nature research

Corresponding author(s): Justin Lakkis, Mingyao Li

Last updated by author(s): Jul 22, 2022

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 our Editorial Policies 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 present study does not involve data collection. We used publicly available datasets and an existing dataset generated by our collaborators to evaluate the performance of our method.

Data analysis

totalVI from scvi-tools package: https://scvi-tools.org, version 0.10.0.

Seurat 4: https://satijalab.org/seurat, version 4.1.0.

sciPENN: https://github.com/jlakkis/sciPENN.

All analyses can be reproduced using this repository (https://github.com/jlakkis/sciPENN_codes).
All packages, including sciPENN, can be installed following the instructions in that repository.

Other software requirements:
Python >= 3.7
torch >= 1.6.1
scanpy >= 1.7.1
pandas >= 1.1.5
numpy >= 1.20.1
scipy >= 1.6.1
tqdm >= 4.59.0
anndata >= 0.7.5
numba >= 0.50.0

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 and 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

We analyzed multiple published CITE-seq datasets throughout the evaluations. These data are available as follows (accession numbers provided where possible): 1) Mucosa-Associated Lymphoid Tissue (MALT) dataset (https://www.10xgenomics.com/resources/datasets/10-k-cells-from-a-malt-tumor-gene-expression-and-cell-surface-protein-3-standard-3-0-0); 2) Seurat 4 human peripheral blood mononuclear cells (PBMCs) (https://atlas.fredhutch.org/data/nygc/multimodal/pbmc_multimodal.h5seurat); 3) H1N1 influenza PBMC dataset (https://nih.figshare.com/articles/dataset/CITE-seq_protein-mRNA_single_cell_data_from_high_and_low_vaccine_responders_to_reproduce_Figs_4-6_and_associated_Extended_Data_Figs_/11349761?file=20706645); 4) human monocyte data (https://upenn.box.com/s/64c9fsex50g1bhv67893cpdg9c5jqjzo). 5) Haniffa COVID Dataset (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10026/); 6) Sanger COVID Dataset (https://covid19.cog.sanger.ac.uk/submissions/release2/vento_pbmc_processed.h5ad).

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 didn't perform analysis at the sample level. Instead, we performed analyses at the cell level and our goal is to predict protein expression for each cell. Thus, no statistical methods were used to predetermine sample size. For publicly available data, the sample sizes were determined in the original publications. For the monocyte CITE-seq data, the sample size was determined based on the number of subjects that were available at the time of the CITE-seq experiment.
Data exclusions	All cells and genes are used, no exclusion prior to analysis.
Replication	We did not perform replication. Instead, we confirmed our findings by comparing to other molecular biology results.
Randomization	No data collection was involved in the present study. All blinding information, if available, were described in the original paper for the used datasets.
Blinding	No data collection was involved in the present study. All blinding information, if available, were described in the original paper for the used datasets.

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 and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Antibodies

Antibodies used	This paper analyzed data generated from CITE-seq experiments, which include protein antibodies involved in each experiment. The antibodies were pre-selected by the investigators of the original papers. For the monocyte CITE-seq data, the antibodies were produced by BioLegend Inc (TotalSeq-A antibodies 99787) and are described in the Methods section on "CITE-seq data generation in the monocyte study".
Validation	For the monocyte CITE-seq data, the antibodies were validated by BioLegend Inc.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Volunteers for human monocyte CITE-seq experiment had a median age of 33.5 and ranged from 30-43. 100% of the individuals were white, non-hispanic males.
Recruitment	Enrollment in the study was based on voluntary participation and collection of corresponding clinical information was performed by a clinical coordinator at Columbia University.
Ethics oversight	Columbia University IRB approved the human monocyte study (IRB protocol AAAR5004).

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