

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

For data collection, we have used the following software tools:

- sra-tools (v3.4.1)
- bamtofastq (v1.4.1)
- dnaio (v1.2.4)

Data analysis

For the analysis of data, we have used the following software tools:

- Python (v3.11)
- SQLite (v3.49.1)
- snakemake (v7.32.4)
- MetaGraph (v0.4.1)
- Fulgor (v4.0.0)
- REINDEER (v1.4.7-1-g0812ad128)
- bcalm2 (v2.3.0)
- STAR (v2.7.11b)
- kraken2 (v2.1.5)
- kallisto-bustools (v0.28.2)
- BLAST+ (v2.16.0)
- scanpy (v1.11.4)
- leidenalg (v0.10.2)
- SPAdes (v4.2.0)
- minimap2 (v2.1.1-r341)

- numpy (v2.2.6)
- pandas (v2.3.2)
- Flask (v3.1.0)
- polars (v1.27.1)
- text2term (v4.5.0)
- ollama (v0.13.4)
- flexiplex (v1.02.3)
- dustmasker (v1.0.0)
- malvacrawl (v1.0.0)

All code for the Malva indexing framework is publicly available on GitHub:
<https://github.com/malva-bio/malva> (v1.0.0)
https://github.com/malva-bio/malva_client

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](#)

Malva Index is publicly accessible through the web interface at <https://malva.bio> and programmatic API at <https://github.com/malva-bio/malva-client>, requiring only ORCID or token-based authentication. The current human Index (v2025.04a) comprises ~51 million cells from 592 studies across 7,966 samples, with an additional ~10 million mouse cells, all derived from publicly available datasets in NCBI SRA, ENA, CNGBdb, and Human Cell Atlas Data Portal with their original accession numbers preserved in metadata. All source datasets used for benchmarking and validation are publicly available with accession numbers provided in the Methods section, including specific examples such as the Stereo-seq mouse liver atlas (CNGBdb: CNP0003447), mixed cell line dataset (SRA: SRR10971813), and various 10x Genomics reference datasets available through their website. No restrictions apply to data access beyond standard repository terms of use for the underlying public datasets. Source code, documentation, and binaries for custom index construction are open and freely available for academic research purposes at <https://github.com/malva-bio/malva>. This code enables users to build and query local indices on their own data.

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

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical power calculations were performed as this study developed a computational platform rather than testing specific hypotheses. Sample sizes were determined by the availability of public datasets, resulting in ~60 million cells from thousands of experiments across major repositories, which provides sufficient statistical power for the platform validation analyses and population-scale genetic variant detection described in the study.
Data exclusions	Quality control filters were applied including removal of sequence probes with extreme GC content (<30% or >70%) for variant analysis, and exclusion of cells with insufficient coverage (e.g., <500 UMIs) for saturation-based correction analyses. Low-complexity and repetitive k-mers were masked using standard tools (Dustmasker) to reduce false positives, as described in the Methods section.
Replication	We have replicated the key points of our manuscript (pseudoquantification accuracy, sensitivity, specificity and indexing/querying speed) across four individual datasets (Open-ST, Visium, and 10x 3' scRNA-seq), as well as two large-scale datasets, one containing samples crawled from public repositories, and another added during peer review (scTML).
Randomization	Randomization is not applicable to this computational platform development study. Datasets were selected based on availability in public repositories and technical criteria (sequencing technology, data quality) rather than experimental group allocation. For validation analyses, datasets were chosen to represent diverse biological conditions, technologies, and species to ensure broad applicability of the platform.
Blinding	This is not applicable as no subjects were recruited, and only public data was downloaded. The datasets were either selected at random, or with a filtering performed based on date (e.g., up to May 2025), or based on a specific technology (e.g., 10x 3' v3 single-cell kit).

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	n/a
Novel plant genotypes	n/a
Authentication	n/a