nature portfolio

Corresponding author(s): Yadong Wang

Last updated by author(s): Apr 29, 2025

Reporting Summary

Nature Portfolio 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 Portfolio 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	No software was used for data collection
Data analysis	Custom software: https://github.com/YuzhiSun/scInfer Public software:python==3.10.16, scipy==1.15.1,torch==2.2.1+cu121,numpy==1.26.0,scanpy==1.11.0,scikit-learn==1.5.2,pandas==2.2.3

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used in this study are publicly available, and their usages are fully illustrated in Methods. The Specht et al. dataset (transcriptome) was downloaded from the Gene Expression Omnibus with accession number GSM4226877. The Wu et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE176078. The Gupta et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE244390. The Laura et al. dataset was

downloaded from the Gene Expression Omnibus with accession number GSE191246. The Bhupinder et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE161529. The Karin et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE211799. The Khan et al. dataset was downloaded from https://scp.slavovlab.net/Khan_et_al_2023. The Leduc2023 et al. dataset was downloaded from https://scp.slavovlab.net/Leduc_et_al_2023. The Leduc2022 et al. dataset was downloaded from https://scp.slavovlab.net/Leduc_et_al_2022. The Derks et al. dataset was downloaded from https://scp.slavovlab.net/Derks_et_al_2022. The Specht et al.dataset(proteome) was downloaded from https://scp.slavovlab.net/Specht_et_al_2019. The Leduc2021 et al. dataset was downloaded from https://scp.slavovlab.net/Leduc_et_al_2021. The Montalvo et al. dataset was downloaded from https://scp.slavovlab.net/Montalvo_et_al_2023. The nanoSPLITS dataset was downloaded from the github https://github.com/Cajun-data/nanoSPLITS. The He et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE253721. The Nettersheim et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE213282. The Cheng et al. dataset was downloaded from https://zenodo.org/record/6348128. The Mimitou et al. dataset was downloaded from the Gene Expression Omnibus with accession number GSE156478.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

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

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	In this study, we utilized three types of datasets: single-cell transcriptomics, single-cell proteomics, and single-cell transcriptomics plus proteomics data. The detailed sample sizes are as follows:1. Single-cell transcriptomics data: The Specht et al. dataset contains 20,274 cells of monocyte cell line, including two cell types: monocytes and macrophages. The measurement technique is 10× Genomics Chromium platform. The Wu et al. dataset contains 130,246 cells of breast. It is an atlas consists of 26 donors, 8 cell types. The measurement technique is 10× Genomics Chromium platform. The Gupta et al. dataset contains 15,081 cells of mouse pancreas. It includes more than ten cell types from two batches, WT and GKO. The measurement technique is 10× Genomics Chromium platform. The Laura et al. dataset contains 737,280 immune cells of breast from 9 donors. It includes two TNBC cell lines (4T1 and EMT6). The measurement technique is Illumina NovaSeq 6000. The Bhupinder et al. dataset contains over 300,000 cells of breast cancer from 55 donors. It includes more than 6 cell types. The measurement technique is 10× Genomics Chromium platform. The Karin et al. dataset contains over 300,000 cells of mouse pancreas from 56 donors. It includes more than 10 cell types. The measurement technique is Illumina HiSeq 4000 and Illumina NovaSeq 6000. 2. Single-cell proteomics data: The Khan et al. dataset contains 421 cells of MCF-10A cell line, including epithelial, mesenchymal, and epithelial–mesenchymal transition cells and 4096 proteins. The measurement technique is SCoPE2. The Leduc2023 et al. dataset contains 121 cells of THP1, WM989, and CPAF pancreatic cancer cell lines, including three cell types (melanoma, monocyte and PDAC) and 4619 proteins. The measurement techniques are plexDIA and pSCoPE. The Leduc2022 et al. dataset contains 872 cells of U-937, Melanoma, and HPAF-II cell lines, including two cell types and 1543 proteins. The measurement techniques are plexDIA and pSCoPE. The Derks et al. dataset contains 165 cells of U-937, Jurkat, and PDAC cell lines, including three cell types and 1475 proteins. The measurement technique is plexDIA. The Specht et al. dataset contains 1490 cells of U-937 cell line, including both monocytes and macrophages and 3042 proteins. The measurement technique is SCoPE2. The Leduc2021 et al. dataset contains of 163 cells of U-937 cell line, including two cell types and 1647 proteins. The measurement technique is SCoPE2. The Montalvo et al. dataset contains 843 cells of mouse pancreas, including five cell types and 501 proteins. The measurement technique is SCoPE2. 3. Single-cell multiomics data: The nanoSPLITS dataset contains single-cell transcriptomic and proteomics data from 106 cells of human islets, encompassing three cell types, two donors and 1745 proteins. The measurement technique is nanoSPLITS. The He et al. dataset contains single-cell transcriptomic and proteomics data from 43,791 cells of blood, encompassing more than nine cell types, four donors and 307 proteins. The measurement technique is CITE-Seq. The Nettersheim et al. dataset contains single-cell transcriptome and proteomics data from 737,280 cells, encompassing five cell types, four donors and 202 proteins of blood. The measurement technique is CITE-Seq. The Cheng et al. dataset contains single-cell transcriptome and proteomics data from 7,108 cells of blood, encompassing 11 cell types, four donors and 52 proteins. The measurement technique is CITE-Seq. The Mimitou et al. dataset contains single-cell ATAC, transcriptome and proteomics data from 13,384 cells of blood, encompassing 14 cell types, two donors and 210 proteins. The measurement technique is DOGMA-seq. The datasets of He et al. and Nettersheim et al. have a total of 131 overlapping proteins for benchmark testing, and the datasets of Mimitou et al. and Nettersheim et al. have a total of 115 overlapping proteins for benchmark testing.

Data exclusions	No data was excluded from the analysis.
Replication	The data includes omics data used for training and testing, which can be utilized to reproduce the experimental results of this study, and the saved model parameters and prediction results are provided on GitHub
Randomization	Based on the principle of random sampling, we fixed the random seed to facilitate the reproducibility of experimental results.
Blinding	The single-cell data used for training and the testing data come from different donors (data collections).

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks	Not applicable.
Novel plant genotypes	Not applicable.
Authentication	Not applicable.