Bayesian inference of the gene expression states of single cells from scRNA-seq data

Jérémie Breda[1,2], Mihaela Zavolan[1,2], and Erik van Nimwegen*[1,2]

[1] *Biozentrum, University of Basel, Basel, Switzerland*
[2] *Swiss Institute of Bioinformatics, Basel, Switzerland*

Abstract

In spite of a large investment in the development of methodologies for analysis of single-cell RNA-seq data, there is still little agreement on how to best normalize such data, i.e. how to quantify gene expression states of single cells from such data. Starting from a few basic requirements such as that inferred expression states should correct for both intrinsic biological fluctuations and measurement noise, and that changes in expression state should be measured in terms of fold-changes rather than changes in absolute levels, we here derive a unique Bayesian procedure for normalizing single-cell RNA-seq data from first principles. Our implementation of this normalization procedure, called Sanity (SAmpling Noise corrected Inference of Transcription activitY), estimates log expression values and associated errors bars directly from raw UMI counts without any tunable parameters.

Comparison of Sanity with other recent normalization methods on a selection of scRNA-seq datasets shows that Sanity outperforms other methods on basic downstream processing tasks such as clustering cells into subtypes and identification of differentially expressed genes. More importantly, we show that all other normalization methods present severely distorted pictures of the data. By failing to account for biological and technical Poisson noise, many methods systematically predict the lowest expressed genes to be most variable in expression, whereas in reality these genes provide least evidence of true biological variability. In addition, by confounding noise removal with lower-dimensional representation of the data, many methods introduce strong spurious correlations of expression levels with the total UMI count of each cell as well as spurious co-expression of genes.

Introduction

In the past decade much effort has been invested in adapting methods for quantifying transcriptome and epigenome state on a genome-wide scale to the single-cell level. This has led to a large number of new methods that are starting to make it possible to track the states of single cells across tissues and embryos as they are developing, measuring transcriptomes, chromatin state, chromatin conformation, and cell lineages, sometimes in parallel [1–22]. Many in the field believe that these single-cell methods will revolutionize our understanding of the ways in

*Corresponding author. E-mail: erik.vannimwegen@unibas.ch

which cell fate, cell identity and developmental processes are regulated, and major consortia are starting to form that aim to comprehensively map single-cells in model organisms [23, 24].

In order to fullfil the promise of these single-cell measurement technologies, it will be crucial that computational methods are available that unambiguously extract what the raw measurements say about the state of the single cells in terms of concrete physical quantities. We not only want to be able to integrate results of single-cell RNA-seq (scRNA-seq) measurements, which we will focus on in this work, from different labs using different protocols, but also across measurements from entirely different measurement technologies such as FISH (e.g. [25]). In order to make that possible, the expression values that we extract from scRNA-seq data should correspond to physically meaningful quantities that can be directly compared with measurements of the same quantities made with other experimental methods. In addition, the estimated values of the concrete physical quantities should follow directly from the experimental data together with as small a number of additional assumptions as possible, and not depend on arbitrary parameters that the user can set at will. Moreover, in order to be able to determine when different measurements are mutually consistent, all estimates should be accompanied by meaningful error bars.

However, although there has been a veritable explosion of scRNA-seq analysis tools in recent years, there has been almost no attention given to satisfying these objectives. Instead of there being a small number of transparent methods that provide unambiguous estimates of quantities with clear physical interpretation, we find a large number of *ad hoc* methods that apply highly complex transformations to the data to perform combinations of tasks including imputation/normalization, clustering, dimensionality reduction, pseudo-time and trajectory inference, and visualization. These methods typically have many tunable parameters, produce outputs in highly abstract spaces that lack clear biological meaning, and are often even stochastic, such that different runs on the same data with the same parameters result in different output. For example, probably the most popular tools for visualizing scRNA-seq data are t-SNE [26] and UMAP [27], which are both stochastic, involve several parameters, and position cells in a lower dimensional space whose dimensions lack biological interpretation.

We here focus on the relatively basic task of normalization/imputation of single-cell gene expression states from raw scRNA-seq transcript counts. Using only minimal assumptions we derive from first principles a Bayesian method that corrects not only for the finite sampling associated with the capture and sequencing of mRNAs, but also for the Poisson noise inherent in the gene expression process itself. Our method, which we call Sanity (SAmpling Noise corrected Inference of Transcription activitY) is deterministic, has zero tunable parameters, and provides error-bars for all its estimates.

After motivating and explaining our method, we compare Sanity with a selection of popular methods for imputation/normalization from the recent literature and show that only Sanity can meaningfully remove Poisson sampling fluctuations and infer the true variation in gene expression intensity of each gene across cells. In addition, we show that all other methods we tested introduce severe distortions of the data such as inducing strong correlations between expression estimates and total UMI count of cells, or inferring strong co-expression between large numbers of genes when none is evident in the data. In addition, we show that Sanity's estimated expression levels outcompete those of other methods on both downstream clustering and differential expression tasks.

Methods

A Bayesian method for inferring gene expression states from count data

We first motivate and explain how we represent gene expression states of single cells, and what concrete physical quantities these gene expression states correspond to. After that, we introduce our method's probabilistic model of a scRNA-seq experiment, calculating how the expression state of the cell determines the probabilities of obtaining particular raw transcript counts, and then discuss how we solve the Bayesian model and the outputs that the method provides.

Defining gene expression states

For any given cell c, we want to represent its 'gene expression state' by a vector $\vec{e}_c$, whose components e_{gc} quantify how strongly each gene g is expressed in the cell. We want these gene expression states to satisfy two basic criteria. First, these gene expression states should have concrete physical interpretation. Second, for downstream processing we want that the differences $e_{gc} - e_{gc'}$ meaningfully reflect the change in expression of gene g between cells c and c' such that the Euclidean distance $d_{cc'} = \sqrt{\sum_g (e_{gc} - e_{gc'})^2}$ between two cells c and c' meaningfully reflects the difference in their gene expression states.

One might think that we could simply take the vector $\vec{m}_c$ of the actual number of mRNAs m_{gc} that exist in cell c for each gene g as the gene epression state of the cell. However, even for cells in the same gene expression state, the number of mRNAs will exhibit stochastic fluctuations. Imagine a gene that is transcribed at a constant rate λ in every cell, and with a constant rate of mRNA decay μ in every cell. The actual number of mRNAs m across cells will then follow a Poisson distribution with mean $a = \lambda/\mu$ which we call its 'transcription activity'. That is, the probability to find m mRNAs is $P_m = a^m e^{-a}/m!$ which has mean $\langle m \rangle = a$ and variance $\text{var}(m) = a$. Thus, instead of assuming any change in mRNA number m reflects a change in gene expression state, it makes more sense to identify changes in gene expression state with changes in the transcription activity a.

Note that mRNA numbers will show Poisson fluctuations in much more general situations than constant rates of transcription and decay [28]. Imagine that, in a particular cell c, both the rate of transcription and mRNA decay of a given gene g has fluctuated in some arbitrary way in time, with $\lambda_{gc}(t)$ the transcription rate a time t in the past, and $\mu_{gc}(t)$ the mRNA decay rate a time t in the past. The expected number of mRNAs $\langle m_{gc} \rangle$ is then given by the transcription activity

$$\langle m_{gc} \rangle = \int_0^\infty \lambda_{gc}(t) \exp\left[-\int_0^t \mu_{gc}(s) ds \right] dt \equiv a_{gc}, \tag{1}$$

which is a weighted average of the transcription rate of the gene in the recent past, i.e. on the time-scale that its mRNAs have turned over. Given this expected mRNA number a_{gc} the distribution of the actual number of mRNAs m_{gc} is still Poisson. That is the probability to obtain m_{gc} mRNAs is

$$P(m_{gc}|a_{gc}) = \frac{(a_{gc})^{m_{gc}}}{m_{gc}!} e^{-a_{gc}}. \tag{2}$$

We thus propose that we should use changes in vectors of transcription activity $\vec{a}_c$ to represent changes in gene expression state.

In addition, we propose to characterize the gene expression state of a cell not by the vector $\vec{a}_c$ of absolute transcription activities a_{gc}, but by the vector $\vec{\alpha}_c$ of *relative* transcription activities, with

$$\alpha_{gc} = \frac{a_{gc}}{\sum_{g'} a_{g'c}}, \tag{3}$$

which we will refer to as *transcription quotients*. First, it has been shown that, as cell volume increases, cells globally upregulate transcription to maintain approximately constant mRNA concentration [29] so that transcriptional activities a_{gc} of all genes are generally expected to scale with cell volume. We argue that a global change in transcriptional activities by a common factor c, i.e. $a_{gc} \to ca_{gc}$ for all genes, does not correspond to a change in gene expression state, but just to a change in cell size. Second, it is well known that, in current scRNA-seq protocols, the rate of capture and sequencing of mRNAs varies significantly across cells [30, 31] so that there is only a weak quantitative relationship between the total number of sequenced mRNA molecules and the true total mRNA content of cells. Although it is possible to estimate capture and sequencing efficiencies, at least to some extent, using RNA spike-in controls [30, 32], most experiments are performed without such controls. Therefore, for most scRNA-seq datasets it is unclear to what extent variations in total sequenced mRNAs across cells represent biological variability, as opposed to technical variability. Consequently, transcription quotients can generally be much more accurately estimated than absolute transcription activities, because they do not directly depend on capture efficiency. Note that quantifying gene expression by quotients, i.e. transcripts per million transcripts, is also the standard approach in bulk RNA-seq experiments.

Finally, we note that if we were to use differences in transcription quotients of mRNAs $\alpha_{gc} - \alpha_{gc'}$ to quantify the change in expression of gene g between cells c and c', then this change would be proportional to overall expression level of the gene. That is, a change from 20 to 40 transcripts per million would be considered ten times as large as a change from 2 to 4 transcripts per million. Since the early days of transcriptomics it has been observed [33] that, as would be expected from the multiplicative effects of fluctuations in rates of various biochemical reactions [34], the relative expression levels of genes in a sample follows a roughly log-normal distribution that covers several orders of magnitude. Consequently, if we were to quantify expression changes directly by the changes $\alpha_{gc} - \alpha_{gc'}$, the expression changes between two cells would be completely dominated by those of the highest expressed genes. Therefore, it has long become standard to instead use *logarithms* of the expression levels. Thus, we propose to quantify the gene expression state of a cell by the *logarithms of the transcription quotients* (LTQs) $\log(\alpha_{gc})$ so that a x-fold change in quotient $\alpha_{gc} \to \alpha_{gc'} = x\alpha_{gc}$ corresponds to the same additive change $\log(\alpha_{gc}) \to \log(\alpha_{gc}) + \log(x)$ in LTQ, independent of the absolute value of the quotient α_{gc}. In summary, we propose to characterize the gene expression state of a cell c by a vector of LTQs $\log(\alpha_{gc})$.

A probabilistic model for a scRNA-seq experiment

The initial steps of scRNA-seq analysis involve basic processing of the raw sequencing reads such as quality control of the reads, identification of the various barcodes that identify the library, the individual cell, the unique mRNA molecule (if available), and mapping each read to the corresponding genome or transcriptome. The methods used in these steps are similar to methods used for bulk RNA-seq and ChIP-seq and have matured to the point that there are accepted methods and little variability in the results from commonly used tools, e.g. [35–38].

The introduction of unique molecule identifiers (UMIs) [39] was an important development in scRNA-seq technology in that it avoids noise in expression measurements due to fluctuations in PCR amplification, and determineds the number of unique mRNA molecules that were captured for each mRNA. Since only protocols that incorporate UMIs allow for a realistic modeling of the statistics of the measurement noise, we will here focus on scRNA-seq protocols that use UMIs.

After the basic processing of the raw data has been performed, the data will consist of a matrix of integers n_{gc} giving the number of captured mRNA molecules for each gene g in each cell c. The key assumption of our probabilistic model is that, in a scRNA-seq experiment, each mRNA molecule in a given cell c has a probability p_c to be captured and sequenced. This capture probability p_c, which varies from cell to cell, has been estimated to be in the range of 10 to 15% [40] and up to 30% with the most recent protocols [41]. Under this assumption, the probability of the observed UMI counts n_{gc} in cell c given the transcription quotients α_{gc} is still given by a product of Poisson distributions (see Supplementary Methods)

$$P(\{n_c\}|\{\alpha_c\}) = \prod_g \left[\frac{(N_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c \alpha_{gc}} \right], \tag{4}$$

where $\{n_c\}$ is the set of UMI counts in cell c, $\{\alpha_c\}$ the set of transcription quotients in cell c, and N_c the total number of UMIs in cell c. Crucially, we see that the convolution of the biological Poisson noise and the sampling noise introduced by the scRNA-seq measurement together still lead to a simple Poisson distribution in terms of the transcription quotients α_{gc}.

Prior probabilities and the Bayesian solution

In order to estimate the $\log(\alpha_{gc})$ from the observed UMI counts n_{gc} a final ingredient that we need is to define a prior distribution over these LTQs. As we aim to minimize the number of assumptions that our inference makes, our model will not assume any dependence structure between the LTQs of different genes, i.e. we will not assume that the gene expression data derives from a low-dimensional manifold. We will also not assume that the LTQs follow a particular distribution. The only thing we will assume is that, for each gene, the prior distribution of LTQs $\log(\alpha_{gc})$ can be characterized by its mean μ_g and variance v_g. We rewrite the transcription quotients α_{gc} in terms of an average quotient α_g and a cell-specific log fold-change δ_{gc}, i.e. $\alpha_{gc} = \alpha_g e^{\delta_{gc}}$. With that reparametrization, the mean μ_g equals $\log(\alpha_g)$ and the δ_{gc} derive from a prior probability distribution with mean zero and variance v_g. Given that we only specify the variance of the distribution of the δ_{gc} to be v_g, we choose the maximum entropy distribution [42] consistent with this constraint, which is a Gaussian distribution. Importantly, this does not mean that we assume that the log fold-changes δ_{gc} follow a Gaussian distribution. It just assumes that, before seeing any of the data, we assume the δ_{gc} are taken from the broadest, least assuming distribution consistent with some (unknown) variance v_g.

In the Supplementary Methods we derive in detail how this model can be solved to estimate, for each gene g:

1. The mean LTQ μ_g and its error-bar $\delta\mu_g$.

2. The estimated variance v_g of the changes in LTQs δ_{gc} across cells.

3. For each cell c, the estimated LTQ δ^*_{gc} and an error-bar ϵ_{gc} on this LTQ.

Note that the LTQs δ^*_{gc} provide estimates for how much the transcription and decay rates of each gene g in cell c differ from their average rates, and thus correct for both the intrinsic biological Poisson fluctuations as well as the finite sampling fluctuations inherent in the scRNA-seq measurement.

Alternative methods for scRNA-seq normalization

To assess the performance of Sanity we will compare it with a number of other methods for normalization/imputation from scRNA-seq data. Apart from a number of other tools from the recent literature, we also include two basic normalization procedures that are widely used. First, the simplest approach to estimating gene expression levels e_{gc} from scRNA-seq data is to simply log-transform the observed number of UMIs n_{gc} after adding a *pseudocount* p in order to avoid problems with zero counts $n_{gc} = 0$, i.e.

$$e_{gc} = \log(n_{gc} + p). \tag{5}$$

A typical choice for the pseudo-count is $p = 1$, because it attenuates fluctuations in n_{gc} on the order of magnitude corresponding to the resolution of the experimental measurements. We will refer to this normalization, with $p = 1$, as the *RawCounts* normalization, since it essentially just log-transforms the raw counts.

However, the total number N_c of mRNAs captured and sequenced from an individual cell c can vary substantially due to fluctuations in capture efficiency and sequencing depth, as well as changes in cell size. Consequently, the RawCounts procedure introduces artificial correlations between the expression levels e_{gc} and the total number of UMIs N_c that were sequenced in the cell c. Thus, the most commonly used normalization approach is to first divide the rawcounts n_{gc} by the total counts N_c and then multiply by a typical total count N before adding a pseudocount and log transforming, i.e.:

$$e_{gc} = \log\left[\frac{n_{gc}}{N_c}N + 1\right], \tag{6}$$

where we will take for the typical total count N the median of the counts N_c across all cells. In a slight abuse of terminology, we will call this normalization the *TPM* normalization because of its close connection to the transcripts per million normalization used in bulk RNA-seq (which corresponds to setting $N = 10^6$).

Beyond these two simple normalization methods, we compare Sanity's performance with that of the following recently published tools:

1. DCA [43], which uses a deep learning based autoenconder.

2. MAGIC [44], which uses diffusion of measured gene expression states between cells with similar expression profiles.

3. SAVER [45], which assumes negative binomial counts distributions n_{gc} and models the underlying rates using Poisson LASSO regression with the expression levels of other genes.

4. scImpute [46], which focuses mainly on correcting 'dropouts', i.e. datapoints for which $n_{gc} = 0$.

5. scVI [47], which uses a deep neural network based autoencoder.

Note that, with the exception of scImpute, all these methods seek to normalize the expression levels for the total UMI count per cell, and seek to remove noise by using lower dimensional representations of the input counts n_{gc}.

We used default parameters for all these methods and, since all methods report expression values in linear space, we log-transformed all expression values. MAGIC sometimes reports 0 or negative values and, as suggested by the authors, we first set all negative values to 0 and then add a pseudocount of 1 to all expression values (including the nonzero ones) before log-transforming. Similarly, scImpute reports some zero values and we added a pseudocount of 1 to all these.

Test datasets

To comprehensively assess the performance of the different methods we used a collection of datasets for which annotation of the sequenced cell types was available. The datasets we used were (labelled by the first author of the publication):

1. *Grün* : 160 mouse embryonic stem cells and 160 corresponding aliquots consisting of, 80 cells from culture in 2i medium, 80 cells from culture in serum, and 80 aliquots for each condition that were created by pooling cells together, and then splitting the pool into single-cell mRNA equivalents [30].

2. *Zeisel*: 3'005 cells from the somatosensory cortex and from the CA1 region of the mouse hippocampus, annotated into 7 cell types [48].

3. *Baron*: 1'937 human pancreatic cells annotated into 14 cell types [49].

4. *Chen*: 14,437 adult mouse hypothalamus cells annotated into 15 clusters [50].

5. Three datasets from *LaManno* [51]:

   (a) *LaManno/Embryo*: 1'977 ventral mid-brain cells from human embryo annotated into 25 classes.

   (b) *LaManno/ES*: 1'715 human embryonic stem cells annotated into 17 classes.

   (c) *LaManno/MouseEmbryo*: 1'907 ventral mid-brain cells from mouse embryo annotated into 26 classes.

In addition to these real datasets we also constructed one simulated dataset as described in the Supplementary Methods. The parameters of the simulation were chosen so as to mimic the statistics of the *Baron* dataset (see Fig. S11).

Results

Sanity accurately corrects for Poisson fluctuations to identify true variance in gene expression

A key aim of Sanity's normalization is to correct for both biological and technical sampling fluctuations in order to quantify the true biological variation in expression of each gene across cells. Testing this is challenging because the true expression variability of each gene is generally unknown. To address this we used a simulated dataset for which the true expression variability of each gene is known, on the one hand, and analyzed a carefully designed study of mouse embryonic stem cells (ESCs) from *Grün et al* [30] on the other hand. In this study, mouse ESCs were culured in both 2i and serum conditions and apart from scRNA-seq measurements on these cells, the same measurement protocol was applied to single-cell equivalent *aliquots* from pooled RNA of multiple cells. Since all aliquots were sampled from the same pool, there is no biological expression variation in this dataset at all, and the expression variation in these aliquots derives solely from technical sampling noise. In addition, the ESCs are highly homogeneous so that also little true expression variation is expected for ESCs in the same condition, and the main expression differences are expected between cells in the 2 different culture conditions.

The amount of expression variability of a gene across a set of cells can be quantified by its coefficient of variation CV, i.e. the ratio of the standard-deviation and the mean of its expression levels. Figure 1A shows box-whisker plots of the distribution of CVs, for each of the 4 datasets, as calculated from the (non log-transformed) expression estimates of each of the normalization methods. Ideally the methods should infer that there is no true variability at all for the aliquots, and relatively little variability for the ESCs. In addition, it is known that variability in 2i conditions is smaller than in serum [30], so that we expect larger CVs in serum. Although, with the exception of scVI, all methods infer that the CVs are larger in serum than in 2i, and smaller for the aliquots than for the cells, the CVs that Sanity infers are at least twofold lower than those of all other methods, and only Sanity infers that the CV is less than 10% for the large majority of the aliquots. Of the other methods, MAGIC and SAVER show distributions of CVs that, while generally larger, are closest to those inferred by Sanity. All other methods show distributions of CVs that are at odds with our prior information in one way or another. For example, due to the Poisson noise, the simple RawCounts, TPM, and scImpute methods infer CVs of at least 0.5 for the large majority of genes in both cells and aliquots. DCA infers very similar distributions of CVs for the ESCs and aliquots and, finally, scVI shows unrealistically high CVs for all genes in both ESCs and aliquots.

The ability for methods to correct for sampling noise can be assessed most clearly by plotting the CV of each gene as a function of its mean expression (Fig. 1B). As is well appreciated in the scRNA-seq literature, e.g. [32], because the variance of a Poisson distribution is equal to its mean, Poisson sampling fluctuations add a term $1/\sqrt{\text{mean}}$ to the CV. Because most genes have low absolute expression values, the CV is dominated by this term for most genes, leading to a strong negative correlation between mean expression and CV. Indeed, the simple RawCounts and TPM methods show an almost perfect negative correlation between CV and mean, showing that sampling noise dominates the observed variability for all but the highest expressed genes. Ideally, the normalization would correct for the Poisson contribution to the CV, and in principle we would

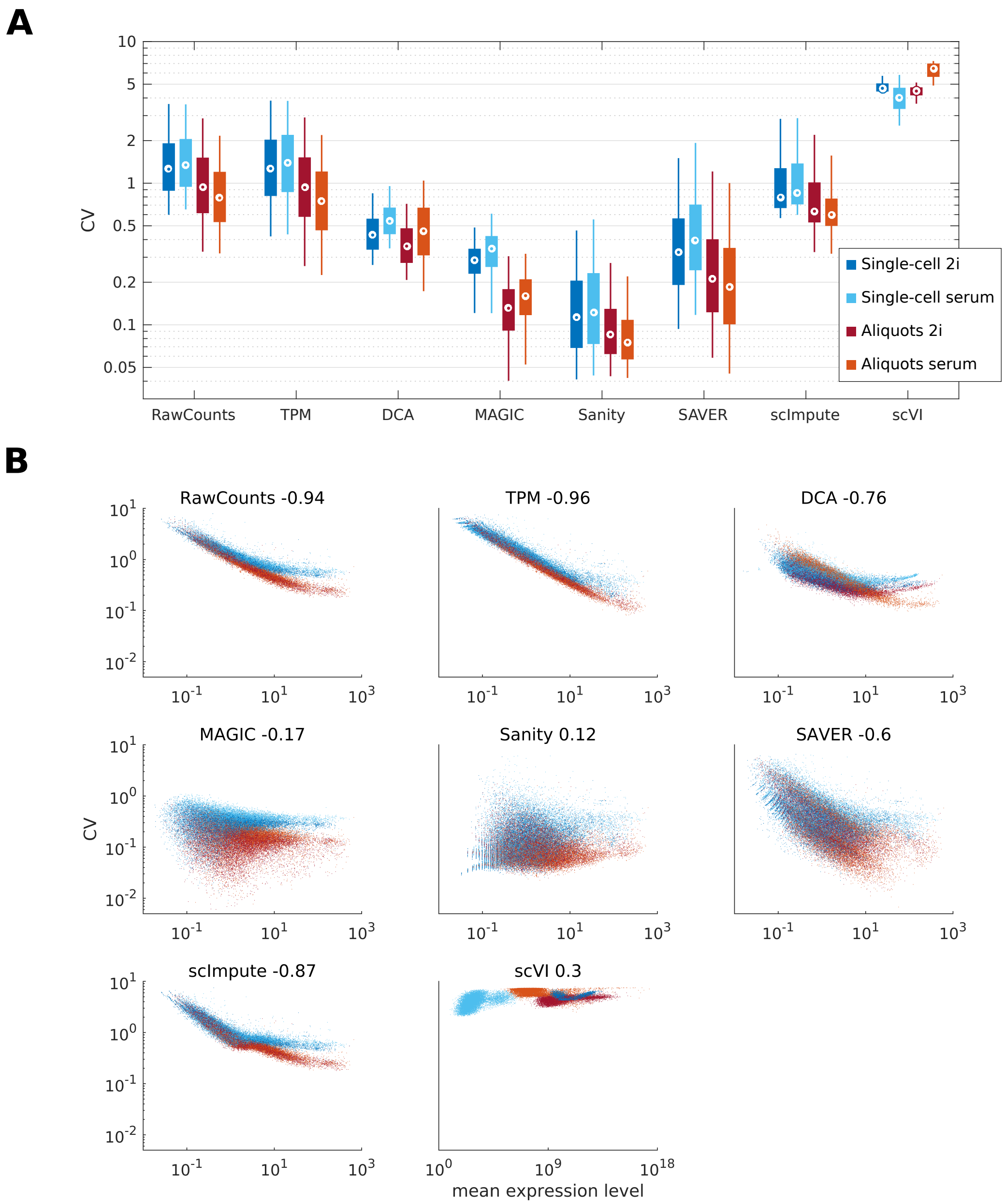

Figure 1: **A**: Box-whisker plots showing the medium (circle) as well as the 5th, 25th, 75th, and 95th quantiles of the distribution of gene expression levels for each of the 4 datasets (see legend) as inferred by each of the normalization methods. **B** Scatter plots of CV (standard-deviation divided by mean) for all genes in each of the 4 datasets (colors as in panel A) as inferred by each of the normalization methods. The Pearson correlation coefficient between $\log CV$ and log mean is shown on top of each plot. The axis are shown on a logarithmic scale and are kept similar across panels, except for *scVI* where the mean expression values are on a very different scale from those of the other methods.

not expect to see a systematic correlation between mean expression and the normalized CV. Indeed, for Sanity the normalized data does not exhibit any correlation between CV and mean. However, with the exception of MAGIC and scVI, a strong negative correlation between CV and mean remains for all other methods, showing that even after normalization the expression variability is dominated by Poisson noise for many genes. We also note that scVI exhibits rather unnatural distributions of CV and mean, with consistently high CV and strongly varying means across datasets. These observations do not only apply to the dataset of [30], but are observed for all test datasets we considered (Suppl. Fig. S1).

We next constructed a simulated dataset (see Supplementary Methods) in which the total number of UMI per cell and mean expression levels where chosen to match those of the dataset of Baron et al. [49]. Each gene was assigned a random true variance in log gene expression, its true expression values were sampled from a Gaussian distribution with corresponding variance, and finally Poisson noise was added to these true expression values. Since, for this simulated dataset, we know exactly the true variability in gene expression for each gene, we directly compared the inferred CV with the true CV used in the simulation (Fig. 2). For the simple TPM and RawCounts methods, there is actually a good correlation between true CV and inferred CV for very highly expressed genes, confirming that for the highest expressed genes the observed CV in the data matches the true CV. However, for the large majority of genes, the Poisson noise causes the inferred CV to be much higher than the true CV, so that there is ultimately almost no correlation between true and observed CVs across the entire set of genes. For most of the other methods, there is very little relation between the true and inferred CVs. MAGIC and DCA predict much lower CVs than the true CVs, scVI predicts consistently high CVs, and there is no correlation between the true and predicted CVs for any of these methods. Only Sanity and SAVER show a good match between the true and inferred CVs across most of the genes. Notably, Sanity accurately estimates CVs for all highly expressed or highly variable genes. For low expressed genes, where there is not sufficient data to reliably detect the true expression variability of a gene, Sanity conservatively infers that the true expression variability is low and these genes will therefore not significantly contribute to any downstream analysis of expression variability across cells. Although SAVER's inferred CVs are reasonable for most genes, they are clearly less accurate than Sanity's predictions, and for a subset of low expression genes SAVER strongly overestimates the CV.

In summary, these results show that Sanity is the only normalization method that can reliably correct for the Poisson sampling noise to quantify the true expression variability of each gene.

Many normalization methods introduce spurious correlations with library size

Due to variations in cell size, mRNA capture efficiency, and sequencing depth, the total number of captured UMIs can fluctuate significantly from cell to cell. Therefore, most scRNA-seq processing methods involve normalize the expression levels of genes in a given cell for the total number of mRNAs (i.e. UMIs) that were sequenced for that cell. For example, whereas the simple RawCounts procedure does *not* correct for total UMI counts per cell, the simple TPM procedure normalizes for total UMI count by dividing the observed counts for each gene by this total count. With the exception of scImpute, all other methods also include methods to normalize for total UMI count.

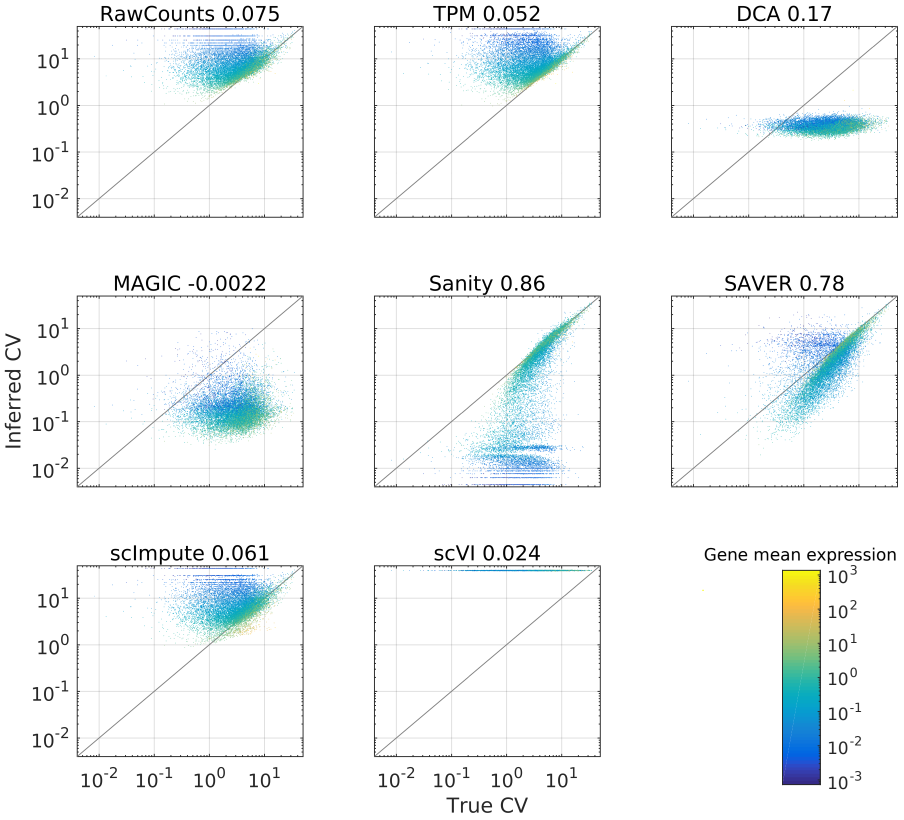

Figure 2: Comparison of the true CVs and those inferred by each of the normalization methods on the simulated dataset. Each panel shows a scatter plot of the true CV (horizontal axis) the CV as inferred by the normalization method (vertical axis) across genes. The color of each datapoint shows the mean expression level of the gene (total UMI in the dataset, see colorbar). The Pearson correlation between the inferred CVs and the true CVs is shown on top of each panel.

To investigate the effects of the normalization for total UMI count we calculated, for each method and each gene, the Pearson correlation between the inferred log expression levels and the logarithm of the total UMI count across cells. Using the Zeisel dataset as an example, Fig. 3 shows the distribution of Pearson correlations for each of the methods as well as the raw scatters of the normalized expression levels as a function of log total UMI count for one example gene (*Zbed3*). Starting with the simple RawCounts method we see, as expected from the fact that this method does not normalize for total UMI count, that for most genes there is a positive correlation between total UMI count of a cell and the expression level of the gene in that cell. The scImpute method shows similar correlations with total UMI count which is consistent with the fact that this method does not normalize for total UMI count either. In contrast both the simple TPM method, and especially Sanity, remove this correlation, confirming that these methods successfully normalize for the fluctuations in total UMI count across cells.

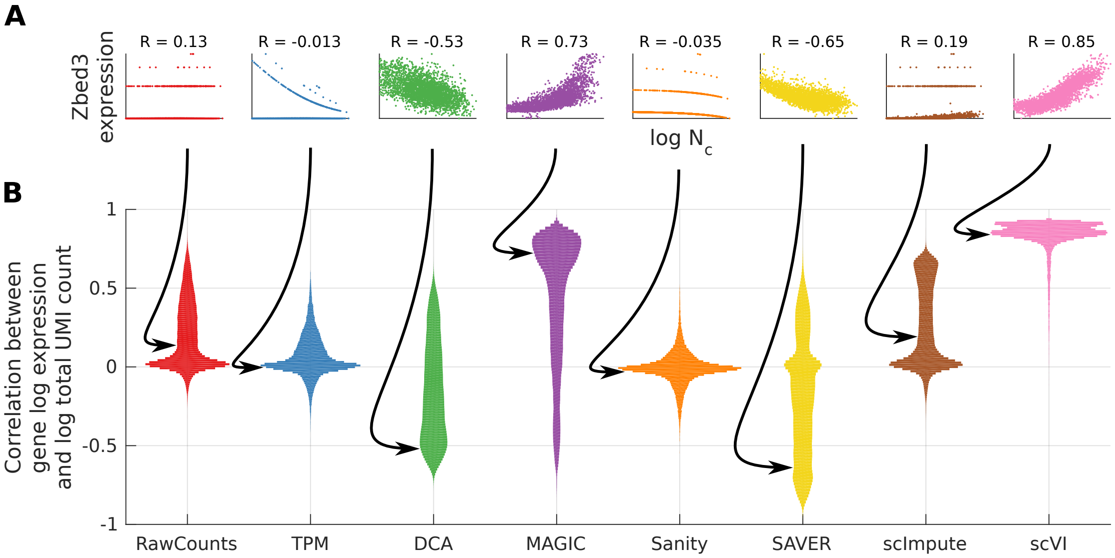

Figure 3: **A**: Scatter plots of the normalized log expression level of the example gene (*Zbed3*) versus the logarithm of the total UMI count $\log(N_c)$ across cells for each of the methods. The Pearson correlation of the dependence is shown above each panel. **B**: Violin plots of the distribution of correlation coefficients between the inferred log expression levels of genes and the log of total UMI count per cell, for the *Zeisel* dataset. Different colors correspond to the different methods, which are indicated below.

We were very surprised to see that for all other methods, rather than removing correlations with total UMI count, the normalized expression levels show even stronger correlations with total UMI count. DCA, SAVER, and MAGIC show a very wide distribution of correlation coefficients with predominantly negative correlations for DCA and SAVER, and predominantly positive correlations for MAGIC. The situation is even more dramatic for scVI which infers that the expression levels of essentially *all* genes are highly correlated with total UMI count. The scatters with the predicted gene expression levels for the gene *Zbed3* as a function of log total UMI count $\log(N_c)$ illustrate how dramatically the various normalization methods transform the input data. The RawCounts show that this gene is fairly low expressed, with either 0 or 1 UMI observed in most cells, and that there is a slightly higher chance to observe one or two UMIs when the total UMI count N_c is larger. However, DCA, MAGIC, SAVER, and scVI completely transform this input data into a scatter of continuously varying expression levels that either correlate strongly negatively (DCA, SAVER) or strongly positively (MAGIC, scVI) with total UMI count. These observations again generalize to all other datasets as shown in Suppl. Fig. S2. In conclusion, only TPM and Sanity reliably normalize for variations in total UMI count, and most other methods introduce strong spurious correlations with total UMI count.

Most normalization methods spuriously infer co-expression between many pairs of genes

One of the most common downstream analyses that are applied to transcriptome data is the identification of co-expressed genes, for example for identifying co-regulated pathways or regulatory modules. In order to perform such co-expression analysis, it is crucial that the pairwise correlations of the normalized expression profiles across genes accurately reflect the co-expression evidence in the data. In order to compare co-expression information across methods we calculated, for each method and for every pair of genes, the Pearson correlation of their normalized expression levels. We then compared these pairwise correlation coefficients across the various methods.

Using the Baron dataset as an example, Fig. 4A shows a scatter of the pairwise correlations as inferred by Sanity and the simple TPM method for all pairs of genes. We see that the inferred pairwise correlations by-and-large agree between the two methods, i.e. most points fall along the diagonal, and there are almost no pairs where the two methods strongly disagree on the strength of the correlation. As shown in Suppl. Fig. S3, this also holds for the comparison of Sanity's pairwise correlations with those of RawCounts and scImpute. However, a very different pattern is observed for the comparison of Sanity with MAGIC (Fig. 4C). For many of the pairs of genes for which Sanity infers no co-expression, i.e. zero correlation, MAGIC infers a broad range of correlations running from almost perfect anti-correlation, to perfect correlation. To assess whether the raw data are more consistent with Sanity's or MAGIC's pairwise correlations, we first focused on a subset of 4360 pairs of genes within the red rectangle of Fig. 4C, for which MAGIC predicts nearly perfect correlation ($r > 0.975$) whereas Sanity predicted none ($-0.03 < r < 0.005$). Summing across all 4360 pairs and all cells, we counted the total number of times $n_{i,j}$ for which i UMIs were observed for the first gene and j for the second. Strikingly, there was not a single example for which both i and j are larger than zero (Fig. 4D). That is, although MAGIC infers that these 4360 pairs of genes are almost perfectly co-expressed, *none* of them are ever observed to be present at the same time in *any* cell. In contrast, for the small set of pairs for which Sanity infers significant co-expression whereas MAGIC does not (magenta box in Fig. 4C), we do generally find evidence of co-expression (Fig. 4E). As shown in Suppl. Fig. S3, the same pattern is observed for the comparisons of Sanity's pairwise correlations with those of DCA and SAVER. That is, MAGIC, DCA, and SAVER all infer large numbers of highly correlated or anti-correlated pairs of genes, whereas there is no evidence at all in the raw data that these pairs of genes are co-expressed. The pairwise correlations predicted by scVI show even more pathological behavior, i.e. scVI predicts that *all* pairs of genes are significantly co-expressed (Fig. 4B).

These observations are confirmed by the overall distributions of pairwise correlations that

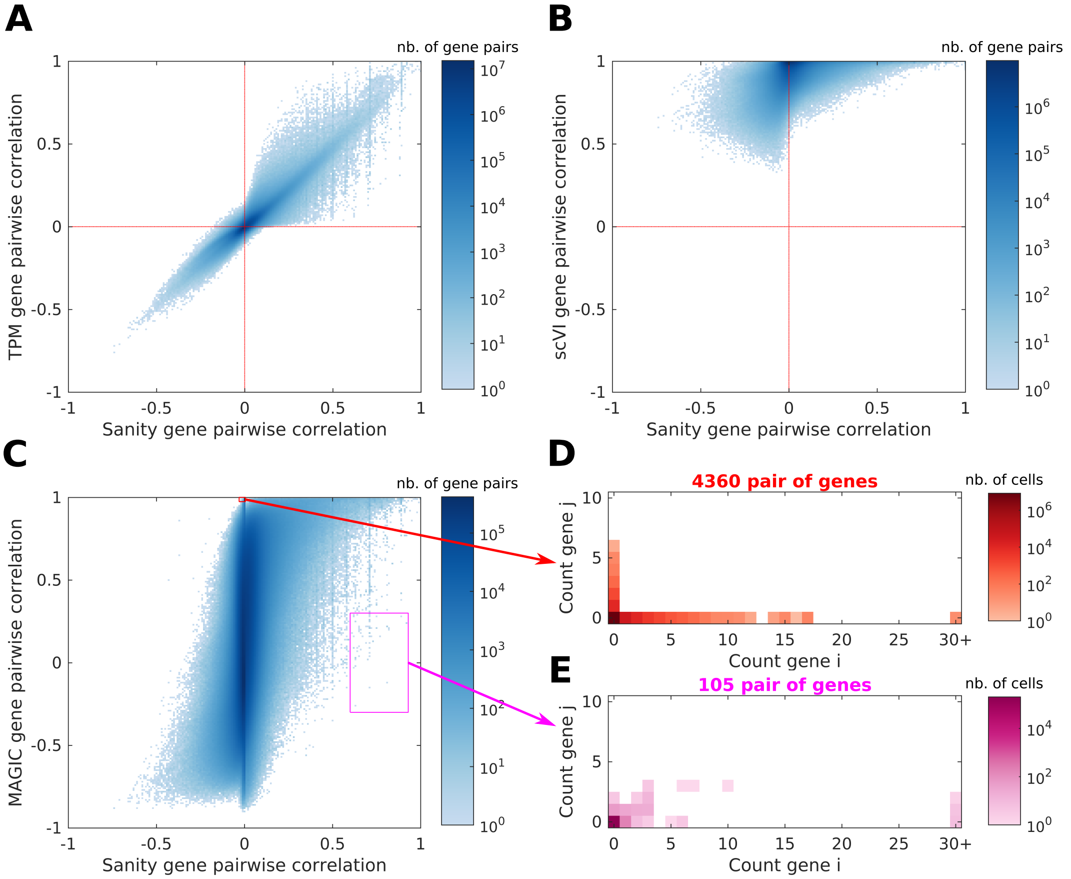

Figure 4: **A**: Density plot of Pearson correlations for all pairs of genes as inferred by *Sanity* (x-axis) against the correlations inferred by *TPM* (y-axis). The color scale shows the density in log_{10} number of gene pairs and values $\log_{10}(0)$ are shown in white. **B**: Density plot as in panel A, but now comparing the correlations inferred by *Sanity* and *scVI*. **C**: Density plot as in panel A but now comparing the correlations inferred by *Sanity* and *MAGIC*. The red and the magenta rectangles indicate the pairs of genes analyzed in panels D and E. The red rectangle contains all pairs of genes with correlation above 0.975 for *MAGIC* and between -0.03 and 0.005 for *Sanity*. The magenta rectangle contains all pairs of genes with correlation between -0.3 and 0.3 for *MAGIC* and between 0.6 and 0.93 for *Sanity*. **D**: 2-dimensional histogram of counts per cell summed over the 4360 pairs of genes from the red rectangle in panel C. The height of the histogram is shown in $\log_{10}$ as a color and values $\log_{10}(0)$ are shown in white. **E**: 2-dimensional histogram of counts per cell summed over the 105 pairs of genes from the magenta rectangle in panel C.

each of the methods predicts on each of the datasets (Suppl. Fig. S4). The results are highly consistent across datasets and show three main behaviors. First, Sanity, TPM, RawCounts, and scImpute have distributions of pairwise correlations that are highly peaked around zero, i.e. these methods predict that most pairs are not co-expressed. Second, instead of a peak at zero, DCA and MAGIC have almost uniform distributions of pairwise correlations. SAVER's distribution of pairwise correlations is somewhat in between these two behaviors, i.e. a very broad distribution with a moderate peak at zero. Third, scVI's pairwise correlations are highly peaked near almost perfect correlation of $r \sim 1$. Notably, even on the simulated dataset that contains no expression correlations at all, MAGIC and DCA also show broad distributions of pairwise correlations, and scVI again predicts almost all pairs to be perfectly co-expressed. This further supports that the correlations that these methods predict are artefactual.

Sanity outperforms other methods on clustering cells into subtypes

One of the main applications of scRNA-seq is to identify (novel) cell types and this is generally done by clustering single cell gene expression patterns using a measure of pairwise distances between cells. Since the pairwise distances between cells will depend on the normalization method, we expect different methods to differ in their ability to recover subpopulations of cells. For six of our test datasets, the corresponding study explicitly reported an annotation of cell types present in the dataset, that was typically obtained using a combination of automated clustering, analysis of marker gene expression, and hand curation. To test the performance of the different normalization methods on cell type identification we investigated to what extent the reported cell type annotations could be recovered by application of simple clustering algorithms to the normalized gene expression data.

Taking the Zeisel dataset as an example, we first obtained a simplified visual indication of the clustering structure implied by the different methods by applying the popular t-SNE algorithm [26] (with the default 50 principal components and a perplexity equal to the average number of cells per annotated cluster) to the normalized expression values of each method, and colored the cells according to the annotation of [48] (Fig. 5A). Although it is well-known that it is difficult to interpret these visualizations beyond the fact that neighboring cells in the visualization are typically also neighboring in the full gene expression space, the visualization does suggest that there is considerable disparity across the normalization methods. For example, it appears that TPM, DCA, Sanity, and SAVER separate the cell types more reliably than MAGIC, and scVI. Similar qualitative observations can be made on the other datasets (Suppl. Fig. S5 - S9).

To quantify the performance of the different normalization methods we applied, for each dataset and each normalization method, simple hierarchical clustering using Ward's method [52]. That is, starting with each cell as its own cluster, at each step two clusters are fused so as to minimize the sum of the variances across all clusters. This is iterated until the number of clusters equalled the number of annotated cell types. We then calculated, for each method and dataset, the similarity between the annotated clusters and the inferred clusters using the normalized mutual information as a similarity measure (see Supplementary Methods). As shown in Fig. 5B, Sanity outperforms all other methods on all datasets except the Zeisel dataset, where the TPM method obtains slightly higher similarity with the annotated clusters. The TPM normalization

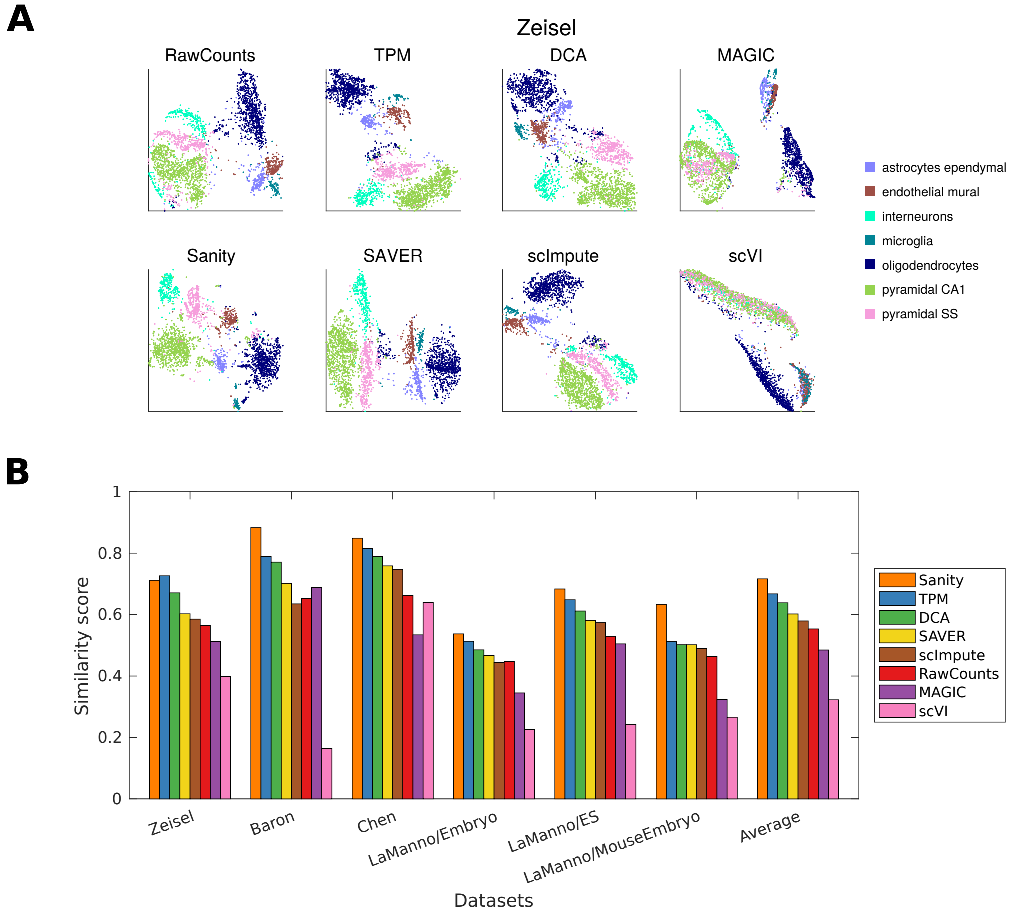

Figure 5: **A**: Each panel shows a t-SNE visualization of the Zeisel dataset using the normalized gene expression values of the method indicated at the top of the panel. Each point represents a cell and is colored according to the cell type annotation given in [48]. **B**: Similarity between the annotated clusters and the clusters inferred by applying hierarchical Ward clustering on the normalized expression values of the different methods. Normalized mutual information, which ranges from 0 (no similarity) to 1 (perfect match) was used as a similarity measure. Each group of bars shows the results for a particular dataset as indicated below it, and the bars are colored according to the normalization method, as indicated in the legend. The last group of bars shows the average similarity per method across all datasets. For ease of viewing, the methods are sorted from left to right according to their average similarity.

generally has second best performance, followed by DCA and then SAVER. We also note that MAGIC and scVI generally perform least well. Very similar results are observed when using a different similarity metric such as the rand index (Supplementary Methods) or a different clustering algorithm such as k-means (Suppl. Fig. S10).

Thus, although the performance differences are not large, Sanity's normalized expression estimates generally outperform the other methods in identifying subtypes of cells.

Sanity outperforms other methods on identification of differentially expressed genes

As a final example of downstream analysis we consider the ability of the normalized expression values to identify genes that are upregulated in particular subtypes of cells. That is, we aim to identify genes whose average expression in a given subtype of cells is significanlty higher than its average in all other cells. A simple and standard statistic for comparing the averages of populations is the t-statistic and we used this to identify upregulated genes for each cell type in a given dataset. In particular, for each gene g and each cell type k annotated in a given dataset, we calculated a t-statistic

$$t_{gk} = \frac{\mu_{gk} - \mu_{g\bar{k}}}{\sqrt{\sigma^2_{gk}/n_k + \sigma^2_{g\bar{k}}/n_{\bar{k}}}}, \tag{7}$$

where μ_{gk} is the average of the normalized expression values of gene g in cells of type k, $\mu_{g\bar{k}}$ is the average in all other cells, σ^2_{gk} and $\sigma^2_{g\bar{k}}$ the corresponding variances in normalized expression levels, and n_k and $n_{\bar{k}}$ the number of cells in type k and the number of all other cells. The t-statistic t_{gk} quantifies that statistical evidence that gene g's average expression in cell type k is higher than in the other cells. To predict a set of upregulated genes, one would then pick a cut-off in t-statistic corresponding to a particular rate of false discovery (FDR), e.g. a 5% FDR. By applying this procedure to the normalized expression values of each method we derived, for each method, a set of upregulated genes for each cell type k of a given dataset of interest.

To test the performance of these predicted sets of upregulated genes we compared these lists with similar lists of predicted upregulated genes from the original publications. For 3 of our test datasets, i.e. the Zeisel and two LaManno datasets, the authors published, for each identified cell type, a list of genes that had higher average expression in the cell type compared to the other types of cells [48, 51]. These lists were obtained using a fairly complex regression procedure and it is of course debatable whether these published lists can be treated as a gold standard. However, since they were obtained using a method that is very different from our simple t-statistic, we reasoned that the match to these reference lists can still be used to assess the relative performance of the different normalization methods.

For each normalization method we calculated a precision-recall curve by producing one sorted list of the t-statistics t_{gc} for all genes in all subtypes and, as a function of a cut-off on t, compared the predicted set of significantly upregulated genes, with the reference lists published in the original study. Figure 6 shows the precision-recall curves obtained for each of the methods on each of the 3 datasets for which reference lists were available. The colored dots indicate the sensitivity and positive predictive values (PPV) that are obtained for each method when using a t-statistic cut-off corresponding to a 5% FDR. We see that, for each dataset, Sanity

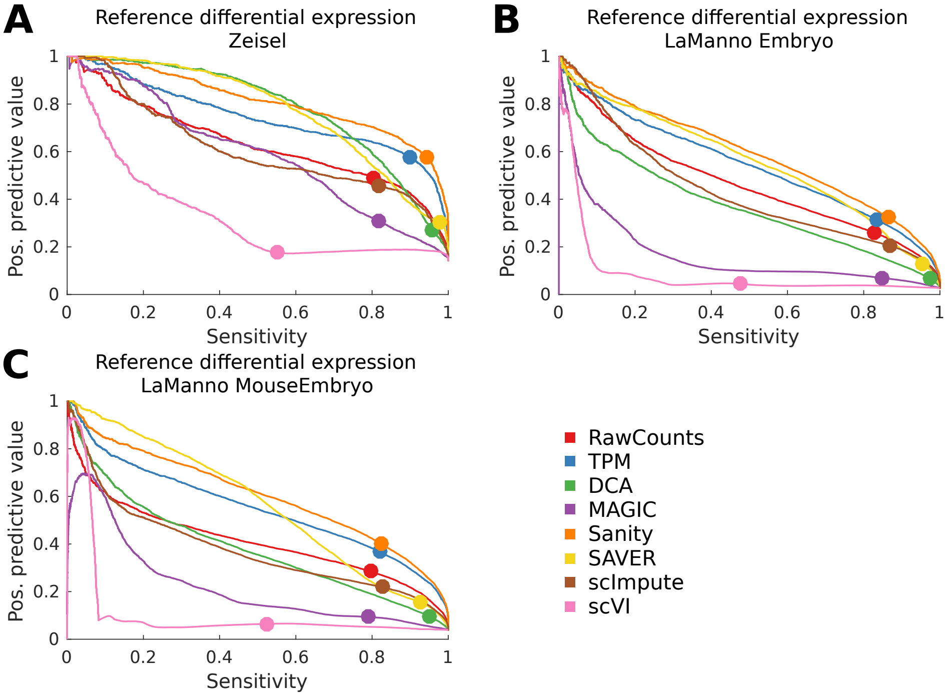

Figure 6: Precision recall curves showing the positive predictive value, i.e. the fraction of predicted upregulated genes that correspond to upregulated genes in the corresponding reference list, as a function of sensitivity, i.e. fraction of all genes in the reference lists that were predicted, as obtained using the t-statistics for each of the normalization methods (colors) for the Zeisel (panel **A**) and two LaManno datasets (panels **B** and **C**). The dots show the values that are obtained when using a cut-off on the t-statistic corresponding to a false discovery rate of 5%.

achieves the highest accuracy of predictions, i.e. a higher PPV at a given sensitivity than all other methods. The simple TPM method achieves the next best performance, MAGIC and scVI perform hardly better than random, and all other methods are somewhere in between. Note that, at a 5% FDR, the more complex DCA, MAGIC, and scVI methods all predict very large numbers of upregulated genes which leads to low PPVs. In summary, these results suggest that Sanity's normalized expression levels also achieve highest accuracy for downstream identification of differentially regulated genes.

The Sanity software

Sanity was implemented in C and is freely available for download at `https://github.com/jmbreda/Sanity`. The raw UMI count tables for each of the scRNA-seq datasets, as well as all the normalized expression values as inferred by each of the methods are available from this website as well.

Discussion

Recent technologies for quantitatively measuring the epigenetic states of single cells are promising to revolutionize our understanding of the mechanisms by which cell fate and identity are regulated in animals, and there has been a surge in the use of these methods. However, even for the most popular scRNA-seq method, there is so far little agreement within the community as to how such single-cell expression data should be processed and normalized. In particular, it has so far been challenging to define a normalization procedure that, on the one hand, deals with the specific artefacts and noise introduced by the scRNA-seq measurement process while, on the other hand, providing quantification of the expression states of cells that have direct biological interpretation. In particular, only when the normalized expression methods provide quantities with a concrete biological interpretation will it be possible to integrate results of scRNA-seq experiments from different protocols employed by different labs with expression data from other experimental techniques. So far, such normalization methods have been lacking.

Here we developed a Bayesian normalization procedure that achieves these objectives, and is derived from first principles using only two basic assumptions. First, we characterize a cell's gene expression state by the vector of log transcription quotients (LTQs) across genes, i.e. the logarithms of the expected fractions of the transcript pool for each gene. Second, estimating these LTQs within a Bayesian setting requires choosing a prior distribution and we chose to characterize the distribution of LTQs of each gene by just its mean and variance across cells. Given only these two assumptions the entire procedure follows from first principles, deterministically, and without any tunable parameters. Given a table of UMI counts for each gene (or transcript) across cells, our Sanity method returns estimates of LTQs and their error bars across all genes and cells. Importantly, these estimates correct both for the Poisson noise that is inherent in the process of transcription, as well as the sampling noise associated with the scRNA-seq measurements, so that the variance in normalized expression levels across cells reflects changes in rates of transcription and mRNA decay rather than biological or technical sampling noise.

Although our normalization method makes only a minimal number of assumptions, one may ask how arbitrary these assumptions are. If one accepts that biological and technical sampling noise do not reflect changes in gene expression state, that expression changes should be measured in terms of fold-changes rather than absolute changes, and that an overall 'cell size' rescaling of the expression levels of all genes by the same amount does not reflect a change in expression state, then LTQs are the most general representation of a cell's expression state. Similarly, our prior distribution over LTQs for a gene is the least assuming, i.e. maximum entropy, distribution that is consistent with only a given mean and variance. This prior thus also aims to minimize the strength of the assumptions that the method makes. In this sense, we think that our method makes the most conservative assumptions that are consistent with current knowledge. To improve on these assumptions we would have to supply specific biological information to determine more informative priors on the gene expression states that cells can take on.

Our comparison of Sanity with other normalization methods showed that the Sanity's normalized expression values outperform other methods on basic downstream processing tasks such as clustering cells into subtypes and identifying differentially expressed genes. More importantly, however, we showed that all other methods produce a representation of the data that is severely distorted in one or more respects. Of all alternative methods that we evaluated, the simple TPM method produces the most reasonable representation of the data and it also performs second best on the downstream processing tasks. The main problem with the TPM method is that the variations in normalized expression levels are dominated by Poisson fluctuations. This not only causes there to be a complete lack of correlation between true biological variability of genes and the variability of the normalized expression values, it also causes low expressed genes to be predicted to be most variable, whereas in reality low expressed have least evidence of true variation in gene expression. The simple RawCounts method, and also the scImpute method that produces results highly similar to those of the RawCounts method, both suffer from this same problem, and additionally have the problem of not correcting for variation in total UMI count across cells.

More striking, however, are the severe problems with the normalized expression values produced by the more sophisticated SAVER, DCA, MAGIC, and scVI algorithms. In particular, these methods produce not only strong artefactual correlations of the normalized expression values with the total UMI count in each cell, they also predict very large numbers of co-expressed genes when there is no evidence for co-expression at all in the raw data. The fact that this even occurs on synthetic data where there are no co-expressed genes at all confirms that such spurious correlations are inherently introduced by these normalization procedures.

We believe that these spurious correlations are introduced because all these methods confound noise removal with fitting the data to a lower-dimensional representation. Although it reasonable to assume that the possible states that cells can take on is much lower-dimensional than the full dimensionality of the transcriptome data, the problem of finding such lower-dimensional representations should be clearly distinguished from the problem of correcting for the biological and technical noise. Not only does this noise affect all genes almost independently, but because Poisson sampling noise scales with absolute expression level, different genes are affected by such noise to different extends and this may be erroneously mistaken for 'structure' in the data. Indeed, even though methods such as SAVER, DCA, MAGIC, and scVI specifically normalize for the total UMI count per cell, their normalized expression levels show strong correlations (and anti-correlations) with total UMI count. Thus, unless the process of noise removal and normalization is carefully separated from fitting of the data to lower-dimensional representations, artefactual correlations are likely to be introduced.

This is not to say that searching for lower-dimensional representations of the transcriptome data is not an important problem. Indeed, finding biologically meaningful lower-dimensional representations of genome-wide gene expression states is a key challenge in this field. However, we believe that this is a very hard problem in general, and it is currently not clear whether this problem is even solvable in principle, i.e. we are not aware of mathematical results that show under what conditions a lower-dimensional manifold embedded in a very high dimensional space can be reliably reconstructed from a limited number of noisy measurements. Our belief is that, rather than black box procedures for dimensionality reduction, progresss in understanding the genome-wide structure of expression data will crucially depend on connecting transcriptome data to the underlying molecular mechanisms, e.g. the folding of the chromosome, chromatin accessibility at enhancers and promoters, and the binding and unbinding of transcription factors.

However, whatever approach is taken to finding lower-dimensional representations of gene expression states, a prerequisite is that the raw data are first carefully normalized and corrected for both biological and technical sampling noise. The Sanity method that we presented here aims to provide such normalization methodology.

Supplementary Methods

Sanity

We denote, for each cell c and each gene g, the transcription rate a time t in the past as $\lambda_{gc}(t)$ and the decay rate of its mRNAs a time t in the past as $\mu_{gc}(t)$. Given these time-dependent transcription and decay rates, we define the *transcription activity* a_{gc} of gene g in cell c as the expected number of mRNAs $\langle m_{gc} \rangle$ which can be written as the following integral

$$a_{gc} = \langle m_{gc} \rangle = \int_0^\infty dt \lambda_{gc}(t) \exp\left[-\int_0^t \mu_{gc}(s) ds \right]. \tag{8}$$

That is, the transcription activity a_{gc} is a weighted time average of the recent transcription rates a time t in the past, with the weight equal to the expected fraction of surviving mRNAs produced at time t.

Conditioned on the transcription activity a_{gc}, the distribution of the actual number of mRNAs m_{gc} for gene g in cell c is given by a simple Poisson distribution

$$P(m_{gc}|a_{gc}) = \frac{(a_{gc})^{n_{gc}}}{n_{gc}!} e^{-a_{gc}}. \tag{9}$$

We now assume that, in the scRNA-seq measurement, each mRNA existing in cell c has a probability p_c to be captured and sequenced. Given this, the probability that precisely n_{gc} unique mRNAs will be sequenced for gene g in cell c is given by

$$\begin{aligned} P(n_{gc}|a_{gc}, p_c) &= \sum_{m_{gc}=n_{gc}}^{\infty} \binom{m_{gc}}{n_{gc}} (p_c)^{n_{gc}} (1-p_c)^{m_{gc}-n_{gc}} P(m_{gc}|a_{gc}) & (10) \\ &= \frac{(p_c a_{gc})^{n_{gc}}}{n_{gc}!} e^{-p_c a_{gc}}, & (11) \end{aligned}$$

which is still a Poisson distribution.

Next, we define the transcription activity a_{gc} as a product of the total transcription activity $A_c = \sum_g a_{gc}$ in cell c and a *transcription quotient* α_{gc}:

$$\alpha_{gc} = \frac{a_{gc}}{\sum_{g'} a_{g'c}} = \frac{a_{gc}}{A_c}, \tag{12}$$

i.e. α_{gc} is the expected fraction of all mRNAs in cell c that are mRNAs for gene g. If we also define the cell dependent constant $\lambda_c = p_c A_c$, then we can rewrite this Poisson distribution as

$$P(n_{gc}|\alpha_{gc}, \lambda_c) = \frac{(\lambda_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-\lambda_c \alpha_{gc}}. \tag{13}$$

The probability for the entire data-set in cell c has the form:

$$P(\{n_c\}|\{\alpha_c\}, \lambda_c) = \prod_g \left[\frac{(\lambda_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-\lambda_c \alpha_{gc}} \right], \tag{14}$$

where the notation $\{n_c\}$ refers to all counts n_{gc} for cell c, and $\{\alpha_c\}$ refers to all transcription quotients α_{gc} for cell c. Next, we remove the dependence on the constant λ_c by setting it to its maximum likelihood value. Noting that $\sum_g \alpha_{gc} = 1$ per definition, the dependence of the likelihood $P(\{n_c\}|\{\alpha_c\}, \lambda_c)$ on λ_c has the form $P(\{n_c\}|\{\alpha_c\}, \lambda_c) \propto \lambda_c^{N_c} e^{-\lambda_c}$, with $N_c = \sum_g n_{gc}$ the total number of sequenced mRNAs in cell c. Thus, the value of λ_c that maximizes $P(\{n_c\}|\{\alpha_c\}, \lambda_c)$ is simply $\lambda_c = N_c$. Substituting this we obtain

$$P(\{n_c\}|\{\alpha_c\}) = \prod_g \left[\frac{(N_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c \alpha_{gc}} \right]. \tag{15}$$

That is, the number of sequenced mRNAs n_{gc} for each each gene g in cell c is still a Poisson distribution with expectation value $N_c \alpha_{gc}$. The probability of the entire dataset of counts $\{n\}$ given all transcription quotients $\{\alpha\}$ is given by simply taking the product of this expression over all cells, i.e

$$P(\{n\}|\{\alpha\}) = \prod_{c,g} \left[\frac{(N_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c \alpha_{gc}} \right]. \tag{16}$$

Instead of trying to estimate the α_{gc} for all genes at once, we will focus on one specific gene g at a time, and infer how α_{gc} varies across the cells c. Note that if we collect all the terms that depend on the α_{gc} of single gene g we obtain

$$P(\{n_g\}|\{\alpha_g\}) = \prod_c \left[\frac{(N_c \alpha_{gc})^{n_{gc}}}{n_{gc}!} e^{-N_c \alpha_{gc}} \right], \tag{17}$$

where $\{n_g\}$ is the set of counts for gene g and $\{\alpha_g\}$ is the set of transcription quotients for gene g.

Finally, without loss of generality, we will write the transcription quotients α_{gc} in terms of the average quotient of the gene α_g and a log-fold change δ_{gc} in a given cell c, i.e.

$$\alpha_{gc} = \alpha_g e^{\delta_{gc}}. \tag{18}$$

In terms of these parameters we have

$$P(\{n_g\}|\alpha_g, \{\delta_g\}) = \left(\prod_c \frac{N_c^{n_{gc}}}{n_{gc}!}\right) \alpha_g^{n_g} \exp\left[\sum_c n_{gc}\delta_{gc} - \alpha_g \sum_c N_c e^{\delta_{gc}}\right], \tag{19}$$

where n_g is the total number of sequenced mRNAs for gene g.

Marginalizing over the average transcription quotient α_g

We now first focus on estimating the log fold-changes δ_{gc}. We return to estimating the overall average transcription quotient α_g once we have determined these. To marginalize expression (19) over α_g we use a simple uniform prior $P(\alpha_g)d\alpha_g \propto d\alpha_g$. Integrating with this uniform prior from 0 to ∞ we obtain

$$P(\{n_g\}|\{\delta_g\}) = \left(\prod_c \frac{N_c^{n_{gc}}}{n_{gc}!}\right) \Gamma(n_g+1) \exp\left(\sum_c n_{gc}\delta_{gc} - (n_g+1)\log\left[\sum_c N_c e^{\delta_{gc}}\right]\right). \tag{20}$$

Note that, because α_g is a fraction, we should have really only integrated from 0 to 1, but as long as each gene is only responsible for a small fraction of all UMIs in the cell, the only contribution to the integral comes from values of α_g much smaller than 1, and we can extend the range of the integral to infinity without loss of accuracy. Note also that the factor $\left(\prod_c \frac{N_c^{n_{gc}}}{n_{gc}!}\right)\Gamma(n_g+1)$ is determined entirely by the counts and does not depend on the δ_{gc}, and we will neglect this prefactor from now on.

Including prior probabilities for the δ_{gc}

We next introduce prior probabilities over the log fold-changes δ_{gc}. Assuming only that the δ_{gc} for gene g have a variance v_g and mean zero, we use the maximum entropy distribution consistent with these constraints, which is a Gaussian

$$P(\delta_{gc}|v_g) = \frac{1}{\sqrt{2\pi v_g}} \exp\left[-\frac{\delta_{gc}^2}{2v_g}\right]. \tag{21}$$

Thus, the prior over the full set $\{\delta_g\}$ of log fold-changes for the C cells is given by

$$P(\{\delta_g\}|v_g) \propto (v_g)^{-C/2} \exp\left(-\frac{1}{2v_g}\sum_{c=1}^{C} \delta_{gc}^2\right). \tag{22}$$

Combining the prior with the likelihood we obtain

$$P(\{n_g\}, \{\delta_g\}|v_g) = (v_g)^{-C/2} \exp\left(-\frac{1}{2v_g}\sum_c \delta_{gc}^2 + \sum_c n_{gc}\delta_{gc} - (n_g+1)\log\left[\sum_c N_c e^{\delta_{gc}}\right]\right), \tag{23}$$

up to a prefactor that does not depend on the parameters δ_{gc} and v_g.

Calculating $P(\{n\}|v_g)$ using the Laplace approximation

We next focus on calculating the probability $P(\{n_g\}|v_g)$ of the data given only the variance v_g. To obtain the probability $P(\{n_g\}|v_g)$, we need to integrate over all possible δ_{gc}. As the integral is close to Gaussian in form, we will assume we can approximate the integral by the Laplace approximation, i.e. by approximating the log-likelihood $L(\{\delta_g\}, v_g) = \log[P(\{n_g\}, \{\delta_g\}|v_g)]$ by expanding it to second order around its maximum. The log-likelihood has the form

$$L(\{\delta_g\}, v_g) = -\frac{C}{2}\log(v_g) - \frac{1}{2v_g}\sum_c \delta_{gc}^2 + \sum_c n_{gc}\delta_{gc} - (n_g+1)\log\left(\sum_c N_c e^{\delta_{gc}}\right). \tag{24}$$

Taking derivatives with respect to the δ_{gc}, the equations for the optimum become

$$-\frac{\delta_{gc}}{v_g} + n_{gc} - (n_g+1)\frac{N_c e^{\delta_{gc}}}{\sum_{\tilde{c}} N_{\tilde{c}} e^{\delta_{g\tilde{c}}}} = 0 \;\forall c. \tag{25}$$

To solve this equation we are going to multiply the equation by v_g and then define the c-independent quantity

$$e^{q_g} = \sum_c N_c e^{\delta_{gc}}, \tag{26}$$

the normalized quantities

$$f_{gc} = e^{-q_g} N_c e^{\delta_{gc}}, \tag{27}$$

which sum to 1, i.e. $\sum_c f_{gc} = 1$, and the c-dependent quantities

$$y_{gc} = v_g n_{gc} + \log(N_c), \tag{28}$$

which are directly determined by v_g and the data.

In terms of these quantities the equations for the optimum become

$$\log(f_{gc}) + v_g(n_g+1)f_{gc} = -q_g + y_{gc} \;\forall c, \tag{29}$$

whose solution is

$$f_{gc} = \frac{W\left[e^{-q_g+y_{gc}} v_g(n_g+1)\right]}{v_g(n_g+1)}, \tag{30}$$

with $W(x)$ the Lambert W-function (also called productlog). Note, however, that the solution depends on q_g, which itself depends on the f_{gc}. However, since $\sum_c f_{gc} = 1$ per definition, we can sum equation (30) over c to obtain the following consistency equation for q_g

$$\sum_c \frac{W\left[e^{-q_g+y_{gc}} v_g(n_g+1)\right]}{v_g(n_g+1)} = 1. \tag{31}$$

In the above equation, everything is determined either by the data (n_{gc}, n_g, and N_c) or the variance v_g, except for the unknown constant q_g, which needs to be solved for numerically. We can perform a binary search to find the value of q_g for which equation (31) is satisfied. Note also that the expression on the left hand side of equation (31) is a monotonically decreasing function of q_g, guaranteeing that there is only a single solution for q_g.

Once q_g has been determined, we obtain the f_{gc} from equation (30) and we obtain the optimal δ^*_{gc} as

$$\delta^*_{gc} = \log(f_{gc}) - \log(N_c) + q_g. \tag{32}$$

Note that these optimal δ^*_{gc} are functions of the variance v_g, which we from now on will express explicitly in our notation.

Substituting the optimal $\delta^*_{gc}(v_g)$ into equation (24) we obtain the optimal log-likelihood $L_*(v_g)$. By expanding the log-likelihood to second order around its maximum, the probability $P(\{n_g\}, \{\delta_g\}|v_g)$ can then be rewritten as

$$P(\{n_g\}, \{\delta_g\}|v_g) = \exp\left[L_*(v_g) - \frac{1}{2}\sum_{c,\tilde{c}}(\delta_{gc} - \delta^*_{gc}(v_g))M^g_{c\tilde{c}}(\delta_{g\tilde{c}} - \delta^*_{g\tilde{c}}(v_g))\right], \tag{33}$$

where the matrix M^g is given by the second derivatives of the log-likelihood around its optimum, i.e.

$$\frac{\partial^2 L}{\partial\delta_{gc}\partial\delta_{g\tilde{c}}}\|_* = -M^g_{c\tilde{c}}. \tag{34}$$

We find

$$M^g_{c\tilde{c}} = \left((n_g+1)f^*_{gc}(v_g) + \frac{1}{v_g}\right)\delta_{c\tilde{c}} - (n_g+1)f^*_{gc}(v_g)f^*_{g\tilde{c}}(v_g). \tag{35}$$

The integral over the likelihood can now be easily written in terms of the determinant of the matrix M^g, given us for the marginal probability of the data as a function of v_g:

$$P(\{n_g\}|v_g) = \int P(\{n_g\}, \{\delta_g\}|v_g)d\{\delta_g\} = \frac{e^{L_*(v_g)}}{\sqrt{\det(M^g)}}. \tag{36}$$

Finally, given the relatively simple structure of the matrix M^g, we use the *matrix determinant lemma* and write the determinant as

$$\det(M^g) = \left(1 - \sum_c \frac{(n_g+1)(f^*_{gc}(v_g))^2}{(n_g+1)f^*_{gc}(v_g) + \frac{1}{v_g}}\right)\prod_c\left((n_g+1)f^*_{gc}(v_g) + \frac{1}{v_g}\right). \tag{37}$$

Posterior $P(v_g|\{n\})$ over variance v_g

To obtain a posterior over the variance v_g we need a prior over the variance v_g, for which we will use a scale prior, i.e. uniform in the logarithm of v_g: $P(v_g)dv_g \propto d\log(v_g)$. Note, however, that our solution of $P(\{n\}|v_g)$ involved a numerical determination of q_g, so that we do not have an analytical formula for $P(v_g|\{n\})$. In order to approximate the full posterior $P(v_g|\{n\})$ we pick a range $[v_{\min}, v_{\max}]$ within which we presume all v_g fall, divide this range into B bins of equal size in $\log(v_g)$, and calculate $P(\{n\}|v_g)$ for each bin b. Per default we choose $[v_{\min}, v_{\max}] = [0.01, 20]$ since this covers the range of observed variances in the datasets we considered. Trading off speed versus accuracy we chose $B = 116$ bins by default, so that the variance increase by about 5% from one bin to the next. However, if desired these values can be changed by the user.

Let v_b denote the variance of bin b and L_b the log-likelihood $\log[P(\{n\}|v_b)]$. We then approximate the full posterior $P(v_b|\{n\})$ by a distribution over a finite number of points:

$$P(v_b|\{n\}) = \frac{e^{L_b}}{\sum_{b'=1}^{B} e^{L_{b'}}}. \tag{38}$$

The posterior $P(\{\delta_g\}|\{n\}, v_g)$ of log-fold changes given a variance v_g

For a given value of the variance v_g, the posterior distribution over the log fold-changes δ_{gc} is given by a multi-variate Gaussian with means $\langle \delta_{gc} \rangle = \delta^*_{gc}(v_g)$ and a covariance matrix C given by the inverse of the matrix M^g. In particular, the variances $\text{var}(\delta_{gc})$ of the log fold-changes across cells are given by the diagonal elements of the inverse of M^g. Fortunately, given the relatively simple structure of the matrix M^g, we can also obtain analytical expressions for these variances. In particular, the components (c, c) of the inverse of M^g are given by the ratio of the minor $[M^g]_{(c,c)}$ (the determinant of matrix M^g with the cth row and column removed) and the determinant of the full matrix. We have

$$\begin{aligned}
\text{var}(\delta_{g\tilde{c}}) &= \frac{[M^g]_{\tilde{c},\tilde{c}}}{\det(M^g)} & (39)\\
&= \frac{\left(1 - \sum_{c\neq\tilde{c}} \frac{(n_g+1){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right) \prod_{c\neq\tilde{c}} \left((n_g+1)f^*_{gc} + \frac{1}{v_g}\right)}{\left(1 - \sum_{c} \frac{(n_g+1){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right) \prod_{c} \left((n_g+1)f^*_{gc} + \frac{1}{v_g}\right)} & (40)\\
&= \frac{\left(1 - \sum_{c\neq\tilde{c}} \frac{(n_g+a){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right)}{\left(1 - \sum_{c} \frac{(n_g+1){f^*_{gc}}^2}{(n_g+1)f^*_{gc}+\frac{1}{v_g}}\right) \left((n_g+1)f^*_{gc} + \frac{1}{v_g}\right)}, & (41)
\end{aligned}$$

where again it should be noted that the f^*_{gc} are themselves functions of v_g.

A technical complication arises in estimating the variance $\text{var}(\delta_{gc})$ when the observed number of UMIs is zero. That is, when $n_{gc} = 0$ the log-likelihood $L(\{\delta_g\}, v_g)$ can be a highly asymmetric function of δ_{gc} around its maximum $\delta^*_{gc}(v_g)$. In particular, whereas the fact that no UMIs were observed, i.e. $n_{gc} = 0$, ensures that the log-likelihood decreases quickly as δ_{gc} increases above $\delta^*_{gc}(v_g)$, it drops only slowly with decreasing δ_{gc}. That is, when no UMI are observed, we can give a reasonably tight upper bound on δ_{gc}, but $n_{gc} = 0$ is consistent with very low δ_{gc}. This asymmetry causes the variance $\text{var}(\delta_{gc})$ to significantly overestimate the error-bar in δ_{gc} toward larger values of δ_{gc}. To fix this problem, we directly set $\text{var}(\delta_{gc})$ from its upper bound for cases with $n_{gc} = 0$. In particular, note that for a Gaussian distribution with mean μ and variance σ^2, the difference between the log-likelihood at the optimum μ and at $\mu + \sigma$ is $L(\mu) - L(\mu + \sigma) = (\mu + \sigma - \mu)^2/(2\sigma^2) = 1/2$. We thus define the $\sigma_{gc} = \sqrt{\text{var}(\delta_{gc})}$ such that the difference between the log-likelihood at δ^*_{gc} and $\delta^*_{gc} + \sigma_{gc}$ is $1/2$, i.e. the solution of

$$L(\delta^*_{gc}) - L(\delta^*_{gc} + \sigma_{gc}) = \frac{\sigma_{gc}(2\delta^*_{gc} + \sigma_{gc})}{2v_g} + (n_g + 1)\log(1 + f^*_{gc}(e^{\sigma_{gc}} - 1)) = \frac{1}{2}, \tag{42}$$

which we determine numerically.

Final estimates $\langle\delta_{gc}\rangle$ and error-bars ϵ_{gc}

For each value of v_g, we have determined the posterior probability $P(v_g|\{n\})$ and given a variance v_g, we have a Gaussian posterior distribution $P(\{\delta_g\}|\{n\}, v_g)$ over the log fold-changes, with means $\delta^*_{gc}(v_g)$ and variances $\text{var}(\delta_{gc})(v_g)$. Using these, we can now calculate final estimates of the fold changes δ_{gc}. In particular, the expectation value $\langle\delta_{gc}\rangle$ is given by the integral

$$\langle\delta_{gc}\rangle = \int dv_g d\{\delta_g\}\delta_{gc} P(\{\delta_g\}|\{n_g\}, v_g)P(v_g|\{n\}) = \int dv_g \delta^*_{gc}(v_g)P(v_g|\{n\}). \tag{43}$$

Similarly, we find for the overall error-bar ϵ^2_{gc}

$$\begin{aligned}
\epsilon^2_{gc} &= \langle(\delta_{gc})^2\rangle - \langle\delta_{gc}\rangle^2 && (44)\\
&= \int dv_g \left[\text{var}(\delta_{gc})(v_g) + \left(\delta^*_{gc}(v_g)\right)^2\right] P(v_g|\{n\}) - \langle\delta_{gc}\rangle^2 && (45)\\
&= \int dv_g \left(\text{var}(\delta_{gc})(v_g) + \left(\delta^*_{gc}(v_g) - \langle\delta_{gc}\rangle\right)^2\right) P(v_g|\{n\}) && (46)
\end{aligned}$$

Sanity returns, for each gene g in each cell c, both the estimated log fold-change $\langle\delta_{gc}\rangle$ and its error-bar ϵ_{gc}.

Mean expression $\langle\log(\alpha_g)\rangle$

Once we have fitted a set of $\delta^*_{gc}(v_g)$ for each v_g, and determined the posterior $P(v_g|\{n_g\})$ we can now easily estimate the mean log quotient $\mu_g = \log(\alpha_g)$ of each gene. Returning to equation (19), and marginalizing over the δ_{gc} we find that the posterior over α_g is proportional to the expression (19) in which the δ_{gc} have been set to $\delta^*_{gc}(v_g)$:

$$P(\bar{\alpha}_g|\{n_g\}, v_g) \propto (\alpha_g)^{n_g} \exp\left[-\alpha_g e^{q_g(v_g)}\right], \tag{47}$$

where n_g is the total number of UMIs captured for gene g, $e^{q_g(v_g)} = \sum_c N_c e^{\delta^*_{gc}(v_g)}$ as defined above, and we have explicitly indicated that q_g is a function of the variance v_g.

Using (47) the expectation value of $\log(\alpha_g)$ at a given value of the variance v_g is given by

$$\langle\log(\alpha_g)\rangle_{v_g} = \psi(n_g + 1) - q_g(v_g), \tag{48}$$

where $\psi(x)$ is the digamma function, i.e. the derivative of the logarithm of the gamma function. Note also that, since n_g is an integer, we have $\psi(n_g + 1)$ is simply related to the Harmonic numbers, i.e. $\psi(n_g + 1) = -\gamma + \sum_{k=1}^{n_g} 1/k$, with $\gamma \approx 0.577$ the Euler–Mascheroni constant.

To get a final estimate $\mu_g = \langle\log(\alpha_g)\rangle$ we obtain the weighted average over the variance v_g, i.e.

$$\mu_g = \psi(n_g + 1) - \int dv_g q_g(v_g)P(v_g|\{n\}) = \psi(n_g + 1) - \langle q_g\rangle. \tag{49}$$

Error bar on mean expression

Going back to equation (47) we find that the variance in $\log(\alpha_g)$, at a given value of the variance v_g, is given by the derivative of the digamma function:

$$\text{var}(\log(\alpha_g))_{v_g} = \psi_1(n_g + 1), \tag{50}$$

with $\psi_1(x)$ the derivative of the digamma function, which is also called the trigamma function. Note that this variance is independent of v_g.

The final error-bar $\delta\mu_g$ for $\log(\alpha_g + 1)$ is then given

$$(\delta\mu_g)^2 = \psi_1(n_g + 1) + \int dv_g \left(q_g(v_g) - \langle q_g \rangle\right)^2 P(v_g|\{n_g\}). \tag{51}$$

Note that, as for the calculation of the log fold-changes, these integrals over v_g are approximated by sums over the same set of B bins.

Simulated dataset

Defining N_{gene} the number of genes, N_{cell} the number of cells, μ_g the mean LTQ of gene g, v_g the variance of the LTQs of gene g across cells, N_c the total number of sequenced mRNAs in cell c, a_{gc} the transcription activity of gene g in cell c, and α_{gc} the transcription quotient of gene g in cell c, we simulated the UMI counts n_{gc} as follows:

$$\begin{aligned}
N_{gene} &= 16'016 \\
N_{cell} &= 1'937 \\
\mu_g & \quad \text{Taken from the measured mean LTQ per gene in the Baron dataset.} \\
v_g & \quad \text{Sampled from a uniform distribution } \mathcal{U}(0,6). \\
N_c &= \text{Taken from the observed total UMI counts per cell in the Baron dataset.} \\
a_{gc} &\sim \exp\left(\mathcal{N}(\mu_g, v_g)\right) \\
\alpha_{gc} &= \frac{a_{gc}}{\sum_g a_{gc}} \\
n_{gc} &\sim \text{Poisson}(N_c \alpha_{gc})
\end{aligned}$$

That is, the mean LTQs μ_g were taken from the mean LTQs measured on the Baron dataset. The variances in LTQs were drawn from a Uniform distribution between 0 and 6. The total number of UMIs per cell was chosen identical to the total number of UMIs per cell observed in the Baron dataset. For each gene g, the transcription activity a_{gc} of each cell c was then drawn from a log-normal with mean of $\log(a_{gc})$ equal to μ_g and variance v_g, and the transcription quotients α_{gc} were set by normalizing to the total transcription activity of each cell. Finally, the observed UMI counts n_{gc} were drawn from a Poisson with mean $N_c\alpha_{gc}$ for each gene g in each cell c.

Figure S11 shows the distributions of the total number of mRNAs per cell, the total number of mRNAs per gene, and the variance in observed mRNA counts for both the Baron dataset and the simulated data. Note that the distributions are highly similarly except for the variances, which are more widely distributed in the simulated data.

Clustering index

Let the sets $\{A\}$ and $\{B\}$ denote two cell classifications where $A_i \in \mathbb{N}_+$ and $B_i \in \mathbb{N}_+$ denote the class numbers of cell i in the two classifications, and $i = 1, ..., C$, with C the number of cells ($i.e.$ the number of elements in sets $\{A\}$ and $\{B\}$).

The distributions and the joint distribution of the two classifications are defined as the frequencies

$$P_A(a) = \frac{|\{A_i = a | 1 \leq i \leq C\}|}{C} \tag{52}$$

$$P_B(b) = \frac{|\{B_i = b | 1 \leq i \leq C\}|}{C} \tag{53}$$

$$P_{AB}(a, b) = \frac{|\{\{A_i = a\} \cap \{B_i = b\} | 1 \leq i \leq C\}|}{C}, \tag{54}$$

where $|\cdot|$ denotes the cardinality of a set.

The entropy of the distributions and the joint distribution are defined

$$H(A) = -\sum_{a \in \mathbb{N}_+} P_A(a) \log P_A(a) \tag{55}$$

$$H(B) = -\sum_{b \in \mathbb{N}_+} P_B(b) \log P_B(b) \tag{56}$$

$$H(A, B) = -\sum_{a,b \in \mathbb{N}_+} P_{AB}(a, b) \log P_{AB}(a, b). \tag{57}$$

The *mutual information* is defined

$$I(A; B) = H(A) + H(B) - H(A, B) \tag{58}$$

$$= \sum_{a \in A, b \in B} P_{AB}(a, b) \log \frac{P_{AB}(a, b)}{P_A(a) B_B(b)}, \tag{59}$$

representing the difference between the summed entropy of the 2 distributions and the entropy of the joint distribution.

We compute the *Normalized mutual information* as

$$NMI(A; B) = \frac{I(A; B)}{\sqrt{H(A)H(B)}}. \tag{60}$$

Alternatively, the confusion matrix being defined

	Inferred classes	
Reference	$B_i = B_j$	$B_i \neq B_j$
$A_i = A_j$	TP	FN
$A_i \neq A_j$	FP	TN

Table S1: Confusion matrix

The *Adjusted rand index* is defined

$$ARI(A,B) \quad = \quad 2\frac{TP \cdot TN - FN \cdot FP}{(TP+FN)(FN+TN)+(TP+FP)(FP+TN)} \tag{61}$$

Differential expression

Let e_{gc} the log-expression of gene g in cell c, C an ensemble of cells, and $\bar{C}$ all other cells in the dataset. The t-statistic t_{gC} quantifies the statistical evidence that the average expression of gene g in the set C differs from the average in all other cells:

$$t_{gC} \quad = \quad \frac{\mu_{gC} - \mu_{g\bar{C}}}{\sqrt{\sigma^2_{gC}/|C| + \sigma^2_{g\bar{C}}/|\bar{C}|}} \tag{62}$$

$$\mu_{gC} \quad = \quad \frac{1}{|C|}\sum_{c\in C} e_{gc} \tag{63}$$

$$\mu_{g\bar{C}} \quad = \quad \frac{1}{|\bar{C}|}\sum_{c\in \bar{C}} e_{gc} \tag{64}$$

$$\sigma^2_{gC} \quad = \quad \frac{1}{|C|}\sum_{c\in C} (e_{gc} - \mu_{gC})^2 \tag{65}$$

$$\sigma^2_{g\bar{C}} \quad = \quad \frac{1}{|\bar{C}|}\sum_{c\in \bar{C}} \left(e_{gc} - \mu_{g\bar{C}}\right)^2, \tag{66}$$

with $|C|$ and $|\bar{C}|$ the number of cells in set C and its complement, respectively.

Given a t-statistic t_{gC}, the p-value under a one-side t-test that the gene is over-expressed in set C is given by

$$P(t_{gC}) = \frac{1}{2}\mathrm{Erfc}\left(\frac{t_{gC}}{\sqrt{2}}\right). \tag{67}$$

Sorting all genes by the t-statistic t_{gC} the list of over-expressed genes at a false discovery rate of f is obtained by picking a cut-off t_c such that average of $P(t_{gC})$ for all genes with $t_{gC} > t_c$ is f.

The reference sets of differentially expressed genes are constructed using a negative binomial generalized linear regression to obtain posterior probability distributions for the class-specific contributions to each gene's expression (also considering contribution of age and sex and a basal expression per gene) (see [48], Supplementary Materials, Gene expression enrichment analysis).

Correlation matrix distance

Given 2 correlation matrix R_1 and R_2, the *correlation matrix distance* (CMD) [53] measures a distance between R_1 and R_2, bound between 0 (equal) and 1 (most different) and is defined as

$$CMD(C_1, C_2) = 1 - \frac{tr(R_1 R_2)}{|R_1|_f |R_2|_f} \tag{68}$$

where $|\cdot|_f$ denotes the frobenius norm, and $tr(\cdot)$ the trace.

Supplementary figures

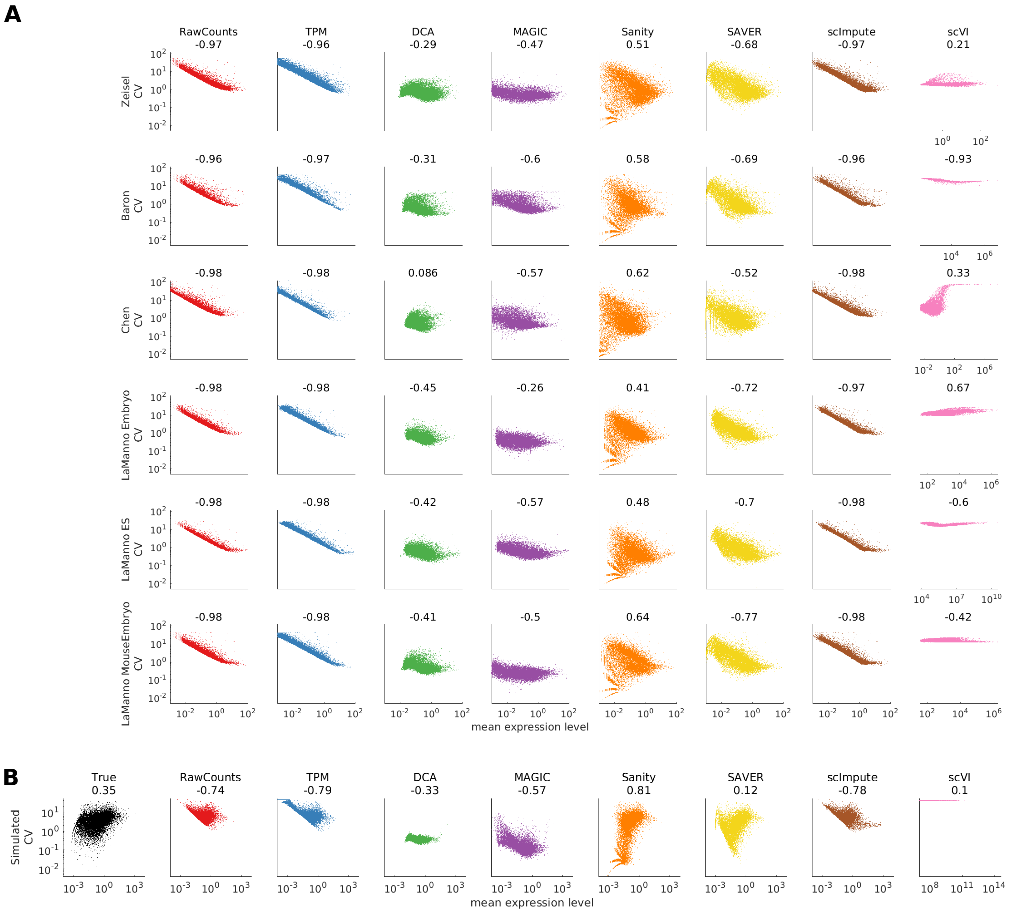

Figure S1: **A** Scatter plots of CV against mean gene expression level. Rows correspond to different scRNA-seq data. Colors and columns correspond to the different methods used to normalize the data.The axis are kept similar across panels, except for *scVI* for which the x-axis is different as the mean expression is on a very different scale compared to the others. **B** Scatter plots of CV against mean gene expression level. The panels and colors correspond to different methods used to normalize the data. The different methods and the correlation between the inferred CV and the true CV is shown on top of each panel.The axis are kept similar across panels, except for *scVI* for which the x-axis is different as the mean expression is on a very different scale compared to the others.

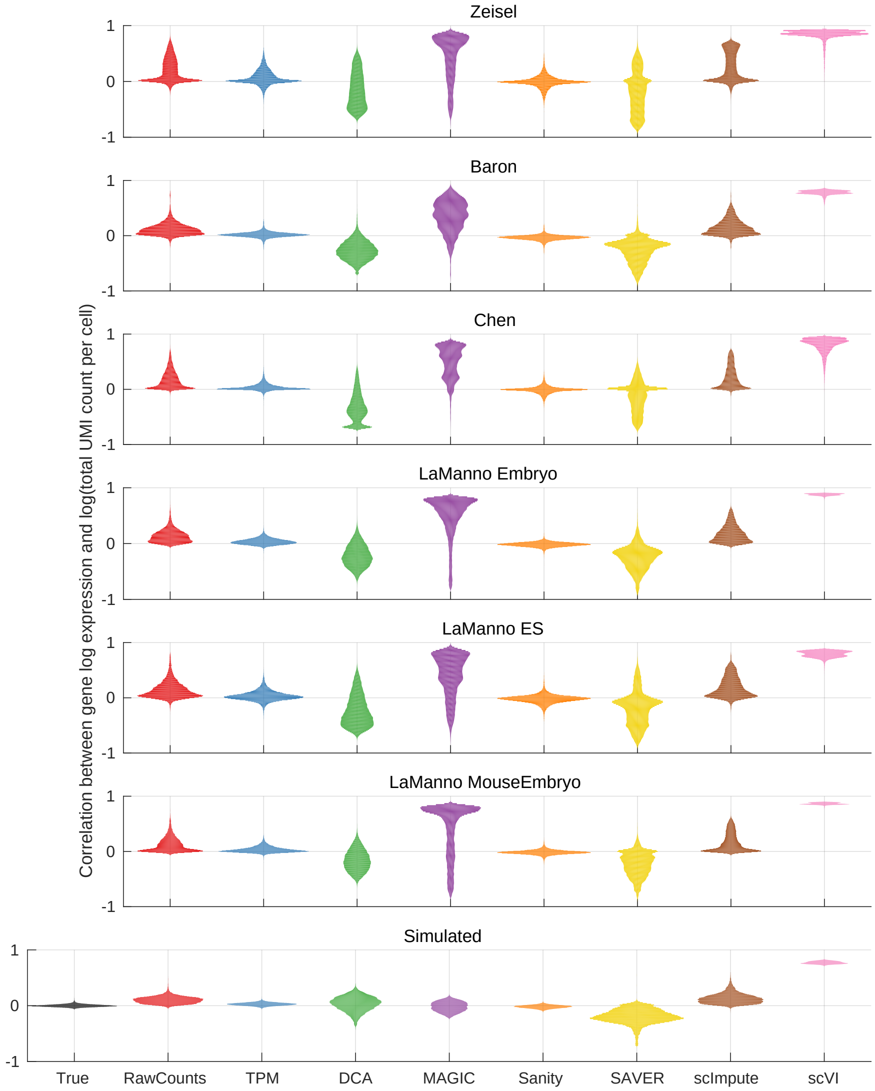

Figure S2: Violin plots of the distribution of correlation coefficients between inferred log expression level of genes and and log of total mRNA molecule count per cell. Rows correspond to different datasets, as indicated on top of each panel. Columns correspond the different methods, as indicated on the x-axis.

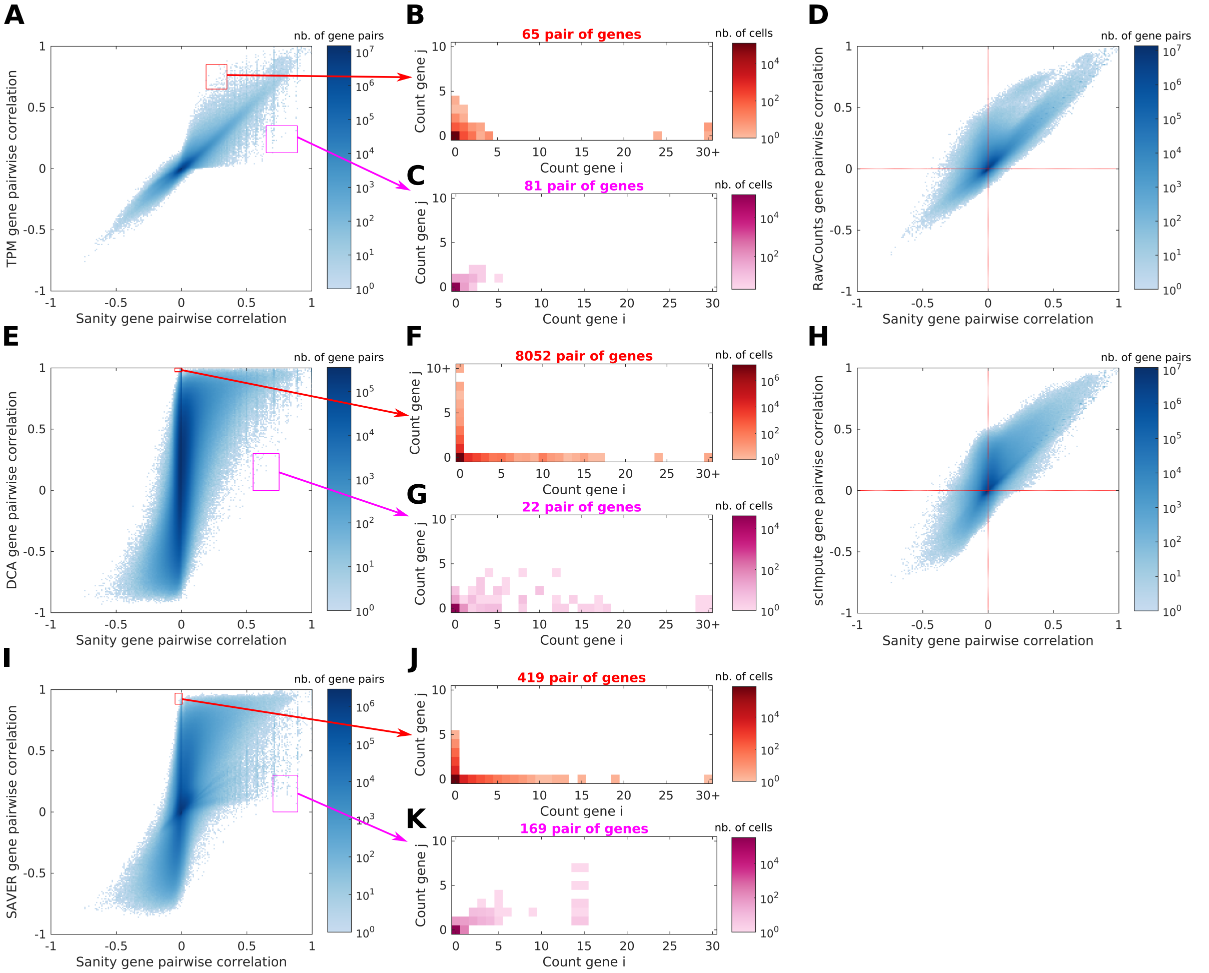

Figure S3: Density plot of the Pearson correlations of normalized expression values of all pairs of genes as inferred by Sanity (x-axis) and the correponding correlation as inferred by TPM (**A**), RawCounts (**D**), DCA (panel **E**), scImpute **H**, and SAVER (**I**)) on the y-axis for the Baron dataset. The color scale shows the density in log_{10} of gene pairs and values $\log_{10}(0)$ are shown in white. For panels **A**, **E**, and **I**, the red and magenta rectangles show selections of gene pairs for which the two methods disagree most strongly on the correlation. For each such set of pairs, we counted the number of $n_{i,j}$ across all pairs and all cells for which i UMI were observed for the first gene and j UMI for the second gene. The panels **B**, **C**, **F**, **G**, **J**, and **K** show the corresponding 2-dimensional histograms $n_{i,j}$ for each selected set with the number of pairs indicated above the panel. The height of the histogram is shown in $\log_{10}$ as a color and values $\log_{10}(0)$ are shown in white.

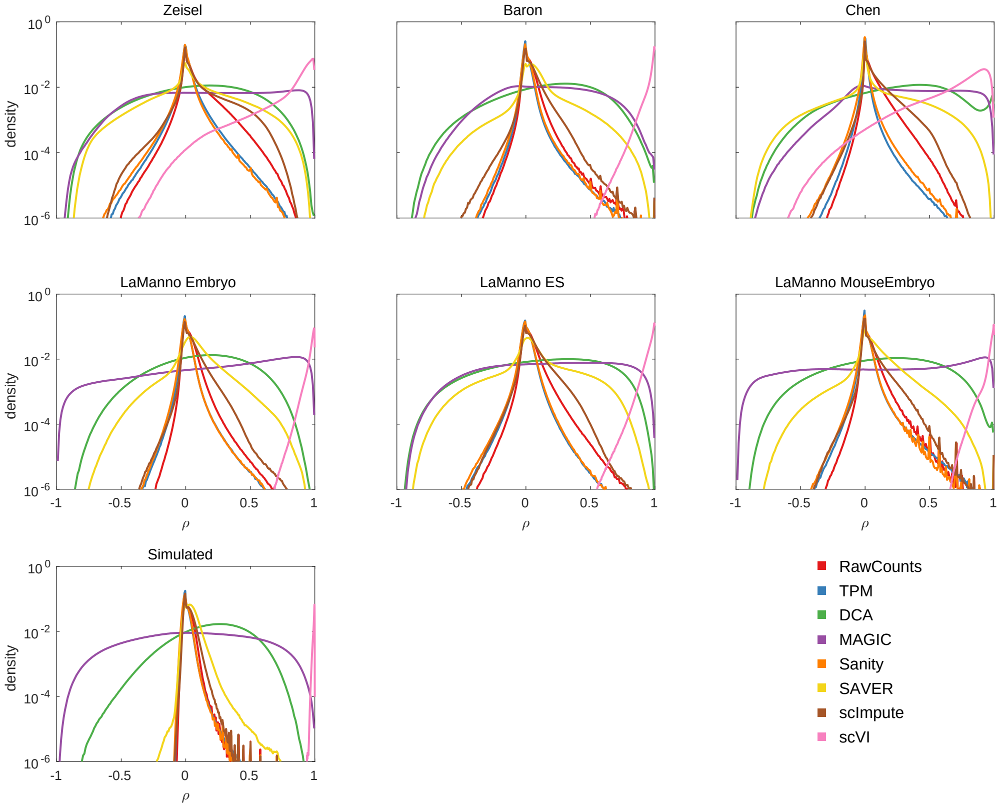

Figure S4: Distributions of the Pearson correlations of all pairs of genes, as inferred by each normalization method. Each panel corresponds to one dataset (indicated at the top of the panel) and each color corresponds to one of the normalization methods, as indicated in the legend. Note that the y-axis is shown on a logarithmic scale.

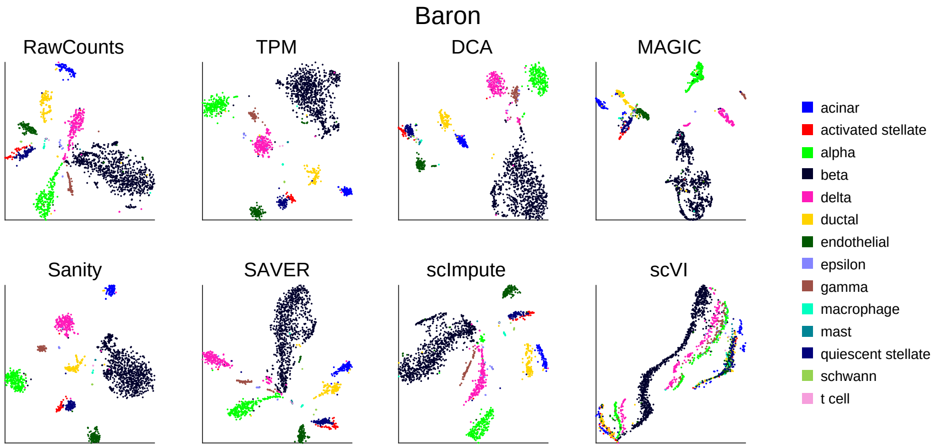

Figure S5: Each panel show a t-SNE visualization of the Baron dataset using the normalized gene expression values of the method indicated at the top. Each point represent a cell and is colored by the cell type annotated in the original publication.

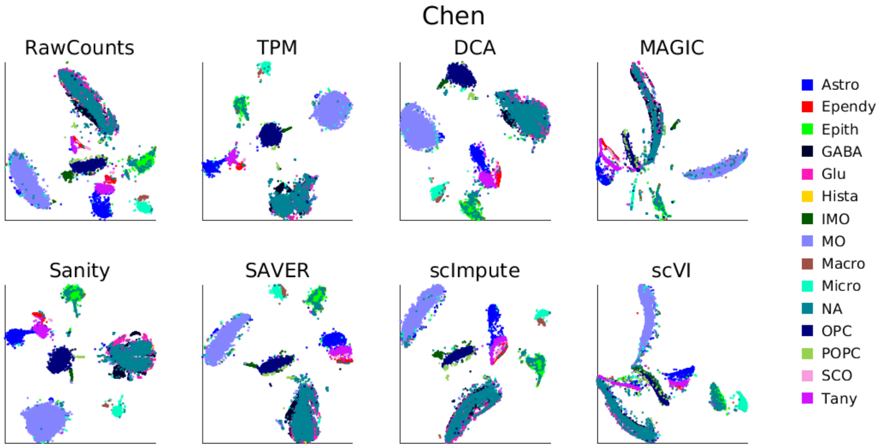

Figure S6: Each panel show a t-SNE visualization of the Chen dataset using the normalized gene expression values of the method indicated at the top. Each point represent a cell and is colored by the cell type annotated in the original publication.

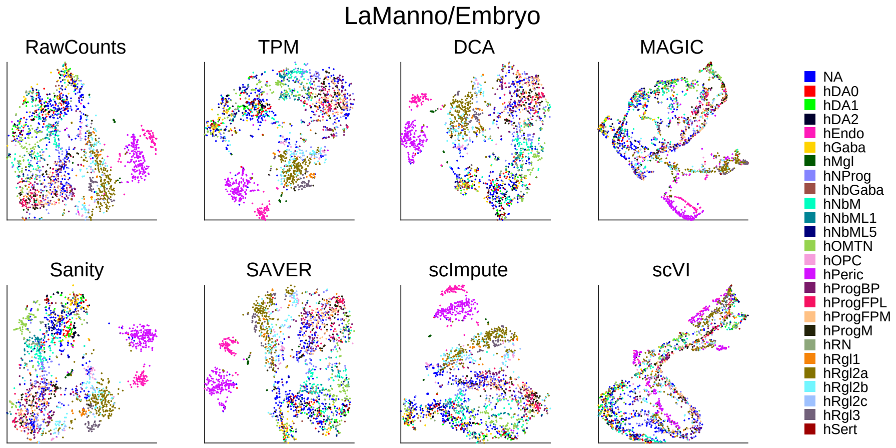

Figure S7: Each panel show a t-SNE visualization of the LaManno/Embryo dataset using the normalized gene expression values of the method indicated at the top. Each point represent a cell and is colored by the cell type annotated in the original publication.

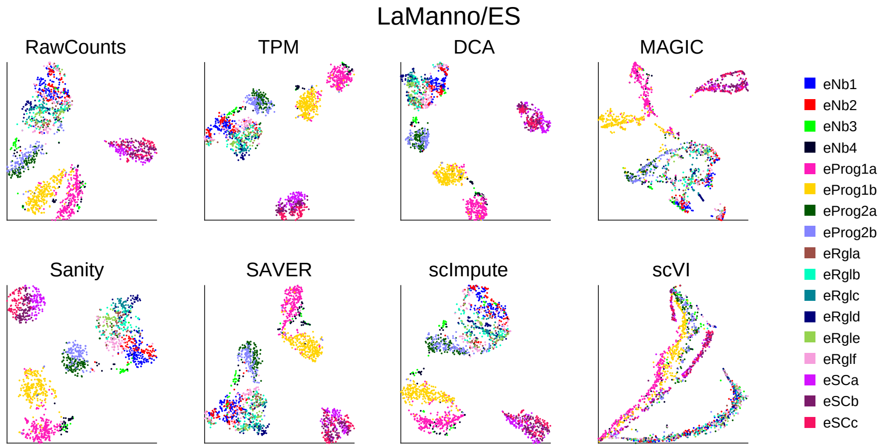

Figure S8: Each panel show a t-SNE visualization of the LaManno/ES dataset using the normalized gene expression values of the method indicated at the top. Each point represent a cell and is colored by the cell type

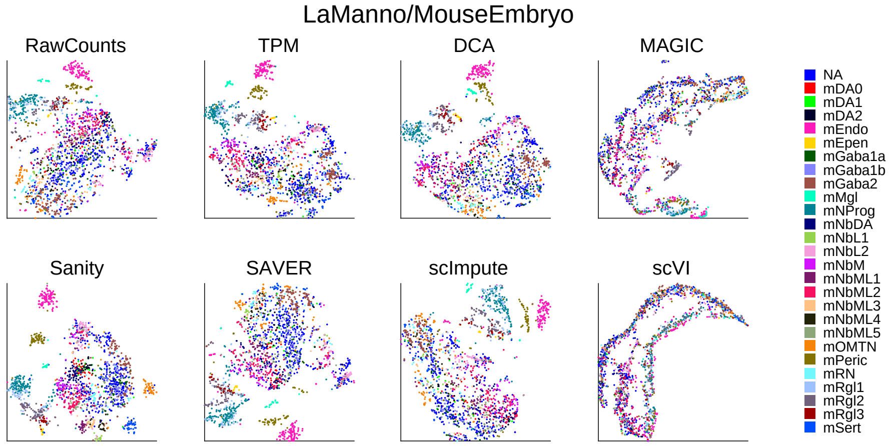

Figure S9: Each panel show a t-SNE visualization of the LaManno/MouseEmbryo dataset using the normalized gene expression values of the method indicated at the top. Each point represent a cell and is colored by the cell type annotated in the original publication.

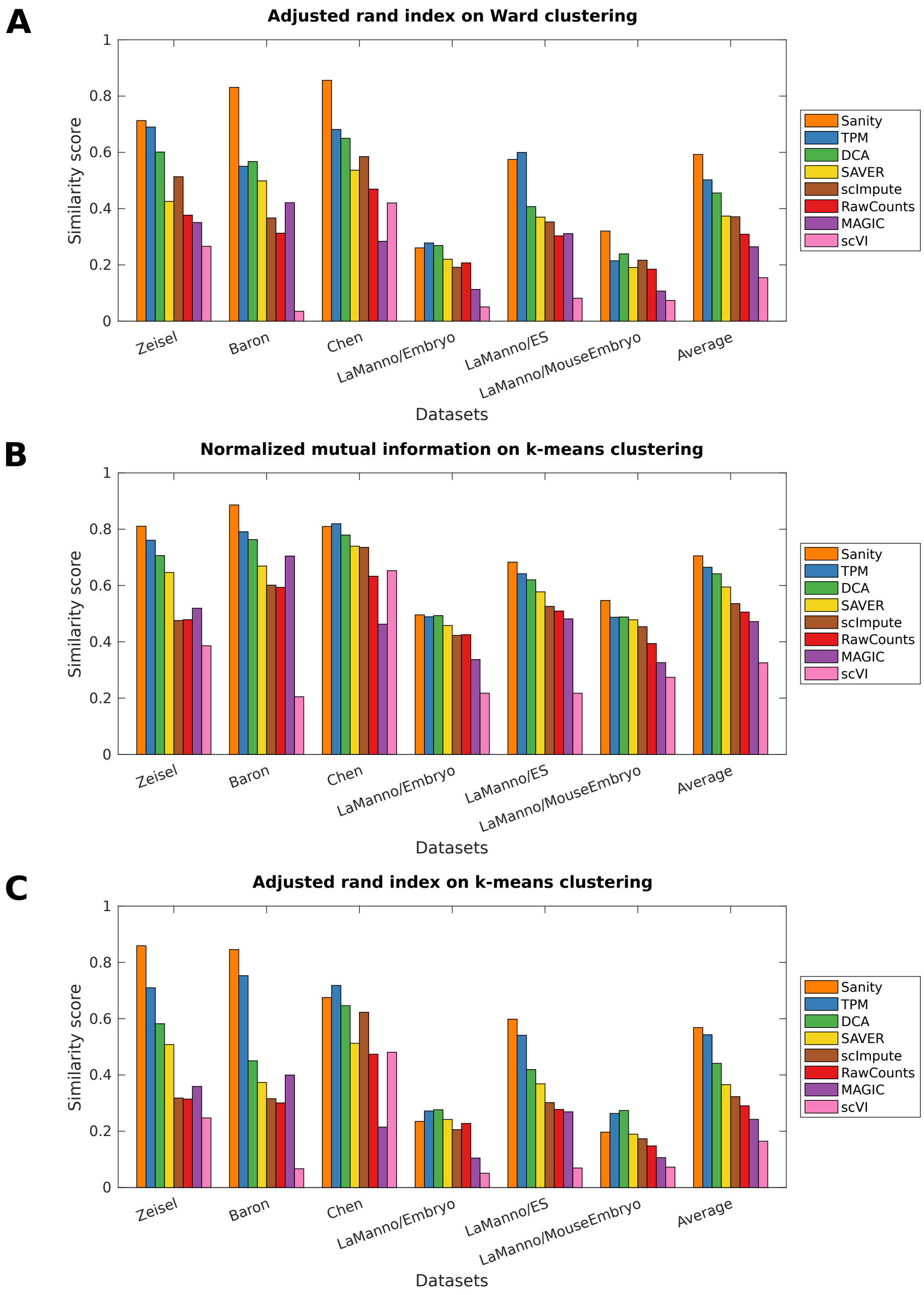

Figure S10: Similarity between the reference clusters and the clusters inferred using the normalized gene expression values of the different methods. Clustering was carried out using either hierarchical clustering with Ward's method (**A**) or using k-means clustering (**B** and **C**). The similarity measures used were the Adjusted rand index (**A** and **C**) and the normalized mutual information (**B**). Both measures take values between 0 (no similarity) and 1 (perfect similarity). Each group of bars shows the results for a particular dataset indicated below) and colors indicate the different methods (see legend). The last group of bars shows the average similarity per method across all datasets. Methods are sorted from left to right according to their average similarity.

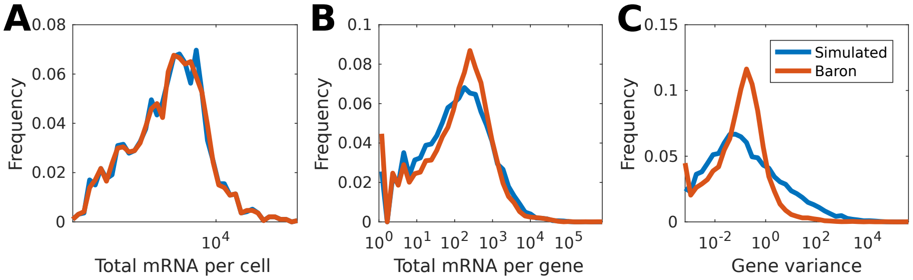

Figure S11: **A** Distribution of total mRNA captured per cell in the simulated dataset (blue) and the *Baron* dataset (red). **B** Distribution of total mRNA captured per gene in the simulated dataset (blue) and the *Baron* dataset (red). **C** Distribution of variance per gene calculated on the raw count matrix obtained from the simulated dataset (blue) and the *Baron* dataset (red).

References

[1] S. Picelli, A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1098, Nov 2013.

[2] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*, 2(3):666–673, Sep 2012.

[3] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.

[4] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

[5] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, Jul 2015.

[6] D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, May 2015.

[7] A. Rotem, O. Ram, N. Shoresh, R. A. Sperling, A. Goren, D. A. Weitz, and B. E. Bernstein. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, 33(11):1165–1172, Nov 2015.

[8] S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*, 11(8):817–820, Aug 2014.

[9] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, Oct 2013.

[10] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, Jul 2016.

[11] R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, and G. M. Church. Developmental barcoding of whole mouse via homing CRISPR. *Science*, 361(6405), 08 2018.

[12] K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, and L. Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, 01 2017.

[13] A. P. Frei, F. A. Bava, E. R. Zunder, E. W. Hsieh, S. Y. Chen, G. P. Nolan, and P. F. Gherardini. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods*, 13(3):269–275, Mar 2016.

[14] B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, and A. F. Schier. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.*, 36(5):442–450, Jun 2018.

[15] B. Spanjaard, B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov, and J. P. Junker. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.*, 36(5):469–473, Jun 2018.

[16] D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, and A. M. Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 06 2018.

[17] C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. Smallwood, C. P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, 13(3):229–232, Mar 2016.

[18] S. J. Clark, R. Argelaguet, C. A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun*, 9(1):781, 02 2018.

[19] B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nunez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev, and J. S. Weissman. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882, Dec 2016.

[20] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866, Dec 2016.

[21] D. A. Jaitin, A. Weiner, I. Yofe, D. Lara-Astiaso, H. Keren-Shaul, E. David, T. M. Salame, A. Tanay, A. van Oudenaarden, and I. Amit. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7):1883–1896, Dec 2016.

[22] P. Datlinger, A. F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, J. Klughammer, L. C. Schuster, A. Kuchler, D. Alpar, and C. Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, 14(3):297–301, 03 2017.

[23] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Gottgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, and N. Yosef. The Human Cell Atlas. *Elife*, 6, 12 2017.

[24] The LifeTime Initiative. https://lifetime-fetflagship.eu/.

[25] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, 5(10):877–879, Oct 2008.

[26] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008.

[27] Leland McInnes and John Healy. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv*, 2018. arXiv:1802.03426.

[28] M. Thattai. Universal Poisson Statistics of mRNAs with Complex Decay Pathways. *Biophys. J.*, 110(2):301–305, Jan 2016.

[29] O. Padovan-Merhar, G. P. Nair, A. G. Biaesch, A. Mayer, S. Scarfone, S. W. Foley, A. R. Wu, L. S. Churchman, A. Singh, and A. Raj. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell*, 58(2):339–352, Apr 2015.

[30] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, 2014.

[31] O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16(3):133–145, Mar 2015.

[32] P. Brennecke, S. Anders, J. K. Kim, A. A. Ko?odziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10(11):1093–1095, Nov 2013.

[33] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass. Making sense of microarray data distributions. *Bioinformatics*, 18(4):576–584, Apr 2002.

[34] Jacob Beal. Biochemical complexity drives log-normal variation in genetic expression. *Eng. Biol.*, 1(1):55–60, jun 2017.

[35] M. I. Love, S. Anders, V. Kim, and W. Huber. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res*, 4:1070, 2015.

[36] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, 05 2016.

[37] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[38] Cell Ranger DNA. https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/what-is-cell-ranger-dna.

[39] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, feb 2014.

[40] Aisha A. AlJanahi, Mark Danielsen, and Cynthia E. Dunbar. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Mol. Ther. - Methods Clin. Dev.*, 10:189–196, sep 2018.

[41] 10X Genomics. What fraction of mrna transcripts are captured per cell? https://kb.10xgenomics.com/hc/en-us/articles/360001539051-what-fraction-of-mrna-transcripts-are-captured-per-cell-, 2018.

[42] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[43] Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, 10(1):390, dec 2019.

[44] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, jul 2018.

[45] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, 15(7):539–542, 2018.

[46] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, 9(1):997, dec 2018.

[47] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, dec 2018.

[48] Amit Zeisel, A.B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Et Al., Ana B. Muñoz Manchado, Simone Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, S. Linnarsson, Peter Lönnerberg, Gioele La Manno, Anna Juréus, and Sueli Marques. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).*, 2015.

[49] Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.*, 2016.

[50] Renchao Chen, Xiaoji Wu, Lan Jiang, and Yi Zhang. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell Rep.*, 2017.

[51] Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E. Borm, Simon R.W. Stott, Enrique M. Toledo, J. Carlos Villaescusa, Peter Lönnerberg, Jesper Ryge, Roger A. Barker, Ernest Arenas, and Sten Linnarsson. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, 2016.

[52] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.*, 58(301):236–244, mar 1963.

[53] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek. Correlation Matrix Distance, a Meaningful Measure for Evaluation of Non-Stationary MIMO Channels. In *2005 IEEE 61st Veh. Technol. Conf.*, volume 1, pages 136–140. IEEE, 2005.