

Ultrafast and reference-free sequence discovery in single-cell data

<https://doi.org/10.1038/s41586-026-10975-w>

Received: 17 September 2025

Accepted: 30 July 2026

Published online: 26 August 2026

Open access

 Check for updates

Daniel León-Perián¹✉, Nikos Karaiskos¹✉ & Nikolaus Rajewsky^{1,2,3,4,5,6,7,8,9}✉

Knowledge of RNA sequences, expression, splicing, isoforms, structure and modifications is central for understanding and targeting cellular processes. Revolutionary single-cell and spatial transcriptomics technologies—for example, as deployed by consortia such as the Human Cell Atlas—partially capture this diversity and generate cellular profiles that expand at petabyte scale each year^{1–5}. Yet researchers cannot search sequences across these datasets: standard pipelines do not scale or rely on references, retaining only gene or isoform counts, whereas accessing raw sequences requires collecting, downloading and processing millions of large files. Here we present Malva, a computational platform that enables ultrafast, species-agnostic and reference-free interrogation of the raw sequence space, enabling searching for any sequence, mutation, splice junction or pathogen, or spatial location of arbitrary transcripts. The continuously expanding Malva Index currently comprises around 74 million cells from thousands of experiments in health and disease. Malva enables reference-free discovery—researchers can, for example, identify cell types and predict cell–cell similarity directly from sequence composition. Building on Malva’s speed and accuracy, we demonstrate how Malva can be flexibly connected to state-of-the-art neural networks and how to execute complex searches and enable automated analyses. Malva transforms single-cell atlases from static gene count tables into dynamic, sequence-resolved resources that may help to bridge human–machine reasoning about biology.

RNA is of fundamental importance for life. Knowledge of RNA sequences, RNA expression, splicing, isoforms, structure and modifications are central for understanding and targeting cellular processes. In the past decade, new sequencing technologies have made possible RNA sequencing (RNA-seq) in single, dissociated cells and subsequently for cells within whole tissue slices. Since cells are the basic units of life, these technologies have transformed life sciences, including medical and clinical research and applications, and annually generate petabytes of sequence information from hundreds of millions of cells^{1,6–9}. The Human Cell Atlas consortium systematically quantifies RNA in healthy human cells^{2,3,10}, for instance, while researchers also apply these technologies to diseases¹¹. Moreover, vast amounts of data are generated for model systems, including organoids, and will be applied to investigate biodiversity on Earth as a whole^{12,13}.

It is therefore of utmost importance to have tools that can both integrate these fragmented datasets into a unified resource and make them efficiently interrogable. Thirty-five years ago, BLAST pioneered this principle for sequence databases, transforming biological research by making it possible to locate a nucleotide sequence of interest in a vast collection of genomic sequences—critically, not only with accuracy but also with unprecedented efficiency¹⁴. These design principles made it broadly usable and ultimately indispensable across the life sciences.

Yet, no existing computational tool—such as cell atlases (CELLxGENE, STOmicsDB, Single Cell Expression Atlas and others^{15–21})—can empower researchers to efficiently search for specific RNA sequences across the recently created vast single-cell, single-nuclei or spatial transcriptomics (SC/ST) data. At best, these platforms provide pre-computed gene counts that reflect the aggregated number of RNA sequences mapped to a reference. Therefore, critical information is missing: about transcripts that do not map to the reference (for example, transcripts generated by genetic lesions in cancer cells or transcripts that do not map to the reference haplotypes), RNA sequence editing, mutations, pathogens present in the cells and more.

Standard single-cell and single-nuclei pipelines remain reference-centric: they process sequencing reads by aligning them to a user-provided reference^{22–26}. Applying these approaches at atlas scale requires downloading and reprocessing petabytes of raw files, which is beyond the computational resources available to most laboratories. By contrast, bulk, sample-centric methods that index short sub-sequences (*k*-mers) directly from the data treat the dataset itself as the searchable unit rather than forcing every read through a reference. They enable ultra-rapid search across large sequencing collections (MetaGraph, REINDEER, Sequence Bloom Tree, Mantis, Fulgor, Themisto and Logan, among others^{27–33}). These approaches have already proved to

¹Laboratory for Systems Biology of Regulatory Elements, Berlin Institute for Medical Systems Biology (MDC-BIMSB), Max-Delbrück-Centrum for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. ²Charité-Universitätsmedizin, Berlin, Germany. ³German Center for Cardiovascular Research (DZHK), Berlin, Germany. ⁴NeuroCure Cluster of Excellence, Berlin, Germany. ⁵German Cancer Consortium (DKTK), Berlin, Germany. ⁶National Center for Tumor Diseases (NCT), Berlin, Germany. ⁷ImmunoPreCept Cluster of Excellence, Berlin, Germany. ⁸Einstein Center for Early Disease Interception, Berlin, Germany. ⁹German Center for Child and Adolescent Health (DZKJ), Berlin, Germany. ✉e-mail: daniel.leonperinan@mdc-berlin.de; nikolaos.karaiskos@mdc-berlin.de; rajewsky@mdc-berlin.de

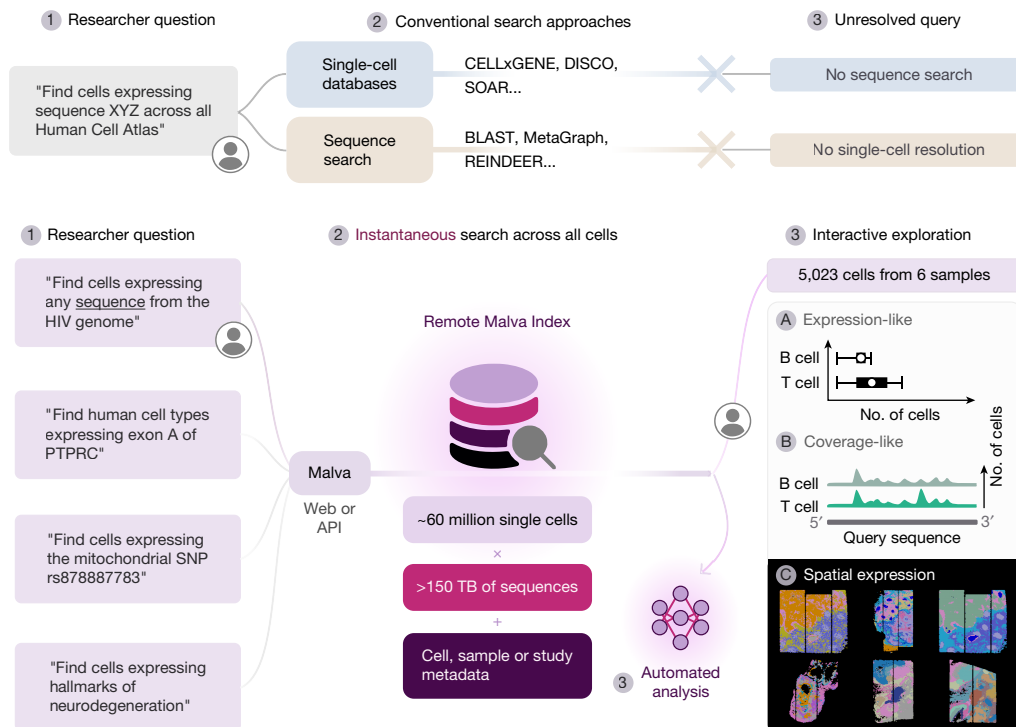


Fig. 1 | Malva enables instantaneous sequence-based queries across single-cell atlases. Conventional single-cell portals lack sequence-level resolution and existing sequence search tools lack single-cell resolution. Malva integrates these tools, providing real-time results at atlas scale. Researchers can query Malva Index using raw nucleotide sequences, gene symbols or natural language descriptions. Malva resolves these inputs to their underlying sequences and returns all matching cells together with rich metadata at both

the cell and sample level (for example, cell types, disease status, sex, age and study identifiers). Results can be fed into computational systems for automated analysis (for example, neural networks) or analysed interactively, for example as expression distributions across cell types (A), coverage profiles along the queried sequence queries (B) or spatial maps when coordinates are available (C).

be very useful: several studies have discovered new plastic-degrading enzymes³³, novel RNAs that shape tumour identity^{34–36}, sequence variation across large cancer cohorts³⁷, disease-associated sequences invisible to alignment-based detection³⁸, functionally related long non-coding RNAs (lncRNAs) that share *k*-mer patterns despite lacking sequence homology³⁹, and even new forms of life⁴⁰. However, these tools were designed for collections of bulk samples (hundreds to hundreds of thousands of genomes or experiments). Their construction typically requires a compacted de Bruijn graph built from the union of all input data, an operation that must be at least partially repeated when new data are added. Annotation is stored either as a label-wide matrix or as inverted lists whose size increases with total label count; in both cases, efficient queries generally require the index to reside in memory, which at single-cell scale (tens to hundreds of millions of individual cells) would demand terabytes of RAM. These properties make existing tools impractical for this regime.

Here we introduce Malva (<https://malva.mdc-berlin.de>), a computational platform that unifies SC/ST data and makes them efficiently searchable at nucleotide resolution. Malva comprises a sequence search algorithm, a continuously expanding searchable corpus (Malva Index) and a public application programming interface (API); just as the success of BLAST rested on speed and accuracy, Malva makes sequence search and analyses practical at atlas scale.

Malva search can start from nucleotide sequences, gene identifiers or natural language prompts. This enables analyses on demand across millions of cells in seconds, from viral genomes and germline variants to isoform usage. Beyond search, Malva can perform reference-free cell-type clustering^{41–43} and de novo assembly of cell-type-specific sequences.

Malva does not replace reference-based quantification pipelines, which remain suitable for gene-level analysis; rather, it extends them

by enabling sequence-level queries that are otherwise inaccessible at very large scale, which is also critical in scenarios where references are not readily available¹³. To our knowledge, Malva is the first platform for large-scale reference-free sequence search and analysis at single-cell resolution.

Fast sequence search at single-cell resolution

Enabling sequence queries across all public single-cell data, beyond pre-computed gene counts, requires addressing three fundamental challenges. First, preserving nucleotide-level resolution across massive datasets means that traditional alignment pipelines, which discard sequence information, are insufficient. Second, comparing data across studies demands harmonization and standardized metadata that bridge different technologies and processing methods. Third, even with aggressive compression, these data are too large to download locally, requiring decentralized access via an API over the Internet.

To address all three challenges, we developed Malva (Fig. 1). Malva is a platform that comprises a new search algorithm, a continuously expanding searchable corpus (Malva Index), and a public API. Malva Index is built on thousands of public datasets and is accessible via the API for both interactive and programmatic access.

From a user's perspective, Malva transforms how researchers explore transcriptomic data. Traditional workflows of a researcher interested in circular RNA (circRNA) expression, for instance, require downloading entire atlases, running specialized circRNA detection pipelines and aggregating results across samples—a process that requires days of computation and specialized expertise. With Malva, searching for a circRNA across millions of cells takes seconds. In general, users can instantly query along four dimensions: the feature that they seek (a raw sequence, gene or pathway), where to search (filters such as tissue,

Table 1 | Example questions supported by Malva and what they return

Mode	User question	Input to Malva Index	Output
Expression-like	TTCCAGTGTCTGCAATATCCAGGGTTCCGATGGCACCTGTG viruses.fa TargetScan.fa dbSNP.fa Which cells carry this KRAS mutation? Where are BCR–ABL fusion reads detected in haematopoietic datasets? Which glioblastoma cells carry the EGFRvIII deletion? Locate Epstein–Barr virus reads in nasopharyngeal carcinoma data Retrieve all cardiomyocytes that express the lncRNA <i>Xist</i>	An interesting backsplice sequence A FASTA file with viral genomes A FASTA file with microRNA targets A FASTA file with common SNPs Short probe with the mutation Short probe for the BCR–ABL junction Short probe spanning the break point Epstein–Barr virus reference genome <i>Xist</i> gene sequence	List of cells or spatial coordinates with at least one match, counts per cell and metadata (cell type, tissue and technology) for aggregation of results (for example, expression level per study, per tissue, per organ, per disease or combinations of these)
Coverage-like	What immune cells include exon A of <i>Ptpnc</i> ? Does 3' UTR lengthen during mouse development? Do cardiomyocytes retain intron 4 of <i>Ttn</i> during post-natal maturation? Compare 5' UTR shortening of <i>Myc</i> between fetal and adult liver	Chromosomal coordinates or an annotated feature Chromosomal coordinates of a 3' UTR Genome coordinates for <i>Ttn</i> 5' UTR coordinates of <i>Myc</i>	Base-level coverage, aggregated by metadata selected by the user
Textual query	Show cells expressing iron-uptake genes in liver atlases Show regions with high hypoxia-response gene activity in solid tumours Highlight neurons expressing markers of neurodegeneration	Free-text prompt (for example, pathway, phenotype or drug target)	Large language model-generated sequence list and single-cell or spatial expression map, plus metadata for aggregation of results

disease state or protocol), how to measure it (base-level coverage or aggregated pseudocounts) and scale (individual cells, whole libraries or aggregated groups). These capabilities are demonstrated through three complementary search modes.

First, researchers can provide a nucleotide sequence (for example, ranging from a short 24-mer splice probe to a full viral genome) and immediately retrieve the cells that contain it. Second, they can specify a genomic interval and view live coverage tracks coloured by lineage or tissue. Third, they can type natural language queries, such as “cells expressing high levels of mitochondrial genes in tumor samples”, which an integrated agent translates to structured searches (Table 1 and Supplementary Fig. 1).

Indexing and query performance

Malva’s search algorithm borrows the *k*-mer indexing principle from bulk sequence search and adapts it to single-cell data: each read is decomposed into *k*-mers (sub-sequences of length *k*) stored together with the cell barcode of origin, enabling constant-time lookups without alignment (Extended Data Fig. 1 and Methods). To overcome the scalability limitations of existing *k*-mer indices at single-cell scale (Supplementary Note 1), Malva makes three design choices: non-overlapping *k*-mer sampling eliminates de Bruijn graph construction; per-*k*-mer sparse inverted lists store only the expressing cell identifiers, decoupling query cost from total index size; and independent per-sample indexing with incremental merging yields an on-disk layout that supports memory-efficient queries and continuous corpus expansion (Supplementary Methods).

To begin, we assessed indexing performance on a Stereo-seq mouse liver atlas containing about 61 billion reads across 20 sections, each with more than 3 billion reads and hundreds of millions of spatial barcodes. Malva completed indexing in 24 central processing unit (CPU) hours, between 4 and 25 times faster than other preprocessing algorithms (that is, *k*-mer indexing and reference mapping), with a peak memory footprint of 8 GB, whereas all other methods were benchmarked without memory constraints and required substantially more resources (Extended Data Fig. 2a and Supplementary Table 1). Indexing speed saturates gzip decompression of FASTQ files, an unavoidable input–output

bottleneck; memory usage scales linearly with the number of unique cell barcodes rather than with sequencing depth, and index sizes scale with sequencing depth per cell rather than cell count or total unique *k*-mers (Extended Data Fig. 2b,c and Methods).

Next, we validated that query performance enabled real-time analysis. When searching for a sequence, Malva tiles it into *k*-mers and counts exact matches per cell (Supplementary Fig. 2c). These raw match totals, termed pseudocounts, emerge directly from *k*-mer matches without alignment (guidelines for parameter choice provided in Supplementary Tables 2 and 3 and Supplementary Note 2). Search latency depends only on disk access: 70 ms for a single *k*-mer, 0.9 s for a 1 kb transcript and approximately 1 min for 1,000 transcripts on a single CPU core, all computed on demand rather than from pre-computed gene matrices (Extended Data Fig. 2d). Other methods had worse time or memory performance, and traditional pipelines would have required read realignment for every new search query (Supplementary Note 1). We additionally selected a large collection of single-cell datasets (about 1.5 TB of raw data) and reindexed them with Malva and Fulgor, the two methods with the best runtime and memory usage; Malva achieved faster indexing with a smaller memory footprint while maintaining comparable query throughput from disk (Supplementary Table 1).

Third, we validated quantification accuracy of Malva on datasets covering different SC/ST technologies, tissue types and disease states (Extended Data Fig. 3 and Methods). Across datasets, pseudocounts showed strong correlation with reference counts at both cellular (Extended Data Fig. 3a) and gene (Extended Data Fig. 3b) levels. We note that pseudocounts are not corrected for unique molecular identifiers (UMIs) and therefore correlate with read counts rather than molecule counts; users should consider these as semi-quantitative measures (Discussion). Systematic offset observed across single-cell counts can be corrected through library-specific normalization based on sequencing saturation (Extended Data Fig. 4 and Supplementary Note 3), analogously to approaches for bulk transcriptomics³⁷. This correction addresses first-moment differences (that is, mean count inflation due to PCR amplification) but does not account for gene- or sequence-specific amplification biases.

Having established quantitative concordance, we assessed whether Malva preserves biological signatures that distinguish cell types.

Using the original cell-type annotations for each dataset, differential expression analysis showed that cell-type-specific enrichment across genes correlated strongly between Malva and reference methods across all datasets (Extended Data Fig. 3c). Moreover, clusters from pseudocounts preserved spatial organization in tissue sections (Extended Data Fig. 3d) and cell-type separation in single-cell data (Extended Data Fig. 3e). Additionally, developmental trajectory analysis on the embryonic day 13.5 (E13.5) mouse midbrain showed that pseudotime inference using pseudocounts closely recapitulated trajectories derived from reference counts (Extended Data Fig. 3f).

We further benchmarked Malva's sequence search accuracy by its ability to identify cell lines through variant detection. The benchmark contains a mixture of seven cell lines and was originally developed to test Flexiplex, a sequence search tool⁴⁴. Malva indexed the more than 200 million reads 4 times faster than Flexiplex's total search time, then resolved each query in milliseconds (Extended Data Fig. 5a). Across all target sequences, Malva maintained both high sensitivity (ability to detect true positives) and specificity (ability to avoid false positives) (Extended Data Fig. 5b–d). Because Malva indexes non-overlapping k -mers, sensitivity for short probes (length = k) depends on alignment with indexed positions, but the sliding window query algorithm guarantees detection for probes longer than k (Extended Data Fig. 5e,f and Supplementary Note 2). Systematic evaluation across k -mer sizes and probe lengths confirmed that our default parameters operate within the optimal detection regime (Supplementary Fig. 2); small deviations from perfect sensitivity reflect methodological differences with Flexiplex, which permits Hamming distance 1 mismatches that exact k -mer matching excludes.

We additionally report that Malva accurately detected tumour driver mutations on a small cohort of lung cancer patients and matched normal samples (Extended Data Fig. 6), and it is robust to sequencing errors; we benchmarked this by simulating error rates from 0.1% to 10% across benchmark datasets. Sensitivity, specificity, pseudocount correlations and downstream analysis results (for example, cell-type clustering) remained stable up to error rates of 2% and degraded gradually up to very high (10%) error rates (Supplementary Fig. 3).

In summary, Malva achieved faster indexing and querying times than existing methods while maintaining a low memory footprint, and showed high concordance with traditional pipelines for expression quantification.

Capturing cellular diversity at scale

Having validated technical performance, we constructed Malva Index by compiling more than 140 TB of public SC/ST data into a searchable resource, in about 70 h of computational wall time (around 9,700 h of CPU time). The build of the human Index at the time of analysis (v2025.04a) encodes more than 100 billion total unique 24-mers from around 51 million cells (592 studies, 7,966 samples) and an additional approximately 10 million cells from mouse samples.

We achieved this by implementing a crawler that continuously collects FASTQ files mainly from the Human Cell Atlas (HCA) Data Portal that is also compatible with the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO), European Nucleotide Archive (ENA) and CNGBdb data portals, extracts barcodes or spatial coordinates across major protocols, indexes and merges each read into Malva Index, a single harmonized container (Fig. 2a and Methods). As of May 2025, Malva Index covered around 80% of the Human Cell Atlas, or around 20% of human, single-cell runs from the SRA. Malva Index has less than 10% the storage footprint of source files, a critical reduction given the continuous expansion of these publicly archived data (Fig. 2b). Malva also maintains individual sample-level indices, enabling local queries when global access is unnecessary. Cell barcodes are matched exactly against technology-specific whitelists and filtered to 'true' cells via an

automated approach (Methods); barcode error correction and doublet removal are omitted to maintain high processing throughput at atlas scale (Discussion).

Malva Index spans embryonic, adult and pathological human material from all major organ systems, with a diverse representation of cell types, donor ages, male–female ratio, disease status and underlying technology, represented in a curated database of sample- and cell-level annotations (Fig. 2c,d and Supplementary Fig. 4a–c).

Notably, we validated that querying remains ultra-rapid on the complete Index. Despite containing tens of billions of k -mers across millions of cells, Malva Index resolves short sequence probes (such as 24-mers) in milliseconds, and returns full-length cDNAs in seconds and catalogues of thousands of transcripts in minutes (Extended Data Fig. 7a). This sub-linear scaling in query time ensures scalability and strengthens future proofness.

Together, Malva Index enables interactive exploration across a continuously expanding, vast corpus of SC/ST data. Users can interrogate Malva Index via the public API, which is freely accessible with ORCID accounts linked to academic users.

Enabling queries beyond current portals

Current atlas-scale single-cell portals excel at interrogating counts from predefined reference genes. Thus, interrogation of unexpected (for example, bacterial, viral, fungal or synthetic) sequences, mutations, isoforms or regulatory variants is not possible without downloading and re-analysing petabytes of data. Malva enables these via instantaneous sequence search (Fig. 3 and Extended Data Fig. 7).

Screening for arbitrary sequences across Malva Index

Upon providing a nucleotide sequence, Malva returns the cells that express it, regardless of whether it is part of a predefined reference (for example, human genome) or not. To demonstrate this, we compiled a small panel of viral genomes, common laboratory contaminants, endogenous retroviruses and experimental vectors, and interrogated Malva Index to identify cells and samples containing these (Fig. 3a).

First, we assessed sensitivity via positive controls (Fig. 3a, left): for instance, PhiX sequencing spike-in, routinely added during library preparation, was detected across most datasets. Similarly, signal for *Mycoplasma* (a well-documented contaminant in cell culture) was detected in 0.5–2% of samples, mostly at the 23S ribosomal RNA (rRNA) locus (Extended Data Fig. 7c). This is consistent with prior reports of *Mycoplasma* contamination in RNA-seq datasets⁴⁵. Similarly, human endogenous retroviral signal (specifically, human endogenous retrovirus K (HERV-K)) was detected, as expected, in embryonic samples where these elements are known to be active. Of note, we observed tissue-specific expression in adult samples—such as healthy kidney and heart—cell lines and diseases including cancer and neurological conditions, supporting reports of HERV-K reactivation in adult tissues and pathological states^{46,47}. We also tested the retrieval of cells carrying specific synthetic constructs. In particular, we queried a lentiviral vector that is used for clonal tracing in a rhabdomyosarcoma study⁴⁸. Top hits returned by Malva, across the whole vector body (Extended Data Fig. 7d), correspond to cells from that study.

Then, we assessed specificity via negative controls (Fig. 3a, right): we did not detect signals for pathogenic viruses that are not represented in the current Malva Index (for example, SARS-CoV-2, hepatitis C, measles, West Nile and Zika viruses). However, false negatives remain possible if they are below the detection threshold. We only observed hits of SARS-CoV-2-positive cells from lung samples that were labelled as healthy. Notably, these correspond to a cell line experiment (derived from healthy donors) in which cells were infected in vitro (Extended Data Fig. 7b), highlighting how Malva overcomes annotation biases.

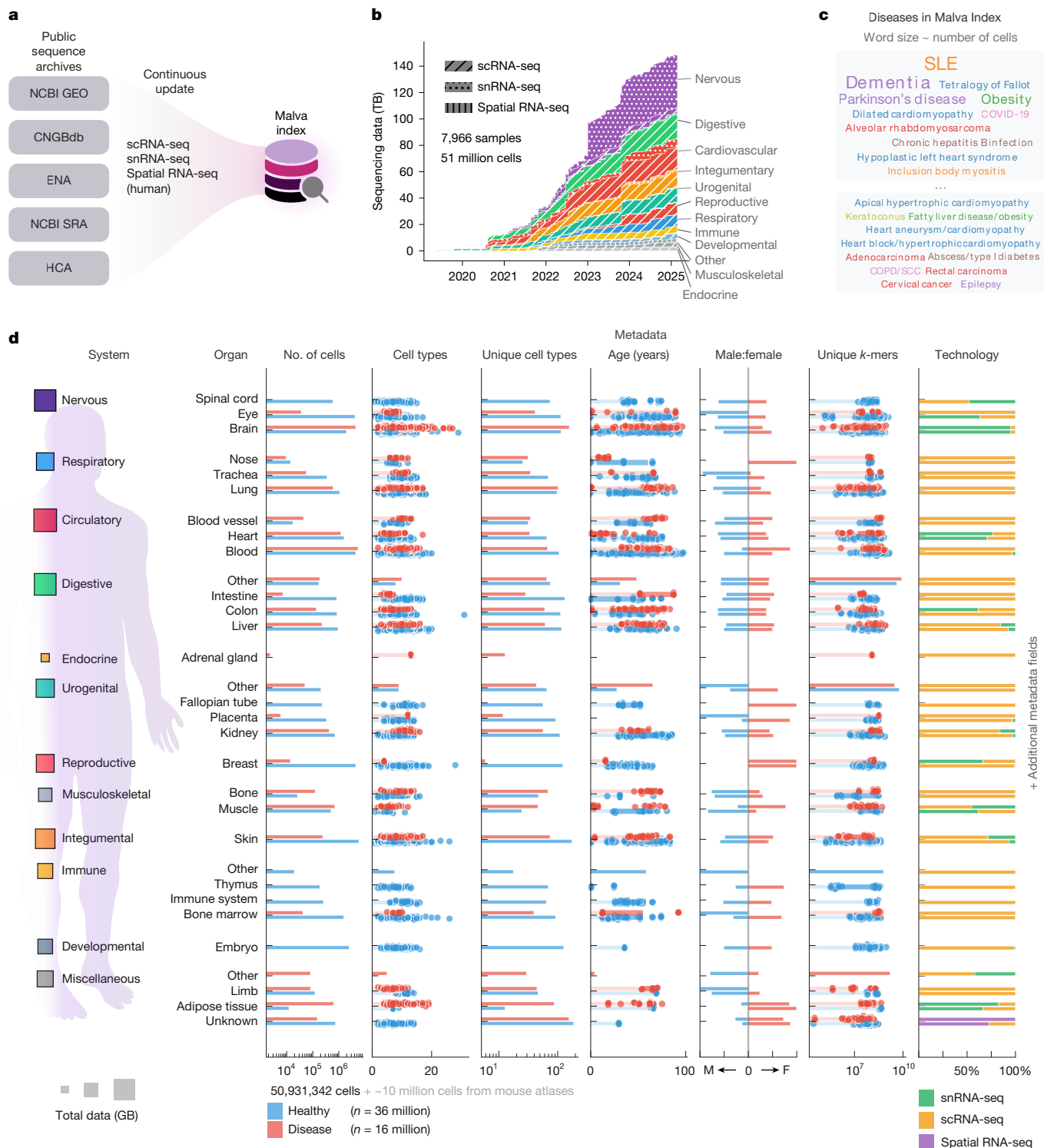


Fig. 2 | Malva Index. **a**, Public sequence archives (NCBI GEO, SRA, ENA, CNGBdb and the HCA data portal) are continuously crawled for single-cell RNA-seq (scRNA-seq), single-nucleus RNA-seq (snRNA-seq) and spatial RNA-seq datasets, which are indexed and integrated into Malva Index. **b**, Cumulative expansion of the human Malva Index (v2025.04a at the time of analysis, spanning around 60 million cells), with sequencing volume stratified by organ system and technology type. **c**, Disease representation in Malva Index, shown as a word cloud in which word size and position in the vertical axis reflect the number of indexed cells per condition, with words at

the top corresponding to the largest number. Autoimmune, neurodegenerative and oncological disorders are among the most prevalent. COPD, chronic obstructive pulmonary disease; SCC, squamous cell carcinoma; SLE, systemic lupus erythematosus. **d**, Key metrics and metadata (part of the query results output) of the human Malva Index, organized by anatomical system and organ. Panels show cell numbers, total and unique annotated cell types, donor age distribution, sex balance, unique *k*-mer counts and technology contributions. Bars distinguish healthy (blue) and disease (red) samples.

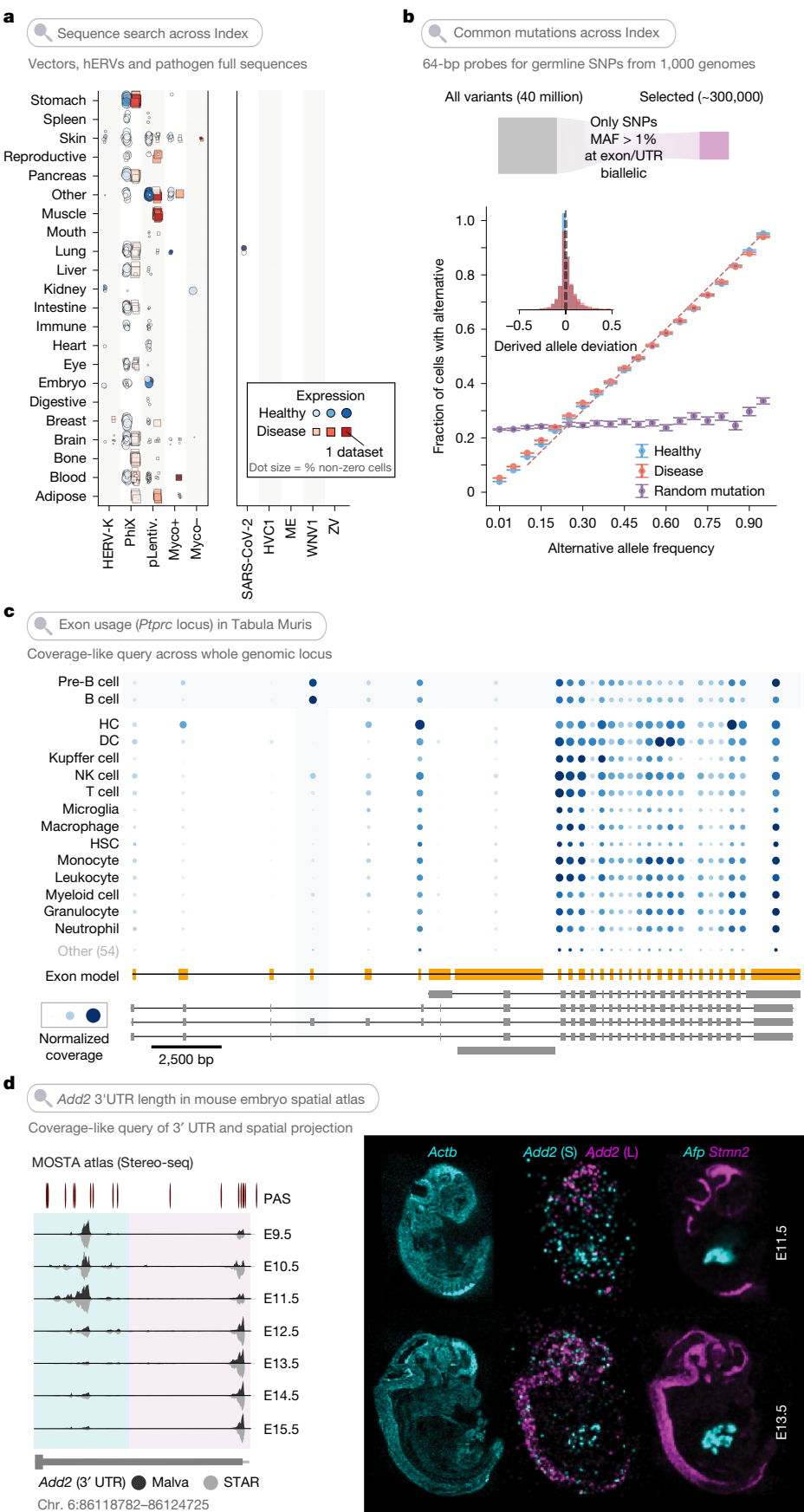


Fig. 3 | See next page for caption.

Fig. 3 | Malva enables sequence-level queries beyond current data portals.

a, Sequence search across Malva Index detects viral, retroviral and contaminant sequences. Columns show genomes or probes (for example, HERV-K, PhiX spike-in, lentiviral vector (pLentiv.), *Mycoplasma* species, SARS-CoV-2, HCV1 (hepatitis C virus probe), measles (ME), West Nile virus (WNV) and Zika virus (ZV)), rows represent human tissues. Dot size indicates the fraction of positive cells; colour denotes normalized expression. *Mycoplasma* species expected in human cell culture (Myco+) and negative-control species (Myco-) are included. hERVs, human endogenous retroviruses. **b**, Malva recapitulates the known distribution of human germline genotypes at single-cell resolution. Systematic screening of 296,985 common SNPs (64-bp probes) from the 1000 Genomes Project. Filtering pipeline from all variants (approximately 40 million) to the selected set (approximately 300,000 coding, biallelic SNPs with MAF $\geq 1\%$). Scatter plots (mean $\pm$ s.e.) show correlation between Malva-observed allele frequencies and population expectations in healthy and disease samples; random mutation controls are generated by altering a non-SNP position

(Methods). Inset, distribution of deviations from the expected alternative allele frequency when subset to derived alleles, confirming no systematic bias in health or disease. **c**, Malva recapitulates isoform switches that remain hidden in conventional gene expression analysis. Exon usage at the *Ptprc* locus in Tabula Muris⁴. Coverage-like queries reveal cell-type-specific inclusion of exon A (encoding the CD45RA isoform in B cells) versus exclusion (encoding CD45RO-like isoform in other immune cells). Rows represent cell types, with dot size and colour reflecting exon inclusion. DC, dendritic cell; HC, haematopoietic cell; HSC, haematopoietic stem cell; NK, natural killer. **d**, Malva analysis explains 3' UTR usage patterns in its spatial context—that is, 3' UTR length regulation of *Add2* during spatiotemporal mouse embryogenesis. Left, coverage profiles across stages E9.5–E15.5 show proximal (cyan) and distal (magenta) polyadenylation site (PAS) usage; Malva-derived profiles (top) closely match STAR alignments (bottom). Right: Stereo-seq spatial maps (MOSTA atlas⁵⁰) highlight distal UTR usage in developing neural tissue versus proximal usage in liver. Chr., chromosome.

Screening for germline variation across single-cell atlases

We next used Malva Index to locate cells with specific mutation signatures. We designed sequence probes targeting 296,985 common germline single nucleotide polymorphisms (SNPs) from the 1000 Genomes Project, focusing on biallelic variants in protein-coding regions with minor allele frequency (MAF) $\geq 1\%$ (Fig. 3b, top and Methods); this ensured that we focused on variants that reached equilibrium rather than rare, deleterious mutations.

Population-based allele frequencies correlated strongly to Malva-observed frequencies across healthy and disease samples (Fig. 3b). Thus, single-cell RNA data in Malva Index recapitulated the underlying population genetic structure despite heterogeneous sequencing technologies and depths. We additionally observed that common coding variants in our panel behave neutrally regardless of disease status, as there was no deviation between observed and expected derived allele frequencies (Fig. 3b, inset). Finally, we validated that allele frequencies at the single-cell level remained around 50% irrespective of population frequency (Extended Data Fig. 7e), supporting how most common coding SNPs segregate at intermediate frequencies in outbred populations, and that individuals are frequently heterozygous.

Together, variant querying recapitulated population genetic structure and aligned with expectations that variants reaching $\geq 1\%$ frequency in coding regions represent evolutionarily stable, predominantly neutral alleles at a single-cell level.

Quantifying isoform usage without mapping

Gene-centric atlases collapse many isoforms into single counts, obscuring cell-type-specific regulatory switches. We leveraged Tabula Muris⁴ (profiled at full-length coverage), to systematically compare expression across gene loci as a proxy for isoform usage. We identified 1,138 loci with cell-type-specific differential coverage patterns (Extended Data Fig. 7f): for example, *Ptprc* (also known as *Cd45*) in B cells predominantly included exon A (encoding CD45RA isoform) while other immune cells mostly excluded it (instead encoding CD45RO), recapitulating this well-known isoform variation across lymphocyte subsets (Fig. 3c). This pattern emerging solely from coverage-like analysis supports the ability of Malva to interrogate exon and isoform usage.

Moreover, we focused on isoforms at the 3' untranslated region (UTR): these are fundamental for regulation of transcript and protein localization and stability and, more importantly, the SC/ST data in the Malva Index are mostly 3'-biased, therefore suitable for exploring this landscape. As an example, we interrogated the 3' UTR of *Add2*, which is known to have tissue- and stage-specific patterns in the mouse embryo⁴⁹, against a high-resolution spatial atlas of mouse organogenesis⁵⁰. Malva validated how distal 3' UTR has the strongest signal in the developing neural tissue, whereas the proximal form accumulates at the developing liver. Malva's coverage-like profiles closely matched those

generated by mapping and alignment with a standard tool (Fig. 3d), confirming that the bulk-level lengthening of *Add2* across developmental stages is likely to represent a shift in cell-type proportions (neuronal versus liver progenitors). More generally, systematic benchmarking against polyASite v3.0⁵¹ confirmed that polyA-specific probes correlate more strongly with database usage scores than control probes, validating probe design for site-specific detection (Extended Data Fig. 8).

Querying circRNAs across cell atlases

circRNAs, produced by back-splicing, remain invisible in gene expression matrices provided by most current atlases. Researchers currently rely on specialized circRNA databases that cover limited samples and often do not cover single-cell or spatial data. Malva enables interactive analysis of circRNA expression by searching for probes specific to their backsplice junctions. For instance, we searched for *CDRIas*, a conserved circRNA that is highly expressed in the brain⁵². Malva confirmed the expected cell-type specificity of *CDRIas* at single-cell resolution (Extended Data Fig. 7g): 95% of positive cells were excitatory neurons, with lower detection in inhibitory interneurons and very low detection across other tissues. This specificity, previously known from bulk studies, emerges from droplet-based data that were not specifically designed to capture circRNAs.

Querying cancer somatic mutations at scale

We next tested whether Malva could detect somatic mutations directly from single-cell RNA-seq data at pan-cancer scale. We used the scTML database⁵³, which provides matched mutation calls across 280 tumour samples from 16 cancer types (Extended Data Fig. 9a), and designed allele-specific probes centred on expressed SNVs in exonic and UTR regions (Methods). Malva recovered mutation-positive cells with high specificity and sensitivity across all cancer types (Fig. 4a), despite differences in tumour type, sample size, sequencing depth and mutation frequency. Sensitivity remained high even for variants detected in few cells (Fig. 4b and Extended Data Fig. 9b), and across variants at different features types (Fig. 4c). Thus, this benchmark shows that Malva can recover cancer-associated mutations from large single-cell cohorts without realignment or running specialized variant calling algorithms.

Together, these analyses demonstrate how, rather than limited to predefined gene sets, Malva allows researchers to interrogate millions of cells across diverse functional layers of RNA biology—expression, mutations and isoforms—via a single sequence search interface.

From cell-by-gene to cell-by-sequence

Searching with Malva enables retrieval of cells containing specific biological signals. However, this does not enable cell–cell comparison, which is critical for single-cell clustering, trajectory inference and

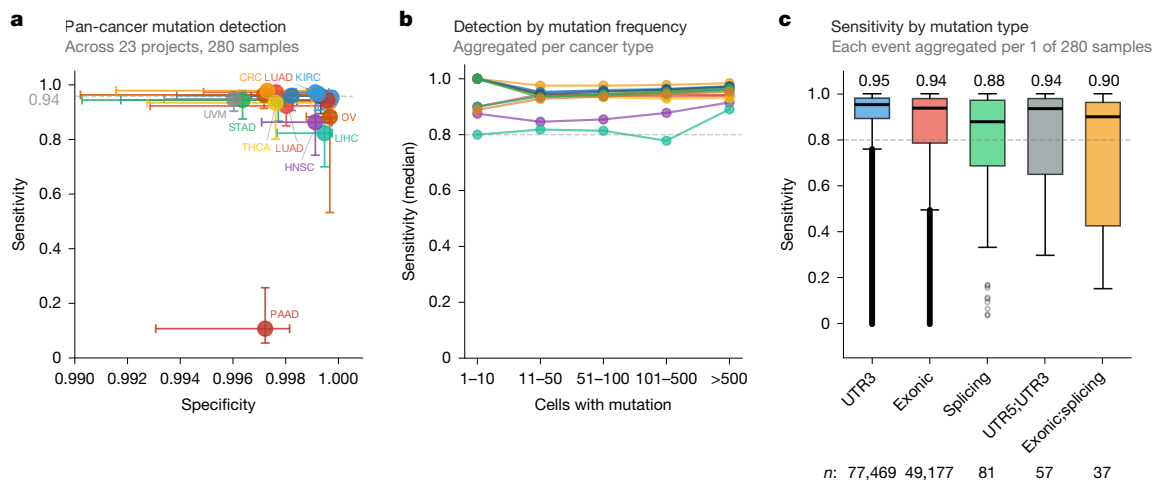


Fig. 4 | Malva detects somatic mutations across single-cell cancer datasets. **a**, Sensitivity versus specificity across 16 cancer types (280 samples) compared to quantification in the scTML database⁵³. Each point represents one study and cancer type; error bars indicate the interquartile range. CRC, colorectal cancer; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; OV, ovarian cancer; PAAD, pancreatic adenocarcinoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UVM, uveal melanoma. **b**, Sensitivity as a function of mutation frequency (number of cells carrying the variant), aggregated by cancer type. **c**, Sensitivity stratified by mutation genomic context.

marker discovery. These operations are carried in a reduced (latent) space computed from cell-by-feature matrices, commonly, cell-by-gene matrices (Fig. 5a).

Malva Index encodes billions of sequences into a large sequence-by-cell matrix. By transposing it, we could in principle perform sequence-based, cell–cell comparison. In practice, these cell-by-sequence matrices are too large to handle (millions of cells times billions of features). Thus, we collapse sequence *k*-mers into buckets of sequence similarity (Methods). This ensures that cell-by-feature matrices are small enough such that dimensionality reduction is tractable.

We tested this idea on approximately 30 million cells from around 5,000 samples from healthy individuals and found that the principal components derived from cell-by-bucket matrices were comparable to those derived from cell-by-gene matrices (Fig. 5b). We quantified this via two metrics: (1) neighbourhood preservation; and (2) clustering similarity (Methods). Across samples and tissue types, the degree of concordance between sequence-based and gene-based embeddings was variable (Fig. 5c and Extended Data Fig. 10a); in some tissues, cells that are neighbours in cell-by-gene space are also neighbours in cell-by-bucket space (high normalized mutual information (NMI)); in others, the two representations captured distinct structures (low NMI). This variation reflects the complementary nature of sequence-based analysis: although it can recover gene-driven neighbourhoods, it may also reveal additional structure. Therefore, these embeddings are an orthogonal layer of information to gene-based analysis.

This approach becomes particularly useful in settings where high-quality reference genomes are lacking, sparse or incomplete; for example, for non-model organisms such as the phylum *Placozoa* (Fig. 5d), showcasing the utility of sequence-based representations for large-scale biodiversity studies¹³. We observed, however, that samples coming from single-nuclei technologies yield consistently lower values across both metrics. We attributed this to the over-representation of low-complexity intronic (nuclei-enriched) RNA (Extended Data Fig. 10b), which is equivalent to yields from low sequence diversity (for example, using 4-mers and below) (Extended Data Fig. 10c and Supplementary Note 4).

UTR3, UTR5, Exonic and Splicing labels indicate the functional annotation of the genomic region of a genetic variant relative to RefSeq genes. Labels joined with semicolons indicate that mutations fall within a region with overlapping features. 3' UTR variants show the highest sensitivity, followed by exonic variants, with splicing-proximal variants showing moderate sensitivity due to junction-spanning probe constraints. Mean sensitivity across cancer types (approximately 0.94) is similar to mutation detection rates in bulk RNA-seq, where *k*-mer indices such as Transipedia (running on REINDEER) report recall values of 0.875–0.945 at high-precision thresholds³⁷. Cancer types in **b,c** are represented by the same colours as in **a**. In **a**, PAAD is an outlier due to mouse-derived samples in the original database; therefore, PAAD is excluded in **b,c**.

Reference-free cell clustering

Cell clusters are often characterized by their marker genes. However, in bucket-based clustering, features (groups of similar sequences) are not directly explainable. We assessed cell-by-bucket representations on a spatial transcriptomics sample (high-resolution profiling of a human tumour), in which tissue architecture provides an orthogonal readout for interpretability; that is, the degree (or lack of) of spatial organization can inform biological relevance. Indeed, gene-based and bucket-based representations captured similar major tissue compartments (tumour and stromal), plus two *k*-mer-specific clusters (KS1 and KS2) with distinct spatial organization, one within the tissue boundaries and another outside the tissue boundaries (Fig. 5e). We hypothesized that there might be sequence signatures that are specific to these two clusters, whose signal is absent from the original reference-based gene representation.

De novo detection of marker sequences

Across all bucket-based clusters, we extracted their specific collections of *k*-mers and used them as the input for de novo assembly (Methods). This yielded thousands of contiguous unique sequences (contigs), which we termed 'marker sequences' (Extended Data Fig. 10d). Upon mapping these to the human genome, around 70% of these contigs were annotated to known genes (mostly within 1 kb of the transcription end site (TES)), around 29% were annotated to intergenic regions, and approximately 1% remained unmapped (Extended Data Fig. 10d–f). Also, by narrowing the *k*-mer search space to those informative of cluster identity, downstream assembly (with SPAdes) operated more efficiently (requiring about 1 h versus about 27 h of wall time).

We then examined the cluster-specific enrichment of all assembled contigs. All clusters (except KS1 and KS2) were enriched in contigs that mapped to human genes. Some contigs mapped to the same gene bodies but across different locations, and a subset of these were significantly detected as markers (Fig. 5f). One example is *KLK10_676*,

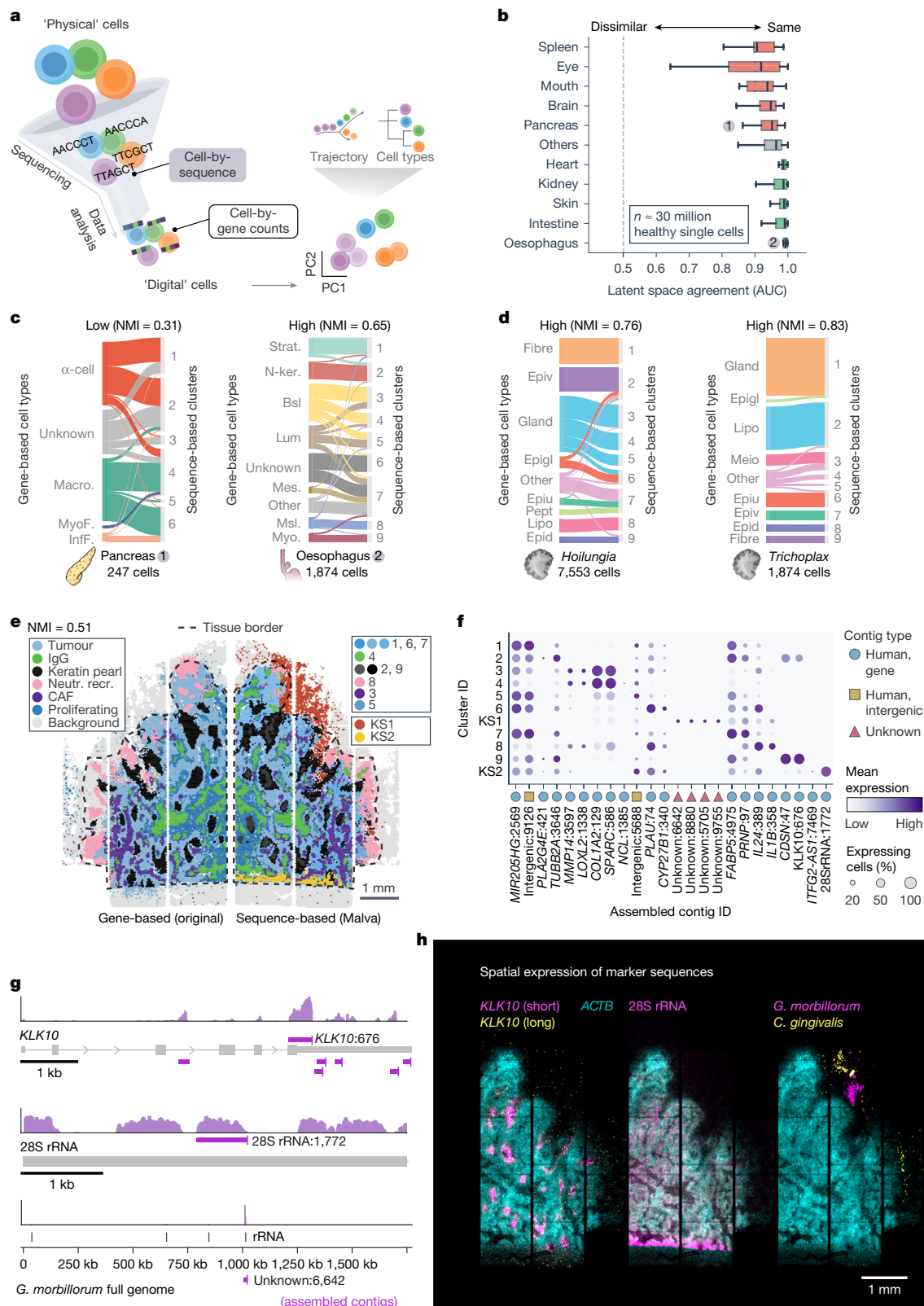


Fig. 5 | See next page for caption.

a contig that maps to *KLK10*, a gene encoding a tumour-enriched serine protease. This was the only *KLK10* contig that was enriched in keratin pearl structures (Fig. 5g,h). Coverage-like profiles showed that this feature may correspond to an unannotated 3' UTR end; in fact, a polyA tract at its 3' end further supports it as a TES. Conversely, KS1 and KS2 were indeed enriched in intergenic or unknown contigs (Fig. 5f).

We first identified that KS2 was enriched in sequences that mapped to human rRNA (Fig. 5g,h), which are often removed during reference preprocessing, and were thus missing from the original cell-by-gene matrix. The KS1 cluster was also enriched in contigs that are compatible with bacterial 23S rRNA from several species that are known to occur in the oral mucosa (Fig. 5g,h). This confirms that this cluster

Fig. 5 | Sequence-based transcriptomic analysis without predefined gene annotations. **a**, Conceptual view of the single-cell information funnel.

Biological ('physical') cells are converted to 'digital' cells via sequencing, which can be represented either as cell-by-gene or cell-by-sequence matrices. These representations enable reconstruction of latent spaces underlying cell types and trajectories. PC, principal component. **b**, Latent space agreement comparing sequence-based (k -mer bucket) versus gene count principal component embeddings across healthy human cells. Values greater than 0.5 indicate better-than-random preservation of local neighbourhoods. $n = 29,240,986$ cells examined over 4,932 sequencing runs. Box plots show the median (centre line) and interquartile range (box edges), and whiskers extend to 1.5 times the interquartile range. AUC, area under the curve. **c**, Example comparisons of clustering based on gene features versus sequence features in human samples. Panels show cases with low (NMI = 0.31) and high (NMI = 0.65) concordance, corresponding to tissues highlighted in **b**. Bsl, basal epithelial cell; InfF., inflammatory fibroblast; Lum, luminal epithelial cell; Macro., macrophages; Mes., mesenchymal progenitor cell; Msl., muscle fibre precursor; Myo., myofibroblast; MyoF., myofibroblast; N-ker., non-keratinized epithelial cell; Strat., stratified epithelial cell. **d**, Sequence-based clustering of non-model organisms (*Hoilungia* and *Trichoplax*). Clustering structure based on k -mer

features closely resembles gene-based annotations. Epid, epithelial dorsal cells; Epigl, epithelial gland-like cells; Epiu, epithelial and unknown cluster cells; Epiv, ventral epithelial cells; Fibre, fibre cells; Gland, glandular cells; Lipo, lipophil cells; Meio, meiotic cells; Pept, peptidergic cells. **e**, Spatial clustering of a head and neck tumour dataset using gene count-based (left) versus sequence-based (Malva, right) embeddings. KS1 and KS2 are two k -mer-specific clusters. CAF, cancer-associated fibroblasts; Neutr. recr., neutrophil-recruiting population. **f**, Cluster-specific de novo assembly of marker sequences. Contigs assembled from discriminative k -mers reveal both annotated human features (genes and intergenic regions) and unaligned sequences. The dot plot shows the fraction of cells per cluster expressing each assembled contig (dot size) and the average log-normalized expression rescaled per contig so that the highest mean expression across clusters is set to 1 (dot colour). **g**, Representative marker sequences assembled from KS1 and KS2 clusters. Examples include a truncated 3' UTR isoform of *KLK10* (human) and bacterial 23S rRNA sequences from *Gemella morbillorum*. **h**, Spatial expression maps of selected marker sequences from **g**, shown as kernel density estimates. Distinct patterns confirm biological specificity: the truncated *KLK10* isoform localizes to tumour regions, whereas microbial rRNA contigs map to extra-tissue areas, co-occurring with oral bacteria such as *G. morbillorum* and *C. gingivalis*.

separated from a genuine microbial signal that is only captured in the sequence-based analysis.

Together, these results show that Malva can unbiasedly detect signals from SC/ST datasets in a reference-agnostic manner, providing insights that remain invisible to gene-centric analysis.

Discussion

We built Malva to make raw single-cell and spatial transcriptomic sequences searchable at (sub)cellular resolution. We indexed public datasets across millions of cells and show that sequences can be queried directly and instantly, without first reducing the data to genes, transcripts or predefined annotations. This makes it possible to search for and compare between splice junctions, mutations, viral sequences, synthetic constructs and non-reference RNAs across many studies, in seconds. This is particularly relevant for artificial intelligence-assisted analysis: neural networks learn from biological data but typically cannot return to the underlying experiments during inference. Malva provides a retrieval layer that allows models and agents to query observed sequences, retrieve cellular evidence and refine their predictions against public data.

Many biological questions begin with such queries. A mutation can define a cancer clone, a splice junction can define a cell state, a viral RNA can define an infected cell, and a barcode can define a perturbation. These signals are not recoverable from processed count matrices. Malva keeps them accessible and links them to cells, tissues, disease states and metadata.

Beyond individual searches, Malva can classify cells from sequence composition, assemble cell-type-specific marker sequences and may serve as a real-time, sequence data source for building artificial intelligence models. It therefore provides a practical discovery layer for asking sequence-defined questions in single-cell biology.

Limitations

Malva currently supports queries of at least 24 nucleotides and uses exact k -mer matching, which enables fast searches of high specificity and sensitivity, with sensitivity being reduced only when the underlying data contain very high error rates, which are not typical in single-cell datasets sequenced with standard technology. Malva reports k -mer-derived pseudocounts rather than absolute molecule counts, and results should be followed by task-specific quality control, quantification or validation when needed. The current Malva Index also reflects the public data available today. Many datasets are short-read

and 3'-biased, and spatial metadata are often incomplete or difficult to standardize. Finally, sequence search in human transcriptomic data can reveal donor-specific genotype information; this risk should be treated similarly to the risks associated with raw genome and exome sequencing data (Supplementary Discussion).

Conclusion

As single-cell and spatial datasets expand, searchable sequence indexes may become a natural way to reuse public data. Future releases of Malva Index could include more organisms, spatial sequencing- and imaging-based assays, long-read transcriptomics, chromatin accessibility, proteomics and RNA modifications, among others. More generally, Malva can enable public single-cell archives to become live molecular resources, queried by sequence and linked back to cells, tissues and biological context. Ultimately, Malva may help to provide a coordinate system for mapping one's own data onto global datasets to understand health and disease mechanisms and trajectories⁵⁴.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-026-10975-w>.

1. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).
2. Amit, I. et al. The commitment of the human cell atlas to humanity. *Nat. Commun.* **15**, 10019 (2024).
3. Hrovatin, K. et al. Considerations for building and using integrated single-cell atlases. *Nat. Methods* **22**, 41–57 (2025).
4. The Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
5. The Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
6. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
7. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
8. Arya, A., Tripathi, P., Dubey, N., Aier, I. & Kumar Varadwaj, P. Navigating single-cell RNA-sequencing: protocols, tools, databases, and applications. *Genomics Inform.* **23**, 13 (2025).
9. Rizvi, S. A. et al. Scaling large language models for next-generation single-cell analysis. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.04.14.648850> (2025).
10. A "Wikipedia for cells": researchers get an updated look at the Human Cell Atlas, and it's remarkable. *Nature* **635**, 523–524 (2024).
11. Pentimalli, T. M., Karaikos, N. & Rajewsky, N. Challenges and opportunities in the clinical translation of high-resolution spatial transcriptomics. *Annu. Rev. Pathol.* **20**, 405–432 (2025).

12. Evolution at the cellular level. *Nat. Ecol. Evol.* **7**, 1155–1156 (2023).
13. Seb  Pedr  s, A. et al. The Biodiversity Cell Atlas: mapping the tree of life at cellular resolution. *Nature* **645**, 877–885 (2025).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
15. CZI Cell Science Program. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).
16. Xu, Z. et al. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res.* **52**, D1053–D1061 (2024).
17. Papatheodorou, I. et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2020).
18. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
19. Franz  n, O., Gan, L.-M. & Bj  rkgren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).
20. Li, Y. et al. SOAR elucidates biological insights and empowers drug discovery through spatial transcriptomics. *Sci. Adv.* **11**, eadt7450 (2025).
21. Li, M. et al. DISCO: a database of deeply integrated human single-cell omics data. *Nucleic Acids Res.* **50**, D596–D602 (2022).
22. Melsted, P. et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* **39**, 813–818 (2021).
23. Sztanka-Toth, T. R., Jens, M., Karaiskos, N. & Rajewsky, N. Spacemake: processing and analysis of large-scale spatial transcriptomics data. *Gigascience* **11**, giac064 (2022).
24. He, D. et al. Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nat. Methods* **19**, 316–322 (2022).
25. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.05.442755> (2021).
26. Xi, J., Lee, J. H., Kang, H. M. & Jun, G. STtools: a comprehensive software pipeline for ultra-high resolution spatial transcriptomics data. *Bioinformatics Adv.* **2**, vbac061 (2022).
27. Karasikov, M. et al. Efficient and accurate search in petabase-scale sequence repositories. *Nature* **647**, 1036–1044 (2020).
28. Marchet, C., Iqbal, Z., Gautheret, D., Salson, M. & Chikhi, R. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics* **36**, i177–i185 (2020).
29. Solomon, B. & Kingsford, C. Fast search of thousands of short-read sequencing experiments. *Nat. Biotechnol.* **34**, 300–302 (2016).
30. Pandey, P. et al. Mantis: a fast, small, and exact large-scale sequence-search index. *Cell Syst.* **7**, 201–207.e4 (2018).
31. Fan, J., Khan, J., Singh, N. P., Pibiri, G. E. & Patro, R. Fulgor: a fast and compact k-mer index for large-scale matching and color queries. *Algorithms Mol. Biol.* **19**, 3 (2024).
32. Alanko, J. N., Vuontoniemi, J., M  klin, T. & Puglisi, S. J. Themisto: a scalable colored k-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. *Bioinformatics* **39**, i260–i269 (2023).
33. Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. & Babaian, A. Logan: planetary-scale genome assembly surveys life’s diversity. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.07.30.605881> (2024).
34. Morillon, A. & Gautheret, D. Bridging the gap between reference and real transcriptomes. *Genome Biol.* **20**, 112 (2019).
35. Nguyen, H. T. N. et al. Reference-free transcriptome signatures for prostate cancer prognosis. *BMC Cancer* **21**, 394 (2021).
36. Wang, Y. et al. The contribution of uncharted RNA sequences to tumor identity in lung adenocarcinoma. *NAR Cancer* **4**, zcac001 (2022).
37. Bessi  re, C. et al. Transpedia.org: k-mer-based exploration of large RNA sequencing datasets and application to cancer data. *Genome Biol.* **25**, 266 (2024).
38. Audoux, J. et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* **18**, 243 (2017).
39. Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* **50**, 1474–1482 (2018).
40. Zheludev, I. N. et al. Viroid-like colonists of human microbiomes. *Cell* **187**, 6521–6536.e18 (2024).
41. Sun, Q., Peng, Y. & Liu, J. A reference-free approach for cell type classification with scRNA-seq. *iScience* **24**, 102855 (2021).
42. Gouri  , H. et al. Scalable single-cell metagenomic analysis with Bascet and Zorn. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.06.20.660799> (2025).
43. Ghaddar, B. et al. Tumor microbiome links cellular programs and immunity in pancreatic cancer. *Cancer Cell* **40**, 1240–1253.e5 (2022).
44. Cheng, O. et al. Flexiplex: a versatile demultiplexer and search tool for omics data. *Bioinformatics* **40**, btac102 (2024).
45. Olererin-George, A. O. & Hogenesch, J. B. Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI’s RNA-seq archive. *Nucleic Acids Res.* **43**, 2535–2542 (2015).
46. Li, W. et al. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl. Med.* **7**, 307ra153 (2015).
47. Dhillon, P. et al. Increased levels of endogenous retroviruses trigger fibroinflammation and play a role in kidney disease development. *Nat. Commun.* **14**, 559 (2023).
48. Patel, A. G. et al. The myogenesis program drives clonal selection and drug resistance in rhabdomyosarcoma. *Dev. Cell* **57**, 1226–1240.e8 (2022).
49. Agarwal, V., Lopez-Darwin, S., Kelley, D. R. & Shendure, J. The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nat. Commun.* **12**, 5101 (2021).
50. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).
51. Moon, Y., Herrmann, C. J., Mironov, A. & Zavolan, M. PolyASite v3.0: a multi-species atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res.* **53**, D197–D204 (2025).
52. Cerda-Jara, C. A. et al. The history and function of a circular RNA. *Nat. Commun.* **17**, 3862 (2026).
53. Li, H. et al. scTML: a pan-cancer single-cell landscape of multiple mutation types. *Nucleic Acids Res.* **53**, D1547–D1556 (2025).
54. Rajewsky, N. et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386 (2020).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

   The Author(s) 2026

Methods

The Malva platform

Malva provides a unified interface to search millions of cells by any RNA sequence. Malva integrates five modular components that operate independently but communicate through a central orchestrator (Extended Data Fig. 1b). When a user submits a query via the public RESTful API—a gene symbol, DNA sequence, or natural language request such as “cells expressing markers of neurodegeneration”—it is routed by a central orchestrator. Natural language queries first pass through an integrated language model that translates biological concepts into gene sets. The search engine then decomposes sequences into *k*-mers and queries our distributed index, leveraging pre-computed results for common genes. Malva Index returns cell identifiers with expression levels or coverage profiles, which are then enriched by a metadata service that appends standardized cell-level (for example, cell-type annotation) and sample-level (for example, tissue, disease and developmental stage) annotations. This enables users to filter results or perform targeted searches on specific sample subsets. The platform can also expand datasets available in Malva Index: a crawler monitors public repositories (GEO, SRA, HCA Data Portal and ENA) for new datasets, which are asynchronously processed into separate Malva Index instances and progressively merged into the searchable corpus. Load balancing across multiple workers ensures consistent performance even as hundreds of users query simultaneously. Results return in standard formats (JSON, h5ad) ready for downstream analysis in familiar tools such as scanpy, with full metadata to enable further filtering and aggregation. The server code operates on a standard, open-source stack of tools (Flask v3.1.0, numpy v2.2.6, pandas v2.3.2, SQLite v3.49.1 and polars v1.27.1).

Data sources, crawling and availability

Malva Index incorporates public single-cell and spatial transcriptomic datasets from major repositories (Human Cell Atlas, GEO, SRA, ENA, ArrayExpress and CNGBdb). Data are ingested into Malva Index via the malvacrawl service. The service maintains a registry of datasets identified primarily from the Human Cell Atlas Data Portal, supplemented by automated parsing of GEO and ENA metadata feeds, and manual ingestion of selected identifiers from other platforms (such as ArrayExpress or CNGB). Each candidate dataset is assigned an internal UUID (universally unique identifier) and checked against existing entries by GEO or ENA accession to prevent duplication. Because deposited metadata are frequently incomplete or inconsistent (for example, the sequencing technology or barcode format is often not specified), malvacrawl utilizes a two-stage annotation strategy. First, a large language model (instruction-tuned Llama 3.1) performs an initial parse of the dataset description, extracting candidate technology, organism and tissue labels. Second, these candidates are manually curated and mapped to a controlled ontology using text2term v4.5.0 (see ‘Metadata standardization’). For raw data acquisition, malvacrawl attempts to download FASTQ files via SRA prefetch, HTTP, or FTP in order of preference, using sra-tools (v3.4.1). For datasets archived as BAM files, we converted them back to FASTQ using bamtofastq (v1.4.1). Barcode and sequence parsing during preprocessing relied on dnaio (v1.2.4). For barcoded technologies (primarily 10x genomics), paired-end data are expected; therefore, if only a single FASTQ file is available, or only BAM files are archived, these BAM files are downloaded and converted back to FASTQ using standard 10x conversion utilities. The technology is identified either from deposited metadata, from the structure of the R1 read (barcode length and whitelist match rate), or, when ambiguous, via manual inspection. Datasets that are not compatible with Malva’s whitelist-based barcode system (see ‘Barcode preprocessing’) are deferred for future integration. For non-barcoded, plate-based assays (for example, Smart-seq2), each FASTQ file comprises one cell, and datasets are ingested in ‘bulk’

mode as per-cell indices, and later merged into a sample-level index that preserves cellular resolution.

Currently spanning ~60 million cells from thousands of experiments, Index includes diverse technologies (10x Chromium, Visium, Open-ST, Stereo-seq, Smart-seq2) across multiple organs, species, and conditions. The platform runs on-premise infrastructure with query limits to ensure consistent performance for all users.

Querying

Malva maintains local databases to translate user queries into searchable sequences. When users query gene symbols or Ensembl IDs, these are mapped to genomic coordinates on standard reference genomes (hg38 for human, mm10 for mouse) with default repeat masking⁵⁵. For expression queries, we retrieve full transcript sequences from Ensembl. For coverage queries that visualize read distribution along genes, we use the genomic coordinates to define the region of interest and return *k*-mer matches at each position.

Several specialized databases support specific semantic biological queries. To support these via natural language, we integrated Llama 3.1 via Ollama as a translation layer between user prompts and gene sets⁵⁶. The language model is optionally augmented with a Retrieval-Augmented Generation (RAG) database built from MSigDB⁵⁷, which includes curated gene sets from Reactome, Gene Ontology (GO) and other pathway databases. When users submit queries such as “inflammatory response genes in macrophages”, the system retrieves relevant gene sets from MSigDB and uses the language model to refine the selection based on the specific query context. This approach provides a convenient interface for exploring biological concepts without requiring users to manually compile gene lists. The resulting gene sets are converted to sequences and searched across Malva Index like any other query.

Pre-computed gene expression matrices

To accelerate common gene-level queries, we pre-compute and store gene expression matrices for frequently accessed gene sets. For each species in Malva Index, we compile comprehensive transcript catalogues from Ensembl, including all protein-coding transcripts, non-coding RNAs, and 3′ UTR sequences. We then query Malva Index with these reference sequences to generate pseudocount matrices for every indexed dataset.

The pre-computation process aggregates *k*-mer counts across all isoforms of each gene, producing a cell-by-gene matrix analogous to standard single-cell expression matrices. These matrices are stored in a distributed database indexed by gene symbol, Ensembl ID, and dataset identifier. When users query common genes through the API, the system retrieves pre-computed values directly from the database rather than searching the *k*-mer index, reducing query latency from seconds to milliseconds. The database is updated incrementally as new datasets are added to Index, ensuring that gene-level queries always reflect the complete corpus of indexed data. For custom non-standard gene symbols, the system falls back to real-time *k*-mer queries, maintaining flexibility while optimizing performance for common use cases.

Metadata standardization

Critical to the ability of Malva to aggregate results across studies is the harmonization of heterogeneous metadata from diverse sources (HCA Data Portal, GEO, SRA, ArrayExpress and CNGB). Raw metadata arrives in various formats with inconsistent naming conventions, where identical biological entities may be described using different terminologies across repositories. Similarly, developmental stages, disease states, and experimental descriptors vary widely between data sources.

We implemented an automated standardization pipeline based on text2term v4.5.0⁵⁸, a specialized biomedical ontology mapping tool that utilizes semantic similarity algorithms to map free-text metadata

terms to established ontology standards. The pipeline first preprocesses metadata by standardizing null values and normalizing common terminology variants. Multi-value fields containing multiple biological conditions are decomposed and mapped individually to ensure comprehensive coverage. The system then applies TF-IDF (term frequency-inverse document frequency)-based semantic matching with hierarchical fallback strategies to map terms to domain-appropriate ontologies: diseases to MONDO, anatomical structures to UBERON, phenotypic traits to PATO, and developmental stages to age-related ontologies. To enable system-level biological analysis, organs are additionally mapped to their hierarchical parent anatomical systems through ontology traversal within UBERON. This multi-layered semantic harmonization approach ensures that functionally equivalent terms from different repositories are mapped to identical standardized identifiers, enabling robust cross-study aggregation and systematic biological queries while preserving both granular and hierarchical biological relationships.

Cell-type classification

To enable aggregation of query results at the cell-type level, we implemented an automated annotation pipeline for all indexed cells. Author-provided cell-type labels are often heterogeneous across studies or entirely absent in many submissions. To address this, we perform de novo annotation using a two-stage approach.

First, we apply a marker-based scoring system. We compiled curated marker gene sets for major cell types following Cell Ontology classifications, with each cell type defined by 10–20 literature-validated markers. We then perform clustering using standard methods (principal component analysis (PCA), k -nearest neighbours, Leiden algorithm via `leidenalg v0.10.2`) and identify differentially expressed genes for each cluster. Across clusters, we calculate enrichment scores for each type signature; thus, clusters are assigned cell types based on the differential expression of marker genes, considering both the number of expressed markers and their specificity to the cluster.

Second, we refine these cluster-based annotations using a language model that integrates biological context, as previously described⁵⁹. The model receives the top differentially expressed genes for each cluster along with metadata (marker-based cell-type annotation, tissue origin, developmental stage, disease state) and produces refined cell-type assignments. To ensure biological accuracy, the model evaluates whether each annotation is plausible given the tissue context, flagging unlikely assignments for correction. We mapped the output cell-type annotations to a predefined, controlled vocabulary based on the cell ontology standard.

Natural language query interface validation. To systematically evaluate the accuracy of natural language query translation, we generated 10,000 synthetic test cases using `gpt-oss:120b` as a test case generator, ensuring independence from our production model (Llama 3.1 8B). Test cases were distributed across eight categories: simple gene queries (15%), gene queries with filters (20%), marker gene requests (15%), sequence searches (10%), database identifier lookups (10%), pathway queries (10%), unsupported queries (7%), and over-inference traps (15%). Over-inference traps were specifically designed to test whether the model inappropriately adds biological context as filters (for example, adding ‘T cell’ filter when querying FOXP3 without explicit cell-type specification).

We evaluated four metrics: query type accuracy (correct classification into the eight categories), gene list F1 (harmonic mean of precision and recall for extracted gene names), filter precision (fraction of returned filters that were explicitly stated in the query), and filter recall (fraction of explicitly stated filters that were captured). The complete set of 10,000 evaluation queries, together with expected outputs and the prompts used during instruction-tuning of the production model, are publicly available (Data availability).

In the production environment, categorical metadata filtering (for example, resolving a user’s tissue or disease term to standardized ontology values) is performed via a hybrid approach. First, the input term is matched against a fuzzy dictionary built from all unique values in the metadata store, using Levenshtein distance with a configurable threshold. If no confident match is found, the system falls back to vector similarity search against a sentence-embedding database (built with the same model used for ontology mapping), returning the nearest ontology term above a cosine similarity threshold.

Malva Index architecture. Malva Index organizes single-cell sequencing data as a searchable collection of k -mers (fixed-length nucleotide sequences) linked to their cellular origins. Each sequencing read is decomposed into non-overlapping k -mers of length k (default $k = 24$), with a final overlapping k -mer to ensure complete sequence coverage. Specifically, for a sequence of length l , we extract k -mers at positions $0, k, 2k, \dots$, plus one k -mer starting at position $l - k$ if needed. The use of non-overlapping k -mers rather than minimizer-based sampling is a purposeful design choice. We acknowledge that minimizers were developed to guarantee shared k -mers between overlapping sequences (essential for assembly); however, Malva’s windowed query strategy already ensures full sensitivity for queries with $\geq 2k$ nucleotides (see ‘Sequence querying and pseudoquantification’). Importantly, non-overlapping sampling greatly reduces indexing time and memory cost, provides uniform detection probability and simplifies index merging to a union operation.

Each k -mer is encoded as a 64-bit integer using 2-bit nucleotide encoding ($A = 00, C = 01, G = 10, T = 11$), enabling efficient storage and comparison. The cellular origin is tracked through a composite identifier that encodes both the cell barcode and dataset source. We allocate the upper bits for dataset identification and lower bits for cell-specific barcodes; for example, using 8 bits for dataset ID and 24 bits for cell barcodes allows indexing up to 256 datasets with ~16 million cells each. This hierarchical encoding enables queries to return both the specific cell and its source dataset.

Malva Index uses a two-tiered structure optimized for sequence search conceptually similar to inverted indices for web search. The sequence array stores all unique k -mers found across datasets, sorted numerically. The location array contains the composite cell-dataset identifiers where each k -mer appears. These are linked via the pointer array mapping each k -mer to its locations, storing start and end positions within the location array. This design enables rapid lookup: finding all cells containing a specific k -mer requires just one binary search in the sequence array followed by a single read from the location array.

Individual datasets are processed independently before being merged into the unified index. FASTQ files are decompressed and parsed in parallel, extracting cell barcodes and sequences at rates exceeding 2 million reads per core per second. For each dataset, reads are streamed through memory-efficient buffers, extracting k -mers and their associated barcodes. The k -mers are lexicographically sorted and deduplicated, keeping track of all cells where each appears. These dataset-specific indices are then merged hierarchically: smaller indices combine into progressively larger ones using k -way merge algorithms that require minimal memory. This approach enables processing of datasets far larger than available RAM while maintaining near-linear scaling with data size. Detailed algorithmic descriptions of the index chunk generation, chunk merging and query scoring procedures, including pseudocode sufficient for reimplementing, are provided in Supplementary Methods.

To handle indices exceeding available memory, k -mers are partitioned into buckets, each stored contiguously on disk (Supplementary Methods). A lookup structure, always resident in memory, maps k -mer prefixes directly to the disk offset and size of its bucket. Locating a k -mer requires one prefix table lookup followed by a binary search within the bucket, typically fewer than 20 comparisons for billion-scale indices.

Article

The prefix table and suffix arrays are accessed via memory-mapped file input–output; cell data blocks are read on demand for each queried bucket.

The final merged index is stored in a compressed binary format optimized for disk-based retrieval. Arrays are partitioned into fixed blocks (512 entries) that are independently compressed using delta-encoding followed by Blosc compression, achieving 10–20× reduction compared to raw FASTQ files (Supplementary Methods). The hierarchical organization enables efficient searches: locating a k -mer on an index with N k -mers in pages of P k -mers requires in the order of $\log_p N$ page accesses, typically 3–4 disk random reads for billion-scale indices. A small least recently used (LRU) cache maintains frequently accessed pages in memory.

Barcode preprocessing. Cell barcodes are extracted according to technology-specific configurations and matched via exact lookup against whitelists (for example, 10x Chromium v2/v3 barcode sets). All barcode– k -mer associations are stored during indexing. Barcodes with Hamming distance >0 from the whitelist are discarded; error correction is not performed. For technologies that do not provide a vendor-supplied barcode whitelist, such as split-seq or sci-seq variants, barcodes must be resolved as a preprocessing step before indexing. Tools such as flexiplex⁴⁴, matchbox⁶⁰, or splitcode⁶¹ can convert non-standard barcode layouts into whitelist-compatible R1/R2 FASTQ pairs, after which Malva indexes the data in the standard way. The datasets currently in Malva Index were generated using protocols for which barcodes can be extracted as fixed-length sub-sequences (for example, 10x Chromium, Visium, Stereo-seq), or, in the case of Smart-seq2, where one index per cell is generated and subsequently merged using the malva combine command.

Prior to cell-type annotation, we apply a UMI-based threshold method to retain barcodes that are likely to correspond to true cells, analogous to the order of magnitude (OrdMag) algorithm used by Cell Ranger⁶²: barcodes are ranked by descending total UMI counts, and an expected number of cells is estimated by minimizing the squared relative difference between a candidate cell count x and the number of barcodes exceeding a UMI threshold derived from x . Specifically, for each candidate x , we compute the 99th percentile UMI count among the top x barcodes, divide this value by 10 to obtain a threshold, and count barcodes exceeding this threshold. The optimal x minimizes $(f(x) - x)^2/x$, where $f(x)$ is the count of barcodes above the threshold. Once the expected cell number is determined, the final UMI threshold is set as the 99th percentile of the top barcodes divided by 10, and all barcodes with UMI counts above this threshold are retained as cells.

Sequence querying and pseudoquantification

Malva identifies query sequences by decomposing them into k -mers and searching Malva Index for matches. For queries longer than k , we utilize a sliding window approach to ensure all possible k -mer alignments are considered. Given a query sequence q and window size w (default $w = 64$), we extract all windows of length w and evaluate the fraction of matching k -mers in each window.

The match score for a window W and cell C is defined as the ratio of matching k -mers to total possible k -mers: $\text{score}(W, C) = \frac{|M(W, C)|}{w - k + 1}$, where $M(W, C)$ represents the set of k -mers in W that appear in cell C . By default, we mask low-complexity k -mers using dustmasker v1.0.0 with default parameters prior to scoring⁵⁵. Optionally, users can further reduce false positives from repetitive sequences by excluding k -mers that appear frequently across the reference transcriptome; this is achieved by counting k -mer occurrences across all transcripts and ignoring those exceeding a user-specified threshold during query, though this option is not enabled by default. Windows with scores exceeding threshold τ (default $\tau = 0.65$) are considered positive matches. For sequences longer than w , presence in a cell is determined as soon as any single window passes the threshold.

This approach tolerates sequencing errors and mutations: a single nucleotide change affects only one k -mer, allowing the remaining k -mers to drive detection. The complete query process aggregates matches across all windows, returning cells ranked by the number of significant windows. For typical use cases ($w = 64$, $k = 24$, $\tau = 0.65$), the method reliably detects sequences with up to ~15% divergence while maintaining low false positive rates. Query optimization includes batching k -mer lookups to minimize disk access and exploiting the sorted index structure for cache efficiency. Complete specifications of the batching strategy, index access patterns, and caching parameters are provided in Supplementary Methods.

Sequence coverage quantification

For coverage analysis, we preserve spatial resolution along transcripts by reporting k -mer matches for each sliding window independently. We apply a sliding average (default window: 24 bp) to smooth the coverage profile, producing visualizations comparable to those from alignment-based tools. As with pseudoquantification, we mask low-complexity k -mers using dustmasker v1.0.0⁵⁵ during coverage calculation to prevent artifactual peaks. Without masking, certain positions can show 100–1,000× spikes due to sequences shared across many genes. This positional information enables detection of alternative polyadenylation sites, splice junction usage, among others.

Benchmarking

We evaluated Malva's performance against established tools for both indexing speed and query accuracy. For indexing benchmarks, we compared Malva with kallisto-bustools v0.28.2²², MetaGraph v0.4.1 (Row-Diff/Multi-BRWT variant, indexing only)²⁷, BLAST+ v2.16.0 (MEGABLAST variant, indexing only)⁶³ and a STAR (v2.7.11b)-based workflow^{23,64}, Fulgor v4.0.0 (meta-diff variant)³¹, and REINDEER v1.4.7-1-g0812ad1²⁸, on various datasets ranging sequencing depths and number of unique labels (spatial locations). For sequence search capabilities, we compared against MetaGraph, BLAST+, Fulgor and REINDEER.

For benchmarking of computational complexity, dataset selection varied by method due to architectural constraints: for MetaGraph and BLAST, we used the Stereo-seq CIRSTA mouse liver atlas (CNCBdb accession: CNP0003447), containing ~60 billion reads (7 TB of raw, compressed data) across 20 samples with ~4 billion unique spatial barcodes⁶⁵. REINDEER and Fulgor require treating each cellular barcode as a separate input file, making analysis of millions of cellular barcodes impractical due to filesystem overhead and preprocessing requirements (REINDEER additionally requires running bcalm⁶⁶ on each sample independently). We therefore benchmarked these methods on a composite dataset from 6 Visium samples (accession numbers: GSM5115019, GSM5115020, GSM5115021, GSM5115022, GSM5115023, GSM5115024) totalling ~1.5 billion reads (~180 GB of raw, compressed data) across ~30,000 unique spatial locations⁶⁷, providing comparable read depth while maintaining feasible file counts. These datasets tested the ability to handle varying spatial resolutions (0.6 μm in Open-ST to 55 μm in Visium), sequencing depths, and total data volumes. For large-scale benchmarking, we additionally selected 1% of datasets at random from Malva Index (~1.5 TB of raw data) and reindexed them with Malva and Fulgor, the two methods with best runtime and memory usage during indexing and querying, using identical settings as in other experiments. Each benchmark was run independently and was allocated at most 96 threads and 512 GB of memory on a multi-socket server with Intel Xeon Platinum 8454H CPUs (256 threads total) and 4 TB RAM, with resource monitoring via snakemake v7.32.4⁶⁸.

Validation of biological signal preservation

We evaluated whether Malva's k -mer-based pseudocounts produce biologically meaningful results comparable to traditional pipelines. We selected representative datasets spanning high- and lower-resolution

spatial transcriptomics, healthy and disease samples: an Open-ST tumour section (high spatial resolution, small scale⁶⁹; SRA accession SRX23008310), 10x v3 embryonic (E13.5) mouse midbrain (mouse single-cell, small scale⁷⁰; SRA accession SRR11947658), 10x v3 human PBMCs (human single-cell, small scale; downloaded from https://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_1k_v2/pbmc_1k_v3_fastqs.tar), and Visium mouse brain (low spatial resolution, small scale, downloaded from https://s3-us-west-2.amazonaws.com/10x.files/samples/spatial-exp/1.3.0/Visium_Adult_Mouse_Brain/Visium_Adult_Mouse_Brain_fastqs.tar). Across datasets, we compared Malva pseudocounts to the original deposited count matrices. For each, we computed per-cell gene counts and calculated correlations between total counts per gene and per cell.

To assess whether pseudocounts support downstream analyses, we performed clustering on all benchmarked datasets using species-specific Ensembl transcript sequences after masking low-complexity and repetitive *k*-mers. We applied standard scanpy processing (normalization, log transformation, highly variable gene selection, PCA, nearest neighbour graph construction, and Leiden clustering via `leidenalg` v0.10.2) and compared cluster assignments between Malva and traditional pipelines using Normalized Mutual Information. Additionally, the clusters from the reference objects were merged to the Malva objects by cell barcode, and used to assess the similarity marker gene detection between Malva and original counts, by applying a Wilcoxon rank-sum tests (one cluster versus the rest), then measured the concordance of the Fold Changes for each gene in each comparison.

For trajectory analysis, we applied the scanpy v1.11.4 implementation of diffusion pseudotime (DPT) on the single-cell E13.5 mouse midbrain dataset^{71,72}. Analyses were restricted to the neuronal subset, and we selected a representative radial glia cell as the root for pseudotime inference. Pseudotime ordering was computed independently on both the Malva pseudocount matrix and the reference UMI-based quantification. To enable direct comparison, we merged results by cell barcode and quantified similarity between pseudotime trajectories.

Simulation of sequencing errors. To evaluate Malva's robustness to sequencing errors, we introduced synthetic errors into the benchmark datasets (Open-ST human tumour, 10x v3 human PBMCs, and Visium mouse brain) at rates ranging from 0.1% to 10%. These synthetic errors are cumulative to the existing errors of the data induced from the current short-read sequencing platforms. For each error rate, we randomly selected positions within reads and substituted nucleotides with uniform probability across the three alternative bases. We then rebuilt Malva indices from the error-containing reads and repeated our standard benchmarking analyses: per-cell and per-gene correlation with reference counts, clustering agreement (NMI), and marker gene detection concordance. For variant detection benchmarks, we additionally assessed sensitivity and specificity of the cell line mixture probes across error rates using both overlapping and non-overlapping *k*-mer indexing strategies.

Sequence-based variant detection validation. To validate Malva's accuracy for sequence-based variant detection, we compared against Flexiplex (v1.02.3), a specialized sequence search tool⁴⁴, using a seven-cell-line mixture dataset (SRR10971813). The dataset contains known sequence variants: MCF7 cells expressing a *BCAS4-BCAS3* fusion gene, HEK293T cells containing adenovirus E1A sequences, and T47D cells carrying mitochondrial SNV rs878887783.

We searched for four target sequences: *BCAS4-BCAS3* SNV variant (CCGATCCTGGGGCCGAGGTACCTTTGACAGGAGC), Adenovirus 5 E1A (TTTGGACTTGAGCTGTAAACGCCCCAGGCCATAA) and a rs878887783 probe (ATTACTCTGCCATCATGACCCCTGGCCATAATATGATTATCT). Compared to the Flexiplex benchmarking set-up, for the *BCAS4-BCAS3*

variant, we excluded the reference (CCGAGCCTGGGGCCGAGGTACCTTTGACAGGAGC), as no MCF7 cells are expected to express it. For the mitochondrial variant, we used a truncated version of the original probe as the full sequence from the Flexiplex paper yielded zero counts when following their supplementary protocol, though our modified probe produced similar detection numbers to their reported results.

Flexiplex was run with default parameters using grep-based detection mode. For output processing, we extracted cellular barcodes by matching read identifiers from Flexiplex results to the corresponding R1 FASTQ file and extracting the first 16 nucleotides as cellular barcodes. For Malva, we used a sliding window size of 24 nucleotides with 100% *k*-mer presence threshold. Malva results directly return cell barcodes and *k*-mer match counts. Wall-clock time was measured for both workflows including index construction (only in the case of Malva, done once) and query processing (for each probe). Sequence search performance metrics (true positives, false positives and false negatives) were calculated by comparing detected cells against expected cell line identities for each target sequence, from the annotated objects deposited by the authors.

To further validate Malva's ability to detect clinically relevant somatic mutations, we analysed a lung adenocarcinoma dataset (PRJCA001731⁷³) containing matched tumour-normal pairs with characterized *EGFR* mutation status. From the archived datasets, four tumours carried *EGFR* exon 19 deletions and one carried the L858R point mutation (c.2573T>G) in exon 21. We first established ground truth by aligning reads with STAR (v2.7.11b) to GRCh38 and visualizing coverage across the *EGFR* locus. We then designed Malva probes targeting: (1) exons 18–21 as positive controls for *EGFR* expression; (2) the wild-type exon 21 sequence; and (3) a 45-bp probe centred on the L858R variant (GTCAAGATCACAGATTTGGGCGGGCCAACTGCTGGGTGCGGAA), with a pair-matched wild-type probe (GTCAAGATCACAGATTTGGGCTGGCCAACTGCTGGGTGCGGAA). Probes were queried against sample-specific Malva indices with parameters $w = 24$, $\tau = 1.0$ for variant-specific detection.

Specialized sequence queries across Malva Index. *Large-scale arbitrary nucleotide sequence search.* To systematically screen for exogenous sequences, we compiled viral genomes from NCBI RefSeq: SARS-CoV-2 (NC_045512.2), hepatitis C virus (NC_004102.1), measles virus (NC_001498.1), West Nile virus (NC_001563.2) and Zika virus (NC_012532.1), as negative controls; HERV-K consensus sequences; PhiX174 (NC_001422.1) as a positive control; and the lentiviral vector pBA439 (AddGene #85967). We queried these sequences against the complete Malva Index with default settings ($k = 24$, window size = 64, threshold = 0.65) and generated tissue- and study-level coverage profiles. We additionally screened *Mycoplasma* sequences because these bacteria are frequent, often undetected cell culture contaminants. Literature and guidance documents consistently report a small set accounting for most laboratory contamination events including *Mycoplasma hyorhinis*, *Mycoplasma hominis*, and *Mycoplasma fermentans* in human and other mammalian cell cultures, and also included *Mycoplasma agassizii*, a respiratory pathogen of turtles, not associated with human or standard mammalian cell culture workflows.

Germline variant query probe design. To systematically screen for germline genetic variants across Malva Index, we generated sequence probes targeting common SNPs in protein-coding regions. We downloaded phase 3 data from the 1000 Genomes Project across human chromosomes 1–22 and filtered for biallelic SNPs with MAF $\geq 1\%$ located within exons or UTRs of protein-coding genes as defined by GENCODE v19 annotations. For each qualifying variant, we extracted the genomic context from the GRCh37 reference genome and constructed 64-nucleotide probes centred on the variant position, generating paired probes containing either the reference or alternative allele.

We annotated variant pathogenicity using ClinVar classifications, categorizing variants as ‘pathogenic’, ‘likely pathogenic’, ‘variant of uncertain significance’ (VUS), ‘likely benign’, ‘benign’ or ‘unknown’ based on clinical significance scores. To enable evolutionary analysis, we incorporated ancestral allele information from the 1000 Genomes Project ancestral sequence alignments, determining whether the reference or alternative allele represented the ancestral versus derived state for each variant. Quality control filters removed probes with extreme GC content (<30% or >70%) to ensure reliable detection. The final probe set contained 296,985 variant pairs spanning all chromosomes.

We queried these variant probes against the complete Malva Index using the standard API, retrieving cell-level match counts for both reference and alternative alleles. For evolutionary analysis, we calculated derived allele frequencies and compared observed frequencies in healthy versus disease samples to expected population frequencies, testing for systematic deviations that might indicate selection effects using the ancestral allele annotations.

Isoform usage. To identify cell-type-specific isoform usage, we analysed the Tabula Muris dataset which provides full-length transcript coverage suitable for exon-level quantification. We used mouse genome annotations (GRCm39.113) and queried individual genes against sample-specific Malva indices with standard coverage analysis parameters. For each gene, we merged overlapping annotated exons to create consensus gene models, then quantified average coverage per exon across cell types.

We implemented a nested marker analysis approach: first identifying marker genes using one-versus-all differential expression testing with the Wilcoxon rank-sum test and Benjamin-Hochberg multiple testing correction (log₂ fold change ≥ 1.0, adjusted *P* value ≤ 0.05). For genes identified as cell-type markers, we performed exon-level differential analysis within marker cell types to identify marker exons: exonic regions showing cell-type-specific expression patterns within broadly expressed genes. This approach enabled detection of exon specificity in specific cell types despite the host gene being a marker gene shared by more cell types.

Additionally, we analysed 3′ UTR length regulation using the MOSTA Stereo-seq mouse embryonic development atlas spanning E9.5–E15.5. For validation, we aligned the same sequencing reads using STAR (v2.7.11b) with default parameters to generate coverage tracks for comparison with Malva results. We performed standard Malva coverage-like analysis targeting the *Add2* locus, focusing on proximal and distal polyadenylation sites previously identified in bulk and single-cell RNA-seq studies. We quantified proximal versus distal 3′ UTR usage by computing coverage ratios in defined UTR regions, then mapped these expression patterns spatially to identify tissue-specific UTR isoform preferences during embryonic development.

Detection of circRNAs. To interrogate the expression of the circRNA *CDRIAs* across Malva Index, we designed a 36-nucleotide probe centred on its backsplice junction. Therefore, the probe sequence spans 18 nucleotides upstream of the splice donor site and 18 nucleotides downstream of the splice acceptor site, capturing the unique junction sequence that distinguishes circular from linear transcripts. We used $w = 24$, $\tau = 0.65$ for querying Malva Index.

Benchmarking polyadenylation site detection against polyASite.

To evaluate Malva’s ability to detect alternative polyadenylation events, we leveraged polyASite v3.0⁵¹, a database of polyadenylation sites identified from a large set of samples from the Human Cell Atlas (also ingested in Malva Index) using 3′ end sequencing protocols. For each annotated polyA site, we designed two probe types: (1) polyA probes consisting of 44 bp of upstream genomic sequence followed by a 20-bp polyA tract (20*A); and (2) control probes consisting of 64 bp of genomic sequence around the same position, but without the polyA signal. We queried both probe sets against the same HCA samples used to build polyASite and computed Spearman correlations

between Malva pseudocounts and polyASite usage scores within each tissue type. Correlations were computed per-probe per-tissue rather than globally, as pseudocounts were not depth-normalized; this approach assesses the concordance of relative usage patterns rather than absolute quantification. Probe sequences are publicly deposited (Data availability).

Large-scale somatic mutation detection and benchmark against scTML.

To systematically benchmark Malva’s mutation detection capabilities at scale, we utilized the scTML database⁵³, which provides uniformly processed single-cell RNA-seq data from 280 tumour samples across 16 cancer types with matched variant calls. The scTML pipeline utilized GATK HaplotypeCaller for variant detection from aligned reads, with subsequent filtering for quality and read support. We extracted SNVs located in exonic regions or UTRs (excluding intronic variants due to lower sequence complexity and repeat content) and constructed 45-bp probes centred on each variant position. For each probe, we queried the corresponding sample’s Malva index and classified cells as variant-positive based on *k*-mer matches ($w = 24$, $\tau = 1.0$). Sensitivity was computed as the fraction of scTML-positive cells detected by Malva; specificity was computed from cells in matched normal samples or variant-negative tumour cells.

Reference-free analysis using *k*-mer composition

Traditional single-cell analysis relies on mapping reads to reference genomes and counting annotated genes. However, this approach cannot capture the full diversity of RNA sequences, particularly in non-model organisms or when studying unannotated transcripts. We developed a reference-free analysis pipeline that clusters cells based on their complete *k*-mer repertoire, revealing relationships invisible to gene-based methods.

Bucketing approach for scalable *k*-mer analysis. Direct analysis of *k*-mer composition faces a fundamental challenge: even modest datasets contain billions of unique *k*-mers, making cell-by-*k*-mer matrices computationally intractable. To address this, we implemented a bucketing strategy that groups similar *k*-mers while preserving sequence relationships.

The bucketing algorithm uses a MinHash-based approach to assign *k*-mers to a fixed number of buckets based on their *w*-mer composition, where $w < k$. For each *k*-mer, we extract all possible contiguous *w*-mers (typically $w = 16$ for $k = 24$) and compute two independent hash functions (MurmurHash3 and xxHash64) for each *w*-mer. We retain the minimum hash value from each function across all *w*-mers in the *k*-mer, then combine these minima into a composite signature. The final bucket assignment is determined by taking the modulo of this signature with the predefined number of buckets (default 100,000).

This approach promotes that *k*-mers with similar *w*-mer composition are assigned to the same or nearby buckets. The resulting cell-by-bucket matrix reduces dimensionality from billions of unique *k*-mers to, for example, 100,000 buckets while preserving sequence relationships. Importantly, bucket assignments are deterministic and consistent across datasets, enabling cross-dataset and cross-species comparisons. We process *k*-mers in chunks and accumulate bucket counts per cell using sparse matrix operations, allowing analysis of datasets with hundreds of millions of cells.

Local structure preservation analysis. To validate that *k*-mer bucket representations preserve biologically meaningful cellular relationships, we compared local neighbourhood structure between sequence-based and gene count PCA embeddings using highly variable features. For each cell in the overlapping dataset, we identified *k*-nearest neighbours in both representations ($k = 15$). We then computed Wasserstein distances between: (1) distances from each cell to its *k*-nearest

neighbours in sequence space; and (2) distances from that cell (in sequence space) to neighbours identified in gene count space. The resulting distribution was compared against a random null model using ROC-AUC analysis, yielding Local Structure Agreement scores where values $\gg 0.5$ indicate better-than-random preservation of local cellular relationships.

Cell clustering and marker sequence discovery. Using the cell-by-bucket matrix, we follow a standard single-cell analysis workflow implemented in scanpy. We normalize bucket counts to account for sequencing depth differences, then select highly variable buckets based on their dispersion relative to mean expression. After dimensionality reduction with PCA (retaining 30 components), we construct a nearest neighbour graph and apply the Leiden algorithm via `leidenalg` v0.10.2 to identify cell clusters with similar k -mer profiles. Clustering similarity between k -mer profiles and gene count inputs were quantified using the Normalized Mutual Information metric:

$$\text{NMI}(Y_1, Y_2) = \frac{2 \times I(Y_1; Y_2)}{[H(Y_1) + H(Y_2)]}$$

where Y_1 and Y_2 are two clusterings to be compared, H is the entropy function and $I(Y_1; Y_2) = H(Y_1) - H(Y_1|Y_2)$ is the mutual information function.

To understand the biological basis of these clusters, we identify differentially abundant buckets between groups using a t -test with Benjamini–Hochberg correction (selecting features with adjusted P value < 0.05 and $\log_2$ fold change > 1). We then retrieve the k -mers assigned to these differential buckets and perform targeted de novo assembly using SPAdes v4.2.0 in RNA mode⁷⁴.

To determine genomic origin and interpretability, contigs were annotated in two passes. First, contigs were aligned to the appropriate host references using `minimap2` (v2.1.1-r341) with spliced presets for genome mapping and ungapped mapping to transcriptomes, against GRCh38. For transcript-level alignment we mapped to Ensembl cDNA/lncRNA catalogues, for alignments with high mapping quality. Based on these, contigs were classified as: (1) genic if overlapping annotated transcripts; (2) intergenic if mapping uniquely outside annotated features; or (3) unmapped if no confident hit was found. For loci with multiple contigs, we compared per-cell Malva coverage tracks across the locus to distinguish isoform-specific versus locus-wide signal. Second, contigs without confident host assignment (or with rRNA signatures) were analysed using `Kraken2` (v2.1.5) with a standard RefSeq PlusPF database (bacteria, archaea, viral, human, UniVec/adaptor sequences). We recomputed the expression using the annotated and unannotated assembled contigs, and compared cell-wise totals and feature-wise abundances.

We also evaluated Malva in a setting with limited genomic reference resources, by processing publicly available *Placozoa* single-cell RNA-seq data⁷⁵ (SRA project PRJNA982113). Reads were indexed with the same k -mer configuration used elsewhere ($k = 24$) and embedded using our sequence-bucket approach. Dimensionality reduction and clustering followed the standard scanpy single-cell analysis workflow outlined above.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Malva platform is accessible at <https://malva.mdc-berlin.de>. Malva Index is continuously updated; the current version and statistics are available through the API. Queries used for evaluation of the large language model engine and sequence probes for alternative polyA site

usage constructed from the polyASite database have been publicly deposited at <https://zenodo.org/records/19455194> (ref. 76). A list of all data sources and software used in context of this work is available as Supplementary Information.

Code availability

The source code and packaged implementation of the Malva indexing and querying algorithms are available on GitHub (<https://github.com/malva-bio/malva>; v1.0.0); Malva is freely available for non-commercial, non-diagnostic, non-clinical, academic research use only. Documentation is available online (<https://malva.readthedocs.io>). A Python client to access Malva Index programmatically (https://github.com/malva-bio/malva_client; v0.3.4) and its documentation (<https://malva-client.readthedocs.io>) are available online.

55. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
56. Dubey, A. et al. The Llama 3 Herd of Models. Preprint at <https://doi.org/10.48550/arxiv.2407.21783> (2024).
57. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
58. Gonçalves, R. S. et al. The text2term tool to map free-text descriptions of biomedical terms to ontologies. *Database* **2024**, baee119 (2024).
59. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods* **21**, 1462–1465 (2024).
60. Schuster, J. et al. Powerful read processing with matchbox. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.11.09.685711> (2025).
61. Sullivan, D. K. & Pachter, L. Flexible parsing, interpretation, and editing of technical sequences with splitcode. *Bioinformatics* **40**, btac331 (2024).
62. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
63. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
64. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
65. Wu, B. et al. A spatiotemporal atlas of cholestatic injury and repair in mice. *Nat. Genet.* **56**, 938–952 (2024).
66. Chikhi, R., Limasset, A. & Medvedev, P. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics* **32**, i201–i208 (2016).
67. Ma, F. et al. The cellular architecture of the antimicrobial response network in human leprosy granulomas. *Nat. Immunol.* **22**, 839–850 (2021).
68. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
69. Schott, M. et al. Open-ST: high-resolution spatial transcriptomics in 3D. *Cell* **187**, 3953–3972.e26 (2024).
70. La Manno, G. et al. Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).
71. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
72. Haghverdi, L., Büttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
73. He, D. et al. Single-cell RNA sequencing reveals heterogeneous tumor and immune cell populations in early-stage lung adenocarcinomas harboring EGFR mutations. *Oncogene* **40**, 355–368 (2021).
74. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
75. Najle, S. R. et al. Stepwise emergence of the neuronal gene expression program in early animal evolution. *Cell* **186**, 4676–4693.e29 (2023).
76. León Perinán, D., Karaiskos, N. & Rajewsky, N. Malva polyA query probes and LLM evaluation [data set]. *Zenodo* <https://doi.org/10.5281/zenodo.19455194> (2026).
77. Purkayastha, A. et al. Direct exposure to SARS-CoV-2 and cigarette smoke increases infection severity and alters the stem cell-derived airway repair response. *Cell Stem Cell* **27**, 869–875.e4 (2020).

Acknowledgements We thank the members of the Rajewsky laboratory, in particular the Open-ST team, for constructive discussions about the applications of Malva, for testing and for sharing unpublished datasets that supported tool development; D. Munteanu for support with computational infrastructure; K. Yperman and N. Shalaby; the Digital Health Accelerator programme from the Berlin Institute of Health; and C. Marchet for insightful discussions. We further acknowledge the Human Cell Atlas Data Portal (Creative Commons Attribution 4.0 International license), NCBI SRA, CNGBdb and the Tabula Muris consortium for providing publicly available datasets that were instrumental for building and validating Malva Index. Figures 1, 2 and 5 and Extended Data Fig. 1 use icons that have been adapted from free with attribution resources made by UIUX Mall, Freepik, Vectorslab, Dinosoftlabs and Ylivdesign downloaded from <https://flaticon.com>.

Author contributions D.L.-P. developed the Malva software, co-generated Malva Index and performed computational analyses. N.K. co-developed the Malva software in regard to testing and maintainability, and co-generated Malva Index. N.K. and N.R. supervised the project,

Article

supported the interpretation of results and conceptualized applications. All authors conceived the project, wrote and approved the manuscript.

Funding D.L.-P. was financially supported by the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS) programme of the Helmholtz Association. N.R. thanks Helmholtz HPC Gateway (grant number 3.11414SFHPC404). N.R. thanks the Leibniz prize of the Deutsche Forschungsgemeinschaft (DFG) (grant number RA 838/5-1). Open access funding provided by Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft (MDC).

Competing interests D.L.-P., N.K. and N.R. are listed as inventors of a patent application relating to the work (WO2026/132228). The patent application was submitted through the Innovation

and Entrepreneurship Office of the Max-Delbrück Center, with the MDC being the patent applicant. N.R. is a member of the scientific advisory board of bioptimus.com.

Additional information

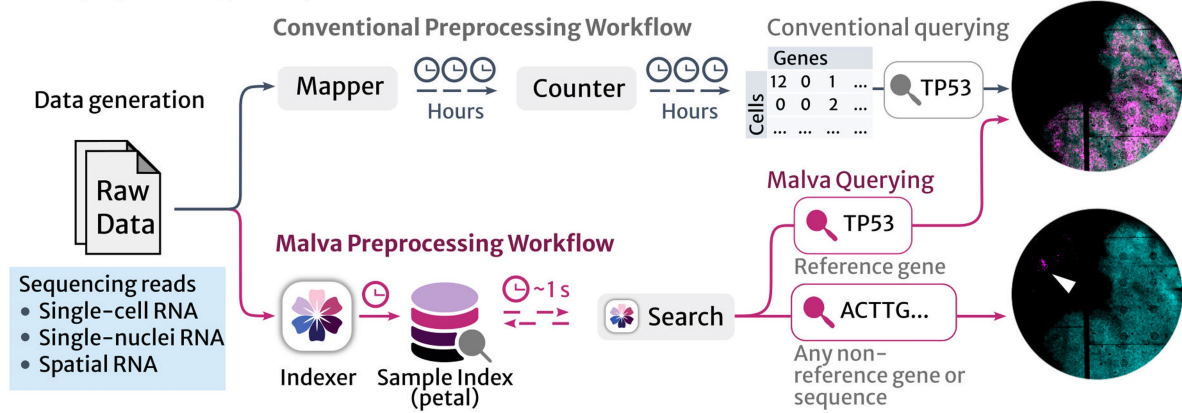
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-10975-w>.

Correspondence and requests for materials should be addressed to Daniel León-Periñán, Nikos Karaiskos or Nikolaus Rajewsky.

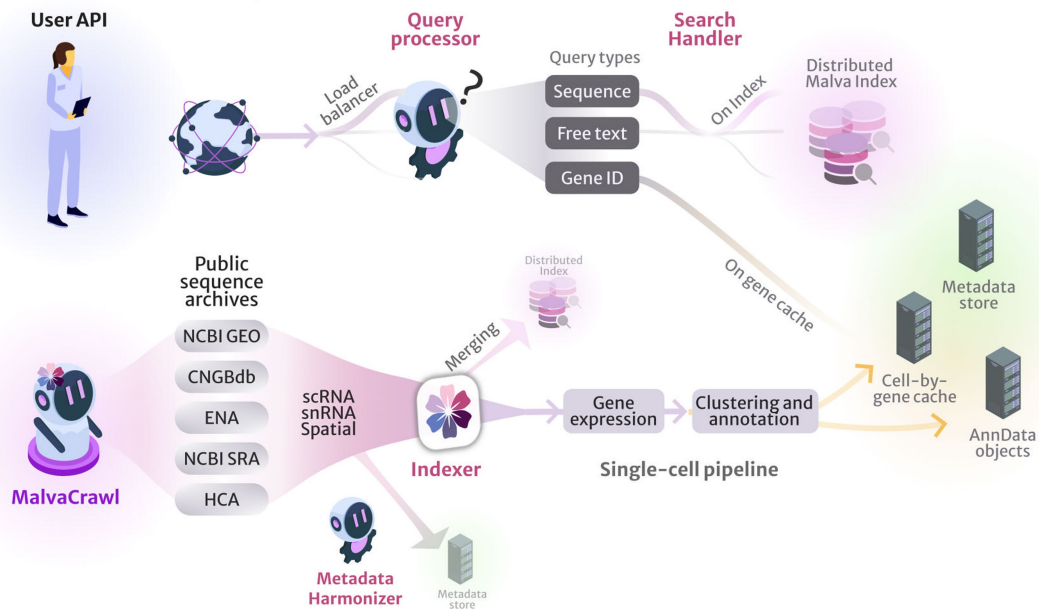
Peer review information *Nature* thanks Rob Patro who co-reviewed with Yuan Gao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

a Sample processing & analysis workflows

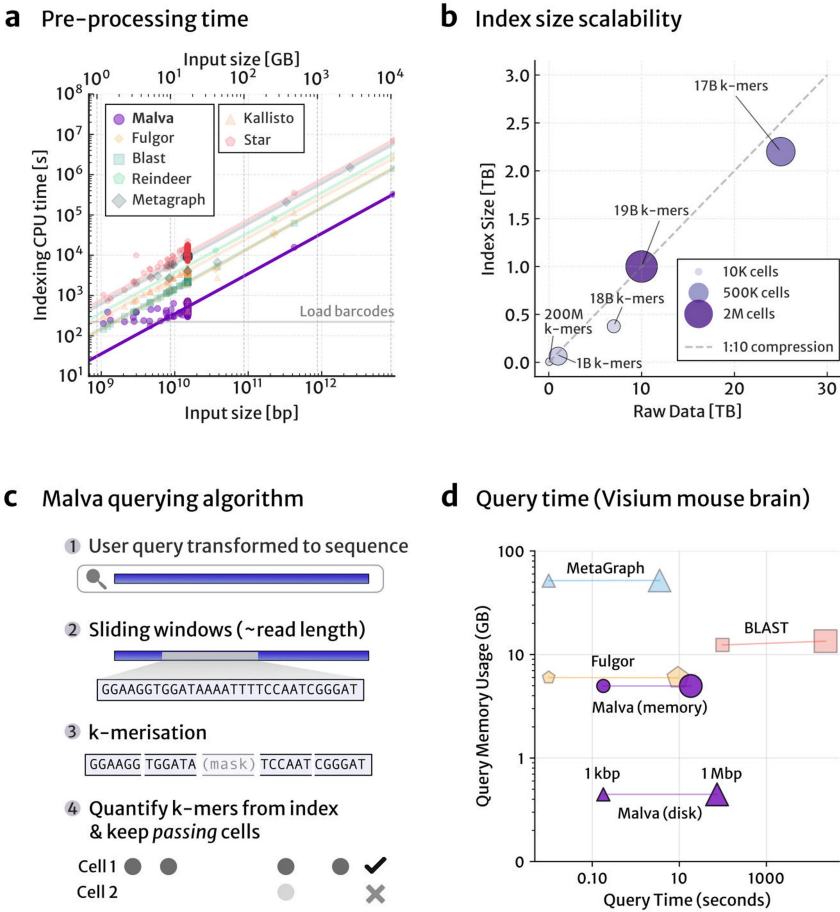


b <https://malva.mdc-berlin.de> infrastructure



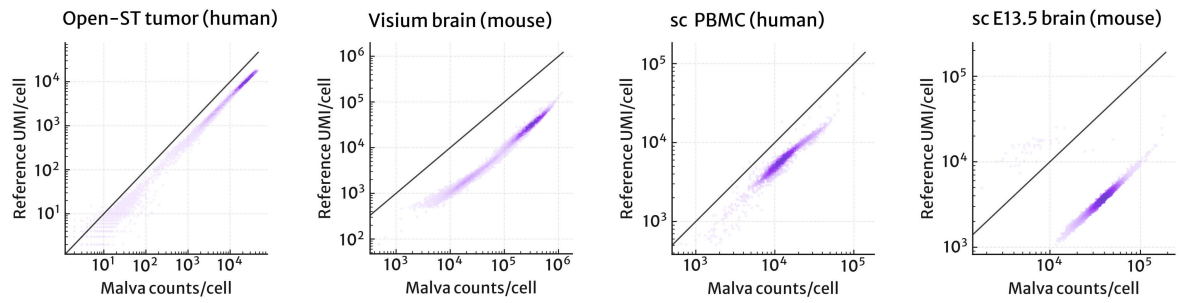
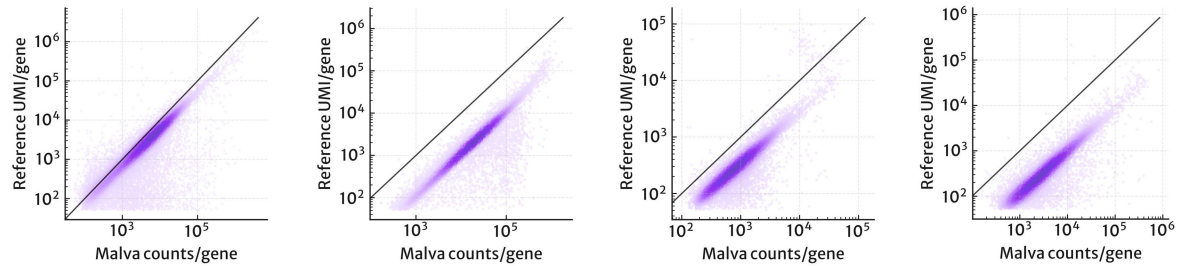
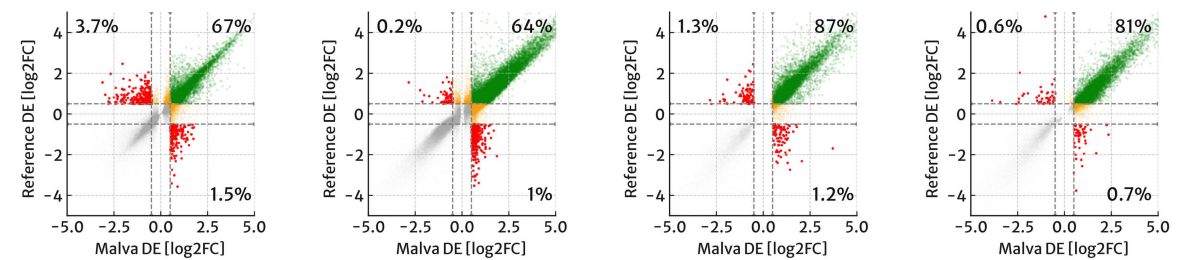
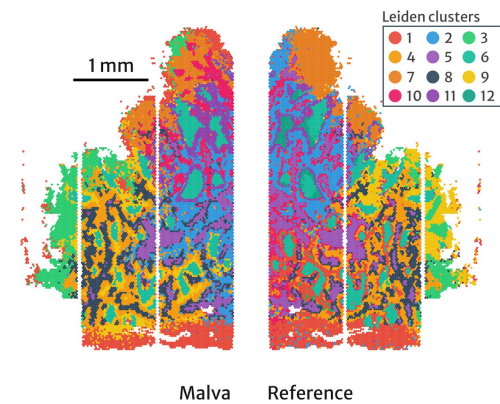
Extended Data Fig. 1 | Malva indexing and querying workflow. (a) Schematic of how sequencing reads (single-cell, single-nuclei, or spatial RNA) are transformed into searchable indices. In contrast to conventional preprocessing (mapping and counting), Malva builds a sample-level index, enabling direct sequence queries without alignment. These sample indices are subsequently merged into the global Malva Index, partitioned for scalability across many datasets. (b) Overview of the <https://malva.mdc-berlin.de> infrastructure. In the upper branch, users submit queries through the API, which are processed by a rule-based engine that routes requests to sequence search, the gene-expression

cache, or the language-model module. In the lower branch, the MalvaCrawl service continuously ingests raw data from public archives (e.g. GEO, SRA, ENA, CNGBdb, HCA), builds sample indices, and merges them into the distributed Malva Index. Metadata are harmonized at the sample level and augmented with cell-level annotations generated by Malva's internal single-cell pipeline. These data are stored as standardized objects (e.g. AnnData) for programmatic access. Abbreviations: API, Application Programming Interface; scRNA, single-cell RNA; snRNA, single-nuclei RNA.

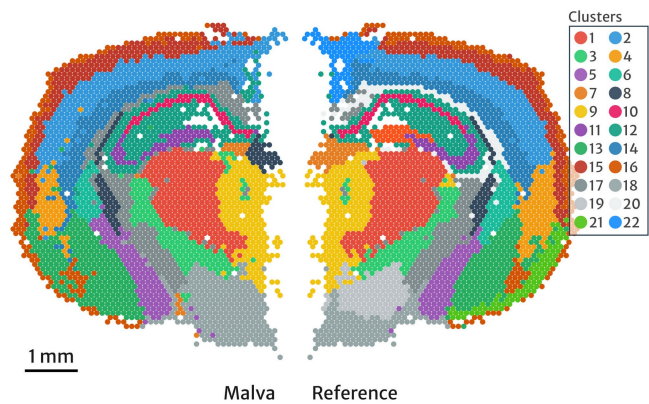
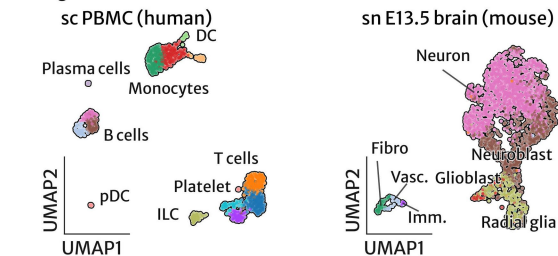
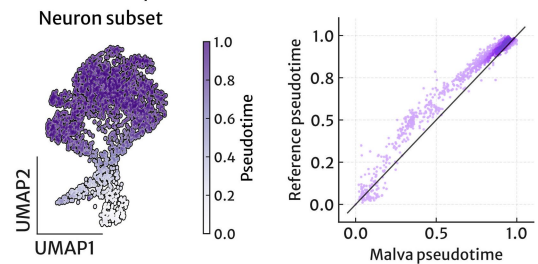


Extended Data Fig. 2 | Malva indexing and query performance.
(a) Preprocessing time comparison across representative datasets comparing alignment-based pipelines (kallisto, STAR) with indexing methods (Malva, MetaGraph, BLAST, REINDEER, Fulgor). Malva achieves the best performance, with runtime dominated by FASTQ decompression rather than algorithmic overhead. (b) Index size scaling across datasets with different cell numbers (bubble size) and sequencing depths. Storage requirements depend on the number of unique k-mer counts and sequencing depth per cell rather than on total cell count. For example, datasets with similar -18 billion k-mers yield

comparable index sizes despite differing in cell numbers (~90,000 vs ~2 million). Compression ratios remain around 1:10. (c) Overview of the Malva querying algorithm. User queries are decomposed into k-mers using sliding windows, transformed into masked k-mers, and matched against Index with cell-barcode filtering to retain only positive cells. (d) Query cost (CPU time and memory usage) across various query lengths (1 kb to 1 Mb), comparing Malva to BLAST, MetaGraph and Fulgor on a single 10x Visium adult mouse brain sample (4,992 unique spatial locations), across on-disk or in-memory modes (Methods).

a Per-cell count accuracy**b** Per-gene count accuracy**c** Marker gene detection similarity**d** Open-ST spatial cluster (NMI: 0.55)

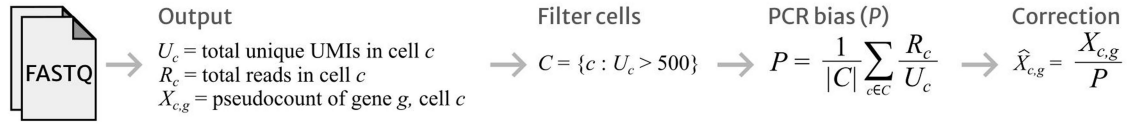
Visium mouse brain (NMI: 0.812)

**e** Original clusters on Malva UMAP**f** Neuronal development E13.5 brain – on Malva UMAP

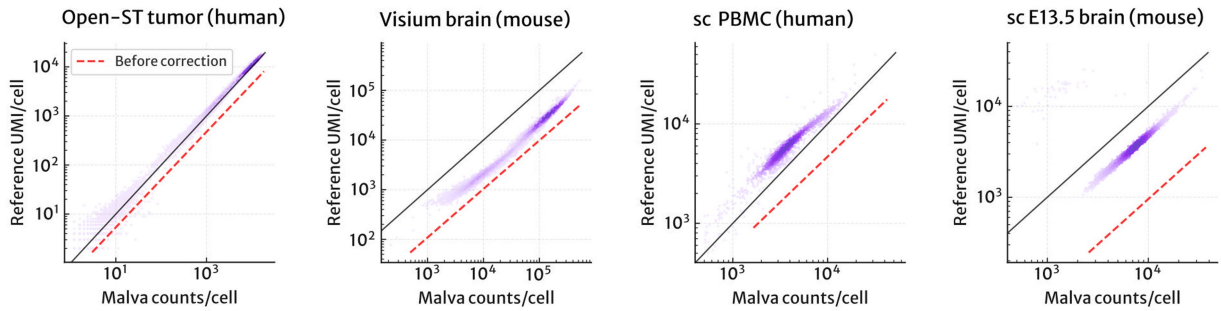
Extended Data Fig. 3 | Validation of Malva quantification accuracy across technologies and contexts. (a) Per-cell count correlation between Malva pseudocounts and reference counts (UMI) across four representative datasets: Open-ST human tumor, Visium mouse brain, single-cell human PBMC, and single-cell embryonic mouse brain. Reported *R* values show Pearson correlations with (*R*_c) and without (*R*_f) RANSAC filtering. (b) Per-gene count correlations across the same four datasets. (c) Concordance of marker gene differential expression between annotated cell types. Scatterplots show log₂ fold-change from Malva pseudocounts versus reference UMs. Green: consistent directionality; red: inconsistent; orange: consistent direction but below the significance threshold in one method. Analyses are restricted to genes significant differential in Malva (adjusted p-value < 0.05). Percentages indicate the fraction of genes with concordant vs. discordant directionality among genes with |log₂FC| > 0.5. (d) Comparison of cluster results for Open-ST

human tumor and Visium mouse brain datasets. Malva-based clustering (left) and reference clustering (right) yield similar structures; spatial coordinates are mirrored for visualization. Leiden clustering used identical parameters, with palettes harmonized for comparison. Reported NMI values quantify agreement. (e) UMAPs computed from Malva pseudocounts, colored by original author-provided cell-type annotations, for PBMC (human) and embryonic mouse brain (sn E13.5). (f) Pseudotime inference on the E13.5 embryonic mouse brain. Left: neuronal subset embedded with Malva pseudocounts; right: correlation between pseudotime computed from Malva versus reference counts (UMIs). Abbreviations: DE, Differential Expression; NMI, Normalized Mutual Information; UMI, Unique Molecular Identifier; PBMC, Peripheral blood mononuclear cell; UMAP, Uniform Manifold Approximation and Projection; sc, single-cell.

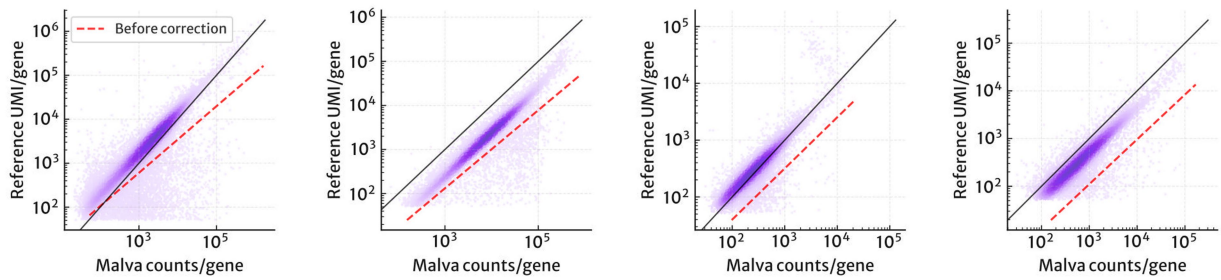
a Saturation-based pseudocount correction factor



b Per-cell count accuracy (corrected by PCR bias)

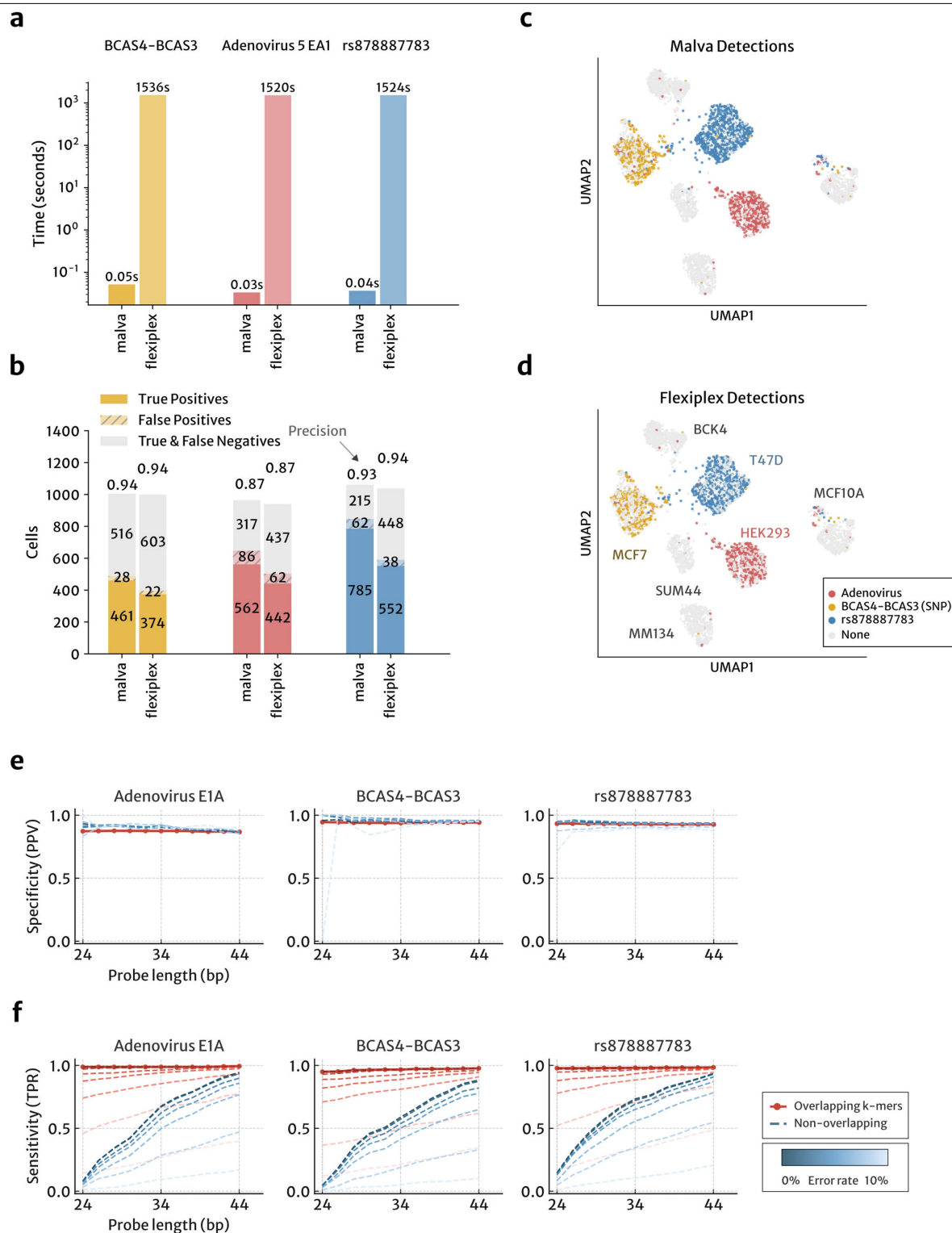


c Per-gene count accuracy (corrected by PCR bias)



Extended Data Fig. 4 | Saturation-based correction of Malva pseudocounts for PCR bias. (a) Workflow for saturation-based pseudocount correction. For each cell, total reads (R) and unique UMIs (U) are counted directly from FASTQ files without alignment. Cells with sufficient coverage (e.g., $U > 500$) are used to compute the PCR bias factor (P), defined as the mean R/U ratio across cells. Malva pseudocounts (X) are then divided by P to yield corrected values ($\hat{X}$), mitigating inflation caused by PCR amplification. (b) Per-cell correlation

between corrected Malva pseudocounts and reference UMI counts across four datasets: Open-ST human tumor, Visium mouse brain, single-cell human PBMC, and single-cell embryonic mouse brain. Red dashed lines show regression fits for uncorrected pseudocounts, highlighting improvement after correction. (c) Per-gene correlations between corrected Malva pseudocounts and reference quantification for the same datasets.



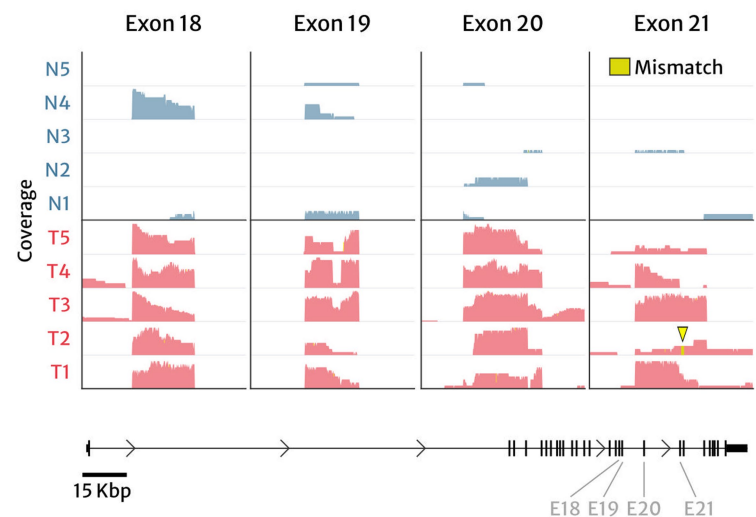
Extended Data Fig. 5 | Benchmarking Malva for sequence-based variant detection in mixed cell populations. (a) Query processing times for four representative targets: BCAS4-BCAS3 fusion, adenovirus 5 E1A, and mitochondrial SNV rs878887783. (b) Detection accuracy comparison between Malva and Flexiplex. Bars show true positives (solid), false positives (hatched), and false negatives (gray) for expected cell lines. Numbers indicate cell counts per category; precision values (True Positives over all Positives) are reported above each bar. (c) UMAP visualization of Malva detections across a mixture of seven cell lines. Target sequences localize to the expected backgrounds: BCAS4-BCAS3 variants in MCF7 (brown), adenovirus E1A in HEK293 (red), and

rs878887783 in T47D (blue). Gray cells show no target sequence detection. (d) Same as (c), but for Flexiplex detections. (e) Precision (PPV) and (f) sensitivity (TPR, computed over the total detections by flexiplex) as a function of probe length. Solid red: computed on a Malva Index generated from overlapping k-mers; dashed red (blue) gradient: Malva Index from overlapping (non-overlapping) k-mers at 0, 0.1, 1, 2, 5, and 10% simulated error rates (light to dark). 0% error rate corresponds to the baseline, intrinsic error rate from the original FASTQ files (Illumina sequencing). None of the benchmark sequences were affected by low-complexity masking.

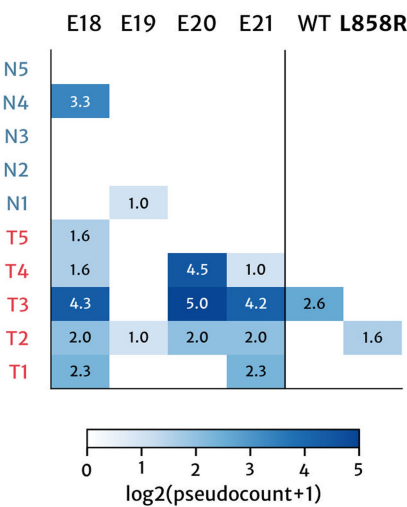
a Lung cancer dataset (PRJCA001731; He *et al.* 2021 Oncogene)

Sample	Driver	Sex	Age	Reads N/Reads T
N5 T5	EGFR 19del	Female	56	245M/271M
N4 T4	EGFR 19del	Male	75	234M/224M
N3 T3	EGFR 19del	Female	58	302M/332M
N2 T2	EGFR L858R	Male	61	264M/273M
N1 T1	EGFR 19del	Female	48	283M/316M

b EGFR locus (STAR mapping)

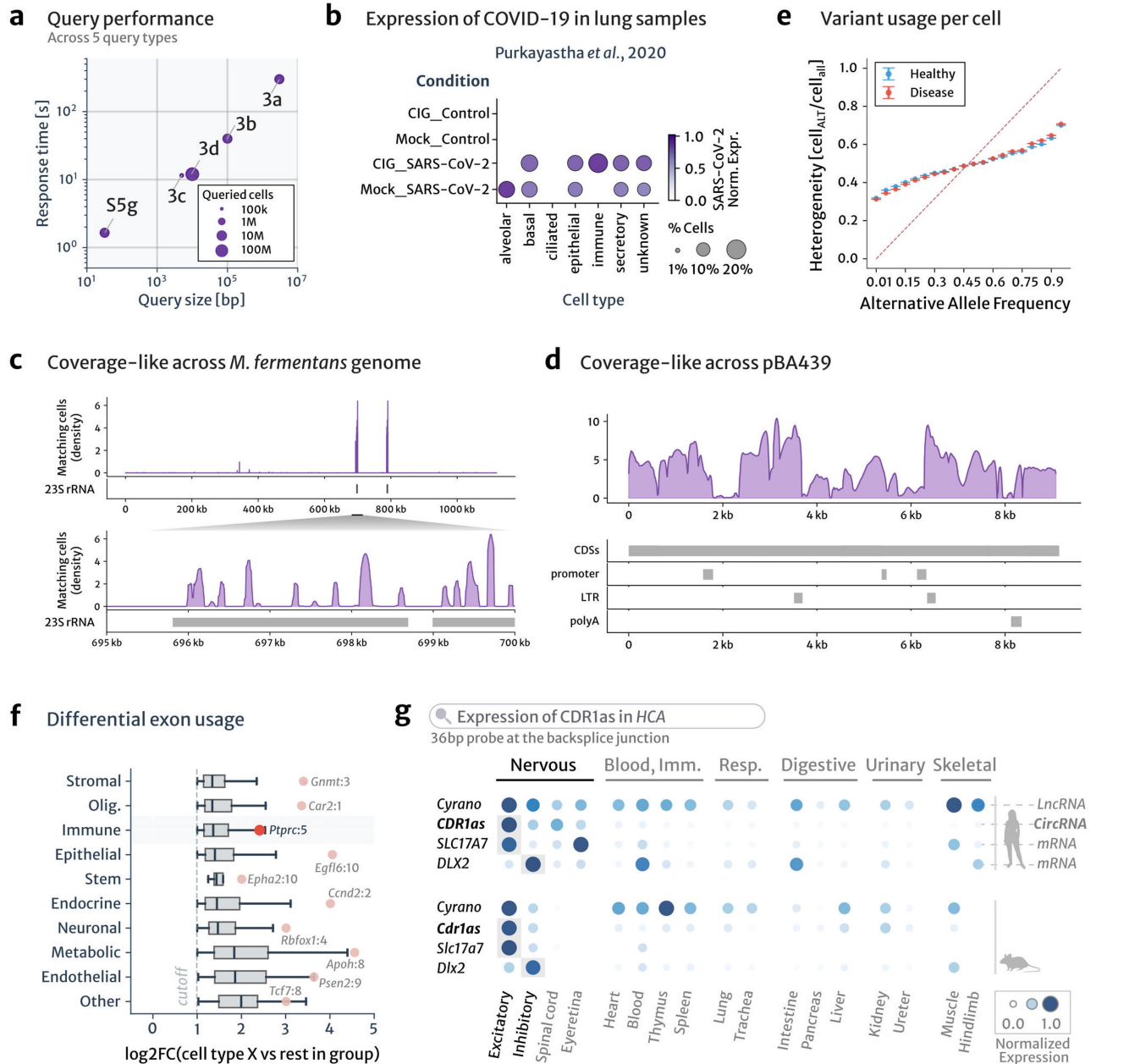


c Probe detection with Malva



Extended Data Fig. 6 | Validation of EGFR mutation detection in lung adenocarcinoma. (a) Sample overview showing five matched tumor-normal pairs from PRJCA001731 (⁷³), including patient sex, age, and sequencing depth. Four tumors carry EGFR exon 19 deletions; one carries the L858R point mutation. (b) STAR alignment coverage across the EGFR locus (exons 18-21) in tumor (T1-T5) and normal (N1-N5) samples. Yellow shading indicates regions

with detected mismatches corresponding to known mutations. (c) Malva probe detection matrix showing pseudocounts for probes targeting exons 18-21, wild-type exon 21 position L858, and the L858R variant. Malva correctly identifies mutation-positive samples while maintaining specificity in normal tissue.



Extended Data Fig. 7 | Malva enables sequence queries beyond predefined gene annotations. (a) Query performance across different sequence types and lengths, ranging from short sequences (24 bp) to large genomic regions (~2 Mbp). Bubble size reflects the number of cells queried (100,000 to 100 million). (b) Dot plot showing the percentage of SARS-CoV-2-positive cells per cell type and sample (dot size) from⁷⁷, with expression values z-score normalized within each cell type to highlight relative enrichment (dot color). (c) Coverage-like analysis of *Mycoplasma* genome across positive samples from Index shows predominant localization to 23S rRNA loci. (d) Coverage-like analysis of lentiviral vector pBA439 across positive samples from Index. (e) Per-sample

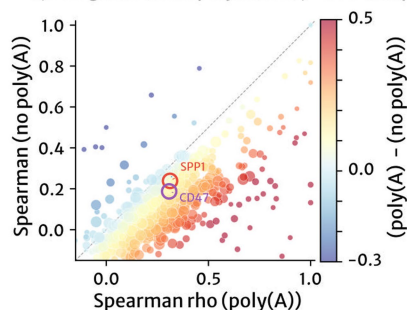
heterozygosity analysis showing the fraction of cells expressing alternative alleles (mean $\pm$ s.e.), centered around 0.5 for individual samples regardless of population allele frequency. (f) Cell-type-specific differential exon usage across the Tabula Muris atlas. Representative examples highlight known regulatory switches, such as *Ptprc* exon inclusion in B cells and alternative splicing of *Cdca2* in neuronal populations. (g) Tissue-specific detection of the circRNA *CDR1as* across the Human Cell Atlas and *Tabula muris*, compared with expression of marker genes *SLC17A7* (excitatory neurons) and *DLX2* (inhibitory neurons).

a Large-scale benchmark of APA detection

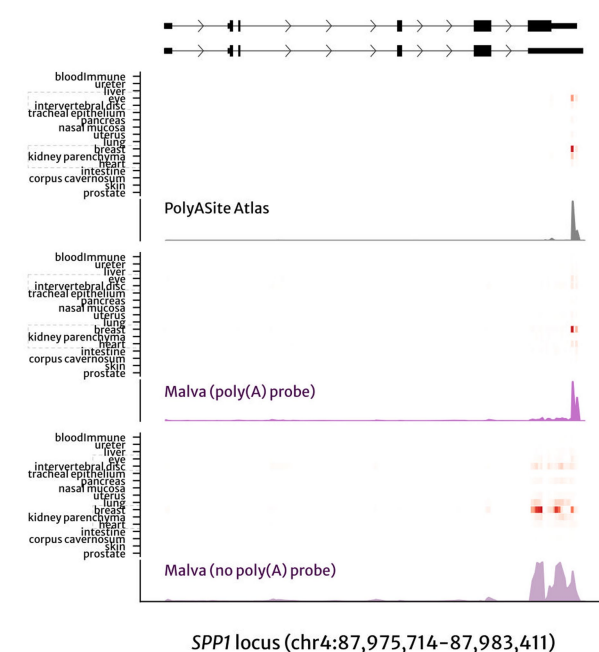


c PolyA usage correlation

~1,000 genes (1.5M polyA sites), ~600 samples

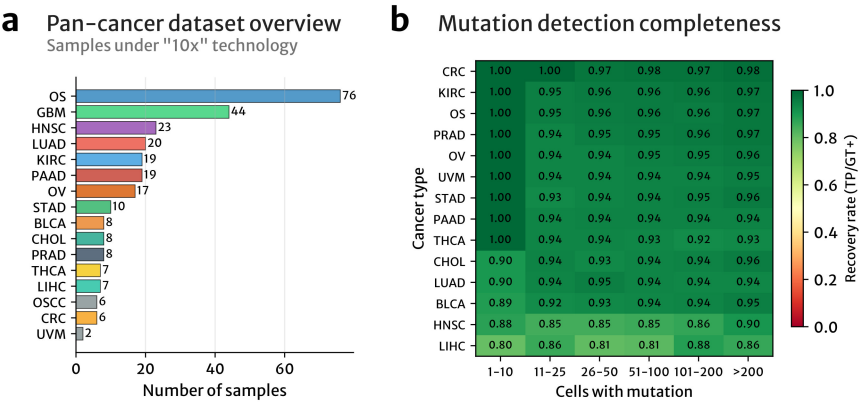


b PolyASite vs Malva polyA site detection comparison



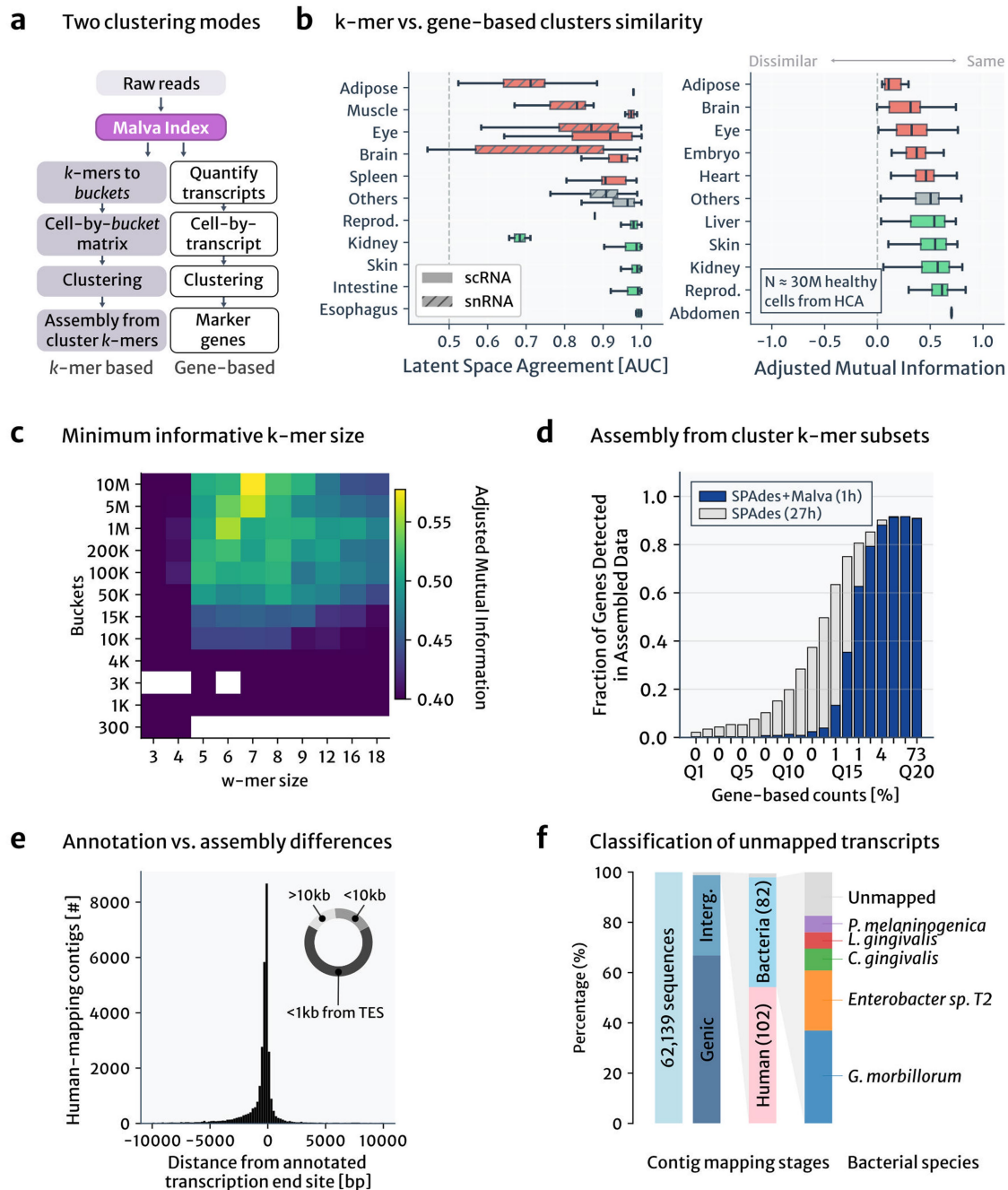
Extended Data Fig. 8 | Large-scale benchmarking of polyadenylation site detection. (a) Workflow schematic. PolyA sites from polyASite v3.0⁵¹ were used to design two probe types: polyA probes (genomic sequence + polyA tract) and control probes (genomic sequence only). Both were queried against overlapping HCA samples in Malva Index (Methods). (b) Example analysis at the *SPP1* locus showing polyASite annotations (top), Malva polyA probe signal (middle), and Malva control probe signal (bottom) across tissues. PolyA probes

recapitulate site-specific usage patterns while control probes show broader coverage-like profiles. (c) Spearman correlation between Malva pseudocounts and polyASite usage scores across ~1,000 randomly selected genes with annotated polyA sites (totaling ~1.5 M probe-tissue pairs across ~600 samples). Point size depicts the relative number of polyA sites per gene; point color shows the difference in correlation (polyA minus control), showing consistent improvement with polyA-specific probe design.



Extended Data Fig. 9 | Pan-cancer dataset composition and mutation detection completeness. (a) scTML dataset composition showing sample counts per cancer type. (b) Detection completeness heatmap showing sensitivity across cancer types and mutation frequency bins. OS, osteosarcoma; GBM, glioblastoma; HNSC, head and neck squamous cell carcinoma; LUAD, lung adenocarcinoma; KIRC, kidney renal clear cell carcinoma; PAAD, pancreatic

adenocarcinoma; OV, ovarian cancer; STAD, stomach adenocarcinoma; BLCA, bladder carcinoma; CHOL, cholangiocarcinoma; PRAD, prostate adenocarcinoma; THCA, thyroid carcinoma; LIHC, liver hepatocellular carcinoma; OSCC, oral squamous cell carcinoma; CRC, colorectal cancer; UVM, uveal melanoma.



Extended Data Fig. 10 | De novo sequence assembly with cluster-specific k-mer subsets. (a) Schematic of the workflow for identifying marker k-mers from Malva-derived clusters. Cluster-enriched k-mers are extracted and used either for transcript quantification (gene-based mode) or for de novo assembly (k-mer-based mode). (b) Comparison of sequence-based versus gene-based clustering across healthy human cells, as shown in Fig. 5b. Latent structure agreement (AUC) and Adjusted Mutual Information (AMI) quantify how well sequence-derived embeddings preserve cluster structure relative to gene counts. Lower agreement in single-nucleus data reflects low-complexity intronic sequences. $N = 29,240,986$ cells examined over 4,932 sequencing runs; boxplots show the median, interquartile range, and whiskers extending to 1.5x the interquartile range. (c) Parameter scan for k-mer bucketing. Varying k-mer

size and bucket definitions shows the bounds of informative sequence length required to recover meaningful latent structures. Color scale is clipped to 0.4 as the cutoff for low-medium quality clustering agreement. (d) Efficiency of cluster-specific assembly. Gene detection rate is shown for transcripts reconstructed with SPAdes using either all reads or cluster-specific Malva k-mer subsets. (e) Annotation versus assembly differences. Histogram shows genomic distance between assembled contigs and annotated transcript 3' ends (TES). Inset: proportion of contigs mapping within different distances from the TES. (f) Mapping of contigs with minimap2 to the human genome, and subsequent classification of unmapped contigs with kraken2. Taxonomic assignment reveals human intergenic sequences, bacterial species and a residual pool of sequences with no confident match.

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	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