DESeq2 normalization should provide reasonably reliable results at absolute FC levels ≥ 2.0. These findings will help guide researchers to normalize TempO-Seq gene expression data for more reliable results.

Keywords: TempO-Seq, normalization, gene expression, mRNA, transcription

INTRODUCTION

Over the past 25 years, interrogation of genome-wide gene expression has taken many forms. cDNA and oligonucleotide microarrays (Millen and Glauser, 1978; Lockhart et al., 1996) analysis methods matured over time whereby preprocessing of the data for single-channel microarrays ultimately defaulted to the de facto Robust Multichip Average (RMA) normalization (Irizarry et al., 2003a,b). The advent of massive parallel signature sequencing (MPSS) and next-generation sequencing by
synthesis for mRNA (RNA-Seq) ushered in a new paradigm for whole transcriptome analysis (Bainbridge et al., 2006; Cloonan et al., 2008; Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). Crowdsourcing bioinformatics analysis of RNA-Seq data through the US Food and Drug Administration MicroArray Quality Control (MAQC), SEquence Quality Control (SEQC) phase effort led to a comprehensive assessment of RNA-Seq analysis including comparison to microarray and normalization using External RNA Control Consortium (ERCC) spike-in controls (Consortium, 2014; Rissou et al., 2014; Wang et al., 2014; Xu et al., 2014). In addition, several studies have compared various normalization approaches for RNA-Seq data (Dillies et al., 2013; Zyprych-Walczak et al., 2015; Lin et al., 2016; Li et al., 2017). Proper normalization of gene expression data is essential to ensure valid and reliable results from downstream analyses (Park et al., 2003).

In the last few years, targeted sequencing of RNA has emerged as a practical means of capturing the totality of the transcriptomic space with less reliance on large resources for consumables and bioinformatics (Li et al., 2012). The TempO-Seq™ technology from BioSpyder™ is a templated, multiplexed RNA-Seq platform that measures the expression of sentinel genes representative of genome-wide transcription (Yeakley et al., 2017; Mav et al., 2018). A few advantages of TempO-Seq over RNA-Seq is that it does not require RNA purification, cDNA synthesis, nor capture of targeted RNA. In addition, by nature of the technology, it lacks 3’ end bias. Recently, several studies utilized the TempO-Seq platform for whole transcriptome profiling, primarily for toxicogenomics (Grimm et al., 2016; House et al., 2017; Yeakley et al., 2017; Bushel et al., 2018; Limonciel et al., 2018; Chappell et al., 2019; Ramaihagari et al., 2019; Simon et al., 2019), but also carcinogenomics (Batai et al., 2018; Hanke et al., 2018), and to profile formalin-fixed paraffin-embedded (FFPE) tissue (Trejo et al., 2019). However, there has not been a comprehensive comparison of normalization methods applied to TempO-Seq data.

Here we utilize control samples from human HepaRG cells interrogated on the TempO-Seq platform to simulate gene expression data at seven-fold change (FC) levels including no interrogated on the TempO-Seq platform to simulate gene expression data. TempO-Seq Analysis

Tempo-Seq analysis was performed as previously described by Biospyder, Inc. (Yeakley et al., 2017). Briefly, frozen lysates were thawed, and sequencing libraries for targeted panels of transcripts were generated. Each detector oligonucleotide (DO) consisted of complementary sequence to specific mRNA targets plus a universal primer binding site. Ligation of detector oligonucleotides via PCR amplification introduces adaptors required for sequencing and well-specific “barcodes” that link sequencing data to a specific well of origin. Barcode sequences flank the target sequence and are inserted into standard illumina adaptors to permit dual-index sequencing and deconvolution of sample-specific reads using standard illumina software. All PCR-amplified and barcoded samples were pooled into a single library for sequencing on a HighSeq 2500 sequencer (Illumina Inc., San Diego, CA) using a 50 cycle single-end read flow cell. Processing of the sequencing data was conducted using Illumina’s bcl2fastq software employing default parameter settings allowing for 1 mismatch per read. Sequencing reads were de-multiplexed using standard instrument software for each sample using barcodes to give FASTQ files linked to each well. Downsampled data was generated to obtain 500 mapped reads per gene on average. The 50 bp reads in the fastq files were aligned using bowtie version 1.2.2 (using parameters: -v 3 -k 1 -m 1 --best --strata --trim3 1) to a manifest of the TempO-Seq target genes sequences (a subset of the human transcriptome [Refseq release #356667]). Differentiated HepaRG cultures (2D-DIFF) were re-differentiated from cryopreserved suspension form over 10 days prior to vehicle exposures. Proliferated HepaRG cultures (PROLIF) were seeded at approximately 2,000 cells/well, grown for 3 days, and proliferated during vehicle exposures. For this, HepaRG were plated using William’s E medium (ThermoFisher, Catalog: A1217601) which was supplemented with MHPIT maintenance additive (Lonza, Catalog: MHPIT). Vehicle exposures durations were 96 h, and at the final time point, incubation media were removed and cultures were washed once with 50 μl phosphate-buffered saline (ThermoFisher, Catalog #10010023). Cells were subsequently lysed for high-throughput transcriptomics using 20 μl of 1 × Tempo-Seq lysis buffer (Biospyder) with a 15-min room temperature incubation with subsequent freezing at −80°C. Edge effects were minimized by excluding lysates in rows A, B, O, P, and Columns 1, 2, 23, and 24.

**Materials and Methods**

**Cell Culture**

HepaRG cells (Lonza, Catalog: NSHPRG) in cryopreserved form were thawed and seeded at approximately 20,000 cells/well onto collagen(I)-coated 384-well plates (Corning, Catalog 35136). Differentiated HepaRG cultures (2D-DIFF) were re-differentiated from cryopreserved suspension form over 10 days prior to vehicle exposures. Proliferated HepaRG cultures (PROLIF) were seeded at approximately 2,000 cells/well, grown for 3 days, and proliferated during vehicle exposures. For this, HepaRG were plated using William’s E medium (ThermoFisher, Catalog: A1217601) which was supplemented with MHPIT maintenance additive (Lonza, Catalog: MHPIT). Vehicle exposures durations were 96 h, and at the final time point, incubation media were removed and cultures were washed once with 50 μl phosphate-buffered saline (ThermoFisher, Catalog #10010023). Cells were subsequently lysed for high-throughput transcriptomics using 20 μl of 1 × Tempo-Seq lysis buffer (Biospyder) with a 15-min room temperature incubation with subsequent freezing at −80°C. Edge effects were minimized by excluding lysates in rows A, B, O, P, and Columns 1, 2, 23, and 24.

**Tempo-Seq Analysis**

Tempo-Seq analysis was performed as previously described by Biospyder, Inc. (Yeakley et al., 2017). Briefly, frozen lysates were thawed, and sequencing libraries for targeted panels of transcripts were generated. Each detector oligonucleotide (DO) consisted of complementary sequence to specific mRNA targets plus a universal primer binding site. Ligation of detector oligonucleotides via PCR amplification introduces adaptors required for sequencing and well-specific “barcodes” that link sequencing data to a specific well of origin. Barcode sequences flank the target sequence and are inserted into standard illumina adaptors to permit dual-index sequencing and deconvolution of sample-specific reads using standard illumina software. All PCR-amplified and barcoded samples were pooled into a single library for sequencing on a HighSeq 2500 sequencer (Illumina Inc., San Diego, CA) using a 50 cycle single-end read flow cell. Processing of the sequencing data was conducted using Illumina’s bcl2fastq software employing default parameter settings allowing for 1 mismatch per read. Sequencing reads were de-multiplexed using standard instrument software for each sample using barcodes to give FASTQ files linked to each well. Downsampled data was generated to obtain 500 mapped reads per gene on average. The 50 bp reads in the fastq files were aligned using bowtie version 1.2.2 (using parameters: -v 3 -k 1 -m 1 --best --strata --trim3 1) to a manifest of the TempO-Seq target genes sequences (a subset of the human transcriptome [Refseq release #356667]). Differentiated HepaRG cultures (2D-DIFF) were re-differentiated from cryopreserved suspension form over 10 days prior to vehicle exposures. Proliferated HepaRG cultures (PROLIF) were seeded at approximately 2,000 cells/well, grown for 3 days, and proliferated during vehicle exposures. For this, HepaRG were plated using William’s E medium (ThermoFisher, Catalog: A1217601) which was supplemented with MHPIT maintenance additive (Lonza, Catalog: MHPIT). Vehicle exposures durations were 96 h, and at the final time point, incubation media were removed and cultures were washed once with 50 μl phosphate-buffered saline (ThermoFisher, Catalog #10010023). Cells were subsequently lysed for high-throughput transcriptomics using 20 μl of 1 × Tempo-Seq lysis buffer (Biospyder) with a 15-min room temperature incubation with subsequent freezing at −80°C. Edge effects were minimized by excluding lysates in rows A, B, O, P, and Columns 1, 2, 23, and 24.

**Simulated Tempo-Seq Data**

Let \( Y_{gs} \) denote the read count of a gene \( g \in \{1, \ldots, G\} \) belonging to a group \( m \in \{1, 2\} \) of a sample \( S_\in \{S_1, S_2, \ldots, S_N\} \) such that:

\[
Y_{gs} \sim NB(\text{mean} = \mu_{gs}, \text{var} = \mu_{gs}(1 + \frac{\mu_{gs}}{\theta_{gs}}))
\]

where \( \theta_{gs} \) is the parameter measuring the dispersion in the data, \( \mu_{gs} \) is the true mean of the data, and NB is the negative binomial distribution. We set the dispersion parameter for each gene to be

\[\text{http://www.htslib.org}\]
the same for all samples. Thus, \( \theta_{gs} = \theta_g \). Here,

\[
\mu_{gsm} = E[Y_{gsm}] = \frac{\lambda_{gs}}{\sum_{g=1}^{G} \lambda_{gs}} M_{sm}
\]

where \( M_{sm} \) is the sequencing depth for the \( m \)th group in sample \( S \). \( M_{sm} = 1 \times 10^6 \) \( U_{sm} \) for \( U_{sm} \sim \text{Unif}[0,2,1.5] \) where Unif is the uniform distribution. The bounds (minimum and maximum limits) of Unif and the estimation of the sample mean \( \lambda^*_{gs} \) and \( \theta_g \) were obtained from a DESeq analysis (Anders and Huber, 2010) of TempO-Seq count data (Supplementary Data 1: 2,680 genes in the 75th percentile of counts and 240 HepaRG control cells in the 90th percentile of counts) generated as previously described (Ramaiahgari et al., 2019) where the read depth was down-sampled to approximately 500 mapped read counts per transcript (Yeakley et al., 2017). Then, to generate sets of simulated genes across the groups of samples, we defined \( \lambda_{gs} = \gamma_{gs} \lambda^*_{gs} \), where \( S_1 \) is considered the untreated group of samples, \( S_2 \) is considered a group of samples with a particular perturbation (treated) and the ratio

\[
\gamma_{gs} = \frac{w_{gs2}}{w_{gs1}}.
\]

The differential expression was simulated at different levels for the respective sets of genes using values of \( w_{gs} \) denoted in

### TABLE 1 | Parameter settings to simulate data FC.

| Set | 1  | 2  | FC  |
|-----|----|----|-----|
| A   | 6  | 1.5| −4  |
| B   | 3  | 1.5| −2  |
| C   | 2.25 |1.5 |−1.5 |
| D   | 1.5 |2.25 |+1.5 |
| E   | 1.5 |3   |+2  |
| F   | 1.5 |6   |+4  |
| G   | 1.5 |1.5 |0   |

Parameters for the simulation model: \( N = 50 \) genes per FC set, \( 35 \) samples per \( m \) group, library size lower limit = \( 0.2 \times 10^6 \), and library size upper limit = \( 1.5 \times 10^6 \). FC, fold change.

![FIGURE 1](image-url) | Representations of the estimates of the fold change (FC) values and simulation of data. (A) Dispersion from the DESeq negative binomial model. Dispersions less than or equal to 3.0 are plotted. (B) Distribution of the log₂ FC estimated from a binomial test of the control samples randomly assigned to two groups (control1 and control2). Values within the range [-0.8, +0.8] are plotted. (C) Mapped reads of the samples from the simulation (untreated and treated). (D) Heat map of the log₂ FC (ratio of treated to average of the untreated with an offset of 1) from 350 simulated genes (50 from each set in Table 1).
Table 1. For each FC set, 50 genes were simulated and for each group $m$, 35 samples were generated (Supplementary Data 2).

**Normalizations**

The following normalizations were applied to the simulated data. Since the TempO-Seq platform is designed of genes that capture greater than 90% of the biological pathways, most of the genes are likely to be differentially expressed. Thus, we normalized the data using the following methods:

**Total Counts**

The counts per gene were normalized to TCs by dividing it by the total number of mapped reads per sample and multiplying by the mean total count across all the samples (Dillies et al., 2013). The TC normalized data were then transformed with log$_2$ using an offset of 1.

**Counts per Million**

The counts per gene were normalized to CPM by dividing it by the total number of mapped reads per sample and multiplying by $1 \times 10^6$ (Robinson et al., 2010). The CPM normalized data were then transformed with log$_2$ using an offset of 1.

**Median**

The counts per gene were Median normalized (Dillies et al., 2013) by dividing it by the median of mapped reads for all the samples and multiplying by $1 \times 10^6$. The Median normalized data were then transformed with log$_2$ using an offset of 1.

**Quantile**

The counts per gene were quantile (Q) normalized using the `normalizeQuantiles` function in the Bioconductor package limma (Ritchie et al., 2015). The method normalizes the counts of the genes in a sample to have the same quantiles across the samples in the data set. If there are ties among the genes for a particular sample, then the ties are normalized to the same value (i.e., the average of the quantiles for the tied values).

**Upper Quartile**

The counts per gene were UQ normalized using the `calcNormFactors` function in the Bioconductor package EdgeR (Robinson et al., 2010; McCarthy et al., 2012) using the 75th percentile of the read counts that are mapped per sample. These scaling factors are then used to adjust the total mapped reads count for each sample.

The following normalization methods have an assumption that the majority of the genes on the platform are unchanged.

**Trimmed Mean of M Values**

The counts per gene were normalized using the “weighted” Trimmed Mean of M-values (TMM) approach (Robinson and Oshlack, 2010) in the Bioconductor package EdgeR. After trimming the data [5% for the A values, log ratio 0.3 for the M values to a reference array (the library whose upper quartile is closest to the mean upper quartile)], scaling factors for each sample were generated using the `calcNormFactors` function. Scaling factors were then used to adjust the total mapped reads count from each sample.

**DESeq2**

The counts per gene were normalized using the `estimateSizeFactors` function in the Bioconductor package DESeq2 (Love et al., 2014). The counts for each gene in each
sample is divided by the geometric mean of the gene across all samples. The median of the ratios for the genes in a sample is the estimated size “scaling” size factor used to adjust the total mapped reads count from each sample.

**Performance of the Normalizations**

**Sensitivity**
Sensitivity is the probability of the normalization of the read counts maintaining the genes’ limma-derived FC values.

\[
\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.
\]

**Specificity**
Specificity is the probability of the normalization of the read counts not falsely altering genes’ limma-derived FC values.

\[
\text{Specificity} = \frac{\text{True Negatives}}{\text{False Positives} + \text{True Negatives}}.
\]

**Precision**
Precision is the proportion of limma-derived FC values predicted correctly.

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}.
\]
RESULTS

Simulation of Fold Change Data

Using the RNA extracted from 240 wells containing HepaRG control samples and interrogated on the TempO-Seq platform, we simulated count data for treated (perturbed) and untreated groups (35 samples in each) using a negative binomial distribution with mean and dispersion estimates randomly sampled from the sentinel genes in the control samples. Table 1 lists the parameters used to generate seven FC sets of genes (A – G) from a comparison of group m = 2 vs. m = 1 denoting simulated treated samples vs. untreated samples. Sets A - F representing -4, -2, -1.5, +1.5, +2, and +4 FC, respectively, contain 50 simulated genes each, while set G with 50 simulated genes also represents no FC. The proportion of genes satisfies the assumption that the majority of genes are differentially expressed in a perturbed biological system with samples assayed on the TempO-Seq targeted platform. Thus, the comparisons of the normalization approaches would be more representative of the FC distribution of the genes in a typical targeted TempO-Seq analysis.

As shown in Figure 1A, the majority of the transcripts from the control samples have dispersion estimates less than 0.2 and thus have variances close to the mean. The 2,680 fitted genes from a binomial test of control samples randomly divided into two groups have predominantly a log₂ FC approximately equal to 0 (Figure 1B). The heterogeneity of the 240 HepaRG samples is observed by the differential expression of some of the genes between the two groups of controls. Following the perturbations to generate the FC sets, the read depth of the samples ranged between $1.3 \times 10^6$ and $1.35 \times 10^6$ for the untreated and $1.47 \times 10^6$ and $1.53 \times 10^6$ for the treated (Figure 1C). The FC range and variability of the genes modeled in the simulation are visualized in a heat map diagram (Figure 1D).

Negligible Normalization Impact on the Distribution and Variance of the Simulated Data

We used seven common normalization methods for RNA-Seq data to compare the effect on the simulated data. Five normalization methods, UQ, Median, CPM, TCs, and Q, adjust the data within sample, whereas two normalization methods, TMM and DESeq2, adjust the data within and between samples. The two latter methods have assumptions that the majority of the genes are not changed. This does not typically hold true for TempO-Seq data since the platform is designed with sentinel genes that capture the totality of the transcriptomic space (Mav et al., 2018). Despite a possible violation of the assumption, we included these two methods in the comparison since at least one has been recently used on TempO-Seq data (House et al., 2017). As shown in Figure 2A, the distributions of the log₂ normalized count data for the untreated samples are relatively tight with short whiskers and consistent across the samples. The Q normalization appears to remove a fair amount of
FIGURE 5 | Comparison of limma fold change (FC) values. (A) Percent of differentially expressed genes as upregulated (red), downregulated (blue), or no change (gray) from the limma model comparing treated to untreated. (B) Correlation of log2 FC values from the limma model.
the variability across the samples. The Median and CPM normalized data shift the median of the data lower than the other normalizations. The distributions of the log$_2$ normalized data shift the median of the data lower than the variability across the samples. The Median and CPM normalization methods were more precise at either one or both extremes of the FC spectrum, but except with Q normalization. Although the Pearson correlation ($r$) values are greater than +0.98 (Table 2), there are visually some simulated genes with FC values that are impacted differently by the normalizations (Figure 5A). Q normalization has the greatest impact on the FC values.

In terms of the performance of the normalizations maintaining the expected FC level, UQ, CPM, TC, and DESeq2 normalizations had high sensitivity (greater than 0.840) at absolute FC levels greater than or equal to 2.0 (Table 3). Other normalizations performed better sensitivity-wise at either end of the FC spectrum. Specificity of the normalizations was reasonably high for all of the methods. DESeq2 had the highest accuracy overall (0.694) in maintaining the FC levels, followed by all the other methods. Q normalization which had the worst accuracy = 0.269. The precision of the normalization methods revealed that all performed well at either one or both extremes of the FC spectrum, but Median and UQ normalization methods were more precise at most FC levels.

To test the agreement of the normalized data with the FC group assignment, we K-means clustered the genes using the log$_2$ ratio values of the treated to the average of the untreated with $K = 7$. The ARI measures the amount of agreement between the genes in the clusters and their FC set assignment. The ARI ranges from 0 to 1, where 0 defines the agreement is essentially random, and 1 indicates that the agreement is perfect. As shown in Table 4, UQ normalization had the highest ARI score of 0.67, followed by TC, CPM, DESeq2, and then Median. TMM normalization was subpar in agreement, and not surprisingly, Q normalization had the worst agreement overall.

**Normalization Impact Affecting Fold Change**

To assess the effect of normalization on the FC estimate of the simulated data, we used limma to test the comparison of treated vs. untreated. The percent of up and down differentially expressed genes varied according to normalization (Figure 5A). TMM normalization had more upregulated genes than downregulated genes, whereas the converse was true for UQ which had more downregulated genes than upregulated genes. CPM, TC, and DESeq2 had relatively the same proportion of differential genes. Q normalization had the most non-changed genes, while Median normalization had approximately the expected proportion of up, down, and non-changed genes.

The limma FC estimates for each normalization were then binned as follows to compare the similarity of the normalizations:

- $\text{bin - 4 : } (-\infty, -3.0]$, $\text{bin - 2 : } (-3.0, -2.0]$, $\text{bin : } (-1.5, -2.0, -1.5)$, $\text{bin 0 : } (-1.5, +1.5)$, $\text{bin + 1.5 : } [+1.5, +2.0])$, $\text{bin + 2 : } [+2.0, +3.0)$, $\text{bin + 4 : } [+3.0, \infty)$.

As shown in Figure 5B, the log$_2$ FC estimates from limma for each normalization method correlated very well with each other except with Q normalization. The various gene expression platforms that researchers rely on for whole genome transcriptomics have their own de facto normalization method that are preferred by analysts. With single-channel microarray, RMA became the standard normalization approach. For bulk RNA-seq, many users gravitated to DESeq2 for normalization. For single-cell RNA-Seq, the path forward continues to emerge. In the case of newer platforms such as TempO-Seq targeted RNA sequencing, there has not been an evaluation of the performance of several normalization methods. Recent publications using the TempO-Seq platform used DESeq2.

## Table 2: Correlations of log$_2$ FC from limma.

| Normalization | TMM | UQ | Median | CPM | TC | DESeq2 | Q |
|---------------|-----|----|--------|-----|----|--------|---|
| TMM           | 1   | 0.997734 | 0.998682 | 0.998683 | 0.998683 | 0.998680 | 0.984968 |
| UQ            | 0.997734 | 1 | 0.999782 | 0.999782 | 0.999782 | 0.999783 | 0.985386 |
| Median        | 0.998682 | 0.999782 | 1 | 1.000000 | 1.000000 | 1.000000 | 0.985765 |
| CPM           | 0.998683 | 0.999782 | 1.000000 | 1 | 1.000000 | 1.000000 | 0.985765 |
| TC            | 0.998683 | 0.999782 | 1.000000 | 1.000000 | 1 | 1.000000 | 0.985765 |
| DESeq2        | 0.998680 | 0.999783 | 1.000000 | 1.000000 | 1.000000 | 1 | 0.985764 |
| Q             | 0.984968 | 0.985386 | 0.985765 | 0.985765 | 0.985765 | 0.985765 | 1 |

*CPM, counts per million; FC, fold change; Q, quantile; TC, total count; TMM, Trimmed Mean of M-values; UQ, upper quartile.*
TABLE 3 | Performance of normalizations.

| FC  | CPM  | TC   | Median | Q   | UQ   | DESeq2 | TMM |
|-----|------|------|--------|-----|------|--------|-----|
| Accuracy | Overall | 0.688 | 0.688 | 0.683 | 0.269 | 0.683 | 0.694 | 0.566 |
| Sensitivity | −4 | 0.980 | 0.980 | 1.000 | 1.000 | 0.962 | 0.980 | 1.000 |
| | −2 | 0.950 | 0.950 | 0.600 | 0.070 | 0.898 | 0.947 | 0.105 |
| | −1.5 | 0.776 | 0.776 | 0.115 | 0.114 | 0.915 | 0.326 | 0.000 |
| | 0.0 | 0.461 | 0.461 | 0.475 | 0.287 | 0.448 | 0.461 | 0.372 |
| | 1.5 | 0.174 | 0.174 | 0.840 | 0.154 | 0.045 | 0.653 | 0.833 |
| | 2.0 | 0.846 | 0.846 | 0.955 | 0.108 | 1.000 | 0.941 | 0.652 |
| | 4.0 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.907 |
| Specificity | −4 | 1.000 | 1.000 | 0.997 | 0.888 | 1.000 | 1.000 | 0.949 |
| | −2 | 0.961 | 0.961 | 0.864 | 0.847 | 0.980 | 0.903 | 0.855 |
| | −1.5 | 0.960 | 0.960 | 0.852 | 0.854 | 0.977 | 0.885 | 0.846 |
| | 0.0 | 0.988 | 0.988 | 0.992 | 0.989 | 0.988 | 0.988 | 0.978 |
| | 1.5 | 0.862 | 0.862 | 0.973 | 0.859 | 0.843 | 0.940 | 0.922 |
| | 2.0 | 0.884 | 0.884 | 0.974 | 0.853 | 0.867 | 0.943 | 0.975 |
| | 4.0 | 0.997 | 0.997 | 0.997 | 0.901 | 0.997 | 0.997 | 0.997 |
| Precision | −4 | 1.000 | 1.000 | 0.980 | 0.240 | 1.000 | 1.000 | 0.680 |
| | −2 | 0.760 | 0.760 | 0.060 | 0.060 | 0.880 | 0.360 | 0.040 |
| | −1.5 | 0.760 | 0.760 | 0.120 | 0.080 | 0.860 | 0.300 | 0.000 |
| | 0.0 | 0.940 | 0.940 | 0.960 | 0.960 | 0.940 | 0.940 | 0.900 |
| | 1.5 | 0.160 | 0.160 | 0.840 | 0.120 | 0.040 | 0.640 | 0.500 |
| | 2.0 | 0.220 | 0.220 | 0.840 | 0.080 | 0.080 | 0.640 | 0.860 |
| | 4.0 | 0.980 | 0.980 | 0.980 | 0.340 | 0.980 | 0.980 | 0.980 |

CPM, counts per million; FC, fold change; Q, quantile; TC, total count; TMM, Trimmed Mean of M-values; UQ, upper quartile.

or CPM for normalization (Grimm et al., 2016; House et al., 2017; Yeakley et al., 2017; Batai et al., 2018; Bushel et al., 2018; Hanke et al., 2018; Limonciel et al., 2018; Chappell et al., 2019; Ramaiahgari et al., 2019; Simon et al., 2019). In this study, we compared seven normalization methods using simulated data from human HepaRG control cells to determine which methods maintained genes at seven assigned FC levels (Table 1). We found that based on sensitivity, specificity, precision and accuracy performance metrics (Table 3) as well as the ARI that assessed the FC group assignments (Table 4), UQ at the 75th percentile of genes performed best. UQ performed well in comparisons of
performed reasonably well in normalizing the simulated TempO-Seq data at an absolute FC level ≥2.0, with UQ being the best overall. Below that threshold, the performances suffered. Furthermore, our analysis is for a two-class comparison of groups. More investigation is needed to determine which normalization is superior when comparing more groups or data with a dose or time dependency. Our analysis focused on simulated data from control HepaRG wells and not treated cells. We simulated various FC levels to mimic perturbations, but we assumed that the dispersion for each gene was the same across the two groups. It may be the case that genes in a perturbed system might have dispersion and other statistical properties different from the control cells which may presumably impact the normalization. Note that the Temp-O-Seq platform that we elevated the normalization methods is a targeted, human version with about 3,000 probes/transcripts as content. Normalizations for data from whole genome Temp-O-Seq platforms and/or targeted content from other species may perform differently. Finally, the simulated data used DESeq, the previous version of DESeq2, for estimating the dispersion and mean of the genes. This may have provided an unfair advantage to the DESeq2 performance in the comparison. Despite the aforementioned limitations, our results shed some light on the utilization of various standard methods for normalization of TempO-Seq data and that if used in a proper way, several choices will hopefully provide reliable analysis results. Future work will investigate more sophisticated normalization methods for TempO-Seq data and concentrate on data generated from whole genome platforms, other species’ gene content, and factorial or series experimental designs.

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
SF and SR generated HepaRG cultures and TempO-Seq data. PB designed the simulation study, implemented the computations, performed the analyses, and wrote parts of the manuscript. SF and SA provided interpretation of the results and wrote part of the manuscript. RP provided interpretation of the results. All authors contributed to the article and approved the submitted version.

FUNDING
This research was supported in part by the National Institutes of Health, National Institute of Environmental Health Sciences.

ACKNOWLEDGMENTS
The authors thank Dr. Raja Jothi and Dr. Thomas Randall for their review of the manuscript and helpful comments.

SUPPLEMENTARY MATERIAL
The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00594/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
Bainbridge, M. N., Warren, R. L., Hirst, M., Romaniuk, T., Zeng, T., Go, A., et al. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7:246. doi: 10.1186/1471-2164-7-246
Batai, K., Imler, E., Pangilinan, J., Bell, R., Lwin, A., Price, E., et al. (2018). Whole-transcriptome sequencing identified gene expression signatures associated with aggressive clear cell renal cell carcinoma. Genes Cancer 9, 247–256. doi: 10.18632 GENESANDCANCER.183
Bushel, P. R., Paules, R. S., and Auerbach, S. S. (2018). A comparison of the Temp-O-Seq S1500+ Platform to RNA-Seq and microarray using rat liver mode of action samples. Front. Genet. 9:485. doi: 10.3389/fgene.2018.00485
Chappell, G. A., Rager, J. E., Wolf, J., Babic, M., Leblanc, K. J., Ring, C. L., et al. (2019). Comparison of gene expression responses in the small intestine of mice following exposure to 3 carcinogens using the s1500+ gene set informs a potential common adverse outcome pathway. Toxicol. Pathol. 47, 851–864. doi: 10.1177/0192623319873882
Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods 5, 613–619. doi: 10.1038/nmeth.1223
Consortium, S. M.-I. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 32, 903–914. doi: 10.1038/nbt.2957
Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeannounguet, 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
Grimm, F. A., Iwata, Y., Sirenko, O., Chappell, G. A., Wright, F. A., Reif, D. M., et al. (2016). A chemical-biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives. Green Chem. 18, 4407–4419. doi: 10.1039/c6gc01147k
Hanke, N. T., Imler, E., Marron, M. T., Seligmann, B. E., Garland, L. L., and Baker, A. F. (2018). Characterization of carfilzomib-resistant non-small cell lung cancer cell lines. J. Cancer Res. Clin. Oncol. 144, 1317–1327. doi: 10.1007/s00432-018-2662-0
House, J. S., Grimm, F. A., Jima, D. D., Zhou, Y. H., Rusyn, I., and Wright, F. A. (2017). A pipeline for high-throughput concentration response modeling of gene expression for toxicogenomics. Front. Genet. 8:168. doi: 10.3389/fgene.2017.00168
Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. 2, 193–218. doi: 10.1007/BF01908075
Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31:e15. doi: 10.1093/nar/gng015
Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003b). Exploration, normalization, and summaries of high density
Mccarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550. doi: 10.1186/s13059-014-0550-8

Mav, D., Shah, R. R., Howard, B. E., Auerbach, S. S., Bushel, P. R., Collins, J. B., et al. (2018). A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. PLoS One 13:e0191105. doi: 10.1371/journal.pone.0191105

Mccarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40, 4288–4297. doi: 10.1093/nar/gks042

Millen, J. E., and Glauser, F. L. (1978). A rapid, simple technic for changing endochondral tubes. Anesth. Analg. 57, 735–736.

Mortazavi, A., Williams, B. A., McCue, K., Wilt, V., Filiposa, M., et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628. doi: 10.1038/nmeth.1226

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349. doi: 10.1126/science.1158441

Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., and Simon, R. (2003). Evaluation of normalization methods for microarray data. BMC Bioinformatics 4:33. doi: 10.1186/1471-2105-4-33

Ramaiahgari, S. C., Auerbach, S. S., Saddler, T. O., Rice, J. R., Dunlap, P. E., Sipes, N. S., et al. (2019). The power of resolution: contextualized understanding of biological responses to liver injury chemicals using high-throughput transcriptomics and benchmark concentration modeling. Toxicol. Sci. 169, 553–566. doi: 10.1093/toxsci/kfa065

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. 32, 102. doi: 10.1038/nbt.2931

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. doi: 10.1093/bioinformatics/btp16

Robertson, 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

Simon, J. M., Pranajpe, S. R., Wolter, J. M., Salazar, G., and Zylka, M. J. (2019). High-throughput screening and classification of chemicals and their effects on neuronal gene expression using RASL-seq. Sci. Rep. 9:4529. doi: 10.1038/s41598-019-39016-5

Trejo, C. L., Babic, M., Imler, J., Gonzalez, M., Bibikov, S. I., Shepard, P. J., et al. (2019). Extraction-free whole transcriptome gene expression analysis of FFPE sections and histology-directed subareas of tissue. PLoS One 14:e0212031. doi: 10.1371/journal.pone.0212031

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., et al. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat. Biotechnol. 32, 926–932. doi: 10.1038/nbt.3001

Xu, J., Su, Z., Hong, H., Thierry-Mieg, J., Thierry-Mieg, D., Krell, D. P., et al. (2014). Cross-platform ultradepth transcriptomic profiling of human reference RNA samples by RNA-Seq. Sci. Data 1:140020. doi: 10.1038/sdata.2014.20

Yeakley, J. M., Shepard, P. J., Goyena, D. E., Vansteenhoven, H. C., Mccomb, J. D., and Seligmann, B. E. (2017). A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling. PLoS One 12:e0178302. doi: 10.1371/journal.pone.0178302

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001). Validating clustering for gene expression data. Bioinformatics 17, 309–318. doi: 10.1093/bioinformatics/17.4.309

Zyprych-Walczak, Z., Szabelska, A., Handschuh, L., Gorczak, K., Klamecka, K., Figlerowicz, M., et al. (2015). The impact of normalization methods on RNA-Seq data analysis. Biomed. Res. Int. 2015:621690. doi: 10.1155/2015/621690

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.

Copyright © 2020 Bushel, Ferguson, Ramaiahgari, Paules and Auerbach. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.