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.