

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

We list all datasets used in the manuscript. All data analyzed within this manuscript are publicly available. No additional software was used for the data collection process.

Data analysis

We performed the data analysis using our developed SANGO software (Version: 1.0.0) in this paper, which is available at the GitHub repository: [<https://github.com/biomed-AI/SANGO>]. The following software packages were used in our code: python v3.8.17, numpy v1.22.4, scikit_learn v1.2.2, scipy v1.10.1, torch v1.13.1, tqdm v4.65.0, h5py v3.8.0, scanpy v1.9.3, bbknn v1.5.1, imbalanced-learn v0.10.1, leidenalg v0.9.1, pySankey v1.0.0, Louvain v0.8.1, anndata v0.7.8, pandas v1.2.3, matplotlib v3.5.1, Signac v1.11.0, cicero v1.18.0, SNPsea v1.0.3, torch-geometric v2.3.1, BEDtools v2.30.0, SingeR v1.0.0, Seurat v4.2.1, EpiAnno v1.0.0, Cellcano v1.0.4, scNym v0.2.0, scJoint v1.0.0.

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 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 datasets used in this study are publicly available by providing either the GEO accession number or the website where they can be downloaded. (1) The datasets Bone Marrow A, Bone Marrow B, Lung A, Lung B, Kidney, Liver, Heart, Large Intestine A, Large Intestine B, Small Intestine, Whole brain A, Whole brain B, Cerebellum, and Prefrontal cortex are derived from the adult mouse atlas data, downloading from either the GEO access number GSE111586 or the website <http://atlas.gs.washington.edu/mouse-atac/data/>. (2) The anterior datasets (Mos-A1, Mos-A2), middle datasets (Mos-M1, Mos-M2), and posterior datasets (Mos-P1, Mos-P2) are from the different sections of the secondary motor cortex in the mouse brain, which can be accessed through GEO accession number GSE126724. (3) The Mouse Brain (10x) dataset and the normal cortex dataset can be downloaded from https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k and <https://www.10xgenomics.com/resources/datasets/fresh-cortex-from-adult-mouse-brain-p-50-1-standard-1-2-0>, respectively. (4) The forebrain dataset can be downloaded through GEO accession number GSE100033. (5) The PBMC atlas data, the Tumor-Infiltrating Lymphocytes atlas from Basal Cell Carcinoma (BCC_TIL), and the BCC sample data can be accessed through the GEO accession number GSE129785 or the download website <https://www.synapse.org/#!Synapse:syn52559388/files/>. (6) The PBMC (10x) data is obtained from the official 10x website: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k. (7) The Healthy Adult Human Large Atlas (HHLA) can be acquired from the GEO accession number GSE184462 or the website <https://www.synapse.org/#!Synapse:syn52559388/files/>.

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

Our study methodology development and all samples followed previous studies, thus sample size calculations are not needed. All data used in this study were obtained from public resources and were used to demonstrate the capability and functionality of SANGO. The datasets including Bone Marrow A, Bone Marrow B, Lung A, Lung B, Kidney, Liver, Heart, Large Intestine A, Large Intestine B, Small Intestine, Whole brain A, Whole brain B, Cerebellum, and Prefrontal cortex, the anterior (Mos-A1, Mos-A2), middle (Mos-M1, Mos-M2), posterior (Mos-P1, Mos-P2), and Mouse Brain (10x) were determined to quantitatively evaluate the performance of our method. This evaluation covered diverse scenarios, including datasets across samples, platforms, and tissues, as well as multi-source reference datasets. The forebrain dataset was determined to quantitatively evaluate the performance of our method in coarse-grained cell type annotation tasks. The PBMC atlas data and the PBMC (10x) data were determined to evaluate the ability of our method on large datasets. The normal cortex data, the Tumor-Infiltrating Lymphocytes atlas from Basal Cell Carcinoma (BCC_TIL), the BCC sample data, and the Healthy Adult Human Large Atlas (HHLA) were used as application cases to test the ability to retain biological meaning and identify unknown cell types and multi-level cell types.

These datasets have been utilized in previous studies such as EpiAnno [1] and Cellcano [2], indicating that they are considered to have appropriate sample sizes.

[1] Chen X, Chen S, Song S, et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding[J]. Nature Machine Intelligence, 2022, 4(2): 116-126.

[2] Ma W, Lu J, Wu H. Cellcano: supervised cell type identification for single cell ATAC-seq data[J]. Nature Communications, 2023, 14(1): 1864.

Data exclusions	We exactly followed previous studies to pre-process data. We performed quality control of scATAC-seq data based on the common used and pre-established criteria in this field
Replication	Due to potential slight variations in outcomes when the random seed is not fixed during the execution of the deep learning model, we conducted ten runs to assess its robustness.
Randomization	The conclusions have been consistently proved on multiple datasets and thus the conclusion is not affected by the randomization.
Blinding	Since all the data used in this study have been previously published, it is not feasible to conduct blind investigations during the reanalysis of the data. However, the allocation information is blinded to the computational algorithms during the 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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging