Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data

Aaron Lun^{1,*}

1 Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom

*** Email: aaron.lun@cruk.cam.ac.uk**

Abstract

Applying a log-transformation to normalized expression values is one of the most common procedures in exploratory analyses of single-cell RNA sequencing (scRNA-seq) data. Normalization removes systematic biases in sequencing coverage between cells, while the log-transformation ensures that downstream computational procedures operate on relative rather than absolute differences in expression. We show that the log-transformation can introduce systematic errors when cells vary in sequencing coverage, leading to spurious non-zero differences in expression and artificial population structure in simulations. We observe similar effects in real scRNA-seq data where the difference in transformed values between groups of cells is not an accurate proxy for the log-fold change. We provide some practical recommendations to overcome this effect and analytically derive an expression for a larger pseudo-count that controls the transformation-induced error to a specified threshold.

1 Background

Log-transformed expression values are widely used in exploratory analyses of single-cell RNA sequencing (scRNA-seq) data. This is driven by the simplicity of the log-transformation and the ease of interpretation of the transformed values. In

particular, differences between log-values can be used as proxies for log-fold changes in expression between cells, which are often more relevant than differences in the counts. This is useful for ensuring that relative rather than absolute differences in the counts are used in distance-based procedures like clustering or trajectory construction. The log-transformation also reduces the impact of stochastic fluctuations in the counts for high-abundance genes. This would otherwise result in large differences between cells that are mostly uninteresting as the fold changes between counts are small.

Despite its popularity, the log-transformation has a number of problems such as suboptimal variance stabilization and arbitrariness in the choice of pseudo-count. One issue of particular interest is that the mean of the log-counts is not generally the same as the log-mean count [1]. This is problematic in scRNA-seq contexts where the log-transformation is applied to normalized expression data. Here, normalization is performed to remove inter-sample biases on the count scale [2,3]. For genes that are not differentially expressed (DE), the expectation of the normalized expression is the same between cells. However, the expectation of the log-transformed normalized values may not be the same, resulting in spurious differences between cells on the log-scale.

The ideal solution would be to use methods that model the counts directly to account for the mean-variance relationship. For example, we could use *BASiCS* [4] for detecting highly variable genes or *zinbwave* [5] for factor analysis. This would avoid the need for any transformation and subsequent distortion of the expected expression. However, most existing methods for scRNA-seq data analysis (e.g., clustering, dimensionality reduction) do not use count-based models. This is due to the difficulties in estimating model parameters, often separately for each gene and/or each cell, compared to algorithms that only need distances between cells as input. Thus, it is still desirable to obtain transformed data for compatibility with existing methods.

Here, we describe the nature and impact of the discrepancy between log-mean and mean-log values in real and simulated data. We suggest some practical strategies to protect against this effect, including the use of DE analyses with count-based models as validation. We also derive an expression for a larger pseudo-count that can be added prior to transformation to control the magnitude of the discrepancy for each gene.

2 Quantifying the discrepancy

Let X_{ig} denote a random variable for a non-negative count of a gene g in cell i . The size factor s_i represents the scaling bias in i , which is removed by computing the normalized expression value X_{ig}/s_i [6]. The log-transformed normalized expression is defined as

$$Z_{ig} = \log \left(\frac{X_{ig}}{s_i} + c \right),$$

where c is a positive “pseudo-count” that ensures that the transformation is defined for $X_{ig} \geq 0$ and $s_i > 0$. This calculation and its variants are widespread in analyses of count data from high-throughput sequencing experiments – for example, a similar approach is used by the `cpm` function in the *edgeR* package [7], differing only in a slight adjustment to the effective library size (equivalent to s_i) for a given prior count (equivalent to c). The second-order Taylor series approximation for the expectation of Z_{ig} is

$$E(Z_{ig}) \approx \log[E(X_{ig}/s_i) + c] - \frac{\text{var}(X_{ig})s_i^{-2}}{2[E(X_{ig}/s_i) + c]^2}.$$

This expectation represents the mean of the log-normalized expression, while the first term on the right represents the log-mean normalized expression (after adding c). The second term on the right represents the discrepancy between these two values.

Now, consider two cells $i = 1$ and $i = 2$ that differ in their s_i . Assume that g is non-DE between these two cells such that $E(X_{ig}/s_i) = \mu_g$ for both. The true log-fold change in the normalized expression values between these two cells is

$$\delta_{12g} = \log \left[\frac{E(X_{1g}/s_1)}{E(X_{2g}/s_2)} \right] = 0.$$

In analyses based on the log-transformed values, we use the expected difference in Z_{ig} as a proxy for δ_{12g} . First, we simplify the expression for $E(Z_{ig})$ to

$$E(Z_{ig}) \approx \log(\mu_g + c) - \frac{\text{var}(X_{ig})s_i^{-2}}{2(\mu_g + c)^2}.$$

This means that the difference in $E(Z_{ig})$ between these two cells is

$$\Delta_{12g} \approx \frac{\text{var}(X_{2g})s_2^{-2} - \text{var}(X_{1g})s_1^{-2}}{2(\mu_g + c)^2}. \quad (1)$$

In general, $\Delta_{12g} \neq 0$ due to differences in the size factors and subsequently variances (driven by the strong mean-variance relationship in count data). This represents a spurious difference on the log-scale as the gene is not actually DE, i.e., $\delta_{12g} = 0$.

We performed simulations to quantify this spurious difference in a range of scenarios for the typical $c = 1$ (Figure 1). We observed that it was most prominent when the size factors were different and either or both of them were less than unity. This is consistent with Equation 1, where Δ_{12g} increases in magnitude with smaller s_i . We note that a 10-fold difference in the size factors across cells is not unusual in real scRNA-seq data [3] due to the variability of capture efficiency and amplification. In this situation, we can easily obtain a difference of 1 in the mean $\log_2$ -values between cells. This corresponds to an artificial 2-fold change in expression that does not correspond to any biological effect.

We verified the existence of this effect in real scRNA-seq data from the 10X Genomics platform [8]. We used the publicly available ERCC data set (see Methods) in which gel emulsion beads containing ERCC spike-in transcripts were captured into droplets. The composition of the ERCC spike-in set should be constant in all droplets, so there should not be any difference in the expression profiles after library size normalization. However, we observed a systematic non-zero difference in the mean log-expression values when comparing the smallest and largest droplets (Figure 2a). This is fully attributable to the log-transformation, as no such difference is present in the log-fold change between the mean normalized expression values (Figure 2b). Discrepancies between the difference in mean log-values and the log-fold change are also present in real scRNA-seq data sets with variable size factors (Figures 2c, d).

The effect of spurious differences in $E(Z_{ig})$ is amplified in procedures that compute distances between cells. Consider the Euclidean distance between the expected log-normalized expression profiles of our two cells $i = 1$ and 2. Ideally, the distance would be zero as there should be no difference in the expected location of these cells after normalization. However, the actual distance will be approximately equal to the square root of the sum of Δ_{12g}^2 across all (non-DE) genes used in the calculation. This

Figure 1. Maximum difference in the mean $\log_2$ -expression values for a non-DE gene due to size factor differences. We simulated counts for non-DE genes in two groups of cells, where the size factor was the same for all cells in the same group (see Methods). Each simulation scenario was defined by the size factor used for the first group (a_1), that for the second group (a_2), and the negative binomial dispersion. For each scenario, we computed the difference between the mean Z_{ig} for each group and determined the maximum difference across genes of varying abundance. Values represent the average of the maximum across 10 simulation iterations. Standard errors were negligible and are not shown. For simplicity, we only show results for scenarios where $a_1 \geq a_2$.

Figure 2. Effect of the log-transformation on comparisons between groups in real scRNA-seq data. (a) Differences in the mean Z_{ig} between the top 20% of libraries with the largest size factors and the bottom 20% in the ERCC 10X Genomics data set (see Methods). Each point represents a single ERCC spike-in transcript. (b) Log-fold change in the mean normalized expression between the same groups of libraries for each ERCC transcript. (c) Differences in the mean Z_{ig} for two clusters of cells with the largest and smallest median size factors in a brain scRNA-seq data set [9], compared to the log-fold change computed from the mean normalized expression. Each point represents a gene. (d) Same as (c) for the peripheral blood mononuclear cell (PBMC) data set [8].

Figure 3. Artificial structures induced by log-transformation in simulations with no true structure. (a) Principal components analysis (PCA) plot of log-normalized expression values for non-DE genes in simulated cells with either very small or very large size factors (see Methods). Each point represents a cell coloured according to its value of s_i . The percentage of variance explained by each principal component (PC) is also shown in parentheses. (b) PCA plot for simulated cells with s_i sampled from a continuous distribution. Each point represents a cell coloured by its value of s_i .

means that the distance between cells with different s_i will be systematically larger than between cells with the same s_i . As a result, spurious clusters or trajectories can form (Figure 3) solely due to the log-transformation. Such artificial structures can be highly misleading when characterizing the heterogeneity of a cellular population.

The discrepancy described in Equation 1 is also consistent with the observations from Hicks *et al.* [1]. In their study, Hicks *et al.* show that the log-normalized expression of many genes is anticorrelated with the “detection rate” (i.e., the percentage of non-zero counts in each cell) in a variety of scRNA-seq data sets. We note that the detection rate is often strongly associated with the size factor due to the sparsity of the data. Any increases to the coverage of a cell (and thus the size factor s_i) will naturally increase the detection rate as fewer zeroes are sampled, such that both are correlated with the error from the log-transformation. However, some caution is required when interpreting these correlations in real data due to the conflation of technical effects with genuine biological factors such as, e.g., differences in RNA content across cell types [10,11].

We wondered whether this problem of artificial differences could be mitigated by using alternative transformations. We considered the square root, which provides variance stabilization for Poisson-distributed counts; and the variance stabilizing

transformation (VST) for negative binomial-distributed counts from *DESeq2* [12]. In simulations involving two groups of cells differing only in their size factors, we observed large differences in the group-specific means for non-DE genes after applying each transformation (Figure 4). This mirrors the mean-dependent nature of Δ_{12g} and indicates that variance stabilization does not inherently protect against spurious differences. In fact, no transformation can truly stabilize the variance at low counts, as the variance is zero at a mean of zero and must always increase. This is problematic as many scRNA-seq data sets are dominated by low counts – 81% and 89% of non-zero counts are below 5 in the brain and PBMC data sets, respectively – which suggests that favouring a transformation on the basis of its variance stabilization is largely misguided.

3 Validating structure with DE analyses

In our opinion, the effect of greatest concern is the formation of spurious clusters and trajectories. This can easily result in incorrect biological conclusions from an exploratory analysis of a scRNA-seq data set. Fortunately, careful further analysis can provide some protection against these artificial structures. Say that we identify spurious clusters like those in Figure 3. We attempt to characterize these clusters by identifying genes that are differentially expressed between them. Here, the key is to use count-based models like *edgeR* [7] to perform the DE analysis. This directly accounts for complex mean-variance relationships in the count data, thereby avoiding the need for any transformation and concomitant artefacts [13]. If the analysis fails to yield any DE genes, we can infer that the separation was driven by the log-transformation and should be ignored. This strategy is motivated by the presence of differences in the observed effect sizes computed from transformed or untransformed data (Figure 2), possibly leading to different biological conclusions. Our results indicate that untransformed data is more appropriate for these comparisons between groups.

While count-based models provide some protection from spurious structure, it is often the case that we still need to use per-gene transformed values for other steps in the analysis. This includes visualization of expression between groups (e.g., in boxplots, or by colouring dimensionality reduction plots), for which the log-transformation is often used to preserve relative differences across a wide range of expression values. It

Figure 4. Difference in the group-specific mean transformed expression values against the mean count for each gene, for Poisson-distributed counts (a, b, c) or NB-distributed counts (d, e, f) after applying a variety of transformations (see Methods). Simulated count data was generated for non-DE genes in each of two groups of cells differing only in their size factors (smaller or larger). Colour intensity is proportional to the density of genes, with outliers shown as points. The red line represents a fitted loess curve.

may also be more convenient to perform DE analyses on the log-transformed values, e.g., using t -tests [14, 15], which are faster and simpler to implement than count-based models. Thus, we would like to reduce Δ_{12g} as much as possible for these applications. The most obvious method of doing so is to increase the pseudo-count. Indeed, we could force Δ_{12g} to zero for all genes by setting c to some arbitrarily large value. However, as $c \rightarrow \infty$, the log-transformation approaches a linear transformation, i.e.,

$$Z_{ig} \approx \log(c) + \frac{X_{ig}}{s_i c}.$$

Thus, using an arbitrarily large pseudo-count defeats the intended purpose of the log-transformation. No variance stabilization is achieved and differences in the transformed values cannot be used as proxies for log-fold changes. Instead, we aim to choose a pseudo-count that restricts Δ_{12g} to an “acceptable” level for each gene.

4 Choosing a larger pseudo-count

We assume that X_{ig} follows a negative binomial (NB) distribution with mean $s_i \mu_g$ and dispersion φ . This means that we can simplify the expression for Δ_{12g} to

$$\begin{aligned} \Delta_{12g} &\approx \frac{\mu_g s_2^{-1} + \varphi \mu_g^2}{2(\mu_g + c)^2} - \frac{\mu_g s_1^{-1} + \varphi \mu_g^2}{2(\mu_g + c)^2} \\ &= \frac{\mu_g (s_2^{-1} - s_1^{-1})}{2(\mu_g + c)^2}. \end{aligned} \quad (2)$$

This reaches its maximum absolute value for positive μ_g when $\mu_g = c > 0$, yielding

$$\tilde{\Delta}_{12g} = \left| \frac{(s_2^{-1} - s_1^{-1})}{8c} \right|.$$

We can control the maximum error $\tilde{\Delta}_{12g}$ below a threshold τ by setting

$$c = \rho |s_2^{-1} - s_1^{-1}|, \quad (3)$$

where $\tau = (\rho 8)^{-1}$. If we set $\rho = 1$, the maximum difference in the mean log-values should be 0.125 (or $[8 \log(2)]^{-1} \approx 0.18$, on the $\log_2$ -scale). The observed maximum difference in simulations lies close to this bound at low dispersions (Figure 5). At higher

dispersions, the bound becomes less accurate but is nonetheless still conservative.

Equation 3 is most obviously applied to specific comparisons between two groups of cells. In such cases, we can set s_1 and s_2 to the average size factor of each group to control $\tilde{\Delta}_{12g}$ (Figure 6a). However, this may not be generally possible if more than two groups are involved or when groups are not well-defined, e.g., trajectories. The exact choices of s_1 and s_2 are not obvious when calculating a single pseudo-count for the entire data set. We suggest setting s_1 to the smallest size factor and s_2 to the largest size factor (or robust equivalents thereof, e.g., the 5th and 95th percentiles). This ensures that the difference in log-values for a non-DE gene does not exceed the theoretical upper bound for most pairs of cells. We thus define our empirical pseudo-count as

$$\hat{c} = \max\{1, \rho |s_{min}^{-1} - s_{max}^{-1}|\},$$

which ensures that the pseudo-count is at least unity when all size factors are equal. We suggest setting $\rho = 1$, which provides a compromise between restricting $\tilde{\Delta}_{12g}$ and keeping the pseudo-count as small as possible. Applying this strategy to the ERCC data set limits the maximum error close to the theoretical upper bound (Figure 6b).

We computed $\hat{c}$ for a number of real scRNA-seq data sets to examine its behaviour in practical settings. With $\rho = 1$, we obtained $\hat{c}$ values of 1.36 for the 416B data set [11], 4.05 for the mouse brain data set [9] and 2.22 for the PBMC data set [8]. These values are within the typical range of pseudo-counts used for exploratory analyses of RNA-seq data [16]. They also lie within the range of mean abundances in each data set – we observed 217, 649 and 12636 genes with mean normalized expression values greater than $\hat{c}$ in the PBMC, brain and 416B data sets, respectively. This indicates that the computed $\hat{c}$ is not so large as to entirely negate the variance stabilization from log-transformation. It also justifies our strategy of controlling the maximum spurious difference $\tilde{\Delta}_{12g}$, which occurs when the mean normalized expression is close to $\hat{c}$. (Otherwise, if all genes had lower abundances than $\hat{c}$, the maximum difference would never be achieved and our procedure would be unnecessarily conservative.)

Figure 5. Maximum difference in the $\log_2$ -expression values for non-DE genes between groups of cells differing only in their size factors, after adding the pseudo-count from Equation 3 with $\rho = 1$ prior to $\log$ -transformation. Simulations and calculations were performed as described for Figure 1. Results are only shown for scenarios where $s_1 > s_2$. The red dashed line represents the theoretical bound on the maximum difference.

Figure 6. Effect of log-transformation with an increased pseudo-count in real scRNA-seq data. (a) Differences in the mean Z_{ig} compared to the log-fold change in the PBMC data set, for the same pair of clusters shown in Figure 2d. The pseudo-count was computed using Equation 3 with $\rho = 1$ and using the median size factor for each cluster as s_1 and s_2 . The unbroken line represents identity while the dashed line denotes the theoretical limit on $\tilde{\Delta}_{12g}$. (b) Differences in the mean Z_{ig} between the top and bottom 20% of libraries with the largest and smallest size factors in the ERCC data set, after adding $\hat{c}$ with $\rho = 1$. The red line denotes the theoretical limit on $\tilde{\Delta}_{12g}$.

5 Practical use of larger pseudo-counts

We recommend using log-expression values computed with $\hat{c}$ during characterisation of population structures identified from exploratory analyses. This considers the worst-case scenario where there are systematic differences in the size factors between cells in different clusters or along trajectories. (Conversely, it is not required when the distribution of size factors is the same across clusters or trajectories.) By controlling $\tilde{\Delta}_{12g} \leq \tau$, we can mitigate the effects of misleading differences during visualization of expression, as previously discussed. Methods like TREAT [17] can also be applied to test whether the observed log-difference is significantly greater than the theoretical limit τ . This provides a statistically rigorous alternative for DE analyses if count-based models are impractical. It is also possible to use log-values computed with $\hat{c}$ in the initial exploratory analysis, which may reduce the risk of detecting spurious structures in the first place. However, this is more difficult to justify as there is no straightforward interpretation for an “acceptable” threshold in the error in the distance between cells.

Pseudo-counts are often set to unity for practical reasons that have little to do with avoiding spurious differences. One reason is to ensure that zero counts always yield zero

log-values, which ensures that sparse matrix representations are still effective for reducing memory usage when analyzing large data sets. A related motivation is to ensure that the log-transformed values are always non-negative, which is necessary for procedures such as non-negative matrix factorization. However, there is no need to conflate these considerations with the choice of pseudo-count. Some arithmetic yields

$$\log \left(\frac{X_{ig}}{s_i} + c \right) = \log \left(\frac{X_{ig}}{cs_i} + 1 \right) + \log c ,$$

where the constant $\log(c)$ on the right hand side can be ignored in any step that involves computing differences between log-values. By computing the first term on the right, we can preserve sparsity and non-negativeness in the log-expression matrix.

As an aside, it is also worth commenting on the widespread use of log-transformed counts-per-million (CPM) values and their relatives, e.g., transcripts-per-million or fragments-per-kilobase-million. Log-CPM values are typically computed by adding 1 to the CPM values, for the same reasons mentioned above. However, this is equivalent to using a pseudo-count equal to the library size (i.e., total count across all genes) in millions. For example, data sets with an average library size of 10 million would use a pseudo-count of 10, while data sets with an average library size of 10,000 would use a pseudo-count of 0.01. This is not desirable as data sets with small library sizes are those with the largest Δ_{12g} (as μ_g is small) and in most need of large pseudo-counts. In contrast, the approximation of the log-fold change by differences in log-values should improve with greater coverage. Increasing the pseudo-count counteracts this effect such that the log-fold change estimates are not consistent with respect to coverage.

6 Discussion

We can speculate on the practical settings in which the log-transformation is most likely to cause distortions. Artifacts are most pronounced when the counts are low and there is large variation in the size factors across cells. Both of these are often observed in data from droplet-based scRNA-seq protocols, as sequencing coverage is shared across a large number of libraries and reaction conditions are not precisely controlled for each droplet. We would also expect the size factors within each experiment to vary across a

continuum, yielding artificial trajectories (Figure 3b) rather than distinct clusters (Figure 3a). This suggests that caution is required when interpreting trajectories that are correlated with the size factors, especially in low-coverage droplet-based data.

In the context of scRNA-seq analysis workflows, additional procedures can be used to reduce transformation-induced artifacts. For example, quality control is typically performed to remove cells with small library sizes [18, 19]. This reduces the variation in the size factors and the potential for artificial differences upon log-transformation. In addition, low-abundance genes can be filtered out, which provides further protection as genes with low counts are most affected by the log-transformation. We note that the use of a larger pseudo-count will also reduce the influence of low-abundance genes. This is a more nuanced approach than filtering, as it ensures that strong biological signal in low-abundance genes is not completely discarded for downstream applications.

Data transformations are powerful tools for exploratory analyses of scRNA-seq data. However, they also have the potential to introduce artificial differences, as we have shown for the log-transformation. Our recommendation is to increase the pseudo-count in cases where size factor variation may be responsible for spurious structure in the data.

7 Methods

7.1 Simulated log-fold changes for non-DE genes

We considered two groups of 10,000 cells where all cells in the same group j were assigned the same size factor $s_i = a_j$. For a non-DE gene g , the count X_{ig} for each cell i was independently sampled from a negative binomial distribution with mean $s_i\mu_g$ and dispersion φ . We computed Z_{ig} for all cells in each group using a pseudo-count c of 1 as previously described. The difference in the mean Z_{ig} between groups represents an estimate of Δ_{12g} . The maximum difference in the means was then determined by testing a variety of gene abundances $\mu_g \in [10^{-3}, 10^3]$. We repeated this procedure across 10 simulation iterations and reported the average of the maximum differences across iterations. This was performed for each simulation scenario defined by the combination of the size factor for two groups (a_1 and a_2 , where $a_1 \geq a_2$) and φ .

7.2 Log-fold changes in real data sets

We obtained the ERCC data set from the 10X Genomics website (<https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/ercc>). We retained all libraries with total counts greater than 100. Size factors were defined as the library sizes, scaled to a mean of unity across libraries. We defined two groups in this data set, with the first group containing the top 20% of libraries with the largest size factors and the second group containing the bottom 20%. For each ERCC transcript, we computed the log-normalized expression values using a pseudo-count of 1. We then computed the differences in the mean log-values between groups to obtain an estimate of Δ_{12g} . For comparison, we also computed the log-fold change (i.e., δ_{12g}) between groups based on the mean normalized expression. Note that log-fold changes were computed after adding a pseudo-count of 1 to the means of both groups.

For the 416B, brain and PBMC data sets, counts were processed according to a published workflow [19]. Briefly, low-quality cells were removed with the *scater* package [20] and size factors were computed using the deconvolution method [3]. PCA was performed on the log-normalized expression values (computed with a pseudo-count of 1) after modelling the mean-variance trend, and cells were clustered based on their PC scores using the shared nearest neighbours method [21]. In each data set, we identified the two clusters with the largest and smallest median size factors. We computed the difference in the mean log-expression values between clusters and compared this to the log-fold change between clusters calculated from the mean normalized expression. Note that log-fold changes were calculated after adding a pseudo-count of 1 to the means of both groups, such that any discrepancy between the difference in mean log-values and the log-fold change is an estimate of Δ_{12g} .

7.3 Simulations with artificial structures

Counts for 1000 genes in 500 cells were sampled independently from a Poisson distribution with a mean of λ_i for each cell i . We set $\lambda_i = 0.1$ for half of the cells and $\lambda_i = 10$ for the other half. The size factor for each cell was defined as λ_i after being scaled to a mean of unity across cells. We computed log-normalized expression values for all cells with a pseudo-count of 1, as previously described. We then performed PCA

and examined the distribution of cells along the first two PCs. We also repeated this simulation where λ_i was sampled from a $\text{Uniform}(0.1, 5)$ distribution instead. Note that all genes are non-DE in these simulations, i.e., $E(X_{ig}/s_i)$ is constant for all g and i .

7.4 Simulations with different transformations

Counts for 10,000 genes in 500 cells were sampled from a Poisson or negative binomial distribution with the mean of $\lambda_i \mu_g$ for each cell i . We sampled $\log_2(\mu_g)$ from a $\text{Uniform}(-5, 10)$ distribution to obtain a range of gene abundances. We partitioned the cells into two groups of equal size, setting $\lambda_i = 0.1$ for all cells in one group and $\lambda_i = 10$ for all cells in the other group. For the negative binomial distribution, we used a constant dispersion of 1 for simplicity. The size factor for cell i was defined as λ_i after being scaled to a mean of unity across cells. We applied the tested transformations – square root, log, and VST – to the normalized expression values, using a pseudo-count of 1 for the log₂-transformation and `fitType="mean"` for the *DESeq2* VST. We then computed the mean of the transformed values in each group for each gene.

7.5 Implementation details

All analyses were performed using R 3.5.0 with packages from the open-source Bioconductor project [22]. We used *scater* 1.9.14, *DropletUtils* v1.1.8, *BiocFileCache* v1.5.5 and *DESeq2* v1.21.16. All code is publicly available on GitHub at <https://github.com/LTLA/PseudoCount2018>.

8 Acknowledgements

We would like to thank Dr. John Marioni for helpful suggestions on the manuscript. This work was supported by core funding from Cancer Research UK (award no. A17197 to Dr. Marioni).

References

1. S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, Nov

- 2017.
2. M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, 2010.
3. A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, Apr 2016.
4. C. A. Vallejos, S. Richardson, and J. C. Marioni. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.*, 17:70, Apr 2016.
5. D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J. P. Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun*, 9(1):284, 01 2018.
6. S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.
7. M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.
8. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, 8:14049, Jan 2017.
9. A. Zeisel, A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Brain structure. *Cell*

- types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, Mar 2015.
10. S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167, Jul 2011.
 11. A. T. L. Lun, F. J. Calero-Nieto, L. Haim-Vilmovsky, B. Gottgens, and J. C. Marioni. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.*, 27(11):1795–1806, Nov 2017.
 12. M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.
 13. R. B. O’Hara and D. J. Kotze. Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122, 2010.
 14. C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, 15(2):R29, Feb 2014.
 15. C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, 15(4):255–261, Apr 2018.
 16. Y. Chen, A. T. Lun, and G. K. Smyth. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res*, 5:1438, 2016.
 17. D. J. McCarthy and G. K. Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, Mar 2009.
 18. T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.*, 17:29, Feb 2016.
 19. A. T. Lun, D. J. McCarthy, and J. C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*, 5:2122, 2016.

20. D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 04 2017.
21. C. Xu and Z. Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, Jun 2015.
22. W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pages, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, 12(2):115–121, Feb 2015.