HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 December 25.

Published in final edited form as:

Nat Methods. 2018 July ; 15(7): 539–542. doi:10.1038/s41592-018-0033-z.

SAVER: Gene expression recovery for single-cell RNA sequencing

Mo Huang[1], **Jingshu Wang**[1], **Eduardo Torre**[2], **Hannah Dueck**[3], **Sydney Shaffer**[2], **Roberto Bonasio**[4], **John I. Murray**[3], **Arjun Raj**[2], **Mingyao Li**[5], and **Nancy R. Zhang**[1,*]

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA

[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA

[3]Department of Genetics, University of Pennsylvania, Philadelphia, PA

[4]Department of Cell and Developmental Biology, University of Pennsylvania, Philadelphia, PA

[5]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA

Abstract

In single-cell RNA sequencing (scRNA-seq) studies, only a small fraction of the transcripts present in each cell are sequenced. This leads to unreliable quantification of lowly and moderately expressed genes which hinders downstream analysis. To address this challenge, we introduce SAVER (Single-cell Analysis Via Expression Recovery), an expression recovery method for UMI-based scRNA-seq data that borrows information across genes and cells to obtain accurate expression estimates for all genes.

A primary challenge in the analysis of scRNA-seq data is the low capturing and sequencing efficiency affecting each cell, which leads to a large proportion of genes, often exceeding 90%, with zero or low read count. Although many of the observed zero counts reflect true zero expression, a considerable fraction is due to technical factors. The overall efficiency of current 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. In differential expression and cell type classification, transcripts expressed in a cell but not detected due to technical limitations are sometimes accounted for by a zero-

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

Author Contributions

This study was conceived of and led by N.R.Z. Jointly with N.R.Z. and M.L., M.H. designed the model and estimation algorithm, implemented the SAVER software, designed the in silico experiments, and led the data analysis. J.W. validated the Poisson noise model in ERCC data. E.T., H.D., S.S., R.B., J.I.M., and A.R. performed the RNA FISH and Drop-seq experiments for the melanoma cell line. M.H. and N.R.Z wrote the paper with feedback from J.W. and M.L.

Competing Financial Interests Statement

A.R. receives consulting income and A.R. and S.S. receive royalties related to Stellaris RNA FISH probes. All other authors declare no competing interests.

inflated model[2–4]. Recently, methods such as MAGIC[5] and scImpute[6] have been developed to directly estimate the true expression levels. Both MAGIC and scImpute rely on pooling the data for each gene across similar cells. However, we demonstrate later that this can lead to over-smoothing and may remove natural cell-to-cell stochasticity in gene expression, which has been shown to lead to biologically meaningful variations in gene expression, even across cells of the same type or of the same cell line[7–9]. In addition, MAGIC and scImpute do not provide a measure of uncertainty for their estimated values.

Here, we propose SAVER (Single-cell Analysis Via Expression Recovery), a method that takes advantage of gene-to-gene relationships to recover the true expression level of each gene in each cell, removing technical variation while retaining biological variation across cells (https://github.com/mohuangx/SAVER). SAVER receives as input a post-QC scRNA-seq dataset with unique molecule index (UMI) counts. SAVER assumes that the count of each gene in each cell follows a Poisson-Gamma mixture, also known as a negative binomial model. Instead of specifying the Gamma prior, we estimate the prior parameters in an empirical Bayes-like approach with a Poisson Lasso regression using the expression of other genes as predictors. Once the prior parameters are estimated, SAVER outputs the posterior distribution of the true expression, which quantifies estimation uncertainty, and the posterior mean is used as the SAVER recovered expression value (Fig. 1a, Online Methods).

First, we assessed SAVER's accuracy by comparing the distribution of SAVER estimates to distributions obtained by RNA FISH in data from Torre and Dueck et al.[10] In this study, Drop-seq was used to sequence 8,498 cells from a melanoma cell line. In addition, RNA FISH measurements of 26 drug resistance markers and housekeeping genes were obtained across 7,000 to 88,000 cells from the same cell line. After filtering, 15 genes overlapped between the Drop-seq and FISH datasets (Supplementary Fig. 1).

Since FISH and scRNA-seq were performed on different cells, the FISH and scRNA-seq derived estimates can only be compared in distribution. Accurate recovery of gene expression distribution is important for identifying rare cell types, identifying highly variable genes, and studying transcriptional bursting. We applied SAVER to the Drop-seq data and calculated the Gini coefficient[11], a measure of gene expression variability, for the FISH, Drop-seq, and SAVER results for these 15 overlapping genes. The Gini coefficient has been shown to be a useful measure for identifying rare cell types and sporadically expressed genes in the original FISH-based study of this cell line[9]. Thus, accurate recovery of the Gini coefficient would allow the same analysis to be performed with scRNA-seq.

For all genes, SAVER effectively recovered the FISH Gini coefficient, which Drop-seq grossly overestimates (Fig. 1b). In addition, we can compare the distributions of each gene's expression across cells and observe that, as compared to Drop-seq, SAVER recovered expression distributions match much more closely with the FISH distributions (Fig. 1c, Supplementary Fig. 2). In comparison, Gini estimates and recovered distributions obtained from MAGIC and scImpute do not match as well with the FISH estimates (Supplementary Fig. 3a-c).

Not only is SAVER capable of recovering gene expression distributions and distribution-level features, it is also able to recover true biological gene-to-gene correlations that are observed in FISH but dampened in Drop-seq. For example, SAVER can recover the strong correlation between housekeeping genes *BABAM1* and *LMNA*, which is lost in the Drop-seq data (Fig. 1d). In comparison, the correlations derived from MAGIC results are much higher than those derived from FISH, suggesting that MAGIC induces spurious correlation. On the other hand, scImpute averages the correlations, leading to biased estimates of the true correlation (Supplementary Fig. 3d). The fact that SAVER does not introduce spurious correlation for gene pairs that have no biological correlation is further demonstrated by a permutation study (Supplementary Note 1), which shows that for such gene pairs, the correlation estimates are shrunk to zero by SAVER, but inflated by MAGIC and scImpute (Supplementary Fig. 4).

Next, we evaluated whether SAVER can accurately recover the true expression level within each individual cell for each gene. Since it is difficult to determine the actual number of mRNA molecules in each cell, we performed down-sampling experiments on four datasets[12–15] to generate realistic benchmarking datasets. For each dataset, we first selected a subset of highly expressed genes and cells to act as the reference dataset, which we treat as the true expression. We then simulated the capture and sequencing process at low efficiencies while introducing cell-to-cell variability in library size (Online Methods). We ran SAVER, MAGIC, and scImpute on each of the observed datasets, as well as conventional missing data imputation algorithms.

To evaluate the performance of each method, we calculated the Pearson gene-wise correlation (p_g^a) across cells and the cell-wise correlation (p_c^a) across genes between the reference and observed data, as well as between the reference and recovered datasets (Supplementary Fig. 5). SAVER improves on both the gene-wise and cell-wise correlations across all datasets, while MAGIC, scImpute, and conventional missing data imputation algorithms usually perform worse than simply using the observed data (Fig. 2a, Supplementary Fig. 6, 7a). Next, we assessed the recovery of gene-to-gene and cell-to-cell correlation matrices, needed, respectively, for gene network reconstruction and cell type identification. To compare, we calculated the correlation matrix distance (CMD)[16] between the reference matrix and the observed/recovered matrix. SAVER lowers the gene-to-gene and cell-to-cell CMD for all datasets, MAGIC and scImpute perform similarly as the observed, and conventional missing data imputation algorithms perform worse than observed (Fig. 2b, Supplementary Fig. 7b).

To investigate the effect of SAVER on downstream analyses, we performed differential expression and cell clustering on the down-sampled data. In the Zeisel study, two subclasses of cells — 351 CAPyr1 and 389 CA1Pyr2 cells — were identified by the original authors. We performed differential expression analysis of these two subclasses using several differential expression methods[2,3,17]. After down-sampling, the number of differentially expressed genes detected is much lower than for the reference, but SAVER detects the most genes in the down-sampled data set while maintaining accurate FDR control (Fig. 2c, Supplementary Table 1).

Next, we performed cell clustering on the reference, observed, and recovered datasets using Seurat[18]. The reference-derived cell type clusters were treated as the truth and clustering accuracy on the observed and recovered datasets was assessed by the Jaccard index and by t-SNE[19] visualization. SAVER achieves a higher Jaccard index than the observed for all datasets, while MAGIC and scImpute have a consistently lower Jaccard index (Fig. 2d, Supplementary Fig. 8). Even though the Jaccard index for SAVER in the Chen and La Manno datasets are only slightly higher than the observed, the t-SNE plots reveal that SAVER clustering of the cells is a more accurate representation of the reference data than the observed. SAVER also gives more stable results across different numbers of principal components, a critical parameter choice for dimension reduction in Seurat prior to the application of t-SNE (Supplementary Fig. 9).

Finally, we demonstrated SAVER in the analysis of a mouse visual cortex dataset where 47,209 cells were classified into main cell types and subtypes through extensive analysis[20]. We applied SAVER to a random subset of 7,387 cells and performed t-SNE visualization of the observed versus the SAVER-recovered cells (Fig. 2e). A population of excitatory neurons is highlighted, and the individual subtypes are colored according to labels given by Hrvatin et al. In the t-SNE plot of the original counts, the subtypes are not well separated and are mostly indistinguishable. SAVER distinguished the individual subtypes with clear separation. This example is common in our general experience with SAVER: It does not affect well-separated cell types but identifies cell types and states for which the evidence in the original data may be weak.

We have shown that SAVER is able to accurately recover both population-level expression distributions and cell-level gene expression values, both of which are necessary for effective downstream analyses. Additional in-depth exploration in Supplementary Note 2 reveals how the performance of SAVER depends on factors such as sequencing depth, number of cells, and cell composition. In almost all scenarios, analyses using SAVER estimates improves upon analyses using the original counts, while in the worst-case scenario, SAVER does not hurt. The robust performance of SAVER is due to its adaptive estimation of gene-level dispersion parameters and its cross-validation-based model selection, which safeguard against unnecessary model complexity. By reducing noise and amplifying true biological relationships, SAVER improves the signal for downstream analyses.

Online Methods

Data Pre-processing and Quality Control

SAVER can be applied to the matrix of raw UMI counts. However, in a standard scRNA-seq data set, many genes have zero total counts across all cells, or have non-zero count in at most 1 or 2 cells. Genes exhibiting such extremely sparse expression would not benefit from the SAVER procedure, since there is little data to form a good prediction; however these genes do not affect the estimates of the other genes, and thus are harmless if left in. As we show in Figure 9 of Supplementary Note 2, SAVER gives the most improvement for genes with medium to low expression, and for these extremely low abundance genes, the SAVER recovered values would be similar to the observed value. Thus, to reduce computational time, we recommend removing these genes at the start. There are several existing

workflows[21–23] that perform a conservative filtering of low abundance genes, which can be applied prior to application of SAVER.

SAVER

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

$$\begin{aligned} Y_{gc} &\sim Poisson\left(s_c \lambda_{gc}\right) \\ \lambda_{gc} &\sim Gamma\left(\alpha_{gc}, \beta_{gc}\right) \end{aligned} \quad (1)$$

where λ_{gc} represents the normalized true expression. The Poisson model has been shown to be a good approximation of the noise in scRNA-seq data using UMIs[24,25]. Datasets without UMIs are subject to strong amplification bias and would violate the Poisson model assumed here. A gamma prior is placed on λ_{gc} to account for our uncertainty about its value. The shape parameter α_{gc} and the rate parameter β_{gc} are reparameterizations of the mean μ_{gc} and the variance v_{gc}, see details in Supplementary Note 3. 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, although other size factors such as those calculated by methods such as scran[26], BASiCS[27], SCnorm[28], or through ERCC spike-ins can be used. SAVER can also accommodate pre-normalized data.

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

We adopt an empirical Bayes-like technique to estimate the prior mean and prior variance. First, we estimate the prior mean μ_{gc}. We let μ_{gc} 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,

$$\log E\left(Y_{gc}/s_c | Y_{g'c}\right) = \log \mu_{gc} = \gamma_{g0} + \sum_{g' \neq g} \gamma_{gg'} \log\left[\frac{Y_{g'c}+1}{s_c}\right]. \quad (2)$$

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. In a Lasso regression, a penalty parameter lambda is added to the likelihood to control the number of predictors that have nonzero coefficients. A large penalty would correspond to a model with very few nonzero coefficients while a small penalty would correspond to a model with many nonzero coefficients. The genes that have nonzero coefficients can be thought of as genes that are

good predictors of the gene that is being estimated. We believe that this accurately reflects true biology since genes often only interact with a limited set of genes.

The regression is fit using the *glmnet* R package version 2.0-5[29]. For gene g, the response is the normalized observed expression $Y_{gc}/_{sc}$ and the predictors are $\log\left[\frac{Y_{g'c}+1}{s_c}\right]$. The regression model at the penalty with the lowest five-fold cross-validation error is selected (Supplementary Fig. 10). We then use the selected model to get our regression predictions $\hat{\mu}_{gc}$, which we treat as the prior mean for each gene in each cell.

The next step is to estimate the prior variance by assuming a constant noise model across cells denoted by a dispersion parameter ϕ_g. We consider three models for ϕ_g: constant coefficient of variation ϕ_g^{cv}, constant Fano factor ϕ_g^F, or constant variance ϕ_g^v. A constant coefficient of variation corresponds to a constant shape parameter $\alpha_{gc} = \alpha_g$ in the gamma prior and a constant Fano factor corresponds to a constant rate parameter $\beta_{gc} = \beta_g$ (see Supplementary Note 3). To determine which model for ϕ_g is the most appropriate, we calculate the marginal likelihood across cells under each model and select the one with the highest maximum likelihood, and then set $\hat{\phi}_g$ to the maximum likelihood estimate. Given $\hat{\phi}_g$ and the choice of noise model, we can derive $\hat{v}_{gc}$.

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

$$\lambda_{gc} \mid Y_{gc}, \hat{\alpha}_{gc}, \hat{\beta}_{gc} \sim Gamma\left(Y_{gc} + \hat{\alpha}_{gc}, s_c + \hat{\beta}_{gc}\right) \quad (3)$$

The SAVER estimate $\hat{\lambda}_{cg}$ is the posterior mean, a weighted combination of the regression prediction and the normalized observed expression:

$$\hat{\lambda}_{gc} = \frac{Y_{gc} + \hat{\alpha}_{gc}}{s_c + \hat{\beta}_{gc}} = \frac{s_c}{s_c + \hat{\beta}_{gc}} \frac{Y_{gc}}{s_c} + \frac{\hat{\beta}_{gc}}{s_c + \hat{\beta}_{gc}} \hat{\mu}_{gc}. \quad (4)$$

As seen from the above equation, the recovered expression $\hat{\lambda}_{gc}$ is a weighted average of the normalized observed counts Y_{gc}/s_c and the prediction $\hat{\mu}_{gc}$. The weights are a function of the size factor s_c and, through the $\hat{\beta}_{gc}$ term, the gene's predictability $\hat{\phi}_g$ and its prediction μ_{gc}. Genes for which the prediction is more trustworthy (small $\hat{\phi}_g$) have larger weight on the prediction $\hat{\mu}_{gc}$. 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. Supplementary Figure 11 shows example scenarios.

Estimating ϕ_g and computing the posterior distribution is fast computationally. The melanoma Drop-seq data with 12,241 genes and 8,498 cells took under 10 minutes total on one core of a standard desktop with an i7-3770 CPU. However, performing the prediction with the Lasso regression is computationally intensive. For the melanoma data, the Lasso regression took on average about 20 seconds per gene. However, this prediction step is highly parallelizable in the SAVER software and gene selection filters can be applied to reduce the dimensionality of the problem. An approximation to the prediction step is the default option, which reduces the computation time of the melanoma data to under an hour over 8 compute cores.

Calculating correlations with SAVER

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

$$\hat{\rho}_{gg'} = Cor\left(\hat{\lambda}_g, \hat{\lambda}_{g'}\right) \times \frac{\sqrt{Var\left(\hat{\lambda}_g\right)}\sqrt{Var\left(\hat{\lambda}_{g'}\right)}}{\sqrt{Var\left(\hat{\lambda}_g\right) + E\left[Var\left(\lambda_g \middle| Z\right)\right]}\sqrt{Var\left(\hat{\lambda}_{g'}\right) + E\left[Var\left(\lambda_{g'} \middle| Z\right)\right]}} \quad (5)$$

where $Var\left(\lambda_g \middle| Z\right)$ is a vector of posterior variances. The same adjustment can be applied to cell-to-cell correlations. See Supplementary Note 4 for derivation of this adjustment factor.

Distribution recovery

SAVER can be used to recover the distribution of either the absolute molecules counts or the relative expression values. 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 if relative expression is desired, we sample the expression level for each gene in each cell from the gene's posterior gamma distribution.

RNA FISH and Drop-seq analysis

The raw Drop-seq dataset contained 32,287 genes and 8,640 cells. Genes with mean expression less than 0.1 as well as cells with library size less than 500 or greater than 20,000 were removed. The filtered dataset contained 12,241 genes and 8,498 cells. RNA FISH measurements of 26 drug resistance markers and housekeeping genes were obtained across 7,000 to 88,000 cells from the same cell line. SAVER, MAGIC, and scImpute were performed on the Drop-seq data. MAGIC was performed using the Matlab version 0.1 with default settings and library size normalization. scImpute version 0.0.2 was used with default

settings. The 16 genes that were left after filtering are: 9 housekeeping genes (*BABAM1, GAPDH, LMNA, CCNA2, KDM5A, KDM5B, MITF, SOX10, VGF*) and 7 drug-resistance markers (*C1S, FGFR1, FOSL1, JUN, RUNX2, TXNRD1, VCL*) (Supplementary Table 2).

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. *GAPDH* read counts have been used as a proxy for cell size[30]. 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 in each dataset as the mean FISH expression divided by the mean dataset expression. We scaled the Drop-seq, MAGIC, and scImpute 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 and MAGIC 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.

Generating reference and down-sampled datasets—To generate a reference dataset from real scRNA-seq data, we selected high quality cells and highly expressed genes from the original dataset to treat as the true expression λ_{gc}. We generated down-sampled observed datasets by drawing from a Poisson distribution with mean parameter $\tau_c\lambda_{gc}$, where τ_c is the cell-specific efficiency loss.

We selected the cells, genes, and efficiency level so that the down-sampled dataset and the original full dataset are similar in mean expression and percentage of zero entries (Supplementary Table 3). We aimed to select roughly 50-60% of the cells with the largest library size and 10-20% of genes with the highest proportion of cells with nonzero expression (Supplementary Fig. 12).

The specific filters used for each dataset are as follows.

> Baron: Human pancreatic islet data contained 20,125 genes and 1,937 cells. Genes with mean expression less than 0.001 and non-zero expression in less than 3 cells were filtered out. The filtered dataset contained 14,729 genes and 1,937 cells. To generate the reference dataset, we selected genes that had non-zero expression in 25% of the cells and cells with a library size of greater than 5,000. We ended up with 2,284 genes and 1,076 cells.

Chen: Mouse hypothalamus data contained 23,284 genes and 14,437 cells. Cells with library size greater than 15,000 were filtered out. Genes with mean expression less than 0.0002 and non-zero expression in less than 5 cells were filtered out. The filtered dataset contained 17,053 genes and 14,216 cells. To generate the reference dataset, we selected genes that had non-zero expression in 20% of the cells and cells with a library size of greater than 2,000. We ended up with 2,159 genes and 7,712 cells.

La Manno: Human ventral midbrain data contained 19,531 genes and 1,977 cells. Genes with mean expression less than 0.001 and non-zero expression in less than 3 cells were filtered out. The filtered dataset contained 19,518 genes and 1,977 cells. To generate the reference dataset, we selected genes that had non-zero expression in 30% of the cells and cells with a library size of greater than 5,000. We ended up with 2,059 genes and 947 cells.

Zeisel: Mouse cortex and hippocampus data contained 19,972 genes and 3,005 cells. To generate the reference dataset, we selected genes that had non-zero expression in 40% of the cells and cells with a library size of greater than 10,000 UMIs. We ended up with 3,529 genes and 1,800 cells. We also filtered out one cell that had abnormally low library size after gene selection to end up with 1,799 cells.

To mimic variation in efficiency across cells, we sampled τ_c as follows,

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

The Baron, Chen, and La Manno datasets were sampled at 10% efficiency and the Zeisel dataset was sampled at 5% efficiency.

Implementation of methods on down-sampled data

We compared the performance of SAVER against using the library-size normalized observed dataset, MAGIC, and scImpute. The missing data 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 on the row and column centered matrix 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 on the matrix transpose with the *missForest* R package version 1.4 with default parameters.

Percentage change over observed was defined as

$$\%\ \ change\ \ over\ \ observed = \frac{r_{method} - r_{observed}}{r_{observed}}.$$

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_{gc}. Correlations were first calculated using the SAVER recovered estimates $\hat{\lambda}_{gc}$ 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)[16]. The CMD for two correlation matrices $R_1, \quad R_2$ is defined as

$$d(\boldsymbol{R}_1, \quad \boldsymbol{R}_2) = 1 - \frac{\mathrm{tr}(\boldsymbol{R}_1 \boldsymbol{R}_2)}{\|\boldsymbol{R}_1\|_f \|\boldsymbol{R}_2\|_f}. \quad (6)$$

Differential expression analysis of down-sampled datasets

For each down-sampled dataset, ten SAVER sampled datasets were generated by sampling from the posterior gamma distribution. A Wilcoxon rank sum test was run on each of the sampled datasets and the combined p-value was obtained via Rubin's rules for multiple imputation[31]. 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. For SAVER, one sampled dataset was considered one permutation.

Cell clustering and t-SNE visualization

Seurat version 2.0 was used to perform cell clustering and t-SNE visualization following the workflow detailed at http://satijalab.org/seurat/pbmc3k_tutorial.html. Briefly, normalization without filtering, identification of highly variable genes, scaling, PCA, jackStraw, cell clustering, and t-SNE were applied to the reference, down-sampled, SAVER, MAGIC, and scImpute datasets. The number of principal components used for cell clustering and t-SNE were identified through the jackStraw procedure. For the reference datasets, 15 PCs were chosen for Baron, Chen, and La Manno and 20 PCs were chosen for Zeisel. The number of principal components chosen for each down-sampled dataset and method is shown in

Supplementary Figure 8. The resolution for each reference dataset was chosen such that the cell clustering had the most agreement with the t-SNE visualization. Resolutions of 0.7, 0.6, 1.1, and 0.8 were chosen for Baron, Chen, La Manno, and Zeisel reference datasets respectively. Cell clusterings were calculated for each observed and recovered dataset at resolutions of 0.4-1.4 at intervals of 0.1. The Jaccard index was calculated at each resolution with the reference dataset, and the maximum Jaccard index was then reported. The Jaccard index was calculated using the R package *clusteval* version 0.1.

Hrvatin Study

Mouse visual cortex data contained 25,187 genes and 65,539 cells. Genes with mean expression less than 0.00003 and non-zero expression in less than 4 cells were filtered out. The filtered dataset contained 19,155 genes and 65,539 cells. 47,209 cells were classified into cell types by the authors. SAVER was run on a subsample of 10,000 cells. Out of these 10,000 cells, 7,387 cells had a subtype label and Seurat was used to cluster these cells. 35 principal components were chosen for the observed data and 30 principal components were chosen for the SAVER results as determined by the jackstraw procedure.

Software availability

SAVER v1.0.0 was used in this study with the setting *do.fast* = FALSE and is provided as Supplementary Software. The newest version of SAVER can be found at (https://github.com/mohuangx/SAVER). Scripts for data and figure generation can be found at (https://github.com/mohuangx/SAVER-paper).

Data availability

RNA FISH data from the melanoma cell line can be found at https://www.dropbox.com/s/ia9x0iom6dwueix/fishSubset.txt?dl=0. Single-cell sequencing data can be found at GSE99330. Five other public datasets were used in this study: Baron (GSM2230757), Chen (GSE87544), La Manno (GSE76381), Zeisel (linnarssonlab.org/cortex), and Hrvatin (GSE102827).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the following funding: NIH R01HG006137 (to N.R.Z. and MH); NIH R01GM125301 (to N.R.Z., M.L., and M.H.); NSF Graduate Fellowship DGE-1321851 (to M.H.); the Wharton Dean's Fund (to J.W.); NIH R21 HD085201 (to J.I.M. and H.D.); the NIH New Innovator Award DP2 OD008514, NIH/NCI PSOC award U54 CA193417, NSF CAREER 1350601, NIH R33 EB019767, NIH P30 CA016520, the NIH 4DN U01 HL129998, the NIH Center for Photogenomics RM1 HG007743, a Penn Epigenetics Program Pilot award, the Charles E. Kauffman Foundation KA2016-85223, and the Tara Miller Melanoma Foundation (to A.R. and E.T.); NIH F30 AI114475 (to S.S.); NIH R01GM108600 and R01HL113147 (to M.L.); 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 (to R.B.). This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (NSF OCI-1053575).

References

1. Svensson V, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017; 14:381–387. [PubMed: 28263961]
2. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11:740–2. [PubMed: 24836921]
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. 2015; 16:278. [PubMed: 26653891]
4. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015; 16:241. [PubMed: 26527291]
5. van Dijk D, et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv. 2017
6. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018; 9:997. [PubMed: 29520097]
7. Wills QF, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat Biotechnol. 2013; 31:748–52. [PubMed: 23873083]
8. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006; 4:1707–1719.
9. Shaffer SM, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017; 546:431–435. [PubMed: 28607484]
10. Torre E, et al. Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. Cell Syst. 2018; 6:171–179.e5. [PubMed: 29454938]
11. Jiang L, Chen H, Pinello L, Yuan G. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016; :1–13. DOI: 10.1186/s13059-016-1010-4 [PubMed: 26753840]
12. Baron M, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. 2016; 3:346–360. [PubMed: 27667365]
13. Chen R, Wu X, Jiang L, Zhang Y. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. Cell Rep. 2017; 18:3227–3241. [PubMed: 28355573]
14. La Manno G, et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell. 2016; 167:566–580.e19. [PubMed: 27716510]
15. Zeisel A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-). 2015; 347:1138–1142.
16. 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. 2005; 1:136–140.
17. Korthauer KD, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016; 17:222. [PubMed: 27782827]
18. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33:495–502. [PubMed: 25867923]
19. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. J Mach Learn Res. 2008; 9:2579–2605.
20. Hrvatin S, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nat Neurosci. 2017; 21

Methods-only References

21. Satija Lab. Seurat – Guided Clustering Tutorial. at <http://satijalab.org/seurat/pbmc3k_tutorial.html>
22. Lun, ATL., McCarthy, DJ., Marioni, JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. Bioconductor. at <https://bioconductor.org/help/workflows/simpleSingleCell/>

23. Kiselev, V., et al. Analysis of single cell RNA-seq data. at <https://hemberg-lab.github.io/scRNA.seq.course/index.html>
24. Wang J, et al. Gene Expression Distribution Deconvolution in Single Cell RNA Sequencing. bioRxiv. 2017:1–17.
25. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for single-cell RNA-Seq data. 2017
26. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016; 17:75. [PubMed: 27122128]
27. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLOS Comput Biol. 2015; 11:e1004333. [PubMed: 26107944]
28. Bacher R, et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods. 2017:1–6.
29. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33
30. 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. 2015; 58:339–352. [PubMed: 25866248]
31. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. John Wiley; 1987.

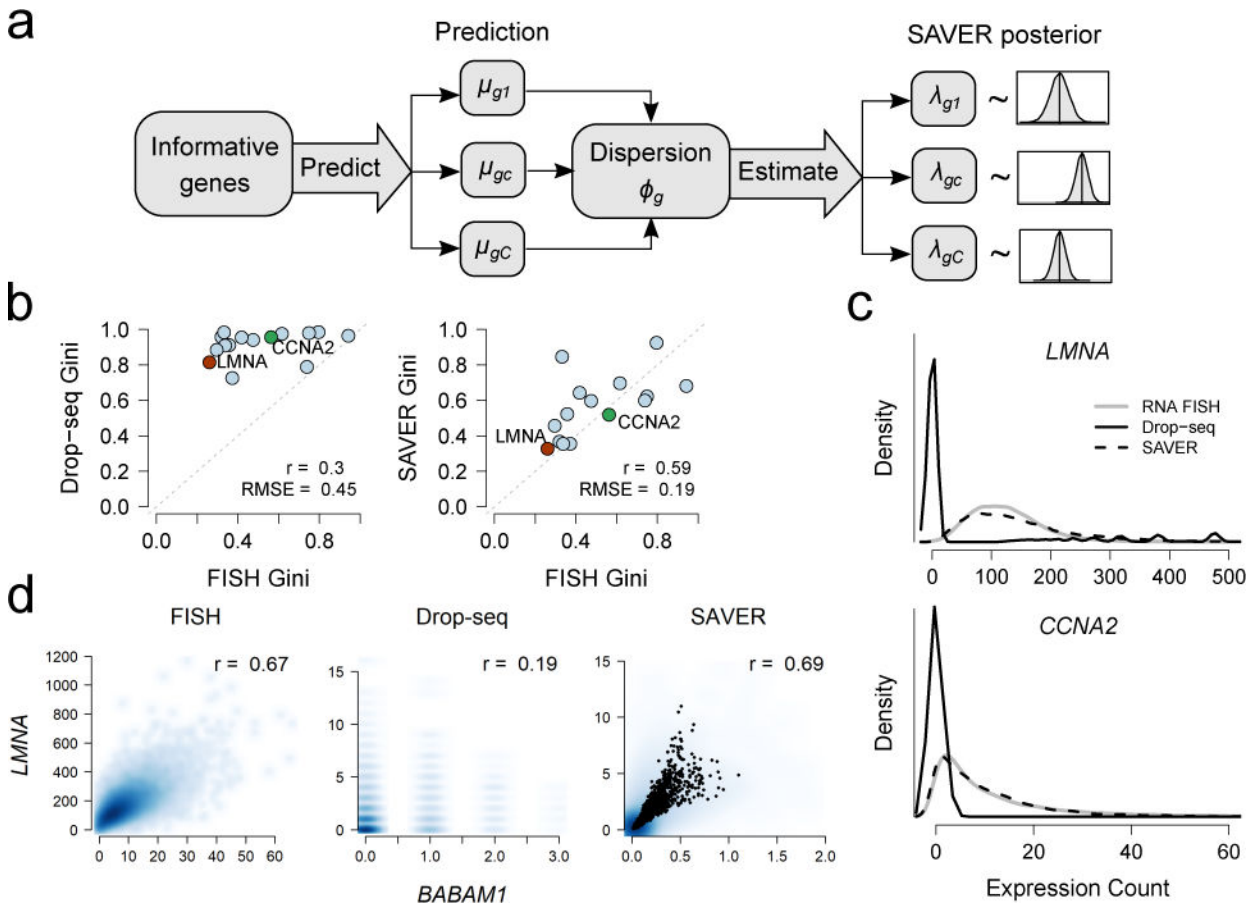

Figure 1.

RNA FISH validation of SAVER results on Drop-seq data. (**a**) Overview of SAVER procedure. (**b**) Comparison of Gini coefficient for each gene between FISH and Drop-seq (left) and between FISH and SAVER recovered values (right) for $n = 15$ genes. (**c**) Kernel density estimates of cross-cell expression distribution of LMNA (upper) and CCNA2 (lower). (**d**) Scatterplots of expression levels between BABAM1 and LMNA. Pearson correlations were calculated across $n = 17{,}095$ cells for FISH and $n = 8{,}498$ cells for Drop-seq and SAVER.

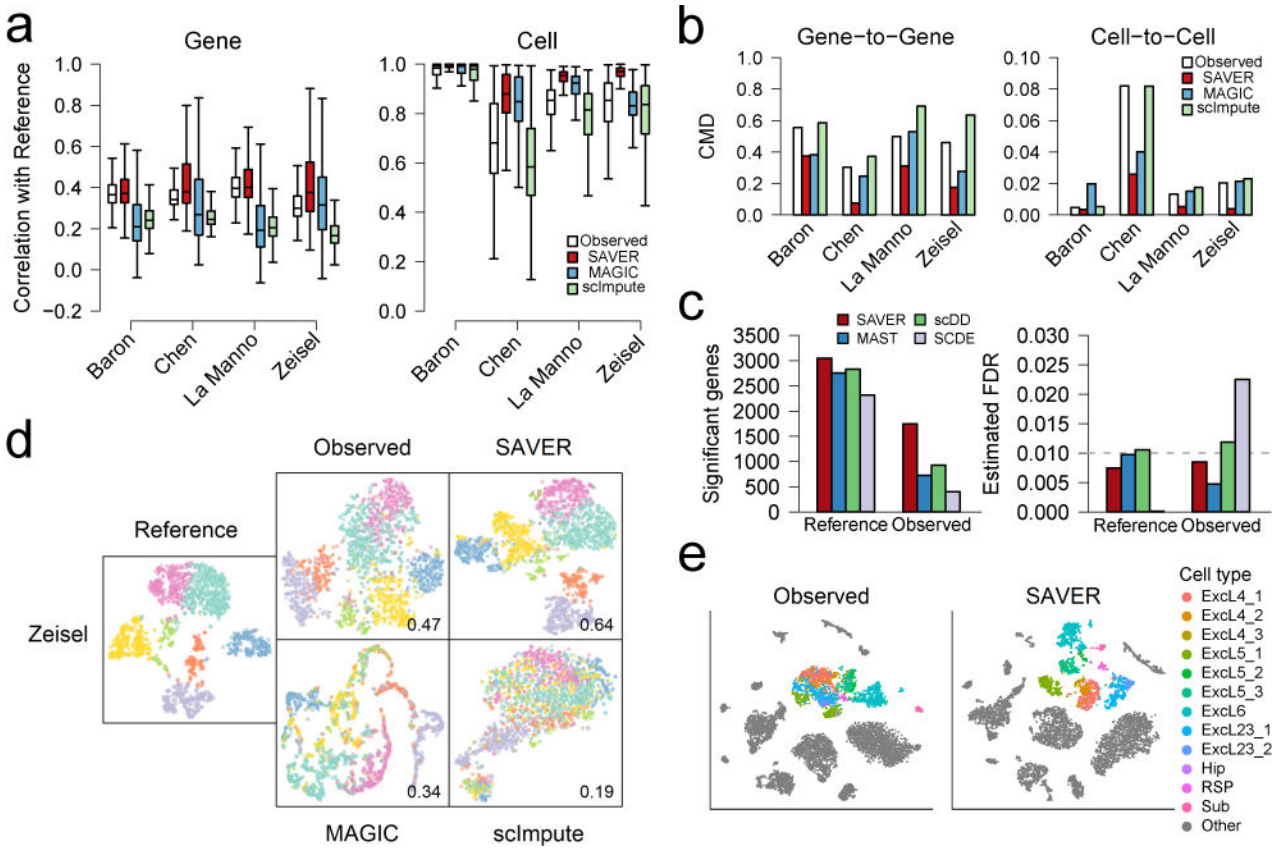

Figure 2.

Evaluation of SAVER by down-sampling and cell clustering. (**a**) Performance of algorithms measured by correlation with reference, on the gene level (left) and on the cell level (right). Number of genes and cells can be found in Supplementary Table 3. Box plots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile range (whiskers); outlier data beyond this range are not shown. (**b**) 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). (**c**) Differential expression (DE) analysis between CA1Pyr1 cells (n = 351) and CA1Py2 cells (n = 389) showing significant genes detected at FDR = 0.01 (left) and estimated FDR (right). (**d**) Cell clustering and t-SNE visualization of the Zeisel dataset (n = 1,799). Jaccard index of the down-sampled observed dataset and recovery methods as compared to the reference classification is shown. (**e**) t-SNE visualization of 7,387 mouse cortex cells for the observed data (left) and SAVER (right) colored by cell types determined by Hrvatin et al.