Published online 24 November 2021

Nucleic Acids Research, 2022, Vol. 50, No. 2 ***e12***
https://doi.org/10.1093/nar/gkab1071

Enhancing biological signals and detection rates in single-cell RNA-seq experiments with cDNA library equalization

Rhonda Bacher [1,*], Li-Fang Chu[2,3], Cara Argus[3], Jennifer M. Bolin[3], Parker Knight[4], James A. Thomson[3], Ron Stewart[3] and Christina Kendziorski[5,*]

[1]Department of Biostatistics, University of Florida, FL, USA, [2]Department of Comparative Biology and Experimental Medicine, University of Calgary, Calgary, AB, Canada, [3]Morgridge Institute for Research, Madison, WI, USA, [4]Department of Mathematics, University of Florida, FL, USA and [5]Department of Biostatistics, University of Wisconsin-Madison, WI, USA

Received September 02, 2020; Revised October 14, 2021; Editorial Decision October 15, 2021; Accepted October 20, 2021

ABSTRACT

Considerable effort has been devoted to refining experimental protocols to reduce levels of technical variability and artifacts in single-cell RNA-sequencing data (scRNA-seq). We here present evidence that equalizing the concentration of cDNA libraries prior to pooling, a step not consistently performed in single-cell experiments, improves gene detection rates, enhances biological signals, and reduces technical artifacts in scRNA-seq data. To evaluate the effect of equalization on various protocols, we developed Scaffold, a simulation framework that models each step of an scRNA-seq experiment. Numerical experiments demonstrate that equalization reduces variation in sequencing depth and gene-specific expression variability. We then performed a set of experiments in vitro with and without the equalization step and found that equalization increases the number of genes that are detected in every cell by 17–31%, improves discovery of biologically relevant genes, and reduces nuisance signals associated with cell cycle. Further support is provided in an analysis of publicly available data.

INTRODUCTION

Single-cell RNA-sequencing (scRNA-seq) protocols have evolved rapidly over the last 10 years, with increased throughput and sensitivity allowing for unprecedented insights into cell type heterogeneity across tissues (1). In spite of the advances, substantial technical variability and biases remain, which present challenges in data analysis and can obscure biological signals (2–5). From mRNA capture, reverse transcription, and PCR amplification, to additional single-cell library preparation and multiplex sequencing, there are numerous opportunities for technical noise to arise in scRNA-seq experiments. Inefficiencies or biases at any of the steps in the protocol may lead to increased technical artifacts and noise affecting expression variability and increase the proportion of dropout (6,7).

Numerous computational approaches including data smoothing and imputation have been developed to address excess variability and zeros in scRNA-seq data (8,9). However, they do so with the risk of introducing or perpetuating bias (8,10), thus making it preferable to optimize experimental protocols when feasible. A few studies have evaluated the downstream effects of various amplification techniques (11) or reverse transcriptases (12) on scRNA-seq data. However, to our knowledge no study has assessed the effect of equalizing cDNA concentrations in single-cell protocols. In bulk RNA-seq experiments, equalization of cDNA concentrations across libraries is a standard procedure that has been shown to reduce sequencing coverage variability and increase transcriptome diversity (13–15) by providing more even sequencing coverage of all samples. Equalization also leads to decreased sequencing of highly abundant transcripts and increases the efficiency at which low and moderately expressed genes are sequenced in bulk experiments (14).

For single-cell RNA-seq we hypothesized that equalization may improve sensitivity by increasing gene detection and thus began our investigation into the technical artifacts in scRNA-seq data by developing a simulation framework that generates counts by modeling each step of the experimental protocol. Simulation frameworks offer a significant advantage to studying sources of variability compared to experimental approaches as they allow an investigator to quickly assess a large number of scenarios at considerably

*To whom correspondence should be addressed. Tel: +1 352 294 5914; Email: rbacher@ufl.edu
Correspondence may also be addressed to Christina Kendziorski. Tel: +1 608 262 3146; Email: kendzior@biostat.wisc.edu

low cost. While a number of good methods are available for simulating scRNA-seq data (16–18), most do not model each step in the experimental protocol, and therefore are not useful for assessing how each step of the process affects the final counts. Two frameworks have attempted to study the data generation process but are limited in scope, either relying on spike-ins (19) or combining all sources of variation into a single parameter (20). Scaffold models the scRNA-seq data generating process by representing each step of the protocol mathematically, from the initial cell-to-cell heterogeneity to the final sequencing (Methods). Here we mainly focus on the SMART-SEQ (21) protocol as it uses oligo-dT priming and template switching as the backbone chemistry to generate cDNA from single cells which is used in multiple major scRNA-seq platforms, including Fluidigm C1 and 10X Chromium. The simulation framework is implemented in the R package, R/Scaffold and freely available at https://github.com/rhondabacher/scaffold/.

Based on our simulation results, which suggest that equalization is a critical step in the scRNA-seq protocol, we designed a set of scRNA-seq experiments in which we varied the extent at which cDNA libraries were equalized. The experiments demonstrate that equalization results in more consistent detection of genes, reduced expression variability, and reduced variability in the count-depth rate (3), the relationship between a gene's observed expression and sequencing depth. Finally, we confirm the effect of equalization in a survey of publicly available scRNA-seq datasets.

MATERIALS AND METHODS

EC and TB cell experiments

We focused on a subset of 96 single cells, from hESC-derived endothelial cells (EC) or trophoblast-like cells (TB) generated using the Fluidigm C1 system. The original data is considered to be unequalized (unEQ), where the single-cell cDNA libraries were first diluted to a range of 0.125–0.375 ng for subsequent library preparation protocols. The unEQ data was published in a previous study (GEO: GSE75748) (22). In the subsequent EQ experiments performed here, including EQ, EQ-Vary and EQ-75%, we retrieved the harvested cDNA, which are amplified full-length single-cell cDNAs identical to those used for the unEQ experiments (Supplementary Figure S1), but further diluted and adjusted so only 0.1 ng of cDNA were used as input across all the cells for subsequent library preparation protocols. In all the experiments, 1.25 μl of indicated input cDNA were used in a 5.0 μl Tagmentation reaction (Nextera XT DNA Sample Preparation Kit, Illumina) followed with a 12.5 μl dual-indexing PCR amplification reaction (Nextera XT DNA Sample Preparation Index Kit, Illumina). In the unEQ, EQ and EQ-75% experiments, 2.0 μl of the amplified/tagmented cDNA were used for pooling. In the EQ-Vary experiment, a single scaling factor was applied to generate variable amounts of the pooling volume. These pooled single-cell libraries were used in an AMpure XP Bead-based Dual Bead Cleanup and Size Selection reaction (Agencourt AMPure XP PCR Purification modified Instructions for Use, Beckman Coulter). In both bead cleanup reactions, 90% of AMPure XP beads were added to the amplified single-cell libraries to select for an approximate size range of 150–700 bp and incubated for 15 min at room temperature. Libraries bound to beads were then placed on a magnet for 5 min, washed twice with 70% ethanol, eluted with Suspension Buffer (Nextera XT DNA Sample Preparation Index Kit, Illumina), and transferred to a new tube. Final amplified and pooled single-cell libraries were quantified with the Qubit dsDNA HS Assay Kit (Q32854, Thermofisher) and Bioanalyzer High Sensitivity DNA Analysis Kit (5067-4626, Agilent). The unEQ libraries were multiplexed with 18–20 samples per lane and sequenced on an Illumina HiSeq2500 with single-end 51 bp reads while the EQ, EQ-75% and EQ-Vary were all pooled with 96 samples per lane and sequenced on an Illumina HiSeq3000 with paired-end 65 or 78 bp reads.

Processing and quality control on cells across equalization experiments

Reads were mapped against the GRCh38 Ensembl reference of protein-coding genes via Bowtie 1.2.3 (22), allowing up to two mismatches. The expected counts were estimated via RSEM 1.2.31 (23). To control for any differences due to differing read lengths all reads were first trimmed to have a length of 51 bp. In the initial unEQ experiment, cells that had fewer than 5000 genes with TPM $>$ 1 or that upon inspection of cell images displayed doublets or appeared dead were removed in quality control.

Using the scater v1.18.6 R package (24) we removed cells from any experiments in which the log10 sequencing depth was $<$5.4 or the percent of counts in the top 50 genes was $>$31%, the thresholds corresponding to being two standard deviations away from the median (Supplementary Figure S2). The expected counts in all experiments were rounded to the nearest whole number for all subsequent analyses.

Comparison of cell-specific and gene-specific detection rates

The cell-specific detection rate was calculated as the proportion of genes with nonzero expression within each cell. Similarly, the gene-specific detection rate was calculated as the proportion of cells with nonzero expression for each gene. When comparing differences in gene-specific detection rates between the experimental datasets, we accounted for differences in the sequencing depth since more sequencing typically results in more genes detected. For each comparison we subset the cells such that the average difference in sequencing depths was zero.

Analysis of highly variable genes

For the analysis of highly variable genes, gene expression estimates were first normalized using SCnorm v1.6.0 (3). We then fit a mean-dependent trend across each gene's mean-variance relationship. The trend represents technical variability and a gene's biological variability was calculated from the residuals using the functions trendVar and decomposeVar in the scran v1.12.1 R package (25). The decomposeVar function tests for nonzero biological variability using an F-test of total variability to technical variability. We considered genes significantly highly variable if they had an

FDR < .10. In order to compare gene variability across datasets, we ranked a gene's relative variability to all other genes in the dataset and calculated the difference in the two ranks.

Estimating the count-depth rate

The gene-specific count-depth rate was estimated within EC and TB separately using a median quantile regression on the log nonzero gene expression versus log sequencing depth using the getSlopes function in the SCnorm v1.6.0 R package. For each condition, we filtered out genes that had fewer than 10 nonzero expression counts across all cells and genes with median nonzero expression less than two. Visualization of the count-depth rate distributions is shown using smoothed density plots of the slopes within gene groups, where genes were split into 10 equally sized groups based on their nonzero median expression. The variability of the count-depth rate is quantified using the median absolute deviation statistic (MAD). First, the mode of the slope distribution was estimated for each gene group, then the MAD was calculated as the median of the absolute differences between the slope modes and one, where one is the expected value of the count-depth rate. All density plots of the slope distribution are done with smoothing parameters adjust = 1, and estimated over the grid (–3, 3) using the density function in R. All analyses were carried out using R version 3.6.3.

Analysis of publicly available datasets

We obtained processed counts from the conquer scRNA-seq database (26) for four single-cell RNA-seq datasets processed identically: Deng *et al.* (27), Grün *et al.* (28), Guo *et al.* (29) and Shalek *et al.* (30). The Chu *et al.* (31) data was obtained from the Gene Expression Omnibus (GEO) with the accession number GSE75748. The Islam *et al.* (32) data was obtained from GEO with the accession number GSE29087. The H1-bulk data from Bacher *et al.* (3) was obtained from GEO with the accession number GSE85917. The Picelli *et al.* (33) was obtained from GEO with the accession number GSE49321. The Smart-seq3 datasets from Hagemann-Jensen *et al.* (34) were obtained from ArrayExpress E-MTAB-8735. The 10× dataset is the pbmc4k dataset from the 10× Genomics website processed by Cell Ranger 2.1.0.

For the non-UMI datasets, cells with fewer than 10 000 total counts were removed and counts were rounded to the nearest whole number. For estimating the count-depth rate, again we filtered out genes that had fewer than 10 nonzero expression counts across all cells and genes with median nonzero expression less than two. In Figure 4, the representative datasets displayed from each study are: EF cells from Islam, Earlyblast-Embryo2 in Deng, M11W-Embryo2 in Guo, Unstim-Rep1 in Shalek and TB2 in Chu. The Picelli and H1-Bulk each only had one dataset in the study. The comparison of properties in Table 1 for the equalized versus unequalized datasets in publicly available studies was done using a two-sided t-test.

For the Grün UMI dataset, the isOutlier function in the scater v1.18.6 R package was used to remove cells having total detected genes greater than five median absolute deviations from the median. The isOutlier function was similarly applied to the Smart-seq3 datasets to remove outlier cells based on total counts and total genes detected per cell. For the 10X dataset, we first used the emptyDrops function in the DropletUtils v1.10.3 R package (35) to remove empty droplets containing ambient RNA and kept cells with an FDR <0.01. Cells were further filtered using the isOutlier function using three quality control metrics: total counts, genes detected per cell, and the percent of mitochondrial counts; outliers were considered as those above three median absolute deviations. Count-depth relationships were not estimated for the 10× or the Smart-Seq3 HCA dataset due to the large number of zeros in the data. For estimating the count-depth relationship in the UMI and Smart-seq3 Fibroblast datasets, all genes having a nonzero mean were included (the median in these datasets is often zero).

Simulation Framework

Here we first describe the data-generating process in Scaffold and the following section contains details on the estimation procedures. Let $M_{g,j}$ be the true number of mRNAs present for gene g in cell j with distribution, $M_{g,j} \sim Poisson(\omega_j \mu_g)$, where $g = 1, \ldots, G$, $j = 1, \ldots, N$, and μ_g is the latent level of gene-specific expression. As not all cells in a population are identical, the parameter ω_j is a cell-specific population heterogeneity parameter $\omega_j \sim Uniform(\omega_{.05}, \omega_{.95})$; scaling factors are applied to each cell to represent the range of cellular heterogeneity.

In an scRNA-seq experiment, a cell is first isolated and its mRNA is captured following cell lysis. A reverse transcription step occurs immediately after to convert the mRNA into cDNA. It is currently not possible to naturally estimate these two steps separately. Thus, here we model both of these events together as a single process. The number of molecules successfully captured for genes in cell j is represented as:

$$Z_{1,j}, \ldots, Z_{G,j} \sim Multinomial \times \left(\lambda_j \sum_{g=1}^{G} M_{g,j}, \frac{M_{1,j}}{\sum_{g=1}^{G} M_{g,j}}, \frac{M_{2,j}}{\sum_{g=1}^{G} M_{g,j}}, \ldots, \frac{M_{G,j}}{\sum_{g=1}^{G} M_{g,j}} \right),$$

where λ_j is the efficiency of conversion, referred to as the capture efficiency. Following this step, the cDNA molecules are exponentially amplified using PCR. The number of successfully amplified cDNA molecules for gene g in cell j is: $A_{g,j} = Z_{g,j}\,(1 + \rho_j)^C$, where C is the number of amplification cycles and ρ_j is the efficiency. When $\rho_j = 1$, all molecules double each cycle. We expect ρ_j to vary across reactions and to be independent across cells.

For droplet/10X protocols, the capture step occurs for each cell independently and all cDNA is then combined for further library preparation (skipping the cell-independent pre-amplification step). For plate-based methods like Smart-seq, the next steps involve re-plating the cells for further library preparation where cells are still processed independently. At this point, cDNA concentrations are typically quantified in part to ensure that quality is high. An optional next step is to equalize the cDNA concentrations to make them as similar as possible. This is first done by determining an acceptable range of concentrations – one may

Table 1. Summary of publicly available datasets. The first column contains the dataset name. Column 2 shows the organism. Column 3 shows the sequencing protocol used. Column 4 shows the number of cells per dataset included in the study. Column 5 is average sequencing depth across all cells. Column 6 is the average cell-specific detection rate across all cells. Column 7 is the average MAD and Column 8 indicates whether cDNA equalization was performed. Datasets above the black line are non-UMI and shown in Figure 4

Dataset	Organism	Protocol	Number of cells	Average sequencing depth (millions)	Average cell-specific detection rate	Average MAD	cDNA equalization
H1-bulk	Human	Bulk	48	3.0	0.73	**0.045**	**Yes**
Picelli	Human	SC–Smart-seq2	35	11.7	0.47	**0.141**	**Yes**
Deng	Mouse	SC–Smart-seq	11–22	13.3	0.65	**0.162**	**Yes**
Guo	Human	SC–Tang *et al.* (44)	12–31	3.5	0.47	**0.247**	**Yes**
Shalek	Mouse	SC–Smart-seq	64–96	3.4	0.39	**0.431**	**No**
Islam	Mouse	SC–STRT-seq	44–48	0.6	0.19	**0.480**	**No**
Chu	Human	SC–Smart-seq	31–87	4.6	0.50	**0.523**	**No**
Grün UMI	Mouse	SC–CEL-Seq (UMI)	562	0.004	0.02	**0.307**	**No**
10X	Human	SC–10X	3735	0.004	0.04	-	**No**
SS3-HCA	Human	SC–Smart-seq3 (UMI)	3112	0.252	0.09	-	**Yes**
SS3-Fibroblast	Mouse	SC–Smart-seq3 (UMI)	369	1.26	0.39	**0.096**	**Yes**

dilute all concentrations to the smallest observed concentration, or alternatively dilute a subset of cells to ensure that all concentrations are within a small target range. In Scaffold, the dilution factor is generated as $S_j \sim Normal(\tau_j,\ 0.01)$, where τ_j is :

$$\tau_j = \begin{cases} 0.95, & \text{if } l_j < q^* \\ \dfrac{q^*}{l_j}, & \text{otherwise} \end{cases}$$

and l_j is the cDNA concentration for cell j; q^* is the upper limit of the acceptable concentration range. For cells having concentrations within the acceptable range, there is no dilution. In this case, Scaffold sets the dilution factor to 0.95 indicating that, on average, 95% of the cDNA molecules will be retained in the next step (100% is not used as some loss of material may occur in the next step when transferring liquids). To mimic the situation in which all concentrations are diluted to the smallest observed, q^* is set to be the concentration of the smallest cell. If a range is chosen, then q^* is set to the midpoint between the lowest concentration and the concentration at a user-specified quantile; all concentrations larger than q^* are then diluted as described above.

The number of cDNA molecules in cell j after equalizing cDNA concentrations is:

$$A^*_{1,j}, \ldots,\ A^*_{G,j} \sim Multinomial \times \left(S_j \sum_{g=1}^{G} A_{g,j},\ \frac{A_{1,j}}{\sum_{g=1}^{G} A_{g,j}}, \frac{A_{2,j}}{\sum_{g=1}^{G} A_{g,j}}, \ldots, \frac{A_{G,j}}{\sum_{g=1}^{G} A_{g,j}} \right)$$

Following the protocols for C1 Fluidigm (Smart-seq and Smart- seq2), the cDNA is fragmented into shorter pieces and sequencing adapters and cell-specific indexes are added. We model this similarly to capture efficiency since the failure of any particular cDNA removes it from further consideration in sequencing. This is commonly referred to as 'tagmentation'. We denote the tagmentation efficiency here as γ_j. The number of cDNA molecules successfully tagmented for genes in cell j is represented as:

$$T_{1,j},\ \ldots,\ T_{G,j} \sim Multinomial \times \left(\gamma_j \sum_{g=1}^{G} A^*_{g,j},\ \frac{A^*_{1,j}}{\sum_{g=1}^{G} A^*_{g,j}}, \frac{A^*_{2,j}}{\sum_{g=1}^{G} A^*_{g,j}}, \ldots, \frac{A^*_{G,j}}{\sum_{g=1}^{G} A^*_{g,j}} \right)$$

Next, the cDNA molecules go through a second round of PCR amplification, where for gene g in cell j the number of amplified molecules is represented as: $B_{g,j} = T_{g,j}\ (1 + \rho_{2,j})^{C_2}$, where C_2 is the number of amplification cycles and $\rho_{2,j}$ is the efficiency per cell. Finally, the observed gene counts per cell, $Y_{g,j}$, are obtained by:

$$Y_{1,1}, \ldots, Y_{G,N} \sim Multinomial\,(R,\ \pi)$$

where $\pi = (\pi_{1,1}, \ldots, \pi_{G,1},\ \ldots,\ \pi_{G,1}, \ldots,\ \pi_{G,N})$, $\pi_{g,j} = \frac{B_{g,j}}{\sum_g \sum_j B_{g,j}}$, and R is the total number of sequences obtained.

To simulate data from UMI protocols, the same steps above are followed, with Scaffold tracking the unique molecules throughout the simulation framework. For 10X/droplet based protocols, there are a few differences in the procedure. Specifically, there is no pre-amplification step, the transcripts from all cells are combined immediately following the capture step, and the tagmentation and PCR amplification steps do not having cell-specific parameters since tagmentation and PCR amplification are not cell-specific.

Estimation of simulation parameters

For the simulation framework described above, a number of parameters must be set or estimated. All parameters that are estimated can also be input by the user if desired so that no input dataset is actually needed to simulate data. The number of genes and cells are estimated from the input dataset. Since the expression means observed in sequence data do not necessarily reflect the total transcripts in the cells, Scaffold first scales all cells to have a total of 300 000 counts to estimate the mean for each gene as human cells have been previously reported to have approximately 300k total transcripts (36). This parameter can be changed in Scaffold, if desired. The estimated means are then used in the Poisson distribution to generate the starting number of mRNA per cell. To estimate the cell heterogeneity parameters, $\omega_{.05}$ and $\omega_{.95}$, Scaffold calculates the ratio of total counts for a random sample of 100 pairs of cells in the observed data, then the 5th and 95th percentile of ratios are used as the lower and upper bound in a Uniform distribution to draw cell-specific scaling factors.

The majority of zeros are thought to occur during the capture step (cell lysis and reverse transcription), thus the capture efficiency has the largest impact on the detection rates. It is not possible, even with spike-ins, to differentiate these two steps; and consequently, default settings in Scaffold treat them as a single step. However, we note that it is possible to simulate these two steps separately. We estimate the capture efficiencies from a Normal($\mu_\lambda, \sigma_\lambda$). The mean capture efficiency, μ_λ, is estimated in Scaffold as follows:

(i) A weighted probability of observing each gene is calculated as $p_g = \mu_g / m$, where $\mu_g$is the gene mean in the initial simulated mRNA counts and m is the total number of genes.
(ii) For a given mean capture efficiency, we calculate $P(X_g = 0)$, where $X_g \sim Binomial(\mu_\lambda * m, p_g)$. To avoid heavy computation, this is estimated for a random representative subset of 100 genes.
(iii) μ_λ is chosen as that which minimizes the difference between $P(X_g > 0)$ and the average detection rate per cell in the observed data. This search is done using the optimize function in R. For UMI and 10× datasets, we found better estimation accuracy using $P(X_g > 1)$, which corresponds to an increased probability of initial counts of one being observed as zeros in the sequenced data. The standard deviation, σ_λ, is calculated from the observed data as the median absolute deviation of the cell detection rates.

The default number of cycles for each PCR step is set to the number used by the Smart-seq protocol as detailed in the Fluidigm user manual (first PCR is 18 cycles and the second PCR is 12 cycles), and these can be adjusted by the user. In our testing we did not identify the PCR or tagmentation steps to have a major influence, and in the literature these are typically regarded as highly efficient procedures (37). Thus, the default distribution of the efficiency parameter for these steps is Normal(0.95, 0.02), but can be adjusted by the user as desired. Finally, the total sequencing depth is set to the sum of all the counts in the observed data. Details on the specific parameters for the simulated datasets is given in Supplementary Methods.

Simulating multiple populations

Following the initial generation of mRNA counts, multiple populations can be simulated by specifying additional parameters. The number of cells per population must be provided, and the first population serves as the reference from which each additional population differs by a proportion of genes having distinct expression. The expression differences are sampled from a Normal distribution with a mean and standard deviation of fold-changes given by the user; and the direction of expression differences for a given gene is chosen at random. For the comparison of cluster visualization between unEQ and EQ simulated datasets, we simulated two populations having 50 and 40 cells using Scaffold. The proportion of DE genes simulated was 10% with fold-changes drawn from a Normal(1.5, .5). All other Scaffold parameters were estimated from the unEQ EC dataset. For generating the UMAP (38), TSNE (39) and EDGE (40) embeddings and plots, 250 simulations were conducted. Within each simulation and for each dimension reduction algorithm, the mean silhouette distances were averaged over 25 iterations using different random seeds. UMAP and TSNE plots were obtained using the scater R package v1.18.6 using the first ten principle components, n_neighbors = 10 for UMAP, and perplexity = 25 for TSNE. EDGE plots were obtained with the EDGE R package v1.0 with the number of weak learners n_wl = 5000, nearest neighbors n_neighs = 10, n_dm = 10, hash table size H = 1000, and optimization opt = TRUE.

Simulating dynamic populations

To simulate datasets from a continuous or dynamic population, Scaffold simulates gene expression via a B-spline for a user-specified proportion of genes. The default spline generation is degree two, with two knots and coefficients sampled from a Normal(5,5); mRNA counts are then generated from a Poisson distribution with latent mean for each cell equal to the value from the B-spline. The spline generation parameters can be user-specified. For the simulation of trajectory analysis, we used the default settings and all other scaffold parameters were estimated from the unEQ EC dataset. The SCORPIUS R package v1.0.8 was used to infer the trajectory using default settings (41). A two-degree polynomial was used to identify genes having a significant dynamic along pseudotime. The pROC R package v1.17.0.1 was used to estimate the area under the receiver operator curve (AUC) (42).

RESULTS

In silico investigation of cDNA equalization using Scaffold

As detailed in Methods and Figure 1A, Scaffold allows for assessment of how each step of the single-cell protocol (cell lysis, amplification, equalization, library preparation, and sequencing depth) affects scRNA-seq measurements. Using an scRNA-seq dataset of unequalized endothelial cells (unEQ EC) as a reference, Scaffold estimated starting parameters and simulated data that reproduced the features of the unEQ EC dataset including gene-specific means, variances, and proportions of zeros (Figure 1B–E). We also simulated data using unequalized trophoblast cells (unEQ TB) as a reference with similar results (Supplementary Figure S3). Systematic variability in the count-depth rate, a feature shown to be unique to scRNA-seq data (3), was also reproduced (Figure 1F and Supplementary Figure S4).

Holding all other parameters constant, we simulated data while varying parameters for equalization and sequencing depth and found that cDNA equalization has the largest effect on the average variability in the count-depth rate (Supplementary Figure S4C, D), while the total sequencing depth (Supplementary Figure S4E) had little effect.

To examine the effect of equalization on other properties of the data, we simulated additional datasets with and without equalization holding all other parameters constant. Specifically, we simulated pairs of unequalized and equalized datasets by adjusting only the equalization parameter. In simulated datasets, gene-specific variation decreased

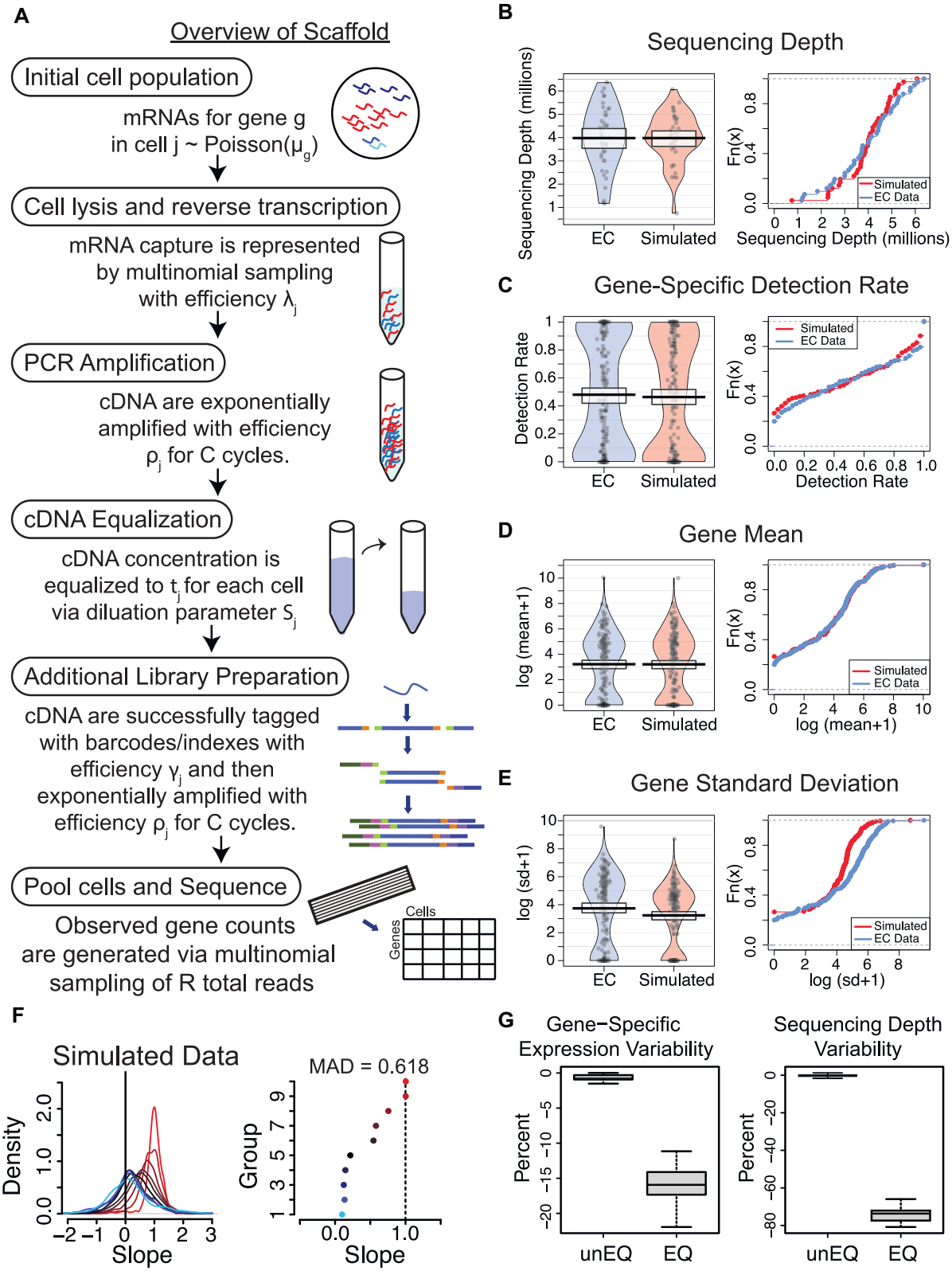

Figure 1. (**A**) Overview of the Scaffold simulation framework. Further details are provided in Methods. (**B–E**) Cell-specific and gene-specific properties of the data simulated based on the unEQ EC dataset. (**F**) Density plots of the distribution of estimated count-depth rates (quantified as the gene-specific slope of a median quantile regression) for the unEQ EC dataset for genes grouped by expression level (left) and the mode of each group's slope distribution (right). The median absolute deviation of the slope modes from one (MAD) is used to quantify the variability in the count-depth rate. (**G**) The percent change in gene-specific variability (left) and sequencing depth (right) is shown for multiple pairs of unequalized and equalized datasets. Multiple pairs of unequalized experiments were also simulated and compared to demonstrate the percent of change due to random sampling.

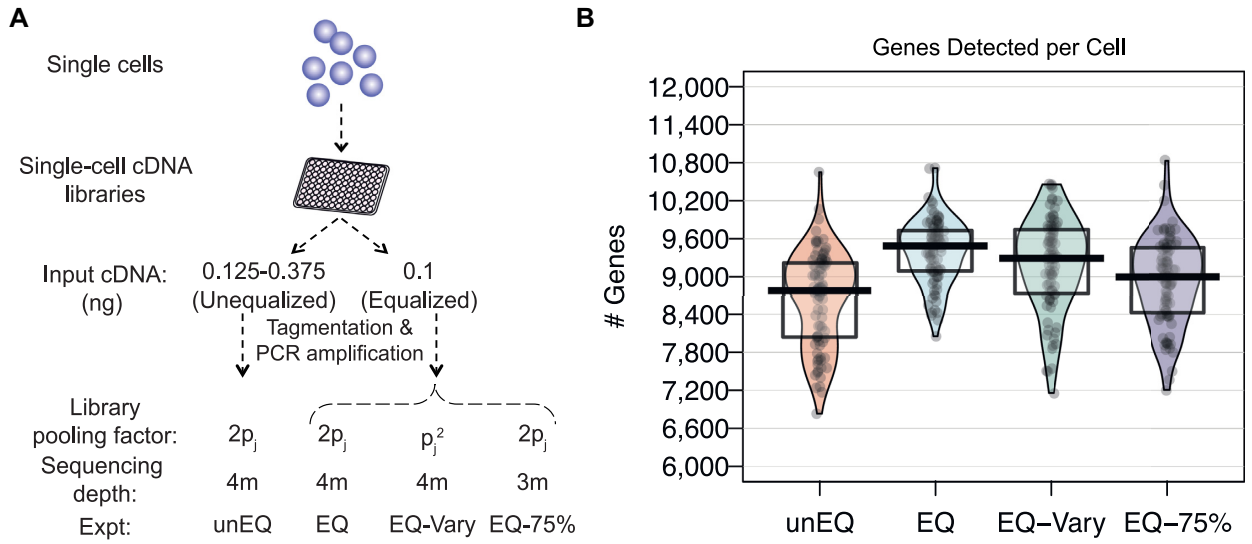

Figure 2. Overview of experiment to assess the effect of cDNA equalization and comparisons of cell-level detection rates. (**A**) Four experiments were conducted involving cells from two different conditions (EC and TB). Using the same initial pools of single-cell cDNA, we created unequalized and equalized sequencing libraries. (**B**) Violin plots with points overlaid of the number of genes detected per cell for all cells in each experiment.

by an average of 16.2% due to equalization alone and the variability in the sequencing depths was reduced by 74.2% despite the simulations having the same average depth (Figure 1G).

Experiments to assess the effect of cDNA equalization

Given results from the simulation study, we hypothesized that a lack of equalization during the preparation of single-cell libraries would increase variation in the amount of input cDNA which in turn could contribute to reduced gene detection and increased variability in expression estimates observed in scRNA-seq data. To test this hypothesis, we applied alternative protocols to full-length single-cell cDNA libraries of identical cells to generate matched scRNA-seq data sets (Figure 2A). The original data includes single endothelial cells (EC) and trophoblast-like cells (TB) derived from human embryonic stem cells (hESC) (31) which were unequalized (unEQ). For these experiments, the cDNA input ranged from 0.125 to 0.375 ng (Materials and Methods). In the next series of experiments, we equalized the same set of single-cell cDNA to a fixed input (0.1 ng) across all the cells. Prior to sequencing, cells were pooled at an equal volume (EQ) or pooled by a scaling factor to produce highly variable sequencing depths (EQ-Vary) (Figure 2A). Finally, we replicated the entire EQ experiment, including equalized cDNA input and pooling, but we sequenced at approximately three-quarters the depth of the previous experiments (EQ-75%). Because these four conditions all derive from identical cells, these experiments provide the most robust investigation to date on how input cDNA variations impact scRNA-seq data.

Equalization increases cell-specific and gene-specific detection rates

A common challenge in scRNA-seq experiments is the high proportions of zeros, or dropouts. Dropouts are due to an incomplete sampling process, stochastic gene expression, and inefficient capture of mRNA, with the probability of dropping out inversely related to a gene's underlying expression level (43). Equalizing cDNA libraries would not recover dropouts that occur upstream in a protocol, but it may recover dropouts that are due to inefficiencies in later preparation steps (e.g. second PCR amplification) or due to underrepresentation in the pooled library. Thus, we first investigated the effect of cDNA equalization on cell-specific detection rates, defined as the proportion of nonzero genes within a cell. Across both EC and TB cells, we observed an increase in the efficiency of gene detection in the equalized experiments (Figure 2B). An average of 745 (8.6%) more genes per cell were detected with expression greater than zero in the EQ versus the unEQ experiments. EQ-vary, which was pooled in a way to reflect possible inefficiencies that might occur after equalization such as during pooling or amplification, reduced the detection efficiency slightly to 534 (6.2%) more genes detected on average. Comparatively, the effect of equalization on gene detection is stronger than the effect of solely increasing total sequencing depth. Between EQ and EQ-75%, in which both experiments were equalized but the latter had three-quarters the sequencing depth, we observed only 470 (5.0%) fewer genes detected per cell in EQ-75%.

We next investigated the gene-level detection rate across experiments, defined as the proportion of cells with nonzero expression for each gene (Figure 3A, B). Here we calculated the difference in gene-level detection rates between EQ and unEQ while accounting for differences in sequencing depth (Methods). The overall increase in detection efficiency due to equalization translates to a 31.1% increase in genes having consistent detection in all EC cells and a 17.9% increase in TB cells (1002 and 622 genes, respectively). We also observed a 10.4% decrease in the number of genes not detected in any cells for EC and an 8.1% decrease in TB (382 and 276 genes, respectively).

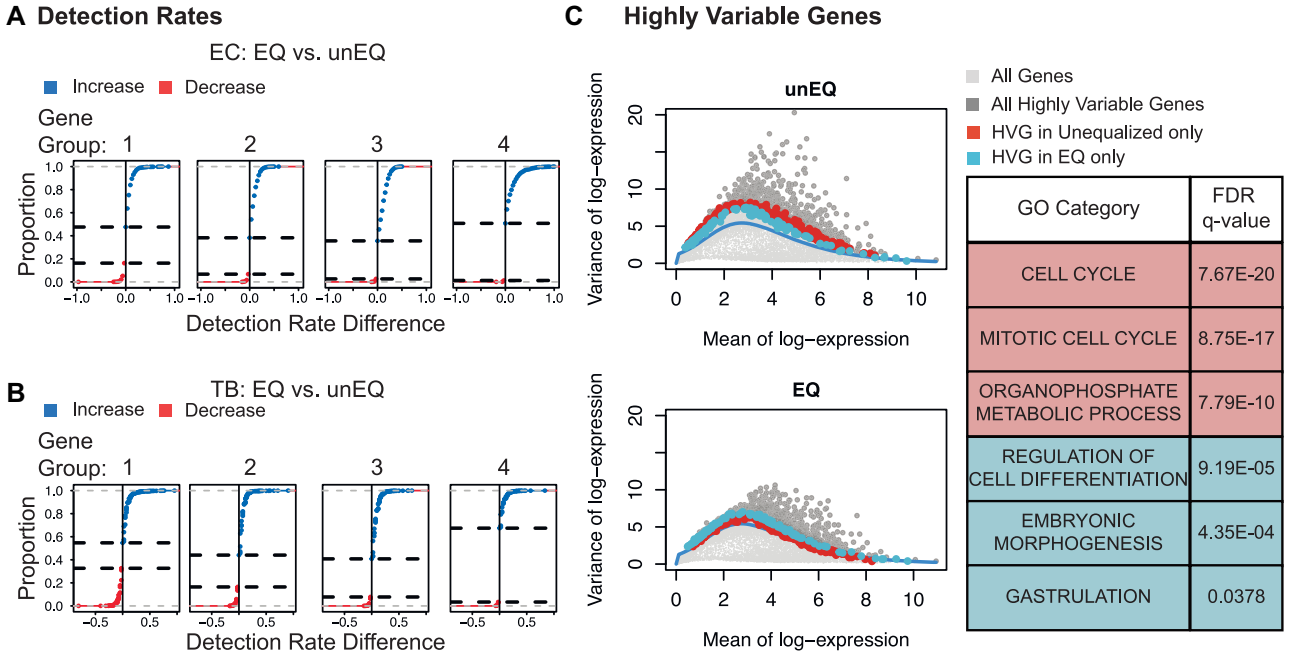

Figure 3. Equalization improves detection rates and decreases expression variability. (**A**) For the EC dataset, genes were divided into four equally sized groups based on their median nonzero expression. For each gene, the difference between the detection rate in the EQ versus the unEQ experiments was calculated. The cumulative distribution curve is shown for the detection rate differences for genes in each expression group. The two horizonal dotted lines indicate the proportion of genes that decrease in detection rate (bottom line) and one minus the proportion of genes that increase in detection rate (top line). (**B**) Same as A for the TB dataset. (**C**) Scatter plot of every gene's mean and variance for the unEQ (top) and EQ (bottom) datasets (light gray). The smoothed fit line represents technical variability. The mean and variance were calculated over all cells, both EC and TB. Genes having significantly high biological variability in either dataset are shown in dark gray. Shown in red are the highly variable genes in the unEQ dataset only, and in blue are the highly variable genes in the EQ dataset only. In the table are the top three GO biological processes enriched for genes that are only HVG in the unEQ (red) or EQ (blue) experiments.

Since a gene's detection rate is related to its expression level, we further analyzed detection differences by splitting genes into four equally sized gene groups based on their nonzero median expression. We assessed what differences would appear due to random chance by randomly splitting the EC or TB cells in the unEQ dataset into two groups and examining the detection rate differences between them. Approximately equal proportions of genes had increased/decreased detection rates across all expression groups for both experimental conditions (Supplementary Figure S5).

Between the EQ data and unEQ datasets, we consistently saw a higher proportion of genes having a higher detection rate in the equalized dataset especially among the moderately expressed genes (62% and 64% for EC gene groups 2 and 3; 56% and 59% for TB gene groups 2 and 3) (Figure 3A&B). The average increase in detection rate in the equalized experiments for the genes in Groups 2–4 is 13.6% in EC2 and 7.9% for TB2. In comparison, we performed the same analysis between the EQ and EQ-Vary datasets which underwent the same equalization procedure and found the ratio of genes with increasing versus decreasing detection rate was stable across expression groups; the increased variability in sequencing depth did not compromise the detection rate in the equalized dataset (Supplementary Figure S6).

To identify any functional relevance of genes with increased or decreased detection rates in the EQ experiment we performed gene-set enrichment using MSigDB's list of GO biological processes on the top 200 genes sorted by their magnitude change in detection. Genes with increased detection rate in the EQ experiment were enriched for important developmental processes including morphogenesis, and tube and epithelium development in both EC and TB (Supplementary Table S1). Genes with decreased detection rates after equalization tended to be among the most lowly expressed genes. Of the 200 genes with the most decreased detection, 142 were in the lowest expression group in EC and 162 such genes in TB. Taken together, these results suggest that equalization improves the detection of biologically relevant genes without compromising signal.

Equalization reduces nuisance variation

Next, we investigated the effect of equalization on gene expression variability. A common first step in single-cell clustering or trajectory inference analysis is to reduce the data to the most informative set of genes, often defined as the most highly variable genes (HVG). However, in the presence of excess nuisance variation, the top ranked HVG may not reflect the most relevant set of genes. Here, we detected HVG by decomposing the total variance of each gene into technical and biological components. To do so, we estimated a mean-dependent trend for the mean-variance relationship across all genes to represent technical variability (Methods). A gene's biological variability was calculated as the differ-

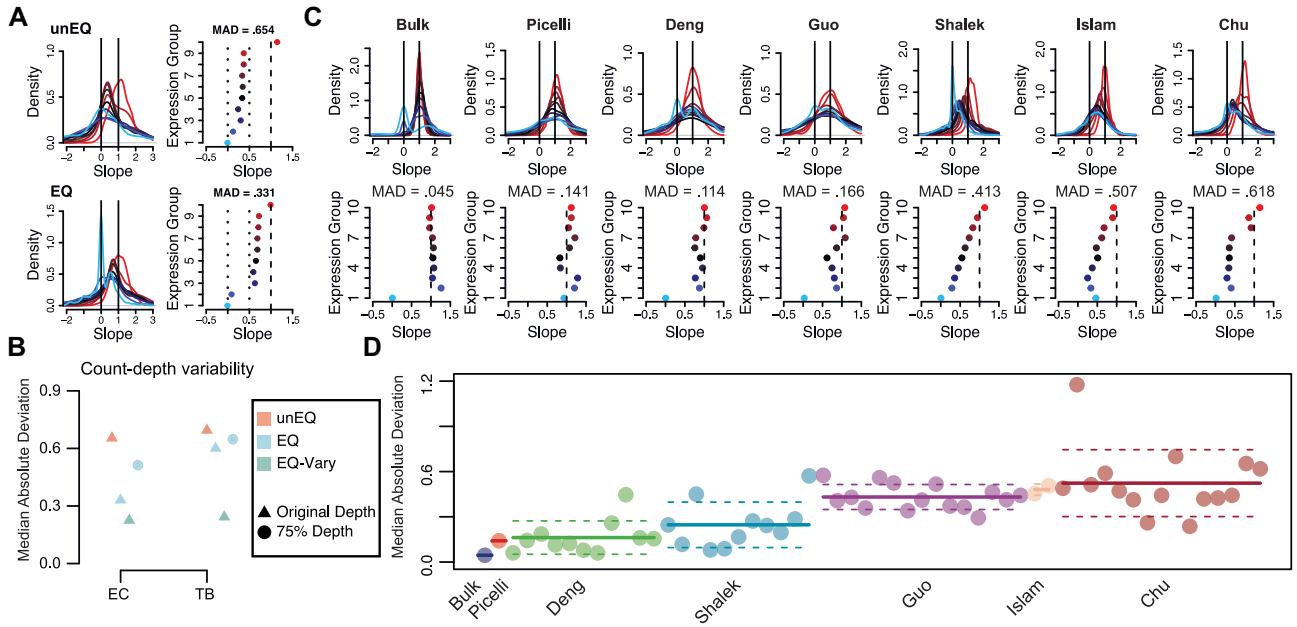

Figure 4. Count-depth rate in equalized scRNA-seq experiments. (**A**) For the unEQ and EQ EC datasets, the count-depth rate was calculated for all genes as the slope of a median quantile regression. Genes were divided into ten equally sized groups based on their median nonzero expression across all cells in the dataset. (**B**) The median absolute deviation (MAD) of the modal slope for each experiment is shown. (**C**) Same as A for seven representative datasets from seven published studies. (**D**) Similar to (B) for all datasets in the seven published studies. The solid line indicates the mean MAD and the dashed line indicates one standard deviation.

ence between a gene's total variability and its fitted trend value. An HVG classification was assigned to genes having biological variability significantly larger than zero (FDR < 0.10). HVG genes in the unequalized experiment were enriched in GO biological processes involving the cell cycle. This is likely due to the fact that cellular mRNA content is directly related to cell cycle stage and, consequently, if cDNA content is not equalized across cells, variability in cell cycle genes is prominent in the resulting data. Following equalization, genes classified as HVG were enriched for biological processes specific to EC cells including gastrulation and cell fate/differentiation (Figure 3C and Supplementary Table S2).

Equalization reduces technical artifacts in the count-depth rate

Previously, we reported that scRNA-seq data display systematic variation in the relationship between a gene's observed expression and sequencing depth (which we termed the count-depth rate), whereby a gene's expected increase in expression with increased sequencing depth fails to materialize (3). Variability in the count-depth rate affects downstream analysis as popular scale-factor based normalization methods assume that the count-depth rate is common across genes and equal to one on the log–log scale (3,25).

As shown in Bacher *et al.*, much of the variability in the count-depth rate arises from under-detection of genes despite increasing sequencing depth since highly expressed genes are over-represented during sequencing. Since equalizing cDNA increases detection rates, we hypothesized that it may also reduce variability in the count-depth rate. To investigate, we quantified the count-depth rate for every gene using median quantile regression, where a slope of one indicates a proportional increase of gene expression with sequencing depth (Supplementary Figure S7). Next, we binned genes into ten equally sized groups based on their median nonzero expression. In the unEQ dataset, we found only highly expressed genes had slopes near one and slopes gradually decreased with gene expression level (Figure 4A). The extent of variability in the count-depth rate was measured using the MAD of the ten groups slope mode from their expected value of one. The EQ experiments had a lower MAD and displayed less variability in the count-depth rates for both EC and TB (Figure 4A, B). EQ-75% was similar to the EQ datasets, indicating the count-depth rate is not affected by total sequencing depth. The EQ-Vary experiment had the most reduction in count-depth variability, with the majority of slopes close to 1 (Supplemental Figure S8), due to its increased dissociation of cell size with sequencing depth.

As more single-cell datasets have become public and identically processed in databases such as conquer (26), we were able to inquire whether systematic variability in the count-depth rate was reduced across scRNA-seq data in published studies. Across seven different studies, we found large heterogeneity in the experiment-specific count-depth rates with the MAD ranging from 0.045 to 1.176 (Figure 4C, D). We found no revealing association between the average MAD within study and various properties of the scRNA-seq data, including the average sequencing depth, cell-specific detection rate, organism, or number of cells (Table 1). However, consistent with our simulated and experimental datasets, the publicly available studies in which equalization was performed had significantly lower MAD values (P-value < 0.001), higher cell-specific detection rates

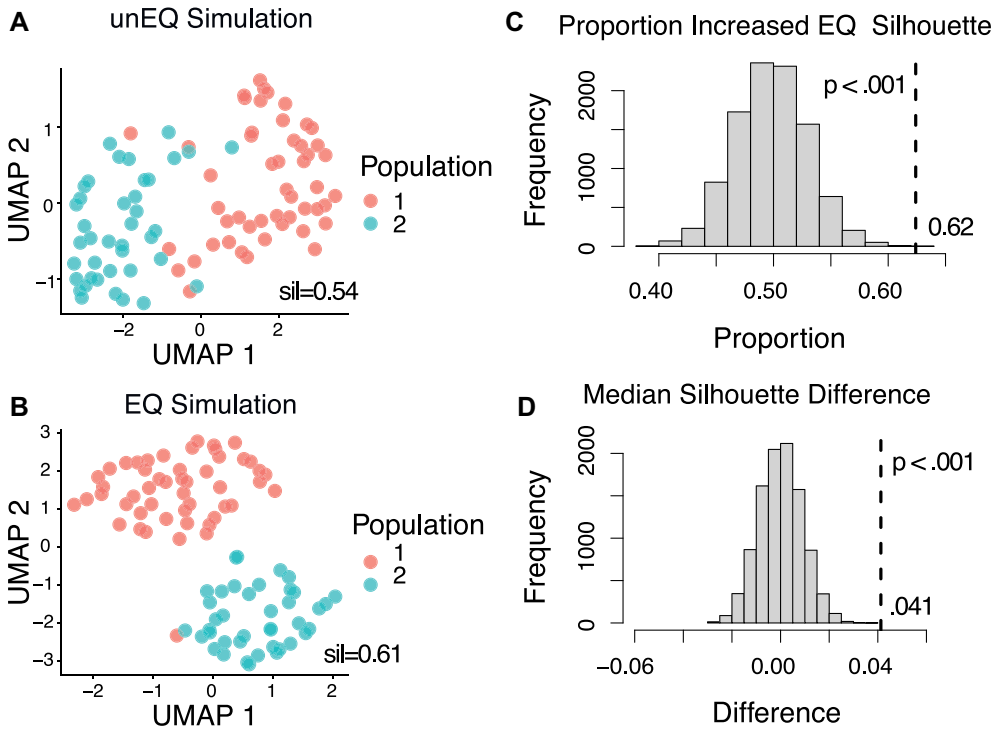

Figure 5. Pairs of unequalized and equalized experiments having two populations were simulated using Scaffold. Datasets were embedded in two-dimensions using UMAP and the silhouette distance was calculated for each dataset. (**A**) UMAP plot of one simulated unequalized dataset. (**B**) UMAP plot of one simulated equalized dataset. (**C**) Across all simulations, 62% had larger equalized silhouette distances compared to those of the paired unequalized distances (P-value < .001). The silhouette distances were permuted for each simulated dataset to obtain a sampling distribution under the null hypothesis of no difference due to equalization. P-values (p) were calculated over 10 000 permutations. The histogram shows the permutation distribution of the proportion of equalized simulated datasets having a larger silhouette distance. (**D**) The permutation distribution of the median silhouette differences. The median differences between unequalized and equalized simulated datasets was 0.041 (P-value < 0.001).

(P-value < 0.001), and higher gene-specific detection rates (P-value = 0.039) (Supplementary Figure S9). On average the equalized datasets contain 2215 additional genes detected consistently in every cell compared to the unequalized datasets (P-value < 0.001 and Supplementary Figure S9).

Equalization improves downstream analyses

To further examine how equalization might affect common downstream analyses, we simulated data for two scenarios – clustering analysis and trajectory analysis. For clustering, we used Scaffold to simulate datasets from multiple populations (Supplementary Figure S10). Here we consider two cell type populations with slight separation—only 10% of genes have distinct expression with an average fold change of 1.5. We simulated pairs of unequalized and equalized datasets and evaluated two-dimensional embeddings of cells using the silhouette distance. On average, equalization had a higher median silhouette distance and improved visible separation of cell populations in UMAP (38) plots (Figure 5). TSNE (39) and EDGE (40) reduced dimension embeddings showed similar trends (Supplementary Figure S11). We also used Scaffold to simulate cells coming from a continuous population, in which we assumed a proportion of genes have dynamic expression across the cells (Supplementary Figure S12). We simulated pairs of unequalized and equalized datasets and inferred a trajectory on the simulated cells (41). We then fit a polynomial regression of each gene's expression to the trajectory to determine the significantly dynamic genes (adjusted P-value < 0.05). Equalization had a slight improvement in the ability to detect dynamically expressed genes, with an AUC of 83.0 versus 81.7 for the unequalized simulations.

Using Scaffold to simulate data from UMI and 10× protocols

Although equalization cannot be applied to 10× protocols, or most UMI protocols, due to the vast number of cells these protocols produce, other aspects of the data generation process can be systematically explored. We applied Scaffold to a 10X dataset and three additional UMI datasets and observed that the simulated data was highly representative of cell- and gene-specific properties of the data (Supplementary Figures S13–S18).

DISCUSSION

Obtaining the highest quality data with minimal technical variability remains a goal for scRNA-seq experiments. Given the competitive nature of the sequencing process, highly expressed transcripts are often overrepresented in the final library and will consume a large proportion of the total reads leading to low detection rates for the majority of genes. Here, we showed that equalizing single-cell cDNA libraries prior to pooling improves detection rates and

decreases nuisance variation such as that attributable to cell cycle.

Our finding of reduced variability in expression for cell cycle genes in equalized experiments is novel, yet not unexpected since cell cycle signals are often the largest drivers of differences in total mRNA. Note that if cell cycle signals are of marked interest, then equalization may not be appropriate. However, reduction of cell-cycle signals has been implemented in most scRNA-seq analysis pipelines as it is considered a hindrance in most downstream analyses (45,46). While different cell types often have different cell sizes, they are also distinguished by relative differences in key marker genes. Equalization preserves these relative differences as the dilution is performed on the entire cell's cDNA and thus, would not remove cell-type specific differences.

In many cases, identified sources of technical variability in downstream analyses have proven to be excellent targets for protocol improvement (47–50). Scaffold, our simulation framework, offers an opportunity to directly and efficiently explore how different steps in a protocol affect scRNA-seq data. Here, we focused the effect of equalizing cDNA concentration across cells. However, Scaffold provides a framework to study other parameters, or to simulate data that recapitulates characteristics of scRNA-seq data (e.g. detection rates and count-depth rate).

In practice, the process of equalizing cDNA concentrations is non-trivial and time-consuming, leading it to be one of the critical limiting points of the library preparation process (51). Automation has alleviated this to some extent, and has been used in large single-cell sequencing projects such as the Tabula Muris (52). However, some state-of-the-art protocols, such as 10×, profile scRNA-seq measurements from thousands to millions of cells using massively parallel sequencing systems with high levels of multiplexing (51) and equalization is not possible since cDNA is pooled early in the experiment. We expect that single-cell protocols will continue to advance and improve with technology. Our study offers insight into one mechanism worth further exploration in protocol design and development.

DATA AVAILABILITY

All R code used for analysis and simulations is available at https://github.com/rhondabacher/scEqualization-Paper. The simulation package Scaffold is available at https://github.com/rhondabacher/scaffold. The unEQ, EQ, EQ-Vary, and EQ-75% datasets are available at the NCBI Gene Expression Omnibus: GSE156494.

For the publicly available datasets, we obtained processed counts from the conquer scRNA-seq database for four single-cell RNA-seq datasets processed identically: Deng *et al.* (27), Grün *et al.* (28), Guo *et al.* (29) and Shalek *et al.* (30). The Chu *et al.* (31) data was obtained from the Gene Expression Omnibus (GEO) with the accession number GSE75748. The Islam *et al.* (32) data was obtained from GEO with the accession number GSE29087. The H1-bulk data from Bacher *et al.* (3) was obtained from GEO with the accession number GSE85917. The Picelli *et al.* (33) was obtained from the GEO with the accession number GSE49321. The Smart-seq3 datasets (34) were obtained from ArrayExpress E-MTAB-8735. The 10X dataset is the pbmc4k dataset from the 10X Genomics website processed by Cell Ranger 2.1.0.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank J. Steill and S. Swanson for initial RNA-seq read processing.
Author contributions: R.B. and C.K. conceived and designed the research and wrote the manuscript. L.-F.C. and J.B. conceived, designed and performed experiments. R.B. processed and analyzed all datasets. P.K. contributed to simulation code development. R.S. and J.A.T. were involved in planning and supervising experiments. All co-authors contributed to the writing of the manuscript.

FUNDING

U.S National Institutes of Health [NIHGM102756 to C.K.]; Morgridge Institute for Research. Funding for open access charge: Corresponding author funds.
Conflict of interest statement. None declared.

REFERENCES

1. Svensson,V., Vento-Tormo,R. and Teichmann,S.A. (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, **13**, 599–604.
2. Hicks,S.C., Townes,F.W., Teng,M. and Irizarry,R.A. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, **19**, 562–578.
3. Bacher,R., Chu,L.-F., Leng,N., Gasch,A.P., Thomson,J.A., Stewart,R.M., Newton,M. and Kendziorski,C. (2017) SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods*, **14**, 584–586.
4. Phipson,B., Zappia,L. and Oshlack,A. (2017) Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res*, **6**, 595.
5. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
6. Vallejos,C.A., Risso,D., Scialdone,A., Dudoit,S. and Marioni,J.C. (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, **14**, 565–571.
7. Brennecke,P., Anders,S., Kim,J.K., Kołodziejczyk,A.A., Zhang,X., Proserpio,V., Baying,B., Benes,V., Teichmann,S.A., Marioni,J.C. *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, **10**, 1093–1095.
8. Hou,W., Ji,Z., Ji,H. and Hicks,S.C. (2020) A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.*, **21**, 218.
9. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 296.
10. Choi,K., Chen,Y., Skelly,D.A. and Churchill,G.A. (2020) Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.*, **21**, 183.
11. Dueck,H.R., Ai,R., Camarena,A., Ding,B., Dominguez,R., Evgrafov,O.V., Fan,J.-B., Fisher,S.A., Herstein,J.S., Kim,T.K. *et al.* (2016) Assessing characteristics of RNA amplification methods for single cell RNA sequencing. *BMC Genomics*, **17**, 966.
12. Zucha,D., Androvic,P., Kubista,M. and Valihrach,L. (2020) Performance comparison of reverse transcriptases for single-cell studies. *Clin. Chem.*, **66**, 217–228.

13. Bogdanova,E.A., Shagin,D.A. and Lukyanov,S.A. (2008) Normalization of full-length enriched cDNA. *Mol. BioSyst.*, **4**, 205.
14. Zhulidov,P.A., Bogdanova,E.A., Shcheglov,A.S., Shagina,I.A., Wagner,L.L., Khazpekov,G.L., Kozhemyako,V.V., Lukyanov,S.A. and Shagin,D.A. (2005) A method for the preparation of normalized cDNA libraries enriched with full-length sequences. *Russ. J. Bioorg. Chem.*, **31**, 170–177.
15. Kooiker,M. and Xue,G.-P. (2014) cDNA Library Preparation. In: Henry,R.J. and Furtado,A. (eds). *Cereal Genomics, Methods in Molecular Biology*. Humana Press, Totowa, NJ, Vol. **1099**, pp. 29–40.
16. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
17. Li,W.V. and Li,J.J. (2019) A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics*, **35**, i41–i50.
18. Zhang,X., Xu,C. and Yosef,N. (2019) Simulating multiple faceted variability in single cell RNA sequencing. *Nat. Commun.*, **10**, 2611.
19. Kim,J.K., Kolodziejczyk,A.A., Ilicic,T., Teichmann,S.A. and Marioni,J.C. (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.*, **6**, 8687.
20. Marinov,G.K., Williams,B.A., McCue,K., Schroth,G.P., Gertz,J., Myers,R.M. and Wold,B.J. (2014) From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.*, **24**, 496–510.
21. Ramsköld,D., Luo,S., Wang,Y.-C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.
22. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
23. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
24. McCarthy,D.J., Campbell,K.R., Lun,A.T.L. and Wills,Q.F. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.
25. L. Lun,A.T., Bach,K. and Marioni,J.C. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.
26. Soneson,C. and Robinson,M.D. (2017) Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data Bioinformatics.
27. Deng,Q., Ramskold,D., Reinius,B. and Sandberg,R. (2014) Single-cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
28. Grün,D., Muraro,M.J., Boisset,J.-C., Wiebrands,K., Lyubimova,A., Dharmadhikari,G., van den Born,M., van Es,J., Jansen,E., Clevers,H. *et al.* (2016) De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, **19**, 266–277.
29. Guo,F., Yan,L., Guo,H., Li,L., Hu,B., Zhao,Y., Yong,J., Hu,Y., Wang,X., Wei,Y. *et al.* (2015) The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell*, **161**, 1437–1452.
30. Shalek,A.K., Satija,R., Shuga,J., Trombetta,J.J., Gennert,D., Lu,D., Chen,P., Gertner,R.S., Gaublomme,J.T., Yosef,N. *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.
31. Chu,L.-F., Leng,N., Zhang,J., Hou,Z., Mamott,D., Vereide,D.T., Choi,J., Kendziorski,C., Stewart,R. and Thomson,J.A. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, **17**, 173.
32. Islam,S., Kjallquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lonnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–1167.
33. Picelli,S., Björklund,Å.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, **10**, 1096–1098.
34. Hagemann-Jensen,M., Ziegenhain,C., Chen,P., Ramsköld,D., Hendriks,G.-J., Larsson,A.J.M., Faridani,O.R. and Sandberg,R. (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.*, **38**, 708–714.
35. participants in the 1st Human Cell Atlas Jamboree, Lun,A.T.L., Riesenfeld,S., Andrews,T., Dao,T.P., Gomes,T. and Marioni,J.C. (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, **20**, 63.
36. Velculescu,V.E., Madden,S.L., Zhang,L., Lash,A.E., Yu,J., Rago,C., Lal,A., Wang,C.J., Beaudry,G.A., Ciriello,K.M. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.*, **23**, 387–388.
37. Ståhlberg,A. and Kubista,M. (2014) The workflow of single-cell expression profiling using quantitative real-time PCR. *Expert Rev. Mol. Diagn.*, **14**, 323–331.
38. McInnes,L., Healy,J. and Melville,J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv doi: https://arxiv.org/abs/1802.03426v2, 18 September 2018, preprint: not peer reviewed.
39. Maaten,L. van der and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
40. Sun,X., Liu,Y. and An,L. (2020) Ensemble dimensionality reduction and feature gene extraction for single-cell RNA-seq data. *Nat. Commun.*, **11**, 5853.
41. Schlitzer,A., Sivakamasundari,V., Chen,J., Sumatoh,H.R.B., Schreuder,J., Lum,J., Malleret,B., Zhang,S., Larbi,A., Zolezzi,F. *et al.* (2015) Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. *Nat. Immunol.*, **16**, 718–728.
42. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.-C. and Müller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
43. Qiu,P. (2020) Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.*, **11**, 1169.
44. Tang,F., Barbacioru,C., Nordman,E., Li,B., Xu,N., Bashkirov,V.I., Lao,K. and Surani,M.A. (2010) RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.*, **5**, 516–535.
45. Barron,M. and Li,J. (2016) Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Sci. Rep.*, **6**, 33892.
46. Hsiao,C.J., Tung,P., Blischak,J.D., Burnett,J.E., Barr,K.A., Dey,K.K., Stephens,M. and Gilad,Y. (2020) Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Res.*, **30**, 611–621.
47. Quail,M.A., Swerdlow,H. and Turner,D.J. (2009) Improved protocols for the illumina genome analyzer sequencing system. *Curr. Protoc. Hum. Genet.*, **62**, 18.2.1–18.2.27.
48. Sanders,J.G., Nurk,S., Salido,R.A., Minich,J., Xu,Z.Z., Zhu,Q., Martino,C., Fedarko,M., Arthur,T.D., Chen,F. *et al.* (2019) Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol.*, **20**, 226.
49. Buchbender,A., Mutter,H., Sutandy,F.X.R., Körtel,N., Hänel,H., Busch,A., Ebersberger,S. and König,J. (2020) Improved library preparation with the new iCLIP2 protocol. *Methods*, **178**, 33–48.
50. Kivioja,T., Vähärautio,A., Karlsson,K., Bonke,M., Enge,M., Linnarsson,S. and Taipale,J. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
51. Lundin,S., Stranneheim,H., Pettersson,E., Klevebring,D. and Lundeberg,J. (2010) Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One*, **5**, e10029.
52. The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.