Gene expression recovery for single cell RNA sequencing

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang*

* Correspondence:
Nancy R. Zhang
nzh@wharton.upenn.edu
(215) 898-8007
Department of Statistics
The Wharton School
University of Pennsylvania

Abstract

Rapid advances in massively parallel single cell RNA sequencing (scRNA-seq) is paving the way for high-resolution single cell profiling of biological samples. In most scRNA-seq studies, only a small fraction of the transcripts present in each cell are sequenced. The efficiency, that is, the proportion of transcripts in the cell that are sequenced, can be especially low in highly parallelized experiments where the number of reads allocated for each cell is small. This leads to unreliable quantification of lowly and moderately expressed genes, resulting in extremely sparse data and hindering downstream analysis. To address this challenge, we introduce SAVER (Single-cell Analysis Via Expression Recovery), an expression recovery method for scRNA-seq that borrows information across genes and cells to impute the zeros as well as to improve the expression estimates for all genes. We show, by comparison to RNA fluorescence in situ hybridization (FISH) and by data down-sampling experiments, that SAVER reliably recovers cell-specific gene expression concentrations, cross-cell gene expression distributions, and gene-to-gene and cell-to-cell correlations. This improves the power and accuracy of any downstream analysis involving genes with low to moderate expression.

Introduction

A primary challenge in the analysis of scRNA-seq is the low efficiency affecting each cell, which leads to a large proportion of genes, often exceeding 90%, with zero or low count. Although many of the observed zero counts reflect true zero expression, a considerable fraction is due to technical factors such as capture and sequencing efficiency. The overall efficiency of scRNA-seq protocols can vary between <1% to >60% across cells, depending on the method used[1].

Existing studies have adopted varying approaches to mitigate the noise caused by low efficiency. Low-abundance genes and low-coverage cells are commonly removed prior to downstream analysis. This is not ideal, as low abundance genes may be of biological importance and stringent filtering of cells exacerbates the biased sampling of the original cell population. In differential expression and cell type classification, transcripts expressed in a cell but not detected due to technical limitations, also known as dropouts, are sometimes accounted for by a zero-inflated model[2–4]. Other methods try to impute the zeros using bulk RNA-seq data[5] or through gene-pair relationships[6]. However, the zero-inflation models do not explicitly recover

low-abundance genes. Imputation methods focus on imputing genes with zero counts, but ignore genes with low counts, which are also unreliably measured. Imputation based on bulk RNA-seq data fail to capture the cell-to-cell stochasticity in gene expression, which has been shown to lead to large variations in true expression, even across cells of the same type[7,8] or of the same cell line[9,10].

The observed variation in scRNA-seq data is due to both biological and technical factors. Biologically, cells vary in type, size, and expression program. Technically, cDNA library construction from the low amount of RNA in each cell inevitably leads to the loss of some transcripts, the amplification of this library introduces more random variation, and the ensuing sequencing step can lead to further loss if sequencing depth is low. Thus, the observed read counts are a poor representation of the true expression.

Here, we propose SAVER to recover the true expression level of each gene in each cell, removing technical variation while retaining biological variation across cells. SAVER starts with gene count data obtained from UMI-based experiments and computes, for each gene in each cell, an estimate of the true expression as well as a posterior distribution quantifying the uncertainty in this estimate. We demonstrate through down-sampling experiments that the observed expression is distorted by efficiency loss, but that the true expression profiles can be recovered using SAVER. We then evaluate the performance of SAVER through comparisons of Drop-seq and RNA FISH on a melanoma cell line[11]. Finally, we apply SAVER to an embryonic stem cell (ESC) differentiation study[12] to recover known gene-to-gene relationships.

Results

Overview of SAVER model and interpretation

SAVER is based on adaptive shrinkage to a multi-gene prediction model (Fig. 1A). Let Y_{cg} denote the observed UMI count of gene g on cell c. We model Y_{cg} as

$$Y_{cg} \sim Poisson(s_c \lambda_{cg}),$$

where λ_{cg} is the true expression level of gene g in cell c, and s_c is a cell-specific size factor which will be described below. The Poisson distribution after controlling for biological variation between cells has been shown previously by bulk-RNA splitting experiments to be a reasonable approximation for observed gene counts[12,13]. Our goal is to recover λ_{cg} with the help of a prediction μ_{cg} based on the observed expression of a set of informative genes S_g in the same cell (Methods). The accuracy of μ_{cg} in predicting λ_{cg} differs across genes — genes that play central roles in pathways are easier to predict, whereas genes that are not coordinated with other genes are harder to predict. To account for prediction uncertainty, we assume for λ_{cg} a gamma prior with mean set to the prediction μ_{cg} and with dispersion parameter ϕ_g. The dispersion parameter quantifies how well the expression level of gene g is predicted by μ_{cg}. After maximum-likelihood estimation of ϕ_g and reparameterization, let $\hat{\alpha}_{cg}$ and $\hat{\beta}_{cg}$ be the estimated shape and rate parameters, respectively, for the prior gamma distribution. Then, the

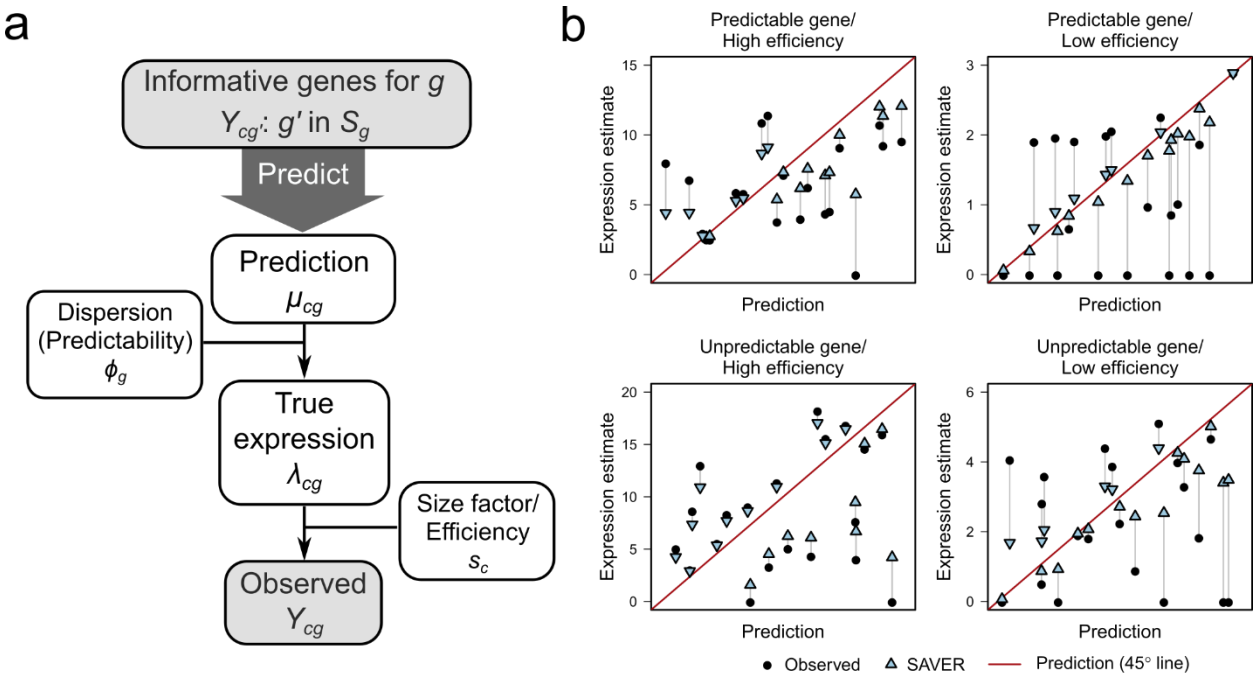

Figure 1 Gene expression model and SAVER recovery. (**a**) A prediction for gene g's true expression level is formed using the expression levels of a set of informative genes $\{Y_{cg'}: g' \text{ in } S_g\}$ in the cell. Our prior belief about the gene's true expression is centered on this prediction and given a prior distribution with dispersion parameter ϕ_g to quantify the gene's prediction uncertainty. The observed expression is derived from the true expression by Poisson sampling with size factor (or efficiency, if known) s_c. (**b**) The SAVER estimate is a weighted average of the normalized observed expression and the predicted expression. The weight is dependent on the predictability of the gene and the cell-specific efficiency. Four scenarios are shown: Predictable (low ϕ_g) versus unpredictable (high ϕ_g) gene, in a high or low efficiency experiment. In each of the scatter plots, each point is a gene, and for each gene, the vertical lines connect the normalized observed expression with the gene's SAVER recovered value, which always lies between the normalized observed expression and the prediction (the 45 degree line).

posterior distribution of λ_{cg} is also gamma distributed with shape parameter $Y_{cg} + \hat{\alpha}_{cg}$ and rate parameter $s_c + \hat{\beta}_{cg}$. The SAVER recovered gene expression is the posterior mean,

$$\hat{\lambda}_{cg} = \frac{s_c}{s_c + \hat{\beta}_{cg}} \cdot \frac{Y_{cg}}{s_c} + \frac{\hat{\beta}_{cg}}{s_c + \hat{\beta}_{cg}} \cdot \mu_{cg}.$$

As seen from the above equation, the recovered expression $\hat{\lambda}_{cg}$ is a weighted average of the normalized observed counts Y_{cg}/s_c and the prediction μ_{cg}. The weights are a function of the size factor s_c and, through the $\hat{\beta}_{cg}$ term , the gene's predictability $\hat{\phi}_g$ and its prediction μ_{cg}. Genes for which the prediction is more trustworthy (small $\hat{\phi}_g$) have larger weight on the prediction μ_{cg}. Genes with higher expression have larger weight on the observed counts and rely less on the prediction. Cells with higher coverage have more reliable observed counts and also rely less on the prediction. Fig. 1B shows example scenarios.

Interpretation of λ_{cg} depends on how the size factor s_c is defined and computed. There are two scenarios. In what is perhaps the simpler scenario, assume that the efficiency loss, that is, the proportion of original transcripts that are sequenced and observed, is known or can be estimated through external spike-ins. If s_c were defined as the cell-specific efficiency loss, then

λ_{cg} would represent the absolute count of gene g in cell c. The second scenario assumes that the efficiency loss is not known, in which case s_c can be set to a normalization factor such as library size or scRNA-seq normalization factors[14,15]. In this case, λ_{cg} represents a gene concentration or relative expression. Which scenario applies depends on the objective of the study and the availability and quality of spike-ins.

It is important to note that SAVER outputs the posterior distribution for λ_{cg}, not just its posterior mean. The posterior distribution quantifies the uncertainty in our estimate of λ_{cg}, and it is crucial to incorporate this uncertainty in downstream analyses. We demonstrate the use of this posterior distribution in two ways. First, to recover the cross-cell expression distribution of a given gene, we sample from the posterior of λ_{cg} for each cell instead of simply using the posterior mean $\hat{\lambda}_{cg}$. Second, in estimating gene-to-gene correlations, the sample correlation of the recovered estimates $\hat{\lambda}_{cg}$ tends to overestimate the true values. We developed an adjusted measure of correlation which takes into account the uncertainty in λ_{cg} (see Methods).

SAVER recovers cell-specific gene expression values

We start with a data down-sampling experiment using the mouse brain scRNA-seq data from Zeisel et al.[16], where we selected a subset of 3,529 highly expressed genes and 1,799 high coverage cells to create a high quality reference dataset. We treat this dataset as the proxy for true expression λ_{cg}. Then, we down-sampled gene counts from this reference dataset at three mean efficiency levels — 25%, 10%, and 5% — to create three observed datasets (Fig. 2A). The reference dataset contains roughly 25% zero counts, while the 25%, 10%, and 5% observed datasets contain roughly 60%, 75%, and 85% zero counts respectively. The down-sampled datasets are analyzed using SAVER and by other quantification methods: library size normalized observed counts, the predictions μ_{cg}, K-nearest neighbor imputation (KNN)[17], singular value decomposition matrix completion (SVD)[18], and random forest imputation (RF)[19]. Using the predictions directly as the recovered values is similar in concept to the strategy used by Satija et al.[20] for spatial reconstruction, although a linear noise model was used instead of a Poisson model for counts. The last three are established missing data imputation algorithms that have been applied to gene expression microarray data[21]. They were applied to the observed datasets treating all zeros as missing data.

First, we considered the recovery accuracy of each gene's expression pattern across cells. For each gene g, under algorithm a, we computed the Pearson's correlation coefficient (ρ_g^a) of its recovered expression across cells with the corresponding values in the library size normalized reference dataset (Fig. 2A). The distribution of ρ_g^a across genes is compared across the three down-sampled datasets and the six algorithms (Fig. 2B left, upper). There is a substantial decrease in correlation between the observed data and the reference data as efficiency decreases. SAVER considerably improves the overall correlation with the reference dataset at each efficiency level, outperforming the other algorithms. We also wanted to measure the percent improvement in correlation for SAVER and other algorithms over using simply the normalized observed counts across genes (Fig. 2B left, lower). We see that for the majority of genes, SAVER has improved correlation with the reference over the observed, with improvements reaching up to 50%. The imputation algorithms KNN, SVD, and RF, all perform worse than using the observed data for the majority of genes. This is not surprising as some of the zeros are true zeros, and eliminating them from the analysis leads to biased estimation.

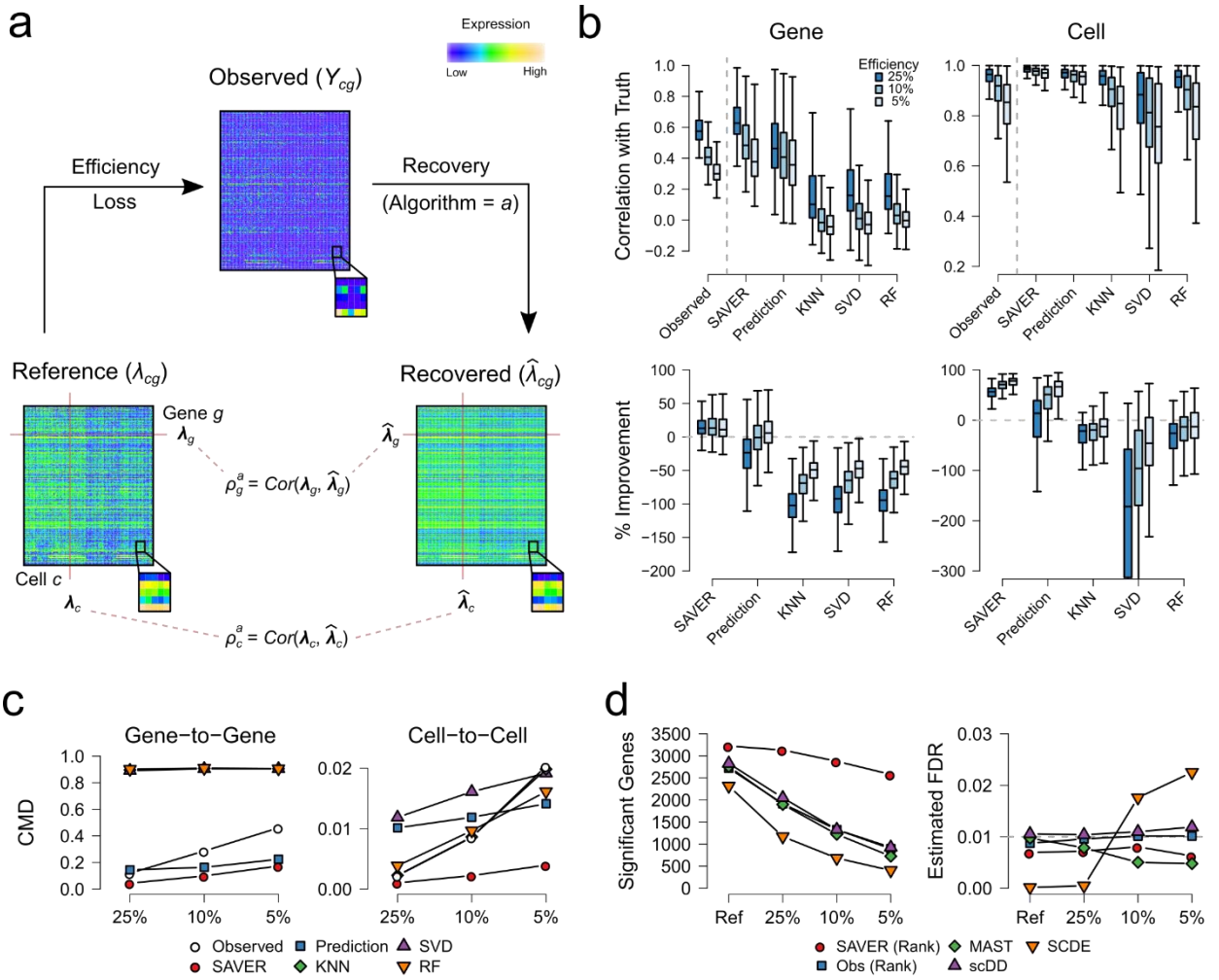

Figure 2 Evaluation of SAVER by down-sampled mouse brain dataset. (**a**) Schematic of down-sampling experiment: Observed data (Y_{cg}) consists of sparse low counts due to low efficiency sampling from the reference (λ_{cg}). Recovery algorithms are applied to the observed and evaluated based on correlation with reference for genes ρ_g and for cells ρ_c. (**b**) Performance of algorithms measured by correlation with reference, on the gene level (ρ_g^a, left) and on the cell level (ρ_c^a, right). For genes, boxplots reflect distribution of ρ_g^a over genes; for cells, boxplots reflect distribution of ρ_c^a over cells. Percentage improvement over simply using the observed data is shown in the lower two panels. SAVER is more closely correlated with the truth across both genes and cells. (**c**) Comparison of gene-to-gene (left) and cell-to-cell (right) correlation matrices of recovered values with the true correlation matrices, as measured by correlation matrix distance (CMD). CMD for each algorithm is plotted against sampling efficiency. SAVER preserves gene-to-gene and cell-to-cell relationships. (**d**) Differential expression (DE) analysis between CA1Pyr1 cells (n = 351) and CA1Py2 cells (n = 389). SAVER (Rank): Wilcoxon rank-sum test on SAVER recovered concentrations, Obs (Rank): Wilcoxon rank-sum test on library-size normalized observed values, MAST: Finak et al. (2015), scDD: Korthauer et al. (2017), SCDE: Kharchenko et al. (2014). Using a rank sum test on SAVER yields more significant genes (left), while still controlling false discovery rate at 0.01 (right).

Importantly, SAVER outperforms simply using the predictions, establishing that SAVER's adaptive weighting is a crucial and effective step.

Next, we considered the accuracy in the recovery of each cell's transcriptome as measured by the Pearson's correlation coefficient (ρ_c^a) of the cell's recovered gene concentrations with their corresponding values in the library size normalized reference dataset (Fig. 2A). Once again, we see that efficiency loss decreases correlation between the observed data and the reference

data at the cell level. SAVER improves the correlation of all cells to their reference values, showing even more substantial gains than at the gene level. Existing imputation algorithms, especially SVD, perform poorly. We believe this is due to the low-dimensional linear representation assumed by SVD, which is too idealistic for scRNA-seq data.

It is interesting to note in the gene-wise analysis that SAVER performs much better than simply using the prediction at 25% efficiency, but only slightly better at 5% efficiency. This is due to SAVER's adaptive weighting of the observed count versus the prediction to essentially only use the prediction when the prediction is trustworthy and the observed counts are not trustworthy. At 5% efficiency, the observed counts are so low that for most genes, SAVER is relying heavily on the predictions, hence the similarity in their performance.

SAVER improves the estimation of gene-to-gene and cell-to-cell relationships

Many downstream analyses, such as gene network analysis or cell type clustering, depend on faithful recovery of pairwise gene-to-gene or cell-to-cell relationships. To evaluate the effect of efficiency loss on these relationships, we computed gene-gene and cell-cell correlation matrices, and then evaluated the distance between the correlation matrices computed using the recovered expression and those computed using the normalized reference dataset (Fig. 2C). In the case of SAVER, we first calculated the correlation matrices on the SAVER estimates $\hat{\lambda}_{cg}$, which was then scaled by a correlation adjustment factor to account for the uncertainty in $\hat{\lambda}_{cg}$ (see Methods). As expected, both the observed gene-to-gene and cell-to-cell correlation matrices stray farther from their reference values as efficiency decreases. However, SAVER is able to recover most of the correlation structure in the reference datasets. In addition, all the three imputation algorithms destroy most gene-to-gene correlations. One interesting observation is that the difference with the reference is almost negligible for cell-to-cell correlations. Even at a 5% efficiency, the cell-to-cell correlation matrix still resembles the original correlation matrix. This is due to the fact that cell-to-cell correlations are driven by the highly expressed genes in each cell, which are not severely affected by low efficiency.

SAVER improves the power of differential expression analysis while controlling FDR

One of the main goals of scRNA-seq is differential expression in comparing gene expression profiles between various conditions or cell types. On the same reference and down-sampled datasets created above, we performed differential expression analysis between two subclasses of cells — 351 CA1Pyr1 cells and 389 CA1Pyr2 cells — identified by Zeisel et al. using a biclustering algorithm. We compared the performance of the following differential expression methods under a FDR of 0.01: Wilcoxon rank sum test using the SAVER recovered expression counts $\hat{\lambda}_{cg}$, Wilcoxon rank sum test on the library size normalized observed counts, MAST[3], scDD[22], and SCDE[2] (Fig. 2D left). The Wilcoxon rank sum test detects shifts in distribution, SCDE detects changes in means, MAST detects changes in means and zero proportions, and scDD detects differential distribution. We see that, as expected, the number of significant genes detected by each method decreases as efficiency decreases. The Wilcoxon test on the observed, MAST, and scDD perform similarly, detecting almost 3,000 genes in the reference dataset but less than 1,000 in the 5% efficiency dataset. SCDE detects fewer genes, but decreases at a similar rate as efficiency decreases. The Wilcoxon test on the SAVER estimates is able to detect 3,224 genes in the reference dataset, while maintaining similar numbers as efficiency decreases, with 2,579 genes detected at the 5% efficiency level (Supp. Table 1).

To examine whether FDR is controlled at the desired 0.01 level, we permuted the cell labels and quantified the number of rejections using each method for the permutated data (see Methods). All methods except SCDE control FDR at the approximate 0.01 level (Fig. 2D right). Thus, using a Wilcoxon test with SAVER estimates detects the most genes and is robust to efficiency loss, all while controlling FDR.

Using RNA FISH as gold standard, SAVER recovers gene expression distributions

SAVER is then evaluated on data from Torre and Dueck et al.[11], where Drop-seq and RNA FISH are applied to the same melanoma cell line. After filtering, 8,498 cells and 12,241 genes were kept for analysis from the Drop-seq experiment, with a median of 1,135 UMI counts per cell. Sixteen of these genes, including resistance markers and housekeeping genes, were profiled using RNA FISH across 7,000-88,000 cells, the exact number of cells depend on the gene. We applied SAVER to the sixteen genes in the Drop-seq dataset to obtain their SAVER estimates and posterior distributions. Counts in both FISH and Drop-seq were normalized by each cell's *GAPDH* expression to ensure comparability.

Torre and Dueck et al. demonstrated that while mean gene expression shows adequate correlation with the FISH counterpart, more subtle features of each gene's cross-cell expression distribution require stringent data filtering to include only cells with high transcriptome coverage. For example, one informative feature is Gini coefficient, which quantifies the inequity of expression across cells of a particular gene and is useful for identifying marker genes for rare cell types. We calculated Gini coefficients in the FISH, Drop-seq, and SAVER datasets for each of the remaining fifteen genes, removing *GAPDH* since it was used for normalization (Fig. 3A). Gini coefficients computed using SAVER recovered expression match extremely well with those computed on the FISH data. In comparison, Gini coefficients computed on the original Drop-seq data show substantial positive bias, in concordance with results of Torre and Dueck et al. If we were to use the filter, proposed by Torre and Dueck et al., of removing cells with less than 2,000 genes detected, then 87% of the cells would have to be removed, leaving 1,135 of the original 8,640 cells for analysis. In comparison, SAVER's analysis used 8,498 of the original 8,640 cells, allowing for more efficient use of the data and a more complete sampling of the original cell population, which is especially important for rare cell analysis. In this analysis, we assumed no knowledge of the efficiency of the Drop-seq experiment.

Another way to compare FISH and Drop-seq is simply to overlay, for each gene, its cross-cell expression distribution obtained by each method. Overlaying the distributions requires knowledge of the relative efficiency between the two technologies. Thus, for each gene, we computed its average relative efficiency as its mean Drop-seq count divided by its mean FISH count. Then, each gene's Drop-seq counts were scaled by its average relative efficiency to produce the Drop-seq raw distributions, which were then normalized by *GAPDH*. Similarly, for each gene, we sampled from its efficiency-adjusted SAVER posterior distribution for each cell, normalized by that cell's *GAPDH* expression, and aggregated across cells to get the SAVER recovered distribution (see Methods). We then compared these distributions to their FISH counterpart via the Kolmogorov-Smirnov (KS) statistic (Fig. 3B), which reveal that the SAVER recovered distributions are much closer to their FISH counterpart, as compared to those derived from the original Drop-seq data. We can also evaluate accuracy by simply plotting, for each gene, its kernel smoothed densities obtained from FISH, original Drop-seq, and SAVER recovery (*LMNA* and *CCNA2* shown in Fig. 3C and all genes shown in Supp. Fig. 1). SAVER substantially improves adherence to the FISH distributions for almost all genes, except for *C1S*.

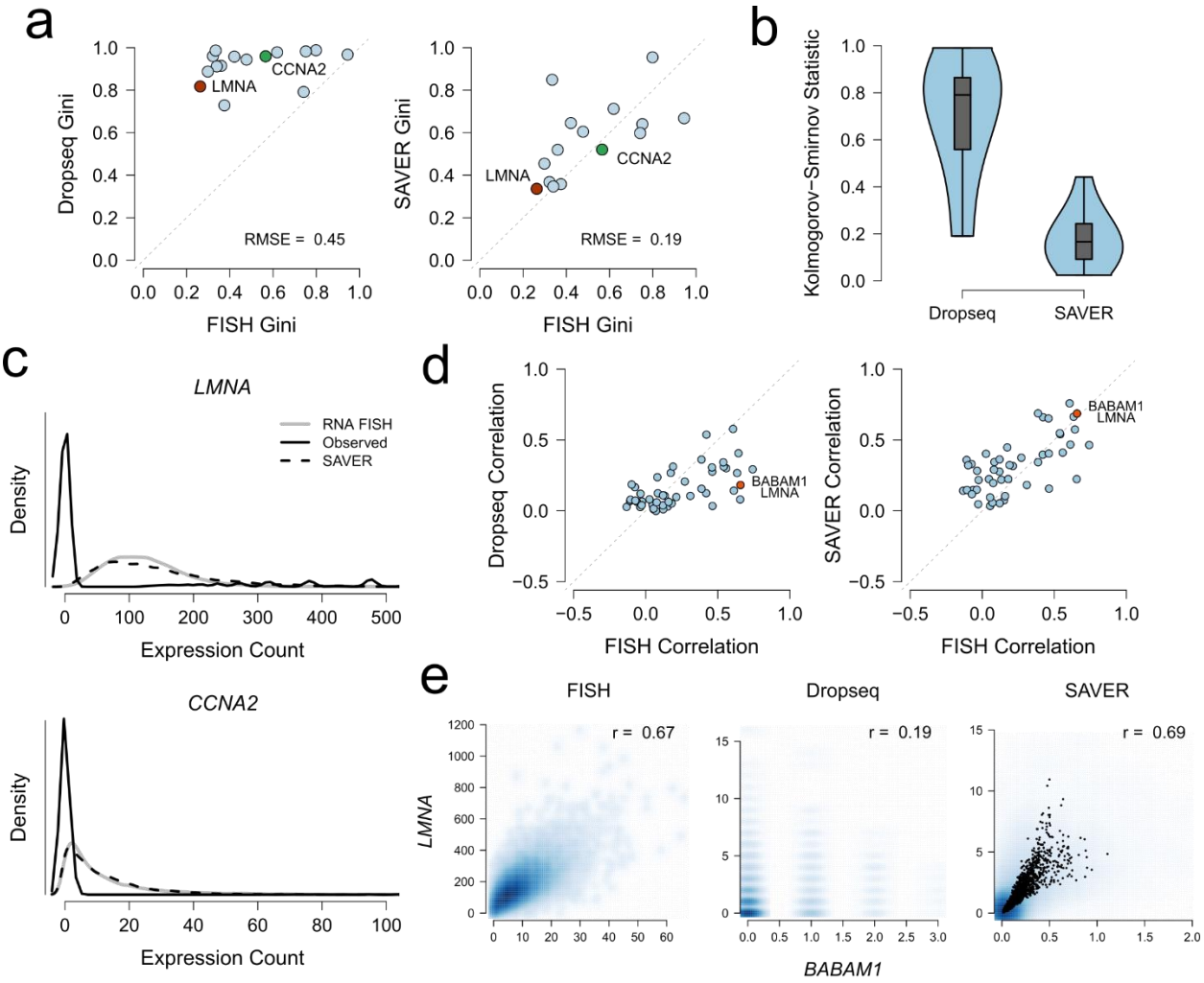

Figure 3 RNA FISH validation of SAVER results on Drop-seq data for 15 genes. (**a**) Comparison of Gini coefficient for each gene between FISH and Drop-seq (left) and between FISH and SAVER recovered values (right). (**b**) Violin plots of Kolmogorov-Smirnov statistics assessing, for each gene, the distance between its FISH density and its Drop-seq density, and between its FISH density and its SAVER density. (**c**) Kernel density estimates of cross-cell expression distribution of *LMNA* (upper panel) and *CCNA2* (lower panel), which are also highlighted in (**a**). Due to sparsity, the densities derived from original Drop-seq counts are heavily skewed towards zero for most genes. The distribution recovered by SAVER is validated by the distribution obtained from FISH data. Kernel density plots for the remaining 13 genes are given in Supplementary Materials. (**d**) Comparison of pair-wise gene correlations computed from Drop-seq original counts (left) and from SAVER recovered values (right) with those computed from FISH counts. SAVER is able to recover the depressed gene correlations in Drop-seq. (**e**) Scatterplots of expression levels between *BABAM1* and *LMNA*, which is also highlighted in (**d**).

The raw Drop-seq distributions, even after scaling by relative efficiency, do not resemble their FISH counterparts. This is because Drop-seq data, comprised mostly of very low counts, is highly discrete, and thus scaling by relative efficiency still results in a highly discrete distribution that is a poor estimate of the true distribution. SAVER, on the other hand, recovers continuous values that matches the true values after scaling.

SAVER recovers gene-to-gene relationships that are validated by RNA FISH

The data down-sampling experiments already suggest that SAVER can improve the inference of gene-to-gene and cell-to-cell relationships. We further evaluate the recovery of gene-to-gene

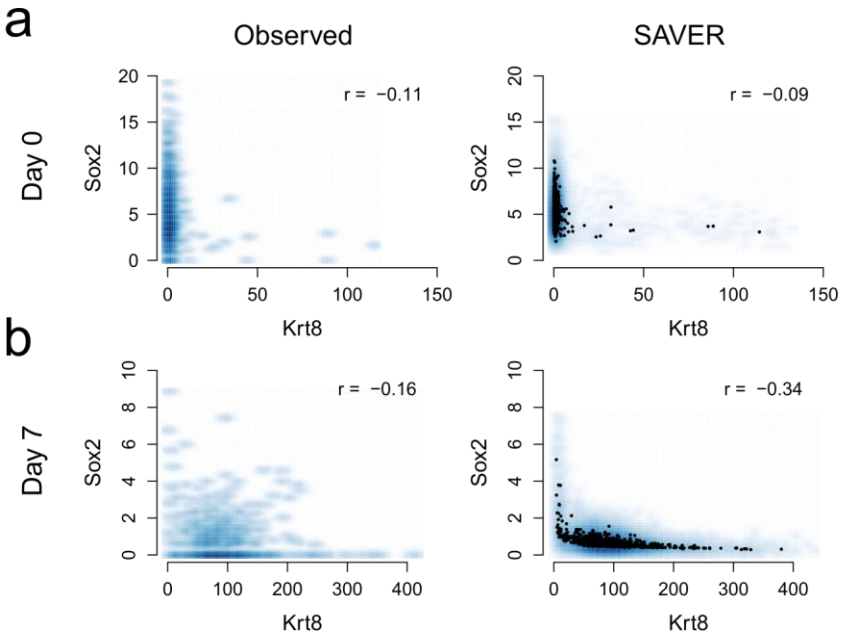

Figure 4 SAVER recovers gene-to-gene correlations in differentiating mouse embryonic stem cells. (**a**) Most day 0 cells express the pluripotency factor *Sox2* but not the epiblast-fate marker *Krt8*. (**b**) At 7 days post-activation, SAVER recovers the negative association between *Sox2* and *Krt8*. This negative relationship is masked in the observed Drop-seq counts by the presence of noise and zeros. Black points in the SAVER panels indicate SAVER estimates while colored gradient indicates density under the posterior distribution.

relationships on the melanoma data, using FISH as the gold standard. For each pair of genes, we computed their correlation for the unnormalized FISH, Drop-seq, and SAVER, making the pre-described correlation adjustment in SAVER (Fig. 3D). Due to low coverage, the correlations computed from the original Drop-seq counts are dampened compared to their FISH counterparts. Correlations computed by SAVER are much closer to their corresponding FISH values. Consider a specific pair of housekeeping genes, *BABAM1* and *LMNA* (Fig. 3E). For this pair, the correlation is 0.67 in FISH, 0.19 in Drop-seq, and 0.69 in SAVER. The SAVER estimates $\hat{\lambda}_{cg}$ are shown as black points while their posterior distributions, quantifying uncertainty, are shown as the blue gradient.

Illustration of SAVER in an ESC differentiation study

To further demonstrate the performance of SAVER, we analyzed the inDrop differentiating mouse embryonic stem cell dataset from Klein et al.[12]. The differentiating population was profiled at four time points: before LIF withdrawal (day 0), and at 2, 4, and 7 days post withdrawal. We focused on comparing gene relationships between known pluripotency factors and differentiation markers at day 0 and day 7 and evaluating how these relationships are recovered by SAVER. For example, consider the relationship between *Sox2*, a pluripotency factor, and *Krt8*, an epiblast differentiation marker in the observed counts and the SAVER estimates (Fig. 4). In the observed counts, there is no marked change in correlation between day 0 and day 7. After SAVER recovery, we see the correlation decrease from -0.09 to -0.34. In addition, in the SAVER day 7 plot, we see a clear trajectory relating decrease in Sox2 to increase in Krt8 as cells acquire epiblast fate. Based on SAVER's day 7 plot, most cells are differentiated and have low *Sox2*, but a few cells remain undifferentiated with high *Sox2* and low

Krt8. This pattern is almost indiscernible in the observed day 7 plot due to the large number of cells with zero *Sox2* expression. Similar observations were made for other gene pairs (Supp. Fig. 2).

Discussion

We have described SAVER, an expression recovery method for scRNA-seq. SAVER aims to recover true gene expression patterns by removing technical variation while retaining biological variation. SAVER uses the observed gene counts to form a prediction model for each gene and then uses a weighted average of the observed count and the prediction to estimate the true expression of the gene. The weights balance our confidence in the prediction with our confidence in the observed counts. In addition, SAVER provides a posterior distribution which reflects the uncertainty in the SAVER estimate and which can be sampled from for distributional analysis.

The down-sampling experiments and the comparisons between Drop-seq and FISH show that true expression patterns across genes and across cells are distorted by efficiency loss. This results in poor performance in downstream analyses such as detecting differentially expressed genes and evaluating gene expression distributions. In addition, imputation methods such as KNN, SVD matrix completion, and random forest imputation perform poorly in recovering the true expression. This is due to the extreme sparsity in the datasets and the fact that zeros do not occur at random, violating the missing-at-random assumptions of these imputation methods. In addition, the existing methods all inevitably remove some of the natural biological variation from the data — SVD by assuming a low-dimensional linear approximation and KNN by averaging across “similar” cells. Such aggressive approaches may be desirable for some analysis, but without limiting downstream analysis goals, SAVER attempts to retain cell-to-cell biological variation. This motivates the important predictability parameter ϕ_g within SAVER, which guards against over-smoothing and maintains natural biological variation.

We believe the closest method to SAVER in published studies of scRNA-seq data is Satija et al.[20], which forms LASSO-based linear predictions of marker genes to stabilize their estimates as part of the *Seurat* pipeline. Improving on Satija et al., SAVER uses a Poisson model with cell-specific size factors, and, more importantly, derives an adaptive-weighting scheme to balance prediction accuracy versus gene- and cell-specific coverage in forming the final estimate of expression. Our benchmark results show that this adaptive-weighting is crucial for accurate gene expression recovery, leading to substantial improvement of SAVER over simply using the initial predictions.

When the efficiency of each cell is unknown, SAVER recovers gene expression *concentrations*, which is always a continuous value. Converting concentrations to discrete expression counts requires knowledge of the efficiency, or equivalently, of the total RNA volume. For a gene with extremely low SAVER recovered concentration, deciding whether its true absolute expression is zero or an extremely low count requires knowledge of the total RNA volume. If the total RNA volume can be estimated, for example through spike-ins, then the recovered concentration can be scaled by this value, and rounded to the nearest integer to yield absolute counts. Alternatively, sampling from the scaled posterior gamma distribution followed by binning gives posterior absolute count probabilities for each gene.

Our analysis of the Drop-seq data with FISH validation demonstrates SAVER's ability to recover important features of gene expression distributions such as the Gini coefficient without knowledge of efficiency. Provided with an estimate of efficiency, SAVER is able to precisely recover not only the GINI coefficient but also the entire expression distribution. By FISH validation in Torre and Dueck et al. data, and by comparing to well-known relationships in the Klein et al. inDrop data, we also demonstrated that SAVER is able to recover true gene-to-gene relationships, while performing minimal pre-filtering of cells. Open-source software for SAVER can be downloaded from: https://github.com/mohuangx/SAVER.

Acknowledgments

M.H. and N.R.Z. acknowledge support from 2-R01-HG-006137. M.H. is also supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321851. J.M. and H.D. acknowledge support from R21 HD085201. A.R. and E.T. acknowledge support from NIH New Innovator Award DP2 OD008514, NIH/NCI PSOC award number U54 CA193417, NSF CAREER 1350601, NIH R33 EB019767, P30 CA016520, NIH 4DN U01 HL129998, NIH Center for Photogenomics RM1 HG007743, and the Charles E. Kauffman Foundation (KA2016-85223). S.S. acknowledges from NIH F30 AI114475. M.L acknowledges support from NIH R01GM108600 and NIH R01HL113147. R.B. acknowledges support from the NIH (DP2MH107055), the Searle Scholars Program (15-SSP-102), the March of Dimes Foundation (1-FY-15-344), a Linda Pechenik Montague Investigator Award, and the Charles E. Kauffman Foundation (KA2016-85223).

Methods

Datasets and pre-processing

The scRNA-seq mouse brain cell data used for simulations was obtained from Zeisel et al.[16]. The dataset contains UMI-labeled reads of 19,972 genes from 3,005 cells from various regions of the mouse brain. To create the true expression dataset, we selected a subset of genes that have non-zero expression in more than 40% of cells and a subset of cells that have a library size of over 10,000 UMI reads. We ended up with the reference dataset containing 3,529 genes and 1,799 cells.

The single molecule RNA FISH and UMI-labeled Drop-seq datasets of a melanoma cell line were acquired from Torre and Dueck et al.[11]. Sets of various genes were profiled across 6 batches of cells in the FISH experiment, totaling 88,040 cells and 26 genes. *GAPDH* expression was measured in all 88,040 cells. The Drop-seq data consisted of 32,287 genes and 8,640 cells. Genes with mean expression less than 0.01 and cells with library size less than 500 or greater than 20,000 were filtered out. 12,241 genes and 8,498 cells remained after filtering. Out of the filtered genes, 16 were profiled in the FISH experiment.

The scRNA-seq mouse embryonic stem (ES) cell data was obtained from Klein et al.[12]. Briefly, UMI-labeled reads for 24,175 genes were obtained at four time points: 935 ES cells before leukemia inhibitory factor (LIF) withdrawal, 303 ES cells 2 days post-LIF withdrawal, 683 ES cells 4 days post-LIF withdrawal, and 798 ES cells 7 days post-LIF withdrawal. The day 0 and day 7 datasets were used in the gene correlation analysis.

Gene expression model

Let Y_{cg} be the observed read count of gene g in cell c. We model Y_{cg} as a negative binomial random variable through the following Poisson-Gamma mixture

$$Y_{cg} \sim Poisson(s_c \lambda_{cg})$$

$$\lambda_{cg} \sim Gamma(\alpha_{cg}, \beta_{cg})$$

where λ_{cg} represents the normalized true expression. A gamma prior is placed on λ_{cg} to account for our uncertainty about its value. The shape parameter α_{cg} and the rate parameter β_{cg} are reparameterizations of μ_{cg} and ϕ_g, see details in Supplementary Materials. s_c represents the size normalization factor. In the following analyses, we use a library size normalization defined as the library size divided by the mean library size across cells. SAVER can also accommodate pre-normalized data.

SAVER procedure

Our goal is to derive the posterior gamma distribution for λ_{cg} given the observed counts Y_{cg} and use the posterior mean as the normalized SAVER estimate $\hat{\lambda}_{cg}$. The variance in the posterior distribution can be thought of as a measure of uncertainty in the SAVER estimate.

As stated before, we let the prior mean μ_{cg} be a prediction for gene g derived from the expression of other genes in the same cell. Specifically, we use the log normalized counts of all other genes g' as predictors in a Poisson generalized linear regression model with a log link function,

$$\log \mu_{cg} = \gamma_{g0} + \sum_{g' \neq g} \gamma_{gg'} \log\left[\frac{Y_{cg'} + 1}{s_c}\right].$$

Since the number of genes often far exceeds the number of cells, a penalized Poisson LASSO regression is used to shrink most of the regression coefficients to zero. We believe that this accurately reflects true biology since genes often only interact with a few other genes. The regression is fit using the *glmnet* R package version 2.0-5[23]. The model with the lowest five-fold cross-validation error is selected. We then use the selected model to get our predictions μ_{cg} for each gene in each cell.

The next step is to quantify the reliability of the prediction μ_{cg} using a dispersion parameter ϕ_g. We consider three models for ϕ_g, depending on what we assume to be constant for gene g across cells: constant coefficient of variation, constant Fano factor, or constant variance. A constant coefficient of variation corresponds to a constant shape parameter $\alpha_{cg} = \alpha_c$ in the gamma distribution and a constant Fano factor corresponds to a constant rate parameter $\beta_{cg} = \beta_c$ (see Theory Supplement). To determine which model for ϕ_g is the most appropriate, we calculate the marginal likelihood across cells under each definition and select the one with the highest maximum likelihood, and then set $\hat{\phi}_g$ to the maximum likelihood estimate.

Now that we have both μ_{cg} and $\hat{\phi}_g$, we can reparametrize, based on the chosen model for ϕ_g, into the usual shape and rate parameters of the gamma distribution, $\hat{\alpha}_{cg}$ and $\hat{\beta}_{cg}$. The posterior distribution is then

$$\lambda_{cg} | \, Y_{cg}, \hat{\alpha}_{cg}, \hat{\beta}_{cg} \sim Gamma(Y_{cg} + \hat{\alpha}_{cg}, s_c + \hat{\beta}_{cg})$$

The SAVER estimate $\hat{\lambda}_{cg}$ is the posterior mean:

$$\hat{\lambda}_{cg} = \frac{Y_{cg} + \hat{\alpha}_{cg}}{s_c + \hat{\beta}_{cg}} = \frac{s_c}{s_c + \hat{\beta}_{cg}} \frac{Y_{cg}}{s_c} + \frac{\hat{\beta}_{cg}}{s_c + \hat{\beta}_{cg}} \hat{\mu}_{cg}.$$

Estimating ϕ_g and computing the posterior distribution is quite fast computationally. The Klein day 0 dataset with 24,175 genes and 935 cells took under 5 minutes total. However, preforming the prediction with the LASSO regression is computationally intensive. For the Klein day 0 dataset, the LASSO regression took on average about 15 seconds per gene. However, this prediction step can be extensively parallelized.

Calculating correlations with SAVER

The SAVER estimate $\hat{\lambda}_{cg}$ cannot be directly used to calculate gene-to-gene or cell-to-cell correlations since we need to take into account its posterior uncertainty. Let the correlation between gene g and gene g' be represented by $\rho_{gg'} = Cor(\boldsymbol{\lambda}_g, \boldsymbol{\lambda}_{g'})$, where $\boldsymbol{\lambda}_g$ and $\boldsymbol{\lambda}_{g'}$ are the true expression vectors across cells. We can estimate $\rho_{gg'}$ by calculating the sample correlation of the SAVER estimate $\hat{\lambda}_{cg}$ and scaling by an adjustment factor, which takes into account the uncertainty of the estimate:

$$\hat{\rho}_{gg'} = Cor(\hat{\boldsymbol{\lambda}}_g, \hat{\boldsymbol{\lambda}}_{g'}) \times \frac{\sqrt{Var(\hat{\boldsymbol{\lambda}}_g)}\sqrt{Var(\hat{\boldsymbol{\lambda}}_{g'})}}{\sqrt{Var(\hat{\boldsymbol{\lambda}}_g) + E[Var(\boldsymbol{\lambda}_g|\boldsymbol{Z})]}\sqrt{Var(\hat{\boldsymbol{\lambda}}_{g'}) + E[Var(\boldsymbol{\lambda}_{g'}|\boldsymbol{Z})]}}$$

where $Var(\boldsymbol{\lambda}_g|\boldsymbol{Z})$ is a vector of posterior variances. The same adjustment can be applied to cell-to-cell correlations. See Supplementary Materials for derivation of this adjustment factor.

Distribution recovery

SAVER can be used to recover the distribution of either the absolute molecules counts or relative expression. Recovery of the absolute counts requires knowledge of the efficiency loss through ERCC spike-ins or some other control. To recover the absolute counts, we sample each cell from a Poisson-Gamma mixture distribution (i.e. negative binomial), where the gamma is the SAVER posterior distribution scaled by the efficiency. If the efficiency is not known or relative expression is desired, we sample each cell from the posterior gamma distribution.

Down-sampling datasets

Using the Zeisel et al. reference dataset as the true transcript count λ_{cg}, we generated down-sampled observed datasets by drawing from a Poisson distribution with mean parameter $\tau_c \lambda_{cg}$, where τ_c is the cell-specific efficiency loss. To mimic variation in efficiency across cells, we sampled τ_c as follows,

1. 25% efficiency: $\tau_c \sim Gamma(10, 40)$
2. 10% efficiency: $\tau_c \sim Gamma(10, 100)$
3. 5% efficiency: $\tau_c \sim Gamma(10, 200)$

Implementation of methods on down-sampled data

We compared the performance of SAVER against using the library-size normalized observed dataset, the regression prediction μ_{cg}, K-nearest neighbors imputation, singular value decomposition imputation, and random forest imputation in recovering the expression profile of the library-size normalized reference dataset. The imputation techniques were performed on the library size normalized observed data treating zeros as missing. KNN imputation was performed using the *impute.knn* function in the *impute* R package version 1.48.0, with parameters *rowmax* = 1, *colmax* = 1, and *maxp* = *p*. SVD imputation was performed using the *soft.Impute* function in the *softImpute* R package version 1.4, with parameters *rank.max* = 50, *lambda* = 30, and *type* = "svd". Random forest imputation was performed with the *missForest* R package version 1.4 with default parameters.

Gene-to-gene and cell-to-cell correlation analysis

Pairwise Pearson correlations were calculated for each library size normalized dataset and imputed dataset. Since the SAVER estimates have uncertainty, we want to calculate the correlation based on λ_{cg}. Correlations were first calculated using the SAVER recovered estimates $\hat{\lambda}_{cg}$ and scaled by the correlation adjustment factor described above.

The correlation matrix distance (CMD) is a measure of the distance between two correlation matrices with range from 0 (equal) to 1 (maximum difference)[24]. The CMD for two correlation matrices $\boldsymbol{R}_1, \boldsymbol{R}_2$ is defined as

$$d(\boldsymbol{R}_1, \boldsymbol{R}_2) = 1 - \frac{\mathrm{tr}(\boldsymbol{R}_1 \boldsymbol{R}_2)}{\|\boldsymbol{R}_1\|_f \|\boldsymbol{R}_2\|_f}.$$

Differential expression analysis of simulated datasets

Differential expression analysis between the CA1Pyr1 and CA1Pyr2 cells for the truth and down-sampled datasets were performed using a Wilcoxon rank sum test on the SAVER estimates, a Wilcoxon rank sum test on the observed expression, MAST, scDD, and SCDE. FDR control was set to 0.01 and no fold change cutoff was used. MAST version 1.0.5 was run on the library size normalized expression counts with the condition and scaled cellular detection rate as the Hurdle model input. The combined Hurdle test results were used. scDD version 1.2.0 was run on the library size normalized expression counts with default settings. Both the nonzero and the zero test results were used. SCDE version 2.2.0 was run on unnormalized expression counts with default parameters, except number of randomizations was set to 100. The p-value was calculated according to a two-sided test on the corrected Z-score.

To calculate the estimated false discovery rate, we first performed a permutation of the cell labels and determined the number of genes called as differentially expressed according to the p-value threshold defined for the unpermuted data. This number divided by the number of differentially expressed genes in the unpermutated data is the false discovery rate for that one

permutation. The final estimated false discovery rate is the average of the false discovery rates over 20 permutations.

RNA FISH and Drop-seq analysis

SAVER was performed on the Drop-seq dataset for the 16 genes that overlapped between Drop-seq and FISH. Since the FISH and Drop-seq experiments have different technical biases, we normalized by a *GAPDH* factor for each cell, defined as the expression of *GAPDH* divided by the mean of *GAPDH* across cells in each experiment. Since some cells have very low or very high *GAPDH* counts, we filtered out cells in the bottom and top 10^{th} percentile. For the Gini coefficient analysis where we assume we do not know the efficiency, we sampled the SAVER dataset from the SAVER posterior gamma distributions. We then filtered out cells in the bottom and top 10^{th} percentile of *GAPDH* expression in the sampled SAVER dataset and normalized the remaining by the *GAPDH* factor. For the distribution recovery, we calculated the efficiency loss for each gene as the mean Drop-seq expression divided by the mean SAVER expression. We scaled the pre-normalized Drop-seq dataset by the efficiency loss, filtered by *GAPDH*, and then normalized by the *GAPDH* factor. We scaled the SAVER posterior distributions by the efficiency loss and sampled from the Poisson-Gamma mixture to get the absolute counts as described above. We then performed the filtering and normalization by the *GAPDH* factor on the sampled SAVER dataset.

Correlation analysis was performed for pairs of genes in unnormalized FISH, Drop-seq, SAVER. Since the SAVER estimates were returned as library size normalized values, we rescaled by the library size to get the unnormalized values and used those to calculate the adjusted gene-to-gene correlations described above.

References

1. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14,** 381–387 (2017).

2. Kharchenko, P. V, Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–2 (2014).

3. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16,** 278 (2015).

4. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16,** 241 (2015).

5. Zhu, L., Lei, J. & Roeder, K. A Unified Statistical Framework for RNA Sequence Data from Individual Cells and Tissue. (2016). at <http://arxiv.org/abs/1609.08028>

6. Lin, P., Troup, M. & Ho, J. W. K. CIDR: Ultrafast and accurate clustering through imputation for single cell RNA-Seq data. *bioRxiv* 68775 (2016). doi:10.1101/068775

7. Wills, Q. F. *et al.* Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31,** 748–52 (2013).

8. Vallejos, C. A., Richardson, S. & Marioni, J. C. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.* **17,** 70 (2016).

9. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4,** 1707–1719 (2006).

10. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*

11. Torre, E. *et al.* A comparison between single cell RNA sequencing and single molecule RNA FISH for rare cell analysis.

12. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).

13. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).

14. L. 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).

15. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Comput. Biol.* **11,** e1004333 (2015).

16. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* **347,** 1138–1142 (2015).

17. O. Troyanskaya *et al.* Missing value estimation methods for {DNA} microarrays. *Bioinformatics* **17,** 520–525 (2001).

18. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Jmlr* **11,** 2287–2322 (2010).

19. Stekhoven, D. J. & Bühlmann, P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118 (2012).

20. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).

21. Aittokallio, T. Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Brief. Bioinform.* **11,** 253–264 (2009).

22. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17,** 222 (2016).

23. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** (2010).

24. Herdin, M., Czink, N., Ozcelik, H. & Bonek, E. Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels. *Veh. Technol. Conf. 2005. VTC 2005-Spring. 2005 IEEE 61st* **1,** 136–140 Vol. 1 (2005).