scientific reports

Check for updates

OPEN

Developing a comprehensive database and search tool for single-cell ATAC-seq data

Arrian Gibson-Khademi[1,2], Jonathan Wang[1,2], Erik Christensen[2,3] & Parisa Shooshtari[1,2,3,4]✉

Single cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) is a technology that allows for analysis of chromatin accessibility at a single cell resolution. scATAC-seq is advancing our understanding of gene regulation, and has found many applications in cell development and disease. Despite the availability of several scATAC-seq datasets and their high applicability, we still lack a definitive scATAC-seq database, where a comprehensive collection of publicly available scATAC-seq datasets has been made available in a consistent format. To solve this, we have created scATAC.Explorer, an R accessible database and search tool, containing a curated collection of 39 currently available scATAC-seq datasets from a wide array of tissues and cell types. scATAC.Explorer allows fast query and retrieval of publicly available scATAC-seq datasets, and it contains accessibility matrices, cell type annotations and other metadata information for each dataset. scATAC.Explorer is available both in GitHub and as an R Bioconductor package, and can easily be integrated into existing R and Python analysis pipelines, allowing researchers to quickly analyze scATAC-seq data from multiple sources.

Keywords Single-cell sequencing, Chromatin accessibility, scATAC-seq, Database and search tool

Assay for transposase-accessible chromatin using sequencing (ATAC-seq) is a method used to capture chromatin accessibility data from cells[1]. Large samples of cells, ranging from 500 to 50,000 can be assayed to identify open chromatin sites in their genome, and this epigenetic data can be used to investigate gene regulation[2]. Peaks representing open regions of chromatin can be identified and likely represent enhancers, promoters, and other genomic elements regulating genes[3,4]. Determining gene regulation in differing cell types[5,6] or differentiating diseased and healthy cells[7] from one another in tissue samples has increasingly become a research question of interest answered using ATAC-seq. Since being first described in 2013, many datasets and publications of ATAC-seq data have become available and have been used to investigate different cell types and tissues[8,9]. Due to ATAC-seq being performed on a sample of cells, its chromatin accessibility data represents an "average" of all cells sampled[10]. Since tissue samples are usually a heterogenous mixture of cells types, and since different sub-populations of cells have differing chromatin accessibility profiles from one another, we would need to assess chromatin accessibility at the single-cell level to capture the heterogeneity in a cell type[11]. Single-cell ATAC-seq (scATAC-seq) is a modified method of ATAC-seq that obtains chromatin accessibility data at a single cell resolution, and with high throughput compared to bulk ATAC-seq[12]. This enables the collection of epigenome data from many individual cells at the same time. Several different sequencing strategies have been developed to perform scATAC-seq, either using a split and pool approach (such as sci-ATAC-seq[10]), or microfluidics approach[12]. Microfluidics devices, such as 10×Genomics Chromium[13] or Fluidigm C1[14] have become increasingly popular, with the 10× device allowing large amounts of cells to be sequenced very quickly. Chromatin accessibility data at the resolution of individual cells can also be processed to identify cell types[15]. This has enabled studies to take samples from complex tissues and compare chromatin accessibility of various cell types within. scATAC-seq has also been used to investigate regulatory changes in the epigenome that occur during cell differentiation in stem and hematopoietic progenitor cells[16,17]. scATAC-seq has found many applications, including identifying relevant cell types in immune-mediated diseases[18], studying epigenomes of differentiation stages by assaying differentiating cells at several time points[19], and investigating chromatin accessibility changes in the tumor microenvironment (TME) where a large variety of malignant, immune, and non-cancerous cells can be individually sequenced[20].

[1]Department of Pathology and Lab Medicine, University of Western Ontario, London, ON N6A 5C1, Canada. [2]Children's Health Research Institute, London Health Sciences Centre Research Institute (LHSCRI), A5-144, 800 Commissioners Road East, London, ON N6C 2R5, Canada. [3]Department of Computer Science, University of Western Ontario, London, ON N6A 5B7, Canada. [4]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. ✉email: pshoosh@uwo.ca

In the past years, many large scATAC-seq datasets have been generated. However, unlike single-cell RNA-seq technology, a comprehensive database of scATAC-seq datasets in a consistent format is still missing[21]. In addition, many studies use previously generated scATAC-seq datasets when testing their new analysis techniques and pipelines or when comparing the existing datasets against their own newly sequenced scATAC-seq data. Searching through literature, determining sequencing technology, or processing pipelines used, and finding datasets containing cell types of interest can be time consuming. Due to scATAC-seq being a relatively new sequencing method, new processing pipelines are being created resulting in new data outputs, and datasets must be curated to confirm the formatting is consistent between datasets. Furthermore, there are multiple formats for scATAC-seq data in use that are not consistently used between pipelines. This means data that has been formatted as an input for one pipeline (e.g. ChromVAR[22]) will not seamlessly work with other pipelines (such as snapATAC[23] or cisTopic[24]) without reformatting them. These issues make it difficult to integrate the growing number of scATAC-seq datasets into analyses and limits new discoveries that could be made by doing so. Other newly developed single cell sequencing technologies, such as scRNA-seq are facing similar issues. Databases and search tools have been created to gather these datasets into easily accessible sources[21,25]. Attempting this for scATAC-seq datasets would be of great benefit to the research community. For this reason, we have created a comprehensive collection of available scATAC-seq datasets in a consistent format, and have made them available in a R package named scATAC.Explorer, which can easily be integrated with scATAC-seq data analysis pipelines. scATAC.Explorer is both a database and a search tool, which allows users to search for curated scATAC-seq datasets using several possible dataset attributes, such as organism, disease type, and cell types present in the sequenced data. This will allow for rapid search and comparison of available datasets.

Results

Data collection

scATAC.Explorer is a database of scATAC-seq datasets. We collected publicly available datasets for inclusion within our database from several resources. These resources include the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), 10 × Genomics' sample sequences, and lab website data repositories. We searched for any scATAC-seq data using search terms such as scATACseq, single cell ATAC-seq, and single cell chromatin accessibility. In total, 39 datasets meeting our search criteria were found and included in scATAC.Explorer (Table 1). 23 of these datasets were acquired from GEO, 13 datasets from 10 × Genomics' website, one from the UCSC Cell Browser[58], one from the Descartes lab's Human Chromatin Accessibility During Development project website[38], and one from a lab's public Gitlab repository (Fig. 1). Cell type and clustering data was also collected if available.

Datasets and metadata

Each dataset in scATAC.Explorer consists of (1) a peak-by-cell matrix containing the chromatin accessibility data per-cell; (2) cell type annotations for the cells in the peak-by-cell matrix when available; and (3) metadata summarizing the dataset-related information. Each dataset retains their original data values as provided by their authors and no further analysis such as batch effect correction were conducted. The metadata information includes which paper and author published the dataset; which sequencing technology was used; what tissues the sequenced cells were sampled from; any diseases the cells sequenced may represent; cell types identified as being present in the dataset by the authors; and more. This metadata allows users of the package to explore what datasets are available, and search and find datasets relevant to their interests. It is accessible as a table through the package and is also attached to the metadata property of the *SingleCellExperiment* R objects that the datasets are downloaded as. To view dataset metadata without downloading the package, see Supplementary Table S1.

Database contents and statistics

Cell types

scATAC-seq datasets in our database represent a wide range of cell types from many different tissues, with some datasets containing several different tissues and cell types (Fig. 2A). 18 of the 39 datasets in scATAC.Explorer contain cells from central nervous system (CNS), including astrocytes, microglial cells, oligodendrocytes, excitatory neurons, and inhibitory neurons. Eight datasets contain immune-related cell types such as CD8 + T, CD4 + T, NK, Naïve T, Basophil, Th17, CD26 + & CD26- T, and memory T cells. Eight datasets contain hematopoietic data containing erythrocytes, hematopoietic stem, common myeloid progenitor (CMP), megakaryocyte-erythrocyte progenitor (MEP), granulocyte-monocyte progenitor (GMP), and other cells differentiated from hematopoietic stem cells. There are also assorted datasets containing scATAC-seq data relating to the kidney cells, epithelial cells, and lung cells. In addition, two datasets contain data from multiple organs, including the heart, intestines, liver, muscle, spleen, and more.

Organisms

Datasets representing several organisms are also present within scATAC.Explorer, with 20 datasets containing human data, 19 datasets containing data from mice (Mus musculus), with 3 of these datasets containing data from both mouse and human cell lines. Individual datasets are also available for the common fruit fly (Drosophila melanogaster), chicken (Gallus gallus), and zebrafish (Danio rerio) (Fig. 2B).

Disease types

Some datasets in scATAC.Explorer contain cells related to diseases, such as cells sampled from cancer patients (Supplementary Table S1). More specifically, of the 39 datasets included in scATAC.Explorer, 32 are from healthy organisms; one dataset contains both normal and acute myeloid leukemia cells; one contains cells sampled from astrocytoma patients; one sampled from a T cell leukemia cell-line model; one sampled from metastasized

Dataset	Organism	Broad cell categories present	Number of cells	Sequencing technology	Cell type annotations available	Cell clustering available
Satpathy et al. Nature Biotechnology[7]	Homo sapiens	Hematopoietic, Immune, TME	110,632	10×Genomics Chromium	N	N
Buenrostro et al. Cell[16]	Homo sapiens	Hematopoietic	2953	C1 Single-Cell Auto Prep System (Fluidigm)	N	N
Corces et al. Nature Genetics[26]	Homo sapiens	Hematopoietic, Stem, Leukemia, Immune	576	C1 Single-Cell Auto Prep System (Fluidigm)	N	N
Cusanovich et al. Cell[5]	Mus musculus	Neuronal, Muscle, Epithelial, Hematopoietic, Immune, Small Intestine, Large Intestine, Heart, Kidney, Liver, Lung, Spleen, Testes	81,173	C1 Single-Cell Auto Prep System	Y	Y
Jansen et al. PLoS Computational Biology[19]	Mus musculus	Immune	227	C1 Single-Cell Auto Prep System (Fluidigm)	N	N
Cusanovich et al. Science[10]	Homo sapiens, Mus musculus	Hematopoietic, Kidney, Stem	1066	MiSeq (Illumina), v2 300 cycle kit	N	N
Cusanovich et al. Nature[27]	Drosophila melanogaster	Embryonic	23,075	NextSeq High output 300 cycle kit	N	Y
Abbasi et al. Cell[28]	Mus musculus	Fibroblast	1578	10×Genomics Chromium	N	N
Abdelsamed et al. Nature Immunology[29]	Homo sapiens	Immune	1,122,623	10×Genomics Chromium, HiSeq4000, NovaSeq 6000	Y	N
Muench et al. Nature[30]	Mus musculus	Hematopoietic, Immune, Stem	41,753	10×Genomics Chromium, NovaSeq 6000	Y	N
Wang et al. Cancer Discovery[20]	Homo sapiens	Neuronal, Immune, Endothelial, TME	23,002	10×Genomics, Chromium	N	N
Rubin et al. Cell[31]	Homo sapiens	Epithelial	288	System (Fluidigm) to preform Perturb-ATAC-seq	N	N
Duren et al. PNAS[32]	Mus musculus	Embryonic	415	C1 Single-Cell Auto Prep System (Fluidigm)	N	N
Williams et al. Developmental Cell[33]	Gallus gallus	Neuronal	1259	10×Genomics Chromium	N	N
Sinnamon et al. Genome Research[34]	Mus musculus	Neuronal	3160	NextSeq 500	N	N
Preissl et al. Nature[35]	Mus musculus	Neuronal	14,554	MiSeq (Illumina)	N	N
Zhu et al. Nature[36]	Homo sapiens, Mus musculus	Neuronal	44,395	custom procedure for Paired-sequencing, HiSeq 2500	N	N
Satpathy et al. Nature Medicine[37]	Homo sapiens	Immune	1344	Modified C1 Single-Cell Auto Prep System (Fluidigm)	N	N
Cao et al. Science[15]	Homo sapiens, Mus musculus	Kidney, Lung	15,401	Custom procedure for sci-CAR ATAC-seq, NextSeq 500	N	N
Domcke et al. Science[38]	Homo sapiens	Adrenal, Neuronal, Eye, Heart, Intestine, Kidney, Liver, Lung, Muscle, Pancreas, Spleen, Stomach, Thymus	720,613	Custom procedure for sci-ATAC-seq3 sequencing	Y	Y
Ziffra et al. Nature[39]	Homo sapiens	Neuronal	77,354	10×Genomics Chromium	Y	N
Corces et al. Nature Genetics[40]	Homo sapiens	Neuronal	70,631	10×Genomics Chromium	N	N
Bella et al. Nature[41]	Mus musculus	Embryonic, Neuronal	12,964	10×Genomics Chromium	N	N
Xu et al. Human Molecular Genetics[42]	Homo sapiens	Breast cancer lymph node	7202	10×Genomics Chromium	N	N
Ranzoni et al. Cell Stem Cell[43]	Homo sapiens	Embryonic, Hematopoietic, Stem	3611	Hiseq 2000	N	N
Avagyan et al. Blood Advances[44]	Danio rerio	Hematopoietic, Stem	12,354	10×Genomics Chromium	N	N
10×Genomics fresh embryonic mouse brain Cell Ranger 1.2.0[45]	Mus musculus	Neuronal	4008	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics fresh embryonic mouse brain Cell Ranger 1.1.0[46]	Mus musculus	Neuronal	5380	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics cryopreserved embryonic mouse brain Cell Ranger 1.2.0[47]	Mus musculus	Neuronal	4751	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics cryopreserved embryonic mouse brain Cell Ranger 1.1.0[48]	Mus musculus	Neuronal	6339	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics flash frozen embryonic mouse brain Cell Ranger 1.2.0[49]	Mus musculus	Neuronal	3789	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics flash frozen embryonic mouse brain Cell Ranger 1.1.0[50]	Mus musculus	Neuronal	4580	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics fresh cortex adult mouse brain Cell Ranger 1.2.0[51]	Mus musculus	Neuronal	3880	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics fresh cortex adult mouse brain Cell Ranger 1.1.0[52]	Mus musculus	Neuronal	5337	10×Genomics Chromium, Illumina NovaSeq	N	Y
10×Genomics 10K human PBMC GEM 1.1 Cell Ranger 1.1.0[53]	Homo sapiens,	PBMC	10,247	10×Genomics 10,247 Chromium, Illumina NovaSeq	N	Y
Continued						

Dataset	Organism	Broad cell categories present	Number of cells	Sequencing technology	Cell type annotations available	Cell clustering available
10 × Genomics 10K human PBMC GEM 1.1 Cell Ranger 2.0.0[54]	Homo sapiens	PBMC	9688	10 × Genomics Chromium, Illumina NovaSeq	N	Y
10 × Genomics 5K human PBMC GEM 1.1 Cell Ranger 2.0.0[55]	Homo sapiens	PBMC	4623	10 × Genomics Chromium, Illumina NovaSeq	N	Y
10 × Genomics 1K human PBMC GEM 1.1 Cell Ranger 2.0.0[56]	Homo sapiens	PBMC	1004	10 × Genomics Chromium, Illumina NovaSeq	N	Y
10 × Genomics 500 human PBMC GEM 1.1 Cell Ranger 2.0.0[57]	Homo sapiens	PBMC	484	10 × Genomics Chromium, Illumina NovaSeq	N	Y

Table 1. scATAC-seq datasets contained within scATAC.Explorer.

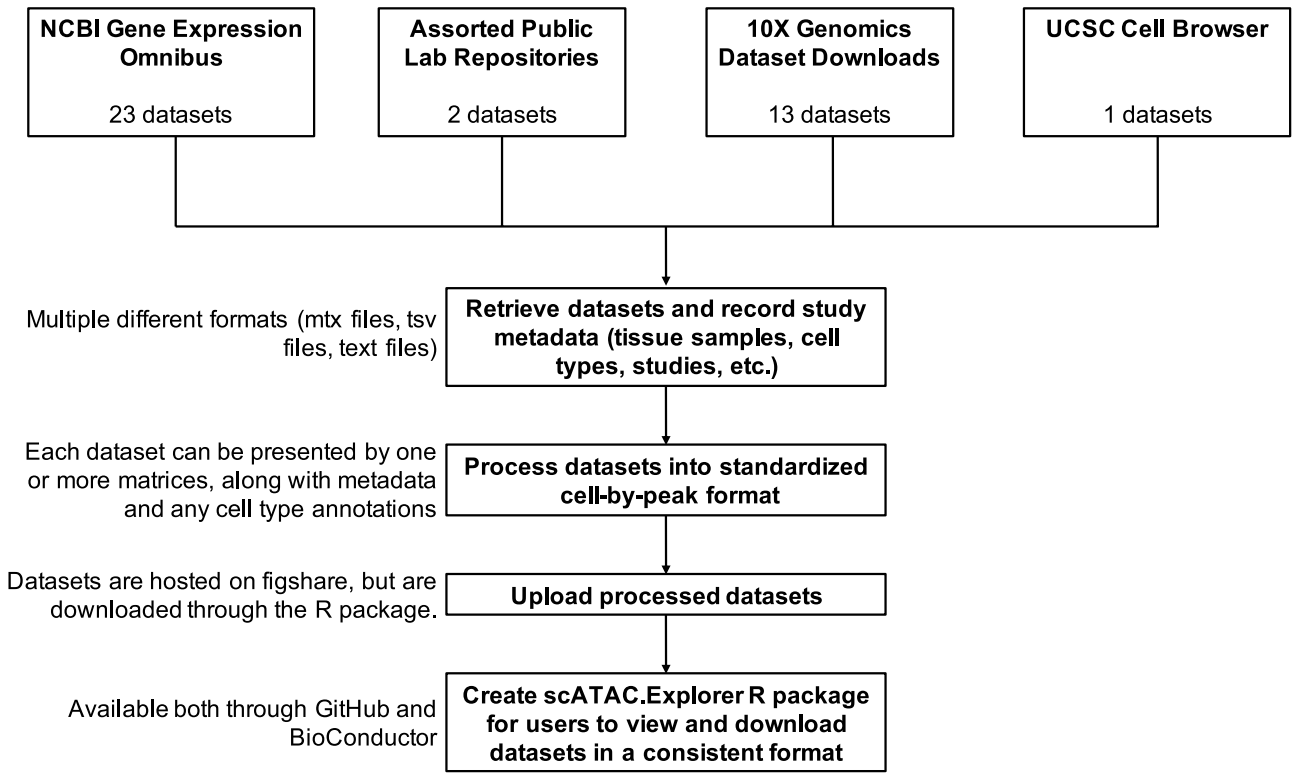

Fig. 1. Our Workflow for Collecting Datasets and Creating the scATAC.Explorer Package. We collected datasets from a variety of sources for inclusion into our package. These datasets were processed through the above workflow for inclusion in our package. Scripts are available on our GitHub for processing datasets in commonly encountered formats retrieved from these sources into our standardized format.

luminal B breast cancer lymph nodes; one from type-1 diabetes patients; one from healing wounds in mice, and one from a GATA2 deficiency syndrome animal model.

Cell type annotations and cluster information

Of the 39 datasets in scATAC.Explorer, 17 datasets provided cluster assignment information, and six datasets provided cell type annotations. We reformatted cluster/cell type annotation data and provide them in our database when available. It is noted that many studies have conducted their own cell type annotations but did not release the labels in their data. These studies will be considered to have no cell type labels available, but their cell type annotation method will still be described in the metadata table.

Sequencing technology

Multiple techniques have been developed to preform single cell ATAC-seq sequencing, with newer techniques generating transcriptome and other multi-omic data for individual cells[59]. The devices used to generate these datasets can also vary from commercial solutions, such as the 10 × Genomics Chromium system[60], to more manual protocols using common sequencing devices and reagents[10,12]. From the 39 datasets included in scATAC.Explorer, 24 of them are generated using 10 × Genomics Chromium, 7 using the C1 Single-Cell Auto Prep System, and the rest were produced using custom procedures (Fig. 2C).

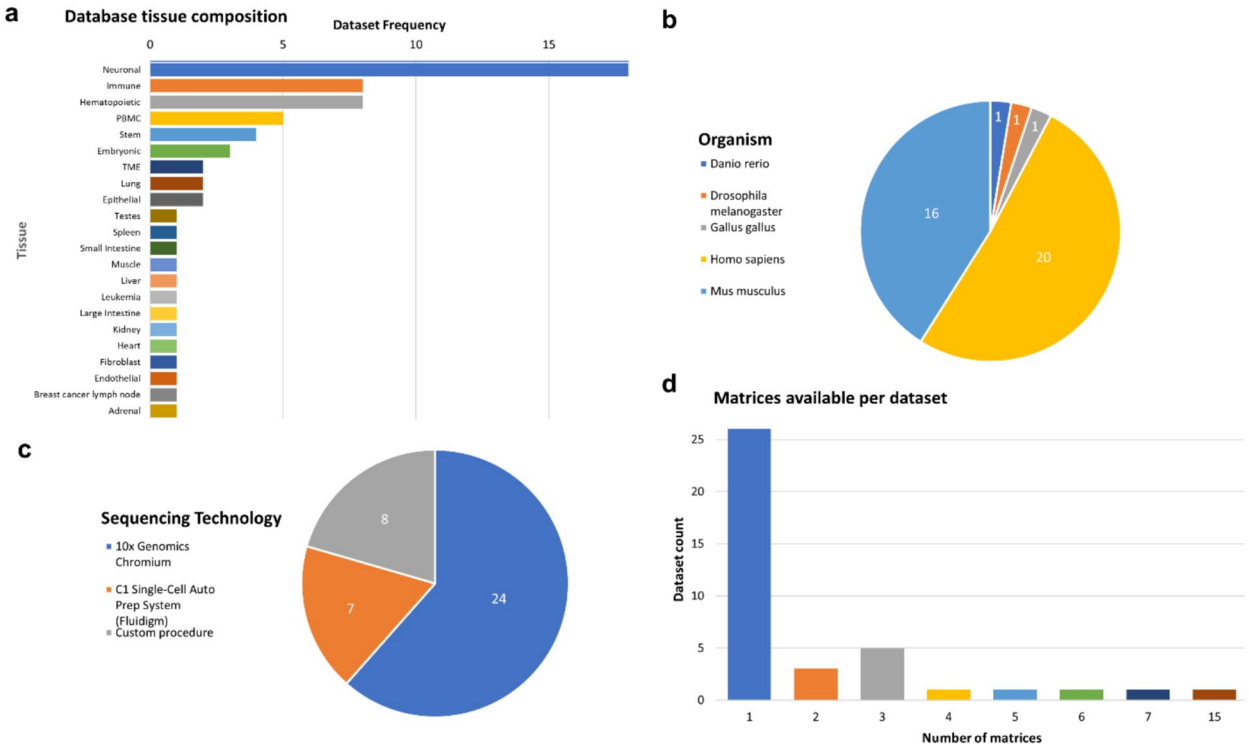

Fig. 2. Database Contents and Statistics. (**a**) Broad cell categories present in datasets included in scATAC.Explorer. The majority of datasets are neuronal, but data from cells in nearly every organ are present. A large amount of neuronal, immune, stem, and hematopoietic datasets have been generated due to the role cell type and cell differentiation has in these systems. (TME: Tumor microenvironment; PBMC: Peripheral Blood Mononuclear Cells). (**b**) Organisms present in datasets included in scATAC.Explorer. Many datasets contain mouse cells due to ease of sampling for disease study. Several datasets used cultured cells from assorted cell lines, such as mouse Patski, A549, and NIH/3T3 cell lines, or human GM12878, HEK293T, and HepG2 cell lines. One dataset sequenced a mix of mouse and human cells to test clustering algorithms, and these datasets are specified as containing both organisms. (**c**) Sequencing technologies used by datasets included within scATAC.Explorer. The majority of datasets used either the 10 × Genomics Chromium[13] or C1 Fluidigm[14] devices. These devices are commercially available automated sequencing devices capable of performing scATAC-seq at much higher throughput than initial non-automated methods, allowing much larger datasets to be gathered in a shorter time frame. Several custom procedures were also used that modified traditional scATAC-seq sequencing methods to concurrently collect scRNA-seq data. (**d**) Number of matrices required to represent datasets. The majority of datasets are represented using a single peak-by-cell matrix, but some datasets require as many as 15 matrices. Matrices are labeled within the returned metadata table if metadata_ only is specified as TRUE. Only data in some matrices may be related to a user's research question, and the matrix labels can be used to determine which matrices the user might keep.

scATAC.Explorer search capability

scATAC.Explorer is not only a database, but also a search tool. scATAC.Explorer provides a search capability, where users can look up and return their datasets of interest, view the peaks by cell accessibility matrices, cell type labels and cluster assignments (if available), and the metadata (Fig. 3). They can continue by either using R for data visualization and analysis, or save the datasets in a format to be analyzed by other programming language, such as Python, MATLAB.

The *queryATAC()* function serves as the main function for querying both metadata to discover relevant datasets within the database, and to query and retrieve the processed data for matching datasets (Fig. 3 and Supplemental Fig. 1). Parameters specified within *queryATAC()* allow the selection of datasets based on several factors (Table 2). This includes parameters such as specifying a disease of interest, a broad cell category or cell type of interest, or datasets sequenced from a specific organism.

queryATAC() can be specified to return an R *dataframe* containing the metadata of matching datasets by changing the *metadata_only* parameter to *True*. For example, if only the metadata of the diabetes dataset was of interest, *queryATAC(disease = "diabetes", metadata_only = TRUE)* would retrieve the metadata of any diabetes datasets. It should also be mentioned that for some scATAC-seq datasets, multiple matrices are available in the original study (Fig. 2D). In such cases, the metadata table has been split and will show the multiple results come from the same original dataset. Here we have assigned a name to each matrix in the metadata table that can be used to differentiate the matrices from one another.

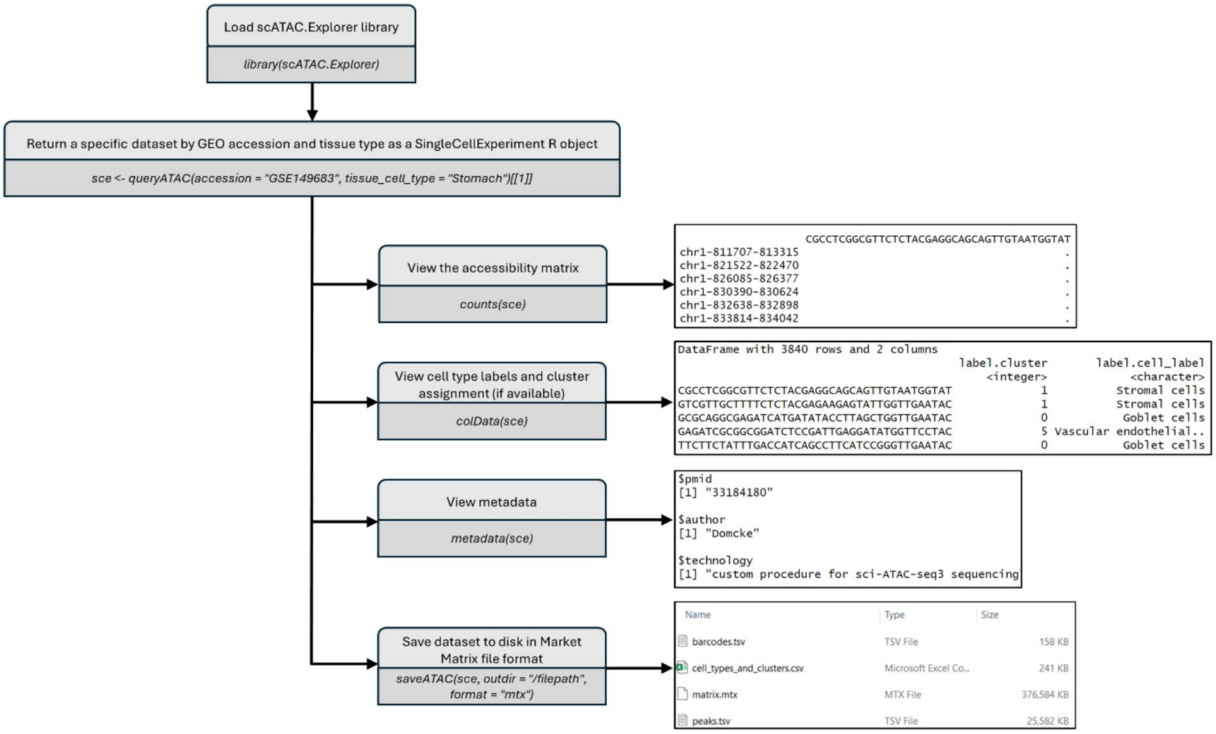

Fig. 3. A flowchart of data query and analysis using scATAC.Explorer. scATAC.Explorer provides a search and analysis capability, where users can find their datasets of interest, view the accessibility matrix, cell type labels, cluster assignments, and the metadata for the retrieved dataset. They can continue by either using R for data visualization and analysis, or saving the datasets in formats compatible with other programming languages, including Python and MATLAB.

Parameter	Example usage	Description
Accession	queryATAC(accession = "GSE129785")	Search by dataset accession code
Organism	queryATAC(organism = "mus musculus")	Search for datasets containing specific organism
Disease	queryATAC(disease = "leukemia")	Search by disease
Journal	queryATAC(journal = "Nature")	Search by journal dataset published in
Sequence_tech	queryATAC(sequence_tech = "10 × Genomics Chromium")	Search by sequencing technology/devices used
Broad_cell_category	queryATAC(broad_cell_category = "immune")	Search by broad cell category
Tissue_cell_type	queryATAC(tissue_cell_type = "astrocyte")	Search by specific cell types found in datasets
Has_cell_type_annotation	queryATAC(has_cell_type_annotati on = TRUE)	Search for datasets with cell type annotations
Has_cluster_annotation	queryATAC(has_cluster_annotation = TRUE)	Search for datasets containing cluster annotations
Metadataonly	queryATAC(metadata_only = TRUE)	Return metadata instead of data

Table 2. Parameters usable within the *queryATAC()* function.

Exporting scATAC.Explorer datasets

Queried datasets can also be exported from the package to files for use outside of R. This is particularly useful for data analysis pipelines built in other programming languages (ex. EpiScanpy[61] in python). *saveATAC* provides this functionality, saving the input objects to disk at the specified location (Supplemental Fig. 2). The files are saved in market matrix (.mtx) and Hierarchical Data Format version 5 Annotated Data (.h5ad) formats, as these were the file formats that are most often used by the scientific community and we encountered most often when gathering datasets. If the user chooses the .mtx format, then the count matrix is saved in a sparse *.mtx* file, with the peaks and cell ID/barcodes saved in separate *.tsv* files.

Example use cases of scATAC.Explorer

Case 1: Non-linear dimension reduction and clustering of adult mouse cortex sample

In this use case, we show how an adult mouse scATAC-seq dataset can be retrieved using scATAC.Explorer (Fig. 4). Then the downstream analysis includes normalization, dimension reduction and clustering of cells. To find our datasets of interest, we first used *queryATAC(metadata = TRUE)* to look over all available datasets

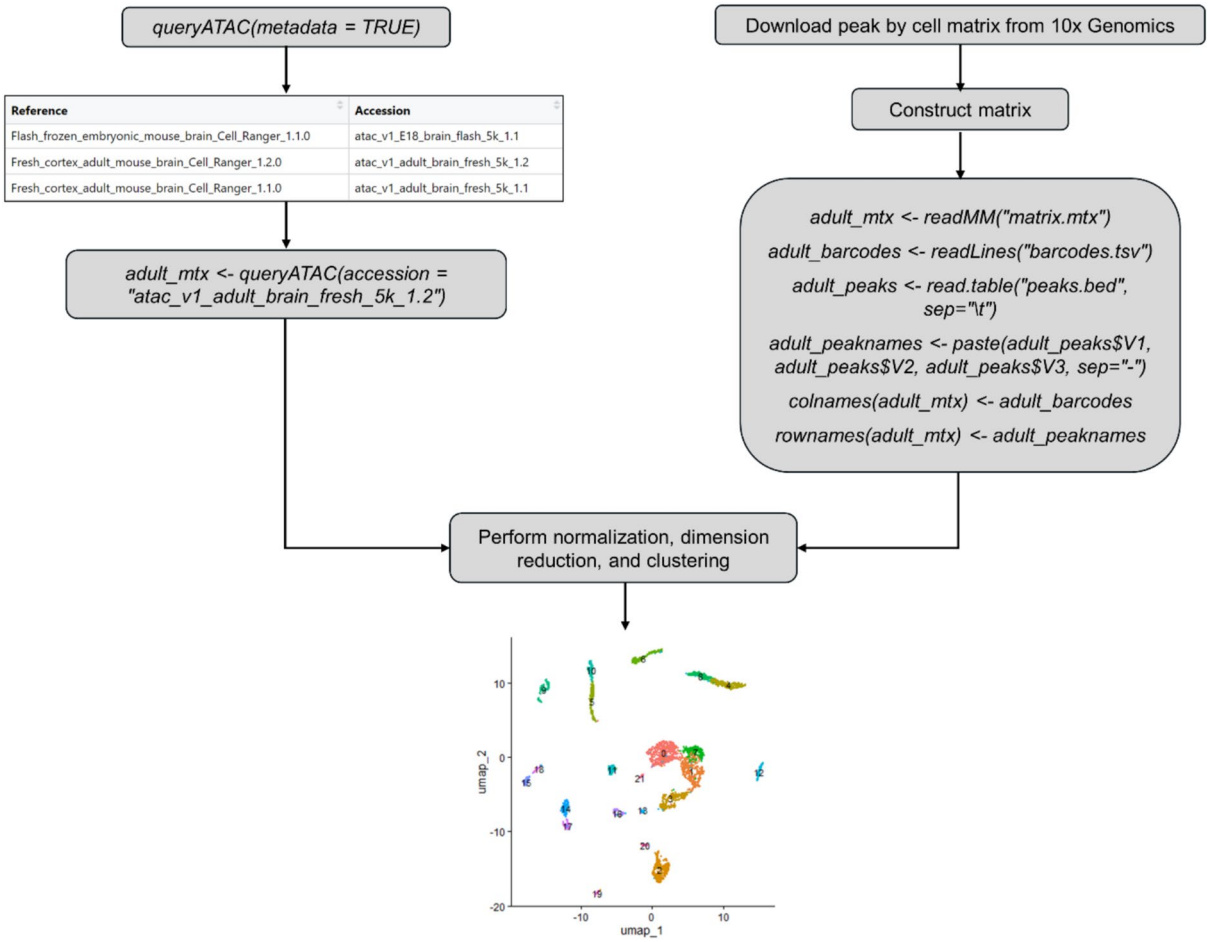

Fig. 4. Comparison of Dataset Retrieval Methods. The left side demonstrates an example of using scATAC.Explorer to retrieve a specified dataset and perform dimension reduction and clustering for visualization. The right side demonstrates how a user might retrieve the dataset without scATAC.Explorer. Both retrieval methods yield matching results after performing the same downstream analysis.

available in the metadata table. One dataset containing a sample from an adult mouse brain was selected. The *SingleCellExperiment* objects representing this dataset were then processed in Seurat[62] for scATAC-seq analysis. We used latent semantic indexing (LSI) to perform dimension reduction within Seurat. We then identified clusters of cells by a shared nearest neighbor (SNN) clustering algorithm which is visualized on a graph. This was then repeated using data collected from the original source and compared with results generated from scATAC.Explorer, showing that scATAC.Explorer can be used to generate the same results with much less data preprocessing required.

Case 2: Performing analysis on datasets with cell type data available

For several datasets within scATAC.Explorer, we provided the cell type annotations released by the authors so that users would not need to repeat the analysis steps needed to generate the cell type labels. In this case study, we show how datasets with cell type annotations can be found, and once a particular dataset of interest is found, how cell type annotations can be visualized in Seurat, and how a UMAP projection by cell type can be performed (Fig. 5). In our example, we looked at a dataset containing a variety of annotated T cells, containing cell types such as naive T cells, short-lived effector memory T cells (TEM cells), and several types of beta cell-specific CD8 + T cells. We generated UMAP projections of the cells based on the cell type annotations available in the scATAC.Explorer (Fig. 5).

Case 3: Using scATAC.Explorer on high-performance computing resources

Datasets in scATAC.Explorer range in size, some containing a few hundred cells, others containing hundreds of thousands of cells. High-performance computing resources allow the analysis of these large-scale datasets that require large amounts of memory to work with. To show that scATAC.Explorer can be used on high performance computing nodes, we retrieved a dataset[5] containing ~ 80,000 cells from scATAC.Explorer and performed rudimentary analysis (Louvain clustering in Seurat at different resolutions). The size of the dataset required the usage of a node with > 60 GB of memory on high-performance computing resources.

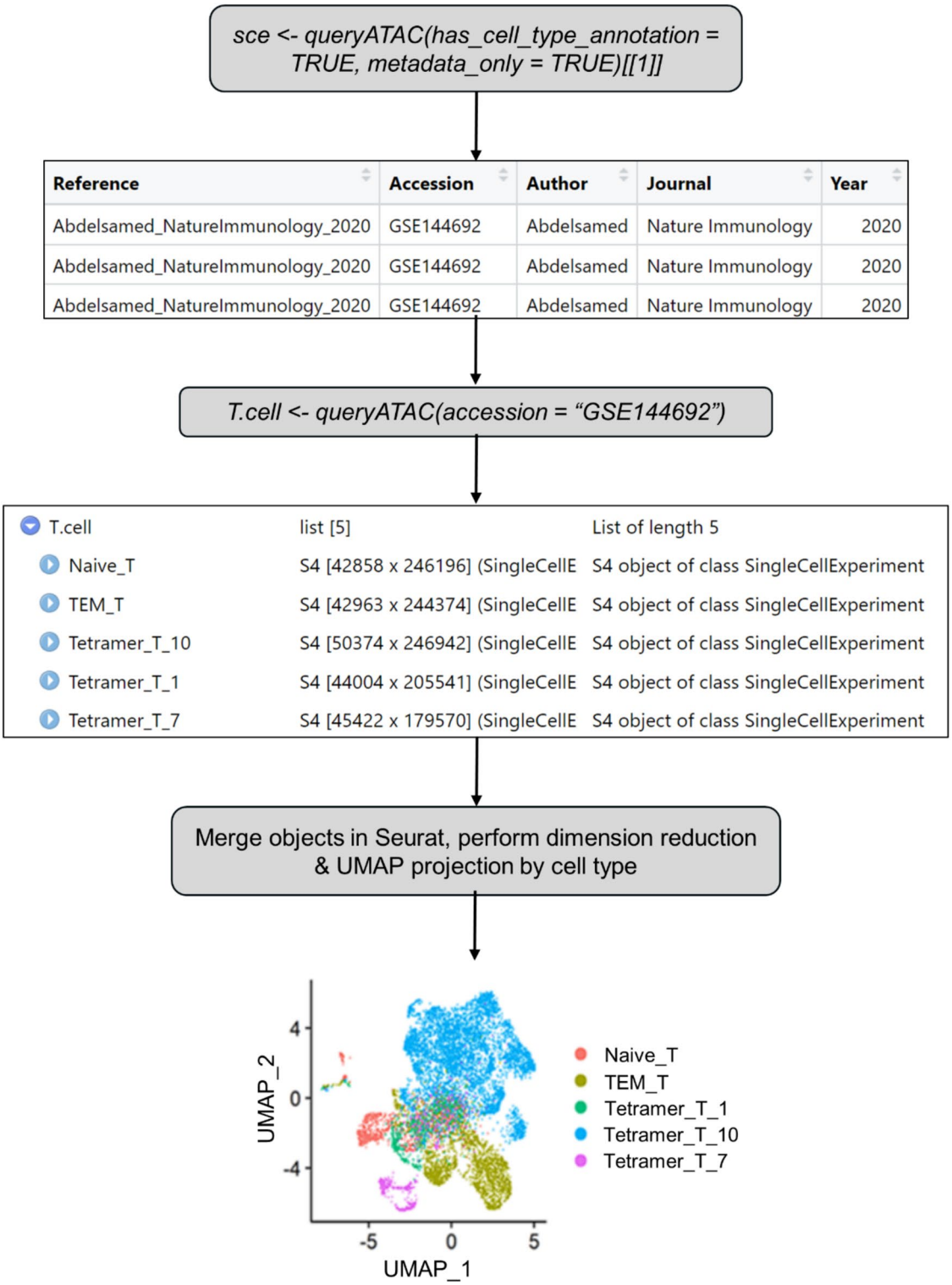

Fig. 5. Workflow for Retrieving Cell Type Annotated Datasets, Followed by UMAP Visualization and Clustering of Dataset Containing Cell Type Annotations. Several cell type annotated datasets are present within scATAC.Explorer and can be retrieved. In this example, we queried for metadata on all datasets with cell type annotations, and selected one based on our interest. The selected dataset contains T cells sampled from type 1 diabetes patients. Naive T cells, short-lived effector memory T cells (TEM cells), and several subpopulations of beta cell-specific CD8 + T cells were identified in the original study, and those annotations have been used in this example. We performed LSI dimension reduction using Seurat, and generated a cell type labelled UMAP projection using the available cell type annotations from the retrieved dataset.

Discussion

scATAC-seq is applicable to investigate gene regulation in a range of cell types, diseases, and developmental processes. Due to its wide range of potential applications in research, many scATAC-seq datasets have been generated for various tissues, cell types, or cell differentiation stages. The number of these datasets will continue to grow as new techniques and pipelines for scATAC-seq analysis are developed. In order to facilitate easy access to a wide range of published scATAC-seq datasets, we have created an R-based database and search tool, named scATAC.Explorer, that allows the rapid survey and collection of currently available scATAC-seq datasets in a consistent format. We searched through currently published literature, as well as NCBI, for publicly available scATAC-seq datasets. We collected and reformatted available datasets into a standardized format, and collected metadata relating to each dataset, as well as cell type and cluster annotations when available. At the current version of scATAC.Explorer, a total of 39 datasets are available, containing data relating to a variety of tissues, cell types, cell developmental pathways. scATAC.Explorer can be used to find, retrieve, and integrate scATAC-seq datasets into the analysis quickly and easily. scATAC.Explorer provides several features that we believe would be quite useful for the scientific community. This includes:

1. Using scATAC.Explorer, researchers can easily query and access one or more datasets of interest from a diverse array of publicly available scATAC-seq datasets.
2. scATAC.Explorer can be used to retrieve a selection of datasets using a combination of search parameters (Table 2). This allows users to retrieve datasets of interest without having to manually search through all available scATAC-seq datasets. This also allows users to quickly find the research studies that generated data relevant to their area of study.
3. Consistent formatting of datasets in scATAC.Explorer allows datasets to be quickly retrieved and accessed, with minimal processing required. This avoids the tedious process of manually formatting several different datasets retrieved from different publications before they can be used for analysis, and also means that any dataset retrieved using scATAC.Explorer can be integrated into analysis pipelines.
4. Cell type annotations are available alongside peak-by-cell matrices for some datasets, when publicly available. This allows extended analysis to be performed without requiring more work from researchers to identify cell types.
5. We regularly maintain our package and add scATAC-seq datasets as they are released, if the dataset is publicly available. If users would like to add their dataset to the package, we have created a Github issue template where they can request a dataset to be added. This consists of a form where users can include a description of the dataset, metadata associated with the dataset, a link to the data, and instructions on formatting the dataset. We have also included R scripts to format datasets from common scATAC-seq formats into our standard format.

As mentioned previously, in datasets that included multi-omic data, such as scRNA-seq or CHIP-seq alongside scATAC-seq data, only the scATAC-seq data was included in our database. Currently next generation sequencing is moving in the direction of integrating multi-omic data to gain new insights into the interplay of the epigenome, transcriptome, proteome, and more. Future work could be done to create a repository of multi-omic single cell data. The creation of scATAC.Explorer through the modification of the previously created TMExplorer[25], which was used for single-cell RNA-seq data, proves the possibility of including multiple modalities of single-cell sequencing in our database. In the future, we plan to investigate the unification of scATAC.Explorer into a multi-omic database containing scATAC-seq, scRNA-seq, and other single cell next generation sequencing technologies.

Methods

Data collection

We collected publicly available datasets for inclusion within our database from several resources. The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) was searched for any scATAC-seq data using search terms such as scATACseq, single cell ATAC-seq, and chromatin accessibility (Fig. 1). Similar keywords were used to find scATAC-seq research articles published in scientific journals, where the data sequenced in them is available on GEO or hosted on their lab's website. Datasets containing processed scATAC-seq data from any organism were included in our database. In cases where multi-omics data was available (such as accompanying scRNA-seq, ChIP-seq, etc.) only the scATAC-seq data was included. 10 × Genomics also makes datasets sequenced using their technology available on their website, and these were included as well[63].

Standard dataset format

For ease of use within analysis pipelines, we modified the obtained datasets and made all datasets available in a standard format in our database (Fig. 1). We chose to format each dataset's scATAC-seq data into two objects. A peak-by-cell matrix was used to store read counts, where the columns of the matrix are cell barcodes, and the row names of the matrix are genomic locations of the peaks represented by chromosome number, start, and end. Sparse representations of the peak-by-cell matrix were created for every dataset in the form of R *dgCMatrix* object class. We also made the dense formats of these matrices available for each dataset, although sparse matrices are likely to be preferred by most users due to having a smaller memory footprint. The second object is an R *dataframe* used to store cell type labels and/or cluster number assignment, if provided within the original dataset. The *dataframe* has three columns: one for the cell ID/Barcode; a column specifying the cluster it was assigned to (or an empty element if this was not provided); and a column specifying the cell type label assigned to each cell. These two objects—the peaks by cells matrix and the cell type/cluster data frame—are accessible within one R object for easy access.

Formatting datasets

The number of steps required to modify the datasets into our standard format varied, as the data was obtained in many different formats. Many of the datasets were from studies that used 10 × Genomics Chromium for sequencing and output their scATAC-seq data using the Cell Ranger ATAC pipeline[60] in Market Exchange (MEX) format. MEX format splits data into 3 separate files, keeping read counts in Matrix Market sparse matrix files (.mtx file extension), using accompanying TSV files to store cell barcodes/IDs, and BED files to store peak region data. Cell Ranger ATAC can also output files in Hierarchical Data Format (HDF5 format), and one of our datasets was in this format. HDF5 stores data in a somewhat similar fashion to MEX format, differing from it by storing the data within a single H5 file structured hierarchically by cell barcode, peaks, and counts objects. MEX and HDF5 format were converted into our standard format by reading in the sparse matrix, assigning the cell barcodes as the column names for the matrix, and assigning the peaks regions as the matrix row names. 17 of our 39 datasets were obtained in MEX or HDF5 format, while the remainder had a wide range of formats. Some studies stored all their data within a single BED file in a dense peak-by-cell matrix, which we converted into a sparse matrix. In some other studies, the reads counts were stored in a CSV file for each cell. We converted it by appending the contents of each CSV file together to form a matrix, resulting in each CSV file serving as a column of the peak-by-cell matrix. Several datasets were obtained in compressed sparse row (CSR) matrices, with their cell barcodes listed sequentially in the header of the file, and with an accompanying peaks text file to specify genomic regions. The conversion was similar to the process for MEX, where we constructed the counts matrix using the CSR matrix, then named the columns using the cell ID's from the CSR file's header, and named the rows using the peak regions from the peaks text file.

Metadata and database query

Previously we have developed an R-based platform for querying and serving single-cell RNA sequencing datasets from tumour microenvironments (TMExplorer)[25]. In the current study, we have used TMExplorer as a platform and modified it to generate our scATAC.Explorer database and search tool of single-cell ATAC-seq data. This allowed us to leverage existing infrastructure and methods for querying and serving datasets in scATAC.Explorer. scATAC.Explorer allows users to search for datasets based on several attributes, such as tissue type, disease, or organism (Table 2). To make this possible, metadata of each dataset featured in the database was gathered into a table for use within the package. To homogenize tissue and cell lines, we have used standardized ontologies in the metadata. More specifically, we used the BRENDA Tissue and Enzyme Source Ontology (BTO) at BioPortal (https://bioportal.bioontology.org/) and the Cellosaurus database (https://www.cellosaurus.org/). For diseases, we used the Mondo Disease Ontology (MONDO) at BioPortal. Per dataset information in the metadata table is used to perform queries and return relevant datasets. To create a metadata table for all available datasets, the GEO accession code, PubMed ID, sequencing technology, score type, clustering method, cluster/cell type annotations, organism, genome build, broad cell categories present, tissue/sample type, disease, and summary of data were recorded for each dataset and included in the metadata table (Supplementary Table S1). After the cell-by-peak matrices were created for all datasets following the previously described format, we also recorded the number of cells, genome regions, and number of matrices in each dataset, and added this extra information to the metadata table. Cell type and cluster annotations were also formatted into the previously described format in R *dataframes* for each dataset. These files were saved in R's native file format (.RDS files) for R objects storing each R object in an individual file. These are the files that are retrieved when a query matches with a dataset within the metadata table. RDS files for each dataset were stored in figshare[64] repository accessed by the R package using URL links. We have made sparse format matrices available for all datasets in the form of R *dgCMatrix* objects. Also, dense matrix formats are available for datasets where the resulting dense matrix is not excessively large in memory. Most scATAC-seq analysis tools and pipelines, such as Signac[65] expect input data to be in sparse matrices, so no sparse to dense conversion is needed for datasets retrieved from scATAC.Explorer.

Multiple matrices

Some of the datasets in our database contained data from different organisms, multiple patient samples, or experimental conditions. When obtained, these were usually stored in multiple separate matrices due to the peak genomic regions being unique from one another, as combining them may degrade data quality. We have chosen to keep these matrices separate during our reformatting. When a query matches with a dataset containing multiple matrices, each matrix will be returned. Within the metadata table, a matrix name field is used to describe the data contained within each matrix, such as containing data from a tumor microenvironment sample (TME) vs from a PBMC sample within the same dataset.

Adding new datasets

To request new scATAC-seq datasets be added to the package, users can create a GitHub issue on the package repository to provide information on the dataset of interest. We have created an issue template for users to submit their dataset through, as well as provided R scripts to process some of the commonly encountered data formats for inclusion into the package. Also included in the issue template is instructions to fill out a provided metadata *.csv* file that is formatted the same as the metadata for all datasets in the package. This will allow users to create a GitHub issue that provides the metadata and processed data for their dataset that they wish to include in the package, which we can then evaluate and incorporate into the package. In addition, we plan to add new scATAC-seq datasets to our package as they get published and become publicly available. As new datasets will become available in subsequent versions of the database, we will host the older versions of the package on Github. With each new version release, any changes to the database will be documented. In order to ensure compatibility of the scATAC.Explorer with the evolutions of the R versions, we will submit each version of the package to R BioConductor which will be tested and verified.

Data availability

All datasets included in scATAC.Explorer are publicly available datasets, and are referenced both within the package by metadata, and within this paper. We also have hosted all these datasets processed into a consistent format on Figshare[66] at: https://figshare.com/projects/scATAC_Explorer_A_Single-Cell_ATAC-Seq_Database_and_Search_Tool/101474. Users can also download the cell type labels and cluster number assignments in the flat text format (i.e., CSV) from there too. A web-based version of the metadata table with a query functionality is available at: https://shooshtarilab.github.io/scATAC.Explorer.

Code availability

The source code of scATAC.Explorer is open source and available on Github at: https://github.com/shooshtarilab/scATAC.Explorer. The package is also hosted and available through Bioconductor at: http://bioconductor.org/packages/release/data/experiment/html/scATAC.Explorer.html. If users would like to add their data to the repository, they can do so by creating a GitHub issue using an issue template we have provided. We will then review the submitted data to be included in the package. The issue can be created at: https://github.com/shooshtarilab/scATAC.Explorer/issues.

Received: 8 August 2023; Accepted: 1 July 2025
Published online: 16 January 2026

References

1. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat. Methods* **10**, 1213–1218 (2013).
2. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).
3. Fullard, J. F. et al. An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).
4. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858-871.e8 (2018).
5. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309-1324.e18 (2018).
6. Liu, C. et al. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci. Data* **6**, 65 (2019).
7. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
8. Nott, A. et al. Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
9. Speedy, H. E. et al. Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. *Nat. Commun.* **10**, 3615 (2019).
10. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
11. Pott, S. & Lieb, J. D. Single-cell ATAC-seq: Strength in numbers. *Genome Biol.* **16**, 172 (2015).
12. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
13. Single Cell ATAC. *10x Genomics* https://www.10xgenomics.com/products/single-cell-atac.
14. Fluidigm | Products | C1. *Fluidigm C1 System* https://www.fluidigm.com/products/c1-system.
15. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
16. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535-1548.e16 (2018).
17. Florian, M. C. et al. Aging alters the epigenetic asymmetry of HSC division. *PLOS Biol.* **16**, e2003389 (2018).
18. Konstantopoulos, D. et al. Op0099 functional mapping of synovial fibroblast populations in health and arthritic disease: Insights into the pathogenic remodeling of synovial microenvironment. *Ann. Rheum. Dis.* **79**, 65–65 (2020).
19. Jansen, C. et al. Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLOS Comput. Biol.* **15**, e1006555 (2019).
20. Wang, L. et al. The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.* **9**, 1708–1719 (2019).
21. Cao, Y., Zhu, J., Jia, P. & Zhao, Z. scRNASeqDB: A database for RNA-seq based gene expression profiles in human single cells. *Genes* **8**, 368 (2017).
22. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
23. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
24. Bravo González-Blas, C. et al. cisTopic: Cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods.* **16**(5), 397–400 (2019).
25. Christensen, E., Naidas, A., Chen, D., Husic, M. & Shooshtari, P. TMExplorer: A tumour microenvironment single-cell RNAseq database and search tool. *PLoS ONE* **17**, e0272302 (2022).
26. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
27. Cusanovich, D. A. et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
28. Abbasi, S. et al. Distinct regulatory programs control the latent regenerative potential of dermal fibroblasts during wound healing. *Cell Stem Cell* **27**, 396-412.e6 (2020).
29. Abdelsamed, H. A. et al. Beta cell-specific CD8 + T cells maintain stem cell memory-associated epigenetic programs during type 1 diabetes. *Nat. Immunol.* **21**, 578–587 (2020).
30. Muench, D. E. et al. Mouse models of neutropenia reveal progenitor-stage-specific defects. *Nature* **582**, 109–114 (2020).
31. Rubin, A. J. et al. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* **176**, 361-376.e17 (2019).
32. Duren, Z. et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci.* **115**, 7723–7728 (2018).
33. Williams, R. M. et al. Reconstruction of the global neural crest gene regulatory network in vivo. *Dev. Cell* **51**, 255-276.e7 (2019).
34. Sinnamon, J. R. et al. The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Res.* **29**, 857–869 (2019).

35. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
36. Zhu, C. et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
37. Satpathy, A. T. et al. Transcript-indexed ATAC-seq for precision immune profiling. *Nat. Med.* **24**, 580–590 (2018).
38. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
39. Ziffra, R. S. et al. Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).
40. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet* **52**, 1158–1168 (2020).
41. Di Bella, D. J. et al. Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **595**, 554–559 (2021).
42. Xu, K. et al. Integrative analyses of scRNA-seq and scATAC-seq reveal CXCL14 as a key regulator of lymph node metastasis in breast cancer. *Hum. Mol. Genet.* **30**, 370–380 (2021).
43. Ranzoni, A. M. et al. Integrative single-cell RNA-Seq and ATAC-Seq analysis of human developmental hematopoiesis. *Cell Stem Cell* **28**, 472-487.e7 (2021).
44. Avagyan, S. et al. Single-cell ATAC-seq reveals GATA2-dependent priming defect in myeloid and a maturation bottleneck in lymphoid lineages. *Blood Adv.* **5**, 2673–2686 (2021).
45. 10x Genomics. atac_v1_E18_brain_fresh_5k -Datasets -Single Cell ATAC-Official 10x Genomics Support. *10x Genomics Embryonic Brain Fresh* https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_E18_brain_fresh_5k.
46. 10x Genomics. atac_v1_E18_brain_fresh_5k -Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics Embryonic Brain Flash* https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_E18_brain_fresh_5k.
47. 10x Genomics. atac_v1_E18_brain_cryo_5k -Datasets-Single Cell ATAC -Official 10x Genomics Support. *10x Genomics Embryonic Brain Cryo* https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_E18_brain_cryo_5k.
48. 10x Genomics. atac_v1_E18_brain_cryo_5k-Datasets -Single Cell ATAC-Official 10x Genomics Support. *10x Genomics Embryonic Brain Cryo* https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_E18_brain_cryo_5k.
49. 10x Genomics. atac_v1_E18_brain_flash_5k-Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics Embryonic Brain Flash* https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_E18_brain_flash_5k.
50. 10x Genomics. atac_v1_E18_brain_flash_5k-Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics Embryonic Brain Flash* https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_E18_brain_flash_5k.
51. atac_v1_adult_brain_fresh_5k-Datasets-Single Cell ATAC-Official 10x Genomics Support. https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k.
52. 10x Genomics. atac_v1_adult_brain_fresh_5k -Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics Adult Brain Fresh* https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k.
53. 10x Genomics. atac_pbmc_10k_nextgem-Datasets-Single Cell ATAC -Official 10x Genomics Support. *10x Genomics PBMC 10K Cell Ranger 1.1.0* https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_pbmc_10k_nextgem.
54. 10x Genomics. atac_pbmc_10k_nextgem-Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics PBMC 10K Cell Ranger 2.0.0* https://support.10xgenomics.com/single-cell-atac/datasets/2.0.0/atac_pbmc_10k_nextgem.
55. 10x Genomics. atac_pbmc_5k_nextgem-Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics PBMC 5K Cell Ranger 2.0.0* https://support.10xgenomics.com/single-cell-atac/datasets/2.0.0/atac_pbmc_5k_nextgem.
56. 10x Genomics. atac_pbmc_1k_nextgem-Datasets-Single Cell ATAC -Official 10x Genomics Support. *10x Genomics PBMC 1K Cell Ranger 2.0.0* https://support.10xgenomics.com/single-cell-atac/datasets/2.0.0/atac_pbmc_1k_nextgem.
57. 10x Genomics. atac_pbmc_500_nextgem-Datasets-Single Cell ATAC-Official 10x Genomics Support. *10x Genomics PBMC 500 Cell Ranger 2.0.0* https://support.10xgenomics.com/single-cell-atac/datasets/2.0.0/atac_pbmc_500_nextgem.
58. Speir, M. L. et al. UCSC cell browser: Visualize your single-cell data. *Bioinformatics* **37**, 4578–4580 (2021).
59. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-cell multiomics: Multiple measurements from single cells. *Trends Genet.* **33**, 155–168 (2017).
60. *10X Genomics Cell Ranger*. (10X Genomics).
61. Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. EpiScanpy: Integrated single-cell epigenomic analysis. *Nat. Commun* **12**, 5228 (2021).
62. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
63. Datasets—Single Cell ATAC-Official 10x Genomics Support. *10X Genomics Single Cell ATAC Datasets* https://support.10xgenomics.com/single-cell-atac/datasets.
64. Figshare—credit for all your research. https://figshare.com/.
65. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).
66. Singh, J. FigShare. *J. Pharmacol. Pharmacother.* **2**, 138–139 (2011).

Acknowledgements

The authors would like to acknowledge Digital Research Alliance of Canada for providing high-performance computing resources for this project. This project was supported by the Government of Canada through the Natural Sciences and Engineering Research Council (NSERC) (RGPIN/04062-2021 and DGECR-2021-00298) and the internal research funds from Western University, Schulich School of Medicine and Dentistry, Children's Health Research Institute, and Lawson Health Research Institute. PS is also supported by an Early Investigator Award from the Ontario Institute for Cancer Research (OICR). The funders had no role in the study design, data collection and analysis, decision to publish, or the preparation of the manuscript.

Author contributions

P.S. has contributed to the conception of the work. A.G.K., E.C., J.W. and P.S. designed the database structure and contributed to the creation of the database. A.G.K. processed and formatted the public datasets and added them to the new database. A.G.K. and P.S. contributed to the interpretation of data and have drafted the work. A.G.K. and J.W. prepared all figures. All authors contributed to writing and editing the paper.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-09962-4.

Correspondence and requests for materials should be addressed to P.S.

Reprints and permissions information is available at www.nature.com/reprints.

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