Feature selection followed by a residuals-based normalization simplifies and improves single-cell gene expression analysis

Amartya Singh[1*], Hossein Khiabanian[1,2†]

1 Center for Systems and Computational Biology, Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, New Jersey
2 Department of Pathology and Laboratory Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers University, New Brunswick, New Jersey
†Current address: Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, New York

***as2197@scarletmail.rutgers.edu/amartya.singh@rutgers.edu**

Abstract

Normalization is a critical step in the computational analysis of single-cell RNA-sequencing (scRNA-seq) counts data. The objective is to reduce systematic biases introduced by technical sources that can obscure underlying biological differences. This is typically accomplished by re-scaling the observed counts to reduce the differences in total counts between the cells and then transforming the scaled counts to stabilize the variances. In the standard scRNA-seq workflow, this is followed by feature selection to identify genes that capture most of the biologically meaningful variation across the cells. Here, we propose a simple feature selection method and show that we can perform feature selection before normalization. We also propose a novel residuals-based normalization method that includes a monotonic non-linear transformation to ensure effective variance stabilization of the residuals. We demonstrate significant improvements in downstream clustering analyses through the application of our feature selection and normalization methods to truth-known biological as well as simulated counts data sets. Based on these results, we make the case for a revised scRNA-seq analysis workflow wherein feature selection precedes and in fact informs our residuals-based normalization. This novel workflow has been implemented in an R package called `Piccolo`.

Introduction

Bulk RNA sequencing (RNA-seq) studies have led to a significant improvement in our understanding of gene expression profiles associated with healthy as well as diseased states of various tissue types. However, these studies only provide an averaged view at the tissue level in which subtle but crucial distinctions of the constituent cell-types and states are obscured. Rapid advances in single-cell RNA-seq (scRNA-seq) protocols and platforms over the past decade have now facilitated investigation of transcriptional profiles at the level of individual cells, thereby enabling identification of distinct cell-types and cell states [1–5], as well as stages of development and differentiation [6, 7].

In contrast to measurements on bulk tissues, single-cell measurements have significantly greater uncertainty due to the low amounts of starting material as well as low capture efficiencies of the protocols (typically, high-throughput protocols only capture between 5% to 20% of the molecules present in each cell [8]). As a result, even deeply sequenced data sets may have up to 50% zeros [9]. The high sparsity poses a significant challenge during the computational analysis. Early attempts to build statistical models to explain the relationship between the observed counts and the true underlying gene expression levels relied on zero-inflation models to explain the excess zeros. However,

data generated using newer scRNA-seq protocols that rely on unique molecular identifiers (UMIs) have been shown to be sufficiently described using simpler statistical models that do not include zero inflation [10, 11]. In this paper, any reference to scRNA-seq data will specifically pertain to UMI counts data.

A critical step in the computational analyses of both bulk RNA-seq and scRNA-seq data sets is that of normalization. The objective of normalization is to reduce the biases introduced by technical sources or even biological sources such as cell cycle state, so that we can confidently identify true biological differences [9, 12–14]. Owing to the small amount of mRNAs captured per cell, the effect of these biases is more pronounced in the case of scRNA-seq data, further underscoring the need to reduce the impact of these biases on downstream analyses. Typically, normalization is performed by re-scaling the observed counts using cell-specific size factors to reduce the differences in sampling depths (total counts) between the cells. The scaled counts are then transformed with the help of a monotonic non-linear function (usually the logarithm function) to stabilize the variances of genes across different mean expression levels.

In the standard scRNA-seq workflow (implemented for instance in Seurat [15–18] and Scanpy [19]) normalization is followed by a feature selection step that focuses on identifying genes that capture most of the biological variation across the cells while eliminating genes that do not exhibit meaningful biological variation. This sequence of steps - normalization followed by feature selection - in the standard workflow appears quite reasonable, especially given the fact that differentially expressed genes can be identified reliably only after reducing the sampling depth differences between the cells. However, this objective: (i) identification of genes that are differentially expressed between groups of cells, is not the same as objective: (ii) identification of highly variable genes (HVGs). While the identification of differentially expressed genes between distinct groups of cells requires that the sampling depth differences be reduced through normalization, we show that it isn't necessary to perform normalization in order to identify HVGs.

In this article, we propose a simple feature selection method that relies on a regression-based approach to estimate dispersion coefficients for the genes based on their observed counts. Using this method, we show that feature selection can be performed before normalization. Importantly, during the feature selection step we not only identify variable genes, but also shortlist stable genes. The variation in the counts of these stable genes is expected to primarily reflect the biases introduced by the technical sources, and can therefore be used to estimate cell-specific size factors in order to perform normalization. During normalization we also need to ensure variance stabilization, especially when relying on principal components analysis (PCA) for dimensionality reduction. Keeping this in mind, we propose a residuals-based normalization method that not only reduces the impact of sampling depth differences between the cells but simultaneously ensures variance stabilization by explicitly relying on a monotonic non-linear transformation (default choice is the *log* transformation).

We demonstrate significant improvements in downstream clustering analyses enabled by the application of our feature selection and normalization methods on biological truth-known as well as simulated counts data sets. Based on these results, we make the case for a revised scRNA-seq analysis workflow in which we first perform feature selection and subsequently perform normalization using our residuals-based approach. We have implemented this novel scRNA-seq workflow in an R package called `Piccolo`.

Results

Genes with small counts exhibit quasi-Poisson variance

Genes with low mean expression levels show Poisson-like variance of their counts

As pointed out by Sarkar and Stephens [11], a good starting point for understanding the nature of the scRNA-seq counts data is to recall that the observed counts reflect contributions from both the underlying expression levels of the genes as well as the measurement errors. This necessitates that the contributions from the two be carefully distinguished in order to better understand and explain

the true biological variation. Based on their analysis, they found that a simple Poisson distribution sufficed to explain measurement error, while a simple Gamma distribution often sufficed to explain the variation in expression levels across cells. The observation model built using these two distributions yields the Gamma-Poisson (or negative binomial (NB)) distribution which is well-known as a plausible model to explain over-dispersed counts. Under this model, the mean-variance relationship is given by,

$$\sigma_{NB}^2 = \mu + \alpha_{NB}\mu^2 \tag{1}$$

where α_{NB} is the NB over-dispersion coefficient ($\alpha_{NB} = 0$ yields the familiar Poisson mean-variance relationship: $\sigma^2 = \mu$).

A more familiar expression for the NB mean-variance relationship is,

$$\sigma_{NB}^2 = \mu + \frac{\mu^2}{\theta} \tag{2}$$

where θ is referred to as the inverse over-dispersion coefficient.

Another mean-variance relationship closely related to the Poisson and the NB is the quasi-Poisson (QP) wherein,

$$\sigma_{QP}^2 = \alpha_{QP}\mu \tag{3}$$

where α_{QP} is the QP dispersion coefficient ($\alpha_{QP} = 1$ yields the familiar Poisson mean-variance relationship; $\alpha_{QP} > 1$ would be associated with counts over-dispersed with respect to the Poisson, while $\alpha_{QP} < 1$ would indicate counts under-dispersed with respect to the Poisson).

We began our investigation into the nature of UMI counts by considering the mean-variance relationship of counts for genes in a technical negative control data set [10](hereafter referred to as `Svensson 1`). This data set consisted of droplets containing homogenous solutions of endogenous RNA as well as spike-in transcripts. The variability of counts in this data arises solely due to technical sources and is not attributable to any underlying biological source. The left panel in Fig. 1A shows the variance (σ^2) vs mean (μ) log-log plot for `Svensson 1`. Each dot in the plot corresponds to a gene. The colors of the dots reflect the local point density, with darker shades (deep blues) indicating low density and brighter shades (bright yellow) indicating high density. The black line depicts the variance expected under the Poisson model ($\sigma^2 = \mu$).

We begin by noting that the UMI counts are heteroskedastic in nature since the variances of the genes depend on the mean (larger variances corresponding to larger means). If we then move on to focus on genes with low mean expression levels (especially $\mu < 0.1$), we can see that the mean-variance relationship appears to be well-approximated by the Poisson model since their observed variances lie close to the black line. It's only for genes with higher mean expression levels (especially $\mu > 1$) that the variance of the observed counts exceed the variance expected under the Poisson model. For comparison with counts data with inherent biological variation, we looked at the NIH/3T3 fibroblast cell line data set [10] (hereafter referred to as `NIH/3T3`) and the 10X Genomics Peripheral Blood Mononuclear Cells (PBMC) 33k data set [20] (hereafter referred to as `PBMC 33k`). Even for these two data sets, most of the genes with low mean expression levels exhibit Poisson-like mean-variance relationship (middle and right panels in Fig. 1A).

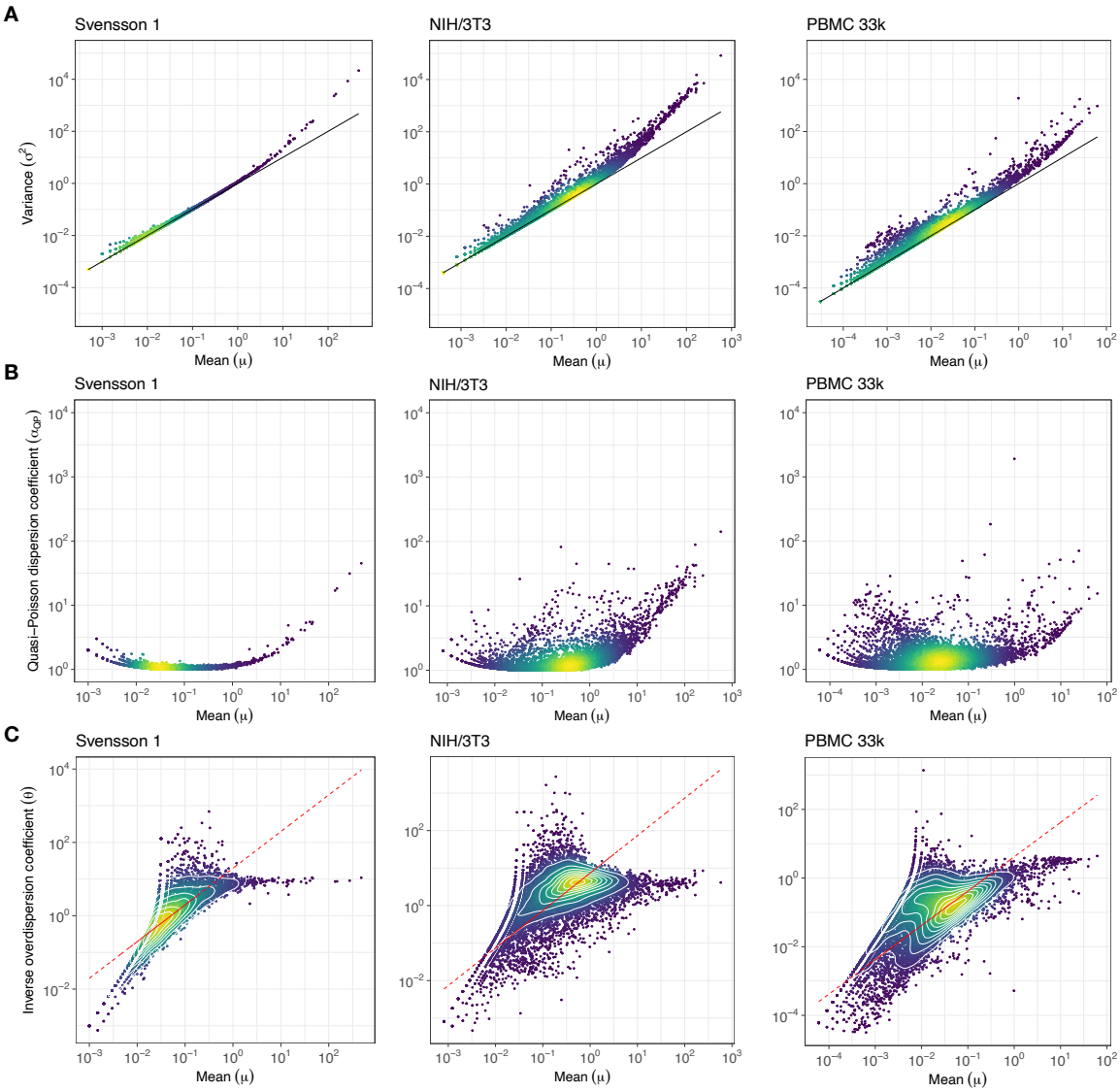

Figure 1. Genes with low mean expression exhibit quasi-Poisson variance. In all the plots, each dot represents a gene and the color of the dots reflect the local point density, with brighter shades (yellow) indicating high density and darker shades (deep blue) indicating low density. **A.** Variance (σ^2) vs mean (μ) log-log scatter plots for the Svensson 1 technical control (left panel), NIH/3T3 fibroblast cell line (center panel), and PBMC 33k (right panel) data sets. The solid black line corresponds to the Poisson model ($\sigma^2 = \mu$). For genes with low mean expression levels, the variance can be adequately described by the Poisson model. **B.** Quasi-Poisson dispersion coefficients (α_{QP}) vs mean (μ) log-log scatter plots for the Svensson 1 (left panel), NIH/3T3 (center panel), and PBMC 33k (right panel) data sets. α_{QP} for each gene were estimated from the observed counts using a regression-based approach. **C.** Inverse overdispersion coefficients (θ) vs mean (μ) log-log scatter plots for the Svensson 1 (left panel), NIH/3T3 (center panel), and PBMC 33k (right panel) data sets. For each gene, θ was obtained based on its α_{QP} using equation. (5). The white curves correspond to the contours lines based on the local point densities. The dashed red line corresponds to estimates of θ obtained by fixing α_{QP} to its value for the gene with the highest point density estimate in the the θ vs μ log-log plot. The monotonically increasing relationship between θ and μ for small μ indicates quasi-Poisson variance at low mean expression levels.

The Poisson-like nature of the mean-variance relationship at low expression levels can be made clearer with the help of a simple example. Consider a gene with $\mu = 0.01$. Let us assume that the counts of this gene are actually NB distributed and that $\alpha_{NB} = \frac{1}{\theta} = 1$ is the true estimate of the over-dispersion coefficient. The variance of the counts will then be $\sigma^2_{NB} = 0.01 + 1 * (0.01)^2 = 0.0101$. In comparison, the expected variance under the Poisson model will be $\sigma^2_{Pois} = 0.01$. The percent difference between the two variances is a mere 1%. Thus, even if we approximated the variance for the counts of this gene with that predicted by the simple Poisson model, the estimated variance would differ by only 1% from the true NB variance. The percentage difference between the Poisson and the NB variances would of course increase for larger values of α_{NB}, however, estimates of α_{NB} for real biological data sets typically range between 0.01 and 1 [21, 22]. Thus, we can conclude that genes with low mean expression levels (especially those that exhibit moderate over-dispersion) exhibit variances that do not deviate significantly from the variances predicted under the Poisson model.

QP dispersion coefficients can be obtained using a regression-based approach

For the NB distribution, the usual approach is to use maximum-likelihood estimation (MLE) to obtain estimates for the over-dispersion parameter (α_{NB}). This approach apart from being computationally intensive has the weakness that if in fact the distribution is not NB, the maximum-likelihood estimator is inconsistent. A more robust alternative is to use an approach that only requires specification of the mean and variance.

One such approach is used in the regression-based test proposed by Cameron and Trivedi [23, 24]. In their test they estimate over-dispersion beyond the null model (Poisson distribution) by specifying the alternate model in the form of a scalar multiple of a function of the mean,

$$Var[x] = \alpha E[x] \tag{4}$$

where the scalar multiple (α) is estimated using least squares regression (see Appendix).

The QP mean-variance relation (see equation. 3) corresponds precisely to such an alternative hypothesis in which the variance is simply a scalar multiple (α_{QP}) of the mean. This enables us to use this simple yet robust regression-based approach to estimate the QP dispersion coefficients (α_{QP}) by relying on just the means and variances.

Quite usefully, the QP variance can be viewed to arise from the NB variance in the special case where α_{NB} exhibits dependence on μ through the following relationship,

$$\alpha_{NB} = \frac{1}{\theta} = \frac{\alpha_{QP} - 1}{\mu} \tag{5}$$

Conversely, this means that we'll observe NB variance when α_{QP} exhibits dependence on μ through $\alpha_{QP} = 1 + \alpha_{NB}\mu$ (obtained from equation. 5).

QP variance for genes with low mean expression is not biological in origin, it is simply attributable to the observed counts being small

We obtained estimates for α_{QP} for all genes using the regression-based approach. Based on these estimates, we filtered out genes that were under-dispersed ($\alpha_{QP} < 1$). For the remaining genes, we plotted α_{QP} vs μ log-log plots (Fig. 1B). Before proceeding to discuss these plots, it will be helpful to lay down some basic expectations based on the relationship between α_{NB}, α_{QP}, and μ encapsulated in equation. 5. First, steady or constant values of α_{QP} across a range of mean expression levels would indicate QP variance of the counts for the corresponding genes. In turn, this will be reflected in the form of a monotonic dependence between θ and μ - larger θ for larger μ in the given range. Second, steady or constant values of θ (or α_{NB}) across a range of mean expression levels would indicate NB variance for the counts of the corresponding genes. In turn, this will be reflected in the form of a monotonic dependence between α_{QP} and μ - larger α_{QP} for larger μ in the given range.

For `Svensson 1` (left panel in Fig. 1B), it is immediately evident that for genes with low mean expression levels the α_{QP} do not exhibit any dependence on μ. As mentioned in our basic expectations

above, this indicates that the counts of these genes exhibit QP variance. Moreover, the values of α_{QP} for most of these genes with low mean expression are quite close to 1 which is consistent with the Poisson-like variance we observe in the mean-variance plot. Since `Svensson 1` is a technical control data set, this observation further validates the simple Poisson distribution as an appropriate model for explaining measurement error, particularly for genes with low mean expression levels. For `NIH/3T3` (middle panel in Fig. 1B) and `PBMC 33k` (right panel in Fig. 1B), despite greater variability in α_{QP} due to the inherent biological variability in the data, the QP nature of the variance of counts is still evident since we don't observe a dependency between α_{QP} and μ for genes with low mean expression levels.

Using equation. 5, we obtained estimates of θ for each gene using the corresponding values of α_{QP} and μ. We plotted the θ vs μ log-log plots for the genes using these values. For `Svensson 1` (left panel in Fig. 1C), we observe a clear monotonic increase in θ with increase in μ, particularly for genes with low mean expression levels. Once again, this clearly indicates that the counts of these genes exhibit QP variance. In order to make the monotonic increase even more apparent, we plotted estimates of θ (the dashed red line) obtained by fixing α_{QP} to the value of α_{QP} for the gene with the highest point density estimate in the the θ vs μ log-log plot. Furthermore, we plotted contour lines (white curves) based on the point densities as a visual aid to infer how the density of the points (genes) varies depending on μ. If we imagine a closed contour loop as an ellipse, the straight line through the center of the ellipse that joins the two points furthest from the center is called the major axis of the ellipse. We refer to analogous lines for the contours as their longer axes. Supposing no dependence of θ on μ, the longer axes of the contours would be parallel to the $x-$axis (zero slope). Instead, we observe that the dashed red line of the estimated θ with fixed value of α_{QP} lies along the same direction as the longer axes of the contours, particularly for genes located in the region with high point density (bright yellow region). For `PBMC 33k` (right panel in Fig. 1C), once again, despite greater variability in θ due to the inherent biological variability, the monotonic increase in θ with μ (reflective of QP variance) particularly for genes with low mean expression levels is still very evident. Particularly for genes located in the region with high point density (bright yellow region), the dashed red line of the estimated θ with fixed α_{QP} closely follows the direction of the longer axes of the contours. The `NIH/3T3` data provides an insightful contrast (middle panel in Fig. 1C) - while the monotonic increase in θ with increase in μ is evident for genes with low mean expression levels, there is a clear discrepancy between the slope of the dashed red line (estimated θ with fixed value of α_{QP}) and the slope of the longer axes of the contours, particularly for genes in the region with the highest point density (bright yellow region). A closer examination reveals that for this data set the mean expression level (μ) of genes in the region of highest point density lie between 0.1 and 1, which is an order of magnitude higher than what we observe for `Svensson 1` and `PBMC 33k` (μ lies between 0.01 and 0.1 for genes in their respective regions of highest density). This difference actually stems from the differences in their respective sequencing depths - while the median total UMIs per cell in `Svensson 1` and `PBMC 33k` are 2309 and 1891 respectively, it is 15560 in `NIH/3T3`. Thus, we expect that higher sequencing depths will lead to more genes exhibiting variances of counts that exceed the QP variance.

So far we have only focused on genes with low mean expression levels. If we now turn our attention to genes with higher mean expression levels, we notice that in `Svensson 1`, genes with $\mu > 1$ do not appear to exhibit a dependence between θ and μ (left panel in Fig. 1C). Given the expectations listed earlier, constant values of θ (or α_{NB}) indicate NB variance of the counts for the corresponding genes. We also expect this to manifest as a monotonic increase in α_{QP} with increase in μ, which is precisely what we observe in the left panel in Fig. 1B. Similar observations hold true for `NIH/3T3` and `PBMC 33k` (middle panels and right panels in Figs. 1B and 1C). For both, we do not observe any pronounced dependence between θ and μ for genes with high mean expression levels ($\mu > 1$), thereby suggesting that the counts of these genes exhibit variance better explained by the NB distribution.

To summarize, we observed that genes with low mean expression levels exhibit QP variance. We showed that the QP variance of genes with low mean expression levels manifests as a monotonically increasing relationship between θ and μ. It is important to highlight that we observed this relationship

not just in the biological data sets (`NIH/3T3` and `PBMC 33k`) but also in the technical negative control data set (`Svensson 1`), clearly suggesting that this relationship is not biological in origin. We must mention here that Lause et. al. [21] simulated a data set with NB distributed counts, relying on means estimated using the `PBMC 33k` data set while keeping the inverse over-dispersion parameter fixed ($\theta = 10$). They obtained MLE estimates for θ and observed a non-decreasing relationship with μ for genes with low mean expression levels (Fig. 1f in their paper). Further, they noted that only for genes with high mean expression levels was the true θ estimated correctly. Based on this observation, they concluded that the positive correlation between θ and μ does not represent an underlying technical or biological cause, but is attributable to estimation bias. As shown by us, however, the explanation is simpler: for genes with low mean expression levels, the variance of the counts can barely exceed the variance expected under the Poisson distribution. It is this Poisson-like nature of the variance that is reflected in the form of a monotonic relationship between θ and μ for genes with low mean expression levels.

Feature selection can be performed before normalization

Standard scRNA-seq workflow is based on an assumption that reflects a confusion between the distinct objectives of identification of variable genes and differentially expressed genes

In the standard scRNA-seq workflow, identification of HVGs - genes that exhibit greater variability in counts compared to other features in the data set - is performed only after normalization. This particular sequence in the workflow is based on the assumption that unless the systematic biases are reduced or eliminated, we cannot reliably identify features that best capture the biological variability inherent in the data. A more careful examination reveals that this assumption reflects a confusion between two fundamentally distinct objectives that on the surface appear similar. The first objective - identification of HVGs - as mentioned above, is about identifying genes that exhibit higher variability of counts compared to other features in the data set; we expect that these genes capture most of the biological variability in the data set. The second objective - identification of differentially expressed genes - is about identifying genes that exhibit differences in their expression levels between distinct sets of cells. For identifying differentially expressed genes, it is indeed imperative that the systematic biases are reduced or eliminated in order to identify genes that truly reflect actual biological differences between the distinct groups of cells being compared. However, it is possible to identify HVGs without first accounting for the systematic biases through normalization. This is so since the biases owing to their systematic nature are expected to manifest as additional but consistent sources of variation for the counts of genes across cells (in fact, this is the assumption underlying the normalization approaches that rely on estimation of cell-specific size factors to re-scale and adjust the observed counts).

We can state the expectation discussed above in terms of the QP dispersion coefficients. Suppose we have gene A and gene B with comparable mean expression levels ($\mu_A \approx \mu_B$) such that,

$$\tilde{\alpha}^A_{QP} > \tilde{\alpha}^B_{QP}$$

where $\tilde{\alpha}^A_{QP}$ and $\tilde{\alpha}^B_{QP}$ are the QP dispersion coefficients of gene A and gene B in the hypothetical case where there is no systematic bias. Our expectation is that even in the presence of systematic biases, the relative magnitudes of the dispersion coefficients for the bias-affected counts will be such that,

$$\alpha^A_{QP} > \alpha^B_{QP}$$

where α^A_{QP} and α^B_{QP} are the respective QP dispersion coefficients obtained from the observed counts of genes A and B. Thus, the expectation is that genes that exhibit high variability in their counts due to underlying biological differences will exhibit high variability even in the presence of systematic biases. We can test whether this is a reasonable expectation by introducing systematic biases into a given data set and then verifying whether the genes we identify as variable for the modified data set

are consistent with the ones identified for the original data set. Before proceeding to perform such a test, however, we first need to lay down a method to identify HVGs.

Feature selection method based on QP dispersion coefficients identifies HVGs and stable genes

As discussed earlier, distinguishing between measurement and biological processes by separately modeling their contributions to observed counts enables a better understanding and interpretation of the observed counts [11]. It also provides a simple and straightforward basis for identifying and filtering out features that are relatively uninformative prior to any downstream analysis: if the measurement model sufficiently describes the observed counts for a given gene, then no meaningful biological inference can be drawn based on the counts of that gene. This is so since all the variation in the counts for that gene is attributable to the measurement process itself, with negligible or no contribution from any biological process. We call such genes whose counts are adequately described by the measurement model as *biologically uninformative*. Since a simple Poisson distribution suffices to explain measurement error, using the estimates for α_{QP} we can easily identify genes that are likely to be biologically uninformative - genes with $\alpha_{QP} \leq 1$ do not exhibit over-dispersion with respect to the Poisson measurement model and can be filtered out.

While we can simply rely on the α_{QP} to identify biologically uninformative genes, identification of HVGs based on just the magnitudes of α_{QP} would result in a bias towards genes with higher mean expression levels (recall that the α_{QP} exhibit a pronounced dependence on μ for genes with higher mean expression levels (see Fig. 1B)). Keeping this in mind, we propose the following approach to shortlist HVGs:

- Group the genes into bins (default is 1000 bins) based on their mean expression levels; each bin contains approximately the same number of genes
- Sort the genes within each bin into quantiles based on their α_{QP}
- Obtain the α_{QP} corresponding to the reference quantile within each bin (default reference is the 10th quantile) - we refer to this as $\alpha_{QP(Reference|Bin)}$
- Calculate $\alpha_{QP} - \alpha_{QP(Reference|Bin)}$ for each gene - larger values would indicate greater over-dispersion

We illustrate the binning process involved in our feature selection method for `Svensson 1` (Fig. 2A). For ease of illustration we show 10 bins, where each bin corresponds to a segment or region between the vertical dashed lines. Since `Svensson 1` is a technical control data set, we don't expect to see much variability in the expression levels of the genes, and indeed in Fig. 2C we can see that there are just two genes (spike-in transcripts ERCC-00074 and ERCC-00130) that exceed the threshold of 20 for $\alpha_{QP} - \alpha_{QP(Reference|Bin)}$ (red horizontal dashed line). For `PBMC 33k`, in contrast, we observe 15 genes that exceed the threshold of 20 for $\alpha_{QP} - \alpha_{QP(Reference|Bin)}$ (Fig. 2D). Clearly, if we lowered the threshold we would identify many more variable genes in `PBMC 33k` compared to `Svensson 1`.

Aside from identifying variable genes, we can also shortlist genes that do not exhibit much variability in their counts. We refer to these genes as *stable genes*. Identification of such genes is very useful since we expect that the variability in the counts of these genes is primarily attributable to the measurement process and is not confounded by biological differences between the cells. Thus, we can rely on these genes to obtain more reliable estimates for the cell-specific size factors in order to perform normalization. Based on our feature selection method, genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} < 0$ are shortlisted as stable genes. For `PBMC 33k`, two of these genes (*PFDN2* and *RPL32*) are labeled in Fig. 2D below the horizontal black line at $\alpha_{QP} - \alpha_{QP(Reference|Bin)} = 0$.

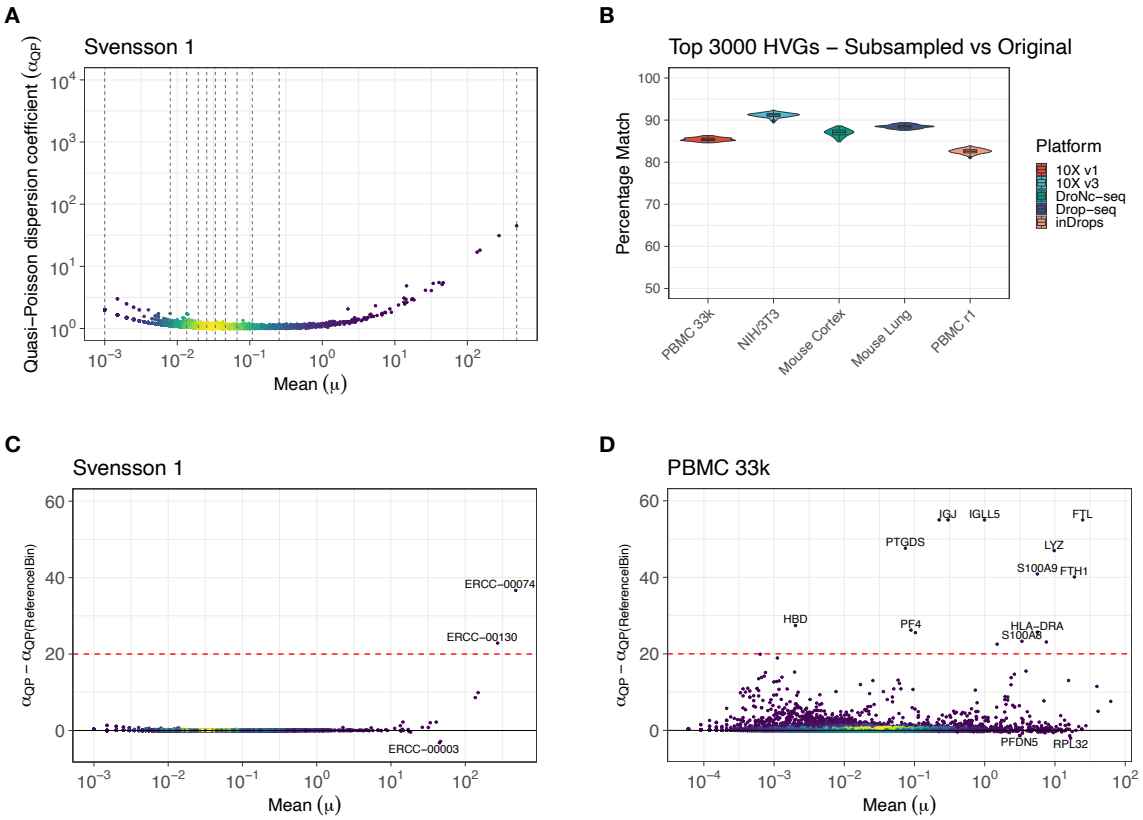

Figure 2. Feature selection can be performed before normalization based on dispersion coefficients estimated using the observed counts. A. Quasi-Poisson dispersion coefficient (α_{QP}) vs mean (μ) log-log scatter plots for genes that exhibited over-dispersion with respect to the Poisson ($\sigma^2 > \mu$) for the Svensson 1 technical control data set. Each dot represents a gene. The dashed gray vertical lines illustrate how the genes are binned based on their mean expression levels (in this figure, there are 10 bins). Genes within adjacent pairs of dashed vertical lines belong to the same bin. For each bin, the default choice of $\alpha_{QP(Reference|Bin)}$ is the α_{QP} corresponding to 10th quantile within the bin. **B.** Violin-box plots showing the percentage match between the top 3000 highly variable genes (HVGs) identified by our feature selection method for the original (unsubsampled) and corresponding 100 subsampled data sets. The original 5 data sets (PBMC 33k, NIH/3T3 cell line, Mouse cortex, Mouse lung, and PBMC r1) were obtained using different single-cell platforms - 10X Chromium v1, 10X Chromium v3, DroNC-seq, Drop-seq, and inDrops respectively. The box/violin plots are colored according to the platforms. For all 5 data sets, more than 80% of the top 3000 HVGs shortlisted for the subsampled data sets matched the top 3000 HVGs of the original data sets. **C.** $\alpha_{QP} - \alpha_{QP(Reference|Bin)}$ vs mean (μ) linear-log scatter plot for genes that exhibited over-dispersion with respect to the Poisson ($\sigma^2 > \mu$) for the Svensson 1 data set. The dashed red horizontal line illustrates the threshold to shortlist HVGs. Genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} > 20$ in this case are shortlisted as HVGs. On the other hand, genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} < 0$ are shortlisted as stable genes. **D.** $\alpha_{QP} - \alpha_{QP(Reference|Bin)}$ vs mean (μ) linear-log scatter plot for genes that exhibited over-dispersion with respect to the Poisson ($\sigma^2 > \mu$) for the PBMC 33k data set. Labeled genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} > 20$ are shortlisted as HVGs. On the other hand, genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} < 0$ are shortlisted as stable genes (some labeled in figure). An important point to note here is that with this feature selection approach, HVGs are shortlisted across the different expression levels with no preferential selection based on the mean expression level of the genes.

HVGs can be consistently identified despite the introduction of systematic biases

Having introduced the feature selection method, we can now test the expectation that the HVGs can be identified despite the presence of systematic biases. To perform the test, we picked 5 UMI counts data sets obtained from different platforms: `PBMC 33k` (10X Genomics Chromium v1), `NIH/3T3` (10X Genomics Chromium v3), `Mouse Cortex` r1 (DroNC-seq) [8], `Mouse Lung` (Drop-seq) [25], and `PBMC r1` (inDrops) [8]; r1 denotes replicate 1. For each of these data sets, we randomly picked 30% of the cells and subsampled the counts in those cells to a fraction of the original total count. The fractions were allowed to take one of the following values - 0.3, 0.4, 0.5, 0.6, 0.7 - and were picked randomly for each cell. We did this 100 times for each of the 5 data sets. Using our feature selection method, we shortlisted the top 3000 HVGs in each of the subsampled data sets and compared with the top 3000 HVGs in the respective original (unsubsampled) data sets to see how many of the top 3000 HVGs matched between the two. We observed that on average more than 80% of the top 3000 HVGs shortlisted for the subsampled data sets matched the top 3000 HVGs obtained from the original data sets across the different platforms and tissue types (Fig. 2B). Thus, despite the introduction of random systematic biases there is good agreement between the HVGs obtained for the original data sets and the HVGs obtained for corresponding data sets with the introduced biases. This supports our assertion that feature selection can be performed prior to normalization.

Normalization can be performed using a residuals-based approach that includes variance stabilization

Cell-specific size factors should be estimated using stable genes

The basic objective of normalization is to reduce systematic biases introduced due to technical or potentially uninteresting biological sources (such as cell size, cell cycle state) before further downstream analyses such as clustering and differential expression. The most common approach to reduce such systematic biases is to scale the counts within each cell by cell-specific *size factors* [13,26]. The simplest estimates are given by,

$$SF_c = \frac{\sum_g X_{gc}}{(\sum_c \sum_g X_{gc})/C} \tag{6}$$

where X_{gc} is the observed count of gene g in cell c (N_c), and C is the total number of cells. The numerator in equation. (6) is the total UMI count in cell c, while the denominator is the mean of the total counts of the cells.

There is an intimate link between the estimates of size factors given by equation. (6) and the estimates for expected means ($\hat{\mu}_{gc}$) under the assumption that the counts are Poisson distributed (see Appendix), which suggests that the size factor estimates obtained from equation. (6) should be viewed as estimates under the approximation that the counts are Poisson distributed. Keeping this in mind, we argue that it is most appropriate to calculate the size factors by relying on the stable genes identified using our feature selection method. The counts of these genes do not exhibit significant over-dispersion compared to the Poisson, and as discussed earlier, we expect that the primary source of the variability of their counts is the measurement process which is what we are trying to account for with the help of the size factors. Therefore, we propose the following refinement to the estimation of the size factors,

$$SF_c = \frac{\sum_{g \in Stable\ g} X_{gc}}{(\sum_c \sum_{g \in Stable\ g} X_{gc})/C} \tag{7}$$

where *Stable* g refers to the set of stable genes.

Limitations of widely used normalization methods

In the standard scRNA-seq workflow, normalization and feature selection are typically followed by dimensionality reduction using principal components analysis (PCA). An important step prior to PCA is to ensure variance stabilization by transforming the size factor adjusted counts using the *asinh*, *log*, *sqrt* functions etc. This ensures that genes with higher expression levels (and as a consequence larger variance) do not contribute disproportionately to the overall variance as evaluated through PCA. The *log*-based variance stabilization transformation (hereafter referred to as `logSF`) is the most popular method according to which the transformed counts are given by,

$$X'_{gc} = log(\frac{X_{gc}}{SF_c} + 1)$$

where the pseudocount of 1 ensures that the *log* transformation works with zero counts, and in fact returns zeros for these counts even after transformation. However, the `logSF` normalization is not very effective since the total counts of the cells (N_c) shows up as a primary source of variation in PCA even after normalization (see Fig. S6). This can be traced to the fact that under this transformation, the zeros remain zeros while only the non-zero counts are scaled according to the size factors. Systematic differences in the number of zero counts between the cells can therefore be identified as a major source of variation even after transformation [27, 28].

Residuals-based approaches proposed by Hafemeister et al. [20], Townes et al. [27], and more recently by Lause et al. [21] provide alternatives that lead to much more effective normalization (see Fig. S7). The Pearson residuals under the NB model are approximately given by,

$$r_{gc} = \frac{X_{gc} - \hat{\mu}_{gc}}{\sqrt{\hat{\mu}_{gc} + \hat{\alpha}_g \hat{\mu}_{gc}^2}} \tag{8}$$

where X_{gc} is the observed count of gene g in cell c, $\hat{\mu}_{gc}$ is the estimated mean of gene g in cell c, and $\hat{\alpha}_g$ is the estimated over-dispersion parameter for gene g.

The increased effectiveness of normalization with the Pearson residuals can be attributed to $\hat{\mu}_{gc}$ taking into account the systematic differences in the total counts between the cells. Unlike the case with `logSF` normalization where zero counts are transformed back to zeros, the zero counts are instead transformed to negative residual values whose magnitudes vary depending on the total counts of the respective cells.

The rationale underlying the residuals-based approach is that the null model should correspond to the measurement process so that the residuals provide estimates for deviations away from the expectations under the measurement model. However, the sparsity and the skewed nature of the counts distributions pose significant challenges to achieving effective variance stabilization with the help of residuals-based methods. The lack of variance stabilization becomes especially noticeable for genes that are robustly expressed in only a subset of cells while showing negligible expression in the rest of the cells (such genes would be considered as *markers* of the specific cell sub-populations in which they are expressed) [22] (see Appendix for more discussion on this). Furthermore, for genes with very low mean expression levels (and as a consequence extremely small estimated standard deviations) even cells with just 1 or 2 UMI counts sometimes end up with unusually large residual values that are then addressed through heuristic approaches [20, 21, 29].

z-score based normalization that includes variance stabilization

Keeping in mind the limitations of the widely used normalization methods discussed above, we propose a conceptually simple residuals-based normalization method that reduces the influence of systematic biases by relying on size factors estimated using stable genes while simultaneously ensuring variance stabilization by explicitly relying on a variance stabilization transformation.

In order to motivate our approach, we begin by pointing out that z-scores are simply Pearson residuals corresponding to the normal distribution. Assuming counts, Y_{gc}, that are normally

distributed,

$$Y_{gc} \sim N(\mu_g, \sigma_g^2)$$

the MLEs for μ_g and σ_g^2 are given by, $\hat{\mu}_g = (\sum_c Y_{gc})/C$ and $\hat{\sigma}_g^2 = (\sum_c (Y_{gc} - \hat{\mu}_g)^2)/C$, respectively. The corresponding Pearson residuals-based on these estimates are given by,

$$r_{gc} = \frac{Y_{gc} - \hat{\mu}_g}{\hat{\sigma}_g} = z_{gc} \tag{9}$$

which as already noted above correspond to z-scores (for simplicity, we assumed that there are no differences in total counts between the cells).

In order to compute z-scores for our data, we first need to apply a variance stabilization transformation to the raw counts (X_{gc}) to bring their distribution closer to the normal distribution. The variance stabilization transformation can be performed using monotonic non-linear functions, $g(X)$, such that the transformed counts are given by,

$$Y = g(X)$$

To compute the residuals, we need estimates for the means and variances of Y based on estimates for means and variances of X. We can arrive at approximations for both using a Taylor expansion around $X = \mu$ (see Appendix). In particular for $g(X) = log(X + 1)$, the first order approximations of the mean and variance are given by,

$$E[log(X + 1)] \approx log(\mu + 1) \tag{10}$$

$$Var[log(X + 1)] \approx \frac{1}{(\mu + 1)^2}\sigma^2 \tag{11}$$

At this point we need estimates for the means ($\hat{\mu}_{gc}$) and the variances ($\hat{\sigma}_{gc}^2$) that account for the systematic biases. Note, that the usual estimate for the mean expression level of gene g based on the observed counts ($\mu_g = \frac{\sum_c X_{gc}}{C}$) includes biological as well as technical effects. However, we are interested in inferring the mean expression level and variance that is primarily reflective of the underlying biology. This can be accomplished by relying on the size factors (equation. (7)) to re-scale the observed counts of the respective cells and then obtaining the estimates for mean and variance based on the scaled counts. Thus,

$$\tilde{\mu}_g = \frac{1}{C}\sum_c \frac{X_{gc}}{SF_c}$$

$$\tilde{\sigma}_g^2 = \frac{1}{C-1}\sum_c (\frac{X_{gc}}{SF_c} - \tilde{\mu}_g)^2$$

Using the estimates for mean and variance for gene g, we get the following estimates for mean and variance of gene g in each cell c,

$$\hat{\mu}_{gc} = SF_c \tilde{\mu}_g \tag{12}$$

$$\hat{\sigma}_{gc}^2 = SF_c^2 \tilde{\sigma}_g^2 \tag{13}$$

With these estimated means ($\hat{\mu}_{gc}$) and variances ($\hat{\sigma}_{gc}^2$) that account for the systematic biases, from equations. (10) and (11) we get,

$$E[log(X_{gc} + 1)] \approx log(\hat{\mu}_{gc} + 1) \tag{14}$$

and

$$Var[log(X_{gc} + 1)] \approx \frac{1}{(\hat{\mu}_{gc} + 1)^2}\hat{\sigma}_{gc}^2 \tag{15}$$

Based on these first-order approximations for means and variances under the $log(X+1)$ transformation, we now define our z-scores based normalization,

$$Z'_{gc} = \frac{log(X_{gc}+1) - log(\hat{\mu}_{gc}+1)}{\hat{\sigma}_{gc}/(\hat{\mu}_{gc}+1)} \tag{16}$$

This log-stabilized z-score transformation is the default normalization method in our R package called `Piccolo` (see Appendix for a discussion on other variance stabilization approaches implemented in `Piccolo`). Hereafter, we refer to it as the `Piccolo` normalization.

Piccolo normalization reduces the impact of sampling depth differences between cells while simultaneously ensuring variance stabilization

As stated earlier, the objective of normalization is to reduce or eliminate the systematic biases in counts between cells that are not reflective of actual biological differences. Since technical control data do not have any biological source of variation, differences in sampling depths are expected to be the major source of variation between the cells (droplets). In terms of PCA, this would translate to sampling depth showing up as a major contributor in one of the first few principal components (PCs).

In Fig. 3A, we show the scatter plots of cells based on their coordinates along PC1 and PC2. The colors of the dots reflect the size factors of the respective cells, with brighter shades (yellow) indicating larger size factors. Recall that larger size factors correspond to cells with larger sampling depths across the stable genes (see equation. (7)). In order to evaluate whether size factors correlate with the first few PCs, we calculated the canonical correlation coefficient (ρ) [30] between the size factors and the top 5 PCs; ρ close to 1 would indicate strong correlation between the size factors and one of the top 5 PCs. For `Svensson 1` raw counts (left panel Fig. 3A), we can clearly observe a color gradient along PC1, with cells with larger size factors lying predominantly on the left and cells with smaller size factors predominantly on the right. Thus for raw counts, sampling depth differences indeed show up as a major source of variation between the cells. The value of the canonical correlation coefficient - $\rho = 0.97$ - supports this further.

Next, we applied the `Piccolo` normalization to `Svensson 1` (right panel in Fig. 3A) and confirmed that the sampling depth differences are no longer identified as a major source of variation by PCA. In fact, not only do we not observe a color gradient along PC1 or PC2, but even the canonical correlation coefficient between the size factors and the top 5 PCs is significantly reduced to $\rho = 0.3$ which suggests a weak correlation. Similar observations were made for another technical control data set [31] (see Fig. S8). These results demonstrate that the `Piccolo` normalization does reduce the impact of the systematic differences in sampling depths.

To examine the effectiveness of variance stabilization, we looked at the variances of the residuals after applying normalization to the raw counts. We compared `Piccolo` with two other residuals-based normalization methods - analytic Pearson [21](`Analytic Pearson`), and the regularized NB regression approach in scTransform [20, 29] (`SCTransform v2`). For reference, we also looked at the simple `logSF` based normalization approach implemented in Scran [12] (`Scran logSF`). Note, Scran relies on pooling of cells to arrive at estimates for the cell-specific size factors that are then used for the `logSF` normalization.

For the `PBMC 33k` data set, we first used our feature selection method to shortlist the top 3000 HVGs, and then applied the `Piccolo` normalization to compute the residuals corresponding to the raw counts of these HVGs. The `Analytic Pearson` residuals were also computed for these top 3000 HVGs identified with our feature selection method. This enables a direct comparison between the two methods since the residuals were calculated using the same set of features. In contrast, for `SCTransform v2` and `Scran logSF`, the top 3000 HVGs were shortlisted using their own respective approaches. For each of the normalization methods, we then calculated the variances of the residuals. These variances are shown in the variance-mean linear-log plots in Fig. 3B. It is immediately apparent that the residuals obtained from `Piccolo` (Fig. 3B, top-left panel) exhibit much better variance

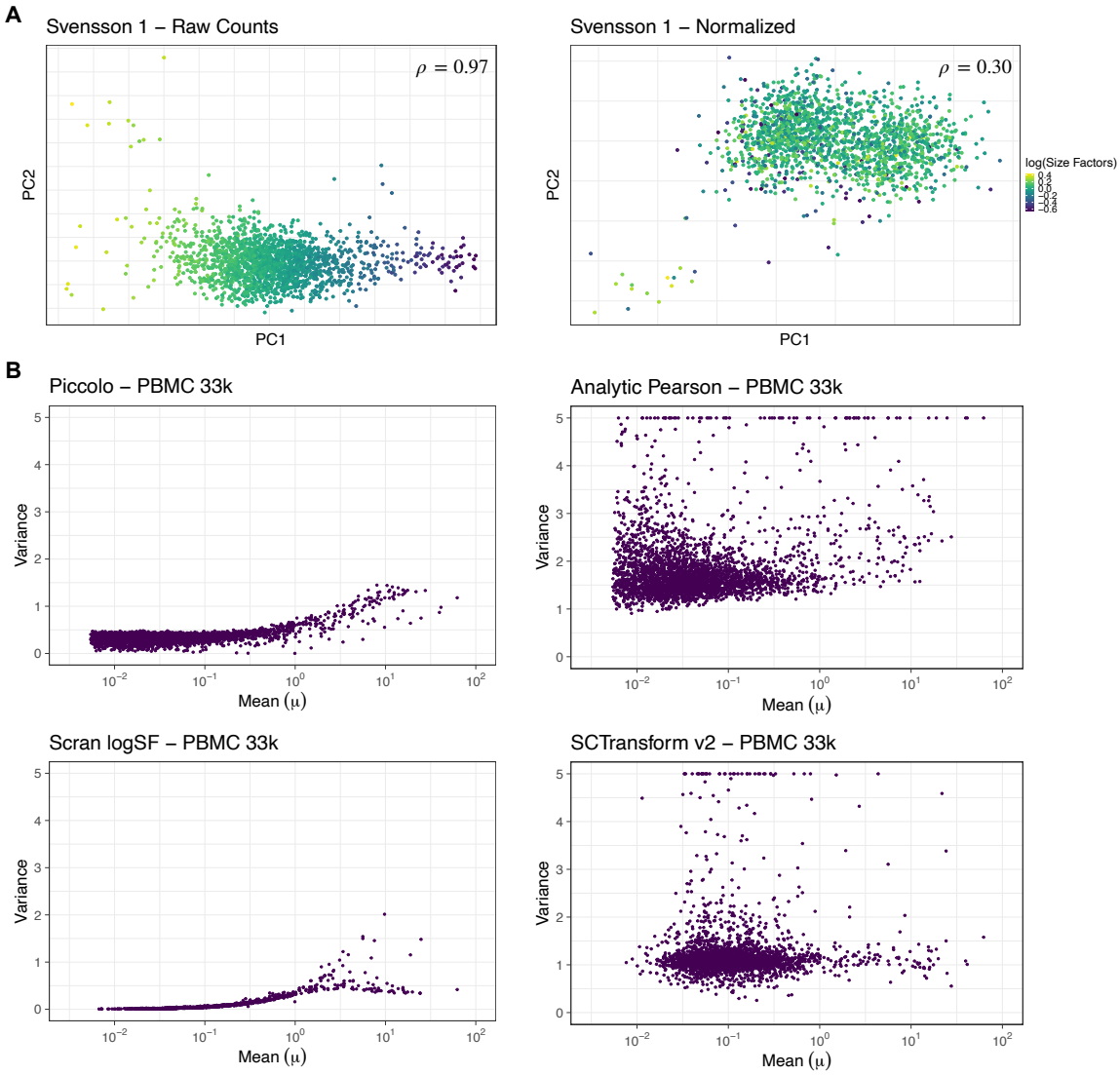

Figure 3. Piccolo normalization reduces sampling depth differences between cells and also ensures effective variance stabilization. A. 2-dimensional (2D) scatter plots based on the first 2 PCs of the Svensson 1 technical control data set. Each dot is a cell and is colored according to the size factors; brighter shades (yellow) correspond to larger size factors and darker shades (deep blue) correspond to smaller size factors. The left panel shows the 2D PC scatter plot for the raw counts, while the right panel shows the 2D PC scatter plot for the z-scores (residuals) obtained from Piccolo normalization. The coefficient (ρ) in the top-right corner of the panels shows the canonical correlation coefficient between the size factors and the top 5 PCs. Smaller values of ρ indicate that the impact of the sampling depth differences has been reduced more effectively by the normalization. Piccolo normalization reduces the impact of sampling depth on the overall variation as evaluated through PCA. **B.** Variance vs mean (μ) linear-log scatter plots for the top 3000 HVGs of the PBMC 33k data set after applying respective normalizations - Piccolo (top-left panel), Analytic Pearson residuals (top-right), Scran logSF (bottom-left), and SCTransform v2 (bottom-right). The y-axis scale was limited to a maximum value of 5 to aid visual comparison; genes with variance greater than 5 were clipped to have the maximum value of 5. The residuals obtained using Piccolo exhibit variances that are much more stable across all mean expression levels compared to the residuals obtained using the other residuals-based methods.

stabilization compared to the other two residuals-based approaches (Fig. 3B, top-right panel and bottom-right panel). In Fig. 3B bottom-left panel, we also show the variance of the *log* transformed counts for reference.

Piccolo feature selection and normalization lead to substantial improvements in cell clustering

Piccolo normalization preserves cell-cell similarities between cells that share cell-type identities

Normalization is typically followed by dimensionality reduction and unsupervised clustering to identify groups of cells with similar expression profiles. Depending on the biological system, the groups of cells may correspond to distinct cell-types, or states. The identification of such groups is a pivotal step since it informs crucial downstream analyses such as differential expression and marker genes identification. The most popular scRNA-seq workflows (for example, Seurat [15–18] and Scanpy [19]) employ PCA to perform dimensionality reduction. Based on the PCs, k nearest-neighbour (kNN) graphs are generated (with cells as nodes) in which communities of cells that are most similar to each other are detected using graph-partitioning algorithms such as Leiden [32] and Louvain [33].

To investigate how well our normalization method preserves cell-cell similarities between cells that share cell-type identities, we began by examining a truth-known data set (data set in which the cell-type identities of the cells are already known) prepared by Duo et al. [34] using cells purified with cell-type specific isolation kits by Zheng et al. [35]. Briefly, they prepared the data set by shortlisting purified cells belonging to 8 PBMC cell-types - B-cells, CD14 monocytes, CD56 NK cells, CD4 T-helper cells, memory T-cells, naive T-cells, naive cytotoxic T-cells, and regulatory T-cells - such that there were roughly equal numbers of cells corresponding to each cell-type in the final data set (between 400-600 cells per cell-type). We refer to this data set as `Zheng Mix 8eq`. Since `Zheng Mix 8eq` consists of a mix of well-separated cell-types (for instance, B-cells vs T-cells) and similar cell-types (different types of T-cells), it provides a simple yet reasonably challenging scenario for evaluating how well cells belonging to the different cell-types can be distinguished after normalizing the counts with the respective normalization methods.

We used `Piccolo` to shortlist the top 3000 HVGs and applied our normalization to obtain the residuals for those HVGs. For `Analytic Pearson`, the residuals were computed for the top 3000 HVGs shortlisted using our feature selection method, while for `Scran logSF` and `SCTransform v2` the transformed counts and residuals were respectively computed for the top 3000 HVGs shortlisted using their own methods. Subsequently, we performed PCA and shortlisted the first 50 PCs. We then used a simple kNN-based classification approach based on the PCs to predict cell-type labels for each cell by relying on the known cell-type labels (see Methods). Finally, we evaluated the extent of the agreement between the predicted labels and the known labels by calculating the following clustering metrics: the Macro F1 score, the adjusted Rand index (ARI), and the adjusted mutual information (AMI).

In Fig. 4A, we show the Uniform Manifold Approximation and Projection (UMAP) [36] plots for `Zheng Mix 8eq` with `Piccolo` normalization (top-left panel), `Analytic Pearson` (top-right panel), `Scran logSF` (bottom-left panel), and `SCTransform v2` (bottom-right panel). In the plots, each dot represents a cell and is colored according to the known cell-type labels (legend provided at the bottom of the 4 panels). Qualitatively, it is apparent from the UMAP plots that the B-cells, CD56 NK cells, and CD14 monocytes can easily be distinguished compared to the rest. As expected, it's more difficult to distinguish between the different kinds of T-cells, with memory and naive cytotoxic T-cells being the only ones that are comparatively easier to distinguish, particularly with the residuals-based approaches. The values of ARI, AMI, and Macro F1 quantifying the extent of the agreement between the predicted and the known cell labels are listed in the bottom-right corner of the panels for the respective normalization methods. While there isn't a significant difference in the metrics between the 4 normalization methods, we do observe the highest values with `Piccolo`.

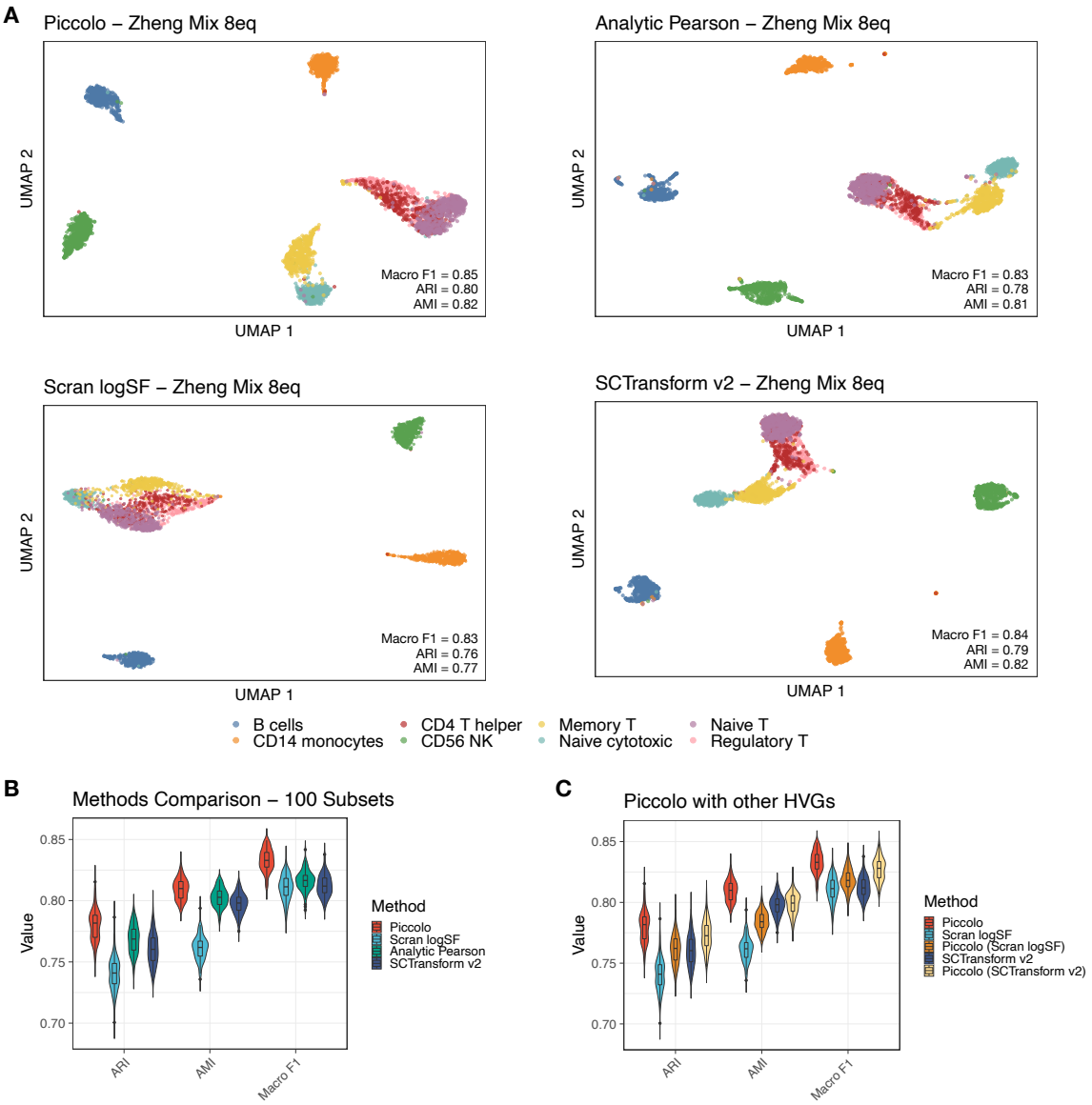

Figure 4. Piccolo normalization preserves cell-cell similarities between cells sharing cell-type identities. A. 2D UMAP plots after applying respective normalizations for the top 3000 HVGs of the Zheng Mix 8eq data set (consists of 3994 cells with roughly equal numbers of cells belonging to 8 distinct PBMC cell-types). Dots represent cells and are colored using the known cell-type labels (legend at the bottom of the panel). Clustering metrics - ARI, AMI, Macro F1 - based on comparisons between predicted cell labels (obtained using a kNN-based classification approach) and known cell labels are listed in the bottom-right corner in each panel. **B.** Violin-box plots of the clustering metrics obtained for 100 subsets of the Zheng Mix 8eq data set. The colors correspond to the respective normalization methods used. For all the 100 subsets, the highest values of the metrics was observed with Piccolo (red). **C.** Violin-box plots comparing the clustering metrics obtained for 100 subsets of the Zheng Mix 8eq data set with Piccolo normalization used with HVGs shortlisted obtained from other methods; Piccolo (Scran LogSF) and Piccolo (SCTransform v2) refer to the methods where the top 3000 HVGs were obtained using Scran logSF and SCTransform v2 respectively, and then the normalization was performed on those HVGs using Piccolo. Piccolo yielded higher values of the clustering metrics despite relying on HVGs shortlisted by other methods.

However, these observations are not sufficient to argue for or against any of the methods. For a more robust comparison between the normalization methods, we created 100 subsets of `Zheng Mix 8eq` by randomly picking 50% of the cells in the original data set 100 times. For each of the 100 subsets, we used the same approach as discussed above for the 4 normalization methods and computed the respective ARI, AMI, and Macro F1 scores for the predicted cell labels based on the kNN-based classification approach. In Fig. 4B, we show the violin-box plots for the ARI, AMI, and Macro F1 scores for the 100 subsets. The colors correspond to the respective normalization methods used (red corresponds to `Piccolo`). We can clearly see that for all 3 clustering metrics, the highest values were consistently obtained with `Piccolo` normalization, reflecting that the best agreement between predicted and known labels is achieved using our feature selection and normalization method. For each clustering metric, we used paired Wilcoxon tests to quantify whether the differences between the values of the metric obtained with `Piccolo` normalization and other normalization methods were statistically significant (see Methods). For all 3 metrics, the values obtained with `Piccolo` normalization were found to be consistently higher than those obtained with other methods (paired Wilcoxon test $p < 1E-11$).

For our analyses on `Zheng Mix 8eq` so far, while `Piccolo` and `Analytic Pearson` normalization were applied on the same set of HVGs, the sets of HVGs for `Scran logSF` and `SCTransform v2` were different. Thus, some of the differences in the results are attributable to the differences in the sets of HVGs. To compare just the normalization methods, we used the top 3000 HVGs shortlisted by `Scran logSF` and `SCTransform v2` respectively, and computed the residuals using `Piccolo` normalization for these respective HVGs. `Piccolo (Scran logSF)` denotes the method wherein the HVGs were obtained from `Scran logSF` and then the `Piccolo` normalization was performed for those HVGs. Similarly, `Piccolo (SCTransform v2)` denotes the method wherein the HVGs were obtained from `SCTransform v2` and then the `Piccolo` normalization was performed for those HVGs. By applying these methods, we arrived at clustering metrics for the 100 subsets which are shown using the violin-box plots in Fig. 4C. The colors correspond to the feature selection and normalization methods used. We used paired Wilcoxon tests to compare the values of the metrics obtained with the normalization method that was used to shortlist the HVGs, with values of the metrics obtained with `Piccolo` normalization using those HVGs. For ARI and Macro F1, the values obtained with `Piccolo` normalization (`Piccolo (Scran logSF)` and `Piccolo (SCTransform v2)`) were consistently higher than the values obtained with the respective normalization methods with which the HVGs were shortlisted (paired Wilcoxon test $p < 1E-09$). For AMI, while the values with `Piccolo (Scran logSF)` were consistently higher than with `Scran logSF` (paired Wilcoxon test $p < 1E-17$), the values obtained with `Piccolo (SCTransform v2)` were not as significantly high compared to the ones obtained with `SCTransform v2` (paired Wilcoxon test $p < 0.06$). Overall, it is clear that even with other HVGs, `Piccolo` normalization is better at preserving cell-cell similarities between cells of the same cell-type compared to the other normalization methods. Moreover, from Fig. 4C what is most striking is that the values of the clustering metrics obtained using our feature selection and normalization are the highest overall (red violin-box plot corresponds to `Piccolo`). This clearly suggests that not only is the performance of `Piccolo` normalization consistently better than the other normalization approaches (assessed with the aid of the clustering metrics), but even our feature selection method is more effective at shortlisting HVGs that better inform the differences between cells with distinct cell-type identities.

Piccolo enables identification of groups containing few cells as well as groups with cells that express fewer differentially expressed genes

To further investigate the performance of `Piccolo`, we utilized Splat [37], a simulation framework that relies on the gamma-Poisson distribution to simulate counts based on estimation of parameters for real UMI counts data sets. Using Splat, we can specify both the number of groups, as well as the extent of differentially expressed genes in each group (this is specified in terms of the probability that a gene will get picked to be differentially expressed in each group). In order to generate the simulated counts, it requires a homogeneous cell population based on which the parameters for the

simulation are estimated. Thus, we used `NIH/3T3` to simulate the counts since the cells in this data set are derived from a cell line and therefore constitute a homogeneous cell population.

We generated simulated counts using `NIH/3T3` for the following two scenarios:

- `Simulated Counts Scenario 1`: 6 groups with different number of cells per group, while keeping the probability that any given gene is picked to be differentially expressed the same for each group. The objective behind this simulation was to examine whether the groups with the fewest cells can be reliably identified after applying the different normalization methods.

- `Simulated Counts Scenario 2`: 6 groups with the same number of cells per group, but with different probability for any given gene to be picked to be differentially expressed within each group. The objective behind this simulation was to examine how well we can distinguish between cells belonging to distinct groups, especially the ones that have fewer differentially expressed genes.

We will discuss each scenario in turn now. For `Simulated Counts Scenario 1`, we simulated 6 groups containing cells in the following proportions: Group 1 - 0.30, Group 2 - 0.30, Group 3 - 0.20, Group 4 - 0.15, Group 5 - 0.04, and Group 6 - 0.01. The probability that any given gene is picked to be differentially expressed was kept fixed at the default value of 0.1 for all the groups. We adopted the same procedure as described in the previous subsection to prepare the UMAP plots for each of the 4 normalization methods - `Piccolo`, `Analytic Pearson`, `Scran logSF`, and `SCTransform v2` respectively (Fig. 5A). The cells are colored according to the group to which they belong. The legend for the group identities is provided at the bottom of Fig. 5A. Using the kNN-based classification approach, we predicted the group labels for every cell by relying on the known group labels. The clustering metrics quantifying the extent of the agreement between the predicted and the known group labels are listed in the bottom-right corner of the panels for the respective normalization methods. From just the qualitative differences in the nature of groupings of the cells in the respective UMAPs, it is clear that `Piccolo` performs quite favorably compared to the other normalization methods. The difference is especially stark for the groups containing the fewest cells (Group 5 - 93 cells, and Group 6 - 35 cells); with our feature selection and normalization (top-left panel in Fig. 5A), we observe that cells belonging to the respective groups form clusters which can be easily distinguished from clusters corresponding to other groups. For the other normalization methods, distinct clusters corresponding to the smallest groups are not evident. The clustering metrics reflect these differences in the overall quality of the clustering as well, with `Piccolo` exhibiting the highest values for all 3 metrics. In Fig. 5B, we show the F1 scores per group based on the concordance between the predicted and the known labels for each of the normalization methods respectively. The tiles of the grid have been colored based on the values of the F1 scores, with brighter shades (yellow) indicating larger values (greater agreement between the predicted and the known group labels), and darker shades (dark blue) indicating smaller values (less agreement between the predicted and the known group labels). The difference between `Piccolo` and other methods is quite stark, with `Piccolo` clearly better at preserving cell-cell similarities between cells sharing group identities. Note, our feature selection also plays a key role by shortlisting variable genes that enable the identification of distinct clusters that correspond to the distinct groups.

For `Simulated Counts Scenario 2`, we simulated 6 groups containing roughly the same number of cells. However, this time we varied the probability that a gene is picked to be differentially expressed in each group as follows: Group 1 - 0.25, Group 2 - 0.2, Group 3 - 0.15, Group 4 - 0.1, Group 5 - 0.05, Group 6 - 0.025. The UMAP plots for each of the 4 normalization methods - `Piccolo`, `Analytic Pearson`, `Scran logSF`, and `SCTransform v2` - are shown in Fig. S10. Using the kNN-based classification approach, we predicted the group labels for every cell by relying on the known group labels. The clustering metrics based on the predicted labels are listed in the bottom-right corner of the panels for the respective normalization methods. Once again, from just the qualitative differences in the nature of clustering of the cells in the respective UMAPs, it is clear that `Piccolo` performs quite favorably compared to the other normalization methods, especially when it comes to distinguishing between the groups with the fewest differentially expressed genes

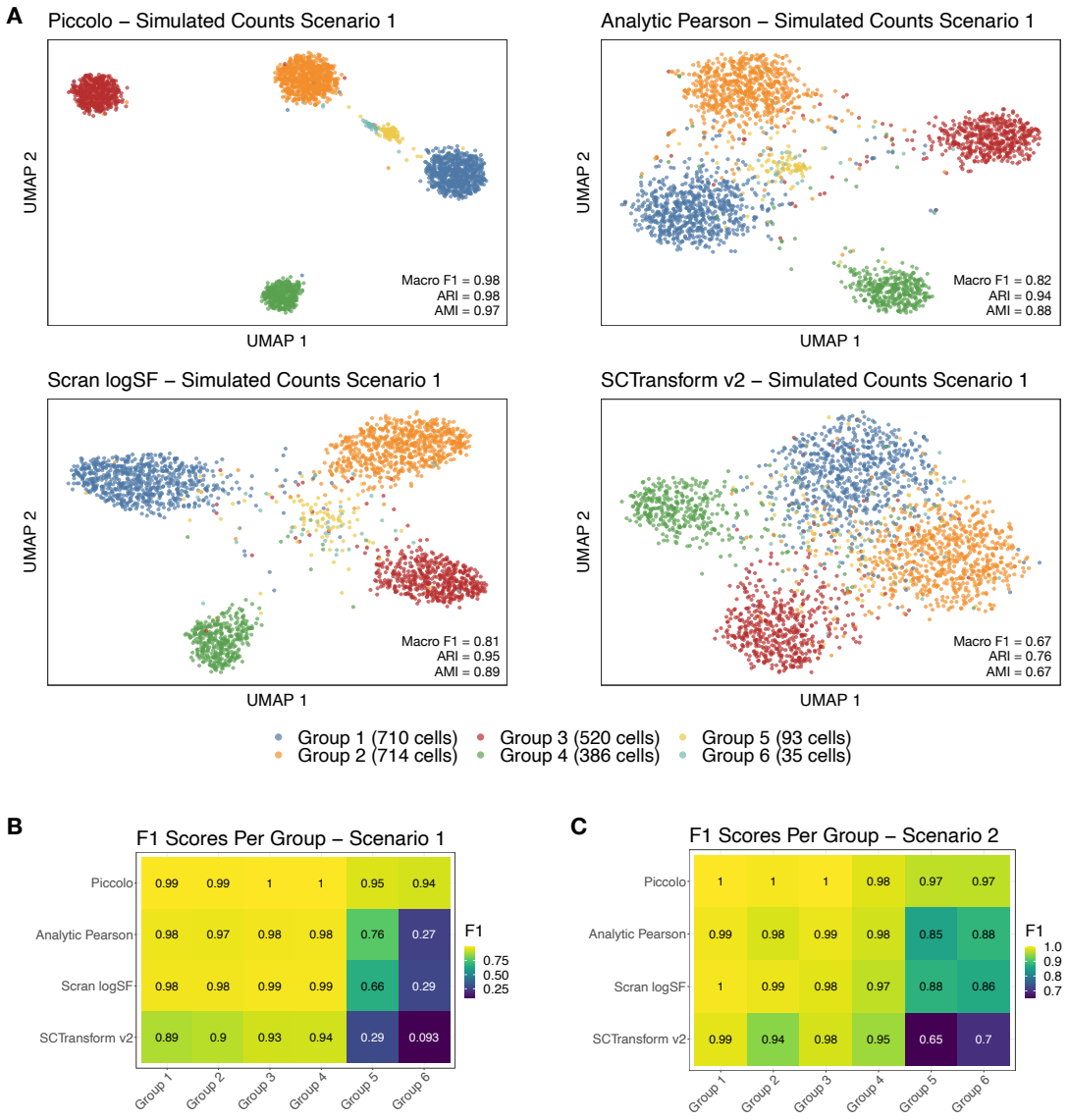

Figure 5. Piccolo enables identification of clusters even for groups containing small numbers of cells. A. 2D UMAP plots after using respective normalization methods for the top 3000 HVGs for the Simulated Counts Scenario 1 data set. The simulated counts were generated by Splat [37] using the NIH/3T3 data set. Dots represent cells and are colored using the known cell-type labels (legend at the bottom of the panel; actual numbers of cells belonging to each group are specified in parentheses). Clustering metrics based on comparisons between predicted cell labels and known cell labels are listed in the bottom-right corner in each panel. Cells belonging to the 6 groups can be more readily distinguished using Piccolo. **B.** Heatmap showing the F1 scores per group for each of the normalization methods applied to Simulated Counts Scenario 1. The tiles of the heatmap are colored according to the F1 values with larger F1 scores corresponding to brighter shades (yellow), and lower F1 scores corresponding to darker shades (deep blue). Piccolo has the highest F1 scores for all 6 groups. **C.** Heatmap of F1 scores per group for the Simulated Counts Scenario 2 data set (also see Fig. S10). Piccolo returns the highest F1 scores for all 6 groups. Note, the colors are scaled differently in panels B and C to bring out the differences between the F1 scores in the respective simulation scenarios.

(Group 5 and Group 6). The clustering metrics reflect these differences in the overall quality of the clustering as well, with `Piccolo` exhibiting the highest values for all 3 metrics. In Fig. 5C, we show the F1 scores per group based on the predicted labels for each of the normalization methods respectively. `Piccolo` clearly does better at preserving similarities as well as distinctions between cells compared to the other normalization methods. In particular, cells belonging to groups with fewer differentially expressed genes can still be reliably distinguished with the help of `Piccolo`. We replicated the above simulation scenarios with another cell line data set (`HEK293T` [10]) with similar results that are summarized in Fig. S11 and Fig. S12.

In summary, with the help of `Piccolo` we observed substantial improvements in the clustering outputs, both in terms of how well the cells sharing cell-type (or group) identities clustered together as well as how differences between cells with distinct cell-type (or group) identities were preserved making it easier to discriminate between them. The simulation scenarios were deliberately kept simple in order to facilitate a straightforward comparison and interpretation. The improvements in the results of clustering analysis enabled by `Piccolo` were especially significant in the case of these simulated counts data sets. With the other normalization methods, groups with fewer cells as well as groups with fewer differentially expressed genes could not be reliably identified after applying PCA and generating the UMAPs.

Discussion

In this article, we propose a revision of the scRNA-seq workflow in which feature selection precedes, and in fact informs normalization (see Fig. 6). Before proceeding to discuss the salient aspects of our revised workflow and the normalization method proposed in this paper, it is worthwhile to discuss the importance of variance stabilization particularly in the context of PCA. In PCA, features that have higher variances are assigned larger loadings/weights. In the context of gene expression, genes with high variance are those that have greater biological variation in their counts across the cells. Therefore, assignment of larger loadings to such genes will help distinguish between distinct cell types or states. However, in case of UMI counts data the variances of marker genes often exceed the variances of other genes by orders of magnitude. Under these circumstances, the lack of variance stabilization can adversely impact downstream clustering analysis based on PCA. In a recent article comparing different transformations for scRNA-seq data, Ahlmann-Eltze and Huber showed that the simple `logSF` normalization outperformed the residuals-based normalization methods particularly in terms of consistency [22]; consistency was evaluated by evenly splitting the data into two such that the two sets contained mutually exclusive sets of genes, and inquiring whether the cells shared the same sets of neighbouring cells between these splits (nearest neighbour cells were identified based on Euclidean distances between the cells in the PC space). This consistency is primarily attributable to the variance stabilization ensured by the log-transformation. Since the variances of the genes after the transformation are comparable, the pairwise Euclidean distances between the cells based on their PC coordinates in the two splits are similar, resulting in consistency in the sets of nearest neighbour cells. In contrast in the existing residuals-based methods, since there is no non-linear transformation of the observed counts to ensure variance stabilization, genes with large variances can alter the PCs significantly and end up reducing the consistency between the sets of nearest-neighbour cells. However, as pointed out earlier, a significant disadvantage of the `logSF` normalization is that it is not very effective at reducing the sampling depth differences between the cells (Fig. S6). In contrast, the residuals-based methods reduce sampling-depth differences between the cells much more effectively (Fig. S7).

Keeping in mind the respective advantages of the `logSF` normalization and the residuals-based methods described above, we proposed a residuals-based (z-scores) based normalization method in which the variance stabilization is achieved with the help of a *log*-transformation (the default monotonic non-linear transformation function used in `Piccolo`, see Appendix for other options). We relied on the first-order approximations for the means and variances of the transformed counts to arrive at the residuals. In addition, the first-order approximations were based on means and variances

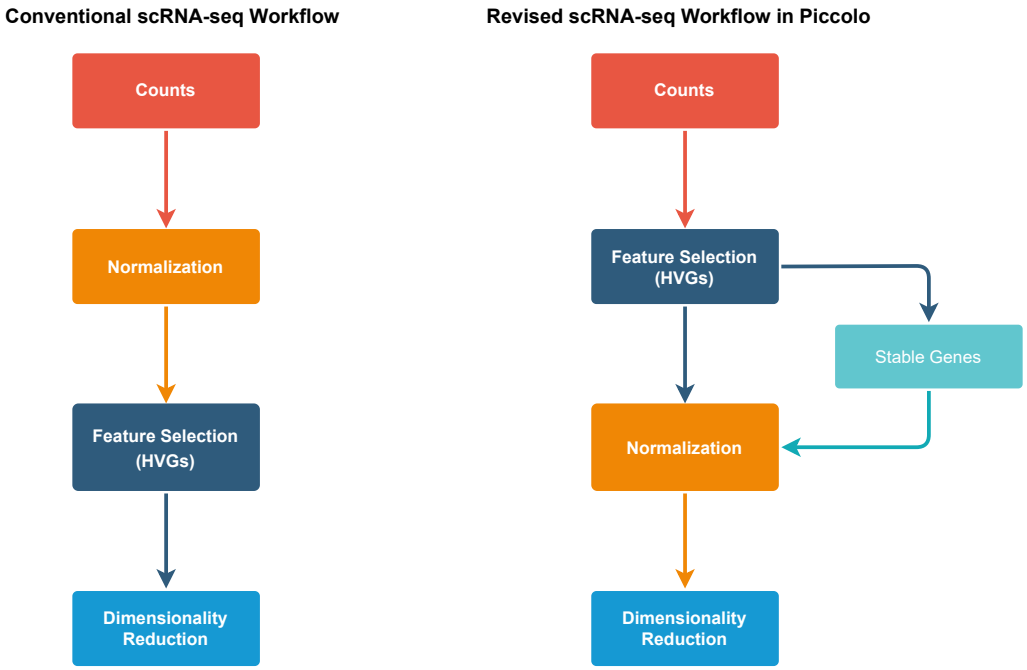

Figure 6. The revised scRNA-seq workflow implemented in Piccolo. We contrast the conventional scRNA-seq workflow (left) with the revised scRNA-seq workflow (right) proposed by us and implemented in Piccolo. The starting point for all analyses are the counts matrices. In the conventional workflow, the counts are first normalized and variance stabilized followed by feature selection to identify HVGs. Dimensionality reduction is then performed on the normalized counts. In contrast, in the revised workflow proposed by us, we first perform feature selection to identify both the HVGs as well as stable genes. We rely on the stable genes to estimate cell-specific size factors for performing normalization. During normalization, we also ensure variance stabilization which in turn leads to significant improvement while performing dimensionality reduction with PCA.

estimated using the size factor adjusted counts in order to reduce the impact of differences in sampling depths. Thus, our conceptually straightforward approach reduces the impact of the sampling depth differences and simultaneously ensures variance stabilization which leads to substantial improvements in the downstream clustering performance.

We first applied our normalization method on technical control data sets (Fig. 3A and Fig. S8) and demonstrated that it does reduce the impact of sampling depth differences between the cells. Using the PBMC 33k data set (and other data sets, see Fig. S9), we showed that our normalization method also ensures effective variance stabilization. We then applied `Piccolo` to a truth-known data set, and showed that it improves downstream clustering analysis and is able to preserve cell-cell similarities between cells that share cell-type identities better than other popular normalization methods. With the help of simulations, we were also able to show that `Piccolo` enables the identification of groups that contain small number of cells, whereas other normalization methods failed to reliably preserve cell-cell similarities between cells in these small groups. In addition, we showed that with `Piccolo` we can better distinguish between groups that express fewer differentially expressed genes. These results are especially significant when viewed from the perspective of the identification of rare cell types, or distinguishing between distinct cell states.

While our method offers significant improvements over the existing workflows, we now discuss some of its limitations. Beginning with feature selection, a fundamental assumption underlying our bin-based approach is that across all expression levels there are always some genes which are not biologically variable. This assumption is not motivated by biological observations and has been made to facilitate ease of computation. It is possible that we overlook some HVGs because of this

assumption. In our current implementation, we mitigate this by keeping the level for the reference dispersion coefficient in each bin relatively low (10th quantile is the default). However, there is definite scope for further improvement of the feature selection process to ensure that the HVGs are more effectively identified. Another point of consideration tied to feature selection is the use of stable genes for estimating cell-specific size factors. Due to the sparsity of the data, it is possible that the counts across the stable genes for some cells are all zero resulting in the size factor estimates to be zero for those cells. In `Piccolo`, we address this by iteratively adding sets of genes from the bottom of the list of HVGs till none of the cells have a size factor estimate of zero. Despite this limitation, we still expect that these size factors will not be confounded by actual biological variation as much as when we estimate them using all the genes (including HVGs). With regard to our normalization method, we relied on the first-order approximations for both the mean and the variance under the variance stabilization transformation. These approximations will work well as long as the non-linear transformation function is approximately linear in the range of the observed counts. For small counts, this is indeed true, however for larger counts (made possible through improvements in capture efficiencies, for instance) these approximations may lead to incorrect estimates. At present, our results suggest that the first-order approximations work quite well. Keeping future improvements in mind however, we point out that in principle it is possible to address this concern under our current framework by relying on higher-order approximations. We also want to note that in this study we did not discuss and elaborate on differential expression analysis which forms a vital component of all scRNA-seq studies. However, we do want to briefly mention here that after applying our normalization method to the observed counts, the distribution of the residuals are brought closer to the normal distribution, thus making it possible to employ the two-sample Student's t-test with the null hypothesis that the means of the two samples are the same. Typically, in scRNA-seq analyses the Wilcoxon rank-sum test, which is a non-parametric alternative to the two-sample t-test, is the preferred choice. However, for genes with low mean expression levels, we expect the t-test to return fewer false positives compared to the Wilcoxon rank-sum test (see Appendix for more further discussion).

Another point of consideration for single-cell workflows is the time taken to perform the normalization, as well as the memory (RAM) that is used. For the latter, since our residuals-based normalization transforms a sparse matrix to a dense matrix, the amount of memory that is used for the post-normalization matrix will increase significantly. However, since we can shortlist the HVGs before normalization, the transformed counts matrix will only be generated for the HVGs thus requiring lesser memory than what would be needed if normalization was performed for all genes. With regards to the computation time, since the transformation is based on a simple analytical relation, it only takes between 30 seconds to 3 minutes to transform the observed counts to the residuals for each of the data sets used in this paper (see Methods for system configuration).

In conclusion, the novel scRNA-seq workflow proposed in this study offers a paradigm shift that not only leads to conceptual simplifications but also leads to substantial improvements in the downstream analyses. We expect that the implementation of this workflow in `Piccolo` will facilitate more cogent and impactful inferences to be drawn from future single-cell gene expression studies.

Methods

Data preparation and preprocessing

The data sets used in the this study are listed in the table in the Code availability section, along with the links to the sources.

Cell and gene filtering

Primarily, the only cell filtering applied for all data sets was to ensure that all the cells had non-zero total counts. However for the evaluation of the effectiveness of our normalization method using PCA, we filtered cells from the `Svensson 1` data set that had total counts 3.5 median absolute deviation away from the median total count. This was done to reduce the impact of outliers on PCA.

For all the data sets, we excluded genes that had fewer than 0.5% cells with non-zero counts. This is the default gene filtering employed in `Piccolo`.

Selection of HVGs

For all the analyses, unless specified otherwise, we shortlisted the top 3000 HVGs. For `Piccolo` and `Analytic Pearson` residuals-based normalization, the identification of HVGs was done using our dispersion coefficient-based feature selection method. For `Scran logSF` and `SCTransform v2`, the top 3000 HVGs were shortlisted using their respective approaches that rely on post-normalization transformed counts/residuals to identify the genes with largest variances of the transformed counts/residuals.

Dimensionality reduction using PCA

After selecting the top 3000 HVGs and performing normalization on the counts of the HVGs, we used PCA for dimensionality reduction and shortlisted the top 50 PCs. Prior to using PCA, for the residuals-based methods (including `Piccolo`) we centered the residuals at 0, but did not scaled to unit variance. For the `logSF` normalization, we did not center or scale the transformed counts.

Benchmarking cell-type clustering and separation using kNN-based classification

Our kNN-based approach to predict cell labels using known cell labels is based on a very simple premise. After we normalize the observed counts and perform PCA, the expectation is that in the PC space the cells that share cell-type (or group) identities are close to each other. Thus, if we examine the nearest neighbours of a given cell belonging to a given cell-type (or group), we expect to find that the nearest neighbours are predominantly cells belonging to the same cell-type (or group). Based on this simple expectation, we predict cell labels for each cell by considering its k (default $k = 15$) nearest neighbours and using the known cell-type (or group) labels to identify the cell-type (or group) that is most over-represented in the nearest neighbour set - the most over-represented cell-type is the predicted cell-type label for the given cell. Over-representation is assessed using the hypergeometric test. By testing for over-representation, we ensure that there is no bias against cell-types (or groups) that have fewer cells while predicting the cell-type identity for any given cell.

Comparing clustering metrics obtained using the different normalization methods

For the 100 random subsets generated using the `Zheng Mix 8eq` data set, we applied the respective normalization methods and performed dimensionality reduction using PCA. We used the top 50PCs for each of them, and with our kNN-based classification approach predicted cell-type labels for each cell that we then compared with the known cell labels using the following clustering metrics: Macro F1 (harmonic mean between precision and recall, averaged across the classes), adjusted Rand index

(ARI), and adjusted mutual information (AMI). We compared the values of each of these metrics obtained with the different normalization methods by using the paired Wilcoxon rank-sum test. The null hypothesis is that there is no difference in the values of the metrics between the two groups being compared. The paired-test is essential in this context since the values of the metrics have to be compared pairwise for the same subset, and cannot be compared across different subsets.

The same 100 subsets were used to also evaluate how well `Piccolo` performs when we rely on HVGs shortlisted based on other normalization methods. Once again, we relied on the paired Wilcoxon rank-sum test to assess whether there are differences in clustering performance as evaluated through the clustering metrics: Macro F1, ARI, and AMI.

System configuration and software used

We used a Macbook Pro with the Apple M1 pro chip and 16GB RAM for all the analyses in this paper. `Piccolo` was developed using R (version 4.2.0), and all the analyses were also performed using R [38]. We used the following R packages in this study: cluster [39], data.table [40], dbscan [41, 42], ggplot2 [43], igraph [44], Matrix [45], matrixTests [46], RSpectra [47], Rtsne [48–50], umap [51], viridis [52]. In addition, the `Piccolo` package also includes a helper function made by Kamil Slowikowski called writeMMgz to help prepare .mtx.gz files.

Datasets

Name	Technology	Tissue	Type	Source	Reference
Svensson 1	10X Chromium v1	Technical Control	Technical Control	Link	[10, 53]
Klein 2015	inDrops	Technical Control	Technical Control	Link	[10, 31]
NIH/3T3	10X Chromium v3	3T3	Cell Line	Link	[10]
HEK293T	10X Chromium v3	HEK	Cell Line	Link	[10]
PBMC 33k	10X Chromium v1	PBMC	Heterogeneous	Link	[20]
PBMC r1	inDrops	PBMC	Heterogeneous	GSE132044	[8]
Mouse Cortex	DroNC-seq	Cortex	Heterogeneous	GSE132044	[8]
Mouse Lung	Drop-seq	Lung	Heterogeneous	GSE124872	[25]
Zheng Mix 8eq	10X Chromium v1	PBMC	Heterogeneous	Link	[34, 54]
Haber 2017	10X Chromium v1	Mouse Small Intestine Epithelium	Heterogeneous	GSE92332	[55]

Data availability

All the data sets used in this study, including the simulated counts data sets, are publicly available at https://github.com/Amartya101/PiccoloPaperData

Code availability

All the code used to perform the analysis and generate the figures in this paper are publicly available at https://github.com/Amartya101/PiccoloPaperData

Availability of software

The development version of `Piccolo` is currently available as an R package through GitHub at https://github.com/Amartya101/Piccolo, and is licensed under the GNU GPLv3 license. In order to facilitate easier adoption with existing tools, the feature selection and normalization methods implemented in `Piccolo` can be used with Seurat (version 4) [18], the instructions are provided at https://github.com/Amartya101/Piccolo-With-Seurat

Acknowledgments

We thank the members of the Khiabanian and Herranz labs for their encouragement and insightful feedback. Special thanks are due to Mona Arabzadeh and Vaidhyanathan Mahaganapathy for carefully reading the manuscript and suggesting improvements. This work was supported by grants from the National Institutes of Health (R01CA233662 and P30CA072720), the V Foundation (T2019-01), and the New Jersey Commission on Cancer Research (COCR23PRG006).

References

1. Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, September 2015.

2. Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L De Jager, Orit Rozenblatt-Rosen, Andrew A Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), April 2017.

3. Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, June 2013.

4. Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, June 2014.

5. Eduardo Torre, Hannah Dueck, Sydney Shaffer, Janko Gospocic, Rohit Gupte, Roberto Bonasio, Junhyong Kim, John Murray, and Arjun Raj. Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. *Cell Syst.*, 6(2):171–179.e5, February 2018.

6. Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375, May 2014.

7. Karsten Bach, Sara Pensa, Marta Grzelak, James Hadfield, David J Adams, John C Marioni, and Walid T Khaled. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.*, 8(1):2128, December 2017.

8. Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, John Y H Kwon, Boaz Barak, William Ge, Amanda J Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z Levin. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.*, 38(6):737–746, June 2020.

9. Tallulah S Andrews, Vladimir Yu Kiselev, Davis McCarthy, and Martin Hemberg. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.*, 16(1):1–9, January 2021.

10. Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, 38(2):147–150, February 2020.

11. Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, 53(6):770–777, Jun 2021.

12. Aaron T. L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Res.*, 5:2122, 2016.

13. Robert A. Amezquita, Aaron T. L. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L. Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C. Hicks. Orchestrating single-cell analysis with bioconductor. *Nature Methods*, 17(2):137–145, Feb 2020.

14. Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, 15(6):e8746, June 2019.

15. Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502, 2015.

16. Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411–420, 2018.

17. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177:1888–1902, 2019.

18. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

19. F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, Feb 2018.

20. Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, 20(1):296, December 2019.

21. Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol.*, 22(1):258, September 2021.

22. Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell rna-seq data. *Nature Methods*, Apr 2023.

23. A. Colin Cameron and Pravin K. Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, 46(3):347–364, 1990.

24. A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge University Press, 2 edition, 2013.

25. Ilias Angelidis, Lukas M. Simon, Isis E. Fernandez, Maximilian Strunz, Christoph H. Mayr, Flavia R. Greiffo, George Tsitsiridis, Meshal Ansari, Elisabeth Graf, Tim-Matthias Strom, Monica Nagendran, Tushar Desai, Oliver Eickelberg, Matthias Mann, Fabian J. Theis, and Herbert B. Schiller. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, 10(1):963, Feb 2019.

26. Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, Dec 2014.

27. F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biology*, 20(1):295, Dec 2019.

28. Peter V. Kharchenko. The triumphs and limitations of computational methods for scrna-seq. *Nature Methods*, 18(7):723–732, Jul 2021.

29. Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scrna-seq. *Genome Biology*, 23(1):27, Jan 2022.

30. Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 12 1936.

31. Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

32. V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, Mar 2019.

33. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

34. Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.*, 7:1141, July 2018.

35. Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8(1):14049, January 2017.

36. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

37. Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome Biology*, 18(1):174, Sep 2017.

38. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

39. Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2022. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).

40. Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2021. R package version 1.14.2.

41. Michael Hahsler and Matthew Piekenbrock. *dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms*, 2022. R package version 1.1-11.

42. Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1):1–30, 2019.

43. Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

44. Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

45. Douglas Bates, Martin Maechler, and Mikael Jagan. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2022. R package version 1.5-1.

46. Karolis Koncevičius. *matrixTests: Fast Statistical Hypothesis Tests on Rows and Columns of Matrices*, 2021. R package version 0.1.9.1.

47. Yixuan Qiu and Jiali Mei. *RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems*, 2022. R package version 0.16-1.

48. L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

49. L.J.P. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

50. Jesse H. Krijthe. *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*, 2015. R package version 0.16.

51. Tomasz Konopka. *umap: Uniform Manifold Approximation and Projection*, 2022. R package version 0.2.9.0.

52. Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Antônio Pedro, Sciaini, Marco, Scherer, and Cédric. *viridis - Colorblind-Friendly Color Maps for R*, 2021. R package version 0.6.2.

53. Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J. Miragaia, Charlotte Labalette, Iain C. Macaulay, Ana Cvejic, and Sarah A. Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature Methods*, 14(4):381–387, Apr 2017.

54. Angelo Duò and Charlotte Soneson. *DuoClustering2018: Data, Clustering Results and Visualization Functions From Duò et al (2018)*, 2022. R package version 1.14.0.

55. Adam L Haber, Moshe Biton, Noga Rogel, Rebecca H Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, Toni M Delorey, Michael R Howitt, Yarden Katz, Itay Tirosh, Semir Beyaz, Danielle Dionne, Mei Zhang, Raktima Raychowdhury, Wendy S Garrett, Orit Rozenblatt-Rosen, Hai Ning Shi, Omer Yilmaz, Ramnik J Xavier, and Aviv Regev. A single-cell survey of the small intestinal epithelium. *Nature*, 551(7680):333–339, November 2017.

56. G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 2023/02/17/ 1964. Full publication date: 1964.

Appendix

Regression-based approach to estimate QP dispersion coefficients

The regression-based test proposed by Cameron and Trivedi [23, 24] relies on the idea that under the null hypothesis - counts, x, are Poisson distributed - the expected value of $(x - E[x])^2 - x$ is zero, while under the alternative hypothesis the expected value would be a scalar multiple of a function of $E[x]$. Since $E[x]$ is unknown, we replace it by the estimate under the null hypothesis and estimate the scalar multiple using least squares regression.

The choice of the QP mean-variance relation corresponds to the alternative hypothesis in which the variance ($Var[x]$) is a linear function of the mean ($E[x]$),

$$Var[x] = \alpha_{QP}.E[x] \tag{17}$$

Given this specification of the alternative hypothesis, the α_{QP} are estimated by auxiliary ordinary least-squares regression using,

$$\alpha_{QP} = 1 + E[\frac{(x - E[x])^2 - x}{E[x]}]$$

with $E[x] = \mu =$ Poisson estimate for the expected value of the counts.

The standard estimate of cell-specific size factors assumes that the counts are Poisson distributed

There is an intimate link between the estimates of size factors given by equation. (6) and the estimates for expected means ($\hat{\mu}_{gc}$) under the assumption that the counts are Poisson distributed. Assuming that each gene g contributes a proportion p_g of the total count N_c in cell c, the counts X_{gc} are modeled as,

$$X_{gc} \sim Poisson(\mu_{gc}) \tag{18}$$

where $\mu_{gc} = p_g N_c$. As shown by Townes et al. [27] and Lause et al. [21], the maximum likelihood estimates for p_g and N_c under the Poisson model are given by,

$$\hat{N}_c = \sum_g X_{gc} \tag{19}$$

$$\hat{p}_g = \frac{\sum_c X_{gc}}{\sum_c \hat{N}_c} \tag{20}$$

Based on these estimates, the maximum likelihood estimate for μ_{gc} is given by,

$$\hat{\mu}_{gc} = \frac{\sum_c X_{gc}.\sum_g X_{gc}}{\sum_c \sum_g X_{gc}} \tag{21}$$

If we divide both the numerator and denominator in equation. (21) by the total number of cells (C), we get,

$$\hat{\mu}_{gc} = \frac{(\sum_c X_{gc}/C).\sum_g X_{gc}}{(\sum_c \sum_g X_{gc})/C}$$

Since $\sum_c X_{gc}/C = \mu_g =$ mean of the observed counts of gene g,

$$\hat{\mu}_{gc} = \frac{\mu_g.\sum_g X_{gc}}{(\sum_c \sum_g X_{gc})/C}$$

Using equation. (6), we can rewrite the above equation more simply as,

$$\hat{\mu}_{gc} = \mu_g SF_c \tag{22}$$

From this we can conclude that the simple estimates of the size factors given by equation. (6) should be more appropriately viewed as estimates under the approximation that the counts are Poisson distributed.

Estimates for mean and variance under variance stabilization transformation

In order to compute z-scores for our data, we first need to apply a variance stabilization transformation to the observed counts (X_{gc}) to bring their distribution closer to the normal distribution. The variance stabilization transformation can be performed using monotonic non-linear functions, $g(X)$, such that the transformed counts are given by,

$$Y = g(X)$$

To compute the residuals, we need estimates for the means and variances of Y based on estimates for means and variances of X. We can arrive at approximations for both using a Taylor expansion around $X = \mu$,

$$Y \approx g(\mu) + g'(\mu)(X - \mu) + \frac{1}{2!}g''(\mu)(X - \mu)^2 + ...$$

where $g'(\mu)$ and $g''(\mu)$ are the first and second order derivatives of $g(X)$ evaluated at $X = \mu$. Considering the expansion up till the 1st order,

$$Y \approx g'(\mu)X + g(\mu) - g'(\mu)\mu$$

The expected value of Y can be approximated to,

$$E[Y] \approx g'(\mu)\mu + g(\mu) - g'(\mu)\mu = g(\mu) \tag{23}$$

In addition, since

$$Y - g(\mu) \approx g'(\mu)(X - \mu)$$

after squaring and taking expectation we get the following approximation for the variance of Y,

$$Var[Y] = Var[g(X)] \approx (g'(\mu))^2 Var[X] \tag{24}$$

For $g(X) = log(X + 1)$, the first order approximations of the mean and variance are then given by,

$$E[log(X + 1)] \approx log(\mu + 1) \tag{25}$$

$$Var[log(X + 1)] \approx \frac{1}{(\mu + 1)^2}\sigma^2 \tag{26}$$

Limitation of residuals without variance stabilization

We discuss the limitation of residuals computed for raw untransformed counts by considering the analytic Pearson residuals approach proposed by Lause et al. [21]. They proposed that the expected means ($\hat{\mu}_{gc}$) can be approximated with the estimates given by equation. (21). Further, they argued that the over-dispersion coefficient ($\hat{\alpha}_g$) can be approximated with a fixed value of $\hat{\alpha}_g = 0.01$ corresponding to the typical over-dispersion observed in technical control data sets. Their rationale for these choices is that the null model should correspond to the measurement process so that the residuals provide estimates for deviations compared to expectations under the measurement model. While conceptually sound, this approach compromises variance stabilization, especially for genes that

are robustly expressed in only a subset of cells while showing negligible expression in the rest of the cells (such genes would be considered as *markers* of the specific cell sub-populations in which they are expressed).

We can illustrate the nature of the problem with the help of a simplified example. Suppose we have a data set consisting of 10000 cells with a gene (gene A) that is only expressed in 100 cells with identical counts of 100 in each of those cells; the rest of the cells have 0 counts. For simplicity, if we assume that all the cells have the same sequencing depth ($SF_c = 1 \, \forall \, c$) then,

$$\hat{\mu}_{gc} = \mu_g$$

Based on this, equation. (8) simplifies to,

$$r_{gc} = \frac{X_{gc} - \mu_g}{\sqrt{\mu_g + \hat{\alpha}_g \mu_g^2}}$$

Given the distribution of counts for gene A stated above, $\mu_A = 1$, and since we assume $\hat{\alpha}_g = 0.01$, the residuals of gene A for the cells with counts of 100 are approximately 98.509, while the residuals for cells with 0 counts are -0.995. The overall variance of the residuals for gene A is approximately 98.03, thus exhibiting significant deviation from the null expectation of 1. The Pearson residuals method proposed by Hafemeister et al. [20] that allows for per gene estimates for $\hat{\alpha}_g$ addresses this issue only to a limited extent.

Variance stabilization transformations implemented in Piccolo

Apart from the *log* transformation as the variance stabilization transformation discussed in the main text and implemented as the default option in `Piccolo`, we also offer the option to apply two other transformations which are described below:

Sqrt

For the *sqrt* transformation, $g(X) = \sqrt{X}$. The first order approximations of the mean and variance under this transformation are,

$$E[\sqrt{X}] \approx \sqrt{\mu}$$

$$Var[\sqrt{X}] \approx \frac{\sigma^2}{1/(4 * \mu)}$$

After accounting for sampling depth differences ($\hat{\mu}_{gc} = SF_c \tilde{\mu}_g$ and $\hat{\sigma}_{gc}^2 = SF_c^2 \tilde{\sigma}_g^2$) we have,

$$E[\sqrt{X_{gc}}] \approx \sqrt{\hat{\mu}_{gc}}$$

$$Var[\sqrt{X_{gc}}] \approx \frac{\hat{\sigma}_{gc}^2}{1/(4 * \hat{\mu}_{gc})}$$

Based on these first order approximations for means and variances under the $\sqrt{X}$ transformation, the residuals are given by,

$$Z'_{gc} = \frac{\sqrt{X_{gc}} - \sqrt{\hat{\mu}_{gc}}}{\hat{\sigma}_{gc}/(1/(2 * \sqrt{\hat{\mu}_{gc}}))}$$

Box-Cox Power Law Transform

The Box-Cox transform [56] belongs to the family of power law transformations. Power law transformations are applicable only for positive variables and are indexed by a parameter λ, such that for an arbitrary observation x, the transformed value is given by x^{λ}. The Box-Cox transformation is a modified power transformation defined as,

$$g(X = x) = x^{(\lambda)} = \frac{x^{\lambda} - 1}{\lambda}, \lambda \neq 0$$

which is a continuous function in λ for $x > 0$ ($g(X) = x^{(\lambda)} = log(x)$, when $\lambda = 0$). Box-Cox proposed that based on the observations x_1, x_2,..., x_n, the appropriate choice of λ corresponds to the value that maximizes,

$$l(\lambda) = -\frac{n}{2}\log[\frac{1}{n}\sum_{j=1}^{n}(x_j^{(\lambda)} - \overline{x^{(\lambda)}})^2] + (\lambda - 1)\sum_{j=1}^{n}\log x_j$$

where,

$$\overline{x^{(\lambda)}} = \frac{1}{n}\sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n}(\frac{x_j^{\lambda} - 1}{\lambda})$$

For the Box-Cox transformation, the first-order approximations of the mean and variance are,

$$E[g(X)] \approx \frac{(\mu + 1)^{\lambda} - 1}{\lambda}$$

$$Var[g(X)] \approx \sigma^2 * (\mu + 1)^{\lambda - 1}$$

where pseudo-counts of 1 have been added to ensure positive (non-zero) values. After accounting for sampling depth differences ($\hat{\mu}_{gc} = SF_c\tilde{\mu}_g$ and $\hat{\sigma}_{gc}^2 = SF_c^2\tilde{\sigma}_g^2$) we have,

$$E[g(X_{gc})] \approx \frac{(\hat{\mu}_{gc} + 1)^{\lambda} - 1}{\lambda}$$

$$Var[g(X_{gc})] \approx \hat{\sigma}_{gc}^2 * (\hat{\mu}_{gc} + 1)^{\lambda - 1}$$

Based on these first-order approximations for means and variances under the Box-Cox transformation, the residuals are given by,

$$Z'_{gc} = \frac{\frac{(X_{gc}+1)^{\lambda}-1}{\lambda} - \frac{(\hat{\mu}_{gc}+1)^{\lambda}-1}{\lambda}}{\hat{\sigma}_{gc} * \sqrt{(\hat{\mu}_{gc} + 1)^{\lambda - 1}}}$$

LogSF

We also provide the popular `logSF` normalization as an option in `Piccolo`. The normalized counts under this transformation are given by,

$$X'_{gc} = \log(\frac{X_{gc}}{SF_c} + 1)$$

where estimates of SF_c are obtained using stable genes (see eq. (7))

We compared these variance stabilization transformations within Piccolo by applying them to the 100 subsets created using the `Zheng Mix 8eq` data set. The results are shown in Fig. S13. The Box-Cox transform performs the best overall. However, the trade-off is that it is also the most computationally intensive. Nevertheless, we would still recommend users to consider applying it for small or medium sized data sets ($10^3 - 10^5$ cells).

Differential expression analysis in Piccolo

After applying our normalization to the observed counts, the distribution of the residuals are brought closer to the normal distribution due to the variance stabilization. This makes it possible to employ the two-sample Student's t-test with the null hypothesis that the means of the two samples are the same. Ideally, the test is applicable only if the variances of the two samples can be assumed to be equal. The variance stabilization during normalization ensures that under most circumstances this assumption holds true.

Typically, the preferred test for single-cell differential expression analyses is the Wilcoxon rank-sum test which is a non-parametric alternative to the two-sample t-test. A key difference between the two is that while the t-test actually tests for location shifts (differences in means) between the samples, the Wilcoxon rank-sum test can be sensitive to shifts in distribution other than a pure location shift. However, in the context of scRNA-seq data sets with the number of samples (cells) typically in the order of $10^4 - 10^6$, this can often lead to false positives, particularly for genes with small counts, since even minor differences (in terms of magnitude) in counts in $10^1 - 10^2$ cells between the groups can result in a significant p-value as assessed by the rank-sum test. Keeping this in mind, currently in `Piccolo` along with the Wilcoxon rank-sum test we also offer the t-test to perform differential expression analyses.

Obtaining corrected counts from the z-scores

While the clustering and the differential expression analyses are performed using the z-scores obtained from `Piccolo`, in some other applications and for the purpose of applying other tools we can also obtain estimates of both the *log*-transformed values as well as the corrected counts. We illustrate how this is done for `Piccolo` normalization (*log*-based variance stabilization).

Recall that,

$$\tilde{\mu}_g = \frac{1}{C}\sum_c \frac{X_{gc}}{SF_c}$$

$$\tilde{\sigma}_g^2 = \frac{1}{C-1}\sum_c (\frac{X_{gc}}{SF_c} - \tilde{\mu}_g)^2$$

are the respective estimates for the gene mean and variance after accounting for sampling depths. Based on these, the estimates for the *log*-transformed values can be obtained using eq. (16) as,

$$\log(\tilde{X}_{gc} + 1) = Z_{gc}^{'} * \frac{\tilde{\sigma}_g}{\tilde{\mu}_g + 1} + \log(\tilde{\mu}_g + 1)$$

Using these *log*-transformed values, we can then obtain estimates for the corrected counts,

$$\tilde{X}_{gc} = round(\exp(Z_{gc}^{'} * \frac{\tilde{\sigma}_g}{\tilde{\mu}_g + 1} + \log(\tilde{\mu}_g + 1)) - 1)$$

where the rounding ensures that the corrected counts have integer values.

Batch effect correction

Batch effects are technical confounders that typically stem from a wide array of non-biological sources such as differences in the reagents, the individuals handling and processing the samples, as well as the time of experiment. If not accounted for, batch effect can be misinterpreted as a biological signal and lead to erroneous conclusions. The z-score based normalization of `Piccolo` allows for a very simplistic approach to perform batch effect correction based on the fundamental assumption that all the biological conditions are processed in all the batches (Note: It is tempting to apply batch correction even when this assumption is not met, however the conclusions drawn from such analyses will be unreliable).

The basic idea of our batch effect correction method is to identify HVGs independently for each batch, and then seek the HVGs that match between the batches. This ensures that we only retain those genes for downstream analysis that are variable in every batch, and eliminate genes that exhibit variability of counts only within specific batches since the latter likely reflect batch-specific effects. In addition, we identify stable genes across all batches as well as stable genes individually for each batch. We retain only those stable genes in each batch that are also stable across the batches. This ensures that the counts are normalized using size factors estimated with stable genes in each batch that we know are less likely to exhibit batch-specific variation. We then perform the z-score normalization independently for each batch. PCA is applied on the composite z-scores matrix obtained by combining the z-scores matrices from all the batches.

We provide an illustration of the application of our simplistic batch effect correction on a mouse intestinal epithelium data set (`Haber 2017`) that consists of cells obtained in 10 batches [55] (see Fig. S14). Further, we used Splat [37] to simulate a data set using `NIH/3T3` that contained 5 groups of cells, with half the cells in one batch, and the other half simulated to belong to another batch (see Fig. S15). With the help of these examples, we are able to show that the simplistic batch correction approach works well when the assumption that all the conditions are present in every batch is reasonably satisfied.

Supplementary Figures

Fig. S1

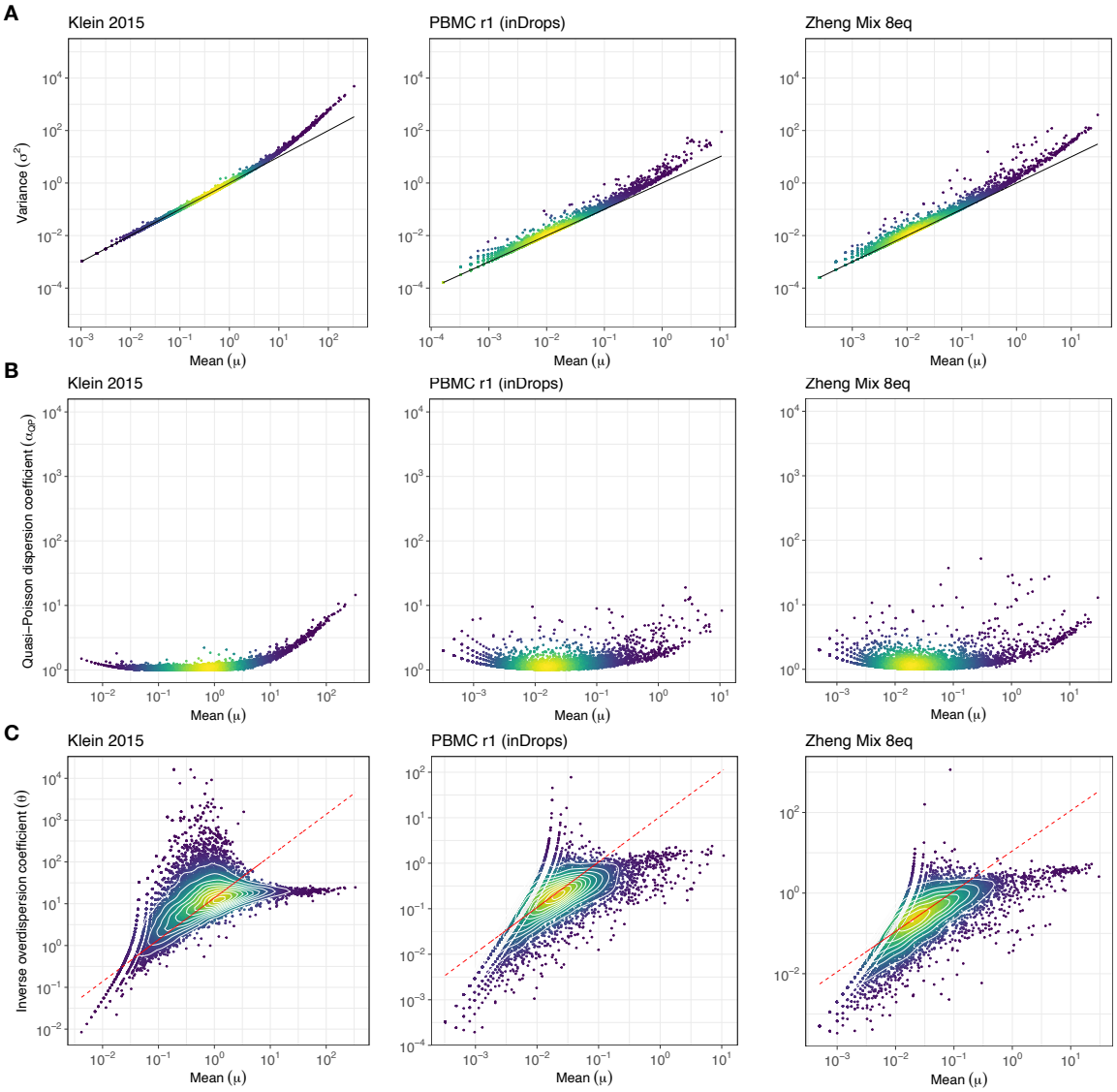

Fig. S1. Genes with low mean expression exhibit quasi-Poisson variance. In all the plots, each dot represents a gene and the color of the dots reflect the local point density, with brighter shades (yellow) indicating high density and darker shades (deep blue) indicating low density. **A.** Variance (σ^2) vs mean (μ) log-log scatter plots for the Klein 2015 technical control, PBMC r1, and Zheng Mix 8eq data sets (left to right). The solid black line corresponds to the Poisson model ($\sigma^2 = \mu$). For genes with low mean expression levels, the variance is adequately described by the Poisson model. **B.** α_{QP} vs μ log-log scatter plots for the Klein 2015, PBMC r1, and Zheng Mix 8eq data sets. **C.** θ vs μ log-log scatter plots for the Klein 2015, PBMC r1, and Zheng Mix 8eq data sets. The white curves correspond to the contours lines based on the local point densities. The dashed red line corresponds to estimates of θ obtained by fixing α_{QP} to its value for the gene with the highest point density estimate in the the θ vs μ log-log plot. The monotonically increasing relationship between θ and μ for small μ indicates quasi-Poisson variance at low mean expression levels.

Fig. S2

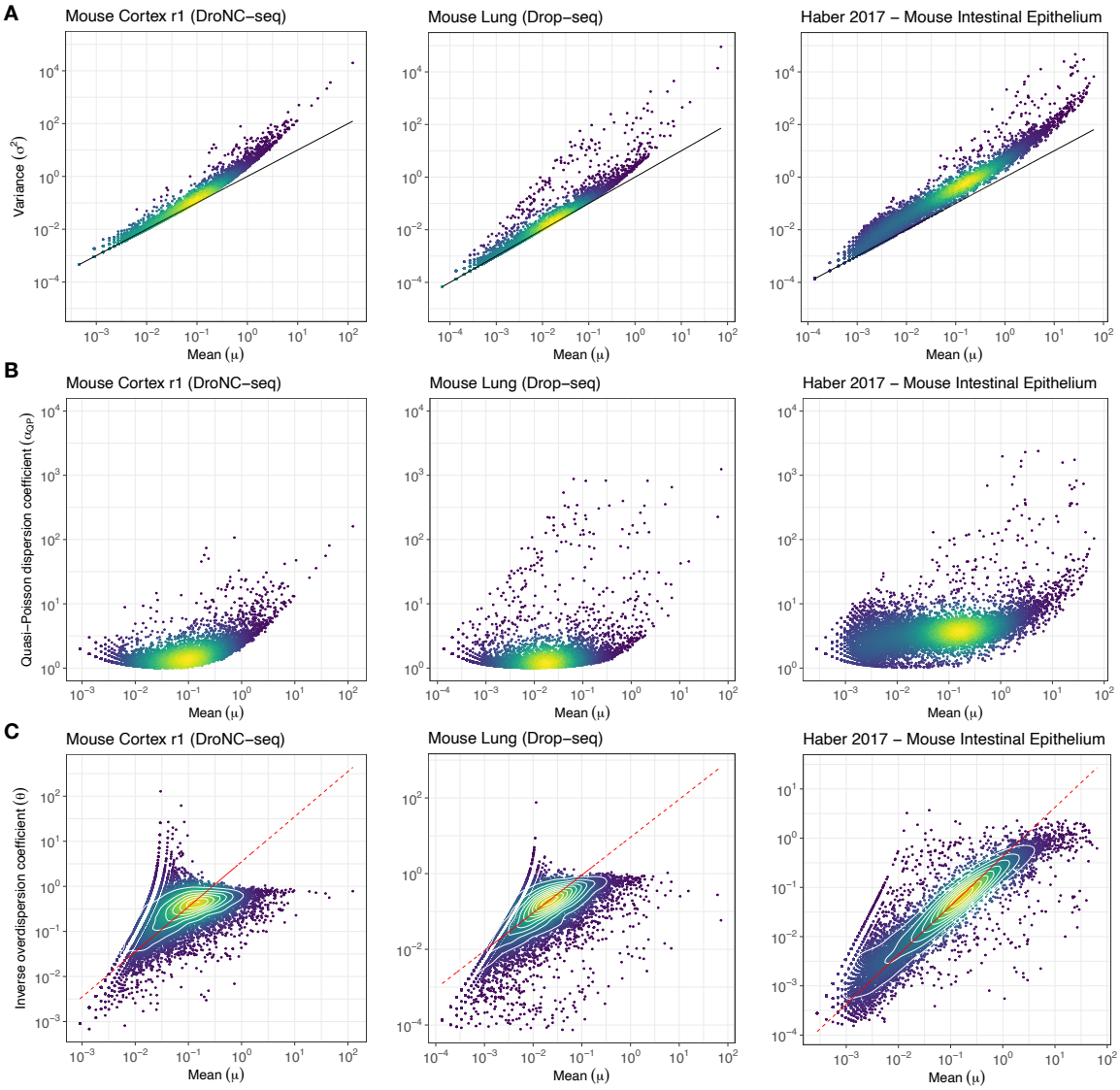

Fig. S2. Genes with low mean expression exhibit quasi-Poisson variance. In all the plots, each dot represents a gene and the color of the dots reflect the local point density, with brighter shades (yellow) indicating high density and darker shades (deep blue) indicating low density. **A.** Variance (σ^2) vs mean (μ) log-log scatter plots for the Mouse Cortex, Mouse Lung, and Haber 2017 data sets (left to right, in that order). The solid black line corresponds to the Poisson model ($\sigma^2 = \mu$). For genes with low mean expression levels, the variance is adequately described by the Poisson model. **B.** α_{QP} vs μ log-log scatter plots for the Mouse Cortex, Mouse Lung, and Haber 2017 data sets. **C.** θ vs μ log-log scatter plots for the Mouse Cortex, Mouse Lung, and Haber 2017 data sets. The white curves correspond to the contours lines based on the local point densities. The dashed red line corresponds to estimates of θ obtained by fixing α_{QP} to its value for the gene with the highest point density estimate in the the θ vs μ log-log plot. The monotonically increasing relationship between θ and μ for small μ indicates quasi-Poisson variance at low mean expression levels.

Fig. S3

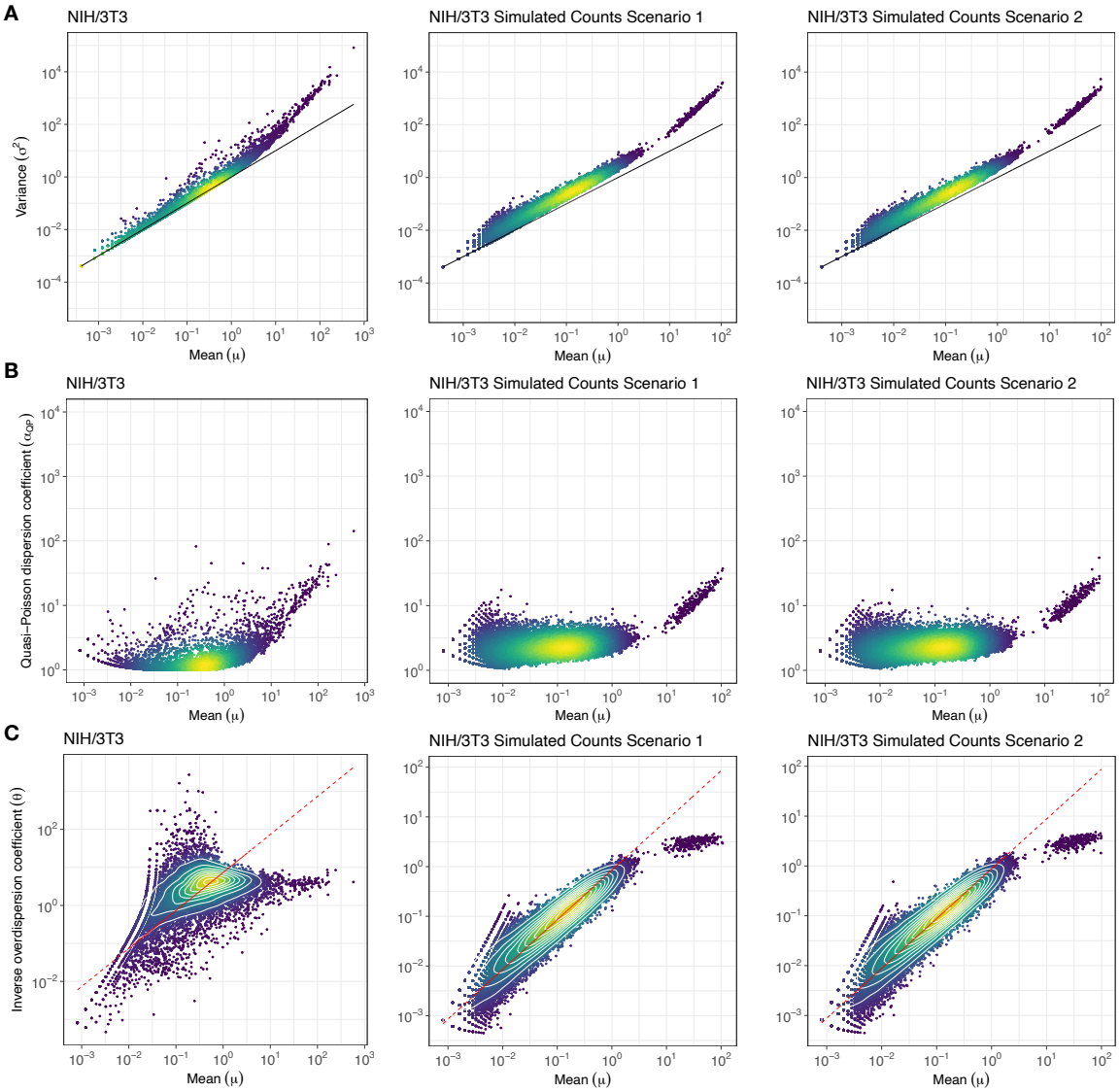

Fig. S3. Genes with low mean expression exhibit quasi-Poisson variance. In all the plots, each dot represents a gene and the color of the dots reflect the local point density, with brighter shades (yellow) indicating high density and darker shades (deep blue) indicating low density. **A.** Variance (σ^2) vs mean (μ) log-log scatter plots for the NIH/3T3, NIH/3T3 Simulated Counts Scenario 1, and NIH/3T3 Simulated Counts Scenario 2 data sets (left to right). The solid black line corresponds to the Poisson model ($\sigma^2 = \mu$). For genes with low mean expression levels, the variance is adequately described by the Poisson model. **B.** α_{QP} vs μ log-log scatter plots for the NIH/3T3, NIH/3T3 Simulated Counts Scenario 1, and NIH/3T3 Simulated Counts Scenario 2 data sets. **C.** θ vs μ log-log scatter plots for the NIH/3T3, NIH/3T3 Simulated Counts Scenario 1, and NIH/3T3 Simulated Counts Scenario 2 data sets. The white curves correspond to the contours lines based on the local point densities. The dashed red line corresponds to estimates of θ obtained by fixing α_{QP} to its value for the gene with the highest point density estimate in the the θ vs μ log-log plot. The Splat [37] simulated counts data sets exhibit pronounced quasi-Poisson variance (monotonically increasing relationship between θ and μ for small μ) compared to the original data set.

Fig. S4

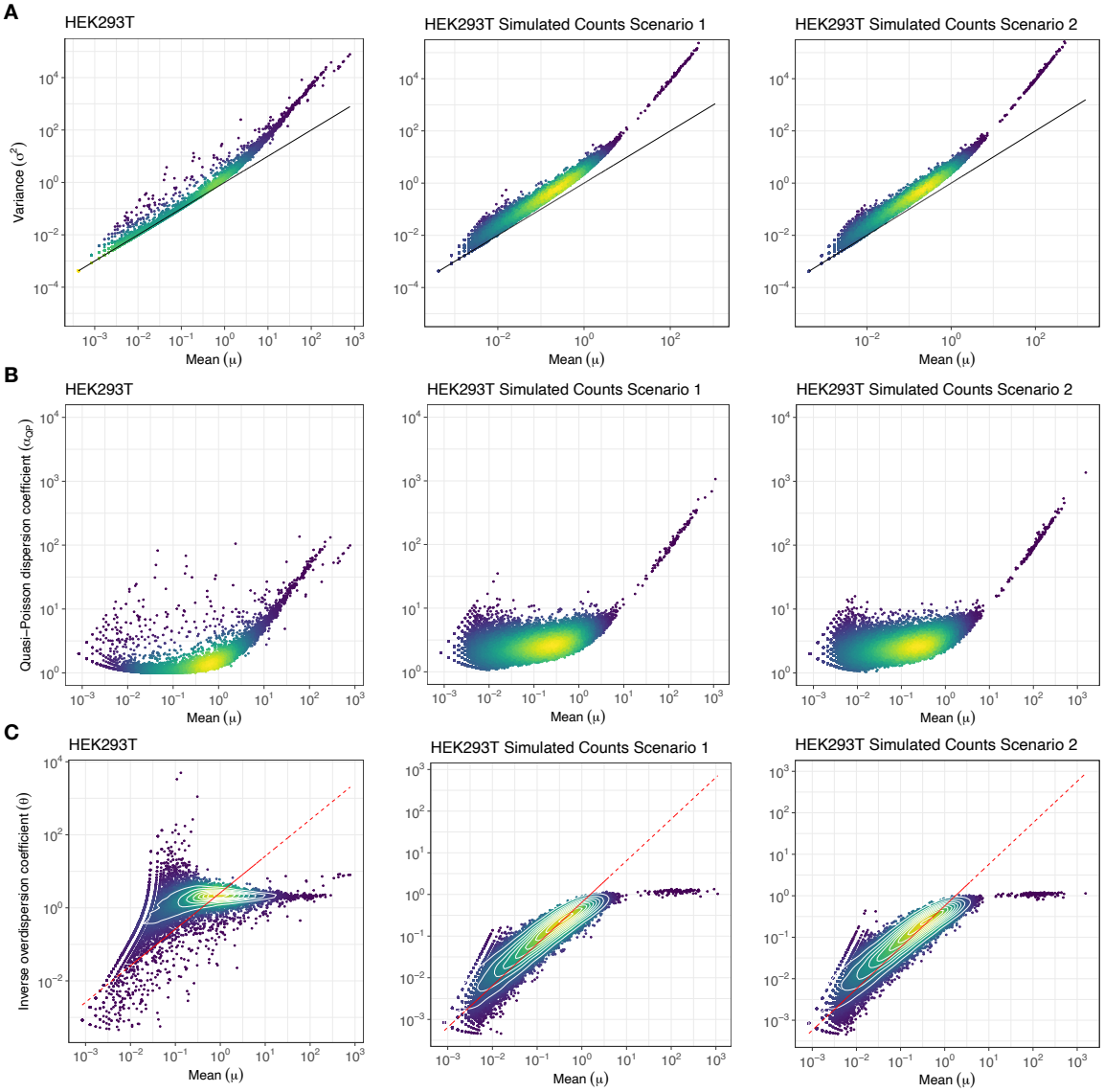

Fig. S4. Genes with low mean expression exhibit quasi-Poisson variance. In all the plots, each dot represents a gene and the color of the dots reflect the local point density, with brighter shades (yellow) indicating high density and darker shades (deep blue) indicating low density. **A.** Variance (σ^2) vs mean (μ) log-log scatter plots for the HEK293T, HEK293T Simulated Counts Scenario 1, and HEK293T Simulated Counts Scenario 2 data sets (left to right). The solid black line corresponds to the Poisson model ($\sigma^2 = \mu$). For genes with low mean expression levels, the variance is adequately described by the Poisson model. **B.** α_{QP} vs μ log-log scatter plots for the HEK293T, HEK293T Simulated Counts Scenario 1, and HEK293T Simulated Counts Scenario 2 data sets. **C.** θ vs μ log-log scatter plots for the HEK293T, HEK293T Simulated Counts Scenario 1, and HEK293T Simulated Counts Scenario 2 data sets. The white curves correspond to the contours lines based on the local point densities. The dashed red line corresponds to estimates of θ obtained by fixing α_{QP} to its value for the gene with the highest point density estimate in the the θ vs μ log-log plot. The Splat [37] simulated counts data sets exhibit pronounced quasi-Poisson variance (monotonically increasing relationship between θ and μ for small μ) compared to the original data set.

Fig. S5

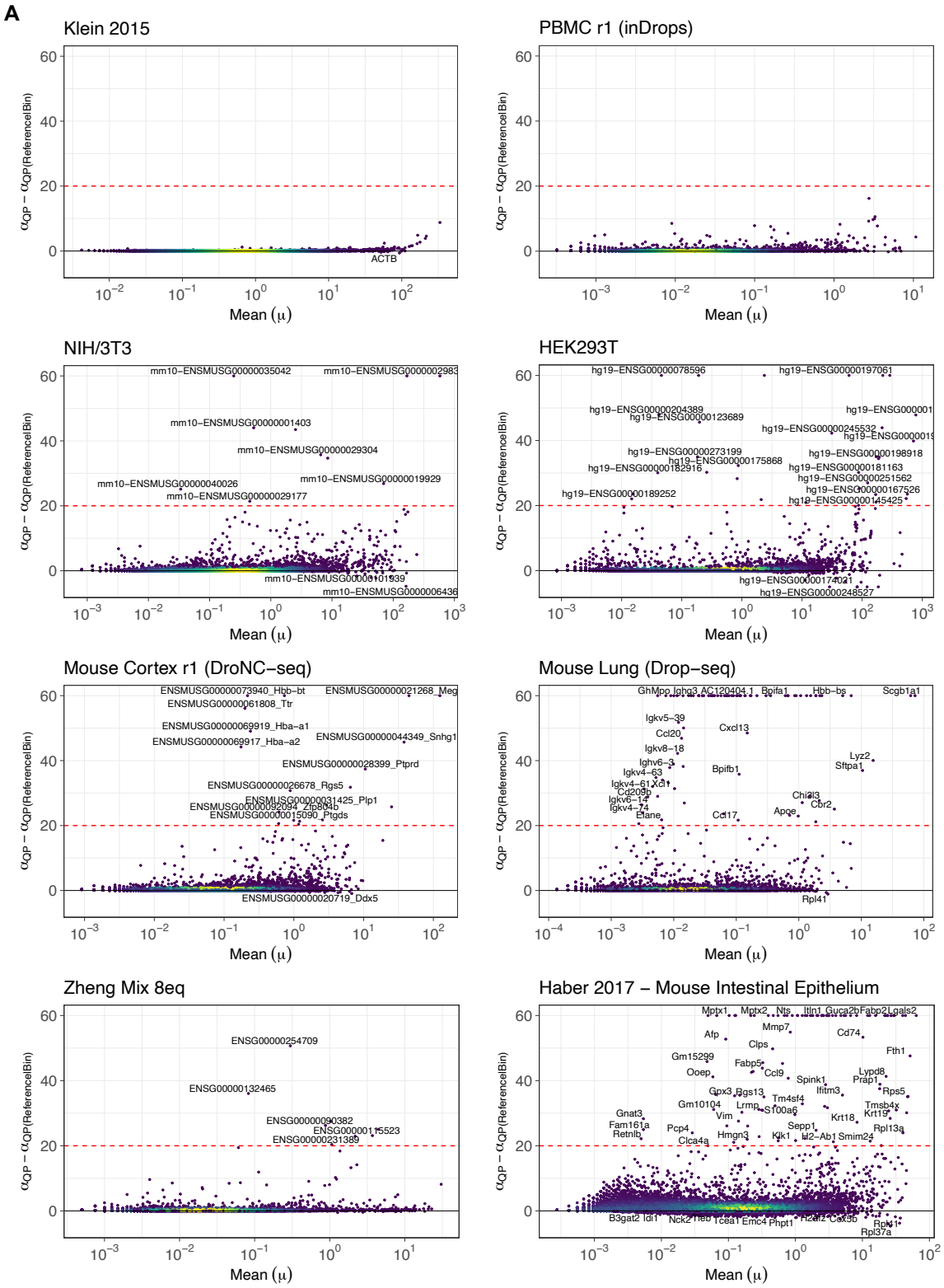
A
Klein 2015
PBMC r1 (inDrops)
NIH/3T3
HEK293T
Mouse Cortex r1 (DroNC−seq)
Mouse Lung (Drop−seq)
Zheng Mix 8eq
Haber 2017 − Mouse Intestinal Epithelium
αQP − αQP(ReferenceBin)
Mean (μ)
ACTB
mm10−ENSMUSG00000035042
mm10−ENSMUSG00000001403
mm10−ENSMUSG00000029304
mm10−ENSMUSG00000040026
mm10−ENSMUSG00000019929
mm10−ENSMUSG00000029177
hg19−ENSG00000078596
hg19−ENSG00000197061
hg19−ENSG00000204389
hg19−ENSG00000123689
hg19−ENSG00000245532
hg19−ENSG00000273199
hg19−ENSG00000175868
hg19−ENSG00000198918
hg19−ENSG00000182916
hg19−ENSG00000181163
hg19−ENSG00000251562
hg19−ENSG00000167526
hg19−ENSG00000189252
hg19−ENSG00000145425
hg19−ENSG00000174021
hg19−ENSG00000248527
ENSMUSG00000073940_Hbb−bt
ENSMUSG00000061808_Ttr
ENSMUSG00000069919_Hba−a1
ENSMUSG00000069917_Hba−a2
ENSMUSG00000044349_Snhg1
ENSMUSG00000028399_Ptprd
ENSMUSG00000026678_Rgs5
ENSMUSG00000031425_Plp1
ENSMUSG00000015090_Ptgds
ENSMUSG00000020719_Ddx5
GhMpp Igfbp3 AC120404.1 Bpifa1 Hbb−bs Scgb1a1
Igkv5−39
Ccl20
Cxcl13
Igkv8−18
Ighv6−3
Igkv4−63
Bpifb1
Lyz2
Sftpa1
Cd209b
Igkv6−14
Igkv4−74
Etfane
Cd17
Apoe
Cbr2
Rpl41
ENSG00000254709
ENSG00000132465
ENSG00000090382
ENSG00000115523
ENSG00000231389
Mptx1 Mptx2 Nts Itln1 Guca2b Fabp2 Lgals2
Mmp7
Afp
Cd74
Clps
Fth1
Gm15299
Fabp5
Ooep
Ccl9
Lypd8
Spink1
Prap1
Gpx3 Bgs13
Ifitm3
Rps5
Tm4sf4
Gm10104
Lrmp
S100a6
Tmsb4x
Krt19
Gnat3
Vim
Krt18
Fam161a
Sepp1
Rpl13a
Retnlb
Pcp4
Hmgn3
Klk1
Il2−Ab1
Smim24
Clca4a
B3gat2 Idi1 Nck2 Heb Tcea1 Emc4 Phpt1 H2afz Cox5b
Rpl41
Rpl37a

Fig. S5. Feature selection using bin-based approach does not lead to a preferential selection of genes based on their mean expression level. In all the panels, each dot is a gene and is colored based on the local density of the dots, with dots in higher density regions shown in brighter shades (yellow) and lower density regions shown in darker shades (dark blue). **A.** $\alpha_{QP} - \alpha_{QP(Reference|Bin)}$ vs mean (μ) linear-log scatter plot for genes that exhibited over-dispersion with respect to the Poisson ($\sigma^2 > \mu$) for 8 data sets (the panels are titled with the names of the respective data sets). The dashed red horizontal line illustrates the threshold to shortlist HVGs. Genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} > 20$ in this case are shortlisted as HVGs. On the other hand, genes with $\alpha_{QP} - \alpha_{QP(Reference|Bin)} < 0$ are shortlisted as stable genes. For all data sets (except the technical control that shows very little variability), it is clear that the HVGs are shortlisted across the different expression levels and there is no preferential selection based on the mean expression level.

Fig. S6

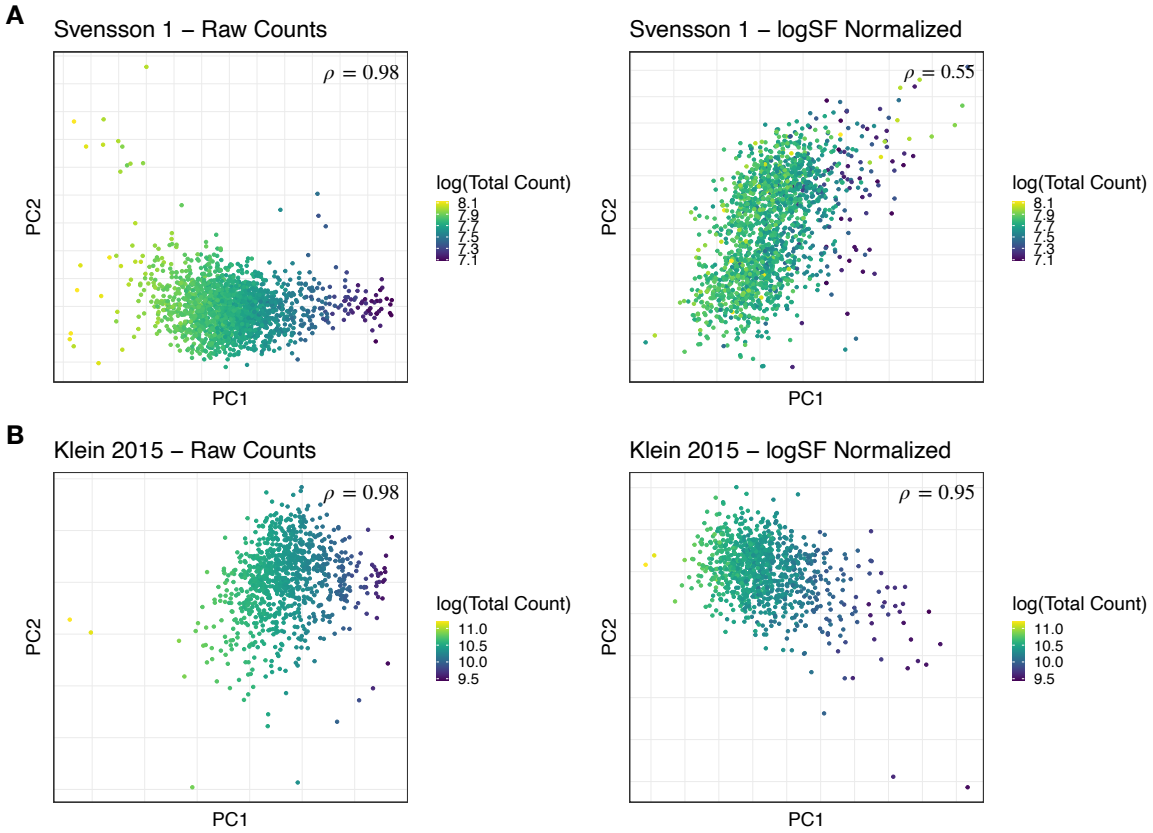

Fig. S6. LogSF normalization does not significantly reduce the impact of sampling depth (total count) differences **A.** 2D scatter plots based on the first 2 PCs of the Svensson 1 technical control data set. Each dot is a cell and is colored according to the total counts; brighter shades (yellow) correspond to larger total counts and darker shades (deep blue) correspond to smaller total counts. The left panel shows the plot for the raw counts, while the right panel shows the plot for the logSF transformed counts. The coefficient (ρ) in the top-right corner of the panels shows the canonical correlation coefficient between the total counts per cell and the top 5 PCs. Smaller values of ρ indicate lesser correlation between the total counts and the top 5 PCs, thus indicating that sampling depth differences have been reduced more effectively by the normalization. LogSF normalization does not reduce the impact of sampling depth on the overall variation as evaluated through PCA since the ρ has not been reduced significantly. **B.** 2D scatter plots based on the first 2 PCs of the Klein 2015 technical control data set. The left panel shows the plot for the raw counts, while the right panel shows the plot for the logSF transformed counts. The coefficient (ρ) in the top-right corner of the panels shows the canonical correlation coefficient between the total counts per cell and the top 5 PCs, and clearly indicates that even after logSF normalization the sampling depth differences are still contributing significantly to the overall variance in the data set.

Fig. S7

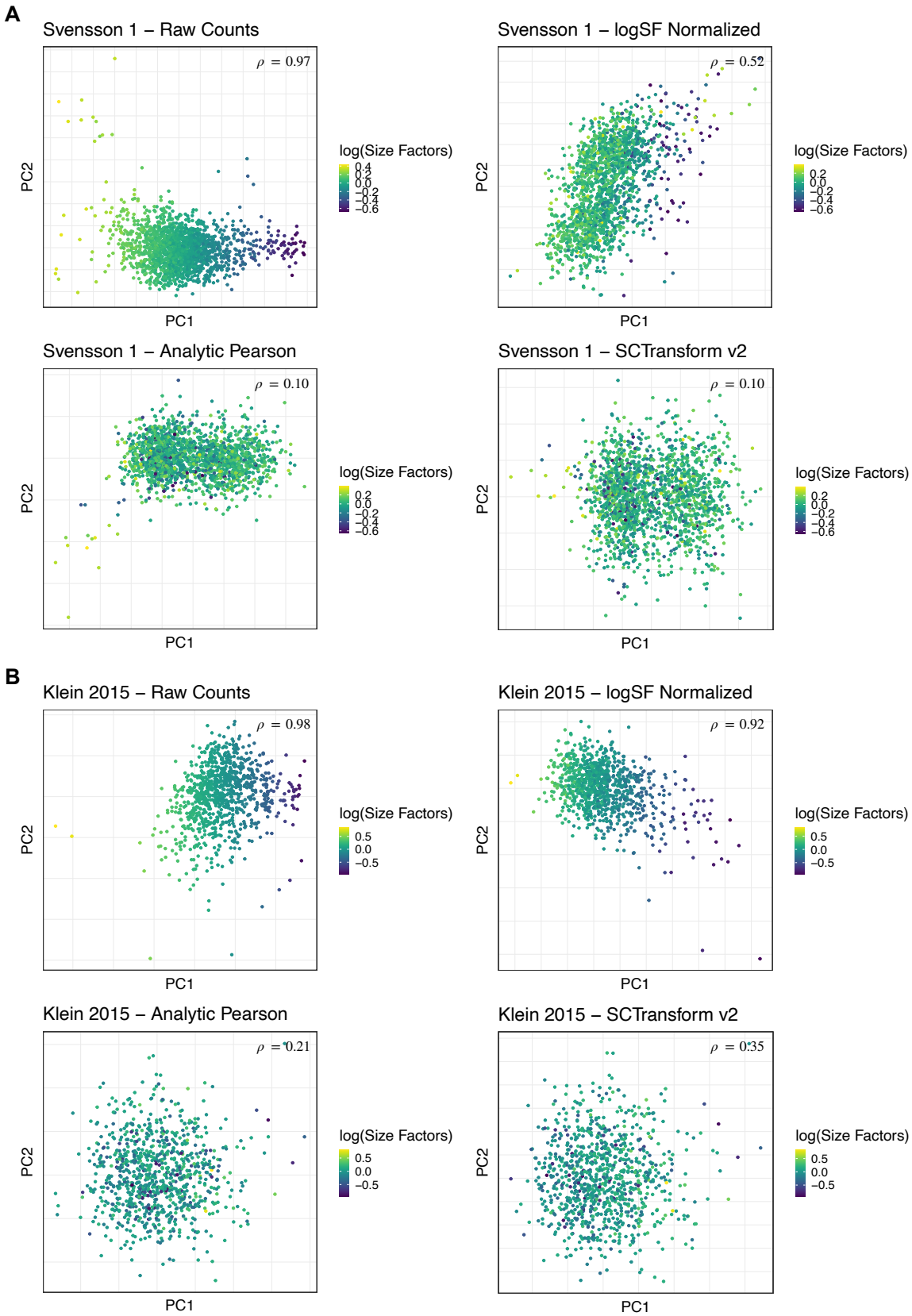
A
Svensson 1 – Raw Counts
ρ = 0.97
Svensson 1 – logSF Normalized
ρ = 0.52
Svensson 1 – Analytic Pearson
ρ = 0.10
Svensson 1 – SCTransform v2
ρ = 0.10
B
Klein 2015 – Raw Counts
ρ = 0.98
Klein 2015 – logSF Normalized
ρ = 0.92
Klein 2015 – Analytic Pearson
ρ = 0.21
Klein 2015 – SCTransform v2
ρ = 0.35
PC1
PC2
log(Size Factors)

Fig. S7. Residuals-based approaches reduce the impact of sampling depth differences more effectively than logSF normalization. A. 2D scatter plots based on the first 2 PCs of the Svensson 1 technical control data set. Each dot is a cell and is colored according to the size factors; brighter shades (yellow) correspond to larger size factors and darker shades (deep blue) correspond to smaller size factors. The top-left panel shows the plot for the raw counts, while the top-right panel shows the plot for the logSF transformed counts, the bottom 2 panels show the plots for the residuals obtained using Analytic Pearson method and SCTransform v2 respectively. The coefficient (ρ) in the top-right corner of the panels shows the canonical correlation coefficient between the size factors and the top 5 PCs. Smaller values of ρ indicate that the impact of the sampling depth differences has been reduced more effectively by the normalization. The values of ρ for the residuals-based methods are significantly smaller than the value obtained using logSF normalization. **B.** Same as in **A.** except this is for the Klein 2015 technical control data set. Once again, the residuals-based methods lead to much lesser correlation between the top 5 principal components and the size factors.

Fig. S8

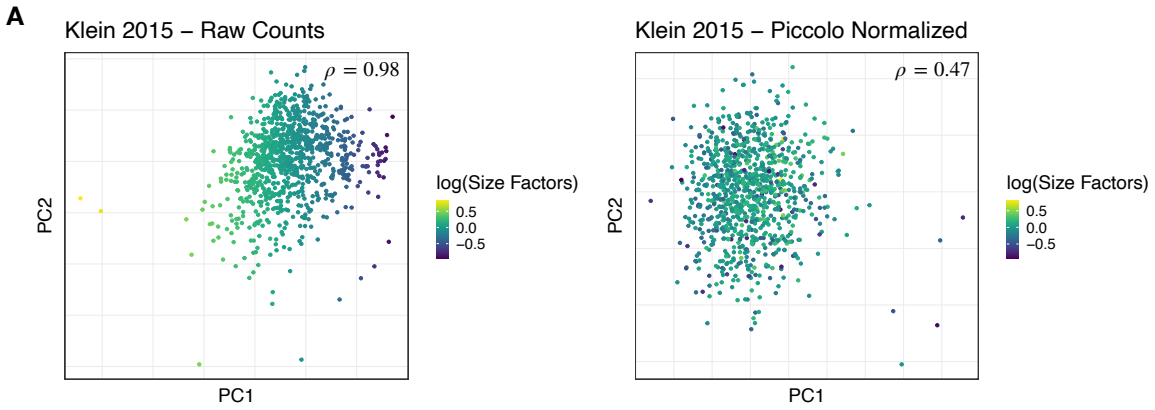

Fig. S8. Piccolo normalization reduces the impact of sampling depth differences. A. 2D scatter plots based on the first 2 PCs of the Klein 2015 technical control data set. Each dot is a cell and is colored according to the size factors; brighter shades (yellow) correspond to larger size factors and darker shades (deep blue) correspond to smaller size factors. The left panel shows the plot for the raw counts, while the right panel shows the plot for the residuals obtained using Piccolo normalization. The coefficient (ρ) in the top-right corner of the panels shows the canonical correlation coefficient between the size factors and the top 5 PCs. The value of ρ for the residuals obtained with Piccolo is significantly smaller than the value for the raw counts.

Fig. S9

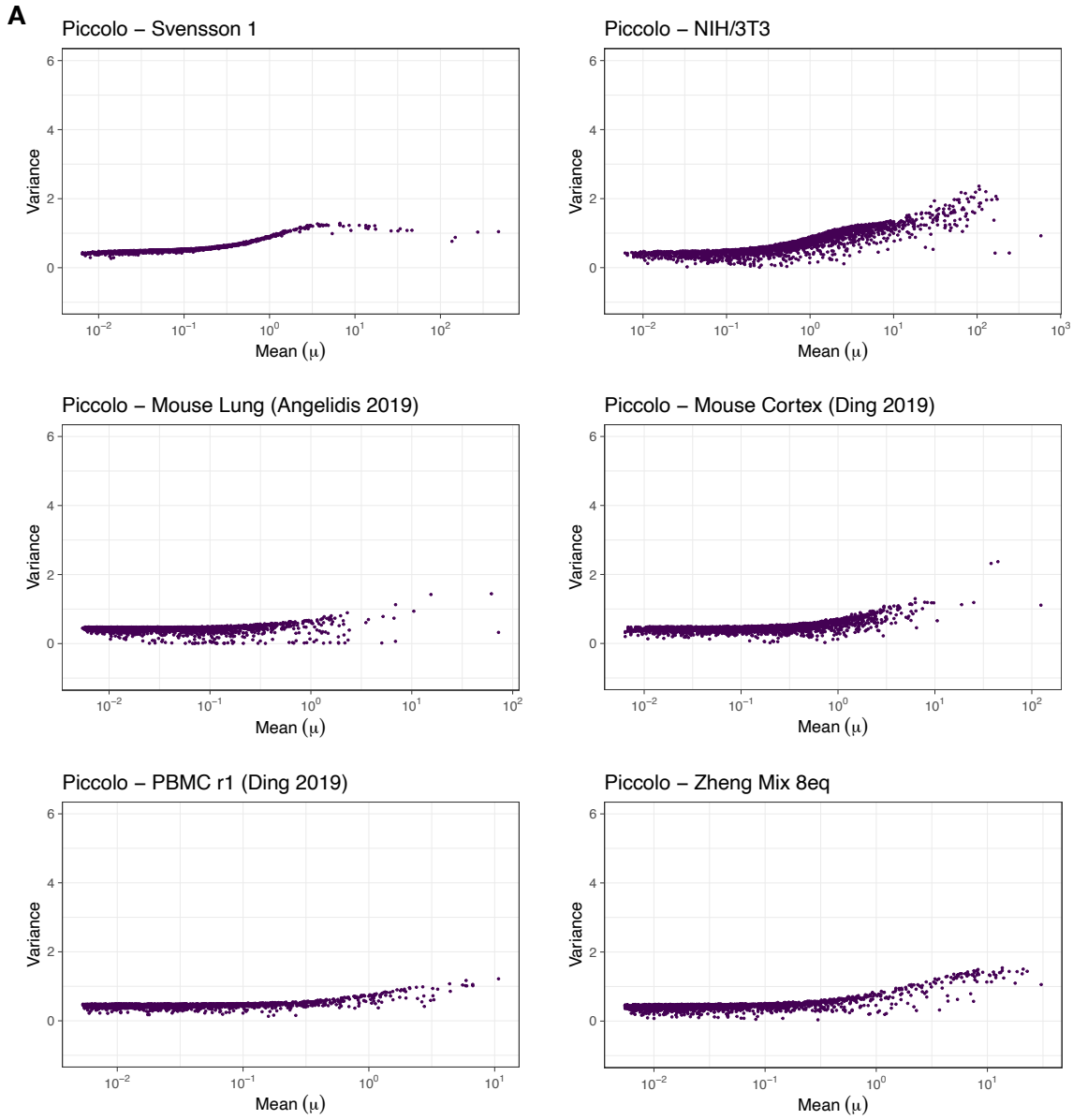

Fig. S9. Piccolo normalization ensures variance stabilization. A. Residual variance vs mean (μ) linear-log scatter plots for the top 3000 HVGs of 6 data sets after applying Piccolo normalization. The z-scores obtained using Piccolo exhibit variances that do not deviate significantly away from 1.

Fig. S10

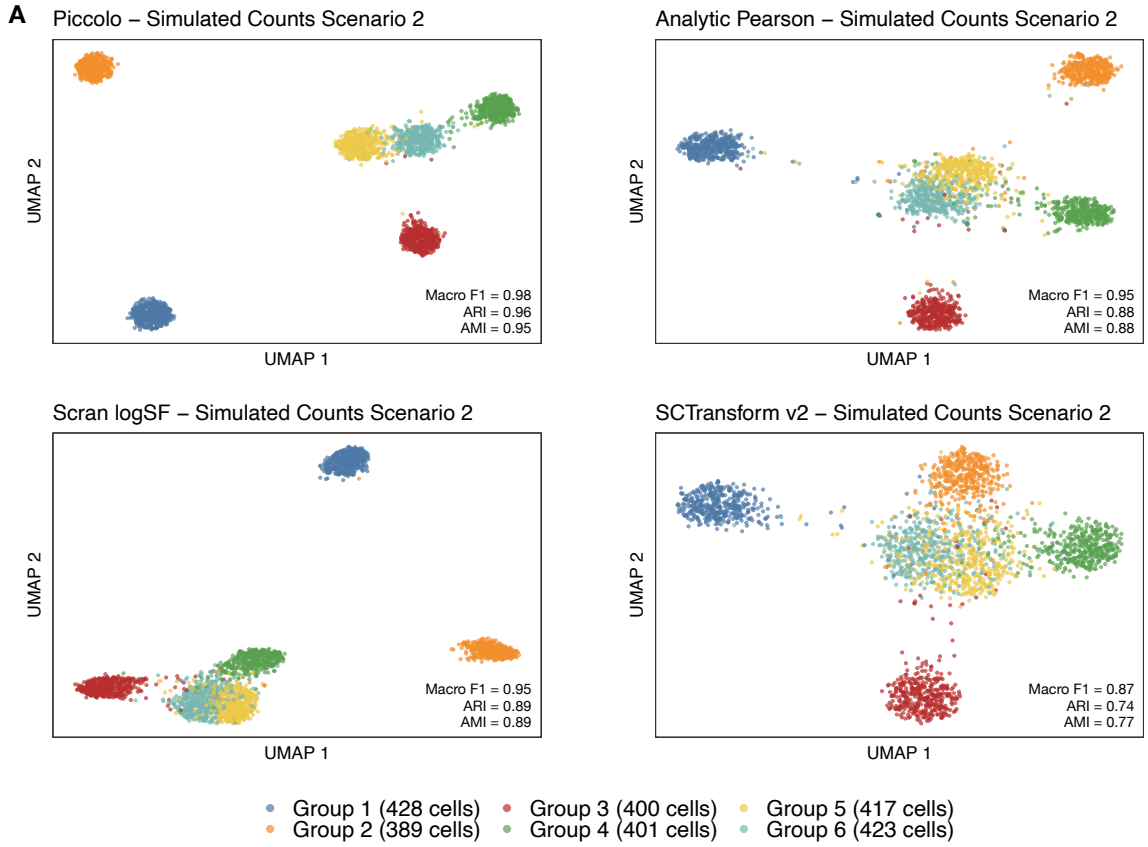

Fig. S10. Piccolo enables identification of cell clusters, including clusters of cells with few differentially expressed genes. A. 2D UMAP plots after using respective normalization methods for the top 3000 HVGs for the Simulated Counts Scenario 2 data set - 6 groups with approximately same number of cells per group but with different number of DE genes per group. The simulated counts were generated by Splat [37] using the NIH/3T3 data set. Dots represent cells and are colored using the known cell-type labels (legend at the bottom of the panel; actual numbers of cells belonging to each group are specified in parentheses). Clustering metrics based on comparisons between predicted cell labels and known cell labels are listed in the bottom-right corner in each panel. With Piccolo, cells belonging to the 6 groups cluster distinctly and can be distinguished better compared to other normalization methods.

Fig. S11

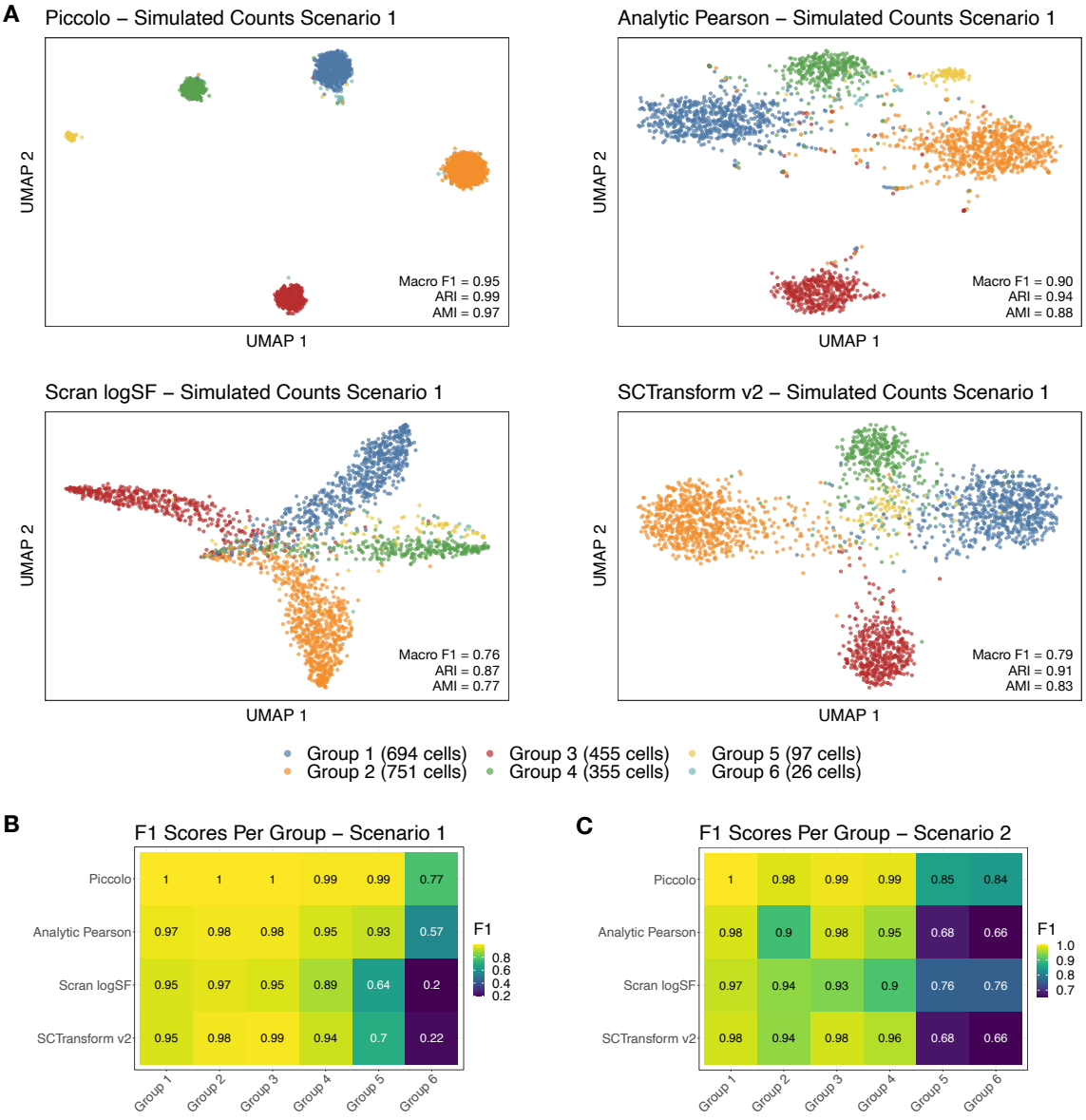

Fig. S11. Piccolo enables identification of clusters even for groups containing small numbers of cells. A. 2D UMAP plots after using respective normalization methods for the top 3000 HVGs for the Simulated Counts Scenario 1 data set. The simulated counts were generated by Splat [37] using the HEK293T data set. Dots represent cells and are colored using the known cell-type labels (legend at the bottom of the panel; actual numbers of cells belonging to each group are specified in parentheses). Clustering metrics based on comparisons between predicted cell labels and known cell labels are listed in the bottom-right corner in each panel. Cells belonging to the 6 groups can be distinguished using Piccolo. **B.** Heatmap showing the F1 scores per group for each of the normalization methods applied to Simulated Counts Scenario 1. The tiles of the heatmap are colored according to the F1 values with larger F1 scores corresponding to brighter shades (yellow), and lower F1 scores corresponding to darker shades (deep blue). Piccolo has the highest F1 scores for all 6 groups. **C.** Heatmap of F1 scores per group for the Simulated Counts Scenario 2 data set (see Fig. S12). Piccolo returns the highest F1 scores for all 6 groups.

Fig. S12

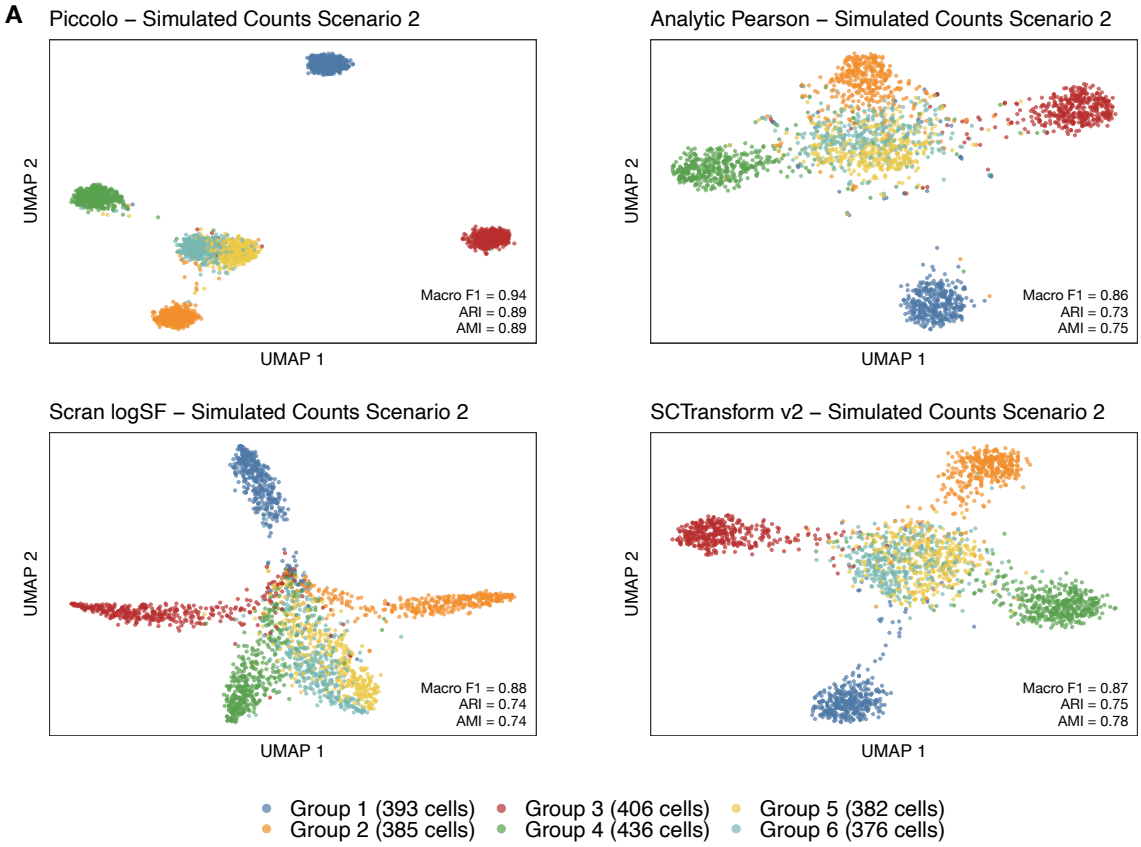

Fig. S12. Piccolo enables identification of cell clusters, including clusters of cells with few differentially expressed genes. A. 2D UMAP plots after using respective normalization methods for the top 3000 HVGs for the simulated counts data set - Simulated Counts Scenario 2 (6 groups with approximately same number of cells per group but with different number of DE genes per group). The simulated counts were generated by Splat [37] using the HEK293T data set. Dots represent cells and are colored using the known cell-type labels (legend at the bottom of the panel; actual numbers of cells belonging to each group are specified in parentheses). Clustering metrics based on comparisons between predicted cell labels and known cell labels are listed in the bottom-right corner in each panel. With Piccolo, cells belonging to the 6 groups cluster distinctly and can be distinguished better compared to other normalization methods.

Fig. S13

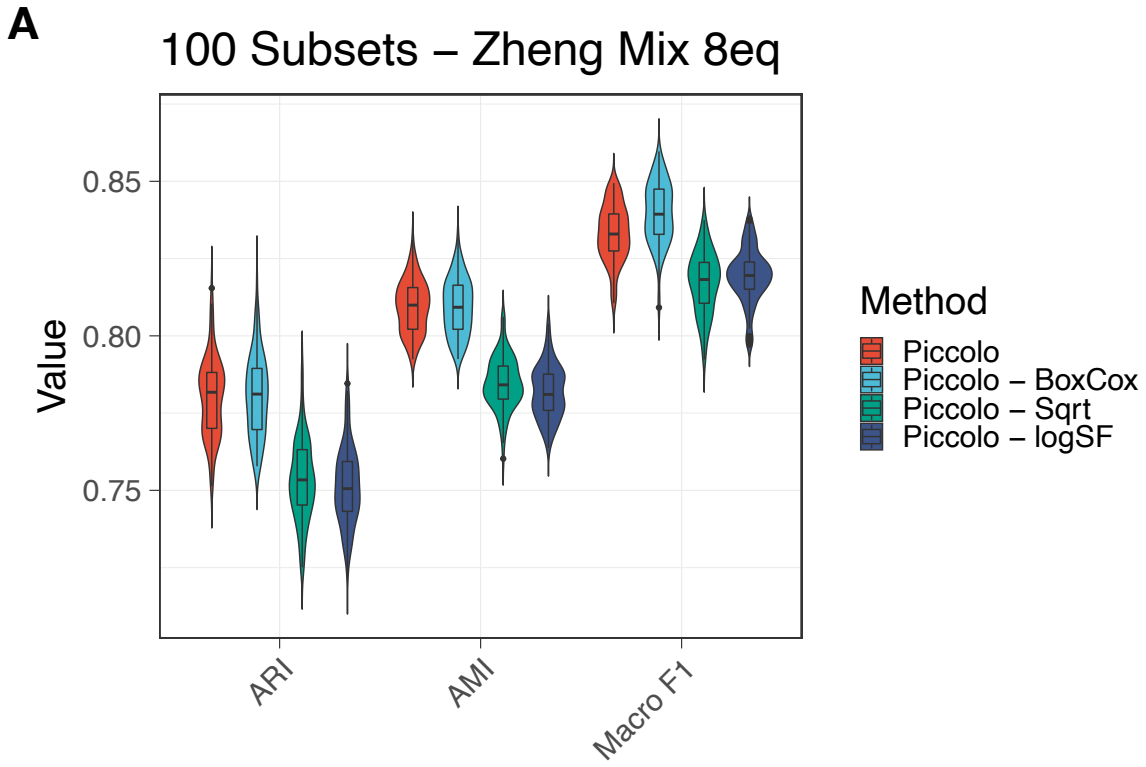

Fig. S13. Comparison between normalizations using the variance stabilization transformations implemented in Piccolo. A. Violin-box plots of the clustering metrics obtained for 100 subsets of the Zheng Mix 8eq data set. The colors correspond to the respective variance stabilization transformation employed during normalization using Piccolo. The default of *log* transformation is not labeled explicitly and is simply labeled as Piccolo. For all the 100 subsets, the highest values of the metrics were observed with Piccolo (red) and Piccolo - BoxCox (blue). In fact, Piccolo - BoxCox yields higher values for the Macro F1 score compared to Piccolo (Paired Wilcoxon test (2-sided): $p = 2.78E - 12$).

Fig. S14

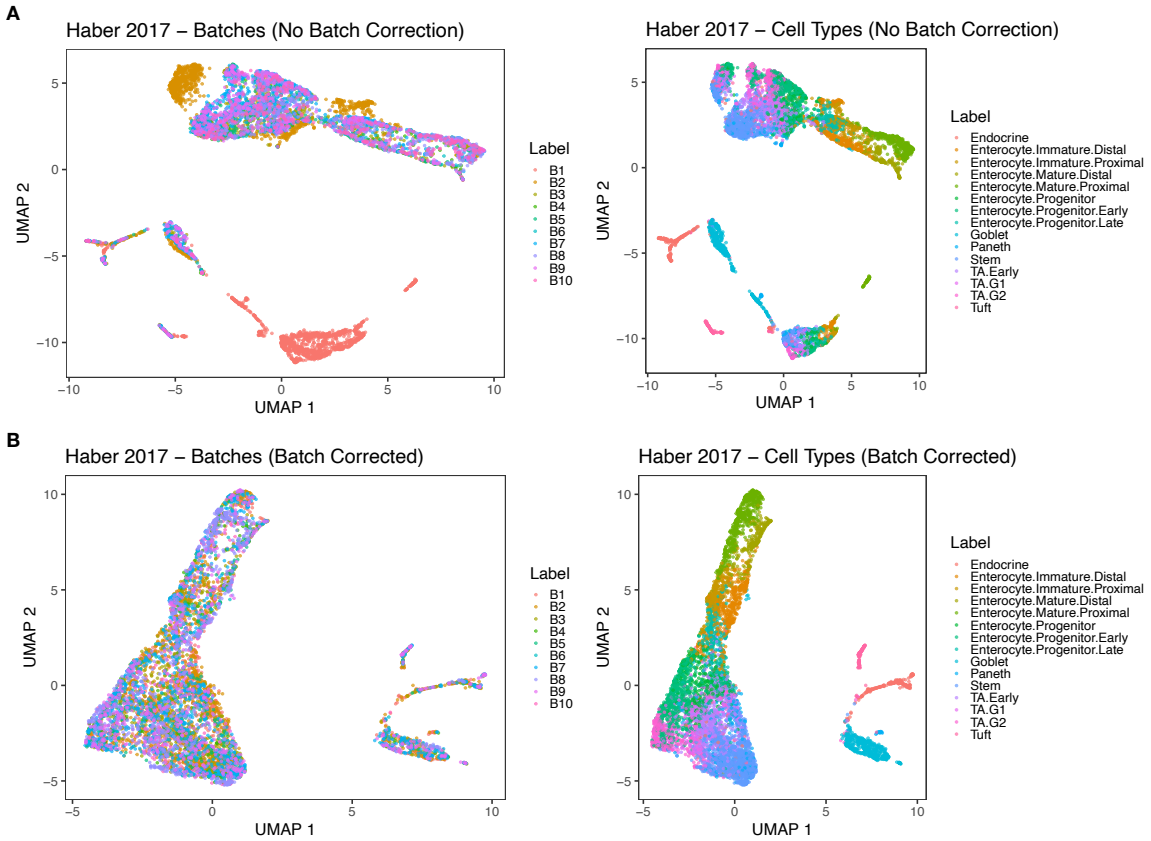

Fig. S14. Simplistic batch effect correction enabled by Piccolo normalization works well when all or most of the biological conditions are present in all the batches. In all the plots, the dots represent cells and are colored based on the batch or cell-type labels. **A.** 2D UMAP plots after applying Piccolo without specifying batches for the Haber 2017 - Mouse Small Intestine Epithelium data set. The left panel shows cells colored by batch. We can observe clusters that clearly correspond to batches (in particular, corresponding to batches B1 (red) and B2 (golden brown)). The right panel shows the cells colored by the cell-type labels provided by the authors of the study, which reveals that within clusters belonging to different batches there are cells belonging to distinct cell-types. **B.** 2D UMAP plots after applying Piccolo wherein we specify the 10 batches. The left panel shows cells colored by batch. We now observe that cells no longer cluster based on the batches. The right panel shows cells colored by the cell-type labels provided by the authors, which clearly reveals that the cells group based on their cell-type identities after batch effect correction.

Fig. S15

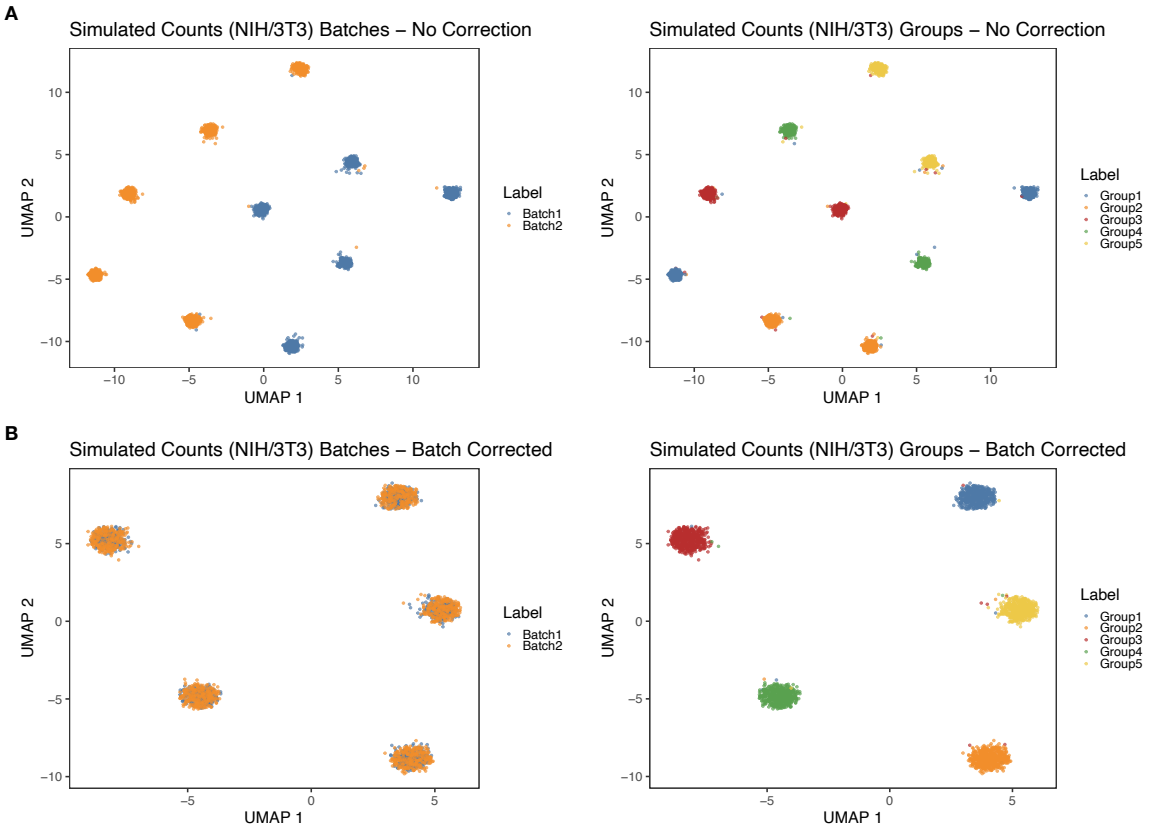

Fig. S15. Simplistic batch effect correction enabled by Piccolo normalization works well when all the conditions (group identities) are present in all the batches. In all the plots, the dots represent cells and are colored based on the batch or group labels. **A.** 2D UMAP plots after applying Piccolo without specifying batches for a simulated counts data set in which we simulated 2 batches, with equal number of cells in each batch. The simulated counts were generated by Splat [37] using NIH/3T3 (5 groups with approximately same number of cells per group). The left panel shows cells colored by batch. We observe 10 clusters with 5 belonging to Batch1 (blue) and the other 5 belonging to Batch2 (orange). The right panel shows cells colored by the group labels, which reveals that there are 2 clusters for each group, corresponding to the 2 batches. **B.** 2D UMAP plots after applying Piccolo wherein we specify the 2 batches. The left panel shows cells colored by batch. We now observe 5 clusters, with each cluster containing cells from both batches (orange overlaid on blue). The right panel shows cells colored by the group labels, which clearly reveals that the 5 clusters correspond to the group identities of the cells.