Published in final edited form as:
Nat Methods. 2017 June ; 14(6): 565–571. doi:10.1038/nmeth.4292.

Normalizing single-cell RNA sequencing data: Challenges and opportunities

Catalina A. Vallejos[#1,2,3,4], **Davide Risso**[#5,^], **Antonio Scialdone**[#2], **Sandrine Dudoit**[5,6,†], and **John C. Marioni**[2,7,8,†]

[1]MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom

[2]EMBL-European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom

[3]The Alan Turing Institute, British Library, London, United Kingdom

[4]Department of Statistical Science, University College London, London, United Kingdom

[5]Division of Biostatistics, School of Public Health, University of California, Berkeley, U.S.A.

[6]Department of Statistics, University of California, Berkeley, U.S.A.

[7]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, United Kingdom

[8]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, United Kingdom

[#] These authors contributed equally to this work.

Abstract

Single-cell transcriptomics is becoming an important component of the molecular biologist's toolkit. A critical step when analyzing this type of data is normalization. However, normalization is typically performed using methods developed for bulk RNA sequencing or even microarray data, whose suitability for single-cell transcriptomics has not been assessed. In this perspective, we discuss commonly used normalization approaches and illustrate how these can lead to misleading results. Finally, we present alternative approaches and provide recommendations for single-cell RNA sequencing users.

Single-cell RNA sequencing (scRNA-seq) has transformed the field of transcriptomics by making it possible to address fundamental questions that are inaccessible to bulk-level experiments [1]. Examples include the study of tumor heterogeneity, the identification of

[†]**Correspondence to**: sandrine@stat.berkeley.edu or marioni@ebi.ac.uk.
[^]Current address: Division of Biostatistics and Epidemiology, Department of Healthcare Policy and Research, Weill Cornell Medicine, 402 E 67th St, New York, NY 10065

Contributions
CAV, DR and AS performed analyses. CAV, DR, AS, SD and JCM wrote the manuscript. SD and JCM supervised the study.

Competing Financial Interests
The authors declare no competing financial interests.

Data Availability
We used datasets previously published in the referenced citations.

Europe PMC Funders Author Manuscripts

novel cell types, and the understanding of cell fate decisions during early embryo development [2–5].

Recent literature has highlighted the need for new computational methods to address the complex features that characterize scRNA-seq data, such as sparsity and technical noise [6–11]. Despite this, little attention has been given to normalization — a critical step in the analysis pipeline that adjusts for unwanted biological and technical effects that can mask the signal of interest. Instead, most tools developed for scRNA-seq rely on normalized expression measures obtained from methods developed for bulk RNA-seq or even microarray data. However, the suitability of such approaches for single-cell transcriptomics has not been rigorously discussed.

In this Perspective, we address normalization and focus on the most widely used strategy, global-scaling, which attempts to remove cell-specific systematic biases by scaling expression measures within each cell by a constant factor. Within this context, we illustrate that using bulk-based normalization methods can have serious adverse consequences for downstream analysis, such as the detection of highly variable genes prior to clustering. Such problems are exacerbated by the high levels of technical noise and dropout typical of scRNA-seq. We also discuss the use of extrinsic spike-in sequences (e.g. [12]) for normalization. To conclude, we summarize state-of-the-art methods for scRNA-seq normalization, including integrated strategies, where normalization is intrinsic to a specific method, and generic tools, which provide normalized data that can be used as input to any downstream analysis pipeline.

From bulk samples to single-cell resolution

Bulk microarray and RNA-seq experiments measure gene expression levels as averages across thousands of cells. While this allows the characterization of population-level differences in overall expression, single-cell-level experiments are required to better understand the dynamics of gene expression patterns. In particular, scRNA-seq experiments can reveal heterogeneity within populations of cells. However, the promise of scRNA-seq comes at the cost of more challenging experimental protocols [13] and higher data complexity [10].

A prominent feature of scRNA-seq is the sparsity of the data, i.e., the high proportion of *zero* read counts [7, 8, 14]. This so-called *zero inflation* arises for both biological reasons (e.g., subpopulations of cells or transient states where a gene is not expressed) and technical reasons (e.g., dropouts, where a gene is expressed but not detected through sequencing). Besides the dropout effect, technical noise in scRNA-seq is also reflected by high variability between technical replicates, even for genes with medium or high levels of expression [6]. Additionally, by capturing individual cells from potentially very different cell types, scRNA-seq data are highly heterogeneous, even in the absence of the technical biases discussed above. Consequently, several assumptions made when analyzing bulk RNA-seq data do not always apply in the context of scRNA-seq.

Systematic biases in scRNA-seq datasets

Data normalization strategies must capture biases that are specific to the technology of interest. For example, two-channel microarrays require normalization to account for differences in dye balance related to intensity and spatial position on the array [15]. By contrast, in sequencing assays, read counts must be adjusted to control for a variety of biases, including sequencing depth [16, 17]. Additionally, in any of these high-throughput assays, one needs to account for possibly more complex and putatively unknown effects, collectively known as “batch effects” [18–20].

While scRNA-seq analysis pipelines routinely include a normalization step, the sources of the systematic biases that this step captures are assay-specific. To illustrate this, we focus on the Illumina sequencing platform, with a simple experimental setup where gene expression is measured in a homogeneous population of cells. The discussion below applies to whole transcript scRNA-seq as well as to 3’ sequencing protocols and Unique Molecular Identifier (UMI) [21] based approaches that use barcodes to obtain molecular counts.

RNA-seq experiments are inherently stochastic, with reads being randomly sampled from a pool of amplified cDNA molecules. Typically, the quantity of interest is the *expression level* of each gene: the relative abundance of mRNA molecules within the population of mRNA molecules in each cell. There are several experimental sources of systematic biases that can affect measurements of gene expression, including gene- and cell-specific features (Fig. 1). Accordingly, we distinguish between two types of normalization: *within-sample* normalization, which removes gene-specific biases (e.g., GC-content); and *between-sample* normalization, which adjusts for effects related to distributional differences in read counts between cells (e.g., sequencing depth). In this Perspective we focus on the latter type of normalization, and in particular, on global-scaling, the most common approach in the literature.

Global-scaling normalization methods assume that the expected value of the read count for a gene in a cell is proportional to a gene-specific expression level and a cell-specific *scaling factor* (also known as a *size factor*), which is an unknown (random) variable representing nuisance technical effects (Box 1). Reverse transcription (RT) efficiency, as well as cell-intrinsic properties, such as endogenous mRNA content, are examples of nuisance effects. Note that, unlike endogenous mRNA content, which is fixed for a given cell, the remaining effects listed in Box 1 are random (if the same cell could be processed twice, these quantities would vary). This implies that scaling factors are inherently random. Nevertheless, most existing methods treat these scaling factors as fixed factors and/or model offsets.

Depending on the experimental protocol, some cell-specific effects cancel out between cells. For example, if library quantification is accurate, the dilution step can remove biases related to differences in the pre-dilution number of amplified cDNA molecules per cell. UMI-based protocols in principle remove amplification and sequencing depth related biases, since multiple reads associated with the same UMI are collapsed into a unique count (Figure 1c). However, this is only true if all libraries are sequenced to saturation (i.e., each uniquely-tagged molecule is observed at least once). If not, some UMI-tagged cDNA molecules will

be lost and, since sequencing depth randomly fluctuates between cells, systematic cell-specific differences between molecule counts can occur. Finally, since UMIs are ligated to each molecule during RT, they cannot account for differences in capture efficiency prior to the RT step, nor for differences in cellular mRNA content.

Normalizing scRNA-seq datasets

scRNA-seq datasets are typically normalized using global-scaling normalization methods inherited from bulk RNA-seq data analysis [7, 8, 24]. In principle, global-scaling factors can be treated as (nuisance) model parameters and jointly estimated with other quantities of interest such as gene-specific expression levels. However, this approach is computationally intensive and necessarily tailored to a specific model (e.g., [19, 25]).

An alternative — and widespread — approach is to compute normalized expression measures based on scaling factors estimates obtained during a pre-processing step (Box 1). Downstream analyses, such as clustering or differential expression, are then typically based on normalized measures (either directly or by treating the estimated scaling factors as model offsets), ignoring uncertainty related to the scaling factor estimation. While this strategy is common, there is no consensus on how to estimate the scaling factors; some popular choices are summarized below. All approaches, however, share the same motivation: to bring cell-specific measures onto a common scale by standardizing a quantity of interest (e.g., total read counts per sample) across cells, while assuming that, e.g., most genes are not differentially expressed.

An intuitive and popular method is *Reads Per Million* (RPM), which standardizes the total number of reads between cells; it is also referred to as library size normalization and is related to RPKM [26] and TPM [27]. However, these estimates can be dominated by a handful of highly expressed genes, which can bias downstream results [16, 22]. Another possibility is to use *Upper-Quartile* (UQ) normalization, which defines scaling factor estimates as proportional to the 75th percentile of the distribution of counts within each cell [16]. An extension of this idea (albeit outside the universe of global-scaling normalization) is *Full-Quantile* (FQ) normalization, where all the quantiles of cell-specific counts are matched to a reference distribution [16]. However, quantile-based normalization methods are problematic in scRNA-seq due to the high frequency of zero counts typically observed. In practice, this can lead to scaling factor estimates being set to 0 in UQ normalization, while the large number of zeros leads to ties in the gene ranking needed by FQ normalization, rendering its interpretation more difficult.

Alternative approaches have been developed in the context of bulk RNA-seq analyses. Two highly popular methods are *DESeq* [22] and *Trimmed Mean of* M*-values* (TMM) [23] normalization. DESeq defines scaling factor estimates based on a pseudo-reference sample, which is based on a geometric mean. TMM trims away extreme log-fold-changes to normalize the counts based on the remaining set of non-differentially expressed genes. Critically, zero inflation is problematic for DESeq, as the calculation of the pseudo-reference sample is only well defined for the potentially very small set of genes with at least one read in every cell.

In the context of bulk RNA-seq, the performance of global-scaling methods was reviewed by [17], where DESeq and TMM were suggested to outperform other methods using a variety of case studies and simulated datasets. However, their performance in the context of scRNA-seq has been given little attention.

Comparing bulk-based approaches: A case study

Using different normalization methods can alter the results of downstream analysis. To illustrate this, we applied the three widely used normalization techniques RPM, DESeq, and TMM to a publicly available dataset (Fig. 2). The data consist of gene expression measures for 933 mouse embryonic stem cells (mESCs) [28]. These cells were processed using a droplet-based protocol, yielding UMI-based counts.

Overall, we observe substantial differences between the methods regarding scaling factor estimation (Fig. 2a, upper right panels). Firstly, due to zero inflation, DESeq scaling factors are based on only 115 genes. As expected, this results in less stable estimation of the scaling factors. Moreover, we observe that — with respect to RPM (and DESeq) — TMM tends to respectively under- and over-estimate large and small scaling factors (this is in line with the simulation results in [14]). This is largely due to the sparsity of the data, with the differences between methods increasing for cells where more zero counts are observed (Fig. 2b, bottom panel).

Crucially, we observe that differences in scaling factor estimation affect gene-specific estimates of variability. This is illustrated using the squared coefficient of variation (CV^2) of the normalized expression measures per gene (Fig. 2a, lower left panels). Thus, analyses whose aim is to uncover heterogeneity within the data are also distorted. For example, studies often start by selecting *highly variable genes* (HVGs) to reduce the dimensionality of the data prior to clustering or other analyses. We observe that HVG selection is sensitive to the choice of normalization, with less than a third of HVGs shared across all normalization methods (Fig. 2c).

We performed the same analyses on additional datasets, showing that these issues are likely general and inherent to scRNA-seq data (Supplementary Data 1). As expected, differences between scaling factors, and consequently between the lists of HVGs, are emphasized in datasets with low sequencing depth. This is critical, as several modern experimental protocols (e.g., droplet-based methods) use shallow sequencing, with less than 50,000 reads per cell, in order to profile a large number of cells. While shallow sequencing has been shown to allow discovery and classification of cell types in complex tissues [29–31], the result of more refined analyses (e.g., pseudo-time ordering [32]) can be distorted by differences between normalization methods.

Given the lack of ground truth in real data, we cannot determine which normalization method, if any, correctly estimates the scaling factors. To shed some light onto the relative merits of each method, we turn to simulations (Supplementary Data 2). We simulated two groups of cells with varying numbers of differentially expressed genes. When the data are simulated with symmetric differential expression, all methods lead to unbiased estimates.

Europe PMC Funders Author Manuscripts

However, with asymmetric differential expression, bulk-based methods lead to biased estimates of the scaling factors (see Figure 2d for an example with 80% up-regulated and 20% down-regulated genes in group 1; see Supplementary Data 2 for other settings). This suggests that great care should be used if bulk-based global-scaling methods are applied to scRNA-seq data.

State-of-the-art

Bulk-based normalization methods are widely applied to scRNA-seq datasets, despite the problems outlined above. However, normalization methods that are specifically tailored to scRNA-seq datasets have recently been introduced. Below, we summarize state-of-the-art methods, provide practical recommendations to scRNA-seq users, and motivate the development of new methodology to address unresolved issues.

We distinguish between two different approaches. Firstly, we consider *bespoke methods* that use pre-normalized expression measures in conjunction with a model that accounts for artifacts specific to scRNA-seq that are not accounted for in the normalization. In the context of differential expression analyses, two examples are SCDE [7] and MAST [8]. To attenuate the effect of technical variation in downstream analysis, SCDE introduces a two-component mixture model to capture dropout events and events where a transcript is faithfully amplified. Alternatively, MAST uses the fraction of genes that are detectably expressed in each cell as a proxy for both technical and biological sources of variation. MAST uses a hurdle model where the expression measure of a detected gene is modeled by linear regression and the probability of detection by logistic regression.

A second strategy for normalizing scRNA-seq datasets is to use *generic methods* that yield normalized expression measures that can be used as input in any subsequent analyses (e.g. [32–34]). A recent example of such an approach, *scran*, pools multiple cells in order to estimate cell-specific size factors more robustly in the presence of zero inflation and unbalanced differential expression of genes across groups of cells (Figure 2d and [14]). In principle, BASiCS [11,25] also provides a generic normalization tool but its implementation has been coupled with specific downstream analysis.

We tested two single-cell motivated methods, BASiCS and scran, on recently published datasets and found that, unlike bulk-based methods, they led to very similar results in terms of scaling factor estimation and HVG selection (Supplementary Data 3). This likely derives from greater robustness to features of single-cell RNA-seq data compared to bulk-based approaches. Other recent examples of normalization methods specifically designed for scRNA-seq include GRM [35] and SAMstrt [36], which both rely on spike-ins, and SCnorm [44], which uses quantile regression to group genes with similar dependence on sequencing depth and estimate different scaling factors for each group. However, it should be noted that GRM is not a between-sample normalization method, but rather a method to de-noise gene expression levels within each cell. In addition, Qiu et al. [45] proposed the Census algorithm to convert relative RNA-seq measurements to relative transcript counts. The Census algorithm can be considered as a normalization method since it rescales TPM values by dividing them by the estimated total number of mRNA molecules.

Finally, we note that although the various global-scaling methods rely on different assumptions, they all fail if the number or fold-change of differentially expressed genes across the cell population is too high. One strategy to alleviate this issue is to pre-cluster the cells into smaller, more homogeneous sets (e.g., using rank-based clustering methods, which are unaffected by global-scaling normalization). Normalization can then be performed separately for each cluster prior to between-cluster normalization to calculate cluster-specific offsets. This approach is used in the “scran” method and has been shown to yield more accurate estimates of scaling factors [14], as also suggested by our simulations (Figure 2d).

Spike-in sequences and normalization

The scaling factors introduced in Box 1 cannot distinguish between technical biases and genuine biological differences between cells, such as total mRNA content. Jiang et al. [12] discussed the benefits of exploiting a set of synthetic control genes — with constant expression level across all samples — to disentangle these effects in bulk RNA-seq. Extrinsic control genes have also been used in the context of scRNA-seq [6, 25, 33, 35, 36], where spike-in sequences are added to each cell’s lysate in a theoretically constant and known amount. The most commonly used set of spike-ins is the set of 92 External RNA Control Consortium (ERCC) molecules [12]. Other examples include the 8 synthetic mRNAs deployed in [37] and the whole transcriptome HeLa RNA spike-in used in [6]. An important question is to understand the utility of synthetic spike-in sequences in the context of global-scaling normalization.

One critical assumption underlying the use of spike-in sequences is that the technical effects summarized in Box 1 equally affect the intrinsic and the extrinsic genes. If this assumption holds, additional technical scaling factors can be defined that capture these shared technical effects [6]. Thus, for any given cell, the ratio between the scaling factor described in Box 1 and the technical scaling factor defined above is equal to the endogenous mRNA content of the cell. As a corollary, normalization based solely on spike-in derived scaling factors does not remove differences in endogenous mRNA content between cells and further normalization is required to remove this effect.

This suggests that spike-in sequences can be used to obtain estimates of endogenous mRNA content per cell. At a coarse level, this is reflected in several scRNA-seq datasets (Figure 3a), consistent with previously described bulk RNA-seq studies [38]. Here, we look at three different datasets [33, 42, 43], for which we can stratify samples according to their expected mRNA content. The ratio of mRNA/spike-in read counts correctly indicates that mRNA content increases as mouse embryonic stem cells progress along the cell cycle [33] and decreases across blastomeres in early mouse embryos at 2-,4-, and 8-cell stages [42] due to their difference in size. Analogously, in an experiment on the Fluidigm C1 instrument, wells including multiple cells are characterized by a higher mRNA content than wells where single cells are captured [43].

However, using spike-in sequences remains challenging. In particular, calibrating the added number of spike-in molecules is non-trivial and depends on intrinsic characteristics of the

studied cells, such as endogenous mRNA content. Poor calibration can invalidate the utility of the spike-ins as control genes: too many spike-ins can overwhelm signal from the intrinsic genes, while the majority of spike-in sequences can be unusable in downstream analysis if too few spike-in molecules are added [19].

Additional issues arise for specific sets of spike-in sequences. In particular, for the widely used set of ERCC controls, the extreme range of concentration of spike-in molecules [38] prevents the use of the entire ERCC set in (single-cell) RNA-seq: Typically only half of the spike-in molecules are detected and the proportion of reads mapped to the spike-in sequences may be extremely variable (Figure 3a).

Moreover, potential biases in the mRNA enrichment process related to gene length and GC-content imply that, overall, technical effects may be different for the ERCC spike-in sequences and the intrinsic genes. In fact, the ERCC set does not reflect the mammalian transcriptome in terms of gene length and GC-content (Figure 3b). Moreover, [19] showed that ERCC spike-in signal can vary considerably between technical replicate samples. Consequently, estimates of endogenous mRNA content derived using ERCC spike-ins have large measurement uncertainties [38].

Developing a set of spike-ins specifically tailored for scRNA-seq experiments could overcome some of these limitations. Ideally, this set should closely resemble intrinsic genes in terms of the distribution of GC-content, total length, and polyA tail length. Ongoing efforts in this context are illustrated by a recent call from the National Institute of Standard and Technology (https://federalregister.gov/a/2015-19742), to design an improved set of controls, which should (i) mimic endogenous RNA and (ii) not interfere with the measurement of endogenous RNA. More recently, [39] introduced *sequins* (sequencing spike-ins) — a set of extrinsic spike-ins designed for (bulk) RNA-seq experiments.

Discussion

One aim of this perspective is to provide a straightforward understanding of the sources of variation that can be captured through global-scaling normalization in the context of scRNA-seq.

Case studies and simulated datasets highlighted that a direct application of bulk RNA-seq normalization methods is not appropriate in the context of scRNA-seq, where — due to biological heterogeneity as well as technical artifacts — we typically observe more heterogeneous and sparser datasets. In particular, we illustrated that the choice of the normalization method affects downstream analyses, such as HVG detection, that aim to uncover heterogeneity within the data. Spike-in sequences can help to disentangle differences in endogenous mRNA content between cells from technical artifacts, though they do carry some caveats. This can be useful in several contexts, such as the whole-transcriptome up-regulation induced by elevated expression of the *c-Myc* transcription factor [40].

A variety of scRNA-seq tailored methods have recently been proposed that outperform bulk strategies. Despite this, bulk-motivated approaches remain widely used in practice. We

Europe PMC Funders Author Manuscripts

therefore suggest that scRNA-seq users update their analysis pipelines — matching advances in technology — to take full advantage of the rich information provided by scRNA-seq datasets. Finally, while the issue of how best to normalize scRNA-seq data has not yet been fully resolved, many efforts are underway to develop additional robust and effective normalization techniques and to systematically assess their performance on individual datasets [44, 45, 46].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank several members of the Marioni laboratory (European Molecular Biology Laboratory - European Bioinformatics Institute, EMBL-EBI; Cancer Research UK - Cambridge Institute, CRUK-CI) for support and discussions throughout the preparation of this manuscript. In particular, we are grateful to Aaron Lun (CRUK-CI) for constructive comments on an earlier version of the manuscript. We are also grateful to UC Berkeley collaborator John Ngai and his group members.

Funding

CAV, AS, and JCM acknowledge core EMBL funding. CAV was supported by core MRC funding (MRC MC UP 0801/1) and by The Alan Turing Institute under the EPSRC grant EP/N510129/1. JCM acknowledges core support from CRUK. AS acknowledges funding from the Wellcome Trust Strategic Award 105031/D/14/Z ‘Tracing early mammalian lineage decisions by single-cell genomics’. DR and SD are supported by the National Institutes of Health BRAIN Initiative grant U01 MH105979 (PI: John Ngai).

References

[1]. Tang F, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods. 2009; 6:377–382. [PubMed: 19349980]

[2]. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics. 2013; 14:618–630.

[3]. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics. 2015; 16:133–145.

[4]. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Research. 2014; 42:8845–8860. [PubMed: 25053837]

[5]. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nature Reviews Genetics. 2016; 17:175–188.

[6]. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods. 2013; 10:1093–1095. [PubMed: 24056876]

[7]. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature Methods. 2014; 11:740–742. [PubMed: 24836921]

[8]. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology. 2015; 16:1–13. [PubMed: 25583448]

[9]. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biology. 2015; 16:241. [PubMed: 26527291]

[10]. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology. 2016; 17:63. [PubMed: 27052890]

[11]. Vallejos CA, Richardson S, Marioni JC. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biology. 2016; 17:70. [PubMed: 27083558]

[12]. Jiang L, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Research. 2011; 21:1543–1551. [PubMed: 21816910]

[13]. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Molecular Cell. 2015; 58:610–620. [PubMed: 26000846]

[14]. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology. 2016; 17:75. [PubMed: 27122128]

[15]. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003; 31:265–273. [PubMed: 14597310]

[16]. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. [PubMed: 20167110]

[17]. Dillies M-A, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics. 2013; 14:671–683. [PubMed: 22988256]

[18]. Hicks SC, Teng M, Irizarry RA. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. bioRxiv. 2015 025528.

[19]. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology. 2014; 32:896–902.

[20]. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research. 2014; 42:e161.

[21]. Islam S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2014; 11:163–166. [PubMed: 24363023]

[22]. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11:R106. [PubMed: 20979621]

[23]. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010; 11:R25. [PubMed: 20196867]

[24]. Grun D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell. 2015; 163:799–810. [PubMed: 26544934]

[25]. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Computational Biology. 2015; 11:e1004333. [PubMed: 26107944]

[26]. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008; 5:621–628. [PubMed: 18516045]

[27]. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010; 26:493–500. [PubMed: 20022975]

[28]. Klein AM, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161:1187–1201. [PubMed: 26000487]

[29]. Pollen AA, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology. 2014; 32:1053–8.

[30]. Zeisel A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015; 347:1138–1142. [PubMed: 25700174]

[31]. Macosko EZ, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161:1202–1214. [PubMed: 26000488]

[32]. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014; 32:381–386.

[33]. Buettner F, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology. 2015; 33:155–160.

[34]. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods. 2016; 13:845–848. [PubMed: 27571553]

[35]. Ding B, et al. Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics. 2015; 31:2225–7. [PubMed: 25717193]

[36]. Katayama S, Tohonen V, Linnarsson S, Kere J. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics. 2013; 29:2943–2945. [PubMed: 23995393]

[37]. Islam S, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research. 2011; 21:1160–1167. [PubMed: 21543516]

[38]. Munro SA, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nature Communications. 2014; 5:5125.

[39]. Hardwick SA, et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nature Methods. 2016; 13:792–798. [PubMed: 27502218]

[40]. Loven J, et al. Revisiting global gene expression analysis. Cell. 2012; 151:476–482. [PubMed: 23101621]

[41]. Kolodziejczyk AA, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 2015; 17:471–85. [PubMed: 26431182]

[42]. Goolam M, et al. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. Cell. 2016; 165:61–74. [PubMed: 27015307]

[43]. Scialdone A, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods. 2015; 85:54–61. [PubMed: 26142758]

[44]. Bacher R, et al. SCnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data. bioRxiv. 2016 090167.

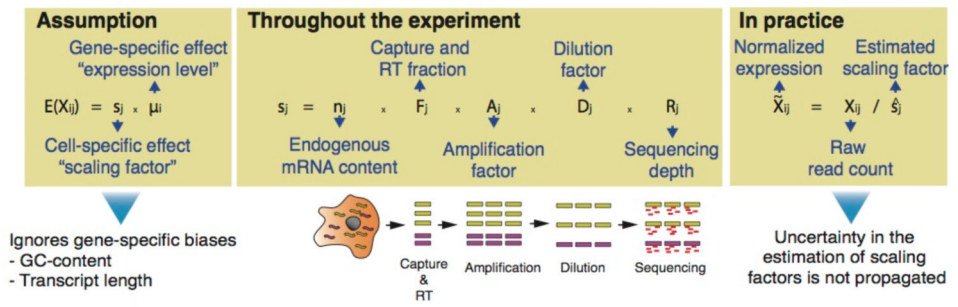
[45]. Qiu X, et al. Single-cell mRNA quantification and differential analysis with Census. Nature Methods. 2017 Advance online publication.

[46]. Cole M, Risso D. scone: Single Cell Overview of Normalized Expression data. 2016 R package version 0.99.6.

Europe PMC Funders Author Manuscripts

Box 1

Global Scaling Normalization for scRNA-seq Datasets



RNA-seq experiments are inherently stochastic, with reads being randomly sampled from a pool of amplified cDNA molecules. Accordingly, let X_{ij} denote a random variable representing the read count of gene i in cell j. Typically, the parameter of interest is the expression level of each gene (see left panel), i.e., the relative abundance of mRNA molecules for a gene within the population of mRNA molecules in each cell. For the sake of simplicity, we consider here the case of a homogeneous population of cells.

Intuitively, a first effect captured through the scaling factor s_j is the *endogenous mRNA content* n_j, the total number of mRNA molecules per cell (middle panel). Indeed, even within a homogeneous population, n_j can vary across cells. Furthermore, after cell lysis, only a fraction of these n_j molecules, F_j, are captured and reverse transcribed into cDNA. Consequently, only $n_j \times F_j$ cDNA molecules can potentially be amplified and subsequently sequenced. Critically, the *capture and reverse transcription efficiency* F_j varies between cells, which introduces cell-to-cell variability that should also be handled by s_j.

Subsequently, due to the minute amount of genetic material contained in a cell, this pool of $n_j \times F_j$ cDNA molecules must be amplified prior to sequencing library preparation. Variability in amplification efficiency can introduce cell- and gene-specific biases in the measurement of expression levels. We denote the cell-specific *amplification factor* by A_j, such that amplification leads to a pool of $n_j \times F_j \times A_j$ molecules.

Unlike microarray experiments, RNA-seq is inherently competitive, meaning that a fixed number of reads are distributed between genes. Given this, the amplified pools are subsequently diluted by a cell-specific factor D_j, so that there are $n_j \times F_j \times A_j \times D_j$ amplified cDNA molecules to be sequenced. In principle, the *dilution factor* D_j can be set so that a library contains the *same number of molecules* from each cell, by carrying out a library quantification step and setting $D_j = m / (n_j \times F_j \times A_j)$, where m is the desired number of molecules per cell. Alternatively, each cell can contribute the *same volume* of amplified cDNA solution to the library, such that each library will contain a different number of amplified cDNA molecules if the concentration of the solution varies between cells. In this case, $D_j = d$, where d is the proportion of amplified molecules used to prepare the sequencing library. This decision is critical for interpreting the scaling factor

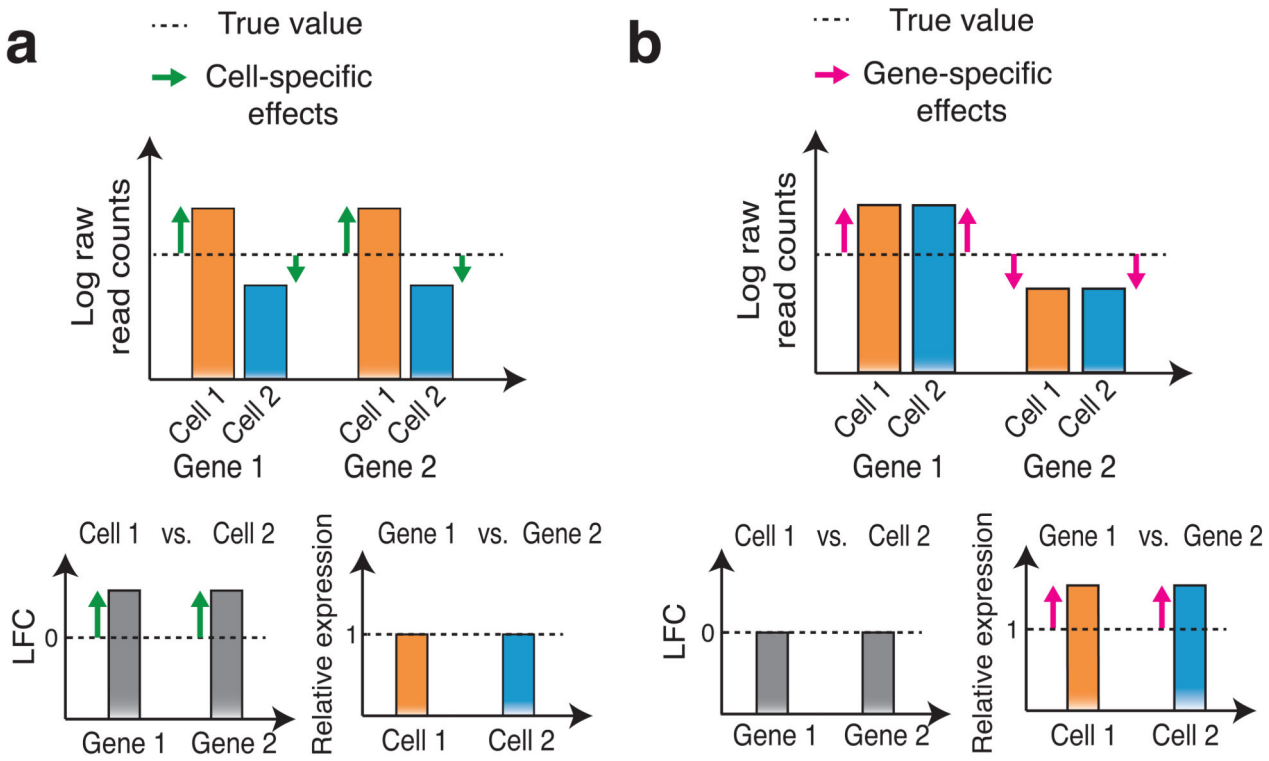
s_j, since it affects the number of molecules that are available for sequencing and, consequently, the scale of cell-specific read counts.

Finally, the number of sequenced reads per molecule from each cell (*sequencing depth*), R_j, also varies stochastically. Consequently, by considering all the above factors, we expect to observe $n_j \times F_j \times A_j \times D_j \times R_j$ reads from cell j. Hence, even within the same sequencing lane, differences in sequencing depth introduce cell-specific artifacts that will be incorporated into the global-scaling factor s_j.

While the above discussion assumes a homogeneous population of cells, this interpretation of scaling factors is still valid for more realistic scenarios - with heterogeneous populations - under specific assumptions, such as that the majority of genes is not differentially expressed or that there are roughly an equal number of up- and down-regulated genes.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



c

	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture & RT efficiency	✓	✓	✓
Gene length		✓	
GC-content	✓	✓	✓
mRNA content	✓		✓

Figure 1. Cell- and gene-specific effects in RNA-seq experiments.

(a) Schematic representation of cell-specific effects. The top panel shows a pair of cells expressing two genes at the same levels. When RNA-seq is performed, cell-specific effects introduce a bias in the estimated log-fold-change (LFC) computed on raw read counts (bottom left panel). (b) Schematic representation of gene-specific effects. The two cells and true gene levels are the same as in (a), but now gene-specific effects are shown to bias the estimation of relative gene expression (bottom right panel). In real situations, both cell-

specific and gene-specific effects are present. (c) List of main cell- and/or gene-specific effects and whether these are removed by unique molecular identifiers (UMIs).

Europe PMC Funders Author Manuscripts

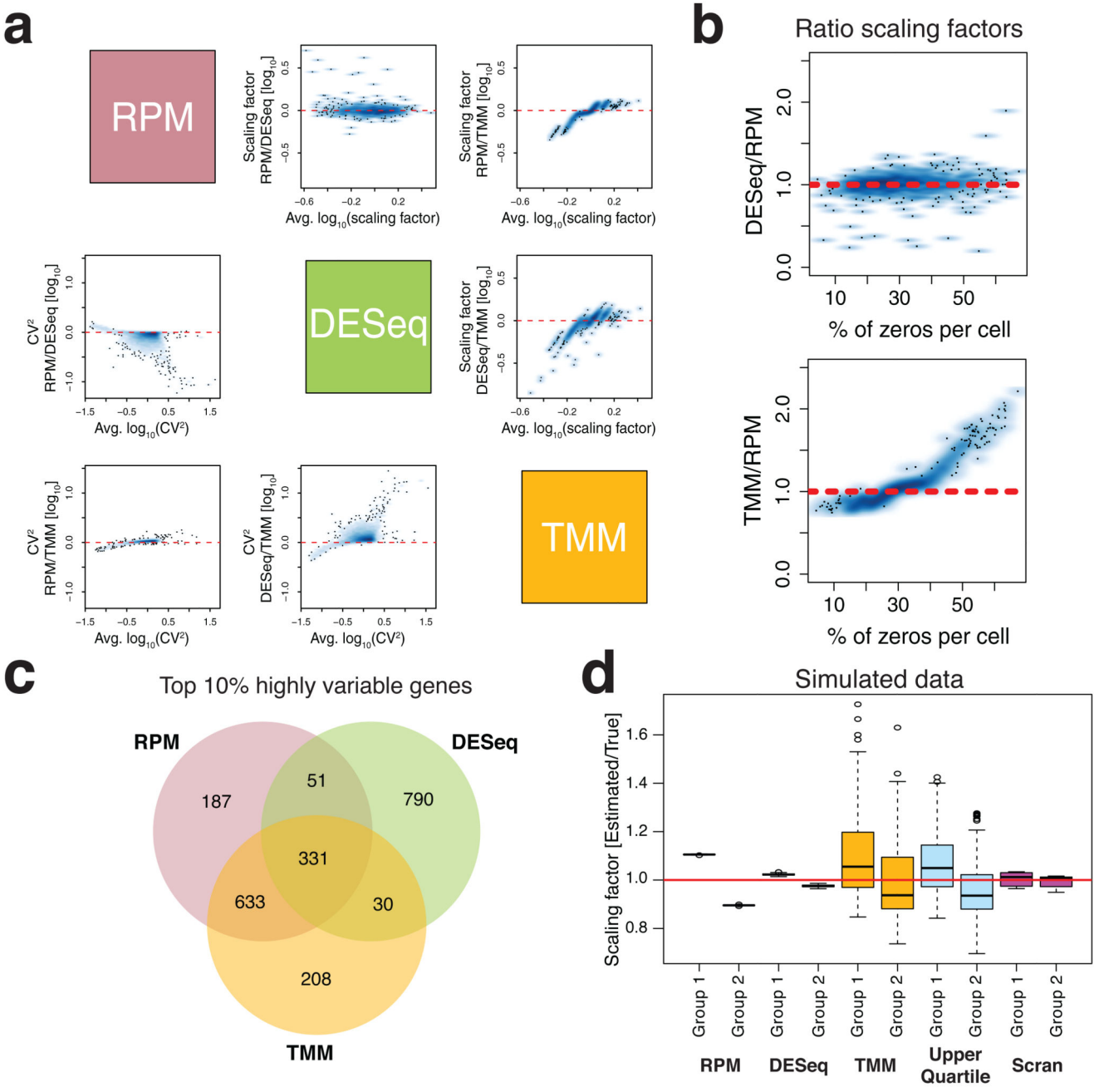


Figure 2. Comparison of bulk-based normalization methods in real and simulated datasets.

(a) Mean-difference plot comparing the estimated scaling factors (upper-triangular panels) and CV^2 of normalized counts (lower-triangular panels) for the dataset published in [28]. (b) Ratio of estimated scaling factors vs. proportion of zero counts per cell for dataset [28]. (c) Top 10% most variable genes identified after normalizing dataset [28] with three different methods. Additional datasets are analyzed in Supplementary Data 1. (d) Ratio between the estimated and the true scaling factors for the most widely used bulk-based normalization methods and a method specifically designed for scRNA-seq ("scran") [14] in a simulated

Europe PMC Funders Author Manuscripts

dataset consisting of two groups of cells. See Supplementary Data 2 for the simulation strategy and additional simulations.

Europe PMC Funders Author Manuscripts

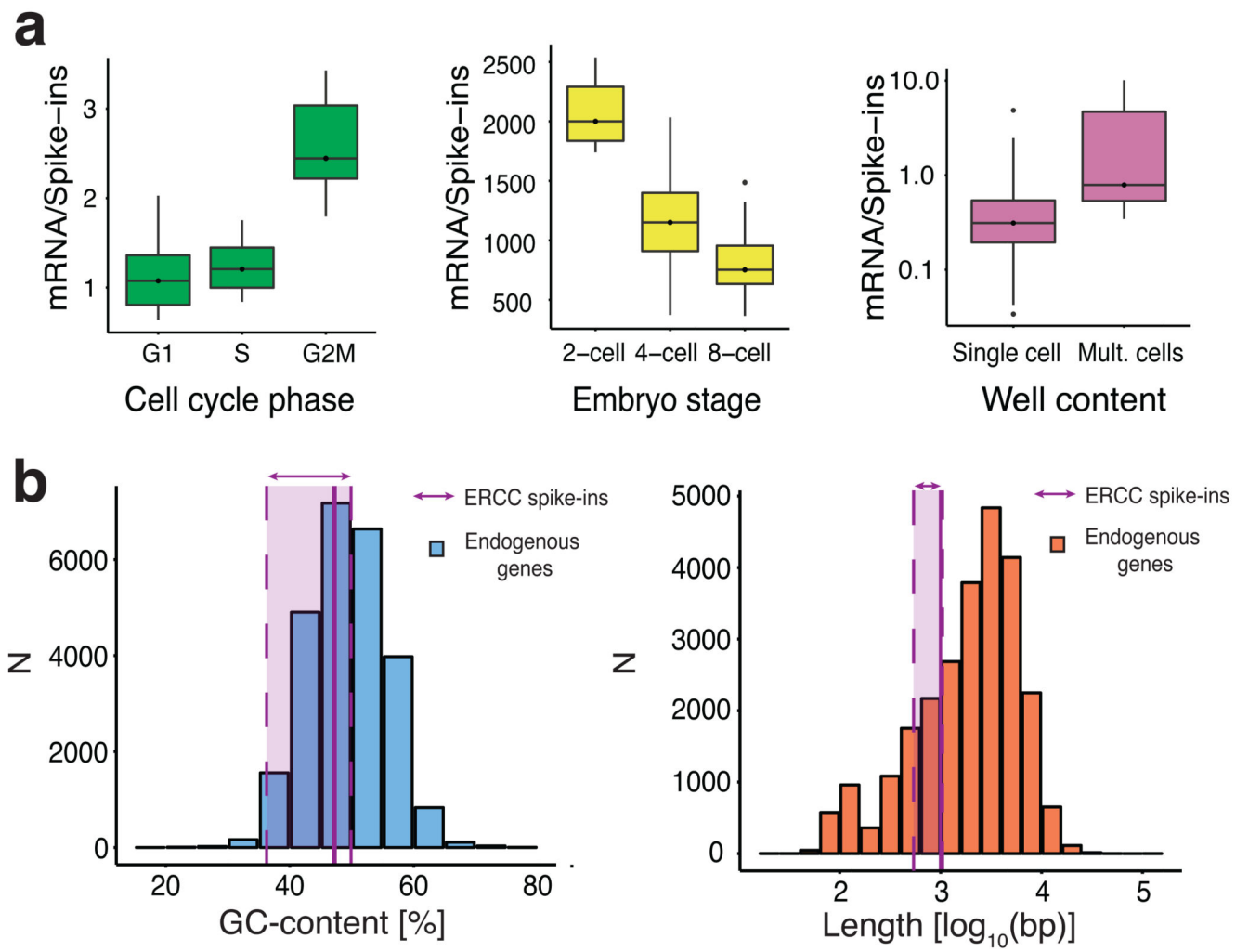


Figure 3. ERCC spike-ins can be used to estimate mRNA content.

(a) Ratio between the number of reads mapped to intrinsic genes and the number of reads mapped to ERCC spike-ins in datasets from [33, 42, 43] (left, central and right panel respectively). (b) Distributions of GC-content (left panel) and length (right panel) for mouse genes with at least one count in one cell in the dataset published in [41]. The purple areas show the interquartile ranges of GC-content and length for ERCC spike-ins, with the medians marked by vertical purple continuous lines.

Europe PMC Funders Author Manuscripts