

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	RareQ is open source and publicly available at GitHub (https://github.com/xiaolab-xjtu/RareQ) and Zenodo (https://zenodo.org/records/17190972). Other tools and packages used in this paper include: R 4.1.3, Seurat 4.3.0, Signac 1.9.0, scan 1.22.1, aricode 1.0.2, SeuratData 0.2.2.9002, python 3.8.15, Scanpy 1.9.1, Pandas 1.4.2, numpy 1.22.3, anndata 0.8.0, h5py 3.11.0, FiRE (v1.0.1, from GitHub: princethewinner/FiRE), GapClust (v0.1.0, from GitHub: fabotao/GapClust/), CellSIUS (v1.0.0, from GitHub: Novartis/CellSIUS), RaceID (v0.3.5, from CRAN), GiniClust2 (v2.0, from GitHub: dtsoucas/GiniClust2), EDGE (v1.0, from GitHub: shawnstat/EDGE), scCAD (v1.0.0, from GitHub: xuy-p-csu/scCAD), MarsGT (v1.0.0, from GitHub: OSU-BMBL/MarsGT).

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](#)

The 20 RNA-seq datasets for benchmarking in this study were obtained from various public websites under accession codes provided in Supplementary Table 2, including NCBI Gene Expression Omnibus (GEO) [<https://www.ncbi.nlm.nih.gov/geo/>], Cell Blast [<https://cblast.gao-lab.org/>], Gut Cell Survey [<https://www.gutcellatlas.org/>], 10X genomics [<https://www.10xgenomics.com/datasets>]. The 10 paired scRNA-seq and scATAC-seq datasets for benchmarking in this study were obtained from various public websites under accession codes provided in Supplementary Table 5, including scMMO-atlas [<https://www.biosino.org/scMMO-atlas/>], scglue example datasets [<https://scglue.readthedocs.io/en/latest/data.html>] and GEO. Human PBMC paired scRNA-seq and scATAC-seq data, 1.4M mouse brain data, Alzheimer's disease data and mouse hippocampal data were obtained from GEO with accession codes GSE194122, GSE212606, GSE303823 and GSE261148. CITE-seq data of human bone marrow mononuclear cells was downloaded with SeuratData package (dataset identifier bmcite). The paired B lymphoma scRNA-seq and scATAC-seq dataset, Xenium spatial datasets of mouse brain and human RCC were obtained from the 10X Genomics website. Mouse brain spatial transcriptomic data on the EEL FISH platform was obtained from the Mouse Brain Atlas [<http://mousebrain.org/adult/downloads.html>]. Bulk AD RNA-seq data was downloaded from GEO with accession code GSE109887. Source data are provided with this paper.

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

Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.

Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.

Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.

Reporting on race, ethnicity, or other socially relevant groupings

Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).

Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)

Please provide details about how you controlled for confounding variables in your analyses.

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

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](https://www.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 analyzed a total of 24 publicly available real single-cell RNA-seq datasets, 15 paired single-cell RNA-seq and ATAC-seq data, 1 CITE-seq data, 3 spatial transcriptomics data in benchmarking and case studies. The number of cells in these datasets ranges from 4,007 to 1.47 million. These datasets we collected commonly used in single-cell clustering analysis methods, as well as a majority of the data analyzed by the comparison methods. These datasets vary in size and represent a broad range of biological contexts, including worm neuron, mouse airway, brain, retina, heart, kidney, human PBMCs, bone marrow, pancreas, liver, immunology data, and clear cell renal cell carcinoma. We believe these datasets are sufficient for the analysis.

Data exclusions	All cells and genes of each single-cell RNA seq dataset were used, and no exclusion was done prior to analysis.
Replication	We performed data shuffling for 30 times to demonstrate the robustness of RareQ. All computational replications for the same dataset were successful using the default parameters from the Github.
Randomization	Sample randomization is not relevant to this study since each data was analyzed separately. The analyses reported in this paper do not involve between-sample comparison.
Blinding	Blinding is not relevant since no data collection was involved in this study. Data information was provided alone with the original publication.

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	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.