Robust Learning of Transfer Functions for Single-Cell Transcriptomics Depth Normalization

Da Kuang [1] **Junhyong Kim** [1 2]

Abstract

Normalization is a critical step in data processing that influences downstream analyses. It aims to adjust for technical variations in data acquisition, facilitating accurate comparisons among observations. In this paper, we identify key challenges in scRNA-seq normalization, including the simplex nature for reads, compositional bias from the mRNA population, technical and biological outliers, and the non-linear relationship between input and output. We introduce a new framework to address these challenges by modeling the measurement function and robust learning of parameters. Empirical validation using real datasets demonstrates the effectiveness of our proposed normalization method, RFNorm, in preserving lower-dimensional mathematical structures, which are crucial for analyzing cell types and states. This effectiveness is assessed through the invariance of k-nearest neighbor graphs and by comparing the performance of RFNorm with established methods.

1. Introduction

Many measurements in empirical sciences are subject to perturbations and noise, both technical and intrinsic. In genomics, the development of high-throughput methods such as next-generation sequencing (NGS) and single-cell RNA sequencing (scRNA-seq) supports the analysis of millions of cells and thousands of features per cell under various conditions, presenting new challenges in data processing (Svensson et al., 2017; Amezquita et al., 2020). These include normalization for instrument noise (L. Lun et al., 2016; Bacher et al., 2017; Lopez et al., 2018; Hafemeister & Satija, 2019), data integration across various sources (Korsunsky et al., 2019; Liu et al., 2020; He et al., 2024), and imputation for sparse data (Huang et al., 2018; Van Dijk et al., 2018; Marouf et al., 2020). Normalization, a cornerstone concept in statistics, aims to transform data into a standardized format to enable accurate comparisons and analyses. Ideally, normalization should: (1) correct for instrumentation-specific biases (Vallejos et al., 2017), (2) manage statistical distributional properties to support analytical techniques (Ioffe & Szegedy, 2015), and (3) enable meaningful comparison with respect to hypotheses of interest (Love et al., 2014).

NGS data are prone to instrumentation-specific biases, notably influenced by sequencing depth—the total number of sequenced reads obtained from each cell(Robinson & Oshlack, 2010; Vallejos et al., 2017; Evans et al., 2018; Booeshaghi et al., 2022). A fundamental aspect of NGS is its ability to quantify the presence of RNA or DNA in terms of counts. However, these counts are not only dependent on the input material but also subject to compositional bias, especially when the library is not sequenced to saturation, a common scenario in scRNA-seq(Love et al., 2014; Evans et al., 2018; L. Lun et al., 2016). Such bias skews the observed counts based on the composition of the RNA population. Furthermore, the intrinsic nature of NGS data is compositional: the abundance of any nucleotide fragment can only be understood in relation to others, constrained by the sequencer's finite capacity. This limitation frames the data within an Aitchison Simplex rather than conventional Euclidean space, where an increase in one fragment's abundance can misleadingly appear as a decrease in another, thereby leading to spurious gene correlations (Su et al., 2023).

Normalization methods must effectively manage the statistical distributional properties of scRNA-seq data to support downstream analyses. These data exhibit inherent complexities such as heteroskedasticity and gene-gene correlations. Heteroskedasticity, the uneven variance across data points, arises from the intrinsic properties of RNA molecules and varies with biological conditions such as environmental influences and cell cycles (Ahlmann-Eltze & Huber, 2023). On the other hand, gene-gene correlations, which stem from regulatory interactions between genes, reflect essential bio-

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA [2]Department of Biology, University of Pennsylvania, Philadelphia, USA. Correspondence to: Da Kuang <kuangda@seas.upenn.edu>, Junhyong Kim <junhyong@sas.upenn.edu >.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

logical functions and should be preserved (Sun et al., 2021). Therefore, normalization should adjust for heteroskedasticity to stabilize variance across samples while maintaining the gene-gene correlations. Current normalization methods often treat genes as independent and identically distributed, or group them into bins based on expression levels. This approach can distort the true biological signal in data that inherently exhibits a mixed distribution.

Normalization must enable meaningful comparisons. In scRNA-seq, we are interested in biological signals to distinguish the functional states of individual cells, track developmental pathways, or identify new cellular phenotypes. A critical and intriguing question emerges: Even if we could precisely count the number of mRNA molecules, how would we compare the genomic profiles between two cells? This comparison must address two primary aspects: the compositional nature of cells and the variability from biological factors such as differential expression (DE). Even if a sequencer could perfectly capture every RNA molecule in a cell, cells themselves are inherently compositional due to constraints like volume and energy, which limit RNA synthesis. This is evidenced by observations that smaller cells of the same type contain proportionally less total mRNA (Quinn et al., 2019). Furthermore, optimal normalization should ensure that non-DE genes maintain consistent normalized counts across conditions, while DE genes display read counts that accurately reflect the true differences in transcripts per cell. Additionally, the comparison of molecular profiles can be likened to high-dimensional geometric shape analysis in Geometric Morphometrics, such as resistant fit Procrustes superimposition followed by thin-plate spline to visualize morphological deformation between two objects (Zelditch et al., 2012).

In this study, we introduce a framework for normalization and explicitly address properties (1)-(3) by modeling measurement functions to mitigate instrument-induced variations, accounting for out-of-distribution values, and robust learning the parameters by minimizing empirical risk. Through our proposed normalization framework and simulation study, we highlight the limitations of existing methods, which often assume that a single scalar factor can adequately correct for most technical variations. We introduce an R package called RFNorm, incorporating more realistic assumptions and robust optimization into our model. To evaluate the effectiveness of RFNorm, we assess its performance in preserving low-dimensional mathematical structures as outlined in (Ahlmann-Eltze & Huber, 2023), and verify accuracy and Type 1 error control using DEs in paired Bulk RNA-seq(Squair et al., 2021) and constructed null datasets(Soneson & Robinson, 2018). Our findings show that RFNorm achieves high k-NN graph overlap, effectively preserving biological relationships across various datasets, and performs comparably in identifying marker genes while accurately managing Type 1 error rates comparing with the other two popular scRNA-seq depth normalization methods such as CP10K and Scran.

2. Related Works

Normalization can be viewed as a two-step transformation (as shown in Fig 1b) applied to the raw counts to produce the normalized number (Ahlmann-Eltze & Huber, 2023; Booeshaghi et al., 2022). In this paper, we focus on depth normalization for single-cell RNA-seq data. Putting the assumptions of different normalization methods under the same framework enables us to see if they resolve some of the key challenges.

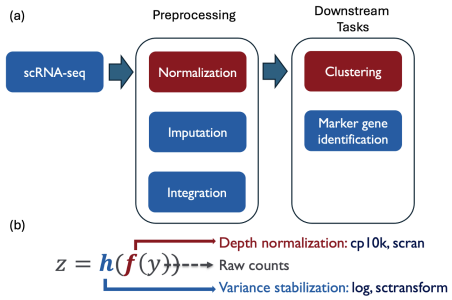


Figure 1. Overview of Normalization in scRNA-seq. (a) Common workflow (b) Two steps of the normalization function.

The concept of variance stabilization has a long research history. Many approaches have been proposed since the era of Bulk RNA-Seq (Anders & Huber, 2010; Robinson & Oshlack, 2010; Love et al., 2014; Hafemeister & Satija, 2019; Lause et al., 2021; Choudhary & Satija, 2022). However, recent research indicates that the simple approach, the logarithm with a pseudo-count (log1p), performs as well or better than more sophisticated alternatives, such as SCTransform (Hafemeister & Satija, 2019), for variance stabilization in terms of the preservation of low dimensional graphs(Ahlmann-Eltze & Huber, 2023). In this work, we choose log1p as the variance stabilization function and focus on depth normalization and focus on the design of depth normalization function.

Most sequencing depth normalization approaches (CP10K (Luecken & Theis, 2019), scNorm(Bacher et al., 2017), BayNorm(Tang et al., 2020)) use the same size factor assumption as SCTransform. However, the scaling transformation has limitations. First, since the observed counts are simplex data, scaling by the sequencing depth does not remove the constant sum constraint and cannot convert closed sequencing data into an "open" unit such as concentration (Quinn et al., 2019). Second, transforming only by sequencing depth is a within-sample transformation, whereas a between-sample transformation (Maza, 2016;

Evans et al., 2018; Love et al., 2014; L. Lun et al., 2016) would be more appropriate due to the compositional bias. To be more specific, since we cannot saturate the library during sequencing most of the time, the observed counts are likely to be biased by the composition of the mRNA population in each cell. Third, the complexity of the depth normalization function $f(y)$ presents a unique challenge that has been overlooked.

3. Normalization Framework

3.1. Preliminaries

In transcriptomics and single-cell RNA sequencing, each biological cell is assumed to contain a certain number of measurable RNA molecules across G different types of transcripts (or genes), denoted as $n_i = x_1 + x_2 + \cdots + x_G$, where n_i may vary for each measured cell. The scRNA-seq instruments aim to recover these counts as integer vectors, producing observations $(y_1, y_2, ..., y_G)$ for each cell, across potentially tens of thousands or even millions of cells. Normalization in this context is a two-step transformation applied to the raw counts y_g to produce the normalized number z_g (Ahlmann-Eltze & Huber, 2023; Booeshaghi et al., 2022). This process can be formally expressed as $z_g = \boldsymbol{h}(\boldsymbol{f}(y_g))$, where $\boldsymbol{h}$ is the variance stabilization function and $\boldsymbol{f}$ is the depth normalization function. For example, SCTransform (Hafemeister & Satija, 2019), a widely used normalization method, employs Negative Binomial (NB) regression to model UMI counts. To account for sequencing depth, SCTransform utilizes the link function $\log \mu = \beta + \log n_i$ in NB regression, where n_i is the total counts of the cell i, and μ and β are parameters specific to each gene. The link function is designed based on the size factor assumption (Vallejos et al., 2017), where the observed counts are the true counts scaled by some cell-specific size factor. Variance stabilization is then achieved using the feature/gene-level Pearson residuals.

3.2. Problem Setup

We propose a framework to characterize the normalization process in scRNA-seq experiments involving N cells and G genes. We define three conceptual subspaces: Biological Space V_B, Measurement Space V_M, and Normalization Space V_N, as illustrated in Fig 2. Biological Space is a subspace of Euclidean space $\mathbb{R}^G$, representing the true molecular profiles. Measurement Space is a subspace of the Aitchison Simplex space because the observed counts are compositional data (Quinn et al., 2019). Ideally, Normalization Space would also be a subspace of Euclidean space $\mathbb{R}^G$, representing the post-normalization profiles.

Data in Biological Space V_B are inherently unknowable to us. We assume that a measurement function ϕ maps vectors from V_B to V_M, where $Y \in V_M$ represents the experimental data collected. A normalization function ω is then defined to map vectors from V_M to V_N. Ideally, ω would serve as the inverse function of ϕ, enabling retrieval of the true biological states. However, inverting this function is complex due to the unknown parameters and nature of ϕ.

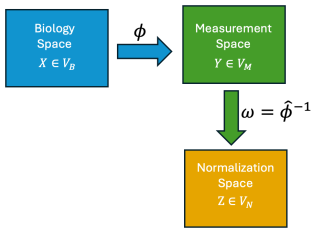


Figure 2. Schema of the normalization framework. Suppose $X = [\mathbf{x}_i], Y = [\mathbf{y}_i], Z = [\mathbf{z}_i]$ are matrices with dimensions $G \times N$. Column vectors $\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i$ represent observations in their respective subspaces, for instance, $\mathbf{y}_i = (y_{1,i}, y_{2,i}, \cdots, y_{G,i}), i \in [N]$.

Normalization methods thus construct the Normalization Space and formulate ω based on an estimated $\hat{\phi}$ under assumptions about the measurement function. For instance, if the observed counts are assumed to be the true counts scaled by a size factor, then $\mathbf{y}_i = \hat{\phi}(\mathbf{x}_i) = s \times \mathbf{x}_i$. The sequencing depth can then be used as the estimator for the size factor, leading to $\mathbf{z}_i = \omega(\mathbf{y}_i) = 1/s \times \mathbf{y}_i = 10^4/n_i \times \mathbf{y}_i$, where $n_i = \sum \mathbf{y}_i$ represents the total number of reads in cell i and 10^4 is included for numerical stability. To this end, we integrate the CP10K method into our normalization framework.

Our framework serves as a conceptual model to address the three fundamental properties of effective normalization, as outlined in the introduction. By now, it is not a normalization method per se, but rather provides a mindset for designing normalization approaches. Property (1), the correction of instrument-related biases, is approached through the inversion of measurement functions ϕ, which helps to minimize inherent variations in the measurement process. For Property (2), we recommend implementing transformations like the log-shift within the normalization function, stabilizing variance across samples and ensuring statistical integrity. With regard to Property (3), which aims to facilitate meaningful comparisons in alignment with research hypotheses, this mindset guides us to optimize empirical risk for data subsets considered representative and within normal distribution. These strategies underscore the framework's utility as a foundation for developing robust normalization methods, further discussed in sections 3.3 and 4.2.

3.3. Normalization Paradigms

This section delineates two primary normalization paradigms: between-sample normalization and within-

sample normalization. Within-sample normalization aims to scale data within a single sample to reduce internal variability and ensure comparability within that sample. However, the simplex nature of scRNA-seq data and inherent compositional bias often render within-sample normalization suboptimal for comprehensive data analysis, as depicted in Fig 3.3(a-c).

Conversely, between-sample normalization, which we advocate for, adjusts measurements across different samples to account for systematic variations. This paradigm is utilized in methodologies such as EdgeR, DESeq2, and SCRAN. The schema of between-sample normalization in our framework is illustrated in Fig 3.3(d-f). Here, we consider the forms of the measurement functions ($\hat{\phi}_i, \hat{\phi}_r$), and then invert them to derive a composite function $\Omega_{i,r} = \omega_i \circ \omega_r^{-1}$. This function maps the counts from a given cell $\mathbf{y}_i$ to those of a reference cell $\mathbf{y}_r$. If we posit that the observed values $\mathbf{y}$ are related to the true values $\mathbf{x}$ by an exponential transformation $\mathbf{y} = A\mathbf{x}^B$, we can then estimate the relationship $\mathbf{y}_r = \lambda \mathbf{y}_i^k$ for unknown parameters λ and k. This allows us to optimize the empirical risk:

$$\arg\min \sum_{g=1}^{G} ||\Omega_{i,r}(y_{i,g}) - y_{r,g}||^2 \tag{1}$$

where $\Omega_{i,r}(\mathbf{y}_i) = \lambda \mathbf{y}_i^k$ and $\mathbf{y}_i = (y_{i,1}, \cdots, y_{i,G})$.

Optimization of the above function accounts for instrument-related bias and achieves between-sample normalization. To this end, we transform the measurements in Simplex Space to Euclidean space using the composed normalization function Ω. To further address property (3), establishing an appropriate comparison between two or more cells is essential. Here, we concentrate on the problem posed by biological perturbations such as differential expression (DE), certain values of the RNA measurements may be out-of-distribution. To approach this problem, we propose optimizing the empirical risk for an appropriate subset of the data, assumed to be in-distribution, as discussed in 4.2.

3.4. Assumptions of measurement functions

Our framework provides a conceptual model for adjusting the measurement function $\hat{\phi}$ based on underlying assumptions about the measurement process. In size-factor-based normalization, commonly used in methodologies like CP10K, DESeq2, EdgeR, and Scran, several critical assumptions underpin the measurement function.

Assumption 1: A single scaling factor is sufficient to model the measurement function.

Assumption 2: The measurement function is cell-specific.

Assumption 3: The majority of genes within each cell share a consistent measurement function.

Assumption 4: The measurement function is expected to be monotonic, meaning that increases in input lead to increases in output without any reversal.

CP10K uses assumption (1) and (2) so that the sum of reads is used as the estimator of the size factor. DESeq2 and EdgeR incorporate assumptions (1), (2), and (3) to have a more robust estimation of the size factor. Here we outline the construction of normalization function and more detailed analysis can be found in the Appendix. By applying Assumption 1, we have $Y = \hat{\phi}(X) = sX$. Based on Assumption 2, we can construct Ω as

$$\Omega(\mathbf{y_i})_{i,r} = \omega_r^{-1}(\omega_i(\mathbf{y}_i)) = \frac{\mathbf{y}_i}{s_i/s_r}. \tag{2}$$

The reference cell can be the geometric mean of the data (DESeq2) or just any other cell (SCRAN). Applying Assumption 3, we have

$$\mathbb{E}[\frac{s_i}{s_r}] = \mathbb{E}[\frac{\sum \mathbf{y}_i}{\sum \mathbf{y}_r}] \propto \mathbb{E}[\frac{y_{g,i}}{y_{g,r}}]. \tag{3}$$

DESeq2, EdgeR, and SCRAN each employ unique strategies to robustly estimate $\mathbb{E}[\frac{y_{g,i}}{y_{g,r}}]$. To this end, we incorporate these between-sample normalization strategies into our normalization theory framework.

Questioning Assumption 1, we recognize that many instruments exhibit lower detection thresholds and upper saturation limits.[1] We propose that a sigmoidal function may offer a more realistic approximation of the NGS measurement function, especially in scRNA-seq where experimental dropouts often result in zero counts for low input values, and the total number of sequences caps the high values. Consequently, ϕ is expected to behave as a piece-wise linear function, registering zero up to a critical lower limit and plateauing at a maximum number beyond an upper saturation limit. In this model, only the input values within a certain "useful range" lead to significant output changes. Within this range, the transfer function is effectively invertible and mimics the scaling transformation typically assumed in size-factor-based normalization. This piece-wise linear behavior can be effectively modeled by a sigmoid function, suggesting that Assumption 1 should be generalized to reflect more complex, realistic behaviors.

It is also crucial to note that all instruments and normalization methods generally assume that the functional relationships are monotonic. Non-monotonic transformations,

[1] For example, consider the input-output relationship of a light sensor that receives photons of a specific wavelength and generates current.

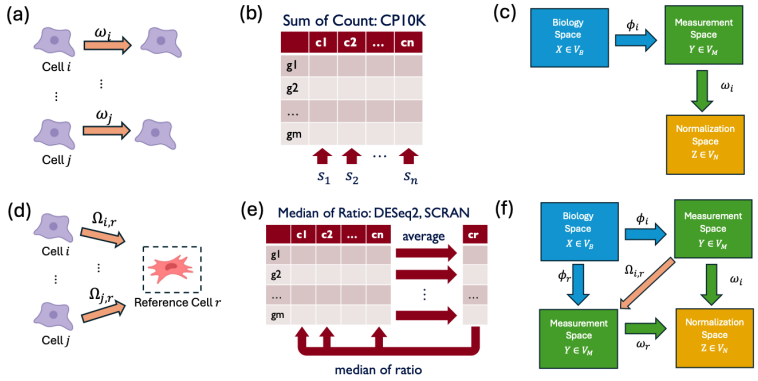


Figure 3. Illustration of normalization paradigms and their components. (a) Represents within-sample normalization where each cell is normalized independently. (b) Shows the application of CP10K which uses the sum of counts across genes for normalization. (c) Depicts the transformation from Biology Space to Measurement Space, then to Normalization Space. (d) Demonstrates between-sample normalization comparing cells to a reference cell. (e) Displays the DESeq2 and SCRAN method where the median of ratio is calculated for normalization. (f) The schema of between-sample normalization and the design of $\Omega_{i,r} = \omega_i \circ \omega_r^{-1}$.

such as gene-level regularization in SCTransform's GLM or quantile scaling in scNorm, can disrupt this order, potentially leading to unwanted variations (Booeshaghi et al., 2022).

4. RFNorm: Resistant Fit Normalization

4.1. Normalization function

In RFNorm, we consider two hypotheses about the measurement function ϕ: a zero-intercept linear scaling function implemented as RFNorm-L using a linear model through the origin, and a non-linear sigmoid function modeled using the Hill function, designated as RFNorm-NL.

For the nonlinear case, the corresponding normalization function ω_{hill}[2] for the non-linear model is formulated as:

$$\tilde{\phi}(\mathbf{x}) = \frac{a\mathbf{x}^n}{b^n + \mathbf{x}^n}, \quad \omega_{\text{hill}}(\mathbf{y}) = \tilde{\phi}^{-1}(\mathbf{y}) = \left(\frac{b^n \mathbf{y}}{a - b}\right)^{\frac{1}{n}}. \tag{4}$$

Here, a represents the upper bound of the function, b describes the x value at which the response is half-maximal, and n is the Hill coefficient, indicating the degree of cooperativity.

As a between-sample normalization approach, a mapping function Ω for every cell i and reference cell r can be constructed as:

$$\Omega_{i,r}(\mathbf{y}_i) = \omega^{-1}_{\text{hill},r}(\omega_{\text{hill},i}(\mathbf{y}_i)) = \frac{\mathbf{y}_i}{\beta \mathbf{y}_i + \alpha} \tag{5}$$

where α and β are reparameterized constants detailed in Appendix A.1. When $\beta = 0$, Ω simplifies to a ratio scaling normalization method.

4.2. Resistant Fit for Robust Optimization

Conventionally, parameters in Ω are estimated by minimizing the empirical risk, as described in Eq 1. However, this approach is vulnerable to outliers or extreme values, necessitating a robust optimization strategy (Diakonikolas et al., 2021; Zelditch et al., 2012). Previous methods such as DESeq2 and Scran use median of ratios (Eq 3) as a form of robust estimate to guard against out-of-distribution values like highly expressed genes and technical zeros. Nevertheless, the median discards informative values that are in-distribution and also do not follow a risk-minimization principle. To address this, we propose using a "resistant fit" risk minimization strategy:

$$\arg\min \sum_{g}^{|B|} ||\Omega_{i,r}(y_{i,g}) - y_{r,g}||^2 \tag{6}$$

where $|B|$ is a subset of features that is smaller than G, the total number of genes. We designate B as the biological feature set and hypothesize that it contains in-distribution points (features) when mapping one cell to another by Assumption 3. Ideally, we would evaluate all subsets B such that $|B|/G = p$ for some fixed proportion p[3]. However, exploring all $\binom{G}{|B|}$ possible subsets relative to the reference cell for every cell is computationally prohibitive. Therefore, we implemented an iterative heuristic solution, as detailed in (Rohlf & Slice, 1990) and (Zelditch et al., 2012), within the `ResistantFit` func-

[2]We analyze three types of common sigmoid functions: Hill function, Geompertz function, and Logistic function. Refer to Appendix A for details.

[3]For example, following the precedent set by EdgeR, which trims 30% of M-values, one might consider $|\mathbf{B}|/G = 0.7$.

tion. Initially, $\Omega_{i,r}$ is fitted to a selected feature set seeded by `InitialFeatureSelection` function. The fitted model is then applied to all gene pairs $(y_{i,g}, y_{r,g})$. These pairs are ranked in ascending order based on their residuals relative to $\hat{\Omega}_{i,r}$. Only the top-ranked $|B|$ points are used in the subsequent iteration to fit $\Omega_{i,r}$. This iterative process continues until the change in deviations fall below a predefined stopping criterion, as depicted in Algorithm 1. This process iteratively refines the normalization by fitting the resistant fit regression between every cell and the reference cell, thus normalizing the entire dataset.

Algorithm 1 RFNorm Algorithm

Input: $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$
$\mathbf{y}_r \leftarrow \text{CreateReference}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$
for $i = 1$ **to** n **do**
  ifs $\leftarrow \text{InitialFeatureSelection}(\mathbf{y}_i)$
  $\mathbf{y}'_i \leftarrow \text{ResistantFit}(\mathbf{y}_i, \mathbf{y}_r, \text{ifs})$
end for
Output: $(\mathbf{y}'_1, \mathbf{y}'_2, \ldots, \mathbf{y}'_n)$

4.3. Initialization

In accordance with our Assumption 4 that measurement instruments should preserve monotonic input-output relationships, we initialize the biological feature set B with a maximally monotonic subset. This strategy involves selecting genes that exhibit consistent orders of expression levels across different cells, effectively reducing noise in the data. This preselection serves as a preliminary data cleaning step, mitigating the risk of normalization artifacts.

We implement this order-preserving strategy by identifying the longest common subsequence (LCS) of genes that display continuous expression between any two cells, using dynamic programming. This ensures that only genes with stable and predictable behavior are included in the initial set B, which is crucial for the effective application of the normalization process.

5. Experiments on Simulation

5.1. Compositional Bias

We adapted the experimental setups from (L. Lun et al., 2016; Evans et al., 2018) to demonstrate compositional bias in scRNA-seq datasets and to illustrate that within-sample normalization methods such as CP10K are incapable of adequately correcting this bias.

We defined two subpopulations, $s = \{A, B\}$, representing different cell types or biological conditions, such as drug treatment or environmental stimulation. For each gene g in cell j, the true counts x_{gj} were sampled from a negative binomial (NB) distribution with mean $\boldsymbol{\lambda}_s$. The factor θ_i represents cell-specific effects during measurement, such as capture efficiency. The term $\sum_g \mathbf{x}_i$ in the denominator accounts for the sampling nature of the NGS procedure and is the source of compositional bias and Simplex relations. Here, we simplify the measurement effect as a scaling factor θ_i, but we discuss its generalization in sections 3.4.

We assumed each cell has $10,000$ genes by the following setup. Parameters $(\mu_g, \varphi_g, \varphi_g, \mu_l, \varphi_l, \mu_d, \varphi_d)$ are estimated by Splatter (Zappia et al., 2017) based on "33k PBMCs from a Healthy Donor, v1 Chemistry" from 10x Genomics.

The variable γ_{gs} represents the group- and gene-specific fold-changes, which indicate changes in some genes' expression levels due to biological conditions. Differential Expression (DE) involves variations in gene expression levels between different conditions or cell types. Genes are classified as differentially expressed based on their fold-changes relative to control conditions, modeled by γ_{gs} as follows:

$$\begin{aligned} \log_2 \gamma_{gs}(\text{regular}) &= 1.2 \\ \gamma_{gs}(\text{extreme}) &= 10 \\ \gamma_{gs}(\text{non-DE}) &= 1 \end{aligned}$$

In group A, 10% of the genes were randomly chosen as DE genes. In group B, 25% were selected as DE genes. Among the DE genes, 85% are regular DEs and 15% are set as extreme responses with 10 times the fold-changes. The rest are non-DEs. To this end, the simulation setup can be written as follows.

$$\begin{aligned} \lambda_{g0} &\sim \text{Gamma}(\mu_g, \varphi_g) \\ \lambda_{gs} &= \gamma_{gs} \lambda_{g0} \\ \mathbf{x}_i &\sim \text{NB}(\boldsymbol{\lambda}_s, \varphi_g) \\ \mathbf{y}_i &= \theta_i \mathbf{x}_i / \sum_g \mathbf{x}_i \\ \log_2 \theta_j &\sim \mathcal{N}(\mu_l, \varphi_l) \end{aligned}$$

We simulate three sets of cells $A1$, $A2$, and B. $A1$ and $A2$ are replicates of Group A. We then form two sets of cell pairs. Set A/A has 100 pairs of cells from group $A1$ and $A2$. Set A/B has 100 pairs of cells from group A and B. We normalize each group by CP10K. For each gene, we calculate the ratio of expression for each pair of cells. For each non-DE genes, the ratios of groups are shown in the violin plot. We show the violin plot of one gene in Fig 4 (b). Ideally, the same non-DE genes in the groups should have same expression levels. However, in Fig 1, we see that the ratio of non-DE genes in group A/B is higher than group A/A with p-val < 0.05.

5.2. Outliers

We employed the same experimental setup as outlined in Section 5.1, with the exception of removing the composi-

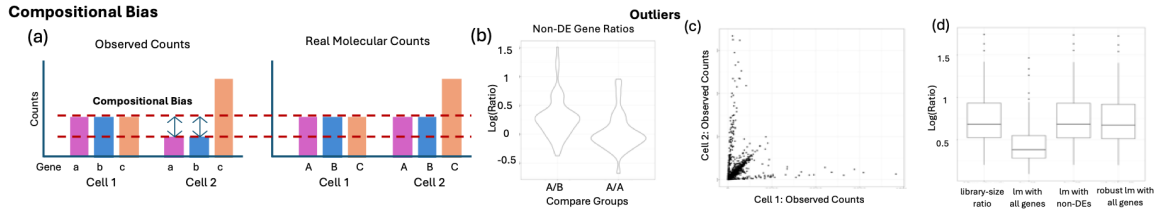


Figure 4. Compositional Bias and Outlier Effects in scRNA-seq Normalization. (a) illustrates the observed counts for three genes in two cells, highlighting the discrepancy between observed and real molecular counts due to compositional bias. (b) presents the distribution of log fold-changes for a non-DE gene between two comparison groups. (c) is a scatter plot illustrating the relationship between observed counts with outliers. (d) shows log fold-change ratios for linear regression and robust linear regression.

tional bias component $\sum_g \mathbf{x}_i$ to focus exclusively on the impact of outliers when fitting composed normalization function Ω between any two pairs of cells. Thus, the outliers would be the pairs of genes $(y_{g,i}, y_{g,j})$, which are DE genes. Figure 4(c) displays a dot plot of gene expression from two simulated cells, where Cell 1 is from group A and Cell 2 is from group B. This experiment was repeated with 100 pairs of cells, and the expression ratios are depicted as a box plot in Figure 4(d).

In this scenario, the ratio of size factors θ_i serves as the ground truth for the relationship between non-DE genes across two cells. Estimating this relationship solely based on non-DE genes allows us to accurately recover the true log-ratios. However, fitting a linear model across all genes typically leads to an underestimation of the log fold-change due to the influence of outliers. By employing a robust linear model, such as RFNorm-linear, we successfully align the log-fold-change ratio with the ground truth, demonstrating the efficacy of robust modeling techniques in managing outliers and ensuring accurate data interpretation.

6. Apply on Real Data

6.1. Invariance of KNN Graph

The efficacy of a preprocessing approach relies on the downstream analysis tasks. The preservation of lower-dimensional mathematical structures, particularly the k-nearest neighbor (KNN) graph, is critical in studying various cell types and states (Ahlmann-Eltze & Huber, 2023; Xia et al., 2023). We adopted the Consistent Benchmark from (Ahlmann-Eltze & Huber, 2023) to evaluate RFNorm alongside the other two commonly used depth normalization methods, CP10K and Scran by comparing the overlap of edges in the KNN graphs post-normalization, as shown in Table 1. It is notable that CP10K might show an inherently higher overlap since count-per-million (CPM) is included in the 22 transformations reviewed in Consistent Benchmark.

The benchmark encompasses three deeply sequenced datasets using msSCRB and Smart-seq3 technologies (Bagnoli et al., 2018; Larsson et al., 2021; Johnsson et al., 2022), each with an average sequencing depth above 50,000 UMIs per cell. A consensus k-NN graph ($k = 50$) is constructed from these datasets after applying 22 transformations, based on the top-50 principal components (PCs) as documented in (Ahlmann-Eltze & Huber, 2023) and Table 1, `Common` column. To emulate the sequencing depth typical of 10x Genomics, the datasets were downsampled to 5,000 counts, with the results presented in the `Reduced` rows of Table 1.

RFNorm demonstrates consistently high overlap in k-NN graphs across all datasets, effectively preserving biological relationships after normalization. This is evident in both full and reduced datasets, indicating RFNorm's capability to accurately fit the composed transfer function Ω. In full datasets, RFNorm-L and RFNorm-NL yielded results comparable to Scran. Particularly in smaller datasets, such as those from mESC cells, RFNorm-NL matched or even surpassed the performance of CP10K, approaching what is considered ground truth. In reduced datasets, RFNorm generally outperformed both CP10K and Scran in the Fibroblasts dataset, with RFNorm-NL showing superior effectiveness over RFNorm-L. However, in the mESC dataset, Scran proved to be the most effective, likely due to its ability to better estimate size factors in smaller, more distinctly clustered cell populations. In contrast, the lack of clear clustering in the Fibroblasts dataset may reduce Scran's effectiveness, as detailed in Appendix Fig. 8. Nonlinear models like RFNorm-NL consistently provided better results than linear models, emphasizing their advantage in handling complex biological data. The siRNA KD dataset, with its larger cell count, showcased the scalability of these normalization methods, with RFNorm maintaining high graph overlap even when data formats were reduced.

6.2. Marker Gene Identification Validation

To assess the effectiveness of normalization in identifying marker genes, we utilized scRNA-seq data from patients with pulmonary fibrosis and healthy controls, previously analyzed by (Reyfman et al., 2019). From (Squair et al.,

Table 1. Comparison of k-Nearest Neighbor (k-NN) Graph Overlap across Different Normalization Methods.The k-NN graphs (k=50) were built based on the top-50 principal components for each normalization method. The downsampled experiments were repeated 5 times.

Dataset	N Cells		Number of Overlapped Edges in KNNs				
			Common	CP10K	SCRAN	RFNorm-L	RFNorm-NL
mESC (mcSCRB)	249	Full	1085	972	941	942	970
		Reduced		782(9.25)	**873**(10.4)	803(20.33)	812(37.35)
Fibroblasts (ss3)	369	Full	1365	1342	1220	1260	1257
		Reduced		603(22.39)	517(12.88)	767(26.48)	**835**(35.91)
siRNA KD (ss3)	4298	Full	7194	7187	7186	7189	7187
		Reduced		6832.9(19.26)	6854.5(22.75)	**6958.5**(18.28)	6499 (29.39)

2021), marker genes identified in bulk RNA-seq were used as benchmarks to evaluate the accuracy of detecting cell-type-specific markers in scRNA-seq data. We processed subsets of 750, 1000, and 1500 cells from each group using CP10K, Scran, RFNorm-L, and RFNorm-NL, with comprehensive results detailed in Appendix D.

Furthermore, to confirm that the normalization methods accurately control for false positives, we conducted a Type 1 error control test using a "null dataset" derived from "16-cell stage blastomeres" (Deng et al., 2014), following the protocol described by (Soneson & Robinson, 2018). This experiment aimed to verify that the methods do not erroneously identify genes as differentially expressed under controlled conditions. Findings from this test are detailed in Appendix E.

Both sets of analyses demonstrated the comparable performance of RFNorm in maintaining accuracy and managing Type 1 error rates in the detection of biologically relevant markers in scRNA-seq data.

6.3. Monotonicity

An effective normalization technique must preserve the data structure by ensuring monotonicity across cells (Booeshaghi et al., 2022). To evaluate how well different normalization methods maintain this property, we employed Spearman's rank correlation to measure the extent of monotonicity in cellular data from the Fibroblast Dataset (Hagemann-Jensen et al., 2020), both before and after normalization. This metric helps in identifying any potential non-monotonic alterations introduced by normalization techniques. As shown in Figure 6.3, the analysis reveals that techniques such as SCTransform's GLM regularization and SCNorm's Quantile scaling may negatively impact the preservation of monotonicity in data structure.

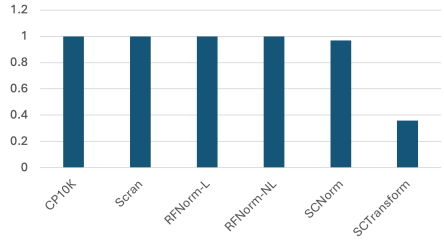


Figure 5. Analysis of Spearman's rank correlation to assess monotonicity preservation in different normalization methods. Lower values indicate greater disruption of data structure.

7. Discussion

Existing literature reveals a wide array of methods for normalizing single-cell transcriptome data. However, these methods often fail to explicitly consider learning an optimal function under a clear objective, particularly with respect to aligning "comparables" (i.e., gene sets considered to be sampled from a common distribution). Here, we propose a new framework that explicitly addresses these issues.

In our study, we established a benchmark to evaluate the effectiveness of depth normalization methods in preserving the structure of k-NN graphs. RFNorm methods consistently showed enhanced performance in maintaining biological relationships across varying complexities and sequencing depths. RFNorm-NL was more proficient in smaller cell populations and in datasets with reduced sequencing depth, whereas Scran showed its strength primarily in the mESC dataset, characterized by a limited number of cells and distinct clusters. In the context of larger and more intricate datasets, such as the siRNA KD dataset, a noticeable performance dip in RFNorm-NL within the reduced dataset was observed, suggesting potential areas for refinement in managing large-scale data. Further investigation is warranted to determine how optimal reference cell selection during optimization can minimize information loss in such complex datasets.

Normalization is surprisingly difficult with many subtle issues in practice (e.g., handling an abundance of zero observations). Here, we suggest that approaching the problem from the perspective of the nature of the measurement function, optimization of an appropriate empirical risk function, and alignment of comparable data (i.e., accounting for out-of-distribution values), provides a broad framework to construct principled methods for normalization.

References

Ahlmann-Eltze, C. and Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, 20(5):665–672, May 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01814-1. URL https://www.nature.com/articles/s41592-023-01814-1. Number: 5 Publisher: Nature Publishing Group.

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., and Hicks, S. C. Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, 17(2):137–145, February 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0654-x. URL https://www.nature.com/articles/s41592-019-0654-x. Publisher: Nature Publishing Group.

Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biology*, 11 (10):R106, October 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-10-r106. URL https://doi.org/10.1186/gb-2010-11-10-r106.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., and Kendziorski, C. SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, 14(6):584–586, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4263. URL https://www.nature.com/articles/nmeth.4263. Number: 6 Publisher: Nature Publishing Group.

Bagnoli, J. W., Ziegenhain, C., Janjic, A., Wange, L. E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., and Enard, W. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature Communications*, 9 (1):2937, July 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05347-6. URL https://www.nature.com/articles/s41467-018-05347-6. Number: 1 Publisher: Nature Publishing Group.

Booeshaghi, A. S., Hallgrímsdóttir, I. B., Gálvez-Merchán, A., and Pachter, L. Depth normalization for single-cell genomics count data, May 2022. URL https://www.biorxiv.org/content/10.1101/2022.05.06.490859v1. Pages: 2022.05.06.490859 Section: New Results.

Choudhary, S. and Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology*, 23(1):27, January 2022. ISSN 1474-760X. doi: 10.1186/s13059-021-02584-9. URL https://doi.org/10.1186/s13059-021-02584-9.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*, 343(6167):193–196, January 2014. doi: 10.1126/science.1245316. URL https://www.science.org/doi/10.1126/science.1245316. Publisher: American Association for the Advancement of Science.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robustness meets algorithms. *Communications of the ACM*, 64(5):107–115, May 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3453935. URL https://dl.acm.org/doi/10.1145/3453935.

Evans, C., Hardin, J., and Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792, September 2018. ISSN 1477-4054. doi: 10.1093/bib/bbx008.

Hafemeister, C. and Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1874-1. URL https://doi.org/10.1186/s13059-019-1874-1.

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J. M., Faridani, O. R., and Sandberg, R. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology*, 38(6):708–714, June 2020. ISSN 1546-1696. doi: 10.1038/s41587-020-0497-0. URL https://www.nature.com/articles/s41587-020-0497-0. Number: 6 Publisher: Nature Publishing Group.

He, Z., Hu, S., Chen, Y., An, S., Zhou, J., Liu, R., Shi, J., Wang, J., Dong, G., Shi, J., Zhao, J., Ou-Yang, L., Zhu, Y., Bo, X., and Ying, X. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology*, pp. 1–12, January 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-02040-y. URL https://www.nature.

com/articles/s41587-023-02040-y. Publisher: Nature Publishing Group.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7): 539–542, July 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0033-z. URL https://www.nature.com/articles/s41592-018-0033-z. Number: 7 Publisher: Nature Publishing Group.

Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, March 2015. URL http://arxiv.org/abs/1502.03167. arXiv:1502.03167 [cs].

Johnsson, P., Ziegenhain, C., Hartmanis, L., Hendriks, G.-J., Hagemann-Jensen, M., Reinius, B., and Sandberg, R. Transcriptional kinetics and molecular functions of long noncoding RNAs. *Nature Genetics*, 54(3):306–317, March 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01014-1. URL https://www.nature.com/articles/s41588-022-01014-1. Number: 3 Publisher: Nature Publishing Group.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0. URL https://www.nature.com/articles/s41592-019-0619-0. Publisher: Nature Publishing Group.

L. Lun, A. T., Bach, K., and Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1): 75, April 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0947-7. URL https://doi.org/10.1186/s13059-016-0947-7.

Larsson, A. J. M., Ziegenhain, C., Hagemann-Jensen, M., Reinius, B., Jacob, T., Dalessandri, T., Hendriks, G.-J., Kasper, M., and Sandberg, R. Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. *PLOS Computational Biology*, 17(3):e1008772, March 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008772. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008772. Publisher: Public Library of Science.

Lause, J., Berens, P., and Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, 22(1):258, September 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02451-7. URL https://doi.org/10.1186/s13059-021-02451-7.

Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E. Z., and Welch, J. D. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature Protocols*, 15(11):3632–3662, November 2020. ISSN 1750-2799. doi: 10.1038/s41596-020-0391-8. URL https://www.nature.com/articles/s41596-020-0391-8. Publisher: Nature Publishing Group.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0229-2. URL http://www.nature.com/articles/s41592-018-0229-2. tex.ids= lopezDeepGenerativeModeling2018a.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL https://doi.org/10.1186/s13059-014-0550-8.

Luecken, M. D. and Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746. URL https://www.embopress.org/doi/full/10.15252/msb.20188746. Publisher: John Wiley & Sons, Ltd.

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., and Bonn, S. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, 11(1):166, January 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14018-z. URL https://www.nature.com/articles/s41467-019-14018-z. Publisher: Nature Publishing Group.

Maza, E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. *Frontiers in Genetics*, 7, 2016. ISSN 1664-8021. URL https://www.frontiersin.org/articles/10.3389/fgene.2016.00164.

Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., and Crowley, T. M. A field guide for the compositional analysis of any-omics data. *GigaScience*, 8 (9):giz107, September 2019. ISSN 2047-217X. doi: 10.

1093/gigascience/giz107. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6755255/.

Reyfman, P. A., Walter, J. M., Joshi, N., Anekalla, K. R., McQuattie-Pimentel, A. C., Chiu, S., Fernandez, R., Akbarpour, M., Chen, C.-I., Ren, Z., Verma, R., Abdala-Valencia, H., Nam, K., Chi, M., Han, S., Gonzalez-Gonzalez, F. J., Soberanes, S., Watanabe, S., Williams, K. J. N., Flozak, A. S., Nicholson, T. T., Morgan, V. K., Winter, D. R., Hinchcliff, M., Hrusch, C. L., Guzy, R. D., Bonham, C. A., Sperling, A. I., Bag, R., Hamanaka, R. B., Mutlu, G. M., Yeldandi, A. V., Marshall, S. A., Shilatifard, A., Amaral, L. A. N., Perlman, H., Sznajder, J. I., Argento, A. C., Gillespie, C. T., Dematte, J., Jain, M., Singer, B. D., Ridge, K. M., Lam, A. P., Bharat, A., Bhorade, S. M., Gottardi, C. J., Budinger, G. R. S., and Misharin, A. V. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 199(12):1517–1536, June 2019. ISSN 1535-4970. doi: 10.1164/rccm.201712-2410OC.

Robinson, M. D. and Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, March 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-3-r25. URL https://doi.org/10.1186/gb-2010-11-3-r25.

Rohlf, F. J. and Slice, D. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. *Systematic Zoology*, 39(1):40–59, 1990. ISSN 0039-7989. doi: 10.2307/2992207. URL https://www.jstor.org/stable/2992207. Publisher: [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.].

Soneson, C. and Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, April 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4612. URL https://www.nature.com/articles/nmeth.4612. Number: 4 Publisher: Nature Publishing Group.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., and Courtine, G. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, September 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25960-2. URL https://www.nature.com/articles/s41467-021-25960-2. Number: 1 Publisher: Nature Publishing Group.

Su, C., Xu, Z., Shan, X., Cai, B., Zhao, H., and Zhang, J. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nature Communications*, 14 (1):4846, August 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40503-7. URL https://www.nature.com/articles/s41467-023-40503-7. Publisher: Nature Publishing Group.

Sun, T., Song, D., Li, W. V., and Li, J. J. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, 22(1): 163, May 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02367-2. URL https://doi.org/10.1186/s13059-021-02367-2.

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, 14(4):381–387, April 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4220. URL https://www.nature.com/articles/nmeth.4220. Number: 4 Publisher: Nature Publishing Group.

Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S., and Shahrezaei, V. bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics*, 36 (4):1174–1181, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz726. URL https://doi.org/10.1093/bioinformatics/btz726.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4292. URL https://www.nature.com/articles/nmeth.4292. Number: 6 Publisher: Nature Publishing Group.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.05.061. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418307244.

Xia, L., Lee, C., and Li, J. J. scDEED: a statistical method for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters. *bioRxiv: The Preprint Server for Biology*, pp. 2023.04.21.537839, September 2023. doi: 10.1101/2023.04.21.537839.

Zappia, L., Phipson, B., and Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome*

Biology, 18(1):174, September 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1305-0. URL https://doi.org/10.1186/s13059-017-1305-0.

Zelditch, M. L., Swiderski, D. L., and Sheets, H. D. Geometric Morphometrics for biologists: A primer. In *Geometric Morphometrics for Biologists*, pp. 1–20. Elsevier, 2012. ISBN 978-0-12-386903-6. doi: 10.1016/B978-0-12-386903-6.00001-0. URL https://linkinghub.elsevier.com/retrieve/pii/B9780123869036000010.

A. Sigmoid Functions

We explored three types of common sigmoid functions. They are Hill function, Gompertz Function, and Logistic function. The corresponding normalization functions are listed in Table 2.

Table 2. Measurement and Normalization transfer functions for different sigmoid functions

Functions	Measurement $\phi(x)$	Normalization $\omega(y)$	Composed Function $\Omega_{i,r}(y_i)$
Hill	$\frac{ax^n}{b^n+x^n}$	$(\frac{b^n y}{a-b})^{\frac{1}{n}}$	$\frac{y_i}{\beta y_i+\alpha}$
Geompertz	$L\exp(-b\exp(-kx))$	$-\frac{1}{k}\log\left(\frac{1}{b}\log\frac{L}{y}\right)$	$\frac{L_r}{L_i^\beta}y_i^\beta$
Logistics	$\frac{a}{1+be^{tx}}$	$\frac{1}{t}\left(\log\left(\frac{a}{y}-1\right)-\log b\right)$	$\frac{y_i}{\beta y_i+\alpha}$

Hill function is commonly used in biochemistry and systems biology to describe the response of a system to a stimulus, typically in terms of the concentration of a substrate. It is especially used to describe cooperative binding of substrates to enzymes or receptors. As discussed in Section 4.1, for scRNA-seq, there are more constrains on the lower end of the range while the observation may fall within the higher-end of the operational range most of the time. So a function with flexiable lower end and a long operational range is perferable. As demonstrated in Fig 6, Hill function has the best flexibility and dynamic range to represent sequencing measurement. This is a qualitative check and needs more exploration to make strict conclusion. In table 2, we reported the composed transfer function based on different sigmoid functions after reprameterization. Hill function and logistic function are reduced to the equivalent format.

A.1. Hill function and Composed normalization function

Set Hill function $\tilde{\phi}$ as the estimated measurement function, we have

$$\tilde{\phi}(x) = \frac{ax^n}{b^n + x^n} \tag{7}$$

Here a is the upper bound of the function, b describes the x whose response is half-maximal, and n is the Hill coefficient, which describes the degree of cooperativity.

Consequently, the corresponding normalization function ω_{hill} is

$$\omega_{\text{hill}}(y) = \tilde{\phi}^{-1}(y) = (\frac{b^n y}{a-b})^{\frac{1}{n}}. \tag{8}$$

As in Fig.7, we construct the composed mapping function Ω for every cell i and r. The expression of Ω is given by:

$$\Omega_{i,r}(y) = \omega^{-1}_{\text{hill},r}(\omega_{\text{hill},i}(y)) \tag{9}$$

After some reformations (see Appendix A.4), we have

$$\Omega_{i,r}(y_i) = \frac{a_r}{1+\exp[\frac{n_r}{n_i}\log\frac{a_i-y_i}{y_i} - n_r\log\frac{b_r}{b_i}]} \tag{10}$$

Set the hill coefficient to be the same, i.e. $n_r = n_i$, we have

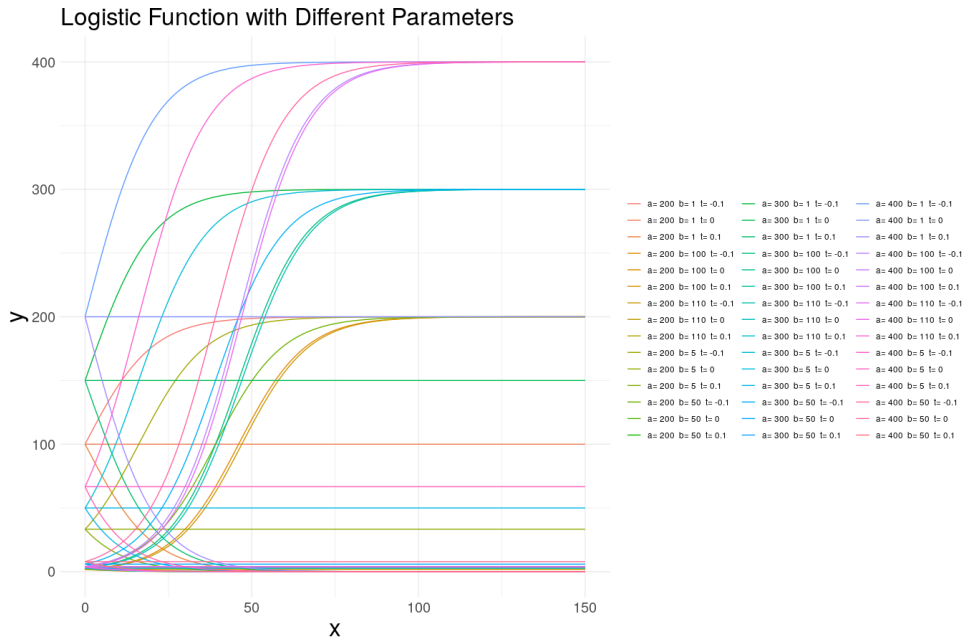
$$\Omega_{i,r}(y_i) = \frac{a_r(b_r/b_i)^{n_r}y_i}{((b_r/b_i)^{n_r}-1)y_i + a_i} \tag{11}$$

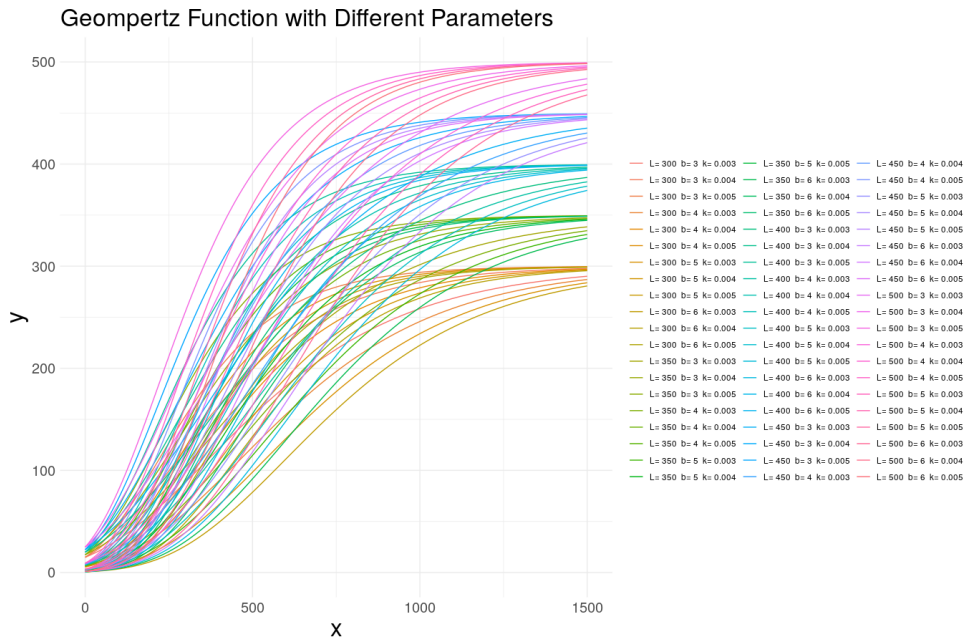
Without loss of generality, we can reparameterize the function as

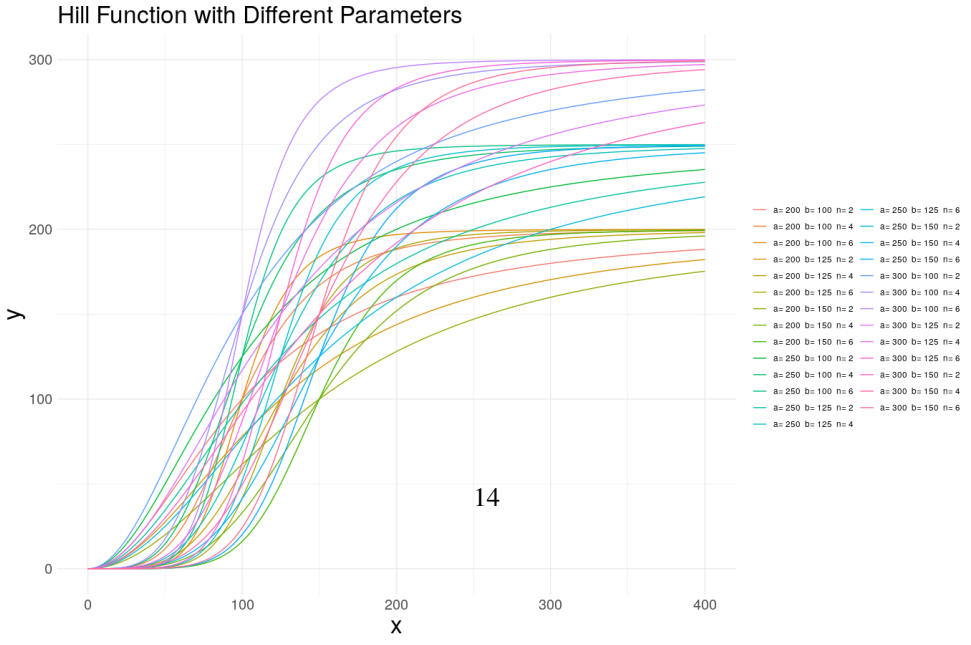
$$\Omega_{i,r}(y_i) = \frac{\lambda_1 y_i}{\lambda_2 y_i + \lambda_3} \tag{12}$$

, where $\lambda_1 = a_r b_i$, $\lambda_2 = b_i - b_r$, and $\lambda_3 = b_r a_i$. It can be further simplified as

$$\Omega_{i,r}(y_i) = \frac{y_i}{\beta y_i + \alpha} \tag{13}$$


Logistic Function with Different Parameters
y
x
0
100
200
300
400
0
50
100
150
a= 200 b= 1 t= -0.1
a= 200 b= 1 t= 0
a= 200 b= 1 t= 0.1
a= 200 b= 100 t= -0.1
a= 200 b= 100 t= 0
a= 200 b= 100 t= 0.1
a= 200 b= 110 t= -0.1
a= 200 b= 110 t= 0
a= 200 b= 110 t= 0.1
a= 200 b= 5 t= -0.1
a= 200 b= 5 t= 0
a= 200 b= 5 t= 0.1
a= 200 b= 50 t= -0.1
a= 200 b= 50 t= 0
a= 200 b= 50 t= 0.1
a= 300 b= 1 t= -0.1
a= 300 b= 1 t= 0
a= 300 b= 1 t= 0.1
a= 300 b= 100 t= -0.1
a= 300 b= 100 t= 0
a= 300 b= 100 t= 0.1
a= 300 b= 110 t= -0.1
a= 300 b= 110 t= 0
a= 300 b= 110 t= 0.1
a= 300 b= 5 t= -0.1
a= 300 b= 5 t= 0
a= 300 b= 5 t= 0.1
a= 300 b= 50 t= -0.1
a= 300 b= 50 t= 0
a= 300 b= 50 t= 0.1
a= 400 b= 1 t= -0.1
a= 400 b= 1 t= 0
a= 400 b= 1 t= 0.1
a= 400 b= 100 t= -0.1
a= 400 b= 100 t= 0
a= 400 b= 100 t= 0.1
a= 400 b= 110 t= -0.1
a= 400 b= 110 t= 0
a= 400 b= 110 t= 0.1
a= 400 b= 5 t= -0.1
a= 400 b= 5 t= 0
a= 400 b= 5 t= 0.1
a= 400 b= 50 t= -0.1
a= 400 b= 50 t= 0
a= 400 b= 50 t= 0.1


Geompertz Function with Different Parameters
y
x
0
100
200
300
400
500
0
500
1000
1500
L= 300 b= 3 k= 0.003
L= 300 b= 3 k= 0.004
L= 300 b= 3 k= 0.005
L= 300 b= 4 k= 0.003
L= 300 b= 4 k= 0.004
L= 300 b= 4 k= 0.005
L= 300 b= 5 k= 0.003
L= 300 b= 5 k= 0.004
L= 300 b= 5 k= 0.005
L= 300 b= 6 k= 0.003
L= 300 b= 6 k= 0.004
L= 300 b= 6 k= 0.005
L= 350 b= 3 k= 0.003
L= 350 b= 3 k= 0.004
L= 350 b= 3 k= 0.005
L= 350 b= 4 k= 0.003
L= 350 b= 4 k= 0.004
L= 350 b= 4 k= 0.005
L= 350 b= 5 k= 0.003
L= 350 b= 5 k= 0.004
L= 350 b= 5 k= 0.005
L= 350 b= 6 k= 0.003
L= 350 b= 6 k= 0.004
L= 350 b= 6 k= 0.005
L= 400 b= 3 k= 0.003
L= 400 b= 3 k= 0.004
L= 400 b= 3 k= 0.005
L= 400 b= 4 k= 0.003
L= 400 b= 4 k= 0.004
L= 400 b= 4 k= 0.005
L= 400 b= 5 k= 0.003
L= 400 b= 5 k= 0.004
L= 400 b= 5 k= 0.005
L= 400 b= 6 k= 0.003
L= 400 b= 6 k= 0.004
L= 400 b= 6 k= 0.005
L= 450 b= 3 k= 0.003
L= 450 b= 3 k= 0.004
L= 450 b= 3 k= 0.005
L= 450 b= 4 k= 0.003
L= 450 b= 4 k= 0.004
L= 450 b= 4 k= 0.005
L= 450 b= 5 k= 0.003
L= 450 b= 5 k= 0.004
L= 450 b= 5 k= 0.005
L= 450 b= 6 k= 0.003
L= 450 b= 6 k= 0.004
L= 450 b= 6 k= 0.005
L= 500 b= 3 k= 0.003
L= 500 b= 3 k= 0.004
L= 500 b= 3 k= 0.005
L= 500 b= 4 k= 0.003
L= 500 b= 4 k= 0.004
L= 500 b= 4 k= 0.005
L= 500 b= 5 k= 0.003
L= 500 b= 5 k= 0.004
L= 500 b= 5 k= 0.005
L= 500 b= 6 k= 0.003
L= 500 b= 6 k= 0.004
L= 500 b= 6 k= 0.005


Hill Function with Different Parameters
y
x
0
100
200
300
0
100
200
300
400
a= 200 b= 100 n= 2
a= 200 b= 100 n= 4
a= 200 b= 100 n= 6
a= 200 b= 125 n= 2
a= 200 b= 125 n= 4
a= 200 b= 125 n= 6
a= 200 b= 150 n= 2
a= 200 b= 150 n= 4
a= 200 b= 150 n= 6
a= 250 b= 100 n= 2
a= 250 b= 100 n= 4
a= 250 b= 100 n= 6
a= 250 b= 125 n= 2
a= 250 b= 125 n= 4
a= 250 b= 125 n= 6
a= 250 b= 150 n= 2
a= 250 b= 150 n= 4
a= 300 b= 100 n= 2
a= 300 b= 100 n= 4
a= 300 b= 100 n= 6
a= 300 b= 125 n= 2
a= 300 b= 125 n= 4
a= 300 b= 125 n= 6
a= 300 b= 150 n= 2
a= 300 b= 150 n= 4
a= 300 b= 150 n= 6

A.2. Geompertz function and Composed normalization function

Set Geompertz function $\tilde{\phi}$ as the estimated measurement function, we have

$$\tilde{\phi}(x) = L\exp(-b\exp(-kx)) \tag{14}$$

Here L is the upper bound of the function, b describes the horizaontal shift, and k affects the growth rate. Consequently, the corresponding normalization function $\omega_{\text{geompertz}}$ is

$$\omega_{\text{geompertz}}(y) = \tilde{\phi}^{-1}(y) = -\frac{1}{k}\log(\frac{1}{b}\log\frac{L}{y}). \tag{15}$$

We construct the composed mapping function Ω for every cell i and r. The expression of Ω is given by:

$$\Omega_{i,r}(y) = \omega^{-1}_{\text{geompertz},r}(\omega_{\text{geompertz},i}(y)) \tag{16}$$

Suppose growth rate k is non-cell specific, $k_r = k_i$ and $Set\beta = \frac{b_i}{b_r}$. After some reformations, we have

$$\Omega_{i,r}(y_i) = \frac{L_r}{L_i^{\beta}}y_i^{\beta} \tag{17}$$

$$\log\Omega_{i,r}(y_i) = \beta\log y_i + \log L_r - \beta\log L_i \tag{18}$$

A.3. Logistic function and Composed normalization function

Set Logistic function $\tilde{\phi}$ as the estimated measurement function, we have

$$\tilde{\phi}(x) = \frac{a}{1+be^{tx}} \tag{19}$$

Here a is the upper bound of the function, b describes the horizontal shift, and t is the growth rate.

Consequently, the corresponding normalization function ω_{logistic} is

$$\omega_{\text{logistic}}(y) = \tilde{\phi}^{-1}(y) = \frac{1}{t}\left(\ln(\frac{a}{y}-1) - \ln b_i\right) \tag{20}$$

We construct the composed mapping function Ω for every cell i and r. The expression of Ω is given by:

$$\Omega_{i,r}(y) = \omega^{-1}_{\text{logistic},r}(\omega_{\text{logistic},i}(y)) \tag{21}$$

Suppose $\lambda = \frac{t_r}{t_i}$, after some reformations (see Appendix A.5), we have

$$\Omega_{i,r}(y_i) = \frac{y_i^{\lambda}}{\frac{b_r}{a_r b_i^{\lambda}}(a_i - y_i)^{\lambda} + \frac{1}{a_r}y_i^{\lambda}} \tag{22}$$

Set the growth rate to be the same, i.e. $t_r = t_i$ and $\lambda = 1$, we have

$$\Omega_{i,r}(y_i) = \frac{y_i}{\frac{a_i b_r}{a_r b_i} + \frac{b_r + b_i}{a_r b_i}y_i} \tag{23}$$

Without loss of generality, we can reparameterize the function as

$$\Omega_{i,r}(y_i) = \frac{y_i}{\beta y_i + \alpha} \tag{24}$$

A.4. Reformation of composed Hill functions

$$
\begin{aligned}
\phi_r(y_r) =& \phi_i(y_i) \\
(\frac{b_r^{n_r} y_r}{a_r - y_r})^{\frac{1}{n_r}} =& (\frac{b_i^{n_i} y_i}{a_i - y_i})^{\frac{1}{n_i}} \\
\frac{1}{n_r} \log(\frac{b_r^{n_r} y_r}{a_r - b_r}) =& \frac{1}{n_i} \log(\frac{b_i^{n_i} y_i}{a_i - y_i}) \\
n_r \log b_r - \log \frac{a_r - y_r}{y_r} =& \frac{n_r}{n_i} \log \frac{b_i^{n_i} y_i}{a_i - y_i} \\
\log(\frac{a_r}{y_r} - 1) =& n_r \log b_r + \frac{n_r}{n_i} \log \frac{a_i - y_i}{b_i^{n_i} y_i} \\
y_r =& \frac{a_r}{1 + \exp[n_r \log b_r + \frac{n_r}{n_i} \log \frac{a_i - y_i}{b_i^{n_i} y_i}]} \\
y_r =& \frac{a_r}{1 + \exp[\frac{n_r}{n_i} \log \frac{a_i - y_i}{y_i} - n_r \log \frac{b_r}{b_i}]}
\end{aligned}
$$

A.5. Reformation of composed logistic functions

$$
\begin{aligned}
\omega_r(y_r) =& \omega_i(y_i) \\
\frac{1}{t_r} \left(\ln(\frac{a_r}{y_r} - 1) - \ln b_r \right) =& \frac{1}{t_i} \left(\ln(\frac{a_i}{y_i} - 1) - \ln b_i \right)
\end{aligned}
$$

Suppose $\lambda_1 = \frac{t_r}{t_i}$,

$$
\begin{aligned}
\ln \left(\frac{a_r - y_r}{b_r y_r} \right) =& \lambda_1 \ln \left(\frac{a_i - y_i}{b_i y_i} \right) \\
\frac{a_r}{b_r y_r} - \frac{1}{b_r} =& \frac{(a_i - y_i)^{\lambda_1}}{b_i^{\lambda_1} y_i^{\lambda_1}} \\
\frac{a_r}{y_r} =& \frac{b_r (a_i - y_i)^{\lambda_1}}{b_i^{\lambda_1} y_i^{\lambda_1}} + \frac{b_i^{\lambda_1} y_i^{\lambda_1}}{b_i^{\lambda_1} y_i^{\lambda_1}} \\
\frac{1}{y_r} =& \frac{b_r (a_i - y_i)^{\lambda_1} + b_i^{\lambda_1} y_i^{\lambda_1}}{a_r b_i^{\lambda_1} y_i^{\lambda_1}} \\
y_r =& \frac{y_i^{\lambda_1}}{\frac{b_r}{a_r b_i^{\lambda_1}} (a_i - y_i)^{\lambda_1} + \frac{1}{a_r} y_i^{\lambda_1}}
\end{aligned}
$$

Assume $t_r = t_i$, i.e. $\lambda_1 = 1$. We have

$$
\begin{aligned}
\frac{a_r - y_r}{b_r y_r} &= \frac{a_i - y_i}{b_i y_i} \\
\frac{a_r}{b_r y_r} - \frac{1}{b_r} &= \frac{a_i}{b_i y_i} - \frac{1}{b_i} \\
\frac{a_r b_i}{y_r} - b_i &= \frac{a_i b_r}{y_i} - b_r \\
\frac{a_r b_i}{y_r} &= \frac{a_i b_r}{y_i} - b_r + b_i \\
\frac{a_r b_i}{y_r} &= \frac{a_i b_r + (b_r + b_i) y_i}{y_i} \\
y_r &= \frac{a_r b_i y_i}{a_i b_r + (b_r + b_i) y_i} \\
y_r &= \frac{y_i}{\frac{a_i b_r}{a_r b_i} + \frac{b_r + b_i}{a_r b_i} y_i}
\end{aligned}
$$

B. Construction of the reference cell

A reference cell need to be selected as the baseline. Different methods make distinct choices in this regard. For example, DESeq2 employs the geometric mean of the dataset as the reference, EdgeR arbitrarily selects one sample for the role, and Scran utilizes an averaged pseudo-cell.

In our proposed approach, we opt for the median of the dataset to serve as the reference cell, denoted as $\mathbf{y}_r$. We argue that using the median offers a balance, providing a central tendency that is robust to outliers. Unlike the geometric mean, which can be sensitive to zero values, or an arbitrary selection that could introduce bias, the median offers a more stable and unbiased point of reference.

C. Invariance of KNN Graph

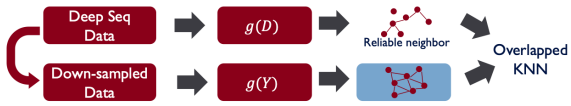


Figure 7. Schema of KNN Consistent Benchmark

mESC dataset (Bagnoli et al., 2018) consist of 249 single mESCs using SCRB-seq. Fibroblasts (Hagemann-Jensen et al., 2020) has 369 individual primary mouse fibroblasts with Smart-seq3. siRNA KD dataset (Larsson et al., 2021) has 4298 fibroblasts siRNA knockdown cells from CAST/EiJ and C57BL/6J mouse strains (5 animals). The tSNE plots of the transformed dataset are in Fig.8.

D. Concordance with Bulk Data

We assessed the efficacy of cell-type-specific marker detection using scRNA-seq data from patients with pulmonary fibrosis and healthy controls, initially analyzed by (Reyfman et al., 2019) and (Squair et al., 2021). This scRNA-seq dataset was compared against a bulk RNA-seq experiment, using the list of marker genes identified in the bulk data as a benchmark. We analyzed randomly selected subsets of 750, 1000, and 1500 cells from each group after normalizing the data using CP10K, Scran, RFNorm-L, and RFNorm-NL. Differential expression was evaluated using a t-test, with significant findings defined by adjusted p-values below 0.05. According to our analysis, summarized in Table 3, all four methods demonstrated similar performance in detecting differentially expressed genes.

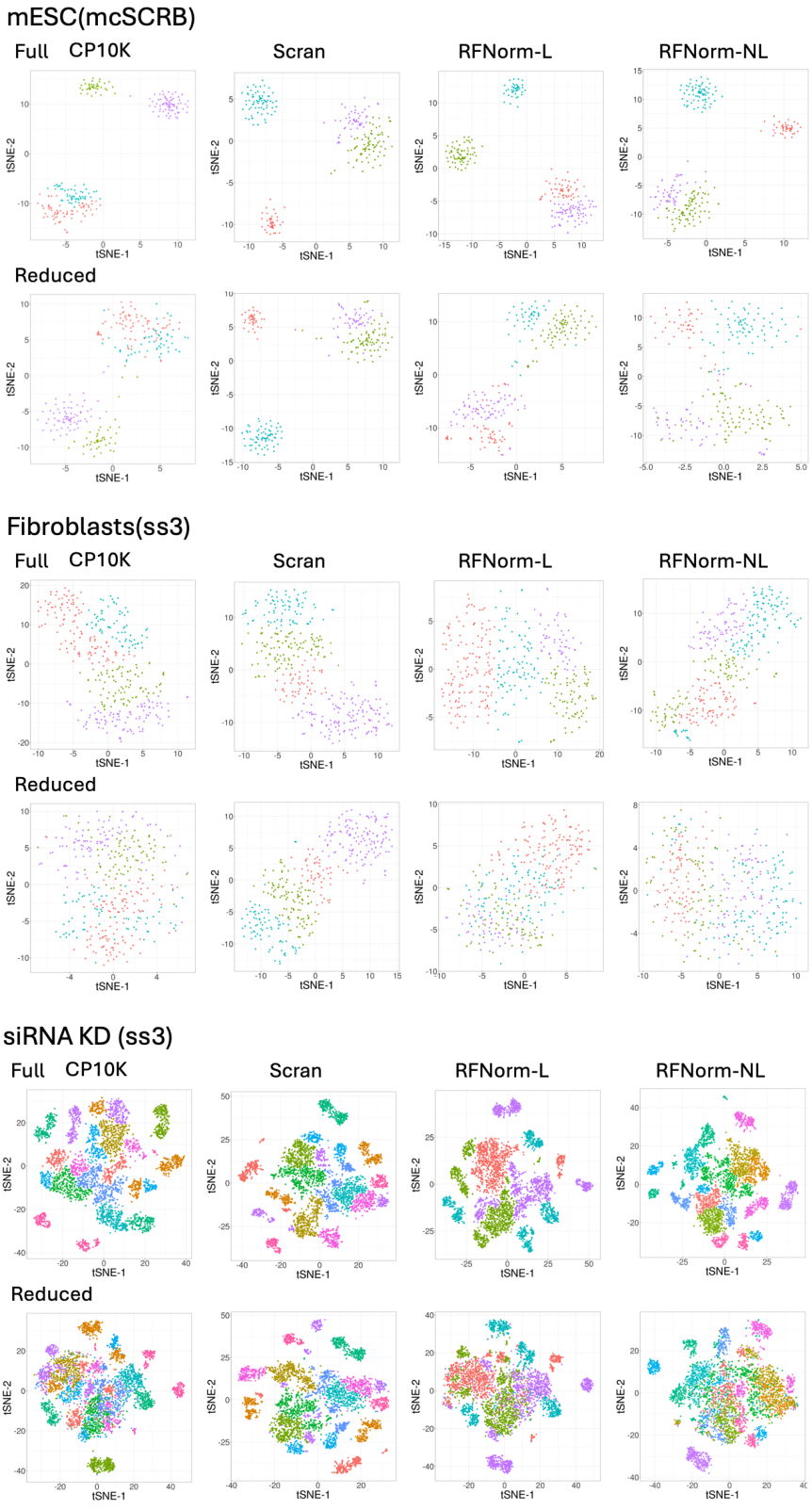


Figure 8. tSNE plots of each dataset after transformations. Cells are colored by clustering using the walktrap clustering algorithm as (Ahlmann-Eltze & Huber, 2023). The reduced dataset is colored by the clusters identified in full dataset.

Table 3. Differential Expression Analysis Results for Various Normalization Methods. Genes with adjusted p-values < 0.05 are considered DE. If the DE is also discovered in the paired bulk-RNA experiment, it is counted as concordance.

N Cells	Type	CP10K	SCRAN	RFNorm-L	RFNorm-NL
1500	# DE	1163.6 (34.98)	1168.4 (24.8)	1206 (34.79)	1166.4 (31.94)
	# Concordance	170.4 (7.19)	169.4 (8.38)	169.4 (7.95)	168.6 (6.58)
1000	# DE	843.6 (50.97)	840.6 (50.48)	884.2 (53.4)	848 (54.55)
	# Concordance	138.4 (11.28)	136.6 (10.11)	139.2 (11.78)	137.6 (11.65)
750	# DE	646.4 (26.14)	648.8 (24.47)	685 (33.41)	653.6 (27.81)
	# Concordance	118.2 (8.22)	117.4 (7.44)	119.6 (7.6)	116 (6.63)

E. Type 1 Error Control

It is essential that the normalization factors ensure genes with identical expression levels in two different samples are not mistakenly identified as differentially expressed (Robinson & Oshlack, 2010). In our study, we utilized a 'null dataset' comprising 50 cells from the '16-cell stage blastomere', as described in (Deng et al., 2014), following the approach introduced by (Soneson & Robinson, 2018). After applying four different transformations, none of the genes exhibited an adjusted p-value (p-adj) below 0.05, indicating no differential expression. This finding aligns with the observations in (Soneson & Robinson, 2018), where both the t-test and Wilcoxon rank sum test were shown to effectively control Type-1 error across various transformations.