nature portfolio

Corresponding author(s): Qi Liu

Last updated by author(s): Feb 5, 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	The custom code for analysis is available at public GitHub repository https://github.com/bm2-lab/SCMMI_Benchmark and https://github.com/bm2-lab/SCMMIB_pipeline . Other packages versions are scanpy (version 1.9.3),Signac (version 1.9.0),Seurat (version 3.2.3),cell ranger ARC pipeline (v2.0.2), NCBI SRA-toolkit.

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

BMMC 10X Multiome and CITE-seq datasets52 analyzed in this study are available at https://openproblems.bio/events/2021-09_neurips/. The raw sequencing files of BMMC Multiome datasets used in this study are available at GEO database with accession number GSE194122. The HSPC 10X Multiome and CITE-seq datasets

are available at https://www.kaggle.com/competitions/open-problems-multimodal/data. The SHARE-seq skin data can be downloaded from GEO database with accession number GSM4156608 and GSM4156597. The COVID19 CITE-seq data is available at E-MTAB-10026 (ArrayExpress). The human WBC CITE-seq data is available at https://atlas.fredhutch.org/nygc/multimodal-pbmc/. The 10X NSCLC CITE-seq, 10X kidney cancer CITE-seq, 10X mouse brain Multiome and 10X PBMC Multiome datasets are downloaded from 10X Genomics website (https://www.10xgenomics.com/datasets/). For spatial multi-omic integration tasks, we obtained SPOTS mouse spleen data from GSE198353(GEO), mouse thymus data from https://zenodo.org/records/10362607, and human lymph node data from https://drive.google.com/drive/folders/1RlU3JmHg_LZM1d-o6QORvykYPoulWWMI. The processed input datasets for all benchmark methods are available at a publicly available Figshare repository (https://figshare.com/projects/Single-cell_multimodal_integration_benchmark_SCMMIB_register_report_Stage_2_study_/221476).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	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	Sample sizes for the publicly available datasets were determined in the original publications.
Data exclusions	No data will be excluded from the analyses.
Replication	There are no experimental findings in this study.
Randomization	To create a mixed batch dataset, we randomly select 10% of cells from each sample batch, ensuring that all cell types within each batch are included, and merge these batches to create a mixed batches dataset for integration analysis.
Blinding	There is no group allocation that the investigators can be blinded to.

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

Methods

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