

Human-scATAC-Corpus: a comprehensive database of scATAC-seq data

Xiaoyang Chen[†], Zijing Gao[†], Keyi Li[†], Zian Wang[†], Qun Jiang, Xuejian Cui, Zhen Li[†], Rui Jiang^{*}

Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

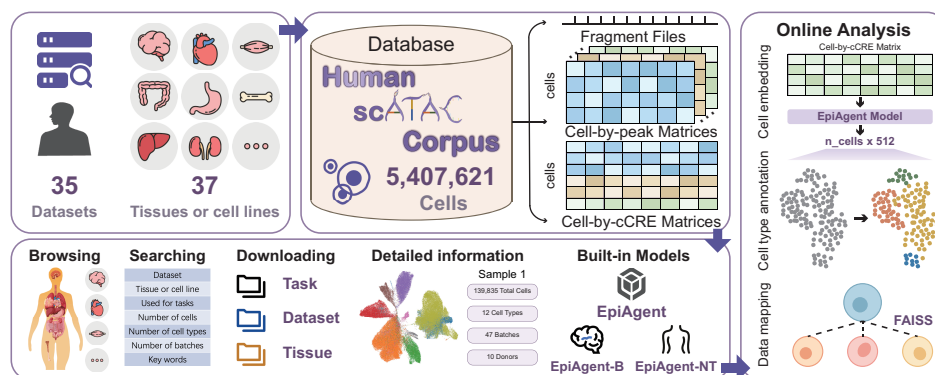
*To whom correspondence should be addressed. Email: ruijiang@tsinghua.edu.cn

[†]The first four authors should be regarded as Joint First Authors.

Abstract

Single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) profiles chromatin accessibility at cellular resolution, making it possible to reveal epigenomic landscapes that govern gene regulation in a variety of cells. Nevertheless, heterogeneous feature spaces and complex processing pipelines have impeded the construction of an ensemble resource capable of supporting diverse downstream analytical scenarios. To address this gap, we present Human-scATAC-Corpus (<https://health.tsinghua.edu.cn/human-scatac-corpus/>), a comprehensive database of human scATAC-seq comprising 5 407 621 cells from 35 datasets across 37 tissues or cell lines. To support complementary use cases, each dataset is distributed in three aligned formats: cell-by-candidate *cis*-regulatory element matrices for cross-dataset integration, raw fragment files for flexible processing, and cell-by-peak matrices for within-dataset analyses. This resource spans diverse biological contexts and includes rich metadata, enabling method benchmarking and development, as well as pretraining of foundation models. The website offers searchable browsing, detailed dataset pages, on-demand downloads, and tutorials. EpiAgent, a foundation model pretrained on Human-scATAC-Corpus, is further integrated to provide online analyses, including reference mapping, embedding extraction, and cell type annotation. Human-scATAC-Corpus establishes a unified and scalable substrate for single-cell epigenomics and is intended to accelerate discovery while standardizing evaluation across tasks.

Graphical abstract



Introduction

Single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) enables the characterization of chromatin accessibility at the resolution of individual cells, thereby revealing the regulatory landscape at the cellular level [1]. In a broad range of biomedical research, including investigations of cellular heterogeneity [2], tissue differentiation [3], and disease pathogenesis [4], scATAC-seq data have been serving as an essential and irreplaceable resource. Recent ad-

vances in sequencing technologies have driven two major developments in this field. First, improvements in throughput have led to the development of high-throughput techniques such as sci-ATAC-seq [5] and sci-ATAC-seq3 [6], enabling the construction of large-scale cell atlases spanning fetal development [6], adult tissues [7], and the adult brain [8]. Second, the scope of applications has been expanding, with techniques such as CRISPR-sciATAC [9] and Spear-ATAC [10] integrating CRISPR-based genome editing to capture cell state

Received: August 15, 2025. Revised: September 29, 2025. Accepted: October 16, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other

permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

changes under various perturbations. Indeed, the growth in both data scale and application diversity has fueled the emergence of task-specific analytical methods and the development of foundation models trained on scATAC-seq data. However, these efforts, whether focusing on cell-level foundation models such as EpiAgent [11], concentrating on genome-level foundation models such as GET [12], or aiming at developing task-specific models [13–20], all require a publicly accessible, ensemble database that encompasses the diversity of scATAC-seq applications and tissue types.

In the single-cell field, large-scale ensemble databases already exist for other omics. For example, the scRNA-seq-based Cell × Gene repository [21] has driven the adoption of large-scale pretraining paradigms for transcriptomic data [22–27]. scRNA-seq features are anchored on a shared gene set, and thus each dataset can be represented as a cell-by-gene matrix that supports most analytic workflows. In contrast, scATAC-seq lacks a universal feature space. Data are typically stored as raw fragment files or as cell-by-peak matrices whose peak definitions differ across datasets [28]. Consequently, no single storage format can satisfy all analytical requirements, and multiple data representations are often necessary. Existing scATAC-seq ensemble resources, such as scATAC-Ref [29], have reduced the time required for users to locate data and facilitated the development of related methods, yet they still present several notable limitations. First, reliance on cell-by-peak matrices yields incompatible feature spaces that impede joint analysis. Second, current scale is insufficient for modern pretraining regimes. Third, data modalities and metadata are narrow, restricting the breadth and fidelity of downstream benchmarking. Fourth, online analysis capabilities are constrained, with few powerful embedded models or reusable applications applied to the hosted data. These limitations underscore the urgent need for a comprehensive database of scATAC-seq data that can address the above challenges.

To facilitate research in single-cell epigenomics, we developed Human-scATAC-Corpus (<https://health.tsinghua.edu.cn/human-scatac-corpus/>), a comprehensive database of scATAC-seq data. The current release contains 5 407 621 cells from 35 datasets across 37 human tissues or cell lines. All data are provided in three complementary formats: cell-by-candidate *cis*-regulatory element (cCRE) matrices for integrative multi-dataset analysis, fragment files for flexible user-defined processing, and cell-by-peak matrices for within-dataset analysis. The collection spans diverse biological contexts, enabling the use in a wide range of downstream tasks, from method development and benchmarking to large-scale pretraining of foundation models. Human-scATAC-Corpus also integrates EpiAgent, a foundation model pretrained on this resource, supporting online analyses including data mapping to reference, cell embedding extraction, and cell type annotation via supervised derivatives EpiAgent-B and EpiAgent-NT. Other functionalities include user-friendly and intuitive browsing, structured and keyword-based search, in-depth dataset detail, on-demand download, customized tutorials, and interactive online analysis. We anticipate that Human-scATAC-Corpus will provide a versatile and scalable resource to accelerate discoveries in single-cell epigenomics and advance the development of analytical methods.

Data collection and database content

Data collection

For the collection of scATAC-seq data, we implemented a two-stage workflow consisting of literature and repository searches followed by eligibility screening and preliminary quality control (Fig. 1). First, we queried the PubMed database and the Gene Expression Omnibus repository using the keywords ‘Human’ and ‘scATAC-seq’, yielding a total of 205 published studies. For these candidate studies, we examined whether the datasets contained scATAC-seq data derived from healthy human samples and excluded datasets that did not meet this criterion. Because the primary objective of our database is to provide high-quality pretraining resources for the foundation model, we excluded cancer and other pathological conditions, as such cells may display unknown regulatory states that could introduce noise and reduce the reliability of training on normal cells.

At the same time, we retained widely used benchmark datasets, such as Buenrostro2018 [30], as well as other cells reported in the studies that contain normal cells, to ensure that our resource remains suitable for both method development and benchmarking. For the studies retained after screening, we downloaded the publicly available data and further examined whether fragment files or high-quality cell-by-peak matrices were provided. Ultimately, we curated and collected 35 datasets spanning 37 tissues or cell lines. For each dataset, we additionally recorded essential metadata, including the reference genome, tissue or cell line of origin, sequencing technology, and PubMed identifier (PMID). We plan to update the database every 6 months, while each update will be additive and will not alter or remove existing data, thereby ensuring that all previously available datasets remain accessible.

Data processing and annotation

For the Human-scATAC-Corpus, only data aligned to the reference genome GRCh38/hg38 are currently provided. Consequently, the first step involves mapping data originally with GRCh37/hg19 into GRCh38/hg38 by liftOver [31]. To facilitate usability and ensure accurate data retrieval, we defined the minimal storage unit (also denoted as a sample) in Human-scATAC-Corpus as a single tissue within a dataset. Each sample is named in the format dataset–tissue/cell line and assigned a unique identifier (ID) for reference. At the sample level, we also annotated their detailed information based on the available metadata.

For each sample, Human-scATAC-Corpus offers three data formats: a cell-by-cCRE matrix, fragment files, and a cell-by-peak matrix. The cell-by-cCRE matrix is provided for all samples in Human-scATAC-Corpus, while the other two depend on the availability of source data. Fragment files, typically stored in formats such as .tsv or .bam, and cell-by-peak matrices, commonly stored in .h5ad or .mtx formats, are the most frequently used data types in current scATAC-seq studies. During the processing stage, we standardized all fragment files into a series of compressed .tsv.gz files, with each file corresponding to a batch. Cell-by-peak matrices were processed into a single .h5ad file.

Since fragment files represent relatively non-processed data and cell-by-peak matrices are dataset-specific due to non-unified peak definitions, they are generally less suitable for

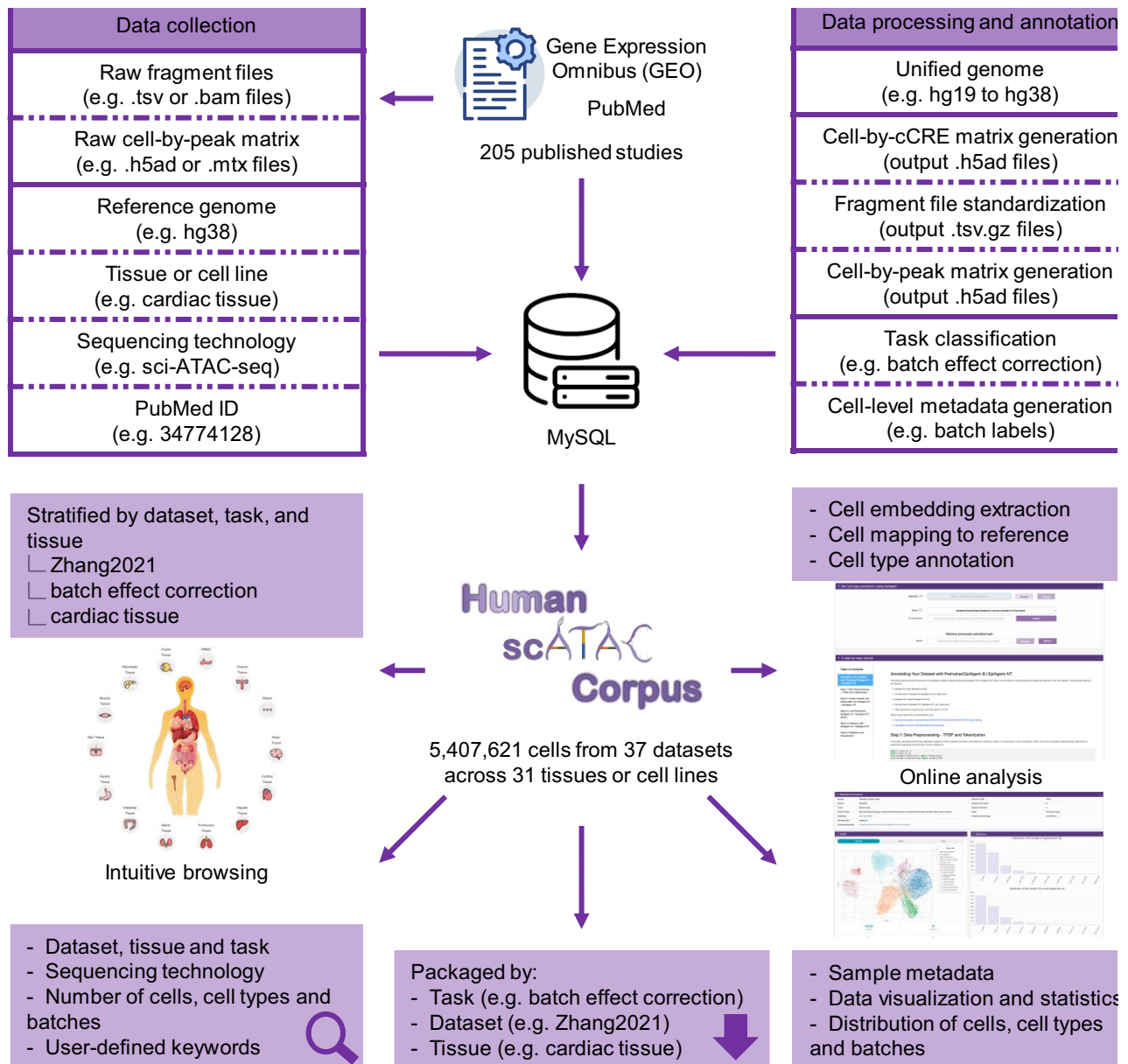


Figure 1. Overview of data collection, data processing and annotation, and database features of Human-scATAC-Corpus.

integrative analysis across datasets. To address this, we constructed a unified set of cCREs by merging three large-scale datasets covering complementary tissue contexts: the Zhang2021 [7] dataset (cCREs from all normal human tissues except brain), the Li2023a [8] dataset (cCREs from brain tissues), and the Lareau2019 [32] dataset (peaks from peripheral blood mononuclear cells, PBMCs). Before merging, we filtered out peaks longer than 1000 bp from the Lareau2019 dataset (all features from the other two datasets had fixed lengths <500 bp) to avoid generating excessively long peaks during the merge. We then applied an additional filter requiring that each retained feature be supported by at least 100 fragments in the Human-scATAC-Corpus pretraining data, resulting in a unified set of 1 355 445 cCREs.

Fragment files (or, when unavailable, a cell-by-peak matrix) were then mapped to this unified cCRE set to generate a cell-by-cCRE matrix of .h5ad format, consistent with the strategy employed in our EpiAgent study [11]. In that

work, we demonstrated that the superior performance of EpiAgent across multiple datasets and tasks supports the reliability of the cCRE definition. Here, we further illustrated that the unified cCRE set exhibits a high degree of overlap with the peak or cCRE sets of the vast majority of datasets, and evaluated the impact of using the unified cCRE set on fragment mapping rates and downstream analyses with five representative datasets [30, 33–36] (Supplementary Note S1 and Supplementary Figs S1–S3). After mapping, we applied quality filtering by removing cells with fewer than 100 non-zero values in the cell-by-cCRE matrix, and we ensured consistency of cell storage across all three data formats. Sample names and batch information were incorporated into file and barcode prefixes, and standardized naming conventions were applied to facilitate downstream use. Standardization and enhanced data usability, particularly in enabling large-scale pretraining, therefore represent core strengths of Human-scATAC-Corpus. For this reason, we provide cell-by-cCRE matrices for all sam-

ples to ensure consistency for pretraining tasks. In contrast, cell-by-peak matrices are primarily intended for reproducing results from the original publications or for benchmarking methods on a limited number of datasets. For these cases, we either use the cell-by-peak matrices provided by the original studies or generate them directly from their provided peaks. We do not perform additional peak calling for datasets where such information is not available. Further clarifications are provided in [Supplementary Note S2](#).

Finally, we classified samples according to the availability of cell-level metadata to indicate the types of tasks they support. Specifically, we designated samples for (i) self-supervised pretraining when all cells are included, (ii) unsupervised feature extraction (also referred to as dimensionality reduction) and supervised cell type annotation when cell type labels are available, (iii) batch effect correction when batch labels are provided, (iv) prediction of cell states under unseen genetic perturbation when CRISPR-based perturbation information is available, and (v) prediction of cell states under out-of-sample stimulated perturbation when external stimulation labels are included. Noted that cell type labels are generally not involved in the training of unsupervised feature extraction methods; rather, they are used only to evaluate whether these methods can produce low-dimensional cell embeddings that adequately capture cellular heterogeneity. The details of the generation, processing, and storage of the cell-level metadata are provided in [Supplementary Note S3](#). To ensure interoperability, all cell-level metadata were standardized: for cell-by-cCRE and cell-by-peak matrices, metadata are stored in the .obs field of the Anndata object [37], while for fragment files, metadata are provided as separate .csv files. This design facilitates the development and benchmarking of analytic methods across diverse tasks (Fig. 1).

Database statistics

As of the current release, the Human-scATAC-Corpus database contains 5 407 621 cells (~5.4 million), organized into 77 samples defined by ‘dataset-tissue/cell line’ combinations and encompassing nine sequencing technologies. Among these, brain tissue accounts for the largest proportion, with 2 549 071 cells, representing 47.14% of the total. By data type, the database provides cell-by-cCRE matrices for all 77 samples (covering all cells), fragment files for 68 samples (4 752 227 cells in total), and cell-by-peak matrices for 65 samples (4 117 861 cells in total). The Human-scATAC-Corpus database supports method development and benchmarking for at least five types of analytic tasks: all cells can be used for the self-supervised pretraining task; 66 samples (5 192 122 cells) contain batch labels, enabling the batch effect correction task; 46 samples (3 021 043 cells) have cell type annotations, supporting unsupervised feature extraction and supervised cell type annotation; four samples (167 064 cells) are suitable for cell states prediction under unseen genetic perturbation; and another four samples (89 723 cells) are suitable for cell states prediction under out-of-sample stimulated perturbation.

Built-in foundation model

Prior databases have largely focused on data provision for download and offline use, with little or no capacity for online analysis, or only limited functionality based on simple plug-ins. As a result, the information embedded in the data

has often been underutilized. In our study, we integrate into the server side of the database a pretrained EpiAgent model, trained on a subset of Human-scATAC-Corpus, together with two supervised derivatives: EpiAgent-B for the annotation of cells from brain tissues and EpiAgent-NT for the annotation of cells from non-brain tissues [11]. This integration enables users to directly perform online analyses, thereby enhancing the accessibility and utility of the database.

Database features and applications

User-friendly browsing

We developed an intuitive web interface that hosts an unprecedented repository of human scATAC-seq data, empowering users to advance research in single-cell epigenomics and to perform online analyses with a built-in foundation model, EpiAgent [11]. The Home page features interactive images of human anatomy and comprehensive statistics detailing the distribution of over 5.4 million cells across file formats, tissues, and analytical tasks, alongside related studies and databases pertinent to single-cell epigenomics. By clicking on the anatomical image, users are directed to the *Browse* page, enabling them to explore and download scATAC-seq data associated with the corresponding tissues (Fig. 2A). At the *Browse* page, the datasets are categorized in parallel by three key attributes: tissue or cell line, dataset, and task. As shown in Fig. 2B, Human-scATAC-Corpus provides an interactive metadata table, a visualized UMAP projection from cell embeddings generated by EpiAgent, and comprehensive statistics for metadata. The metadata table presents basic information for each sample, including tissue, organ, PMID, hyperlinks to source data and publication, sequencing technology, and raw data type. Users can directly download the cell-by-peak matrix, cell-by-cCRE matrix, and fragment files via clickable icons in the table. Each sample is also assigned an interactive unique ID that directs to a new webpage with detailed sample information. The UMAP visualization displays up to 100 000 sampled cells of the selected item, labelled by cell type, batch, or donor, with interactive label switching. The statistics comprise pie charts illustrating the distribution of cell counts by sequencing technology, raw data type, tissue, and dataset. Detailed usage instructions for the website are available on the *Tutorial* page. In addition, we provide a case study to guide users on how to search, browse, and download data of interest through the website ([Supplementary Fig. S4](#)). We also provide a schematic illustrating the complete pipeline, including data collection, data standardization, data generation, submission to the inference interface (EpiAgent), and result download ([Supplementary Fig. S5](#)).

Detailed information

Detailed information for each sample selected from the *Browse* or *Search* page is compiled and presented on a dedicated webpage, with each sample ID linked to its corresponding *Detail* page. The *Detail* page encompasses comprehensive metadata for the selected sample, including sample name, dataset, tissue, downstream task, raw data type, organ, sequencing technology, number of cells, number of cell types, number of batches, as well as links to the source data (e.g. NCBI [38], BioStudies [39]) and related publication (Fig. 2C). An interactive UMAP projection of cell embeddings generated from EpiAgent is set on the page, enabling users to explore in-

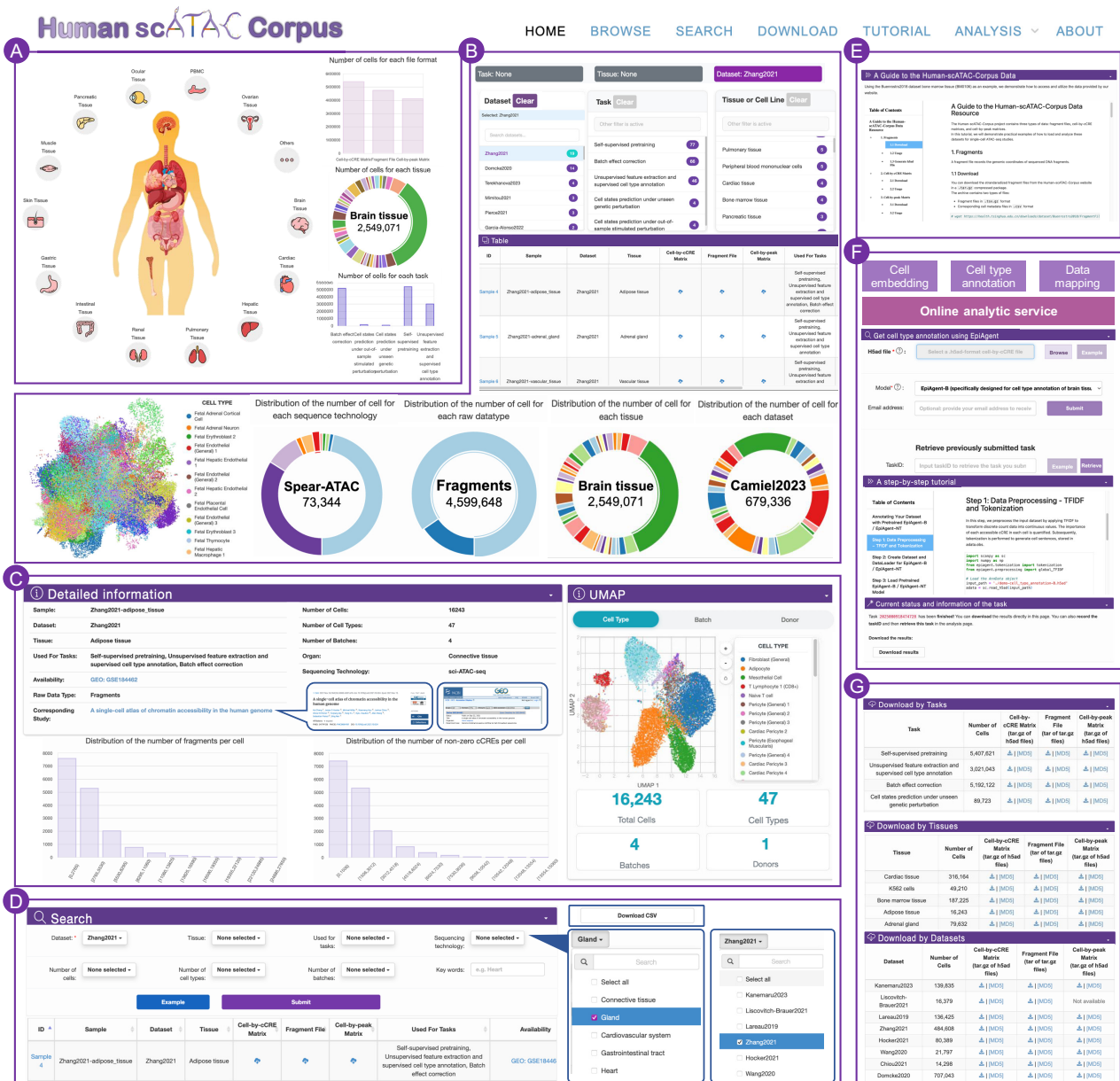


Figure 2. The schematic features on various webpages of Human-scATAC-Corpus. **(A) Home:** interactive images of human tissue-level navigation and overview data statistics. **(B) Browse:** hierarchical organization of datasets by tissue, task, and dataset, with metadata and visualizations. **(C) Detail:** comprehensive sample-level metadata, UMAP projections, and statistics. **(D) Search:** metadata- and keyword-based retrieval of datasets. **(E) Download:** task-, tissue-, and dataset-based data packages with MD5 checksums. **(F) Tutorial:** guidance on data formats, labels, and website usage. **(G) Analysis:** online tools for cell embedding, cell type annotation, and data mapping.

dividual cells and toggle labels between cell type, batch, and donor, with associated count distribution displayed for cells, cell types, batches, and donors via pie charts. Additionally, detailed statistics including distributions of fragment counts, non-zero peak counts, and non-zero cCRE counts per cell are presented via histograms.

Advanced searching

Structured metadata-based filters and keyword-based searches are available at the *Search* page to facilitate flexible queries for sample data (Fig. 2D). Users can specify criteria from drop-down menus such as dataset and tissue,

and further refine searches by selecting downstream tasks, sequencing technologies, and defined ranges for the numbers of cells, cell types, and batches. A fuzzy search function is also supported, enabling keyword entry to identify potentially relevant samples. Upon submission of a searching query, an interactive metadata table will be presented, enabling direct exploration and download of three types of files for each sample: the cell-by-cCRE matrix, fragment files, and the cell-by-peak matrix. Users can also click on the sample ID in the table to access the detailed web page. In addition, users can click the *Example* button to obtain an exemplary search entry along with corresponding results. Alternatively, a *Download* button located at the bottom right corner of the

table allows users to download a .csv file containing both basic dataset information and download links for datasets within the search results, thereby supporting batch data retrieval via command-line tools.

Data download

In addition to downloading sample files via the tables on the *Search* and *Browse* pages individually, we have developed the *Download* page to facilitate convenient access to prepackaged data (Fig. 2E). Specifically, for each task, tissue, and dataset, three file types (cell-by-peak matrix, cell-by-cCRE matrix, fragment file) are archived and compressed into downloadable packages. To ensure data integrity, corresponding MD5 checksum files are provided alongside the downloadable files. We also provide guidance on the usage of various data formats and labelling conventions on the *Tutorial* page, where users can also find a comprehensive tutorial on navigating and utilizing the website (Fig. 2F).

Online analysis and case study

Human-scATAC-Corpus integrates the pretrained EpiAgent model and its supervised derivatives, EpiAgent-B and EpiAgent-NT, into the backend to support three types of online analyses: cell embedding extraction, cell type annotation, and data mapping (Fig. 2G). All online analyses require the input data to be provided in .h5ad format as a cell-by-cCRE matrix. To ensure stable performance, the system restricts each individual analysis to a maximum of 10 000 cells, thereby preventing memory overload or excessive computational time for a single task. For larger datasets, users are advised to partition their data into smaller subsets for batch analysis. We have verified that splitting large datasets into multiple subsets does not affect the reliability of results, with detailed results provided in [Supplementary Note S4](#) and [Supplementary Figs S6--S9](#). Under this constraint, the typical runtime for each type of online analysis is summarized in [Supplementary Note S5](#) and [Supplementary Fig. S10](#). For each analysis task, users can specify hyperparameters and submit their data; the backend automatically generates a task ID and executes the process. Upon completion, users are notified via the email address they provided and can retrieve the results from the online analysis portal using the assigned task ID. The detailed workflows of the three online analyses and representative case studies are described below.

Cell embedding extraction. For this analysis, Human-scATAC-Corpus automatically executes a series of built-in scripts, including data preprocessing, model inference with EpiAgent, and results aggregation (Fig. 3A). The generated cell embeddings are stored in the AnnData object under the .obs field with the key 'cell_embeddings_zero_shot', and the full AnnData object is returned to the user in .h5ad format. As a case study, we applied this workflow to a randomly downsampled subset of the Buenrostro2018 [30] dataset (1000 cells), which is also provided as example data on our website. As shown in Fig. 3A, our database successfully produced embeddings that capture cellular heterogeneity without requiring local computation, thereby greatly facilitating downstream analyses.

Cell type annotation. The pipeline of cell type annotation closely parallels that of cell embedding extraction, with the exception that the backend model is replaced by one of the supervised derivatives, EpiAgent-B or EpiAgent-NT, according

to the user's selection (Fig. 3B). Predicted cell type labels are stored in the AnnData object under the .obs field with the key 'predicted_cell_types'. As a case study, we used a randomly downsampled subset of the Li2023b [40] dataset (1000 cells), also available as example data, and selected EpiAgent-B as the annotation model. As illustrated in Fig. 3B, the predicted cell type labels exhibit strong concordance with the ground-truth labels, demonstrating the potential of this function as a powerful tool for preliminary annotation of scATAC-seq data. In addition, we used the Kanemaru2023 [41] and Li2023b datasets to demonstrate the advantages of performing this online analysis with EpiAgent-B/NT compared to other scATAC-seq-based methods [15,42,43] ([Supplementary Note S6](#) and [Supplementary Fig. S11](#)).

Data mapping. The data mapping function leverages EpiAgent to associate query cells with reference data, enabling comparisons between disease and normal cells or between *in vivo* and *in vitro* cell states, thereby providing insights into pathogenic mechanisms and developmental processes [44]. On the backend, reference cells with known type annotations are selected and processed through EpiAgent to obtain their embeddings. For each dataset, the embeddings of cells belonging to the same type are averaged to obtain reference samples.

This function supports two levels of resolution (Fig. 3C). In cell-level mapping, each query cell is assigned to the nearest reference sample in embedding space based on Euclidean distance, and the results are returned in a .csv file. In cell type-level mapping, the query cells are first grouped by a user-specified annotation key, which refers to one of the input hyperparameters for online analysis ('Cell-type key' on the Analysis page). For each group, the 10 closest reference samples, defined as dataset-cell type pairs, are identified. During the mapping procedure, the Faiss library [45] is employed to accelerate computation. The outputs include an Excel file (with one sheet per group) and distance visualization plots (one per group), all packaged in a compressed archive.

As a case study, we applied cell type-level mapping to a randomly downsampled subset of clear cell renal cell carcinoma (ccRCC) cells from the Terekhanova2023 [46] dataset (1000 cells). The results indicated that ccRCC cells were mapped to astrocytes and normal proximal tubule epithelial cells, respectively. These mappings are consistent with the original analyses reported in Terekhanova2023 [46], underscoring the accuracy and reliability of the data mapping function. Furthermore, we used ccRCC cells from the Terekhanova2023 [46] dataset to show the superiority of the pretrained EpiAgent over cisTopic [17] for this online analysis task ([Supplementary Note S6](#) and [Supplementary Fig. S12](#)). In addition, to illustrate the similarities and differences between the two mapping strategies, we provide both a conceptual workflow and an applied example to explain their relative advantages ([Supplementary Note S7](#) and [Supplementary Fig. S13](#)).

System design and implementation

The website is maintained on a Linux-based Nginx web server v1.28.0 (<https://nginx.org/>) with the backend running PHP (v7.4.33). All metadata are stored in a MySQL database (v8.4.0). We utilize various jQuery and JavaScript plugins, including DataTables (v1.10.19), morris.js (v0.5.0), D3.js (v7.8.5), and Chart.js (v3.9.1), to implement and render tables, charts, and UMAP visualizations. The web interface is

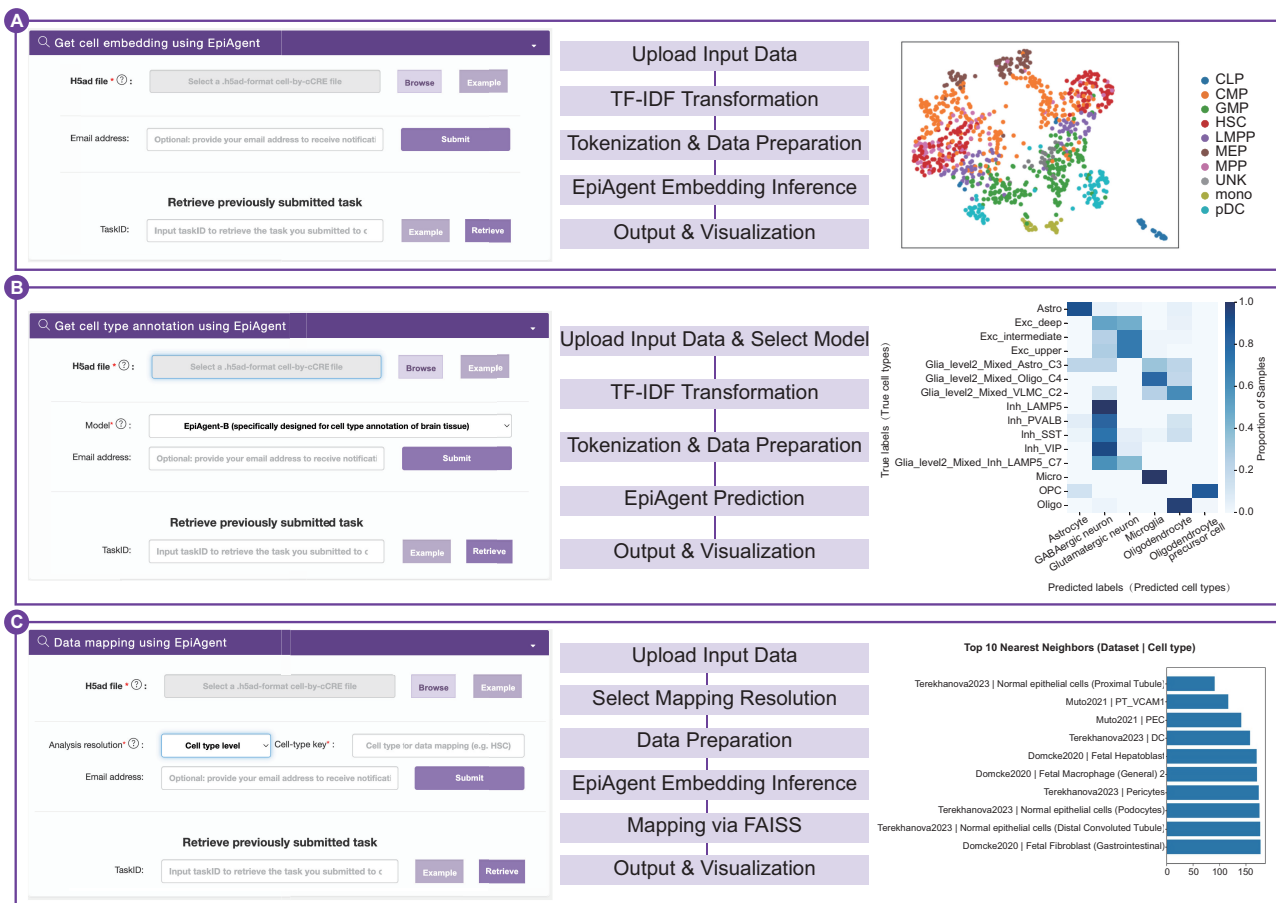


Figure 3. Online analysis in the Human-scATAC-Corpus database. **(A)** Cell embedding extraction: extraction of cell embeddings using the EpiAgent model pretrained on parts of cells in Human-scATAC-Corpus. **(B)** Cell type annotation: supervised prediction of cell types with EpiAgent-B (brain tissue) or EpiAgent-NT (other tissues). **(C)** Data mapping: matching query cells or cell types to normal tissue references within Human-scATAC-Corpus using the EpiAgent model and Faiss library. TF-IDF denotes as Term Frequency-Inverse Document Frequency.

optimized using the Bootstrap framework (v3.3.7). Besides, the online analysis service powered by the built-in foundation model is currently deployed on a computing node equipped with 4 NVIDIA GeForce RTX 4090 GPUs within a high-performance computing cluster. Considering the lengthy computation time required for the tasks, we designed a parallelization strategy to enable efficient inference across these GPUs. The current Human-scATAC-Corpus website supports most mainstream web browsers, including but not limited to Google Chrome, Firefox, Safari, Microsoft Edge, and Opera. With regard to API/CLI endpoints, the website currently provides direct download links for all datasets but does not offer programmatic API access.

Discussion

Recent advances in scATAC-seq profiling have produced datasets that are both vast in scale and broad in biological scope, developmental stages [6], spanning diverse tissues [7], and perturbation contexts [9, 10, 32]. This expansion creates the conditions for foundation models to become a central paradigm in scATAC-seq [11, 12], yet progress has been constrained by the absence of a curated database that reconciles divergent feature spaces, standardizes metadata, and supports reproducible computation. Recognizing this gap, we de-

veloped Human-scATAC-Corpus, a unified ensemble database designed to reconcile these challenges by integrating multiple data representations, spanning broad tissue and perturbation contexts, and embedding a built-in foundation model for online analysis. Compared with other ensemble databases [29, 47–49] that include scATAC-seq data, our resource not only surpasses them in richness and diversity but also introduces novel functional features (Supplementary Note S8 and Supplementary Table S1). In doing so, our database addresses the technical fragmentation derived from characteristics of scATAC-seq data that has hindered integrative analysis, while providing an accessible and scalable resource to accelerate the development and application of large-scale foundation models in single-cell epigenomics.

Following the initial release of Human-scATAC-Corpus, we plan to advance the resource along three directions. First, we will extend beyond human data by incorporating data from mouse and other species, establishing a multi-species database. Second, we will systematically curate additional experimental contexts to broaden the range of downstream applications and provide stronger support for method development and benchmarking. Third, we will integrate additional single-cell epigenomic modalities, such as chromatin interactions and DNA methylation, with harmonized metadata and interoperable representations. These expansions are intended to en-

able large-scale pretraining on comprehensive, multi-omic regulatory signals and to facilitate joint representation learning across species and modalities.

Acknowledgements

Not applicable.

Author contributions: R.J. conceived the study and supervised the project. X.C. designed the data storage format of Human-scATAC-Corpus and the frontend of the database. X.C., Z.W., and K.L. collected, processed, and annotated the data. Z.G., K.L., Z.W., Q.J., X.C., and Z.L. contributed to the development of the database front-end. X.C., Z.G., Z.W., K.L., and R.J. wrote the manuscript with input from all authors.

Supplementary data

[Supplementary data](#) is available at NAR online.

Conflict of interest

None declared.

Funding

National Key Research and Development Program of China [2025YFC3409300, 2023YFF1204802, 2021YFF1200902]; National Natural Science Foundation of China [62273194]; Beijing Natural Science Foundation [L242026]. Funding to pay the Open Access publication charges for this article was provided by the National Key Research and Development Program of China [2025YFC3409300, 2023YFF1204802, 2021YFF1200902]; National Natural Science Foundation of China [62273194]; and Beijing Natural Science Foundation [L242026].

Data availability

Users can access all features of Human-scATAC-Corpus without requiring registration or login. Users can freely access all data hosted in Human-scATAC-Corpus at <https://health.tsinghua.edu.cn/human-scatac-corpus/>. All backend code, data processing scripts, and the EpiAgent inference pipelines are provided on GitHub (<https://github.com/xy-chen16/Human-scATAC-Corpus>) and archived on Zenodo (<https://doi.org/10.5281/zenodo.17219448>) to ensure reproducibility and public accessibility.

References

- Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015;16:716–26. <https://doi.org/10.1038/nrg3980>
- Buenrostro JD, Wu B, Litzenburger UM *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;523:486–90. <https://doi.org/10.1038/nature14590>
- Pijuan-Sala B, Wilson NK, Xia J *et al.* Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nat Cell Biol* 2020;22:487–97. <https://doi.org/10.1038/s41556-020-0489-9>
- Chiou J, Zeng C, Cheng Z *et al.* Single-cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. *Nat Genet* 2021;53:455–66. <https://doi.org/10.1038/s41588-021-00823-0>
- Cusanovich DA, Daza R, Adey A *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015;348:910–4. <https://doi.org/10.1126/science.aab1601>
- Domcke S, Hill AJ, Daza RM *et al.* A human cell atlas of fetal chromatin accessibility. *Science* 2020;370:eaba7612. <https://doi.org/10.1126/science.aba7612>
- Zhang K, Hocker JD, Miller M *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* 2021;184:5985–6001. <https://doi.org/10.1016/j.cell.2021.10.024>
- Li YE, Preissl S, Miller M *et al.* A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* 2023;382:eadf7044. <https://doi.org/10.1126/science.adf7044>
- Liscovitch-Brauer N, Montalbano A, Deng J *et al.* Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol* 2021;39:1270–7. <https://doi.org/10.1038/s41587-021-00902-x>
- Pierce SE, Granja JM, Greenleaf WJ. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat Commun* 2021;12:2969. <https://doi.org/10.1038/s41467-021-23213-w>
- Chen X, Li K, Cui X *et al.* EpiAgent: foundation model for single-cell epigenomic data. *Nat Methods* 2025. <https://doi.org/10.1038/s41592-025-02822-z>
- Fu X, Mo S, Buendia A *et al.* A foundation model of transcription across human cell types. *Nature* 2025;637:965–73. <https://doi.org/10.1038/s41586-024-08391-z>
- Yuan H, Kelley DR. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat Methods* 2022;19:1088–96. <https://doi.org/10.1038/s41592-022-01562-8>
- Cui X, Chen X, Li Z *et al.* Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. *Nat Comput Sci* 2024;4:346–59. <https://doi.org/10.1038/s43588-024-00625-4>
- Chen X, Chen S, Song S *et al.* Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nat Mach Intell* 2022;4:116–26. <https://doi.org/10.1038/s42256-021-00432-w>
- Tang S, Cui X, Wang R *et al.* scCASE: accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data. *Nat Commun* 2024;15:1629. <https://doi.org/10.1038/s41467-024-46045-w>
- Gonzalez-Blas B, C. M, L. P *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* 2019;16:397–400. <https://doi.org/10.1038/s41592-019-0367-1>
- Ashuaq T, Reidenbach DA, Gayoso A *et al.* PeakVI: a deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods* 2022;2:100182. <https://doi.org/10.1016/j.crmeth.2022.100182>
- Li Z, Kuppe C, Ziegler S *et al.* Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat Commun* 2021;12:6386. <https://doi.org/10.1038/s41467-021-26530-2>
- Xiao Y, Jin W, Ju L *et al.* Tracking single-cell evolution using clock-like chromatin accessibility loci. *Nat Biotechnol* 2025;43:784–98. <https://doi.org/10.1038/s41587-024-02241-z>
- Program CCS, Abdulla S, Aevermann B *et al.* CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2025;53:D886–900. <https://doi.org/10.1093/nar/gkac1142>
- Theodoris CV, Xiao L, Chopra A *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;618:616–24. <https://doi.org/10.1038/s41586-023-06139-9>
- Cui H, Wang C, Maan H *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;21:1470–80. <https://doi.org/10.1038/s41592-024-02201-0>

24. Hao M, Gong J, Zeng X *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;21:1481–91. <https://doi.org/10.1038/s41592-024-02305-7>
25. Szalata A, Hrovatin K, Becker S *et al.* Transformers in single-cell omics: a review and new perspectives. *Nat Methods* 2024;21:1430–43. <https://doi.org/10.1038/s41592-024-02353-z>
26. Lotfollahi M. Toward learning a foundational representation of cells and genes. *Nat Methods* 2024;21:1416–7. <https://doi.org/10.1038/s41592-024-02367-7>
27. Yang X, Liu G, Feng G *et al.* GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res* 2024;34:830–45. <https://doi.org/10.1038/s41422-024-01034-y>
28. Minnoye L, Marinov GK, Krausgruber T *et al.* Chromatin accessibility profiling methods. *Nat Rev Methods Primers* 2021;1:10. <https://doi.org/10.1038/s43586-020-00008-9>
29. Qian FC, Zhou LW, Zhu YB *et al.* scATAC-Ref: a reference of scATAC-seq with known cell labels in multiple species. *Nucleic Acids Res* 2024;52:D285–92. <https://doi.org/10.1093/nar/gkad924>
30. Buenrostro JD, Corces MR, Lareau CA *et al.* Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* 2018;173:1535–48. <https://doi.org/10.1016/j.cell.2018.03.074>
31. Haeussler M, Zweig AS, Tyner C *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* 2019;47:D853–8. <https://doi.org/10.1093/nar/gky1095>
32. Lareau CA, Duarte FM, Chew JG *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* 2019;37:916–24. <https://doi.org/10.1038/s41587-019-0147-6>
33. Muto Y, Wilson PC, Ledru N *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat Commun* 2021;12:2190. <https://doi.org/10.1038/s41467-021-22368-w>
34. Jin C, Wang X, Yang J *et al.* Molecular and genetic insights into human ovarian aging from single-nuclei multi-omics analyses. *Nat Aging* 2025;5:275–90. <https://doi.org/10.1038/s43587-024-00762-5>
35. Lee AJ, Kim C, Park S *et al.* Characterization of altered molecular mechanisms in Parkinson's disease through cell type-resolved multiomics analyses. *Sci Adv* 2023;9:eabo2467. <https://doi.org/10.1126/sciadv.abo2467>
36. Granja JM, Klemm S, McGinnis LM *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* 2019;37:1458–65. <https://doi.org/10.1038/s41587-019-0332-7>
37. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15. <https://doi.org/10.1186/s13059-017-1382-0>
38. Schoch CL, Ciufo S, Domrachev M *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020;2020:baaa062. <https://doi.org/10.1093/database/baaa062>
39. Sarkans U, Gostev M, Athar A *et al.* The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Res* 2018;46:D1266–70. <https://doi.org/10.1093/nar/gkx965>
40. Li J, Jaiswal MK, Chien JF *et al.* Divergent single cell transcriptome and epigenome alterations in ALS and FTD patients with C9orf72 mutation. *Nat Commun* 2023;14:5714. <https://doi.org/10.1038/s41467-023-41033-y>
41. Kanemaru K, Cranley J, Muraro D *et al.* Spatially resolved multiomics of human cardiac niches. *Nature* 2023;619:801–10. <https://doi.org/10.1038/s41586-023-06311-1>
42. Ma W, Lu J, Wu H. Cellcano: supervised cell type identification for single cell ATAC-seq data. *Nat Commun* 2023;14:1864. <https://doi.org/10.1038/s41467-023-37439-3>
43. Zeng Y, Luo M, Shangquan N *et al.* Deciphering cell types by integrating scATAC-seq data with genome sequences. *Nat Comput Sci* 2024;4:285–98. <https://doi.org/10.1038/s43588-024-00622-7>
44. Heimberg G, Kuo T, DePianto DJ *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* 2025;638:1085–94. <https://doi.org/10.1038/s41586-024-08411-y>
45. Douze M, Guzhva A, Deng C *et al.* The faiss library. arXiv, <https://arxiv.org/html/2401.08281v4>, 23 October 2025, preprint: not peer reviewed
46. Terekhanova NV, Karpova A, Liang WW *et al.* Epigenetic regulation during cancer transitions across 11 tumour types. *Nature* 2023;623:432–41. <https://doi.org/10.1038/s41586-023-06682-5>
47. Jiang Q, Chen X, Gao Z *et al.* CASHeart: a database of single cells chromatin accessibility for the human heart. *Quant Biol* 2025;13. <https://doi.org/10.1002/qub2.90>
48. Zhao Y, Yu ZM, Cui T *et al.* scBlood: a comprehensive single-cell accessible chromatin database of blood cells. *Comput Struct Biotechnol J* 2024;23:2746–53. <https://doi.org/10.1016/j.csbj.2024.06.015>
49. Huang K, Gong H, Guan J *et al.* AgeAnno: a knowledgebase of single-cell annotation of aging in human. *Nucleic Acids Res* 2023;51:D805–15. <https://doi.org/10.1093/nar/gkac847>