nature portfolio

Corresponding author(s): Laura Cantini

Last updated by author(s): Jul 15, 2024

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	Only open-source software was used to perform our data analysis. We used scconfluence (v0.1.0), seurat (v4.3.0), rliger (v1.0.0), MultiMAP (v0.0.1), Uniport (v1.2.2), scglue (v0.3.2), scvi-tools (v0.16.4), to perform the benchmark. We used Signac, Cicero and Maestro to compute gene activities. Additionally to the dependencies of scconfluence we used the packages scikit-misc, louvain and scib to compute the benchmark metrics. We also used the NeuroM package to extract and visualize neuron morphologies. Finally, the code we developed and whith which we obtained our results is hosted at https://github.com/cantinilab/scconfluence.

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

- Cell lines. We retrieve a scCAT-seq (RNA+ATAC) dataset with 205 cells from three cancer cell lines (HCT116, HeLa-S3, K562) from Liu et al. Data is available in the Supplementary Materials of the original publication.
- PBMC 10X. We retrieve a 10X Genomics Multiome (RNA+ATAC) dataset available at
https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-no-cell-sorting-10-k-1-standard2-0-0.
- OP Multiome and OP Cite. We retrieve a Multiome (RNA+ATAC) and a Cite-seq bone marrow dataset from the 2021 NeurIPS Open Problems challenge. The GEO accession number is GSE194122 and the data is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122.
- BMCITE. We retrieve a CITE-seq (RNA+ADT) bone marrow dataset from Stuart et al, the GEO accession number is GSE128639 and the data is available at
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128639.
- Smartseq cortex. We retrieve a scRNA-seq mouse somatosensory cortex dataset from Zeisel et al. using scvi-tools's helper
function scvi.data.cortex. The data is available at https://storage.googleapis.com/linnarsson-lab-wwwblobs/blobs/cortex/expression_mRNA_17-Aug-2014.txt.
- smFISH. We retrieve an osmFISH mouse somatosensory cortex dataset from Codeluppi et al. using scvi-tools's helper function scvi.data.cortex. The data is available at
http://linnarssonlab.org/osmFISH/osmFISH_SScortex_mouse_all_cells.loom.
- 3omics RNA. We retrieved a scRNA-seq dataset of PBMCs from the Covid study of Wilk et al. and selected cells from all healthy patients. The GEO accession number is GSE150728 and the data is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150728.
- 3omics ATAC. We retrieve a scATAC-seq dataset of PBMCs and Bone marrow cells from the hematopoietic study of Satpathy et al. in which we select the four batches of PBMCs ("PBMC_Rep1", "PBMC_Rep2", "PBMC_Rep3", "PBMC_Rep4"). The GEO accession number is GSE129785 and the data is available at
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129785.
- 3omics CyTOF. We retrieve a CyTOF dataset of PBMCs from the Covid study of the Combat Consortium in which we select an experimental batch of healthy cells (Batch B). The data is available at 10.5281/zenodo.5139560 under the name "CBD-KEYCYTOF-WB.tar.gz".
- Patch neurons. We retrieve a Patch-seq dataset of mouse primary motor cortex cells published by Scala et al. The scRNA counts are available with GEO accession number GSE163764 at
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163764 and neuronal morphological reconstructions are available at https://download.brainimagelibrary.org/3a/88/3a88a7687ab66069/.

We used 11 different datasets which were chosen based on their quality, previous uses in the literature and availability. They represent a large enough number of biological conditions and measured modalities to provide a thorough study of what our method can accomplish.

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	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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	Every method in the benchmark was run with 5 different initialization random seeds (except Seurat whose outputs don't vary across seeds) to account for the stochasticity of the methods.

Data exclusions	There were no data exclusions.
Replication	All attempts are replication were successful, experimental configurations, scripts and outputs are available at https://github.com/cantinilab/scc_reproducibility.
Randomization	There were no experimental groups.
Blinding	There were no experimental groups but all computational methods were blinded to ground-truth cell type labels/pairing for performance evaluation.

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	N/A
Novel plant genotypes	N/A
Authentication	N/A