ANALYSIS
https://doi.org/10.1038/s41587-021-00875-x

nature biotechnology

Check for updates

Bayesian inference of gene expression states from single-cell RNA-seq data

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

Despite substantial progress in single-cell RNA-seq (scRNA-seq) data analysis methods, there is still little agreement on how to best normalize such data. Starting from the basic requirements that inferred expression states should correct for both biological and measurement sampling noise and that changes in expression should be measured in terms of fold changes, we here derive a Bayesian normalization procedure called Sanity (SAmpling-Noise-corrected Inference of Transcription activitY) from first principles. Sanity estimates expression values and associated error bars directly from raw unique molecular identifier (UMI) counts without any tunable parameters. Using simulated and real scRNA-seq datasets, we show that Sanity outperforms other normalization methods on downstream tasks, such as finding nearest-neighbor cells and clustering cells into subtypes. Moreover, we show that by systematically overestimating the expression variability of genes with low expression and by introducing spurious correlations through mapping the data to a lower-dimensional representation, other methods yield severely distorted pictures of the data.

In the past decade, much effort has been invested in adapting methods for quantifying transcriptomic and epigenomic states on a genome-wide scale to the single-cell level. This has led to the development of 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[1–22]. It is widely expected that these methods will revolutionize our understanding of the ways in which cell fate and cell identity are regulated, and large consortia are being formed with the aim to comprehensively chart single-cell landscapes in model organisms[23,24].

To fulfill the promise of these single-cell measurement technologies, it will be crucial that computational methods are available to unambiguously determine what the raw measurements say about the states of individual cells. We not only want to be able to integrate results of scRNA-seq measurements from different labs and protocols but also want to integrate scRNA-seq measurements with results derived from different measurement technologies, such as fluorescence in situ hybridization (for example, see ref. [25]). To make this possible, 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 these concrete physical quantities should follow directly from the experimental data with as small a number of additional assumptions as possible and not depend on arbitrary parameters that the user can set at will. Moreover, to determine when different measurements are mutually consistent, estimates should be accompanied by error bars.

However, although there has been a veritable explosion of scRNA-seq analysis tools in recent years, little attention has been given to satisfying these objectives. Instead of a few methods that estimate quantities with clear physical interpretation in a transparent manner, scientists are faced with a large number of ad hoc methods that apply complex transformations to the data to perform combinations of tasks, including imputation/normalization, clustering, dimensionality reduction, pseudotime and trajectory inference and visualization. These methods often have many tunable parameters, produce outputs in abstract spaces that lack clear biological meaning and are often even stochastic, giving varying outputs when run on the same data with the same parameters. For example, the popular t-SNE[26] and UMAP[27] visualization tools are both stochastic and highly dependent on parameter settings, and position cells in a space whose dimensions lack biological interpretation.

Here, we focus on the 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 no tunable parameters and provides error bars for all of its estimates.

We compare Sanity with a selection of popular methods for imputation/normalization from the recent literature[28–34] (see Methods) and show that only Sanity can effectively remove Poisson sampling fluctuations to infer the true variation in gene expression 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 counts of cells or inferring strong coexpression between large numbers of genes when none is evident in the data. Finally, we show that expression levels estimated by Sanity outcompete those of other methods in downstream analysis tasks, such as finding nearest-neighbor cells and clustering.

Results

Sanity's approach, which is detailed in the Methods, is summarized in Fig. 1. Although it is tempting to simply consider the gene expression state of a cell to correspond to the vector of its mRNA counts, these mRNA counts will exhibit Poisson fluctuations from cell to cell, even if the rates of transcription and mRNA decay are constant across cells and time. We thus argue that changes in expression state should only reflect changes in transcription and decay rates of mRNAs, and should correct for intrinsic noise in gene expression.

[1]Biozentrum, University of Basel, Basel, Switzerland. [2]Swiss Institute of Bioinformatics, Basel, Switzerland. ✉e-mail: erik.vannimwegen@unibas.ch

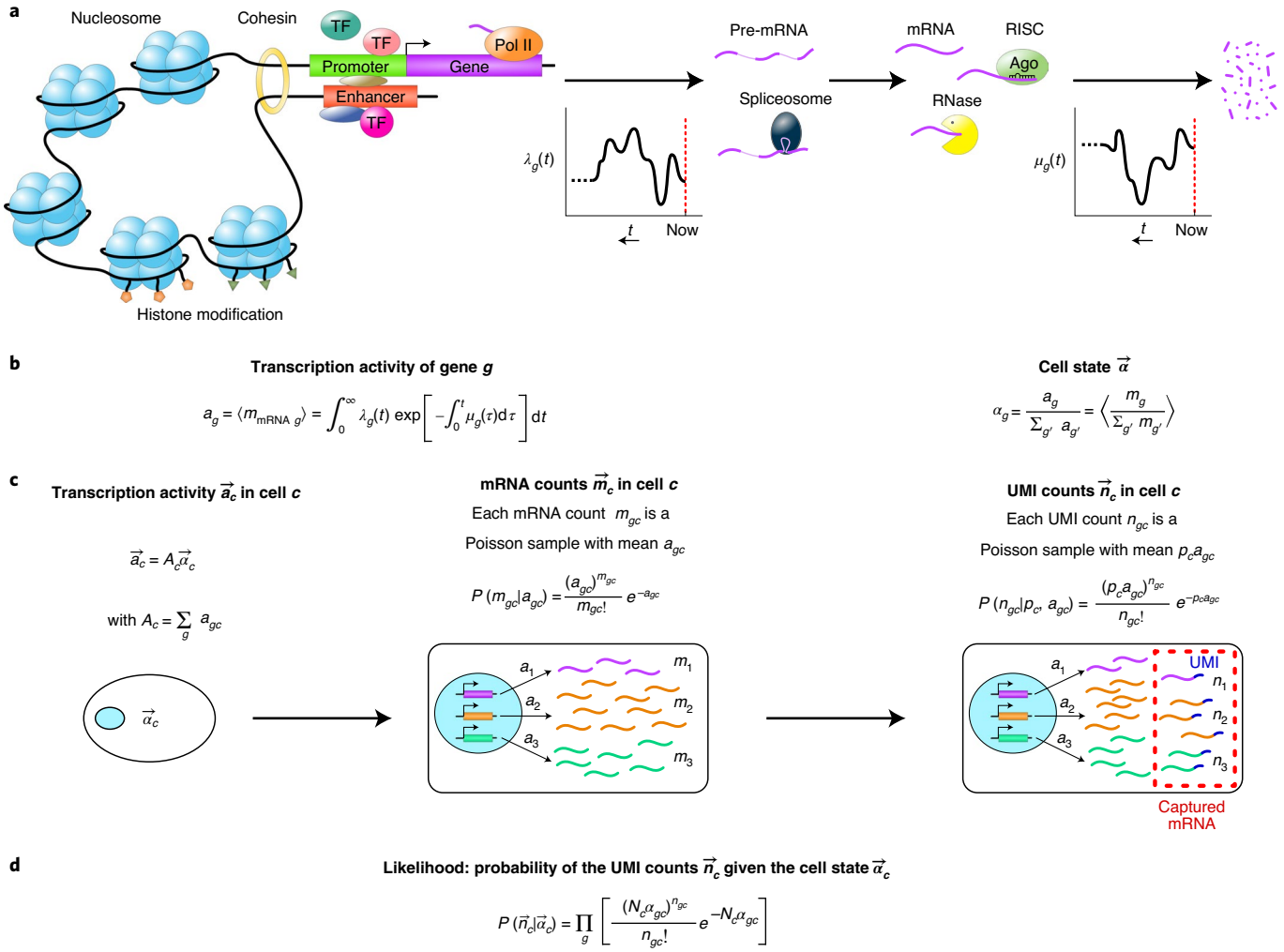

Fig. 1 | Summary of the Sanity approach. a, Cartoon of the flow of causality from the physical state of the cell to gene expression patterns. The concentrations of transcription factors (TFs), chromatin modifiers and other regulatory factors determine changes in chromatin state, three-dimensional (3D) organization of the chromosomes, binding and unbinding of TFs to promoters and enhancers, and so on. These determine the time-dependent rate $\lambda_g(t)$ at which gene g was described a time t in the past. Similarly, the concentrations of microRNAs, RNases and other RNA-binding proteins determine the time-dependent rate $\mu_g(t)$ at which mRNAs of gene g decayed at time t in the past. **b**, The transcription activity a_g of gene g is defined as the expected number of mRNAs and is a weighted average of its transcription and decay rates in the past. We define the expression state of the cell as the vector $\vec{\alpha}$ of relative transcription activities of all genes. **c**, Logical flow from expression state $\vec{\alpha}_c$ to observed UMI counts $\vec{n}_c$. The expression state $\vec{\alpha}_c$ and total transcription activity A_c determine the transcription activities a_{gc}. For each gene g, the probability $P(m_{gc}|a_{gc})$ of having m_{gc} mRNAs is a Poisson distribution with mean a_{gc}. Assuming each mRNA in cell c has a probability p_c of being captured and sequenced, the probability $P(n_{gc}|p_c, a_{gc})$ of obtaining n_{gc} UMIs is a Poisson distribution with mean $p_c a_{gc}$. **d**, The probability of obtaining the UMI counts $\vec{n}_c$ given the cell state $\vec{\alpha}_c$ is a product over genes of Poisson distributions with means $N_c \alpha_{gc}$, where N_c is the total UMI count in cell c.

The crucial insight is that even if transcription and mRNA decay rates vary with time in an arbitrary way in a given cell, the mRNA count m_g of each gene g is still a Poisson sample of a single effective 'transcription activity' a_g, which is a weighted average of its recent transcription and mRNA decay rate in the cell (Fig. 1a,b). Sanity represents the expression state of a cell by a vector of transcription quotients α_g corresponding to these relative transcription activities (Fig. 1b,c). As shown in the Methods and Supplementary Methods, the probability of the raw UMI counts of a cell given its transcription quotients is a product of Poisson distributions (Fig. 1d).

To infer the log transcription quotients (LTQs) of each gene in each cell from the UMI counts, Sanity makes as few prior assumptions as possible about how LTQs might vary across genes and cells. In particular, it only assumes that, for each gene g, the distribution of its LTQs across cells can be characterized by an unknown mean μ_g and variance v_g. Given this, the entire inference procedure follows from first principles, without any tunable parameters.

As detailed in the Methods, we use seven real and two simulated scRNA-seq datasets to compare Sanity's performance with those of two basic normalization methods that simply log transform raw or normalized UMI counts (called RawCounts and TPM, respectively) and seven other recently proposed normalization methods (DCA[28], Deconvolution[29], MAGIC[30], SAVER[31], scImpute[32], sctransform[33] and scVI[34]).

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 noise to quantify the true biological variation in gene expression across cells. Testing this is challenging because the true expression

variability of each gene is generally unknown. To address this issue, we first analyzed a carefully designed study of mouse embryonic stem cells (ESCs) from Grün et al.[35] in which not only were scRNA-seq measurements taken for cells cultured in both 2i and serum conditions, but the same measurement protocol was applied to single-cell equivalent aliquots from pooled RNA. The expression variation in these aliquots thus solely derives from technical sampling noise. In addition, these ESCs are highly homogeneous, so little true expression variation is expected for ESCs in the same condition.

Fig. 2a shows box plots of the distributions of coefficients of variation (CVs) across genes for each of the four datasets, as calculated from the expression estimates of each of the normalization methods (except for sctransform, which does not report estimated expression values). Analogous results using s.d. in LTQ (which is equivalent to the CV when the CV is small; Supplementary Text 1) are shown in Supplementary Fig. 1.

Ideally, the methods should infer that there is no true variability at all for the aliquots and relatively little variability for the ESCs. However, although all methods infer that CVs are slightly larger in cells cultured in serum than in 2i, which is in line with a previous analysis[35], most methods infer substantial variability for most genes. In particular, methods that do not correct for Poisson noise (RawCounts, TPM, Deconvolution and scImpute) infer CVs larger than 0.5 for the large majority of genes in both cells and aliquots. By contrast, the CVs that Sanity infers are at least twofold lower than those of all other methods, and only Sanity correctly infers that there is no expression variability in the aliquots, that is, with CVs less than 10% for almost all genes.

There is no reason to expect that CVs in expression should correlate with mean expression, and, in bulk RNA-seq, there is indeed no correlation between mean log expression and variance in log expression across conditions (Supplementary Fig. 2). However, at the single-cell level, the intrinsic Poisson fluctuations will add a term $1/\sqrt{\text{mean}}$ to the CV, as is well appreciated in the scRNA-seq literature (for example, see[36]). Thus, systematic correlations between CVs and the mean of normalized expression levels reflect to what extent a method has failed to correct for Poisson sampling noise. Fig. 2b shows scatter plots of CVs against mean expression for all methods, and we see that, with the exception of Sanity and MAGIC, all other methods show a strong negative correlation between CV and mean, indicating that Poisson sampling noise dominates the observed variability for all but the most highly expressed genes. These observations apply to all datasets (Supplementary Fig. 3), including the simulated dataset that we discuss next.

To more directly test the accuracy with which different methods estimate the expression variance of each gene, we constructed a simulated dataset for which the true mean and variation in LTQ across cells is known for each gene (see Methods and Supplementary Fig. 4). Comparing the true CVs of genes with those inferred by each method (Supplementary Fig. 5) shows that only Sanity and, to a lesser extent, SAVER exhibit a good correlation between true and inferred CVs. A comparison of true and inferred variances in LTQs confirms this overall picture (Supplementary Fig. 6). Notably, for all methods except Sanity, the Poisson noise causes the inferred CVs of genes with low expression to be systematically higher than their true CVs, resulting in an almost complete loss of correlation between true and inferred CVs across genes for most methods. For genes with very low expression, the expression data are so sparse that it is only possible to estimate an upper bound on expression variability (see Supplementary Text 1), and Sanity conservatively infers that the true expression variability is low so that these genes will not substantially contribute to most downstream analyses.

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

The accuracy of gene expression estimates strongly depends on the depth of coverage. Gene expression measurement noise is expected to scale inversely proportional to absolute expression; that is, for a gene with $\langle n \rangle$ expected UMIs in a cell, the Poisson noise will cause the measured log expression $\log(n)$ of a gene to differ from the true log expression $\log(\langle n \rangle)$ by a term of order $1/\sqrt{n}$. We thus used the same simulated dataset to compare the accuracy of gene expression estimates of the different methods as a function of depth of coverage. In particular, we stratified all genes into bins according to their absolute expression (average number of UMIs per cell) and calculated the accuracy of various expression estimates for each method and each bin (Fig. 3).

First, while most methods accurately estimate mean log expression levels for genes with at least 0.1 UMIs per cell, DCA, scVI, sctransform and scImpute never do (Fig. 3a). Second, although Sanity is essentially the only method that can accurately estimate the true variance in log expression levels across cells, even Sanity can only reliably estimate the true variance in LTQ for genes that have at least 1 UMI per cell on average (Fig. 3b). Third, Pearson correlations between true and estimated log fold changes quantify how accurately each method identifies in which cells a gene has highest and lowest expression (Fig. 3c). We observed that Pearson correlations systematically increase with absolute expression, with Sanity performing best at each expression level, followed closely by TPM, Deconvolution and SAVER. By contrast, the log fold changes predicted by MAGIC, DCA and scVI show almost no correlation with the true log fold changes, even for highly expressed genes, suggesting that these methods systematically distort expression levels. However, even for the best methods, correlations are only consistently high for genes with at least 1 UMI per cell and are consistently low for genes with less than 0.1 UMI per cell.

As discussed in Supplementary Text 1, with current capture efficiencies, the vast majority of genes have less than 1 UMI per cell (Supplementary Fig. 26). As accurate estimates of expression levels are only guaranteed for genes with at least 1 UMI per cell (Fig. 3), this implies accurate estimates of expression patterns for only a few hundred genes. Consequently, if it were possible to substantially raise capture and sequencing efficiencies, then the number of genes for which we would be able to obtain accurate expression estimates could be dramatically increased (Supplementary Fig. 26).

Many normalization methods introduce spurious correlations. Due to variations in cell size, mRNA capture efficiency and sequencing depth, the total number of UMIs can fluctuate significantly from cell to cell. Therefore, most scRNA-seq processing methods normalize expression levels for the total number of mRNAs (that is, UMIs) that were captured from a given cell. The simple TPM procedure does so by dividing the observed counts for each gene by the total UMI count of the cell, and Deconvolution accomplishes the same normalization using a more sophisticated approach. With the exception of RawCounts and scImpute, all other methods normalize for total UMI count.

If the normalization for total UMI count were successful, we would expect no systematic correlation between inferred expression levels and total UMI counts across cells for most genes. However, this is not what we observe. For each method and gene, we calculated the Pearson correlation between the inferred log expression levels and log total UMI counts. Using the Zeisel dataset as an example, Fig. 4a,b shows the distribution of Pearson correlations, as well as raw scatters of the normalized expression levels as a function of log total UMI count for one example gene (*Zbed3*).

As expected, because RawCounts and scImpute do not normalize for total UMI count N_c, most genes show a positive correlation between the inferred expression levels and $\log(N_c)$ with these methods. By contrast, the simple TPM method, Deconvolution and especially Sanity and sctransform successfully remove this

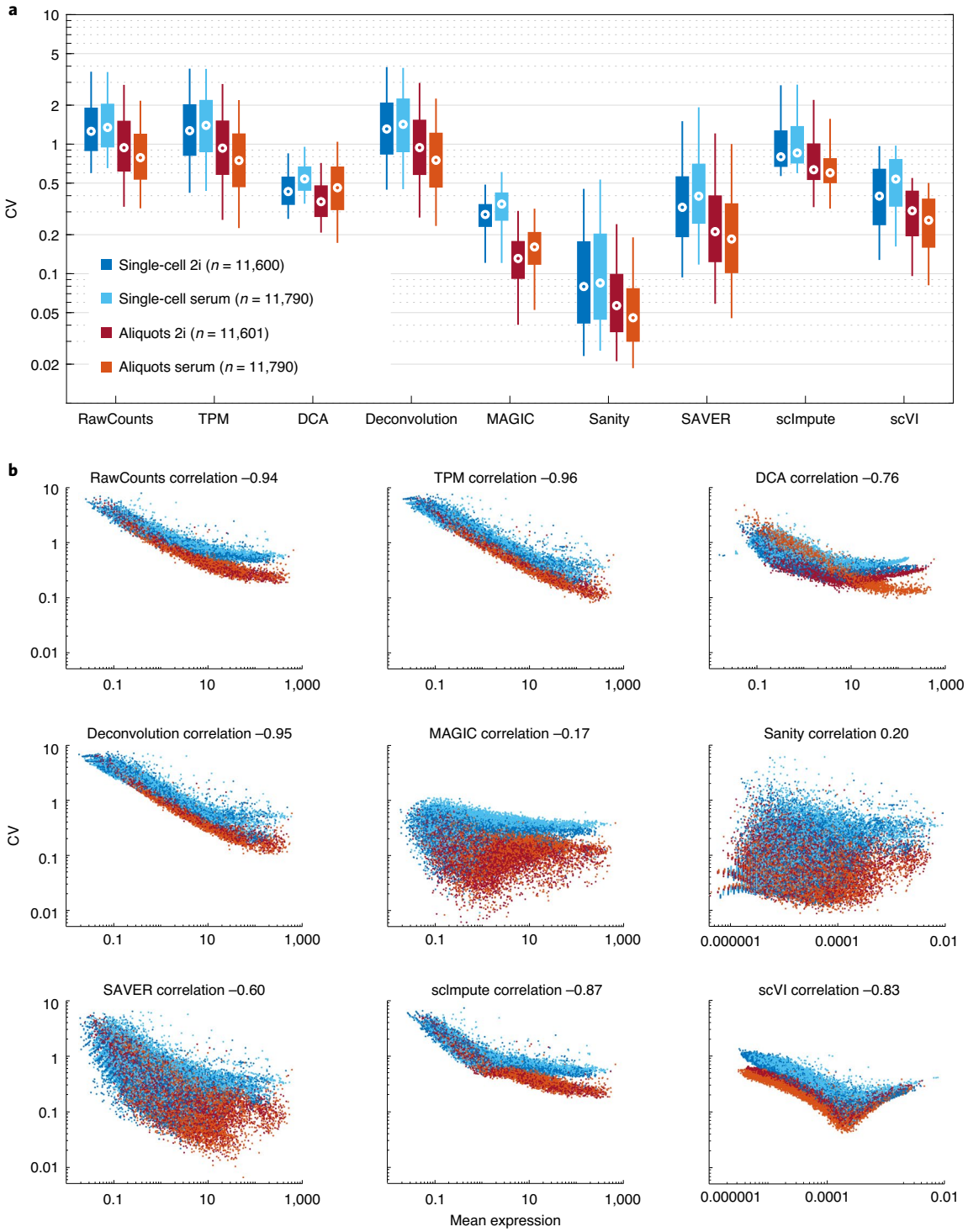

Fig. 2 | Effects of Poisson fluctuations on gene expression variance. a, Box plots showing the median (circle) and the 5th, 25th, 75th and 95th quantiles of the distribution of CVs of gene expression levels across genes for each of the four datasets as inferred by each of the normalization methods. **b**, Scatter plots of CVs (s.d. divided by mean) against mean expression for all genes in each of the four datasets 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. Axes are shown on logarithmic scales and are kept similar across panels except for Sanity and scVI, where the mean expression values are on a very different scale from those of the other methods. Color coding of data sets is as indicated in **a**.

correlation. However, although DCA, SAVER, MAGIC and scVI also intend to normalize for total UMI counts, their normalized expression levels show even stronger correlations with $\log(N_c)$ than the non-normalized RawCounts. Scatters with inferred expression levels for the gene *Zbed3* as a function of $\log(N_c)$ illustrate how dramatically some normalization methods transform the input data. The RawCounts results show that this gene has fairly low expression, with 0 or 1 UMIs observed in most cells, and with a slightly higher

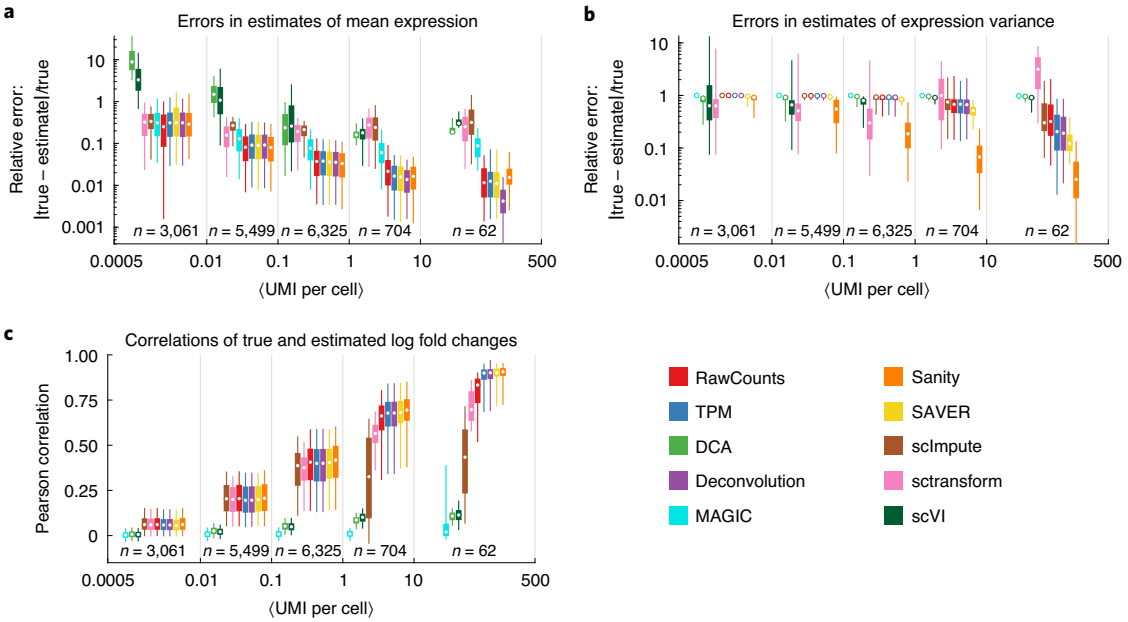

Fig. 3 | Accuracy of gene expression estimates as a function of depth of coverage. a–c, Genes were stratified into five bins of absolute expression (in average number of UMIs per cell), and, for each bin, the distribution of relative errors in estimated mean log expression (**a**), estimated variance in log expression (**b**) and Pearson correlations between true and estimated log fold changes across cells (**c**) were calculated for each method. Distributions are visualized as box plots showing the median, interquartile range, 5th percentile and 95th percentile for the genes in each expression bin. Note that the vertical axes are shown on a logarithmic scale for the top two panels. Methods are sorted from right to left in each panel in approximate order of their accuracy.

chance to observe 1 or 2 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 negatively (DCA, SAVER, scVI) or strongly positively (MAGIC) with total UMI count. These observations again generalize to all other datasets (Supplementary Fig. 7).

In many studies, systematic analysis of coexpression of pairs of genes is used to identify co-regulated pathways or regulatory modules. For such applications, it is thus crucial that the pairwise correlations of the expression profiles accurately reflect the coexpression evidence in the data. To investigate this, we calculated Pearson correlations of the normalized log expression levels of all pairs of genes and then compared these pairwise correlation coefficients across the various methods, using the Baron dataset as an example (Fig. 4c–g). The pairwise correlations by and large agree between Sanity and the simple TPM method (Fig. 4c), and this agreement is also observed for Deconvolution and sctransform (Supplementary Fig. 8). Although Sanity and scImpute also largely agree on which pairs of genes are most strongly positively or negatively correlated (Fig. 4d), scImpute predicts moderate positive correlations for many gene pairs for which Sanity predicts no correlation at all. This behavior results from scImpute not normalizing for total UMI count and is indeed also observed for RawCounts (Supplementary Fig. 8).

A very different pattern is observed for the comparison between Sanity and MAGIC (Fig. 4e). For many of the pairs of genes for which Sanity infers no coexpression (that is, zero correlation), MAGIC infers a broad range of correlations, running from almost perfect anticorrelation to perfect correlation. To further investigate this, we focused on a subset of 4,360 pairs of genes within the red rectangle of Fig. 4e for which MAGIC predicted nearly perfect correlation and Sanity almost none. Summing across all 4,360 pairs of genes and all cells, we found that there was not a single example for which both genes in a pair were observed in the same cell (Fig. 4f). That is, although MAGIC infers that these 4,360 pairs of genes are almost perfectly coexpressed, *none* of them are ever observed to be present at the same time in *any* cell. By contrast, for the small set of pairs for which Sanity infers coexpression whereas MAGIC does not, we do generally find evidence of coexpression (Fig. 4g). This same pattern is observed for the comparisons of Sanity's pairwise correlations with those of DCA, SAVER and scVI (Supplementary Fig. 9). That is, these methods all infer large numbers of highly correlated or anticorrelated pairs of genes, whereas there is no evidence at all of coexpression in the raw counts of these pairs. Consistent with these observations, these methods show very wide distributions of pairwise correlations on each dataset, whereas correlations are highly peaked around zero for Sanity, TPM, Deconvolution and sctransform (Supplementary Fig. 10). Moreover, although our simulated dataset contains no correlations by construction, DCA, MAGIC, scVI and to a lesser extent SAVER also predict a wide range of correlations on these data (Supplementary Fig. 11).

We believe that these pervasive spurious correlations result from the fact that these methods map the expression data to a lower-dimensional manifold. Indeed, if we project the TPM-normalized results from the simulated data on the first n principal component analysis (PCA) components, the amount of spurious correlations systematically increases with decreasing n (Supplementary Fig. 12). Comparison of Supplementary Figs. 11 and 12 shows that the amount of spurious correlations in SAVER's results is equivalent to projecting on the first 100–200 PCs, the first 20–30 PCs for DCA and scVI, and the first 5–10 PCs for MAGIC.

Sanity outperforms other methods on identifying nearest-neighbor cells. Many downstream scRNA-seq analyses, including clustering and trajectory reconstruction, require estimating the distances between cells in gene expression space. In particular, many methods involve identifying the k nearest neighbors of each cell with the most similar expression profiles (with k typically in the range of 3–30). Assessing the accuracy of different methods in identifying nearest-neighbor cells on real data is challenging because it is not known which cells are truly nearest neighbors. We

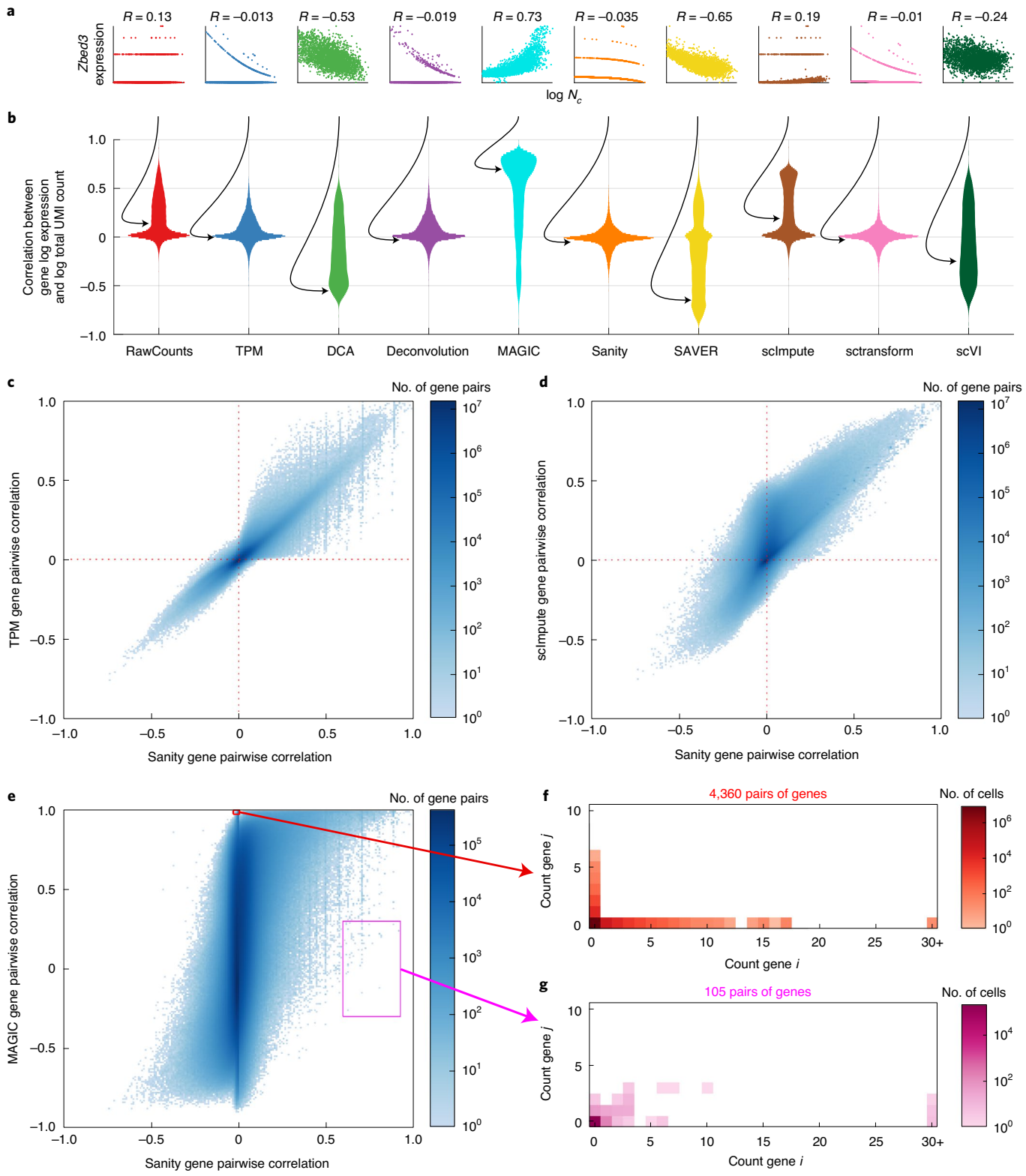

Fig. 4 | Correlations between inferred gene expression levels and library size and between pairs of genes. a, Scatter plots of the normalized log expression $\log(e_c)$ of the example gene *Zbed3* versus the logarithm of the total UMI count $\log(N_c)$ across cells for each method. The Pearson correlation coefficient of the dependence is shown above each panel. **b**, Violin plots of the distribution of correlation coefficients between $\log(e_c)$ and $\log(N_c)$ for all genes for the Zeisel dataset. Each color corresponds to a method, as indicated below each plot. **c**–**e**, Density plots of Pearson correlations for all pairs of genes as inferred by Sanity (*x* axes) against the correlations inferred by TPM, scImpute and MAGIC (*y* axes). The color scale shows the density in $\log_{10}$ number of gene pairs, and zero counts are shown in white. Red and magenta rectangles in **e** indicate the pairs of genes with a correlation above 0.975 for MAGIC and between −0.03 and 0.005 for Sanity (red) and all pairs with a correlation between −0.3 and 0.3 for MAGIC and between 0.6 and 0.93 for Sanity (magenta). **f**, Two-dimensional (2D) histogram of counts per cell summed over the 4,360 pairs of genes from the red rectangle in **e**. The height of the histogram is shown in $\log_{10}$ as a color, and zero counts are shown in white. **g**, Analogous 2D histogram of counts for the 105 pairs of genes from the magenta rectangle in **e**.

thus created a simulated dataset in which cells are distributed along a tree that was constructed by performing a branched random walk through gene expression space, that is, setting the true LTQs of each cell equal to those of the previous cell plus a small random perturbation to the LTQ of each gene (see Supplementary Methods).

For each method, we calculated the Euclidean distances between the normalized log expression vectors of all pairs of cells and determined the k nearest neighbors of each cell. For Sanity, we also estimated cell-to-cell distances using a Bayesian method that incorporates Sanity's error bars, which automatically causes genes with large error bars ϵ_{gc} to contribute less to the distance estimate (Supplementary Methods). For each method, we then calculated the fraction of predicted k nearest neighbors that belong to the set of true k nearest neighbors as a function of k (Fig. 5a).

Sanity clearly performs best in identifying the k nearest neighbors, but when its error bars are ignored, the performance is much reduced, highlighting the value of incorporating error bars. This reduction in performance is due to the noisy estimates of the LTQs of genes with low expression because, if we calculate distances based only on the genes with at least 1 UMI per cell on average, Sanity's performance without error bars is dramatically improved, approaching the performance incorporating error bars for large k (Fig. 5b). Other normalization methods (for example, TPM and sctransform) also perform much better when distances are only estimated from genes with at least 1 UMI per cell. By contrast, the performance of scVI and DCA is not sensitive to excluding low-expression genes, suggesting that, for these methods, the expression levels of low-expression genes are effectively determined by the expression levels of high-expression genes. Notably, whereas DCA and scVI performed poorly on previous tests concerned with the accuracy of inferred gene expression levels, here, they are the best-performing methods after Sanity and also perform well at estimating distances between all pairs of cells (Supplementary Figs. 13 and 14). This shows that these methods are optimized to correctly estimate distances between cells at the expense of severely distorting the expression patterns of individual genes.

To give a visual impression of the accuracy with which different methods are able to capture the local structure in the data, Supplementary Fig. 15 shows t-SNE visualizations of the matrices of true cell-to-cell distances and cell-to-cell distances as estimated by each of the methods. It is notable that, even though the data correspond to a complex tree structure of 149 branches with 13 cells each, Sanity's estimates of the cell-to-cell distances allow for a reasonably accurate reconstruction of this complex structure.

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 cells based on their gene expression patterns. For six of our test datasets, the corresponding study reported an annotation of cell types, which was typically obtained by combining automated clustering with analysis of marker gene expression and hand curation using prior knowledge. Taking the Zeisel dataset as an example[37], Fig. 5c visualizes the clustering structure implied by the different methods by applying the popular t-SNE algorithm[26] to the normalized expression values of each method. Although it is well known that, beyond reasonably conserving which cells are nearest neighbors, it is difficult to interpret these visualizations, the visualization does suggest that there is considerable disparity between normalization methods. In particular, Sanity, TPM and Deconvolution appear to separate the cell types more reliably than MAGIC, RawCounts and scImpute, and similar observations can be made on the other datasets (Supplementary Figs. 16–20).

Rigorously benchmarking the performance of normalization methods on clustering is challenging because the ground truth is again not known. While the provided reference annotations are likely reasonable, it is by no means clear that these annotations are optimal. In addition, clustering performance will also depend on what clustering algorithm is used and even what similarity measure is used to compare clusterings. We thus chose to assess the quality of each normalization method by its performance across all six datasets using three different clustering algorithms (K-means[38], Ward[39] and Louvain[40]) and using four different similarity measures (Supplementary Methods), giving 72 comparisons of similarity scores across methods (Supplementary Fig. 21). To summarize these results, we calculated the number of times each method was the best-performing method (Fig. 5d). In addition, we calculated how close each method comes to the best-performing method across the 72 combinations (Fig. 5e). Sanity clearly outperforms the other methods; it was the best-performing method on more than half of the combinations and scored close to the best-performing method on a large majority of combinations. TPM, Deconvolution, DCA and scVI also perform robustly, typically scoring within 10% of the best method.

As a final example of downstream analysis, we tested the ability of the normalized expression values to identify genes that are significantly upregulated in particular subtypes of cells, as detailed in Supplementary Text 1. Here too, we found that Sanity performed best, although sctransform, TPM and Deconvolution achieved almost equal performance, whereas MAGIC, DCA and scVI typically performed poorly on this task (Supplementary Fig. 22). Supplementary Text 1 provides additional in-depth discussion of Sanity's features and limitations, including its performance for genes with very low or multimodal expression (Supplementary Figs. 23 and 24) and how observed absolute expression distributions (Supplementary Fig. 25) and sequencing depth determine the accuracy of expression estimates across genes (Supplementary Fig. 26).

Discussion

In this work we developed a new normalization procedure for scRNA-seq data from first principles using only two basic assumptions. First, we characterize a cell's gene expression state by the vector of LTQs across genes, that is, the logarithms of the expected fractions of the transcript pool for each gene. Second, to estimate these LTQs from the raw UMI count data, we characterize the prior distribution of LTQs of each gene only by its mean and variance across cells. Given these two assumptions, the entire procedure follows from first principles without any tunable parameters and returns estimated LTQs that correct both for the Poisson noise that is intrinsic to the process of transcription and for the sampling noise of the scRNA-seq measurements. Consequently, variation in the inferred LTQs reflects changes in the rates of transcription and mRNA decay of each gene.

Although our procedure makes only minimal assumptions, one may still 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 rescaling the expression levels of all genes by a common factor does not change expression state, then LTQs naturally follow as the most general representation of a cell's expression state. Similarly, our prior distribution over LTQs of a gene also aims to minimize the strength of our method's assumptions by using the least-assuming (that is, maximum entropy) distribution consistent with a given mean and variance. Improving on these assumptions would require specific biological information to determine more informative priors on the gene expression states that cells can take on.

Our benchmarking tests indicate that Sanity's normalized expression values outperform those of other methods on basic downstream processing tasks, such as clustering cells into subtypes and identifying nearest-neighbor cells. More importantly, we show that all other methods produce a representation of the data that is distorted in one or more respects.

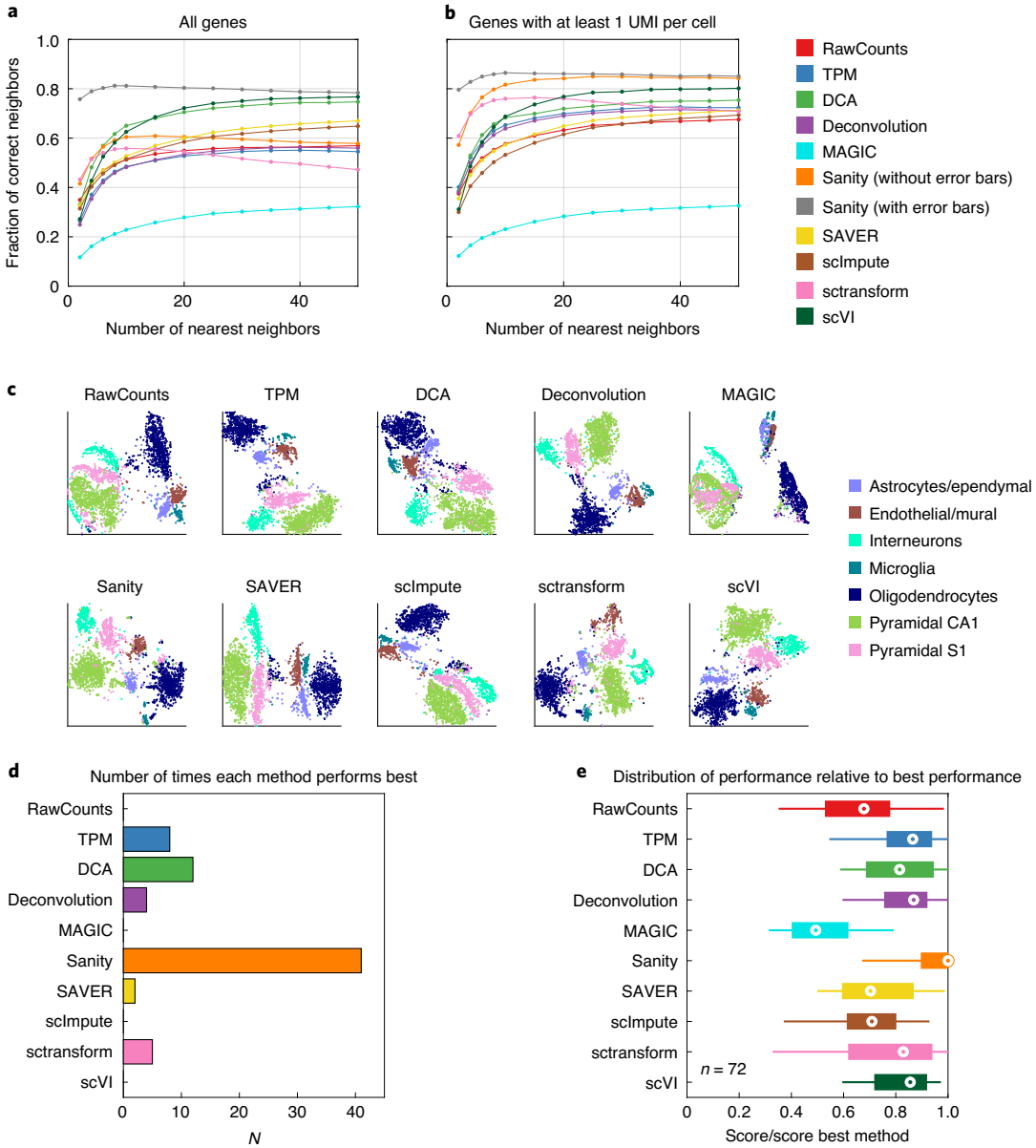

Fig. 5 | Accuracy of the *k* nearest-neighbor and clustering predictions. a, For each method, we calculated the Euclidean distances between the log expression profiles of all pairs of cells to predict the nearest neighbors of each cell, on the simulated dataset for which cells lie along a branched random walk in gene expression space. The curves show the fraction of predicted *k* nearest neighbors for each method that are members of the set of true *k* nearest neighbors as a function of *k*. **b**, The same as in **a**, but calculating distances using only highly expressed genes (at least 1 UMI per cell on average). **c**, 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 ref. [37]. **d**, Similarity scores between annotated and predicted clustering were calculated for each method across 72 combinations of 6 annotated datasets, 3 clustering algorithms and 4 similarity metrics. The bars show, for each method, the number of combinations for which it performed best. **e**, For each method *m*, the distribution (across the 72 combinations) of the ratio $s_m/s_\star$ of its similarity score s_m relative to the similarity score $s_\star$ of the best-performing method (on that combination) is shown as a box plot indicating 5th percentile, first quartile, median, third quartile and 95th percentile.

The simple TPM and closely related Deconvolution methods produce representations of the data that are generally reasonable and perform quite well on downstream tasks, such as clustering and identification of differentially expressed genes. The main problem with the TPM method is that the variation in normalized expression levels is dominated by Poisson fluctuations for most genes and that genes with low expression are predicted to be the most variable, whereas in reality these have the least evidence of true variability. This also causes the TPM and Deconvolution methods to perform poorly in identifying nearest-neighbor cells, although this can be mitigated to some extent by only considering highly expressed genes. The simple RawCounts method and the similarly performing scImpute method suffer from these same problems, and additionally have the problem of not correcting for variation in total UMI count across cells.

The sctransform method outputs *z*-statistics rather than gene expression estimates. Although this has some advantages (for example, the method performs well in identifying differentially expressed genes), the clear drawback is that it cannot accurately predict log fold changes in expression levels and performs quite poorly in identifying nearest-neighbor cells.

The sophisticated scVI and DCA methods that use autoencoders to map the data to a low-dimensional manifold perform well on estimating distances between cells but do this at the cost of strongly distorting the expression levels of individual genes. These methods poorly estimate log fold changes of genes across cells, produce strong artifactual correlations of the normalized expression values with the total UMI count in each cell and spuriously predict large numbers of coexpressed genes. Although SAVER performs better in estimating the variances and log fold changes of genes across cells, it suffers from the same spurious prediction of correlations as does MAGIC, which, in our hands, performed poorly on most tests.

The fact that such spurious correlations are also induced when the TPM-normalized expression values are projected onto the top PCA components suggests that they generically result from fitting the data to a lower-dimensional representation. Although it is reasonable to assume that the space of gene expression states that cells take on has much lower dimensionality than the full dimensionality of the transcriptomic data, the task of finding such lower-dimensional representations should be clearly distinguished from normalization and noise correction. Because Poisson sampling noise scales with absolute expression levels, different genes and cells are affected to different extents, and this may be erroneously mistaken for 'structure' in the data. Thus, unless the process of noise removal and normalization is carefully separated from fitting of the data to lower-dimensional representations, artifactual correlations are likely to be introduced.

Finding biologically meaningful lower-dimensional representations of genome-wide gene expression states is one of the most important challenges in the field. However, it is likely a very hard problem in general, and it is unclear to us whether the problem is even solvable with current data. For example, 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. We believe that rather than black box procedures for dimensionality reduction, progress in understanding the genome-wide structure of expression data will crucially depend on connecting transcriptomic data to the underlying biophysical mechanisms (for example, the dynamics of the chromosome, chromatin accessibility at enhancers and promoters, the binding and unbinding of transcription factors, recruitment of the transcriptional machinery and the mechanisms of transcription initiation). However, whatever approach is taken to finding lower-dimensional representations of gene expression states, a prerequisite is that the raw data are carefully normalized and corrected for both biological and technical sampling noise. The Sanity method that we present here aims to provide such a normalization methodology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00875-x.

Received: 13 January 2020; Accepted: 26 February 2021;
Published online: 29 April 2021

References

1. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
2. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
3. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
5. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
6. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
7. Rotem, A. et al. Single-cell ChIP–seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
8. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
9. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
10. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
11. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
12. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
13. Frei, A. P. et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat. Methods* **13**, 269–275 (2016).
14. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
15. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
16. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
17. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
18. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
19. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
20. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
21. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896 (2016).
22. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
23. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
24. Rajewsky, N. et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386 (2020).
25. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
26. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
27. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* http://arxiv.org/abs/1802.03426 (2018).
28. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
29. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
30. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
31. Huang, M. et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).
32. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
33. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
34. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
35. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
36. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
37. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
38. Lloyd, S. P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
39. Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
40. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

A Bayesian method for inferring gene expression states from count data. After motivating how we represent gene expression states of single cells and to what concrete physical quantities these gene expression states correspond, we introduce our probabilistic model of an scRNA-seq experiment and calculate the probabilities of the observed raw transcript counts as a function of each cell's expression state. We then explain the Bayesian procedure by which the gene expression states are inferred from the sequencing data and the outputs that the method provides. Additional discussion of the properties and limitations of Sanity's model are provided in Supplementary Text 1, including a discussion of how Sanity can be used to correct for technical batch effects.

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. These gene expression states should satisfy two basic desiderata. First, gene expression states should have a concrete physical interpretation. Second, for each gene g, the difference $e_{gc} - e_{gc'}$ should meaningfully reflect the change in its expression between cells c and c'.

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 expression state of the cell. However, the gene expression process is inherently stochastic due to thermal noise and the low molecule numbers involved; for example, there are only 1 to 2 copies of each promoter in a given cell, causing mRNA counts to fluctuate even between cells that are in the same state. To illustrate this, let us imagine a gene that is transcribed at a constant rate λ and whose mRNAs decay at a constant rate μ in every cell. This is the closest that one can come to having no variation in expression state across cells. However, even in this case, the actual number of mRNAs m for this gene will fluctuate across cells according to a Poisson distribution with mean $a=\lambda/\mu$. That is, the probability of finding m mRNAs is $P_m = a^m e^{-a}/m!$, which has mean $\langle m\rangle = a$ and variance $\mathrm{var}(m) = a$. Thus, instead of interpreting any change in mRNA number m as a change in gene expression state, it makes more sense to identify changes in gene expression state with changes in the transcription and decay rates λ and μ, respectively.

In general, for a given gene g in a given cell, the transcription rate λ_g and decay rate μ_g of its mRNAs will vary with time t in a potentially complex manner. As illustrated in Fig. 1a, a large array of different biophysical processes can affect the transcription rate of a given gene, including changes in the chromatin state around its locus, the binding and unbinding of TFs to promoters and enhancers, changes in the 3D organization of the chromosome and so on. Together, these processes will determine a time-dependent transcription rate $\lambda_g(t)$. Similarly, the rate $\mu_g(t)$ at which mRNAs for gene g decay will depend on the concentrations of RNases, microRNAs, various RNA-binding proteins and so on. If, at some point in time, cell c is sampled and its mRNAs are extracted, then the number of mRNAs m_{gc} that one finds for gene g will depend on what the transcription rate $\lambda_g(t)$ and decay rate $\mu_g(t)$ were in the recent past of this cell.

In particular, if we denote the time point at which the cell is sampled as $t=0$ and denote by $\lambda_g(t)$ and $\mu_g(t)$ the transcription and decay rates at a time t in the past of the cell, then the expected number of mRNAs $\langle m_{gc}\rangle$ is given by

$$\langle m_{gc}\rangle = \int_0^\infty \lambda_{gc}(t)\exp\left[-\int_0^t \mu_{gc}(\tau)\mathrm{d}\tau\right]\mathrm{d}t \equiv a_{gc} \qquad (1)$$

We call a_{gc} the transcription activity of gene g in cell c (Fig. 1b). Note that a_{gc} is a weighted average of the transcription rates in the past of the cell, where the weights correspond to the probability that an mRNA that was described a time t in the past has survived until now.

Crucially, independent of how $\lambda_{gc}(t)$ and $\mu_{gc}(t)$ have fluctuated in time, the probability of seeing m_{gc} mRNAs for gene g in cell c is still given by a Poisson distribution with mean a_{gc} (ref. [41]) (Fig. 1c); that is:

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

Thus, independent of how $\lambda_{gc}(t)$ and $\mu_{gc}(t)$ have fluctuated in the cell's past, the number of mRNAs m_{gc} depends on these rates only through transcription activity a_{gc}. Vice versa, all information about the time-dependent rates $\lambda_g(t)$ and $\mu_g(t)$ that is contained in measurements of mRNA counts in cell c is contained in the transcription activities a_{gc} for each gene. Thus, we propose to characterize the expression state of a cell by the vector $\vec{a}_c$ of its transcription activities. Note that, as discussed in Supplementary Text 1, it is, in principle, possible to learn more about the functions $\lambda_{gc}(t)$ and $\mu_{gc}(t)$ by also incorporating information from intronic UMIs of each gene g (for example, as done in the RNA velocity approach[42,43]). Although this is an interesting direction for future extensions of Sanity, here, we do not yet incorporate information from intronic UMIs.

Next, we propose that, rather than directly representing the gene expression state of the cell by the vector $\vec{a}_c$ of absolute transcription activities a_{gc}, it is beneficial to use the vector $\vec{\alpha}_c$ of relative transcription activities, defined as

$$\alpha_{gc} = \frac{a_{gc}}{\sum_{g'} a_{g'c}}, \qquad (3)$$

which we will refer to as transcription quotients and which correspond to the expected proportions of mRNAs in the cell (Fig. 1b). First, it has been shown that as cell volume increases, cells globally upregulate transcription to maintain approximately constant mRNA concentrations[44] 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 scale factor S, that is, $a_{gc} \to S a_{gc}$ 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[35,45] so that there is only a weak quantitative relationship between the total number of sequenced mRNA molecules and the true total mRNA content of a cell. Although it is possible to estimate capture and sequencing efficiencies, at least to some extent, using RNA spike-in controls[35,36], 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 α_{gc} can generally be much more accurately estimated than absolute transcription activities a_{gc} because they do not directly depend on capture efficiency. Note that quantifying gene expression by quotients (that is, 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 the 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[46] that, as would be expected from the multiplicative effects of fluctuations in rates of various biochemical reactions[47], the relative expression levels of genes in a sample follow a roughly log-normal distribution that covers several orders of magnitude, and the variance in absolute expression of a gene across conditions scales with the square of its means expression (Supplementary Fig. 2). Consequently, if we were to quantify expression changes directly by the changes $\alpha_{gc} - \alpha_{gc'}$, the expression changes between two cells would be dominated by those of the most highly expressed genes. Therefore, it has long become standard to instead use logarithms of expression levels. Indeed, in bulk RNA-seq experiments, one also generally finds that the variance in log expression of a gene across conditions is uncorrelated with its mean expression (Supplementary Fig. 2).

Thus, we propose to quantify the gene expression state of a cell by LTQs $\log(\alpha_{gc})$ so that an 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}.

To define an overall change in expression state between two cells, we still have to combine the changes in LTQ of all genes into a total 'distance'. As motivated in more detail in Supplementary Text 1, we will follow the generally accepted practice of calculating simple Euclidean distances in the space of LTQ vectors, that is, the squared distance $d^2_{cc'}$ between a pair of cells c and c' is defined as

$$d^2_{cc'} = \sum_g \left[\log(\alpha_{gc}) - \log(\alpha_{gc'})\right]^2 \qquad (4)$$

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, identification of barcodes to identify the library, the individual cell and 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 is little variability in the results from commonly used tools (for example, see[48–51]).

The introduction of UMIs[52] was an important development in scRNA-seq technology in that it avoids PCR amplification noise and allows for the determination of the number of unique mRNA molecules that were captured for each gene. It is currently unclear how to realistically model the noise statistics of protocols that do not incorporate UMIs, and we will here focus on scRNA-seq protocols that use UMIs.

After basic processing of the raw sequences, 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 an scRNA-seq experiment, each mRNA molecule in a given cell c has the same probability p_c to be captured and sequenced. This capture probability varies from cell to cell and has been estimated to be in the range of 10% to 15% (ref. [53]) and up to 30% with most recent protocols[54]. Under this assumption, the probability of the observed UMI counts n_{gc} in cell c given the transcription quotients α_{gc} is given by a product of Poisson distributions (Fig. 1c and Supplementary Methods). Finally, if we marginalize over the unknown capture efficiency p_c, we obtain (see Supplementary Methods)

$$P(\vec{n}_c|\vec{\alpha}_c) = \prod_g \left[\frac{(N_c\alpha_{gc})^{n_{gc}}}{n_{gc}!}e^{-N_c\alpha_{gc}}\right] \qquad (5)$$

where $\vec{n}_c$ is the vector of UMI counts in cell c, $\vec{\alpha}_c$ is the vector of transcription quotients in cell c and N_c is the total number of UMIs in cell c (Fig. 1d). Crucially, 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} (Supplementary Methods).

Prior probabilities and the Bayesian solution. Having argued that we want to characterize each cell's gene expression state by the vector of LTQs $\log(\alpha_{gc})$ and having determined how likely it is to observe UMI counts $\vec{n}_c$ given the transcription quotients $\vec{\alpha}_c$ (that is, equation (5)), we now want to invert this relation and estimate the LTQs from the observed UMI counts. The uniquely consistent set of mathematical procedures for doing this is generally referred to as Bayesian probability theory[55].

This calculation requires that we specify a prior probability distribution that represents the prior information we want to assume about how LTQs may vary across cells before obtaining expression data. As we aim to minimize the number of assumptions, our model will not assume any dependence structure between the LTQs of different genes, that is, we will not assume a priori that the gene expression data derive from a low-dimensional manifold. We will also not assume that the LTQs of a gene across cells follow a particular distribution. The only thing that 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.

Without loss of generality, we rewrite the transcription quotients α_{gc} in terms of an average log quotient μ_g and cell-specific log fold changes δ_{gc}, that is, $\alpha_{gc} = e^{\mu_g+\delta_{gc}}$. With this reparametrization, the δ_{gc} values 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, this implies that the prior corresponds to the maximum entropy distribution consistent with this constraint, which is a Gaussian distribution[55]. Importantly, this does not mean that we assume that the log fold changes δ_{gc} follow a Gaussian distribution. Indeed, as we demonstrate in Supplementary Text 1, the δ_{gc} values that our method infers upon seeing the data do not necessarily follow Gaussian distributions. For example, if a gene is bimodally distributed, the method correctly infers this in general (Supplementary Fig. 23).

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

1. The mean LTQ μ_g and its error bar $\delta\mu_g$
2. The variance v_g of the changes in LTQs δ_{gc} across cells
3. For each cell c, the log fold changes δ^*_{gc} and an error bar ϵ_{gc} on each of these.

Note that the δ^*_{gc} values 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 scRNA-seq measurements.

Other methods for scRNA-seq normalization. To assess the performance of Sanity, we compared it to a number of other methods for normalization/imputation from scRNA-seq data. Here we introduce these other methods and highlight the ways in which their approaches differ from Sanity's. Apart from tools from the recent literature, we 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 to avoid problems with zero counts $n_{gc}=0$, that is,

$$e_{gc} = \log(n_{gc} + p) \tag{6}$$

A typical choice for the pseudocount is $p=1$ because it attenuates fluctuations in n_{gc} on the order of magnitude corresponding to the resolution of the experimental measurements. We refer to this normalization with $p=1$ as the RawCounts normalization because it essentially just log transforms the raw UMI 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 differences in cell size. Consequently, the RawCounts procedure introduces systematic correlations between the expression levels e_{gc} and the total number of UMIs N_c that were sequenced from cell c. Thus, the most commonly used normalization approach is to first divide the raw UMI counts n_{gc} by the total count N_c and then multiply by a typical total count N before adding a pseudocount and log transforming, that is,

$$e_{gc} = \log\left[\frac{n_{gc}}{N_c}N + 1\right] \tag{7}$$

Here, we 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 $N=10^6$).

Given the definition of the LTQs as logarithms of relative expression levels, a reader may wonder how our approach is different even from this standard TPM procedure. Indeed, for a cell with total count N_c and LTQs $\log(\alpha_{gc}) = \mu_g + \delta_{gc}$, the expected number of UMIs is

$$\langle n_{gc}\rangle = N_c e^{\mu_g+\delta_{gc}} \tag{8}$$

This might suggest that if we simply divide n_{gc} by N_c and log transform the result, we would end up with the LTQ $\mu_g+\delta_{gc}$. However, the actual UMI counts n_{gc} are not the same as the expectations $\langle n_{gc}\rangle$. That is, the n_{gc} are measured quantities that contain Poisson noise due to both the intrinsic stochasticity of gene expression and the measurement process. Importantly, instead of n_{gc} differing from $\langle n_{gc}\rangle$ by noise of a constant size, the size of the Poisson noise depends on the expected count $\langle n_{gc}\rangle$ itself. In addition, because UMI counts n_{gc} are very small for most genes, the noise is typically larger than the true variation in LTQ across cells. Therefore, to estimate the LTQ of each gene in each cell, it is crucial to account for this Poisson noise, and this is one of Sanity's main aims.

Beyond the simple RawCount and TPM normalization methods, we compare Sanity's performance with those of the following recently published tools:

1. DCA[28], which uses an autoencoder based on deep learning together with a zero-inflated negative binomial noise model
2. Deconvolution[29], which is similar to the TPM method but uses a more sophisticated approach to normalize for the variation in sequencing depth across cells
3. MAGIC[30], which uses diffusion of measured gene expression states between cells with similar expression profiles
4. SAVER[31], which assumes negative binomial count distributions n_{gc} and models the underlying rates using Poisson LASSO regression with the expression levels of other genes
5. scImpute[32], which focuses mainly on correcting 'dropouts', that is, data points for which $n_{gc}=0$
6. sctransform[33], which uses regularized negative binomial regression and reports Pearson residuals of this regression rather than estimated expression values
7. scVI[34], which uses an autoencoder based on deep neural networks together with a zero-inflated negative binomial noise model

Note that, with the exception of RawCounts and scImpute, all these methods seek to normalize the expression levels for the total UMI count per cell. In contrast to Sanity, RawCounts, TPM and Deconvolution, all other methods seek to remove noise by fitting the data to lower-dimensional representations. Specifically, in SAVER and sctransform, the parameters of each gene's negative binomial model are fitted by using information from other genes; in scImpute, zero values are corrected for by using information from neighboring cells; in MAGIC, the entire expression profile of each cell is estimated using information of neighboring cells and in DCA and scVI the autoencoders effectively force a lower-dimensional representation of the distribution of cells in gene expression space.

Many of the models above use a negative binomial or zero-inflated negative binomial to model the distribution of UMI counts of a gene across cells, and the reader may wonder how Sanity's noise model relates to these models. In Supplementary Text 1, we explain why, as discussed recently[56], no zero inflation is necessary and discuss the relationship of Sanity's model with negative binomial noise models.

We used default parameters for all methods except for scVI, where we adapted settings based on direct feedback from the scVI developers (the default parameter n_epochs=20 was increased to 400, and we used the recently added get_sample_scale instead of the imputation method to get predicted expression values).

Because all methods report expression values in linear space, we log transformed all expression values. MAGIC sometimes reports 0 or even negative values and, as suggested by its developers, we first set all negative values to 0 and then added a pseudocount of 1 to all expression values (including the non-zero ones) before log transforming. Similarly, scImpute reports some zero values, and we added a pseudocount of 1 to all the expression values.

Directly comparing the results of sctransform with those of the other methods is complicated by the fact that, in contrast to all other methods, sctransform does not provide estimated gene expression values but z-statistics z_{gc} that quantify how significantly the expression of gene g in cell c deviates from what would be expected from the negative binomial model. The authors of the sctransform paper suggest that these z-statistics should be used for downstream analyses. Because the z-statistics are variance normalized and centered around zero, we use the z-statistic z_{gc} equivalently to the log fold changes δ_{gc}. Finally, in the negative binomial fit, sctransform fits the expected mean log expression μ_{gc} of gene g in cell c to a function of the form $\mu_{gc} = \beta_0 + \beta_1\log(N_c)$, with N_c representing the total UMI count of cell c. To calculate a predicted average expression for gene g, we use $\mu_g = \beta_0 + \beta_1\log(N)$, with N representing the median total UMI count.

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

1. Grün: 160 mouse ESCs 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 RNA from the cells and splitting the pool into single-cell mRNA equivalents[35]
2. Zeisel: 3,005 cells from the somatosensory cortex and from the CA1 region of the mouse hippocampus annotated into seven cell types[37]
3. Baron: 1,937 human pancreatic cells annotated into 14 cell types[57]
4. Chen: 14,437 adult mouse hypothalamus cells annotated into 15 clusters[58]
5. Three datasets from LaManno[59]:
   (a) LaManno/Embryo: 1,977 ventral midbrain cells from human embryo annotated into 25 classes
   (b) LaManno/ES: 1,715 human ESCs annotated into 17 classes
   (c) LaManno/MouseEmbryo: 1,907 ventral midbrain cells from mouse embryos annotated into 26 classes

In addition to these real datasets, we also constructed two simulated datasets, as detailed in the Supplementary Methods. The distributions of means and variances in log expression as well as the distribution of total UMI count per cell were chosen so as to mimic the statistics of an arbitrarily chosen real dataset, for which we chose the Baron dataset (see Supplementary Fig. 4). In the first simulated dataset, the expression profiles of all genes were drawn randomly and independently so that there were no expression correlations by construction. We used this dataset to test the ability of different methods to correctly estimate true means, variances and log fold changes in the expression of each gene and to assess the extent to which different methods spuriously predicted coexpression of genes. The second simulated dataset was constructed by performing a branched random walk in the high-dimensional gene expression space so that the true expression profiles of the cells fall on a tree. We used this dataset to test the ability of different methods to identify the k nearest-neighbor cells of each cell.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw UMI count tables for each of the scRNA-seq datasets as well as all normalized expression values as inferred by each of the methods are freely available from https://doi.org/10.5281/zenodo.4009187.

Code availability

Sanity was implemented in C and is freely available for download at https://github.com/jmbreda/Sanity. Besides Sanity itself, we also provide code for estimating pairwise distances between cells. In addition, at the same GitHub site, we provide a collection of scripts and supplementary files that should allow other researchers to reproduce the results presented in this publication.

References

41. Thattai, M. Universal Poisson statistics of mRNAs with complex decay pathways. *Biophys. J.* **110**, 301–305 (2016).
42. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
43. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
44. Padovan-Merhar, O. et al. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
45. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
46. Hoyle, D. C., Rattray, M., Jupp, R. & Brass, A. Making sense of microarray data distributions. *Bioinformatics* **18**, 576–584 (2002).
47. Beal, J. Biochemical complexity drives log-normal variation in genetic expression. *Eng. Biol.* **1**, 55–60 (2017).
48. Love, M. I., Anders, S., Kim, V. & Huber, W. RNA-seq workflow: gene-level exploratory analysis and differential expression. *F1000Res* **4**, 1070 (2015).
49. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
50. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
51. Cell Ranger DNA. https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/what-is-cell-ranger-dna
52. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
53. AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An introduction to the analysis of single-cell RNA-sequencing data. *Mol. Ther. Methods Clin. Dev.* **10**, 189–196 (2018).
54. 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).
55. Jaynes, E. T. *Probability Theory: The Logic of Science* (Cambridge Univ. Press, 2003).
56. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).
57. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
58. Chen, R., Wu, X., Jiang, L. & Zhang, Y. Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep.* **18**, 3227–3241 (2017).
59. La Manno, G. et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).

Acknowledgements

This work was supported by the Swiss National Science Foundation, grant 310030_184937. Calculations were performed at sciCORE (http://scicore.unibas.ch/), the scientific computing core facility of the University of Basel.

Author contributions

E.v.N. developed the theoretical formalism. J.B. and E.v.N. developed the implementation and designed the benchmarking. J.B. performed all computations, analyses and simulations. J.B., M.Z. and E.v.N. interpreted the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00875-x.

Correspondence and requests for materials should be addressed to E.v.N.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Erik van Nimwegen

Last updated by author(s): Aug 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	Custom code, developed for this study, is available from Github at https://github.com/ jmbreda/Sanity.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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 GitHub page of the algorithm presented in the manuscript (https://github.com/jmbreda/Sanity) and are accessible on Zenodo (https://doi. org/10.5281/zenodo.3996271)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	7 datasets were used containing different number of cells. Grün 320 cells, Zeisel : 3005 cells, Baron: 1937 cells, Chen: 14437, LaManno/Embryo: 1977 cells, LaManno/ES : 1715 cells, LaManno/MouseEmbryo: 1907 cells.
Data exclusions	No data were excluded
Replication	n/a
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging