nature portfolio

Corresponding author(s): James Zou

Last updated by author(s): Jul 20, 2026

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 in this study. All data analysed were publicly available from previously published sources, including GTEx, 10x Genomics, Zenodo, the ADHD GWAS dataset, the MPRA-coupled scCRISPRi data for cis-regulatory element perturbation, and CD4+ T-cells perturb-seq data
Data analysis	Data analysis was performed using Python v3.12.1, Paper2Agent (17 September 2025; commit 37ecd69), AlphaGenome v0.2.0, TISSUE (10 March 2024; commit ffc3599), and Scanpy v1.11.4. All software is open source and publicly available through the corresponding GitHub repositories.

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

No new datasets were generated in this study. All datasets analysed are publicly available from the following sources:
10x Genomics single-cell RNA-seq datasets:
http://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_1k_v2/pbmc_1k_v2_filtered_feature_bc_matrix.h5
http://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_1k_v3/pbmc_1k_v3_filtered_feature_bc_matrix.h5
http://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_1k_protein_v3/pbmc_1k_protein_v3_filtered_feature_bc_matrix.h5
http://cf.10xgenomics.com/samples/cell-exp/6.0.0/Brain_Tumor_3p_LT/Brain_Tumor_3p_LT_filtered_feature_bc_matrix.h5
http://cf.10xgenomics.com/samples/cell-exp/3.0.0/neuron_1k_v3/neuron_1k_v3_filtered_feature_bc_matrix.h5
http://cf.10xgenomics.com/samples/cell-exp/3.0.0/neuron_10k_v3/neuron_10k_v3_filtered_feature_bc_matrix.h5
http://cf.10xgenomics.com/samples/cell-exp/3.0.0/heart_1k_v3/heart_1k_v3_filtered_feature_bc_matrix.h5
Mouse somatosensory-cortex spatial transcriptomics data (Dataset 15):
https://zenodo.org/records/8259942
GTEx eQTL data for chr1:109274968:G>T:
https://gtexportal.org/home/snp/chr1_109274968_G_T_b38
ADHD GWAS summary statistics:
https://www.ebi.ac.uk/gwas/studies/GCST90568440
https://www.ebi.ac.uk/gwas/studies/GCST90568441
CD4+ T-cell Perturb-seq data:
https://virtualcellmodels.cziscience.com/dataset/genome-scale-tcell-perturb-seq
MPRA-coupled scCRISPRi data for cis-regulatory-element perturbation:
https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-025-02301-3/MediaObjects/41588_2025_2301_MOESM4_ESM.xlsx

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	This study did not involve the collection or analysis of individual-level human participant data. Therefore, sex and gender variables were not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	No human participant or socially categorized data were used in this study. All analyses were performed on publicly available, de-identified genomic and transcriptomic datasets (e.g., GTEx, 10x Genomics, and ADHD GWAS summary statistics).
Population characteristics	Not applicable. The study analyzed only publicly available, aggregated datasets without individual-level identifiers or demographic variables.
Recruitment	Not applicable. No participants were recruited; all data were obtained from previously published, publicly accessible sources.
Ethics oversight	Not applicable. The study used only publicly available, de-identified datasets and did not involve any new human data collection. Ethics approvals for the original datasets (e.g., GTEx, 10x Genomics, ADHD GWAS) are described in their respective publications.

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	The number of datasets was not determined by a statistical power calculation because this study evaluated software functionality rather than

Sample size	biological effect sizes. Public datasets were selected to cover each prespecified case study and provide diverse data modalities, tissues, dataset sizes and reference outputs. They were considered sufficient because they enabled evaluation of every intended workflow and comparison with human-executed or published reference results. No datasets were excluded based on the observed outcomes.
Data exclusions	No data were excluded. All publicly available datasets were used in their entirety as provided by their original sources.
Replication	Stochastic agent evaluations were performed in five independent runs per method. The AlphaGenome benchmark, TISSUE reproducibility analysis, non-biology benchmark and repaired-server evaluations were each repeated five times. The Scanpy workflow was evaluated once on each of 11 independent datasets (four primary and seven additional datasets) against human-executed references. The monolithic-agent ablation was repeated three times. Deterministic POP-TOOLS validation was performed once for each of five analysis tasks because the outputs were byte-for-byte identical to the human-executed references.
Randomization	Not applicable. This study did not involve experimental groups or participant assignment. All analyses were computational and based on existing datasets
Blinding	Blinding was not applicable to dataset selection or automated computational analyses because no participants, treatment groups or experimental allocations were involved. Benchmark responses were assessed against predefined ground-truth answers and scoring rubrics. Two human experts graded the responses independently; formal blinding to system identity was not feasible because the response formats and execution traces could reveal the evaluated system. Grading discrepancies were resolved by consensus.

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	Not applicable. This study did not involve the use of any plant materials or seed stocks. All analyses were computational and performed on existing genomic and transcriptomic datasets.
Novel plant genotypes	Not applicable. No novel plant genotypes were generated in this study. The work focused entirely on developing and evaluating the Paper2Agent computational framework.
Authentication	Not applicable. No plant genotypes or biological materials were used. All results are based on publicly available datasets analyzed using the Paper2Agent system.