nature research

Corresponding author(s): Ge Gao

Last updated by author(s): Jun 16, 2020

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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	Code used for data collection is deposited at https://github.com/gao-lab/Cell_BLAST.
Data analysis	Code and environment configuration necessary for reproducing analysis results are deposited at https://github.com/gao-lab/Cell_BLAST. Specific software used include R packages: irlba (v2.3.3), Rtsne (v0.15), zinbwave (v1.6.0), sva (v3.32.1), scran (v1.6.9), Seurat (v2.3.3 and v3.0.2), harmony (v1.0), scmap (v1.6.0); Julia package: CellFishing.jl (v0.3.0); Python packages: gseapy (v0.9.16), umap-learn (v0.3.8), scPhere (v0.1.0), ZIFA (v0.1), Dhaka (v0.1), DCA (v0.2.2), scVI (v0.2.3), scScope (v0.1.5), SAUCIE (commit c2e59683ddf401f07d4c226a420b367181934715), Cell-BLAST (v0.3.7).

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

All scRNA-seq datasets used in this study were obtained from public data repositories, with detailed information including accession codes and URLs available in Supplementary Data 2. Source data for the benchmark experiments are available in Supplementary Data 4. Curated datasets in ACA are available through our Web portal https://cblast.gao-lab.org/download.

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	All computational experiments were repeated multiple times using different random initializations (N values indicated in figure legends), to evaluate algorithmic stability.
Data exclusions	No data were excluded from analysis.
Replication	All computational experiments were assembled using Snakemake and environment configuration files are provided to ensure exact reproducibility. All attempts at replication were successful.
Randomization	Not applicable since no biological experiment was involved.
Blinding	Not applicable since no biological experiment was involved.

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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