nature research

Corresponding author(s): Joshua Z. Levin

Last updated by author(s): 02/18/2020

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	The demultiplexed FASTQ files from the sequencing center were shared with us via FTP
Data analysis	The scumi (v0.1.0, available from https://bitbucket.org/jerry00/scumi-dev/src/master/) Python package was used to profile single-cell datasets to generate gene by cell UMI (read) count matrices. We used STAR v2.6.1a to align reads to genomes and featureCounts v1.6.2 to annotate each alignment with a gene tag. RSEM v1.3.0 was used to generate TPM matrices for Smart-seq2 data and bulk RNA-Seq data. HISAT2 v2.0.5 was used to alignment scRNA-seq data from Smart-seq2 and CEL-Seq2. For de-novo transcript construction, we first aligned bulk RNA-Seq data to the genome using HISAT2 v2.0.5 and then used StringTie v0.1.18 for transcript construction. We also used Cell Ranger v2.0.0, Drop-seq pipeline v1.3, the inDrops pipeline (https://github.com/indrops/indrops, accessed on May 23 2018), and CEL-Seq2 pipeline (https://github.com/yanailab/CEL-Seq-pipeline, accessed on May 23 2018) to profile the data from their corresponding platforms. We used Cell Ranger 3.0.2 to profile 10x Chromium (v3) datasets. We used Seurat v2.3.4 R package for clustering analysis and t-SNE dimension reduction, and harmony (https://github.com/immunogenomics/harmony, accessed on Nov 9 2018) to merge datasets from different runs/platforms. Custom R scripts for filtering low-quality cell barcodes, assigning cell types to clusters, and optimizing clustering analysis parameters were available from bitbuckt repo: https://bitbucket.org/jerry00/scumi-dev/src/master/. Trimmomatic v0.39 was used to trim adapter sequences and low quality bases from unmapped reads. Differential expression between methods and between batches was performed with the glm function in R, using the binomial family.

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

Data

Policy information about availability of data

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

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

RNA-Seq (both single cell/nucleus data and bulk) data generated in this project are available from the Gene Expression Omnibus with accession number GSE132044.
Figures 2-6 have associated source data included.
No restrictions on data availability.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We chose the number of cells to profile based on the cell capture rate of lab methods, the frequencies of different cell types in the studied tissue, the costs, and our sense of the sample size of a typical experiment. In each experiment, we aimed to collect data from ~384 cells (or nuclei) for the low-throughput methods and ~3,000 cells (or nuclei) for the high-throughput methods. We aimed for 750,000 to 1,000,000 reads per cell for low-throughput methods and 50,000 to 100,000 reads per cell for high-throughput methods.
Data exclusions	No data was specifically excluded, but filtering was performed as described in the Methods section to exclude cells (cell barcodes) of low quality. In some datasets (cell line mixtures), we used a mixture of two Student's t distribution model to model the read or UMI (log10 transformed) count distributions of each cell, and removed the cells that were likely from the mixture component with fewer reads or UMIs. For other datasets (like PMBCs), we selected more cell barcodes than expected, clustered the data, and removed clusters likely to have resulted from low quality cells. The second strategy was chosen to avoid a bias against cells with lower RNA content. We used a combination of these approaches for the cortex nuclei datasets. In addition, for the standard computational pipeline analyses, we excluded cells using a filter of a minimum number of reads per cell -- see Supplementary Table 3.
Replication	For each sample type, we generated two replicates of data using each lab method. Generally, the results (the ranking of methods) from replicates were consistent.
Randomization	Not relevant to our study. The cells for each replicate were prepared on different days or from different sources. In our comparisons of scRNA-seq methods, we used essentially identical aliquots from the same source of cells. There were no human participants to randomize.
Blinding	Not relevant to our study. We used computational analysis methods that were intended to be unbiased in their evaluation of methods.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293 and NIH3T3 were obtained from the American Type Culture Collection.
Authentication	None of the cell lines were authenticated.
Mycoplasma contamination	Both cell lines were tested for mycoplasma contamination using a PCR-based assay (ATCC, #30-1012K) three days prior to each RNA-Seq experiment and were negative.
Commonly misidentified lines (See ICLAC register)	None of these misidentified lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mus musculus, C57BL/6, male, 1 month old animals.
Wild animals	None used
Field-collected samples	None used
Ethics oversight	All animal-related work was performed under the guidelines of the Division of Comparative Medicine, with the protocol (# 0416-050-1) approved by the Committee for Animal Care of the Massachusetts Institute of Technology, and was consistent with the Guide for Care and Use of Laboratory Animals (1996 edition).

Note that full information on the approval of the study protocol must also be provided in the manuscript.