SCnorm: robust normalization of single-cell RNA-seq data

Rhonda Bacher[1,5], Li-Fang Chu[2,5], Ning Leng[2], Audrey P Gasch[3], James A Thomson[2], Ron M Stewart[2], Michael Newton[1,4] & Christina Kendziorski[4]

The normalization of RNA-seq data is essential for accurate downstream inference, but the assumptions upon which most normalization methods are based are not applicable in the single-cell setting. Consequently, applying existing normalization methods to single-cell RNA-seq data introduces artifacts that bias downstream analyses. To address this, we introduce SCnorm for accurate and efficient normalization of single-cell RNA-seq data.

Methods used to quantify mRNA abundance introduce systematic sources of variation that can obscure signals of interest. Consequently, an essential first step in most mRNA-expression analyses is normalization, whereby systematic variations are adjusted to make expression counts comparable across genes and/or samples. Within-sample normalization methods adjust for gene-specific features, such as GC content and gene length, to facilitate comparisons of a gene's expression within an individual sample; whereas between-sample normalization methods adjust for sample-specific features, such as sequencing depth, to allow for comparisons of a gene's expression across samples[1]. In this work, we present a method for between-sample normalization, although we note that the R implementation of our method, R/SCnorm, also allows gene-specific features to be adjusted (**Supplementary Software** and http://www.biostat.wisc.edu/~kendzior/SCNORM/).

A number of methods are available for between-sample normalization in bulk RNA-seq experiments[2,3]. Most of these methods calculate global scale factors (one factor is applied to each sample, and this same factor is applied to all genes in the sample) to adjust for sequencing depth. These methods demonstrate excellent performance in bulk RNA-seq, but they are compromised in the single-cell setting because of an abundance of zero-expression values and increased technical variability[4].

Recent methods have been developed specifically for single-cell RNA-seq (scRNA-seq) normalization[5,6]. Like bulk methods, they calculate global scale factors, and therefore they cannot accommodate a major bias in scRNA-seq data that has not been recognized and reported in previous studies. Specifically, scRNA-seq data show systematic variation in the relationship between transcript-specific expression and sequencing depth (which we refer to as the count–depth relationship) that is not accommodated by a single scale factor common to all genes in a cell (**Fig. 1** and **Supplementary Fig. 1**). Global scale factors adjust for a count–depth relationship that is assumed to be common across genes. When this relationship is not common across genes, normalization via global scale factors leads to overcorrection for weakly and moderately expressed genes and, in some cases, undernormalization of highly expressed genes (**Fig. 1**).

To address this, SCnorm uses quantile regression to estimate the dependence of transcript expression on sequencing depth for every gene. Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group. Within-group adjustment for sequencing depth is then performed using the estimated scale factors to provide normalized estimates of expression. Although SCnorm does not require experimental RNA spike-ins, performance may be improved if spike-ins that span the range of expression observed in endogenous genes are available (**Supplementary Note 1**).

We evaluated SCnorm and compared it with MR[3], transcripts per million (TPM)[7], scran[5], SCDE[8], and BASiCS[6] using simulated and case study data. In the first simulation (SIM I), two scenarios are considered where the number of gene groups having different count–depth relationships (K) is set to one (to mimic a bulk experiment) or four (**Supplementary Fig. 2**). Each simulated data set contains two conditions, the second condition having approximately four times as many reads as the first; 20% of the genes are defined to be differentially expressed (DE). Prior to normalization, counts in the second condition will appear four times higher on average given the increased sequencing depth. If normalization for depth is effective, fold-change estimates should be near one, and only simulated DE genes should appear to be DE. When $K = 1$, with the exception of TPM, fold-change estimates are consistently robust among methods (**Supplementary Fig. 2a**), and all normalization methods provide data that result in high sensitivity and specificity for identifying DE genes (**Supplementary Fig. 2b**). However, when $K = 4$, only SCnorm maintains good operating characteristics, whereas approaches based on global scale factors overestimate fold changes for weakly to moderately expressed genes on account of overcorrection of sequencing depth (**Supplementary Fig. 2c,d**).

In the second simulation (SIM II) counts are generated as in Lun *et al.*[5], following their simulation study scenarios 1, 2, 3, and 4. Briefly, scenario 1 contains no DE genes; scenarios 2, 3, and 4 contain moderate DE, strong DE, and varying magnitudes of DE genes, respectively. We found that SCnorm is similar to scran with

[1]Department of Statistics, University of Wisconsin–Madison, Madison, Wisconsin, USA. [2]Morgridge Institute for Research, Madison, Wisconsin, USA. [3]Laboratory of Genetics, University of Wisconsin–Madison, Madison, Wisconsin, USA. [4]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to C.K. (kendzior@biostat.wisc.edu).

RECEIVED 22 AUGUST 2016; ACCEPTED 22 MARCH 2017; PUBLISHED ONLINE 17 APRIL 2017; DOI:10.1038/NMETH.4263

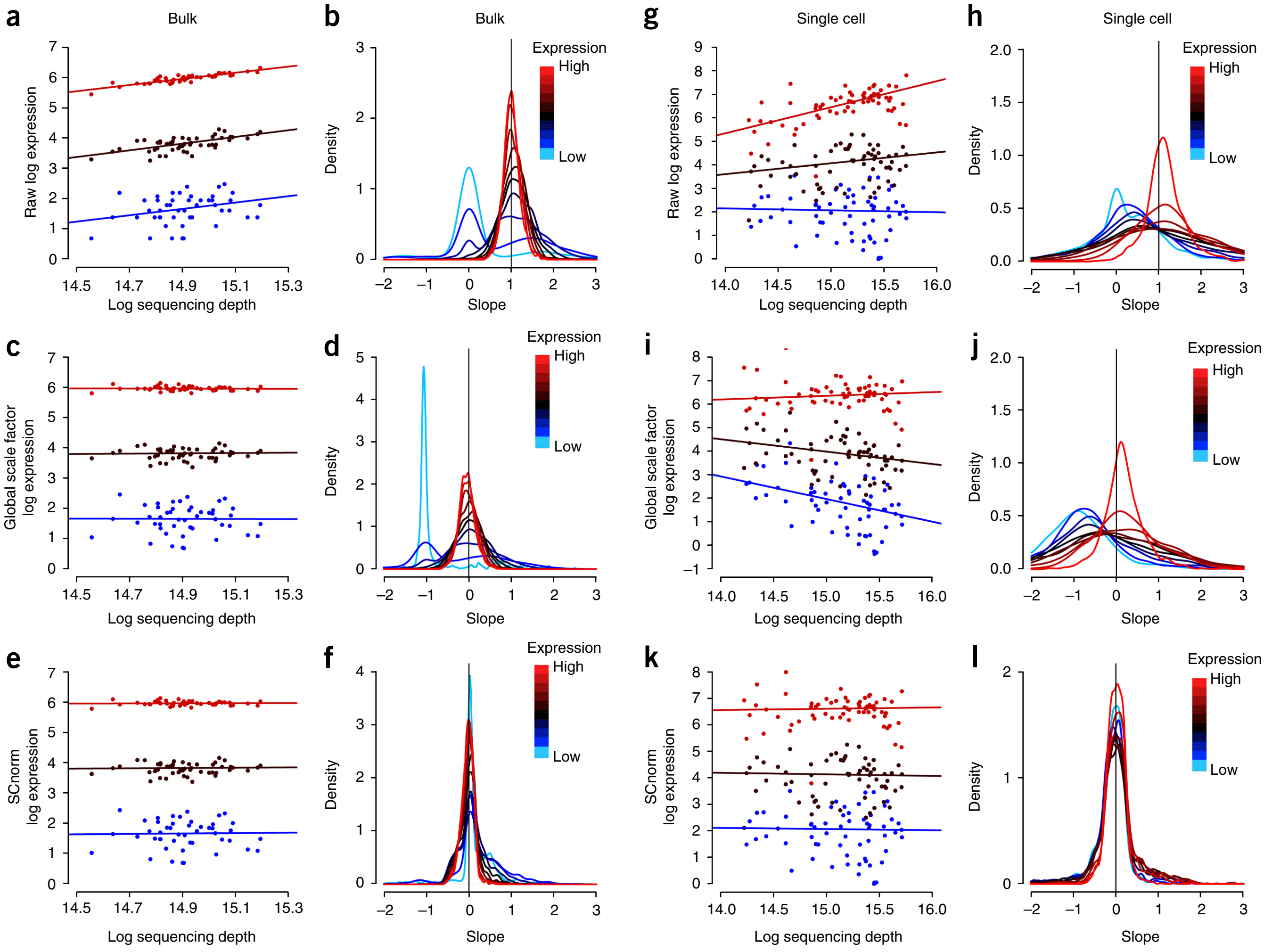

Figure 1 | Count–depth relationships in bulk and single-cell data sets before and after normalization. For each gene, median quantile regression was used to estimate the count–depth relationship before normalization and after normalization via MR or SCnorm for the H1 bulk RNA-seq data set (**a**–**f**) and the DEC scRNA-seq data set (**g**–**l**). (**a**) Log expression versus log depth and estimated regression fits for three genes containing no zero measurements and having low, moderate, and high expression defined as median expression among nonzero un-normalized measurements in the 10th to 20th quantile (blue), 40th to 50th quantile (black), and 80th to 90th quantile (red), respectively. (**b**) Densities of slopes within each of ten equally sized gene groups where a gene's group membership is determined by its median expression among nonzero un-normalized measurements. (**c**,**d**) The data in panels **a** and **b** normalized via MR and (**e**,**f**) by SCnorm. (**g**–**l**) Identical to **a**–**f** but for the DEC scRNA-seq data set. Qualitatively similar results are observed if slopes are calculated via generalized linear models (**Supplementary Note 2** and **Supplementary Fig. 1**).

respect to fold-change estimation and retains relatively high sensitivity and specificity for identifying DE genes (**Supplementary Fig. 3**).

To further evaluate SCnorm, we conducted an experiment that, similar to the simulations, sequenced cells at very different depths. We used the Fluidigm C1 system to capture 92 H1 human embryonic stem cells (hESCs). Each cell's fragmented, indexed cDNA was split into two groups before pooling for sequencing. The first group (H1-1M) was pooled at 96 cells per lane and the second (H1-4M) at 24 cells per lane, resulting in approximately 1 million and 4 million mapped reads per cell in the two groups, respectively. Prior to normalization, counts in the second group will appear four times higher on average given the increased sequencing depth. However, if normalization for depth is effective, fold-change estimates should be near one; and all genes should appear to be EE, since the cells between the two groups are identical. SCnorm provides normalized data that result in fold-change estimates near one, whereas other methods show biased estimates (**Fig. 2a**).

To evaluate the extent to which biases introduced during normalization affect the identification of DE genes, we applied MAST[9] (false discovery rate, FDR = 0.05) to identify genes that are DE between the H1-1M and H1-4M conditions. Normalization with SCnorm resulted in the identification of no DE genes; whereas normalization with MR, TPM, scran, SCDE, and BASiCS resulted in the identification of 530; 315; 684; 401; and 1,147 DE genes, respectively. The majority of DE calls made using data normalized from these latter approaches are weakly expressed genes (**Fig. 2b**), which appear to be overnormalized (**Fig. 2a**; see **Supplementary Fig. 4** for similar results using H9 cells).

We also evaluated the impact of normalization on downstream analyses such as principal component analysis (PCA) and on the identification of DE genes in case study data. Specifically, we considered the H1-FUCCI data from Leng *et al.*[10] where 247 H1

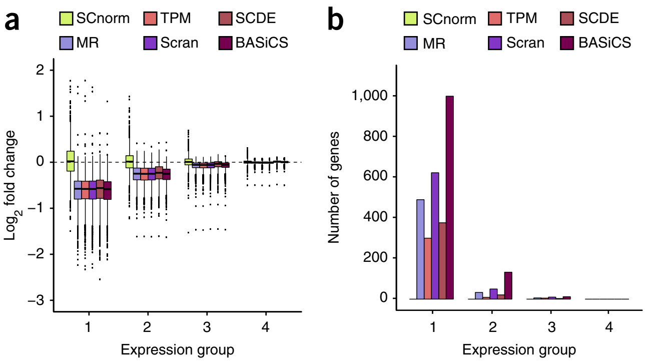

Figure 2 | SCNorm removes bias from fold-change estimates. Fold changes and DE genes were calculated from the H1 case study data. For each gene, the fold change of nonzero counts between the H1-4M and H1-1M groups was computed for data following normalization via SCnorm, MR, TPM, scran, SCDE, and BASiCS. (**a**) Gene-specific fold changes for data normalized by each method. (**b**) Number of genes identified as DE using MAST. Genes are divided into four equally sized expression groups based on their median among nonzero un-normalized expression measurements, and results are shown as a function of expression group (see **Supplementary Note 3** for why nonzero counts are used for fold-change calculations).

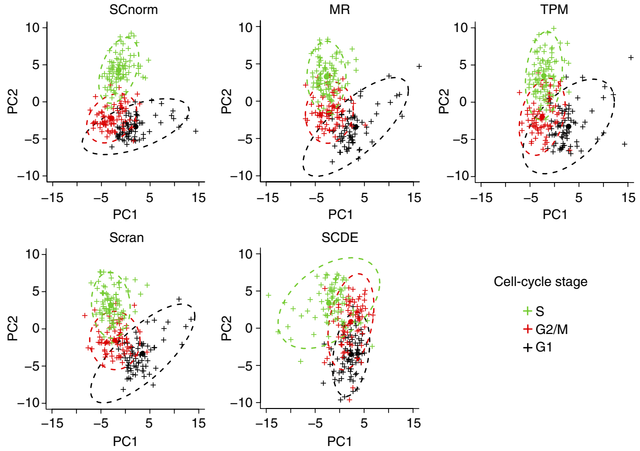

Figure 3 | Normalization by SCnorm improves researchers' ability to resolve cell populations. PCA applied to the H1-FUCCI case study. The upper left panel shows the first two principal components (PC1 versus PC2) from a PCA analysis using 578 cell-cycle genes normalized via SCnorm. The other panels show similar results for data normalized using MR, TPM, scran, and SCDE. Cells are colored according to cell-cycle phase. 95%-confidence ellipses are shown for each method. Misclassification rates for SCnorm, MR, TPM, scran, and SCDE averaged across the three cell-cycle phases are 0.26, 0.32, 0.38, 0.29, and 0.45, respectively.

hESCs were labeled with fluorescent ubiquitination-based cell-cycle indicators[11] to enable identification of cells as being in G1, S, or G2/M phase. PCA was applied to the H1-FUCCI data following normalization via SCnorm, MR, TPM, scran, and SCDE. SCnorm shows some advantage in distinguishing at least one of the groups and has the lowest misclassification rate (**Fig. 3**). As a second positive control, we evaluated the ability of each normalized data set to be used to identify DE genes (**Supplementary Fig. 5**). Specifically, we considered the S and G2/M phases from the H1-FUCCI data. For these two phases, we subsampled cells so that there were negligible differences in cellular detection rates (CDRs) between the two conditions, and on average there was a 1.5-fold increase in sequencing depth. Without differences in CDR, we would expect an EE gene expressed at level x in S to be expressed at level $1.5x$ in G2/M. Given this, we defined a gold standard list to be those genes showing a fold change bigger than a threshold (or smaller than one over that threshold) for varying thresholds, adjusting for the expected increase in expression caused by increased sequencing depth. SCnorm provides improved sensitivity over other methods (**Supplementary Fig. 5**).

We also evaluated the performance of SCnorm on a number of other case study data sets. For these evaluations, a data set was considered well normalized if the relationship between counts and depth was negligible following normalization. SCnorm allows for robust normalization of scRNA-seq data when the count–depth relationship is common across genes, as in a bulk RNA-seq experiment (or a deeply sequenced scRNA-seq experiment); and SCnorm outperforms other approaches when this relationship varies systematically, as in a typical scRNA-seq experiment (**Fig. 1** and **Supplementary Figs. 6–11**).

Single-cell RNA-seq technology offers an unprecedented opportunity to address important biological questions, but accurate data normalization is required to ensure that results are meaningful. Our approach allows investigators to accurately normalize data for sequencing depth and improve downstream inference.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by NIH GM102756 (C.K.), NIH U54 AI117924 (C.K. and M.N.), 1T32LM012413-01A1 (M.N.), NIH 5U01HL099773 (J.A.T.), and the Morgridge Institute for Research. We thank J. Bolin, A. Elwell, and B.K. Nguyen for the preparation and sequencing of the RNA-seq samples and P. Jiang and S. Swanson for performing the RNA-seq read processing.

AUTHOR CONTRIBUTIONS

R.B. and C.K. designed the research, developed the method, and wrote the first version of the manuscript. L.-F.C. performed experiments and quality control on scRNA-seq data generated from H1 and H9 hESCs. R.B. analyzed all data sets. L.C., N.L., A.P.G., J.A.T., R.M.S., and M.N. analyzed results from early versions of the method, which helped during method refinement. All authors contributed to the writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Conesa, A. *et al. Genome Biol.* **17**, 13 (2016).
2. Robinson, M.D. & Oshlack, A. *Genome Biol.* **11**, R25 (2010).
3. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
4. Bacher, R. & Kendziorski, C. *Genome Biol.* **17**, 63 (2016).
5. Lun, A.T., Bach, K. & Marioni, J.C. *Genome Biol.* **17**, 75 (2016).
6. Vallejos, C.A., Marioni, J.C. & Richardson, S. *PLoS Comput. Biol.* **11**, e1004333 (2015).
7. Li, B. & Dewey, C.N. *BMC Bioinformatics* **12**, 323 (2011).
8. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. *Nat. Methods* **11**, 740–742 (2014).
9. Finak, G. *et al. Genome Biol.* **16**, 278 (2015).
10. Leng, N. *et al. Nat. Methods* **12**, 947–950 (2015).
11. Sakaue-Sawano, A. *et al. Cell* **132**, 487–498 (2008).

ONLINE METHODS

Filter. Genes without at least ten cells having nonzero expression were removed before all analyses. They are not shown in plots.

SCnorm. SCnorm requires estimates of expression, but it is not specific to one approach. Estimates may be obtained via RSEM[7], HTSeq[12], or any method providing un-normalized counts per feature. Let $Y_{g,j}$ denote the log nonzero expression count for gene g in cell j for $g = 1,\ldots,m$ and $j = 1,\ldots,n$; X_j denote log sequencing depth for cell j. Motivation for considering nonzero counts is provided in **Supplementary Note 3** and **Supplementary Figures 12** and **13**.

The number of groups for which the count–depth relationship varies substantially, K, is chosen sequentially. SCnorm begins with $K = 1$. For each gene, the gene-specific relationship between log un-normalized expression and log sequencing depth is represented by $\hat{\beta}_{g,1}$ using median quantile regression with a first degree polynomial: $Q^{0.5}\left(Y_{g,j} \mid X_j\right) = \beta_{g,0} + \beta_{g,1} X_j$. The overall relationship between log un-normalized expression and log sequencing depth for all genes in the $K = 1$ group is also estimated via quantile regression. Since the median might not best represent the full set of genes within the group, and since multiple genes allow for estimation of somewhat subtle effects, in this step SCnorm considers multiple quantiles τ and multiple degrees d:

$$Q^{\tau_k, d_k}\left(Y_j \mid X_j\right) = \beta_0^{\tau_k} + \beta_1^{\tau_k} X_j + \ldots + \beta_d^{\tau_k} X_j^{d_k}$$

The specific values of τ_k and d_k, τ_k^* and d_k^*, are those that minimize $\left|\hat{\eta}_1^{\tau_k} - \overset{mode}{\underset{g}{}}\hat{\beta}_{g,1}\right|$, where $\hat{\eta}_1^{\tau_k}$ represents the count–depth relationship among the predicted expression values as estimated by median quantile regression using a first-degree polynomial: $Q^{0.5}\left(\hat{Y}_j^{\tau_k} \mid X_j\right) = \eta_0^{\tau_k} + \eta_1^{\tau_k} X_j$. Scale factors for each cell are defined as

$$SF_j = \frac{e^{\widehat{Y_j}^{\tau_k^*, d_k^*}}}{e^{Y^{\tau_k^*}}}$$

where $Y^{\tau_k^*}$ is the $\tau^{*\text{th}}$ quantile of expression counts in the k^{th} group. Normalized counts $Y'_{g,j}$ are given by

$$\frac{e^{Y_{g,j}}}{\mathrm{SF}_j}$$

To determine if $K = 1$ is sufficient, the gene-specific relationship between log normalized expression and log sequencing depth is represented by the slope of a median quantile regression using a first-degree polynomial as detailed above. $K = 1$ is considered sufficient if the modes of the slopes within each of ten equally sized gene groups (where a gene's group membership is determined by its median expression among nonzero un-normalized measurements) are all less than 0.1. Any mode exceeding 0.1 is taken as evidence that the normalization provided with $K = 1$ is not sufficient to adjust for the count–depth relationship for all genes and, consequently, K is increased by one and the count–depth relationship is estimated within each of the K groups using equation (1). For each increase, the K-medoids algorithm is used to cluster genes into groups based on $\hat{\beta}_{g,1}$; if a cluster has less than 100 genes, it is joined with the nearest cluster.

When multiple biological conditions are present, SCnorm is applied within each condition, and the normalized counts are then rescaled across conditions. During rescaling, all genes are split into quartiles based on median expression among nonzero un-normalized measurements. Within each group and condition, each gene is scaled by a common scale factor defined as the median of the gene-specific fold changes between each gene's condition-specific mean and the gene-specific mean across conditions, where means are calculated over nonzero counts. Motivation for considering nonzero counts during rescaling is discussed in **Supplementary Note 3**. Although the focus of SCnorm is on between-sample normalization, gene-specific features may also be adjusted using the R/SCnorm package. As in Risso *et al.*[13], we implemented a two-step procedure where gene-specific effects may be adjusted for before between-sample normalization using SCnorm. It should be noted that SCnorm is not designed to adjust for batch effects; methods such as ComBat[14] or sva[15] may be used for this purpose following normalization.

SCnorm.SI. SCnorm does not require spike-ins, since we find that the performance of spike-ins in scRNA-seq is often compromised (**Supplementary Figs. 14** and **15**), and many labs do not use them for normalization[16,17]. However, if good spike-ins are available, performance of SCnorm may be improved in the post-normalization scaling step, which is required when multiple conditions are available. Recall that in SCnorm, during rescaling, all genes are split into quartiles based on median expression among nonzero un-normalized measurements. In SCnorm.SI, the same is done with spike-ins and, if the spike-ins are representative of the full range of expression, we expect them to be approximately evenly divided among the four groups. Within each group and condition, each gene is scaled by a common scale factor defined as the median of the spike-in-specific fold changes between each spike-in's condition-specific mean and the spike-in's specific mean across conditions, where means are calculated over nonzero counts. For more on SCnorm.SI, see **Supplementary Note 1**.

Application of comparable methods. All analyses were carried out using R version 3.3.0 unless otherwise noted. The method MR, originally described by Anders and Huber[3], was implemented using the DESeq R package version 1.24.0 using the default settings of the estimateSizeFactorsForMatrix function. TPM estimates were obtained as output from RSEM version 1.2.3. Expected counts were used in SCnorm and TPM was evaluated separately. The method scran was implemented with the scran R package version 1.0.0; size factors were obtained using the function computeSumFactors. The pool sizes were set to 5, 10, 15, and 20; and size factors were constrained to be positive. SCDE was implemented in R version 3.2.2 using the SCDE R package version 1.99.1 with default parameter settings, and normalized counts were obtained using the function scde.expression.magnitude. BASiCS was implemented using the BASiCS R package version 0.4.1 using R version 3.2.2, obtained from Github at https://github.com/catavallejos/BASiCS; and normalized expression estimates were obtained using the function BASiCS_DenoisedCount, where BASiCS_MCMC was run with $N = 20{,}000$; Burn = 10,000; and default parameters were used otherwise. Because BASiCS requires spike-ins, results are only shown for data sets where spike-ins are available. Finally, we also evaluated NODES[18] (**Supplementary Figs. 16–18**), an unpublished approach, version 0.0.0.9010.

Evaluation of methods. Gene-specific count–depth relationships were estimated using median quantile regression as well as regression with a negative binomial generalized linear model (glm). The quantreg package in R was used with the Barrodale and Roberts algorithm to carry out the median regressions; MASS in R was used to fit the glms. Zeros are not included in the fits since our goal is to estimate the count–depth relationship present in data before and after normalization; and that relationship is obscured by dropouts, which are largely technical. Because glms are sensitive to outliers, an initial glm to estimate the count–depth relationship is fit on the un-normalized data, and the top two and bottom two residual gene expression values were removed from each gene before estimating the final count–depth relationship via glm. Since the same set of putative outliers was removed for every method, excluding these values will not bias results in favor of any one method.

MAST was used to identify DE genes, using the MAST R package version 0.933, obtained from Github at https://github.com/RGLab/MAST. The continuous component test was considered; and differential zeros were not used to evaluate performance of normalization methods, since all normalization methods leave zeros un-normalized. *P* values from MAST were adjusted using Benjamini and Hochberg's method[19]. Unless otherwise noted, a DE gene was defined as a gene with corrected *P* value <0.05, which controls the false discovery rate at 5%. ROC curves were plot using the R package ROCR. The false positive and true positive rates were calculated by ROCR, with a positive representing a DE gene. Average ROC curves show the average true positive rate. PCA was conducted using the prcomp function in R, and confidence ellipses were drawn using the dataEllipse function in the car package in R. Outlier adjustment (values in the upper 0.995th percentile were set to the 0.995th percentile) was done before applying PCA for each data set. The misclassification rate for the S phase was calculated as the percentage of G1 or G2/M cells present within the 95% confidence ellipse for S; misclassification rates for the other phases were calculated similarly.

Simulation SIM I. Data were simulated to match characteristics of the H1-1M and H1-4M data sets. For each gene *g*, gene-specific intercepts $\hat{\beta}_{g,0}$, slopes $\hat{\beta}_{g,1}$, and variance intercepts $\widehat{\sigma^2}_g$ were estimated using median quantile regression on the H1-1M data. Two SIM I simulation scenarios were generated: $K = 1$ and $K = 4$. In the $K = 1$ simulations, only genes having at least 75% nonzero expression values and $\hat{\beta}_{g,1} \in (.9,1.1)$ were used. For the $K = 4$ simulations, genes were split into four equally sized groups based on $\hat{\beta}_{g,1}$. The medians of $\hat{\beta}_{g,1}$ were calculated within each group; these were denoted by $\beta_{\text{med},1}$, $\beta_{\text{med},2}$, $\beta_{\text{med},3}$, and $\beta_{\text{med},4}$, respectively. For genes in the k^{th} group, genes having $\hat{\beta}_{g,k} \in (\beta_{\text{med},k} - 0.1,\ \beta_{\text{med},k} + 0.1)$ were used, where $\beta_{\text{med,k}}$ is the median $\hat{\beta}_{g,k}$ over all genes.

For a given gene, counts were simulated on the log scale as

$$\hat{\beta}_{g,1} \log\left(X_j\right) + \hat{\beta}_{g,0} + \epsilon_{g,j}$$

and then exponentiated, where

$$\epsilon_{g,j} \sim N\left(0, \widehat{\sigma^2}_g\right).$$

Two biological conditions were simulated: one condition with 90 cells simulated from sequencing depths ranging from 500,000 to 1.5 million reads (X_j was sampled uniformly between 500,000 and 1.5 million) and a second condition with 90 cells simulated with depths ranging from 2 to 6 million reads (X_j was sampled uniformly between 2 and 6 million). For a randomly selected set of cells, counts were set to zero, where the proportion set to zero was defined to match the proportion observed empirically. Each simulated data set contained 1,200 genes—80% EE and 20% DE. For approximately half of the DE genes, fold changes were sampled uniformly between 2 and 4, and counts in the second condition were multiplied by the sampled fold change. The other (approximately) half of DE genes were simulated similarly, but counts in the first condition were multiplied by the sampled fold change to keep the DE balanced. **Supplementary Figure 19** shows that basic summary statistics are well preserved between the simulated and case study data.

Simulation SIM II. Counts are generated as in Lun *et al.*[5] following their simulation study scenarios 1, 2, 3, and 4. In that simulation setup, three populations were simulated. We here consider populations 1 and 2.

H1 bulk data. The data set contains 48 samples of H1 hESCs as described in detail in Hou *et al.*[20]. The H1 bulk RNA-seq data have an average sequencing depth of 3 million mapped reads per sample.

H1 and H9 case studies. Undifferentiated H1 or H9 hESCs were cultured in E8 medium[21] on Matrigel-coated tissue-culture plates with daily media feeding at 37 °C with 5% (v/v) CO_2. Cells were split every 3–4 d with 0.5 mM EDTA in 1× PBS for standard maintenance. Immediately before preparing single-cell suspensions for each experiment, hESCs were individualized by Accutase (Life Technologies), washed once with E8 medium, and resuspended at densities of $5.0–8.0 \times 10^5$ cells/mL in E8 medium for cell capture. The H1 hESCs are registered in the NIH Human Embryonic Stem Cell Registry with approval number NIHhESC-10-0043. Details of the H1 cells can be found online (http://grants.nih.gov/stem_cells/registry/current.htm?id=29). The H9 hESCs are registered in the NIH Human Embryonic Stem Cell Registry with approval number NIHhESC-10-0062. Details of the H9 cells can be found online (http://grants.nih.gov/stem_cells/registry/current.htm?id=414). All the cell cultures performed in our laboratory have been routinely tested and have been found negative for mycoplasma contamination and authenticated by cytogenetic tests.

Single-cell loading, capture, and library preparations were performed following the Fluidigm user manual[22]. Briefly, 5,000–8,000 cells were loaded onto a medium-size (10–17 μm) C1 Single-Cell Auto Prep IFC (Fluidigm), and cell-loading script was performed according to the manufacturer's instructions. The capture efficiency was inspected using EVOS FL Auto Cell Imaging system (Life Technologies) to perform an automated area scanning of the 96 capture sites on the IFC. Empty capture sites or sites having more than one cell captured were first noted, and those samples were later excluded from further library processing for RNA-seq. Immediately after capture and imaging, reverse transcription and cDNA amplification were performed in the C1 system using the SMARTer PCR cDNA Synthesis Kit (Clontech) and the Advantage 2 PCR kit (Clontech) according to the instructions in the Fluidigm user manual. Full-length, single-cell cDNA libraries

doi:10.1038/nmeth.4263

were harvested the next day from the C1 chip and diluted to a range of 0.1–0.3 ng/μL. Diluted single-cell cDNA libraries were fragmented and amplified using the Nextera XT DNA Sample Preparation Kit and the Nextera XT DNA Sample Preparation Index Kit (Illumina). Libraries were multiplexed either at 24 or 96 single-cell cDNA libraries per lane to target 4 or 1 million mapped reads per cell, respectively, and single-end reads of 67 bp were sequenced on an Illumina HiSeq 2500 system. We refer to the data obtained from 24 libraries per lane as the H1-4M set, since approximately 4 million mapped reads per cell were generated. For similar reasons, H1-1M is used to refer to the data obtained from 96 libraries per lane.

Reads were mapped against the Hg19 Refseq reference via Bowtie 0.12.8 (ref. 23), allowing up to two mismatches and up to 20 multiple hits. The expected counts and TPMs were estimated via RSEM 1.2.3 (ref. 7). Cells that had less than than 5,000 genes with expected counts >1 or that upon inspection of cell images displayed doublets or appeared dead were removed in quality control. 92 H1 cells passed the quality control. 91 H9 cells passed quality control.

H1-FUCCI case study. Single-cell RNA-seq data were downloaded from GSE64016 (ref. 10). In this experiment, 247 H1 human embryonic stem cells were labeled with fluorescent ubiquitination-based cell-cycle indicators[11] to enable identification of cell-cycle phase for each cell. For the PCA analysis, cell-cycle genes were defined from GO:0007049 and from Cyclebase[24]. Specifically, we took genes from GO:0007049 that showed strong evidence of cell-cycle association by having a rank within the top 400 Cyclebase genes (giving a total of 578 genes). For the S versus G2/M DE analysis, we sampled 50 cells from the S phase and 50 cells from the G2/M phase to match on cellular detection rate (CDR). For the S condition, the 25th, 50th, and 75th percentiles of CDR were 0.62, 0.63, and 0.64, respectively; for the G2/M condition they were 0.61, 0.63, and 0.64, respectively. Sequencing depth was approximately 1.5× higher in the G2/M condition (4 million reads on average in S and 6 million on average in G2/M; medians 4.05 and 6.1 million reads, respectively). Without differences in CDR, we would expect an EE gene expressed at level x in S to be expressed at level $1.5x$ in G2/M. Given this, we define a gold standard list to be those genes showing a fold change bigger than a threshold (or smaller than one over that threshold) for varying thresholds, adjusting for the expected increase in expression due to increased sequencing depth. For example, genes with two-fold change or greater are defined as those with empirical fold change of three or greater.

Buettner case study. Single-cell RNA-seq expression data were downloaded from ArrayExpress E-MTAB-2805 (ref. 25). In this experiment, *Mus musculus* embryonic stem cells were sorted using fluorescence-activated cell sorting (FACS) to determine cell-cycle phase; cells were then captured using the C1 Fluidigm system. Libraries were multiplexed and sequenced across four lanes using an Illumina HiSeq 2000 system. Gene-level read counts were generated by HTSeq version 0.6.1. Here we consider the three data sets, each of which had 96 cells in either G1, S, or G2M phase of the cell cycle. The data have average sequencing depths of 4.9, 6.5, and 4.5 million, respectively. Cells with sequencing depths less than 10,000 were removed before analysis, which resulted in 95 G1, 88 S, and 96 G2M cells.

Islam case study. Single-cell RNA-seq expression data were downloaded from GEO GSE29087 (ref. 26). In this experiment, *Mus musculus* R1 embryonic stem cells (ES) and embryonic fibroblasts were captured using a semiautomated cell picker on a 96-well capture plate; libraries were generated using the STRT protocol and sequenced using on a Genome Analyzer IIx system. Gene-level counts were obtained by counting reads mapped using Bowtie[23] for each feature. Here we consider two data sets—one having 48 ES cells, and the other having 44 EF cells. These data sets have average sequencing depths of 180,000 reads and 800,000 reads, respectively.

DEC case study. The data set contains 64 H1 cells consisting of the first batch of experiments studying H1 differentiation toward definitive endodermal cells as described in detail in Chu *et al.*[27]. The DEC scRNA-seq data have an average sequencing depth of 4 million mapped reads per cell. The data can be downloaded from GEO GSE75748.

Data availability statement. The H1 bulk and the H1-1M, H1-4M, H9-1M, H9-4M case study data sets are available at the NCBI Gene Expression Omnibus: GSE85917. The R package R/SCnorm is available at http://www.biostat.wisc.edu/~kendzior/SCNORM/.

12. Anders, S., Pyl, P.T. & Huber, W. *Bioinformatics* **31**, 166–169 (2015).
13. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. *BMC Bioinformatics* **12**, 480 (2011).
14. Johnson, W.E., Li, C. & Rabinovic, A. *Biostatistics* **8**, 118–127 (2007).
15. Leek, J.T. & Storey, J.D. *PLoS Genet.* **3**, 1724–1735 (2007).
16. Lin, Y. *et al. BMC Genomics* **17**, 28 (2016).
17. McDavid, A., Finak, G. & Gottardo, R. *Nat. Biotechnol.* **34**, 591–593 (2016).
18. Sengupta, D., Rayan, N.A., Lim, M., Lim, B. & Prabhakar, S. Preprint at http://biorxiv.org/content/early/2016/04/22/049734 (2016).
19. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. B Stat. Methodol.* **57**, 289–300 (1995).
20. Hou, Z. *et al. Sci. Rep.* **5**, 9570 (2015).
21. Chen, G. *et al. Nat. Methods* **8**, 424–429 (2011).
22. Fluidigm Corporation. *Using the C1 Single-Cell Auto Prep System to Generate mRNA from Single Cells and Libraries for Sequencing* (Fluidigm Corporation, 2017).
23. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
24. Santos, A., Wernersson, R. & Jensen, L.J. *Nucleic Acids Res.* **43**, D1140–D1144 (2015).
25. Buettner, F. *et al. Nat. Biotechnol.* **33**, 155–160 (2015).
26. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
27. Chu, L.-F. *et al. Genome Biol.* **17**, 173 (2016).

doi:10.1038/nmeth.4263