nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-018-0033-z

SAVER: gene expression recovery for single-cell RNA sequencing

Mo Huang[1], Jingshu Wang[1], Eduardo Torre[2,3], Hannah Dueck[4], Sydney Shaffer[3], Roberto Bonasio[5], John I. Murray[4], Arjun Raj[3,4], Mingyao Li[6] and Nancy R. Zhang[1]*

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 genes with low or moderate expression, which hinders downstream analysis. To address this challenge, we developed SAVER (single-cell analysis via expression recovery), an expression recovery method for unique molecule index (UMI)-based scRNA-seq data that borrows information across genes and cells to provide accurate expression estimates for all genes.

A primary challenge in the analysis of scRNA-seq data comes from the low transcript capture and sequencing efficiency of current methods. This leads to a large proportion of genes—often >90%—with zero or low read counts. Although many of the observed zero counts reflect a true absence of expression, a considerable fraction are due to technical factors. The overall efficiency of current scRNA-seq protocols can vary between <1% and >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 because of technical limitations are sometimes accounted for by a zero-inflated count distribution model[2–4]. Recently, methods such as MAGIC[5] and scImpute[6] have been developed to directly estimate true expression levels. Both MAGIC and scImpute rely on pooling of the data for each gene across similar cells. However, we found that this can lead to oversmoothing and may remove natural cell-to-cell stochasticity in gene expression. This stochasticity can represent biologically meaningful variation 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.

We developed SAVER, 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 uses a post-quality-control scRNA-seq dataset with UMI counts as input. 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 and Methods).

We assessed SAVER's accuracy by comparing the distribution of SAVER estimates to distributions obtained by RNA fluorescence in situ hybridization (FISH) in data from a previous study[10]. In that 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 7,000–88,000 cells from the same cell line. After filtering, 15 genes overlapped between the Drop-seq and FISH datasets (Supplementary Fig. 1).

Because different cells were used for the FISH and scRNA-seq analyses, the estimates derived via these two approaches can be compared only in terms of 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 was 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 done with scRNA-seq.

For all genes, SAVER effectively recovered the FISH Gini coefficient, which Drop-seq grossly overestimated (Fig. 1b). In addition, we were able to compare the distributions of each gene's expression across cells and observed that, compared with Drop-seq, SAVER recovered expression distributions that matched much more closely to the FISH distributions (Fig. 1c and Supplementary Fig. 2). Gini estimates and recovered distributions obtained from MAGIC and scImpute did 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, but 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 recovered the strong correlation between the housekeeping genes *BABAM1* and *LMNA*, which was lost in the Drop-seq data (Fig. 1d). In comparison, the correlations derived from MAGIC results (Supplementary Fig. 3d) were much higher than those derived from FISH, which suggests that MAGIC induces spurious correlation. In contrast, scImpute averages the correlations, leading to biased estimates of the true correlation (Supplementary Fig. 3d). The fact that SAVER

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [5]Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. *e-mail: nzh@wharton.upenn.edu

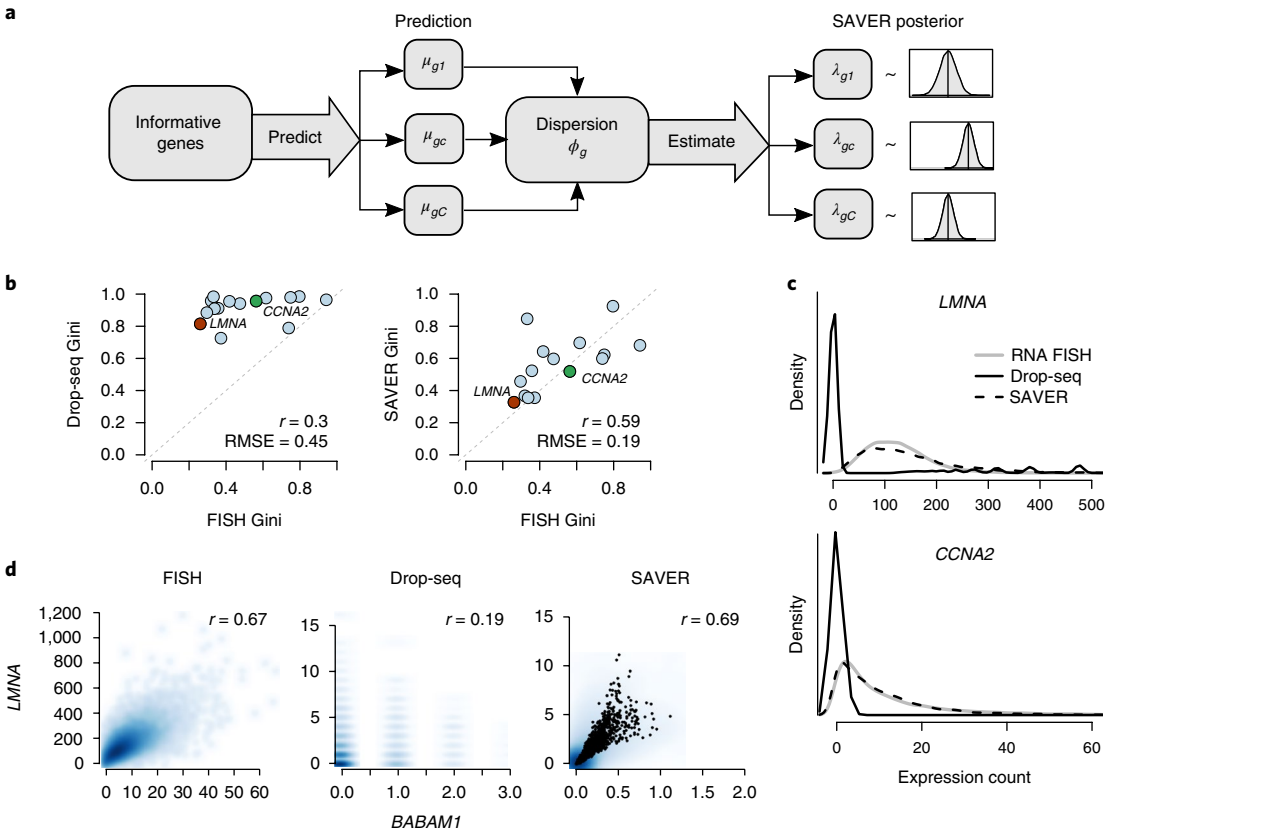

Fig. 1 | RNA FISH validation of SAVER results on Drop-seq data. a, Overview of SAVER procedure. **b**, Comparison of the Gini coefficient for each gene between FISH and Drop-seq (left) and between FISH and SAVER recovered values (right) for $n = 15$ genes. **c**, Kernel density estimates of cross-cell expression distributions of *LMNA* (top) and *CCNA2* (bottom). **d**, Scatter plots of expression levels for *BABAM1* and *LMNA*. Pearson correlations were calculated across $n = 17{,}095$ cells for FISH and $n = 8{,}498$ cells for Drop-seq and SAVER.

does not introduce spurious correlation for gene pairs that have no biological correlation was further demonstrated in a permutation study (Supplementary Note 1), which revealed that for such gene pairs, the correlation estimates were shrunk to zero by SAVER but were inflated by MAGIC and scImpute (Supplementary Fig. 4).

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

To evaluate the performance of each method, we calculated the Pearson gene-wise correlation (ρ_g^a) across cells and the cell-wise correlation (ρ_c^a) across genes between the reference and observed data, as well as between the reference and recovered datasets (Supplementary Fig. 5). SAVER improved on both the gene-wise and cell-wise correlations across all datasets, whereas MAGIC, scImpute, and conventional missing data imputation algorithms usually performed worse than use of the observed data (Fig. 2a and Supplementary Figs. 6 and 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 reduced the gene-to-gene and cell-to-cell CMD for all datasets, MAGIC and scImpute performed similarly as the observed data, and conventional missing data imputation algorithms performed worse than the observed data (Fig. 2b and Supplementary Fig. 7b).

To investigate the effect of SAVER on downstream analyses, we performed differential expression and cell clustering on the downsampled data. In a previous study[15], two subclasses of cells, 351 CAPyr1 and 389 CA1Pyr2 cells, were identified. We performed differential expression analysis of these two subclasses using several differential expression methods[2,3,17]. After downsampling, the number of differentially expressed genes detected was much lower than for the reference, but SAVER detected the most genes in the downsampled dataset while maintaining accurate false discovery rate (FDR) control (Fig. 2c and Supplementary Table 1).

Next, we carried out cell clustering on the reference, observed, and recovered datasets with Seurat[18]. We treated the reference-derived cell-type clusters as the truth, and we assessed clustering accuracy on the observed and recovered datasets by the Jaccard index and *t*-distributed stochastic neighbor embedding (t-SNE)[19] visualization. SAVER achieved a higher Jaccard index than that observed for all datasets, whereas MAGIC and scImpute had a consistently lower Jaccard index (Fig. 2d and Supplementary Fig. 8). Even though the Jaccard index for SAVER obtained with previously published datasets[13,14] was only slightly higher than that for the observed dataset, the t-SNE plots revealed that SAVER clustering of the cells was a more accurate representation of the reference data than the observed data. SAVER also yielded more stable results across different numbers of principal components, a critical

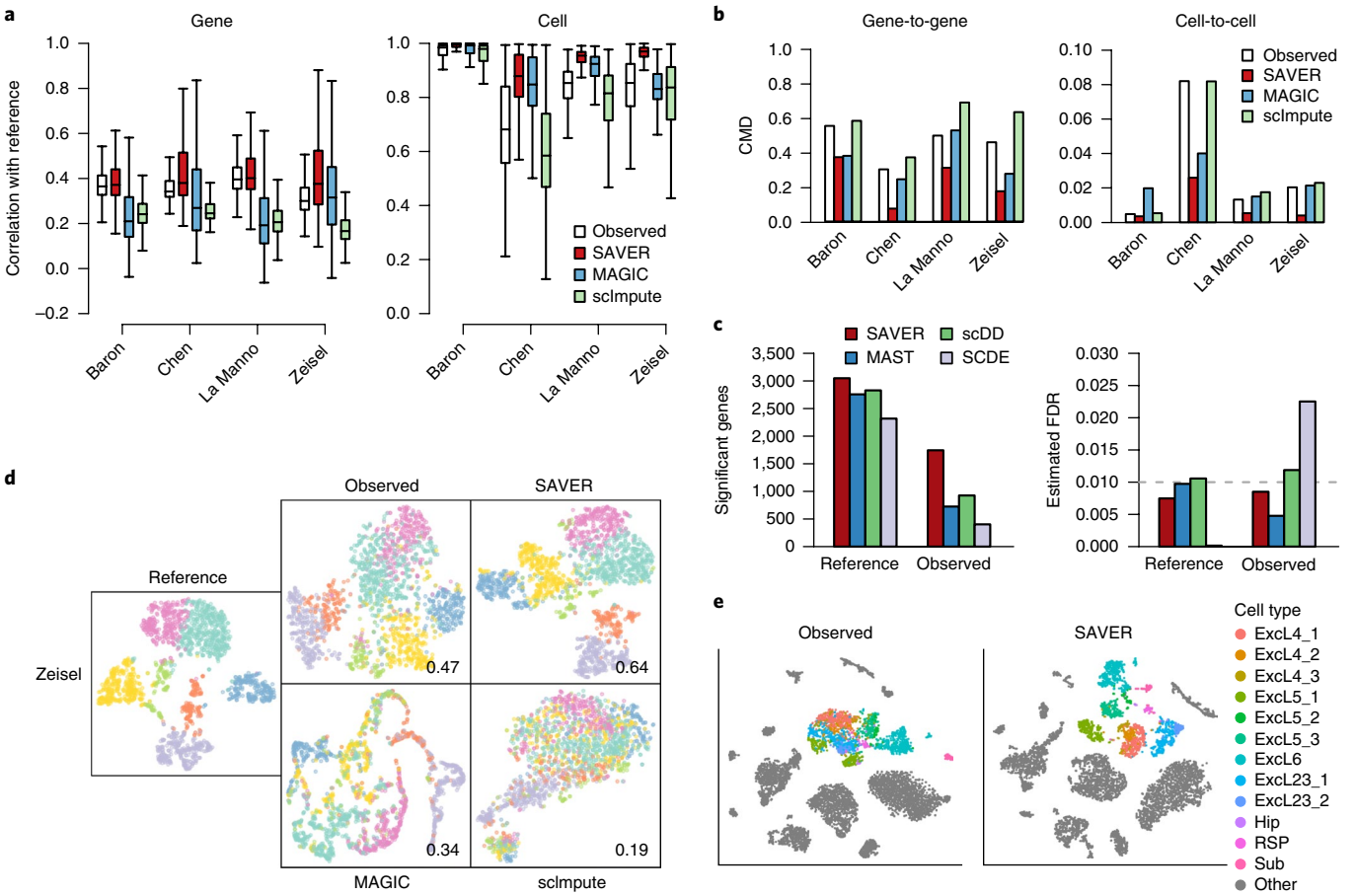

Fig. 2 | Evaluation of SAVER by downsampling and cell clustering. a, Performance of algorithms measured by correlation with reference data, on the gene level (left) and on the cell level (right). The numbers of genes and cells analyzed can be found in Supplementary Table 3. Box plots show the median (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 and cell-to-cell correlation matrices of recovered values with the true correlation matrices, as measured by CMD. **a**,**b**, Baron, Baron et al.[12]; Chen, Chen et al.[13]; La Manno, La Manno et al.[14]; Zeisel, Zeisel et al.[15]. **c**, Differential expression 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 data from Zeisel et al.[15] ($n=1,799$). The Jaccard index of the downsampled observed dataset and recovery methods as compared with the reference classification is shown. **e**, t-SNE visualization of 7,387 mouse cortex cells for the observed data and SAVER, color-coded by previously determined cell type[20].

parameter choice for dimension reduction in Seurat before the application of t-SNE (Supplementary Fig. 9).

Finally, we used SAVER to analyze a mouse visual cortex dataset in which 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 carried out t-SNE visualization of the observed versus the SAVER-recovered cells (Fig. 2e). A population of excitatory neurons was highlighted, and the individual subtypes were colored according to labels from a previous study[20]. In the t-SNE plot of the original counts, the subtypes were not well separated and were 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 it 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 shows how the performance of SAVER depends on factors such as sequencing depth, the number of cells, and cell composition. In almost all scenarios, analyses using SAVER estimates were improved compared with analyses using the original counts, and even in the worst-case scenario, SAVER did not lead to worse results. The robust performance of SAVER is a result of 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.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41592-018-0033-z.

Received: 25 September 2017; Accepted: 30 April 2018;
Published online: 25 June 2018

References

1. Svensson, V. et al. *Nat. Methods* **14**, 381–387 (2017).
2. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. *Nat. Methods* **11**, 740–742 (2014).
3. Finak, G. et al. *Genome Biol.* **16**, 278 (2015).
4. Pierson, E. & Yau, C. *Genome Biol.* **16**, 241 (2015).
5. van Dijk, D. et al. *bioRxiv* preprint at https://www.biorxiv.org/content/early/2017/02/25/111591 (2017).
6. Li, W. V. & Li, J. J. *Nat. Commun.* **9**, 997 (2018).

7. Wills, Q. F. et al. *Nat. Biotechnol.* **31**, 748–752 (2013).
8. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. *PLoS Biol.* **4**, e309 (2006).
9. Shaffer, S. M. et al. *Nature* **546**, 431–435 (2017).
10. Torre, E. et al. *Cell Syst.* **6**, 171–179 (2018).
11. Jiang, L., Chen, H., Pinello, L. & Yuan, G. C. *Genome Biol.* **17**, 144 (2016).
12. Baron, M. et al. *Cell Syst.* **3**, 346–360 (2016).
13. Chen, R., Wu, X., Jiang, L. & Zhang, Y. *Cell Rep.* **18**, 3227–3241 (2017).
14. La Manno, G. et al. *Cell* **167**, 566–580 (2016).
15. Zeisel, A. et al. *Science* **347**, 1138–1142 (2015).
16. Herdin, M., Czink, N., Ozcelik, H. & Bonek, E. *Proc. 2005 IEEE 61st Vehicular Technology Conference* 1, 136–140 (IEEE: Piscataway, NJ, 2005).
17. Korthauer, K. D. et al. *Genome Biol.* **17**, 222 (2016).
18. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. *Nat. Biotechnol.* **33**, 495–502 (2015).
19. Van Der Maaten, L. J. P. & Hinton, G. E. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
20. Hrvatin, S. et al. *Nat. Neurosci.* **21**, 120–129 (2018).

Acknowledgements

This work was supported by the NIH (grant R01HG006137 to N.R.Z. and M.H.; grant R01GM125301 to N.R.Z., M.L., and M.H.; R21 HD085201 to J.I.M. and H.D.; NIH New Innovator Award DP2 OD008514 to A.R.; R33 EB019767, P30 CA016520, and 4DN U01 HL129998 to A.R. and E.T.; F30 AI114475 to S.S.; R01GM108600 and R01HL113147 to M.L.; DP2MH107055 to R.B.), the NSF (Graduate Fellowship DGE-1321851 to M.H.), the Wharton Dean's Fund (to J.W.), the NCI (NIH/NCI PSOC award U54 CA193417 to A.R. and E.T.), the NSF (CAREER award 1350601 to A.R. and E.T.), the NIH Center for Photogenomics (RM1 HG007743 to A.R. and E.T.), a Penn Epigenetics Program Pilot award (A.R. and E.T.), the Charles E. Kauffman Foundation (KA2016-85223 to A.R. and E.T.), the Tara Miller Melanoma Foundation (to A.R. and E.T.), the Searle Scholars Program (15-SSP-102 to R.B.), the March of Dimes Foundation (1-FY-15-344 to R.B.), a Linda Pechenik Montague Investigator award (R.B.), 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).

Author contributions

N.R.Z. conceived and led this work. M.H., N.R.Z., and M.L. 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 Interests

A.R. receives consulting income and A.R. and S.S. receive royalties related to Stellaris RNA FISH probes.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0033-z.

Correspondence and requests for materials should be addressed to N.R.Z.

Methods

Data preprocessing and quality control. SAVER can be applied to the matrix of raw UMI counts. However, in a standard scRNA-seq dataset, many genes have zero total counts across all cells, or have a nonzero count in at most one or two cells. Genes exhibiting such extremely sparse expression would not benefit from the SAVER procedure, as there are few data with which 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 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 apply conservative filtering of low-abundance genes, which can be implemented before the 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

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

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 with 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} (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], and SCnorm[28] or through ERCC spike-ins can be used. SAVER can also accommodate prenormalized 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,

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

As 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 λ 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, whereas 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, as genes often interact with only a limited set of genes.

The regression is fit using the glmnet R package version 2.0–5 (ref. [29]). For gene g, the response is the normalized observed expression Y_{gc}/s_c and the predictors are $\log\left[\left(Y_{g'c}+1\right)/s_c\right]$. The regression model at the penalty with the lowest fivefold 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, and 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 $\beta_{gc}=\beta_g$ (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, according to 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

$$\lambda_{gc}|Y_{gc},\hat{\alpha}_{gc},\hat{\beta}_{gc} \sim Gamma\,(Y_{gc}+\hat{\alpha}_{gc}, s_c+\hat{\beta}_{gc}) \tag{3}$$

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

$$\hat{\lambda}_{gc} = \frac{Y_{gc}+\hat{\alpha}_{gc}}{s_c+\hat{\beta}_{gc}} = \frac{s_c}{s_c+\hat{\beta}_{gc}}\frac{Y_{gc}}{s_c} + \frac{\hat{\beta}_{gc}}{s_c+\hat{\beta}_{gc}}\hat{\mu}_{gc} \tag{4}$$

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 $\hat{\phi}_g$) are weighted more toward the prediction $\hat{\mu}_{gc}$. Genes with higher expression are weighted more toward 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.

Estimation of ϕ_g and computation of the posterior distribution are fast computationally. The melanoma Drop-seq data with 12,241 genes and 8,498 cells took under 10 min 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 s 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 reduced the computation time for the melanoma data to under an hour over eight 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 because its posterior uncertainty needs to be accounted for. 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 that takes into account the uncertainty of the estimate:

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

where $Var\,(\lambda_g|\boldsymbol{Z})$ 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 molecule 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., a 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. We removed genes with mean expression less than 0.1 and cells with library size less than 500 or greater than 20,000. 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 7,000–88,000 cells from the same cell line. SAVER, MAGIC, and scImpute were performed on the Drop-seq data. MAGIC was performed using 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 were 9 housekeeping genes (*BABAM1*, *GAPDH*, *LMNA*, *CCNA2*, *KDM5A*, *KDM5B*, *MITF*, *SOX10*, and *VGF*) and 7 drug-resistance markers (*C1S*, *FGFR1*, *FOSL1*, *JUN*, *RUNX2*, *TXNRD1*, and *VCL*) (Supplementary Table 2).

Because 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 *GAPDH* expression across cells in each experiment. *GAPDH* read counts have been used as a proxy for cell size[30]. Because some cells had very low or very high *GAPDH* counts, we filtered out cells in the bottom and top tenth percentiles. For the Gini coefficient analysis, where we assumed we did 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 tenth percentiles 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 datasets by the efficiency loss, filtered by *GAPDH* expression, 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 carried out filtering and normalization by the *GAPDH* factor for the sampled SAVER dataset.

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

Generating reference and downsampled datasets. To generate a reference dataset from real scRNA-seq data, we selected high-quality cells and genes with high expression from the original dataset to treat as the true expression λ_{gc}. We generated downsampled 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 downsampled dataset and the original full dataset were similar in mean expression and the 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 et al.[12]: Human pancreatic islet data contained 20,125 genes and 1,937 cells. Genes with mean expression less than 0.001 and nonzero expression in fewer than three 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 nonzero expression in 25% of the cells and cells with a library size greater than 5,000. We ended up with 2,284 genes and 1,076 cells.

Chen et al.[13]: Mouse hypothalamus data contained 23,284 genes and 14,437 cells. Cells with a library size greater than 15,000 were filtered out. Genes with mean expression less than 0.0002 and nonzero expression in fewer than five 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 nonzero expression in 20% of the cells and cells with a library size greater than 2,000. We ended up with 2,159 genes and 7,712 cells.

La Manno et al.[14]: Human ventral midbrain data contained 19,531 genes and 1,977 cells. Genes with mean expression less than 0.001 and nonzero expression in fewer than three 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 nonzero expression in 30% of the cells and cells with a library size greater than 5,000. We ended up with 2,059 genes and 947 cells.

Zeisel et al.[15]: Mouse cortex and hippocampus data contained 19,972 genes and 3,005 cells. To generate the reference dataset, we selected genes that had nonzero expression in 40% of the cells and cells with a library size greater than 10,000 UMIs. We ended up with 3,529 genes and 1,800 cells. We also filtered out one cell that had an abnormally low library size after gene selection, thus ending 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 datasets from Baron et al.[12], Chen et al.[13], and La Manno et al.[14] were sampled at 10% efficiency, and the dataset from Zeisel et al.[15] was sampled at 5% efficiency.

Implementation of methods on downsampled data. We compared the performance of SAVER against use of the library-size-normalized observed dataset, MAGIC, and scImpute. We applied missing-data-imputation techniques on the library-size-normalized observed data, treating zeros as missing. KNN imputation was carried out with 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 done 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 transposed 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. Because the SAVER estimates have uncertainty, we wanted to calculate the correlation on the basis of λ_{gc}. We first calculated correlations by using the SAVER recovered estimates $\hat{\lambda}_{gc}$ and scaled by the correlation adjustment factor described above.

The CMD is a measure of the distance between two correlation matrices and ranges from 0 (equal) to 1 (maximum difference)[16]. The CMD for two correlation matrices $\boldsymbol{R}_1, \boldsymbol{R}_2$ is defined as

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

Differential expression analysis of downsampled datasets. For each downsampled dataset, we generated ten SAVER sampled datasets 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 non-normalized expression counts with default parameters, except the 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 FDR, 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 FDR for that one permutation. The final estimated FDR is the average of the FDRs over 20 permutations. For SAVER, one sampled dataset was considered one permutation.

Cell clustering and t-SNE visualization. We used Seurat version 2.0 to perform cell clustering and t-SNE visualization according to 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, downsampled, SAVER, MAGIC, and scImpute datasets. The number of principal components (PCs) used for cell clustering and t-SNE was identified through the jackStraw procedure. For the reference datasets, 15 PCs were chosen for Baron et al.[12], Chen et al.[13], and La Manno et al.[14], and 20 PCs were chosen for Zeisel et al.[15]. The number of PCs chosen for each downsampled dataset and method is shown in Supplementary Fig. 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 reference datasets from Baron et al.[12], Chen et al.[13], La Manno et al.[14], and Zeisel et al.[15], 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[20]. Mouse visual cortex data contained 25,187 genes and 65,539 cells. Genes with mean expression less than 0.00003 and nonzero expression in fewer than four 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 PCs were chosen for the observed data and 30 PCs 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.

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

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 et al.[12] (GSM2230757), Chen et al.[13] (GSE87544), La Manno et al.[14] (GSE76381), Zeisel et al.[15] (https://linnarssonlab.org/cortex), and Hrvatin et al.[20] (GSE102827). Source data for Figs. 1 and 2 are available online.

References

21. Satija, R. et al. Seurat: guided clustering tutorial. *Satija Lab* http://satijalab.org/seurat/pbmc3k_tutorial.html (2018).
22. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. Workflow package: simpleSingleCell. *Bioconductor* https://bioconductor.org/help/workflows/simpleSingleCell/ (2016).

23. Kiselev, V. et al. Analysis of single cell RNA-seq data. *Hemberg Lab* https://hemberg-lab.github.io/scRNA.seq.course/index.html (2018).
24. Wang, J. et al. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2017/12/01/227033 (2017).
25. Wagner, F., Yan, Y. & Yanai, I. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2017/11/21/217737 (2017).
26. Lun, A. T., Bach, K. & Marioni, J. C. *Genome Biol.* **17**, 75 (2016).
27. Vallejos, C. A., Marioni, J. C. & Richardson, S. *PLOS Comput. Biol.* **11**, e1004333 (2015).
28. Bacher, R. et al. *Nat. Methods* **14**, 584–586 (2017).
29. Friedman, J., Hastie, T. & Tibshirani, R. *J. Stat. Softw.* **33**, 1–22 (2010).
30. Padovan-Merhar, O. et al. *Mol. Cell* **58**, 339–352 (2015).
31. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys* (John Wiley, Hoboken, NJ, 1987).

nature research

Corresponding author(s): Nancy R. Zhang

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

SAVER was applied to 6 scRNA-seq datasets produced by inDrop, Drop-seq, and STRT-seq. One human brain, one human pancreas, and three mouse brain datasets were analyzed. The melanoma cell line scRNA-seq data from Torre & Dueck was supported by FISH validation of gene expression. No statistical methods were used to predetermine sample size.

2. Data exclusions

Describe any data exclusions.

Low quality cells and genes from Baron, Chen, La Manno, Torre & Dueck, and Hrvatin datasets were excluded prior to analysis.

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.

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.

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.

Torre & Dueck: 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.

Hrvatin: 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.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

SAVER recovery of distributional characteristics and gene-pair correlations was validated with RNA FISH gene expression measurements of the same melanoma cell line. Down-sampling experiments demonstrated the improvements of SAVER in estimating the true reference expression levels and cell clustering across four datasets. Finally, SAVER was able to recover validated cell subtypes using a fraction of the cells in the Hrvatin data analysis. SAVER results were consistent across all datasets.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was performed as there were no experiments performed.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Cell subtypes validated in the Hrvatin study were not revealed until after SAVER analysis and clustering.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ 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 any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

SAVER v1.0.0 (https://github.com/mohuangx/SAVER), MAGIC v0.1 (Matlab), and scImpute v0.0.2 were used for gene expression recovery. impute v1.48.0, softImpute v1.4, and missForest v1.4 were used for missing data imputation. Seurat v2.0 was used for cell clustering and t-SNE visualization. clusteval version 0.1 was used for calculating the Jaccard index. MAST v1.0.5, scDD v1.2.0, and SCDE v2.2.0 were used for differential expression analysis. reldist v1.6.6 was used to calculate the Gini coefficient.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

Melanoma cell lines (WM989-A6, WM989-A6-G3) were obtained from Meedhard Herlyn and grown in the laboratory of A.R.

b. Describe the method of cell line authentication used.

The laboratory of Meedhard Herlyn performed short tandem repeat profiling using AmpFLSTR Identifiler PCR Amplification Kit (Life Technologies), in Tu2% media containing 78% MCDB, 20% Leibovitz's L-15 media, 2% FBS, and 1.68 mM CaCl2 and primary melanocytes isolated from human neonatal foreskin (Fom217-1 from the laboratory of M.H.) in Medium 254CF (Life Technologies, M254500) sup- plemented with Human Melanocyte Growth Supplement (Life Technologies, S0025).

c. Report whether the cell lines were tested for mycoplasma contamination.

Cell lines tested negative for mycoplasma.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.