NAR Genomics and Bioinformatics, 2024, **6**, lqae135
https://doi.org/10.1093/nargab/lqae135
Advance access publication date: 8 October 2024
Pipelines and Workflows for Biological Data Analysis

scATAcat: cell-type annotation for scATAC-seq data

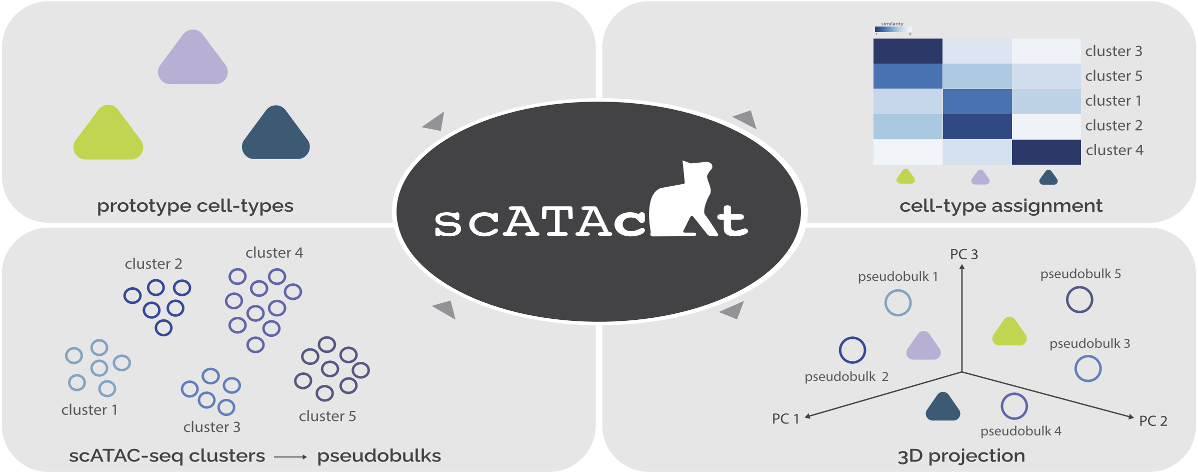
Aybuge Altay and Martin Vingron*

Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany
*To whom correspondence should be addressed. Tel: +49 30 8413 1152; Email: vingron@molgen.mpg.de

Abstract

Cells whose accessibility landscape has been profiled with scATAC-seq cannot readily be annotated to a particular cell type. In fact, annotating cell-types in scATAC-seq data is a challenging task since, unlike in scRNA-seq data, we lack knowledge of 'marker regions' which could be used for cell-type annotation. Current annotation methods typically translate accessibility to expression space and rely on gene expression patterns. We propose a novel approach, scATAcat, that leverages characterized bulk ATAC-seq data as prototypes to annotate scATAC-seq data. To mitigate the inherent sparsity of single-cell data, we aggregate cells that belong to the same cluster and create pseudobulk. To demonstrate the feasibility of our approach we collected a number of datasets with respective annotations to quantify the results and evaluate performance for scATAcat. scATAcat is available as a python package at https://github.com/aybugealtay/scATAcat.

Graphical abstract



Introduction

Chromatin structure can control the accessibility of potential gene regulatory elements in a dynamic and cell-type-specific manner and therefore plays a critical role in gene regulation (1). It has been shown that accessibility measurements offer valuable additional information to gene expression and have been demonstrated to be more cell-type specific than expression data (2). This is due to much of the cell-type-specific information within the genome being located in enhancer regions, which are captured by accessibility assays. Today, even single-cell technologies are applied to accessibility measurements. One of the most commonly used techniques is scATAC-seq (3). Further advances in single-cell genomics have facilitated the profiling of thousands of cells simultaneously, even at a multi-modal level (4). While these protocols lead to an enormous growth in the volume of data generated, the large datasets also allow deeper insights into complex biological systems.

Numerous types of cells can be found in an organism. Traditionally, these cell types have been defined based on phenotypic characteristics (5). With the advent of single cell RNA sequencing (scRNA-seq) it has become a common practice to cluster cells based on their transcriptome profile, in the expectation that these clusters correspond to cell types (6). The results, e.g. of the Human Cell Atlas (7), rely on this assumption and cell-types are frequently assigned to individual cells based on the determined transcriptome.

In this work, we deal with the problem of assigning cell-types to the cells for which a single cell ATAC sequencing (scATAC-seq) experiment has been performed. This is an important problem if we want to exploit the detailed cell-type

Received: January 19, 2024. **Revised:** September 11, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 23, 2024

information that is assumed to be contained in the accessibility of the chromatin. Since most data and most cell-type annotation is available for scRNA-seq data, many existing approaches capitalize on the RNA level by predicting a proxy to gene expression from the accessibility landscape. This proxy is frequently referred to as 'gene activity score' (8,9). Given this gene activity score, annotation methods for scRNA-seq data can be carried over. These fall roughly into marker-based and reference-based methods. Marker-based annotation usually is a manual process which requires expert knowledge and testing of multiple markers. Recent work (10,11) has tried to better support this process. Transferring labels from a known cell-type to the query cells by way of reference-based annotation can be achieved either by the statistical similarity metrics (12–16) or by machine learning models (17–21). (22) provides a benchmarking study comparing these techniques. A recent tool, Cellcano (23) is developed specifically for cell-type annotation in scATAC-seq data and demonstrated superior performance over the existing approaches.

An alternative and potentially more accurate way to annotate cell-types in scATAC-seq data is to perform within-modality annotation, that is to say, annotating the cell types using annotated-ATAC-seq data. Ideally, this will rely on an annotated scATAC-seq reference. One of the recent efforts to tackle this problem is EpiAnno (24) which leverages existing annotated scATAC-seq data and employs a Bayesian neural network framework for supervised cell-type annotation. However, this method is not computationally scalable (23). Often times the scATAC-seq reference itself is a product of annotation via expression markers on RNA levels. As a surrogate to annotated scATAC-seq references, an alternative approach is to use characterized-bulk ATAC-seq data as a reference. This approach has been suggested by (25).

Here, we put forward *scATAcat–scATAC-seq cluster annotation tool* for annotation of cell-types in scATAC-seq data based on characterized bulk ATAC-seq data. scATAcat provides results comparable to or better than many approaches that perform cell-type annotation. Rather than using the genes and their predicted activity as the features for assignment, we focus on the regulatory elements in the chromatin. We explore the use of FAC-sorted scATAC-seq data as a demonstration of the methodology. We further apply our annotation method and compare it to other approaches using six more datasets of blood and brain cells. For comparison we study four approaches, namely, marker-based annotation, reference-based label-transfer, Cellcano and EpiAnno. We will discuss the challenges and biases in cell-type annotation in scATAC-seq data.

Materials and methods

scATAcat

Method outline

Given a set of cells for which a scATAC-seq experiment has been performed, scATAcat seeks to annotate cells with their corresponding cell-type. To be more precise, we first cluster the scATAC-seq data yielding what we call 'pseudobulk' clusters. This serves the purpose of aggregating the sparse counts from several cells and to obtain better accessibility profiles at the pseudobulk level. All the cells in a pseudobulk cluster will inherit the assignment of the pseudobulk cluster. Furthermore, for annotation we require prototypes of the possible cell-types. Typically, for each cell-type there may be several replicate prototype samples available. Such prototypes can, e.g. come from existing characterized bulk ATAC-seq data. scATAcat takes as input these bulk accessibility profiles of distinct cell-types to which the computed pseudobulk clusters can be matched. scATAcat co-embeds the prototype accessibility profiles with the pseudobulk clusters in a principal component analysis (PCA) space and exploits the distance in this space to assign the cell-type labels.

Preprocessing of reference bulk ATAC-seq data

To allow for integration of different data sets we rely on candidate cis-regulatory elements (cCREs) provided by ENCyclopedia Of DNA Elements (ENCODE) project (26) (version 2) as a feature space for all the datasets in this study. We calculate the cCRE coverage of the bulk ATAC-seq datasets. Next, we identify the differentially accessible cCREs in pairwise manner. This approach is adopted based on our hypothesis that these particular cCREs hold the most discriminative information and are cell-type specific. We use the DiffBind R package (v3.0) (27) with DESeq2 (28) as the underlying method after applying sequencing depth normalization provided by the package. We identify significantly differential accessible regions by filtering for FDR ≤ 0.05.

This analysis results in variable number of differential regions depending on the similarly between the compared cell-types. To ensure equal contribution of various comparisons and mitigate the potential bias, we gather the same number of features, by default 2000 regions, unless a comparison contains fewer differential regions, from each pairwise comparison to derive a final feature space. When choosing this number we took into consideration the lower bound for the differential regions, as well as the total number of regions for the final feature set. Note that this number may be adjusted based on the specific cell-types under consideration. These features, hereafter named *differential cCREs*, are used as the final feature space for the rest of the analysis. We apply library size normalization and $\log_2$ transformation to the original data and finally subset the matrix to *differential cCREs*.

Preprocessing of scATAC-seq data

Due to the sparsity of scATAC-seq data, some preproccessing is needed. For scATAC-seq data, we calculate the coverage of cCREs by counting the number of fragments within each cCRE region for each single cell. This yields a *cell-by-cCRE matrix*. Features which occur in less than k (default $k = 3$) cells get eliminated. Additionally, we get rid of the Y chromosome to avoid gender biases. On the level of cells, we filter out cells with <1000 and >80 000 non-zero features, as well as doublet cells detected by AMULET (29).

Analysis suites for scATAC data (8,9) process scATAC-seq data using TF-IDF. More specifically, we apply TF-logIDF normalization (30). This results in re-weighted features (cCREs) by assigning greater weight to more important features. Then we subset the data to the *differential cCREs* as defined by the reference bulk ATAC-seq data. We then reduce the dimension via PCA and continue with the standard scanpy clustering pipeline (31). We determine the nearest neighbors with *neighbors* function (n_pcs = 50, n_neighbors = 30) and compute a UMAP (32) embedding. UMAP provides a low dimensional, non-linear embedding in which one can visualize a clustering. Next, we apply Leiden clustering (33) with *leiden* function. In this study we set the Leiden resolution parameter to 1 for all the datasets used.

The cells in one cluster form a *pseudobulk*. To represent this pseudobulk we compute its accessibility profile by adding the read coverage for each feature across the cluster member cells. Just like for the bulk data, we apply library size normalization and $\log_2$ transformation to the pseudobulk matrix and subset the matrix to *differential cCREs*.

Co-embedding prototypes with pseudobulks

To make datasets comparable we apply a z-score transformation. First consider the *bulk-by-cCRE matrix* of prototypes. We substitute each matrix entry by its z-score with respect to its column, i.e. mean and standard deviation are computed over the cells in the column. Call this matrix X. For purposes of integrating the data, the z-score transformation on the *pseudobulk-by-cCRE matrix* is performed with the very same column-means and column-standard deviations of the first matrix. Call this transformed pseudobulk-by-cCRE matrix Y. Both matrices have as many columns as there were differential features. The number of rows, n_X, of the transformed bulk-by-cCRE matrix is the number of annotated prototype bulk samples given. These may describe a number k of different cell-types. For one cell-type there may be several (replicate) bulk samples. The number of rows of Y, n_Y, is just the number of pseudobulk clusters.

After this transformation step, we proceed to co-embed prototype bulk samples and pseudobulk clusters into one space using PCA. To this end, we obtain the eigenvectors of the matrix X. Depending on the number of dimensions one wants to keep, concatenate this number of eigenvectors into a matrix W. The data can then be represented in a lower dimensional space by transforming the original matrix as follows:

$$\hat{X} = W^t \times X \quad (1)$$

where W^t corresponds to the transpose of the matrix W. The pseudobulk samples get projected onto the same PCA space as follows:

$$\hat{Y} = W^t \times Y \quad (2)$$

As a result, we have projection with both pseudobulk and prototype bulk samples embedded into the same space.

Annotating pseudobulk clusters

Co-embedding $\hat{X}$ and $\hat{Y}$, and visualizing the projection in 3D PCA space facilitates a simplified interpretation of cell-type relationships. However, determining the annotations solely based on a visualisation may be tedious. In particular, the first three PCs used in the projection might not suffice to fully capture the inherent structure of the high dimensional data. Following common practice, we therefore keep a larger number of dimension (typically 30 if there are enough samples available) to compute Euclidean distances in this high-dimensional embedding space.

To obtain a matching between pseudobulk clusters and prototype samples, we first compute centroids for each of the prototype cell-types. Next, we compute the Euclidean distances (in many dimensions) between the pseudobulk clusters and the k centroids, yielding a $n_Y \times k$ matrix D. Finally, we annotate each pseudobulk cluster by its closest centroid of a cell-type.

Going beyond the mere assignment, inspection of the matrix D can provide valuable information about the data. This will already show when decisions are ambiguous. Additionally, we also compute a matrix of size $(k + n_Y) \times (k + n_Y)$ with all the distances among pseudobulks and prototypes. Hierarchical clustering on this matrix helps to understand the grouping among pseudobulks and prototypes, as well as to detect possible outliers.

scATAcat is implemented in Python and made compatible with scanpy (31) and anndata (34) libraries for an easy integration.

Marker-based annotation

Marker-based annotation is one of the standard ways to annotate cell-types in scRNA-seq data. This method includes exploiting the expression of so-called marker genes, which are the genes exclusively expressed in a known cell-type, as a predictor of the associated cell-type. Application of this method for scATAC-seq data is enabled by the use of a metric called a 'gene activity score' as a stand-in for the expression of genes. The gene activity score typically measures the level of accessibility around the gene body and nearby enhancer regions with the assumption that expression can be inferred from accessibility. Therefore, this value is also referred as *predicted* expression. In this study we use the gene score calculations as defined by Signac (8), which considers the number of fragments within the gene body and 2 kb upstream region of each gene to determine for each gene's activity. Here, we refer to the resulting matrix as *gene-score matrix*. Note that the scATAC-seq data is processed as outlined by Signac with default parameters to obtain this *gene-score matrix*. Once the gene activity scores are calculated, one can look at the predicted expression levels of the marker genes to determine the cell-type of a cluster. In this work, CellMarker 2.0 (35) is used as the marker gene resource, if not specified otherwise.

Reference-based label-transfer

Another commonly used way to annotate cells in single-cell data includes *transferring* the cell labels from existing annotated reference atlases. Here, we use one of the widely adapted tools, Seurat (v3) (13,36) for this purpose. Briefly, Seurat (v3) employs canonical correlation analysis (CCA) to determine so-called 'anchors' between the query and reference datasets. Anchors represent the cells with the highest similarity and therefore serves as a correspondence between reference and the query. The assumption is that the anchor cells define matching cell states and create a shared space between the query and reference. Each anchor pair is then scored based on the correspondence in connecting cells' shared nearest neighbor (SNN) graphs (37). Anchors serve as bridges to *transfer* information, such as cell-type label, from reference to query cells.

It is important to note that in order for reference and query to align and for bridges to be defined, (i) there needs to be a shared feature space between the reference and query and (ii) both the data need to be processed similarly. To satisfy the first requirement, we use *gene-score matrix* of scATAC-seq data described above, effectively defining our features as genes. For the second requirement, we use Seurat toolkit and apply 'log-Transformation' for both the datasets. We next determine the variable features using Seurat's 'vst' method and select top 3000 variable features. Subsequently, we determine anchors using 'FindTransferAnchors' method with 'cca' as the reduction method. Note that, in our experience, the latest Seurat v4 integration with reduction method 'spca' typically results in limited and insufficient number of anchors. Consequently, we chose to use 'cca' reduction method to attain more informed annotations. Lastly, we use 'TransferData' function to transfer cell-type information from reference to query data. We apply this approach for cell-type annotation, which we referred to as *label-transfer*.

Cellcano

Cellcano is a supervised cell-type annotation method specifically designed for scATAC-seq data (23). Cellcano relies on the ArchR (9) defined gene scores for both reference and query datasets. ArchR calculates these gene scores by considering the accessibility around the gene body and 5 kb upstream of the transcription start site. This value is then scaled by incorporating the *regulatory* accessibility signal around the *gene boundary*, which is 100 kb on either side of the gene, with exponential decay function. Cellcano performs the cell-type annotation task in two rounds. In the first round, reference gene scores of the cells with known cell-labels are used to train a multi-layer perceptron (MLP), which then predicts the cell-types in the query data. In the second round, query cells annotated with high confidence are selected as 'anchor' cells. These cells serve as the training data for the second training round where a self-Knowledge Distiller (KD) model predicts the cell labels of non-anchor cells.

For all the used datasets, we first used 'Cellcano preprocess' to obtain gene scores as per ArchR's default settings. Next, we trained the Cellcano model with the scATAC-seq reference dataset using 'Cellcano train' option. Finally, we applied 'Cellcano predict' option with the trained model to annotate cell types in the query scATAC-seq data. The datasets we used as query and reference datasets are listed in Table 1. Cellcano is employed with default parameters.

EpiAnno

EpiAnno is another supervised cell-type annotation method specifically designed for scATAC-seq data (24). This methods employs a Bayesian neural network for its supervised learning strategy. Unlike other methods, EpiAnno does not rely on any gene scoring approximation and uses peaks or genomic regions as features.

The first step of the method is preprocessing starting with feature selection. Peaks occupied by less than three percent of the cells, which is controlled by 'peak_rate' parameter, are eliminated to reduce noise. The resulting dataset undergoes TF-IDF transformation, followed by z-normalization to standardize the data. After reprocessing, the training data is fed into the Bayesian neural network along with the cell-type labels. Besides cell-type annotation, EpiAnno generates latent representation for each cell based on its label, which is then mapped back to original feature space through the nonlinear Bayesian neural network. This enables interpretable embedding.

For the trained model to make predictions, it is essential that the feature sets of both the training and test data are the same. EpiAnno achieves this by *unifying* the peak set to those of the test dataset. Specifically, it counts the reads of the training dataset that fall into the peaks of the test dataset, ensuring that both datasets share the same features. This enables inter-dataset predictions. In this study, we used the peaks identified by the default ArchR pipeline as the features for our datasets. Following the instructions in original paper, we *unified* the training data features to the peaks of test data. The datasets we used as query (test) and reference (training) are listed in Table 1. For all cell-type annotation tasks, we employed the 'run_crossdataset_projection' script with default settings, except for PBMC and FACS bone marrow scATAC-seq datasets where we adjusted the 'peak_rate' parameter to 0.05 due to memory constraints.

Performance assessment

In the assessment of annotation methods, we first adapted the performance metrics introduced in (38). Given an experiment with a set of cells, each cell has a ground-truth annotation, whereby each cell is assigned a label (e.g. B cell). While each method provides an annotation, the annotations may not cover all cells due to variations in the preprocessing filters used by each method. For a fair comparison, we only consider cells annotated by both the ground-truth and *all* compared methods. More importantly, the annotation labels may not encompass all the labels that are used by the ground-truth annotation as each method uses different reference/training data with a distinct set of labels. To address this discrepancy, we implemented two different performance evaluations. In the first evaluation, we determine the set of common cell-type annotations across all methods and the ground-truth, here referred as 'common cell-types'. We consider only those cells whose ground-truth annotation is one of the common cell-types. Despite being conservative, this approach enables unbiased comparison and minimizes the impact of difference in reference data labels. The second strategy differs in that we evaluate each method individually such that the common cell-types are defined as cell-types shared between the method in question and the ground-truth. This strategy allows for the evaluation of a possibly larger number of cell-type annotation labels and the corresponding cells, depending on the number of common cell-types annotated by a method.

Let the set of annotation labels be $\mathcal{L} = 1, 2, ..., K$. Further, let C be the set of cells c_i. We only use cells that have been annotated by the ground-truth as one of the common cell-types. Let m_i denote method i, where we adopt the convention that m_1 is the ground-truth. m_i is a mapping from cells to labels, i.e. it is of the form $m_i(c_j) = l$ with $l \in \mathcal{L}$. This would mean that mapping m_i labels cell j as type l. Note that we only use those labels that are predicted by all methods, i.e. we reduce the label set to $\bigcap_i Im(m_i)$. Accordingly, we also reduce the cells in the ground-truth and only keep those cells for which the label is a member of this reduced label set. The first metrics we use, *accuracy* (Acc), quantifies the proportion of cells with accurately assigned cell-type annotations across cells and calculated as:

$$Acc = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(m_i(c_j) = m_1(c_j)) \tag{3}$$

where n is the number of labeled cells, c_j is the jth cell, m_i is the mapping from cell to predicted cell-type and m_1 is the mapping from cell to true (ground-truth) cell-type, and $\mathbb{I}$ is the indicator function that returns 1 when its argument is true and 0 otherwise.

Balanced accuracy ($BAcc$) measures the average accuracy across cell-types and calculated as:

$$BAcc = \frac{1}{r} \sum_{i=1}^{r} Acc_i \tag{4}$$

Table 1. Summary of the query and reference datasets utilized for cell-type annotation methods compared in this study

Query dataset	scATAcat prototypes	Seurat label-transfer reference	Cellcano reference	EpiAnno reference
FACS BM scATAC-seq	**feasibility study:** aggregated single cells for BMMC progenitors **application:** bulk ATAC-seq of sorted BMMC progenitors	BMMC CITE-seq of hematopoietic progenitors	CD34+ BM progenitors scATAC-seq	CD34+ BM progenitors scATAC-seq
10X PBMC sc-multiome **PBMC scATAC-seq**	bulk ATAC-seq of sorted PBMCs	PBMC CITE-seq	PBMC scATAC-seq	PBMC scATAC-seq
NeurIPS BMMC sc-multiome **BMMC scATAC-seq**	bulk ATAC-seq of sorted BMMCs	BMMC CITE-seq	BMMC scATAC-seq	BMMC scATAC-seq
Corces brain scATAC-seq **Brain cortex scATAC-seq**	bulk ATAC-seq of sorted brain cell-types	brain primary motor cortex snRNA-seq	brain cerebral cortex sc-multiome	brain cerebral cortex sc-multiome

Abbreviations: BM, bone marrow; BMMC, bone marrow mononuclear cells; PBMC, peripheral blood mononuclear cells. A comprehensive version of this table with detailed dataset references is available in Supplementary Table S1.

In this equation, r denotes the number of cell-types in reduced cell-type subset $\bigcap_i Im(m_i)$. Acc_i refers to the accuracy of the ith cell-type.

Another variant of measuring accuracy is *cluster accuracy* (*CAcc*). The clusters are defined by scATAcat pipeline using the Leiden clustering as explained above (see section *Preprocessing of scATAC-seq data*). For each annotation method, as well as the ground-truth, every cluster is annotated as the most abundant cell-type of its constituent cells. We only consider the clusters with more than 10 cells. Let p_i denote method i, where we adopt the convention that p_1 is the ground-truth cluster labels. p_i is a mapping from clusters to labels, i.e. it is of the form $p_i(s_j) = l$ with $l \in \mathcal{L}$. This would mean that mapping p_i labels cluster j as type l. The cluster accuracy is calculated as follows:

$$CAcc = \frac{1}{n_Y} \sum_{j=1}^{n_Y} \mathbb{I}(p_i(s_j) = p_1(s_j)) \qquad (5)$$

where n_Y refers to the number of clusters as defined by the Leiden clustering, s_j is the jth cluster, p_i is the mapping from clusters to predicted cell-type and p_1 is the mapping from clusters to true (ground-truth) cell-type.

To define true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for a specific label $l \in \mathcal{L}$, we use the following definitions:

$$\mathrm{TP}_l = \sum_{j=1}^{n} \mathbb{I}(m_i(c_j) = l \wedge m_1(c_j) = l)$$

$$\mathrm{FP}_l = \sum_{j=1}^{n} \mathbb{I}(m_i(c_j) = l \wedge m_1(c_j) \neq l)$$

$$\mathrm{TN}_l = \sum_{j=1}^{n} \mathbb{I}(m_i(c_j) \neq l \wedge m_1(c_j) \neq l)$$

$$\mathrm{FN}_l = \sum_{j=1}^{n} \mathbb{I}(m_i(c_j) \neq l \wedge m_1(c_j) = l)$$

Using these definitions, we calculate the median Precision, median Recall, median and macro $F1$ scores.

Matthews correlation coefficient (MCC) is a metric for assessing the performance of binary classification models. MCC takes into account true and false positives and negatives and a high MCC score is only achieved when the majority of both negative and positive cases, regardless of their proportions in the dataset are accurately prediced. Therefore, MCC score effectively addresses the class imbalance that affects the accuracy measure.

For the multiclass case, the MCC can be defined as follows (39):

$$\mathrm{MCC} = \frac{c \times s - \sum_k (t_k \times p_k)}{\sqrt{(s^2 - \sum_k p_k^2) \times (s^2 - \sum_k t_k^2)}}$$

where c represents the total number of correct predictions, s is the total number of samples, t_k and p_k represent the total number of samples in the true class k and the predicted class k, respectively. In multiclass scenario, MCC score takes 1 as the maximum value while the minimum value ranges between -1 and 0.

Cohen's kappa (κ) calculates the level of concordance between annotations by taking the agreement occurring by chance into account. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement ratio and p_e is the expected agreement in case of a random label assignments. κ ranges between -1 and 1 with 1 referring to the full agreement between annotators.

Rand index (RI) measures the similarity between two set of clusters by examining the pairs of elements that match and differ within the clusters. It is defined as number of agreeing pairs divided by number of pairs. Adjusted rand index (ARI) *adjusts* this measure for chance with the following formula:

$$\mathrm{ARI} = (\mathrm{RI} - \mathrm{Expected\ \ RI})/(\max(\mathrm{RI}) - \mathrm{Expected\ \ RI})$$

This similarity score ranges between -0.5 and 1, with 1 being the perfect match and random labelling ranging between 0 and 1.

Datasets

Query scATAC-seq datasets

FACS human hematopoiesis (bone marrow) scATAC-seq

The bone marrow scATAC-seq data is acquired from (40). The dataset includes scATAC-seq data of 2210 human hematopoietic progenitor cells which are obtained from bone marrow and isolated via fluorescence-activated cell sorting (FACS). FACS enables sorting of single cells into a well plate using cell-type specific cell-surface markers and thereby provides a true cell-type annotation for each cell in the data.

Owing to this unique property, we use this data both as single-cell data as well as bulk prototype data. When using as a bulk prototype, we combine the single cells of the same cell-type by summing the features across cells and form *bulk-like* data. As the data is by default annotated, here we refer to this dataset as annotated bulk prototype dataset. When using as a single-cell data, as the name suggests, we use the data obtained from cells individually and do not leverage the FACS annotations.

10X human PBMC sc-multiomics

Human peripheral blood mononuclear cell (PBMC) sc-multiomic data has been obtained from 10X website (41). The dataset includes paired scRNA-seq and scATAC-seq profiles of 10 661 cells, obtained from PBMCs of a healthy donor. Inference of the cell-type annotations of this dataset is introduced in the Methods section.

NeurIPS human bone marrow sc-multiomics

This dataset has been provided as part of the NeurIPS challenge (42). The dataset serves as a valuable resource for benchmarking studies and claimed to be the largest and most realistic multimodal data available (43). The datasets are also carefully annotated by experts considering various marker genes and across scRNA-seq and scATAC-seq data. We focused on one sample from this study, namely, *s1d1* for the sake of simplicity. The data consist of 19 039 bone marrow cells obtained from a healthy donor. In this study, we decoupled the paired measurements and exclusively used the scATAC-seq data. For this dataset, we use the annotations provided by the challenge organizers as the ground-truth, therefore, do not perform cell-type annotation for the scRNA-seq part.

Granja human PBMC scATAC-seq

Human PBMC scATAC-seq data of sample D10T1 comes from (44) and was downloaded from GEO. The fragments file was aligned to the hg19 genome version and, using the liftOver utility, was transferred to hg38 genome assembly. The dataset includes 2891 healthy PBMCs. We used the cell-type annotations provided in the original publication.

Granja human BMMC scATAC-seq

Human BMMC scATAC-seq data of sample D6T1 is also acquired from (44). Like PBMC data, we obtained the fragments file and used liftOver utility to align the fragment files to hg38 genome assembly. The dataset includes 12 394 healthy BMMCs. We rely on the cell-type annotations from the original publication.

Corces brain scATAC-seq

Corces brain scATAC-seq data (45) comes from caudate nucleus of a cognitively healthy individual (control 1). The raw data is aligned to GRCh38 genome version using cellranger-atac-2.0.0. We use 10 016 brain cells whose cell-type annotations are provided in the original publication.

Morabito brain cortex scATAC-seq

The Morabito dataset (46) includes scATAC-seq data from the postmortem prefrontal cortex of an individual with no cognitive impairments. We focus only on Sample-90 for the sake of simplicity. The raw data is obtained from GEO database and and aligned to the GRCh38 genome version using cellranger-atac-2.1.0. It comprises of 5933 brain cells, with cell-type annotations derived from the original publication.

Reference datasets

Human PBMC CITE-seq

Human PBMC CITE-seq (36) comprises the transcriptomic measurements of 211 000 PBMCs along with 228 cell-surface proteins. The reference includes two levels of cell-type annotations with increasing granularity, namely, cell-type.l1 and cell-type.l2. We used the coarse annotation, celltype.l1, which includes the main blood cell types, namely; B, CD4 T, CD8 T, natural killer, monocytes, dentritic, other T and other cells.

Human BMMC CITE-seq

The BMMC (bone marrow mononuclear cell) dataset (13) comprises single-cell transciptomics measurements of 33 454 bone marrow cells along with 25 cell-surface proteins. This reference as well, includes two levels of cell-type annotations with increasing granularity, namely, cell-type.l1 and cell-type.l2. The used fine-grained annotation, celltype.l2, which includes progenitor cells and hematopoietic stem cells, along with the PBMCs.

Satpathy human PBMC sc-multiome

This dataset is obtained from (47) and includes the paired scRNA-seq and scATAC-seq measurements of 9616 PBMCs. We only used the scATAC-seq measurements of a PBMC sample, referred to as Rep_1. We obtained the fragments file from GEO which was aligned to hg19 genome version. We used liftOver utility to align the fragments file to hg38 genome assembly. The dataset annotation encompasses a range of blood cell types, including B cells, CD4 T cells, CD8 T cells, natural killer cells, dendritic cells, monocytes, basophils and various subtypes thereof.

Satpathy human CD34+ progenitor sc-multiome

This dataset also comes from (47) and consists of paired scRNA-seq and scATAC-seq measurements. We use only the scATAC-seq measurements of 10 056 CD34+ cells, identified as Rep_1. The processing of the fragments file was conducted in the same manner as with the Satpathy Human PBMC sc-multiome dataset. In comparison to the cell types found in the Satpathy Human PBMC sc-multiome dataset introduced above, the CD34+ progenitor dataset includes annotated progenitor cell-types such as hematopoietic stem cells (HSCs), monocyte-dendritic cell progenitors (MDPs), megakaryocytic-erythroid progenitors (MEPs), lymphoid-primed multipotent progenitors (LMPPs), granulocyte-monocyte progenitors (GMPs), multipotent common myeloid progenitors (CMPs) and common lymphoid progenitors (CLPs).

Satpathy human BMMC sc-multiome

Similarly, this dataset is obtained from (47) and consists of paired scRNA-seq and scATAC-seq measurements. We use only the scATAC-seq measurements of 6011 BMMCs, identified as Rep_1. The processing of the fragments file was conducted in the same manner as with the Satpathy Human PBMC sc-multiome dataset. The BMMC dataset encompasses cell-types from both the PBMC and CD34+ progenitor datasets, offering a comprehensive view of the cellular landscape.

Zhu human cerebral cortex sc-multiome

This dataset is acquired from (48) and consists of the paired expression and accessibility measurements of human cerebral cortex across time. We focus on only one adult brain sample referred to as Adult_1 which consists of 1395 cells with corresponding cell-type labels. These labels includes oligodendrocytes, astrocytes, oligodendrocyte progenitor cells (OPCs), excitatory neurons (ENs), pericytes, inhibitory neurons (INs) and microglia.

Allen Brain Map primary motor cortex snRNA-seq

This dataset (49,50), obtained from the Allen Brain Atlas, comprises single-nucleus transcriptomes of 76 533 nuclei obtained from two post-mortem human brain samples. The dataset explores the cell-type composition of primary cortex and it is commonly referred to as ‘Human M1 10x’. It encompasses a variety of cell-types, including astrocytes, endothelial cells, excitatory neurons, inhibitory neurons, microglia, oligodendrocyte precursor cells, oligodendrocytes and vascular and leptomeningeal cells.

Human hematopoietic differentiation bulk ATAC-seq

The human hematopoietic differentiation bulk ATAC-seq dataset (2) includes bulk ATAC-seq profiles of 13 human primary blood cell-types extending over diverse hematopoiesis layers. Samples are obtained from peripheral blood and bone marrow. In our study, we exploit distinct subsets of this data to match potential cell-types in the given query dataset. As the FACS-profiled scATAC-seq bone marrow data inherently specifies the cell-types, we match the bulk prototypes to those cell-types. When integrating the bulk samples bone marrow data we use all 13 cell-types. Lastly, when integrating with PBMC data, we consider only the terminal cell states (unipotent cells), which constitutes the largest portion of a typical PBMC samples. Notably, this dataset lacks the dendritic cell population. To account for this, we merge the the dataset with the plasmacytoid dendritic cells data from (51). All the data is preprocessed according to ENCODE ATAC-seq analysis pipeline (52).

BOCA2: lineage-specific brain open chromatin atlas

The BOCA2 dataset (53) comprises of bulk chromatin accessibility profiles of four brain cell-types across three brain regions: anterior cingulate cortex, dorsolateral prefrontal cortex, and primary visual cortex. These cell nuclei were isolated using fluorescence-activated nuclei sorting (FANS), ensuring that each ATAC-seq dataset reflects the chromatin accessibility of homogeneous cell populations. The dataset contains a total of 94 samples, including 23 glutamatergic neurons, 22 GABAergic neurons, 24 oligodendrocytes and 25 microglia/astrocytes. Given the consistent profiles of cell types across different regions (53), samples from various brain regions were treated as replicates. The data was processed in accordance with the ENCODE ATAC-seq analysis pipeline standards (52).

A summary of the datasets and in which combination they are used for assessing different methods is outlined in Table 1.

Results

Testing scATAcat on FACS bone marrow scATAC-seq data

To evaluate the performance of scATAcat, we first used FACS-profiled scATAC-seq data as both the prototype reference and the query. Therefore, this part of the study serves as a feasibility study. Essentially, we aimed to test if scATAcat could accurately match individual cells to prototypes when the datasets are identical, before applying it to real-world scenarios. The data is inherently annotated by cell-types due to the cell-surface markers of sorted cells. We leveraged these annotations to assess the efficacy of our scATAcat method. scATAcat requires bulk prototype data to provide cell-type annotation for query scATAC-seq data. We created *bulk-like* prototypes by aggregating the same cell-types from the sorted scATAC-seq. DiffBind requires minimum two replicate per cell-type to provide reasonable statistics regarding differential regions. Therefore, for each cell-type we created two bulk-like prototypes by randomly splitting the cells per cell-type into two. We next identified differentially accessible regions using DiffBind. We considered pairwise combinations of the progenitor cell-types when determining differentially accessible regions.

Figure 1A shows the clustering of the single cells into 11 clusters. From this clustering we compute 11 pseudobulk samples. The clustering is not in full agreement with annotation, as can be seen in Figure 1B. For example, hematopoietic stem cells (HSCs) and multipotent progenitors (MPPs) do not show a clear separation. Based on the annotation (Figure 1B), we computed bulk-like prototypes. Next, we integrate the pseudobulks from the clustering with the prototypes in one PCA space. Figure 1C represents the 3D PCA projection obtained through scATAcat. The visualization clearly illustrates a distinct separation among prototypes. The pseudobulk clusters closely align with these prototypes.

While the 3D projection provides an intuition about the relationships between prototypes and pseudobulks, the Euclidean distances from high dimensions provide more accurate proximity information. Therefore, we determined the Euclidean distances between the prototypes and pseudobulk samples in high dimensional PCA space consisting of 30 principal components (PCs). A heatmap representation of these distances is depicted in Figure 1D. For the purpose of assignment, each pseudobulk gets assigned to its closest prototype. Accordingly, clusters 0, 3 and 4 correspond to HSC; Cluster 1 to granulocyte-monocyte progenitor (GMP); clusters 2, 5 and 6 to common myeloid progenitor (CMP); cluster 7 to lymphoid-primed multipotent progenitor (LMPP); cluster 8 to common lymphoid progenitor (CLP) and clusters 9 and 10 to megakaryocytic-erythroid progenitor (MEP).

We use the metrics introduced in the Methods section to evaluate scATAcat’s performance, with results presented in Figure 2L. In this feasibility study, scATAcat demonstrated promising results across various metrics, particularly in cluster accuracy with a perfect score of 1.0, and high median precision and recall scores of 0.87 and 0.91, respectively. Notably, the ARI score is lower than the other evaluation metrics indicating that the clustering of predicted labels does not fully align with those of ground-truth.

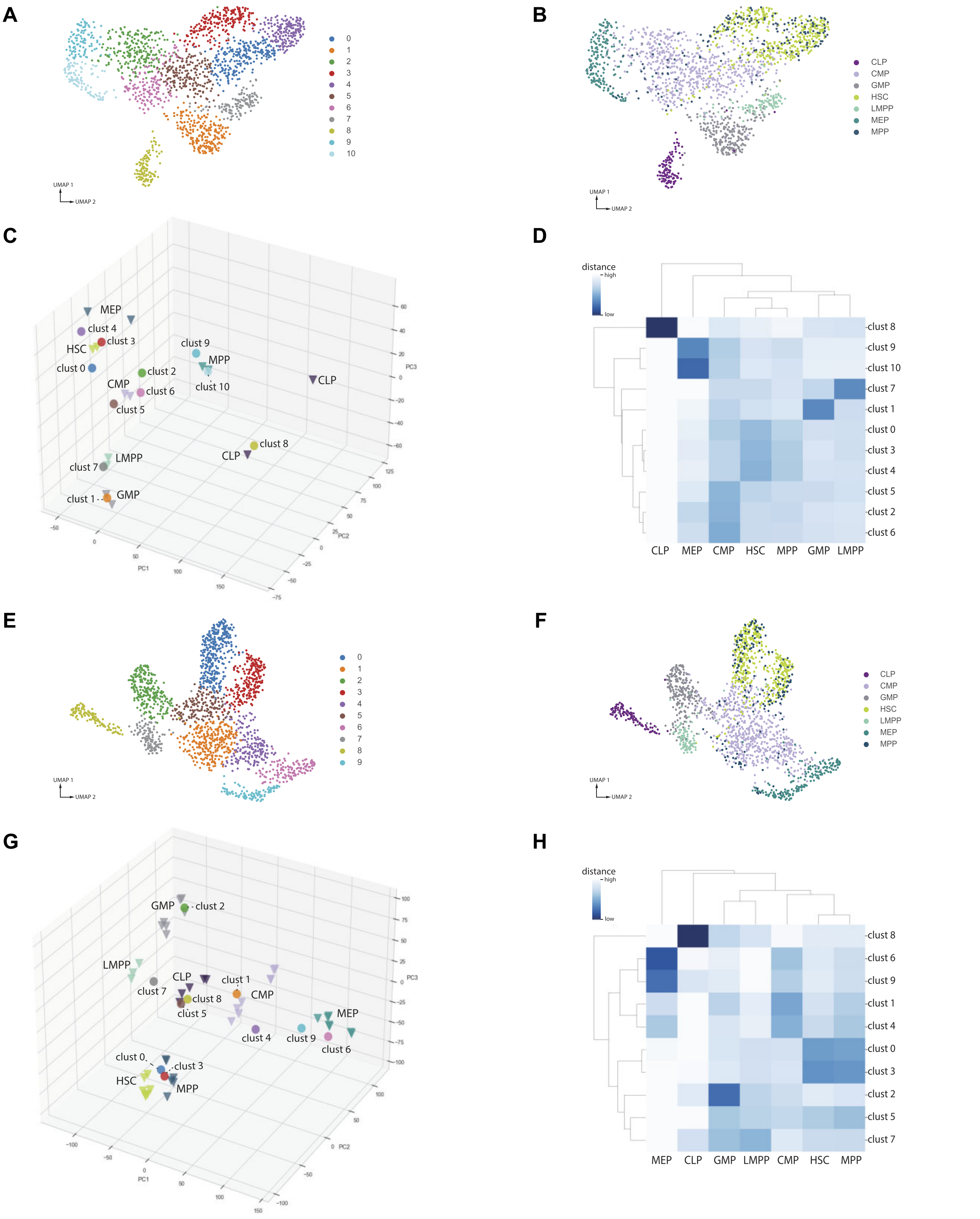


Figure 1. (**A**–**D**) Testing scATAcat on FACS bone marrow scATAC-seq data as part of the feasibility study. (**A**) UMAP embedding colored by the clustering of FACS scATAC-seq cells, and (**B**) by true (ground-truth) cell identities derived from FACS labels of common lymphoid progenitor (CLP), common myeloid progenitor (CMP), granulocyte-monocyte progenitor (GMP), hematopoietic stem cell (HSC), lymphoid-primed multipotent progenitor (LMPP), megakaryocytic-erythroid progenitor (MEP) and multipotent progenitor (MPP). (**C**) 3D PCA projection of the feasibility study. Prototypes, depicted as triangles, are created by aggregating single cells from (B) based on cell-type, with two prototypes generated for each cell-type. Circles represent pseudobulks, the aggregated forms of clusters from (A). (**D**) Heatmap of the high-dimensional Euclidean distances between prototypes and pseudobulks. (**E**–**H**) Application of scATAcat in annotating FACS bone marrow scATAC-seq data. (**E**) UMAP embedding colored by the clustering of FACS scATAC-seq cells, and (**F**) by true (ground-truth) cell identities derived from FACS labels. (**G**) 3D PCA projection of the pseudobulks together with prototype bulk samples. Triangles represent prototypes while circles represent pseudobulks. (**H**) Heatmap of the high-dimensional Euclidean distances between prototypes and pseudobulks.

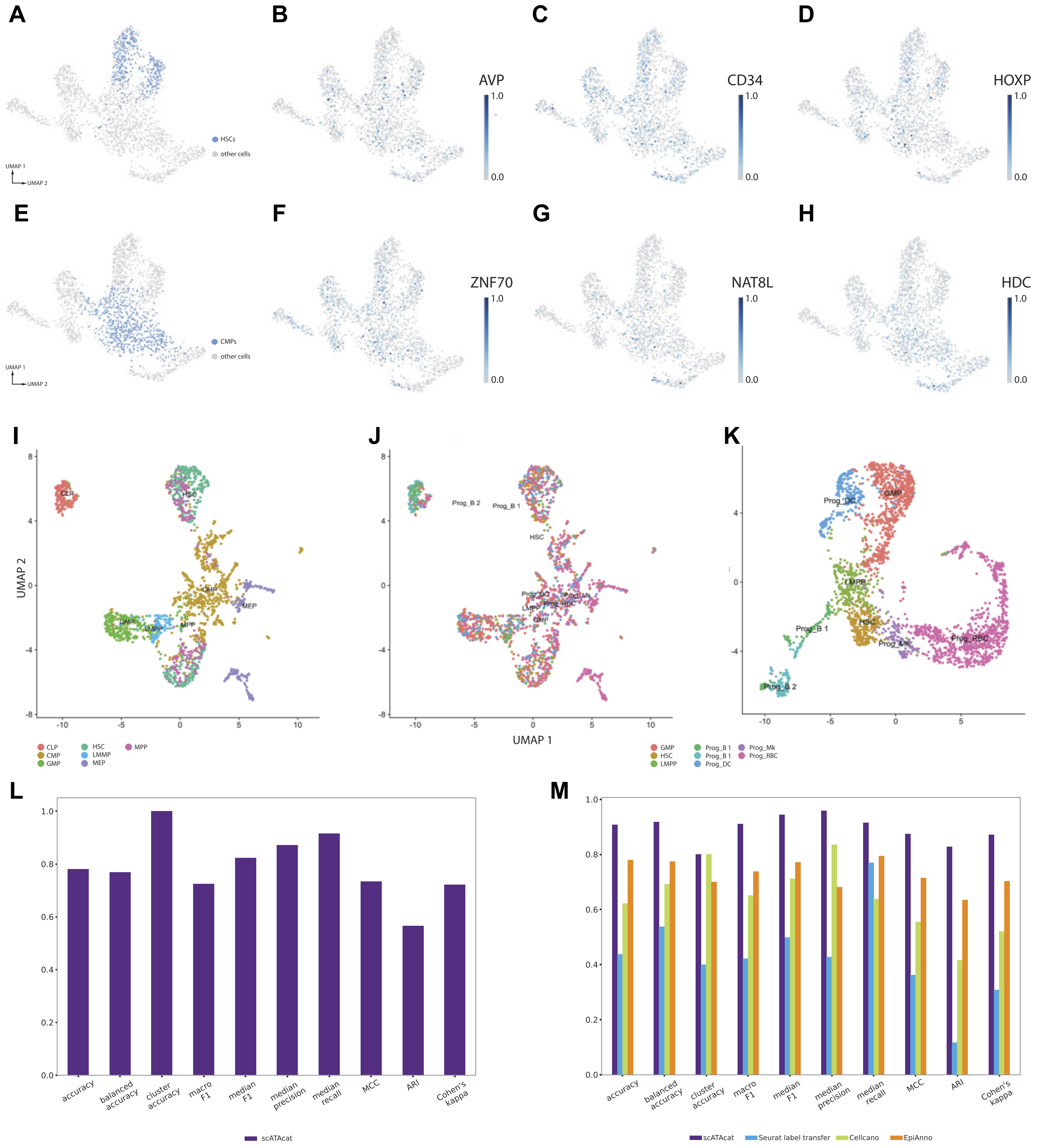


Figure 2. Performance evaluation of methods in annotating FACS bone marrow scATAC-seq data. (**A**–**H**) Marker-based annotation of hematopoietic stem cells (HSCs) and common myeloid progenitors (CMPs). (**A**) depicts UMAP embedding of FACS-characterized bone marrow scATAC-seq data. The blue colored cells represents ground-truth HSCs. The figures (**B–D**) show the predicted expression levels of HSC marker genes, AVP, CD34 and HOXP, with darker shades indicating higher expression. (**E**) depicts UMAP embedding of FACS bone marrow scATAC-seq data. The blue colored cells represents ground-truth common myeloid progenitors (CMPs). The figures (**F–H**) depict the predicted expression levels of CMP marker genes, ZNF70, NAT8L, HDC, with darker shades indicating higher expression. (**I**–**K**) Cell-type annotation via label-transfer. (**I**) UMAP representation of predicted gene expression matrix colored by ground-truth cell-types. (**J**) shows the same UMAP embedding however, this time the color-code indicates the predicted cell-type identity by the label-transfer approach as shown by the legend. (**K**) scRNA-seq UMAP embedding of the reference bone marrow data, used as a reference when applying label-transfer approach, colored by cell-type identities. (**L**, **M**) Evaluations of cell-type annotations performances for FACS bone marow scATAC-seq data. (**L**) shows the performance of scATACat on feasibility study. (**M**) Performance comparison of all the methods across various evaluation metrics. MCC, Mathew's correlation coefficient; ARI, adjusted rand index.

The matrix of Euclidean distances between pseudobulks and prototypes already shows that some decisions are almost a tie. An alternative visualization is provided by clustering all the distances among pseudobulks and prototypes. Such a heatmap with the associated single-linkage dendrogram is shown in Supplementary Figure S1B. In this figure, e.g., the group of clusters 0, 4 and the HSC cells form a cluster in the dendrogram.

Performance of scATAcat in annotating FACS bone marrow scATAC-seq data

After demonstrating the effectiveness of our method in the feasibility study, we applied scATAcat again to the same FACS scATAC-seq data. However, this time, we integrated external prototypes from bulk ATAC-seq data (2). These prototypes align with the cell-types present in the scATAC-seq data. We define differentially accessible regions using the prototypes and use these features to represent the data. Therefore, the data representations differ despite using the same scATAC-seq data.

Figure 1E shows the clustering of cells in a UMAP embedding, revealing 10 clusters. We form pseudobulks by aggregating the cells in each cluster. Figure 1F depicts the real identities of the cells. Clusters show remarkable consistency with annotations. Nevertheless, HSCs and MPPs do not exhibit a clear separation. Our next step was to co-embed pseudobulks with prototypes. Figure 1G depicts the 3D PCA projection produced by scATAcat. The figure contains colored triangles for the prototype replicates and circles for the pseudobulks. Mostly there is a clear association between the correct prototypes and pseudobulks, although there is larger variability among the CMP and GMP prototypes. HSCs and MPPs are less clearly distinguishable but mix with each other, as can also be seen in the UMAP.

Subsequently, we compute the Euclidean distances between the prototypes and pseudobulks in high dimension to compute more informative cell-type annotation. Figure 1H shows the heatmap of these distances. Using this distance matrix we annotate each pseudobulk by the closest prototype cell-type. In this way we can correctly associate clusters 0 and 3 to HSC, clusters 1 and 4 to CMP, cluster 2 to GMP, cluster 5 to MPP, clusters 6 and 9 to MEP, cluster 7 to LMPP and cluster 8 CLP. This shows that using the high-dimensional Euclidean distances scATAcat again provides accurate annotations for each cluster.

We next applied marker-based annotation to the same FACS scATAC-seq data and aimed at annotating cell-types by the *predicted* expression profiles of the marker genes in cells. Marker-based annotation requires collection of cell-type specific marker genes which in itself may be a nontrivial task. We obtain marker genes from a curated data table from (54) (Supplementary Table S3). As these genes are by definition cell-type specific, we expect also their predicted expression to be specific for the same cell-type. Figure 2 shows the UMAP embedding of the cells with the color code depicting ground-truth cell identities for HSCs and CMPs (A and E, respectively). However, as shown in Figure 2B–D for HSCs (respectively Figure 2F–H for CMPs), the marker genes do not show the distinctive predicted expression profiles. In fact, in both cases marker genes are expressed in a dispersed set of cells over the UMAP. Some marker genes show ubiquitous predicted expression (Figure 2C) while others show unspecific characteristics. We conclude that transcriptome based marker genes cannot be directly carried over to predicted gene expression space to annotate cell-types in scATAC-seq data.

Following this, we employed Seurat's label-transfer approach on the same FACS-characterized bone marrow scATAC-seq data for cell-type annotation. Label-transfer essentially uses Canonical Correlation Analysis to transfer cell-type labels from a well-annotated reference dataset to a query dataset. In order to use this approach we rely on the BMMC CITE-seq data (see Datasets) as a reference and predicted gene expression values of FACS bone marrow scATAC-seq data as the query. Since the CITE-seq data includes measurements of cell-surface proteins, the cells in the reference dataset have reliable and independent annotations. The BMMC CITE-seq data comprises a range of blood cell types including both terminal and progenitor cells. As our FACS bone marrow scATAC-seq data consist of only the progenitor cells, we subset the CITE-seq data accordingly. Figure 2K shows the UMAP embedding of the scRNA-seq part of reference CITE-seq data which serves as the source for transferred cell-type labels. Figure 2I shows the UMAP representation of the query scATAC-seq data and the color code depicts the real cell-type identities. Some of the cell-types form a distinct cluster, e.g. CLPs, GMPs. However, some cell-types display inconsistency between the clustering and their real cell-type identities. Notably, HSCs and MPPs cannot be distinguished. Additionally, HSC and MPPs, as well as MEPs are split into two clusters. Finally, Figure 2J shows the annotated FACS bone marrow scATAC-seq data where the color code denotes cell-type annotations obtained with the label-transfer approach. These annotations do not show uniformity within clusters and annotations do not overlap with the original cell-type identities.

Next, we applied Cellcano for cell-type annotation of the FACS-characterized bone marrow scATAC-seq dataset. We trained the Cellcano model with a gene score matrix obtained from the Satpathy human CD34+ progenitor scATAC-seq dataset using ArchR default parameters, incorporating cell-type annotations as reported in the original study. Subsequently, we generated a gene score matrix for our FACS-characterized bone marrow scATAC-seq query dataset and used the trained model to assign cell types to the query data.

The last cell-type annotation method to compare is EpiAnno. EpiAnno uses peaks as the reference frame which we define using ArchR. Given that peak regions can vary between datasets, a preprocessing step is necessary to align features for supervised learning. Following the authors' guidelines, we adapt the reference data to match the peak regions of the query dataset, creating a unified dataset ready for learning. This implies that, for the reference data, we count the reads falling into the peak regions of the query data. After these preprocessing steps, the datasets are ready for supervised learning.

Just like in Cellcano, we use Satpathy human CD34+ progentior scATAC-seq dataset as the reference scATAC-seq data and unify it with the ArchR-defined peaks of our FACS-characterized bone marrow scATAC-seq dataset. We then train EpiAnno's Bayesian neural network algorithm with this unified dataset and its cell-type annotations. The method failed when using the default 'peak_rate' parameter, due to the high memory demands from the large amount of peaks. By modifying this parameter to 0.05, we were able to successfully execute EpiAnno. Finally, we apply the model to predict cell-types in the query dataset.

Performance assessment

In order to quantitatively assess the performances of the above mentioned methods, we compare the annotations obtained from each approach to the real cell-type identities obtained from FACS labels. The annotations in reference data sets cover a broader spectrum of bone marrow cells than those of query FACS labels. Additionally, the annotations of reference and query datasets disagree on parts of the nomenclature, even though they may correspond to similar cell types. To increase the concordance between annotations, we modified the annotations to a coarser annotation scheme whenever possible, for example, we used 'HSC/MPP' instead of using 'HSC' and 'MPP' separately, following the annotation style of the Satpathy dataset. Similarly, nomenclature is aligned between the annotations, for example, we used 'CLP' instead of 'B cell progenitor (Prog_B)' in CITE-seq reference data and 'B cell progenitor (Pro_B)' in Satpathy scATAC-seq reference. Original annotations along with their corresponding simplified annotations are provided in Supplementary Table S2. Different preprocessing and filtering parameters result in distinct sets of annotated cells. We thus focus on those cells that have an annotation across all methods, as well as a ground-truth label.

We follow the two strategies for performance evaluation presented in Methods subsection 'Performance Assessment'. In the conservative approach we determine the set of common annotations across methods and the ground-truth. The common cell-type annotations between CITE-seq data, Satpathy scATAC-seq reference, and FACS labels are MEP, GMP, HSC/MPP, CLP and LMPP. Therefore we focus on these common annotations. Evaluation of FACS bone marrow dataset across these metrics is summarized in Figure 2M. Except for cluster accuracy where three methods have the same score, scATAcat consistently shows the best performance across all the metrics, followed by EpiAnno and Cellcano. The label-transfer approach performs poorly except when assessed by median recall with a score of 0.77. Despite relatively high accuracy scores, the methods other than scATAcat show lower ARI, indicating that although most of the data points are labeled correctly, the overall clustering structure does not align well with the true label structure. Additionally, if one cell-type dominates the annotations, a method might achieve high accuracy by predominantly predicting the majority cell-type, but the ARI could be low if the method fails to correctly identify the structure of smaller cell-type classes. Similarly, for the methods except scATAcat, a similar low score trend is observed for Cohen's kappa. This suggests that these methods may not be as effective in datasets with imbalanced class distributions.

In the second strategy, we define the common cell-types in a pairwise manner by comparing the cell-type annotations of each method independently with ground-truth labels. The resulting performance scores are shown in Supplementary Figure S1D. In this second strategy as well, scATAcat maintained its leading position among the methods. Except for median recall, where label-transfer secured the second position, EpiAnno was next, followed by Cellcano and then label-transfer across the metrics. In comparison to the first strategy, the scores for label-transfer remained unchanged in this second strategy, while most of the scores for other methods saw a slight decrease.

Since the granularity of Leiden clustering results depends on the resolution parameter, we tested the impact of this parameter on the performance of scATAcat. We applied scATAcat with Leiden clustering resolutions varying from 0.1 to 2.5 increasing by increments of 0.1. This was done on the same FACS bone marrow scATAC-seq data and the prototypes introduced above for scATAcat. We compared the resulting annotations with ground-truth cell identities obtained from FACS labels. Resulting performance metrics are presented in Supplementary Figure S1E and Supplementary Table S9. With increasing resolution parameter, more cell-types are being annotated and evaluated resulting varying performances. The highest overall performance is observed at the lowest tested resolution (0.1) and higher resolutions generally lead to a decrease in most performance metrics, such as accuracy, balanced accuracy, $F1$, median precision, and median recall score. This suggests that as the clustering becomes finer (more clusters), the method's ability to accurately classify and recall the data points decreases slightly.

We also tested the effect of number of features in reference frame on the performance of scATAcat. The results are presented in Supplementary Figure S12B and Supplementary Table S16, and detailed in Supplementary Material Section 1. 1.

Additionally, to assess the effectiveness of scATAcat methodology, we investigated various co-embedding strategies within the scATAcat approach and different embedding algorithms. We then performed a comparative analysis relative to the scATAcat framework which are presented in Supplementary Material Section 1.2 and Supplementary Figure S13.

Performance of scATAcat in annotating 10X PBMC data

Single-cell transcriptome technologies have been frequently demonstrated and used on PBMCs and thus provided deep understanding of blood cell-types. For purposes of this study it is particularly helpful that a sc-multiome PBMC dataset is available combining nuclear scATAC-seq and scRNA-seq of the same cells (see Datasets).

Since the PBMC sc-multiome data has not been externally annotated, we are leveraging the scRNA-seq part of the data to create a surrogate ground-truth. Relying on the assumption that cell-type annotation in scRNA-seq data is more straightforward (55) than for scATAC-seq, we can evaluate our scATAcat annotation with respect to this ground-truth.

To obtain a ground-truth cell-type annotation, scRNA-seq data gets processed using Seurat's standard pipeline (36). Subsequently, cell-type annotation is performed via the label-transfer approach using PBMC CITE-seq data as the reference and applying *SCTransform* for both the reference CITE-seq and query scRNA-seq data. Anchors are determined using *spca* as the reduction method and cell-type identities are extracted via the *MapQuery*. The cell-type annotations of the scRNA-seq data is shown in Supplementary Figure S2A.

Having a ground-truth annotation of the cells available, we proceeded to annotate the scATAC-seq part of the multiome data using scATAcat, via the marker genes, by label-transfer, Cellcano and EpiAnno. To apply scATAcat, scATAC-seq data is processed as introduced earlier. For the prototypes we again use the sorted hematopoietic bulk ATAC-seq data (see Datasets), albeit now focusing only on those terminal cell states (CD4 T cells, CD8 T cells, B cells, NK cells, monocytes, dendritic cells) that are also part of PBMCs (56). The differentially accessible regions between the bulk prototypes are identified and used as the reference frame for clustering. A UMAP representation of the clustering of the cells is shown in Figure 3A. We formed 15 pseudobulks out of these clusters and applied scATAcat to integrate prototypes and pseudobulks.

Figure 3B depicts the 3D PCA projection obtained by scATAcat. The bulk prototypes roughly cluster into three groups, akin to the UMAP plot. Pseudobulks closely align and co-embed with bulk samples except for cluster 11. The replicates of CD8 T cell and dendritic cells are more scattered in 3D space indicating a higher variability among them. Clusters 1 and 4 project in between CD4 and CD8 T cells. It is important to highlight that the 3D PCA plot, which is based on the variation within the first three PCs, might not capture the complex similarity relationships within the data. Therefore, we calculate the distances between between bulk prototypes and pseudobulks considering additional (higher) dimensions. The heatmap showing these distances is presented in Figure 3C. The heatmap depicts a clear clustering structure for most of the clusters resulting in clear annotations. However, the high similarity between CD4 and CD8 T cells (also visible in Supplementary Figure S2C) makes it harder to annotate the clusters surrounding them. Nevertheless, cluster 4, which presents the most ambiguous signal in the 3D PCA plot, aligns more closely with CD4 T cells than with CD8 T cells. The same is true for cluster 1, although it is less pronounced. Consequently, we annotated each pseudobulk by the closest bulk sample, hence, clusters 0, 3, 8, 11 and 12 correspond to monocyte; clusters 1, 2, 4 and 6 to CD4 T cell; clusters 5 and 13 to CD8 T cell; cluster 7 and 10 to B cell; cluster 9 to natural killer (NK) cell and cluster 14 to plasmacytoid dendritic cells (pDC).

Although cluster 11 is annotated as a monocyte, it shows similarity with other prototypes, too. This observation could be attributed to a biological characteristic, such as the presence of highly similar cell types in the scATAC-seq data. In such intricate cases, it is helpful to further investigate the data. An alternative visualization provided as part of scATAcat, depicting the clustering of pairwise Euclidean distances between pseudobulks and prototypes, can be helpful in interpreting such cases. Supplementary Figure S2C shows the heatmap of these distances. Cluster 11 shows affinity to both the monocyte prototype together with other monocyte pseudobulks (clusters 0, 3, 8, 12) as well as to the CD4 T-cells with clusters 1, 2, 4 and 6. Possible explanations for this phenomenon include heterogeneity in cluster 11 or possible doublet cells. The annotation process will respect the closest distances in high dimensions and annotate cluster 11 as monocytes. Yet, the example shows that visual inspection of the high-dimensional distances can indicate potential biological or technical problems.

To put these results in context, we generated the predicted expression levels of PBMC scATAC-seq cells and carried out marker-based annotation. Predicted expression levels of a well-established B cell marker gene, PAX5 (57), is depicted in Figure 3D. Clusters 7 and 10 show highest expression of PAX5 suggesting these clusters as potential B cells. Notably, we also observe a low-level but ubiquitous expression of PAX5 across other cell types. Supplementary Figure S2D-N illustrates more examples of the marker-based annotation which further demonstrates the non-quantitative property of this annotation approach.

Next, we employed Seurat's label-transfer approach to the scATAC-seq part of the multiome data. We used the same procedure as in the label-transfer for the FACS bone marrow scATAC-seq data, although now based on PBMC CITE-seq data as reference. As a query we use the predicted expression values of PBMC scATAC-seq data. Figure 4E shows the predicted annotations of the query cells and Supplementary Figure S2B represents the scRNA-seq UMAP embedding of the reference CITE-seq data (Supplementary Figure S2). Annotations shows high concordance with clusters, although dendritic cells appear close to monocytes indicating a lack of separation. Similarly, CD4 T cells cluster closely with CD8 T cells. This pattern is further reflected in cell-type annotation - some of the cells forming a cluster with CD8 T cells are annotated as CD4 T cells or NK cells. Likewise, certain cells within the CD4 T cell cluster are annotated as monocytes.

Subsequently, we annotate the cell-types in PBMC data via Cellcano, using Satpathy PBMC scATAC-seq data as the reference data. Just like in the FACS bone marrow dataset, we generate gene score matrices from both the query and reference scATAC-seq datasets, which serve as inputs for Cellcano. Cellcano is then trained with the reference data and the cell-type annotations for the query data were derived by applying this trained model.

As a final annotation strategy, we employed EpiAnno. For this method as well, the Satpathy PBMC scATAC-seq data was selected as the reference. We perform peak calling for both the query and reference scATAC-seq data using ArchR. These peaks are then unified as introduced above to generate the common feature space. The unified reference peaks and the cell-type labels allow for training of EpiAnno. Notably, with the default 'peak_rate' paramater of 0.03, the method failed due to excessive memory requirements of the high number of peaks. We adjusted this parameter to 0.05 to successfully run EpiAnno. We use the trained model to predict the cell-types of the query data.

Performance assessment

Lastly, we compared the annotations obtained from scATAcat, label-transfer, Cellcano and EpiAnno. We excluded the marker-based annotation from this comparison because choice of marker genes is subjective and making clear decisions about the annotation is difficult. We again adapt a coarser cell-type annotation scheme to accommodate more cells, for example, use use 'CD8 T cell' instead of 'CD8 effector memory cells' or 'CD8 central memory cells'. Supplementary Table S3 includes both the original annotations and their simplified versions. For the first performance evaluation strategy, we define the common cell-type annotations across methods and the ground-truth annotations and consider only those cells with the cell-type corresponding to these common cell-types in their ground-truth label. We then calculate the performance metrics based on the ground-truth derived from the RNA part of the multiome data. The results of these metrics are presented in Figure 3F. In this dataset, label-transfer approach demonstrated superior performance, especially in accuracy, balanced accuracy and cluster accuracy, among other metrics. scATAcat and Cellcano were competitive, with scATAcat high scores in median precision and recall scores. Cellcano showed a notable performance for balanced accuracy and ARI score highlighting its ability to handle imbalanced datasets. EpiAnno showed the lowest performance across metrics.

The results of the second evaluation strategy, where we determine the common cell-types by comparing the methods' annotations to ground-truth annotations individually, are pre-

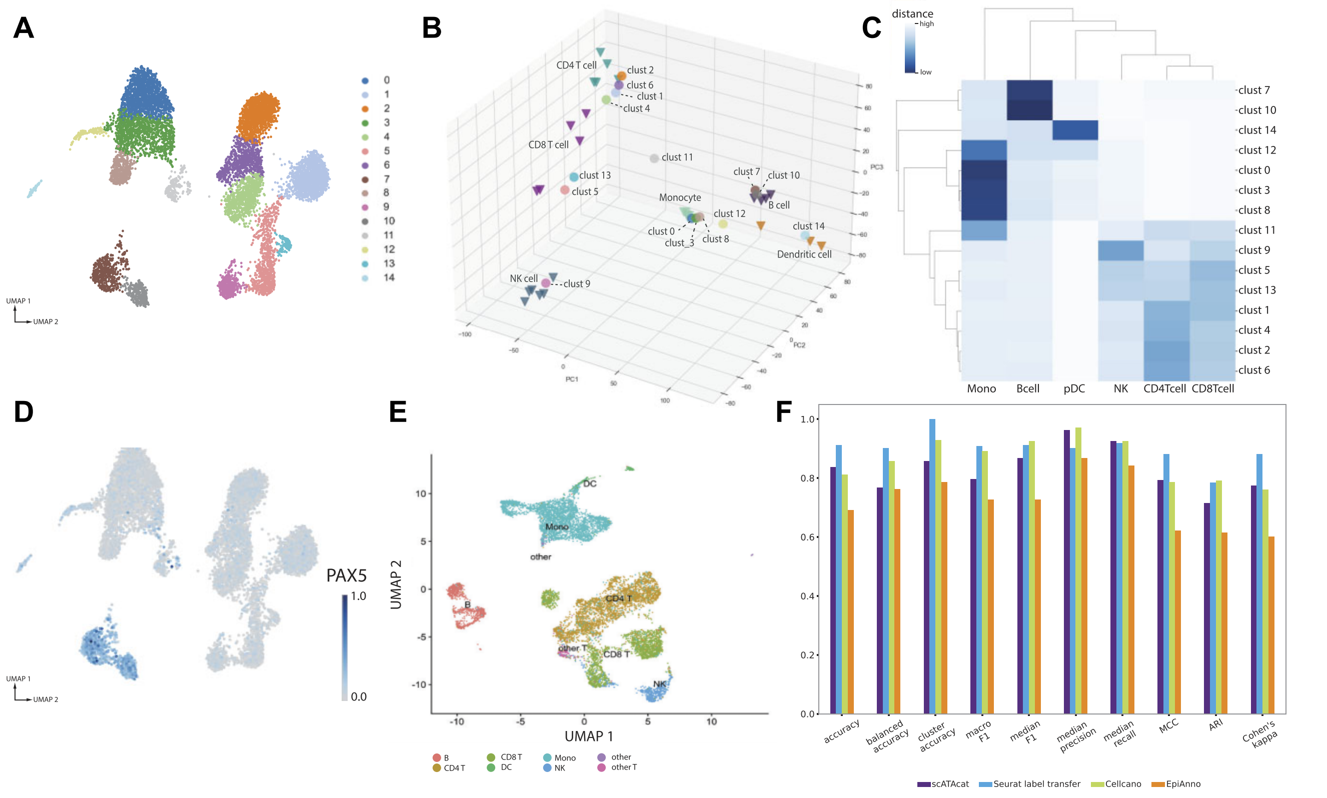


Figure 3. Performance evaluation of methods in annotating 10X PBMC data. (**A**) UMAP embedding colored by the clustering of PBMC scATAC-seq cells. (**B**) 3D PCA projection of the PBMC scATAC-seq pseudobulks together with prototypes. Triangles represent prototypes while circles represent pseudobulks. (**C**) Heatmap of the high-dimensional (50 PCs) Euclidean distances between the pseudobulks and prototypes shown in (B). Monocyte (Mono), B cell (Bcell), plasmacytoid dendritic cell (pDC), natural killer cell (NK), CD4 T cell (CD4Tcell), CD8 T cell (CD8Tcell). (**D**) UMAP embedding PBMC scATAC-seq cells colored by predicted marker gene expression of B cell marker PAX5. (**E**) Cell-type annotation via label-transfer. UMAP representation of predicted gene expression matrix colored by predicted cell-type labels based on label-transfer. (**F**) Evaluation of cell-type annotations performances for PBMC scATAC-seq data across methods.

sented in Supplementary Figure S3A. Despite a slight reduction in its scores, label-transfer remains the top performer across the accuracy metrics, macro F1, MCC and Cohen's kappa scores. scATAcat performs similarly to before now showing better results than Cellcano in terms of all the accuracy metrics and macro *F*1 score. EpiAnno continues to show lower performance across the metrics.

Finally, we assessed the impact which different Leiden clustering parameters have on the performance of scATAcat. The results across varying clustering parameters (0.1–2.5) are shown in Supplementary Figure S3B and Supplementary Table S10. For this dataset our method showed stable performance with respect to changes in clustering parameters, with only minor changes in most metrics. Cluster accuracy, however, improved with increasing clustering parameter.

Performance of scATAcat in annotating NeurIPS bone marrow sc-multiome data

Next, we applied the five annotation strategies to NeurIPS bone marrow sc-multiome data (see Section Datasets). Like in PBMC sc-multiome data, we focus on the scATAC-seq part of the data and simply ignore the scRNA-seq part. This dataset has been provided as part of a benchmarking competition and it comes with expert cell-type annotation.

We first applied our scATAcat method for cell-type annotation. As prototype, we used both the progenitor and terminal cell-types in the bulk data and determined differentially accessible regions between pairs of cell-types. Up-to 2000 most differential regions per comparison get combined to obtain the final feature set. Then we preprocess the scATAC-seq data and perform clustering.

A UMAP of the clustered cells is shown in Figure 4A. We form 13 pseudobulks out of these clusters and integrate them with prototypes using scATAcat. The result of this projection is depicted in Figure 4B. Most of the pseudobulk-clusters project closer to terminal cell-types than to progenitors. Besides, as expected by their high prevalence in bone marrow (58), many clusters project closer to erythrocytes. Most of the lymphoid cells (CD8 T cell, CD4 T cells and NK cells) show rather indistinctive positioning in the 3D projection.

Figure 4C shows a heatmap of the high-dimensional Euclidean distances between prototypes and pseudobulks. This representation of the distances shows a similar profile to 3D projection. The progenitors form a separate cluster and the lymphoid cells, while not well separated in the 3D PCA, form a cluster in the heatmap. Pseudobulks 1, 2, 3, 5 and 6 are nearby the prototypes for CD4 T cells, CD8 T cells NK cells in the 3D PCA and are clearly clustered together with them in the heatmap. In fact, this is expected considering the lineage tree shown in Supplementary Figure S1A. We again annotate each pseudobulk by considering its closest prototype cell-type in terms of high-dimensional Euclidean distance. Cluster 0 corresponds to monocyte, clusters 1 and 5 to CD4 T cell; cluster 2 to NK cell; clusters 3 and 6 to CD8 T cells; clusters 4 and

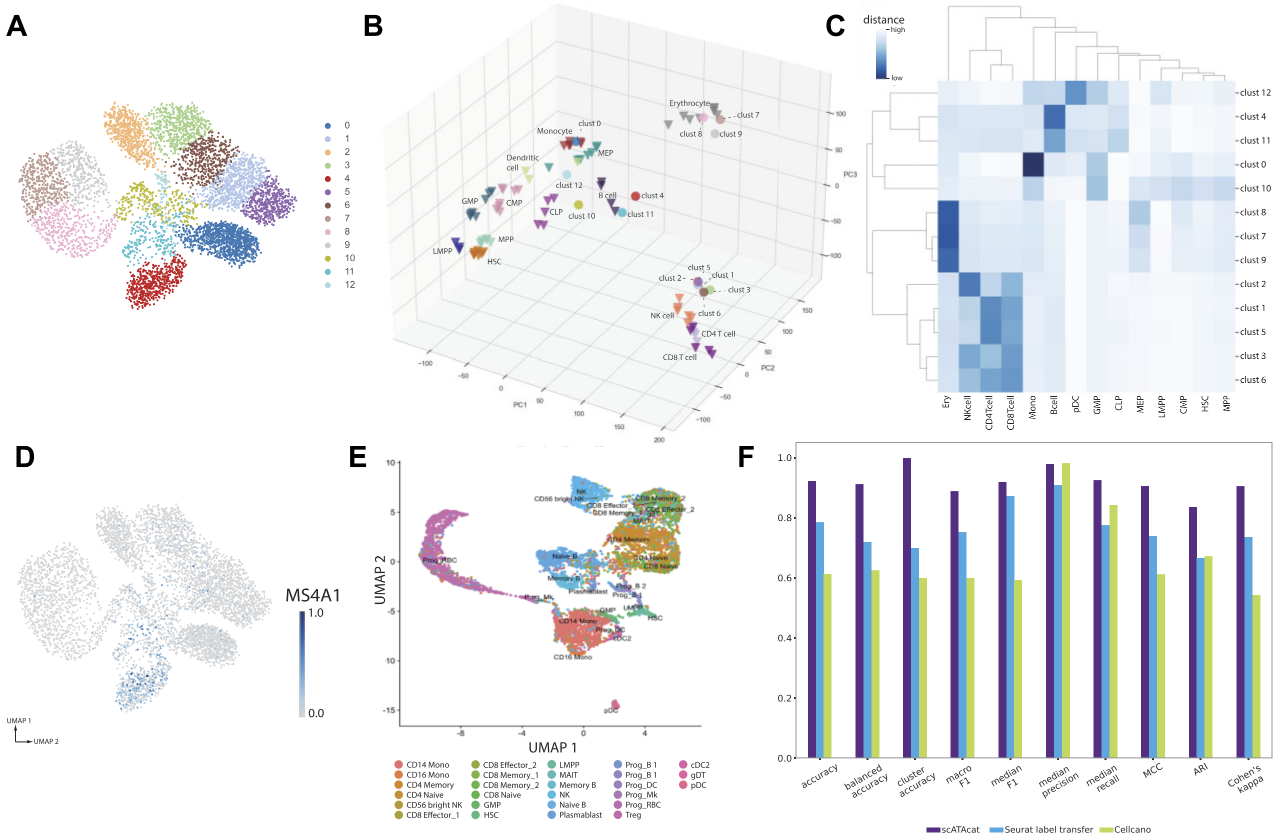


Figure 4. Performance of scATACcat in annotating NeurIPS bone marrow data. (**A**) UMAP embedding of scATAC-seq cells of bone marrow sc-multiome data colored by the clustering. (**B**) 3D PCA projection of the bone marrow scATAC-seq pseudobulks together with prototypes. Triangles represents bulk samples while circles represent pseudobulk-clusters. (**C**) Heatmap of the high-dimensional Euclidean distances between the pseudobulks and prototypes shown in (B). Monocyte (Mono), B cell (Bcell), plasmacytoid dendritic cell (pDC), natural killer cell (NK), CD4 T cell (CD4Tcell), CD8 T cell (CD8Tcell), erythrocyte (Ery), common lymphoid progenitor (CLP), common myeloid progenitor (CMP), granulocyte-monocyte progenitor (GMP), hematopoietic stem cell (HSC), lymphoid-primed multipotent progenitor (LMPP), megakaryocytic-erythroid progenitor (MEP) and multipotent progenitor (MPP). (**D**) Cell-type annotation via marker-genes. UMAP embedding of scATAC-seq cells colored by predicted marker gene expression of B cell marker MS4A1. (**E**) Cell type annotation via label-transfer. UMAP representation of predicted gene expression matrix colored by predicted cell-type labels. (**F**) Evaluation of cell-type annotations performances for NeurIP bone marrow data across methods.

11 to B cell; clusters 7, 8 and 9 to Erythrocyte; and finally, cluster 10 to GMP.

Next, we generated the predicted expression levels of the scATAC-seq profiles and employed marker-based annotation. We use the same marker genes here as we used in annotating FACS-profiled bone marrow scATAC-seq data (Figure 2). While progenitor cell-type marker genes only provided a fuzzy picture of cell identity, terminal cell-types display better predictive performance for the sc-multiome bone marrow data studied now (Supplementary Figure S4). For example, B cell marker gene MS4A1 clearly highlights the B cell clusters (4 and 11) in the UMAP. Nevertheless, marker-based annotation remains the most subjective approach.

Subsequently, we carried out label-transfer approach. We followed the same procedure as employed in label-transfer for the FACS bone marrow scATAC-seq data. This time we used the complete BMMC CITE-seq dataset without subsetting to the progenitor cell-types as the reference, and predicted expression levels of sc-multiome (scATAC-seq) data as the query. Figure 4E shows the predicted labels of the query cells in the UMAP embedding. Generally, clusters are predominantly annotated by one cell type, although the annotations get mixed near cluster boundaries. Also, subsets of CD4 and CD8 T cell co-embed without a clear separation.

For cell-type annotation with Cellcano we use the Satpathy human BMMC sc-multime data as the reference. The gene score matrices of the query and the reference datasets are generated via ArchR. Using the reference dataset and its cell-type annotations we train Cellcano and use the resulting trained model to predict the cell-types in the query dataset.

As the last cell-type annotation method, we use EpiAnno with the same Satpathy human BMMC sc-multime data as the reference dataset. We follow the same approach conducted for other datasets, identifying peaks in both the reference and query datasets. EpiAnno gets trained using the peaks from the Satpathy human BMMC dataset. By default, EpiAnno trains the network for 50,000 epochs. For this dataset, the training process's loss was reported to be 0 after the 20 000th epoch.

Performance assessment

Next, we compare the cell-type annotations obtained across methods to the provided ground-truth annotations. All the annotations include similar cell-types at varying granularity. To increase the overlap between different annotation schemes and provide more inclusive comparison we simplified the annotations to a coarser annotation scheme, if possible. Original cell-type annotations along with their corresponding simplified annotations are provided in Supplementary Table S4. In the first evaluation strategy, we identify common cell types by considering both the predicted annotations from the methods and the ground-truth. Due to EpiAnno's identifying only three distinct cell types, this common cell-type set is reduced to a single cell-type. Since some performance metrics require at least two groups of cell types for calculation, we excluded EpiAnno from the evaluation of this dataset. The performance metrics of other methods are presented in Figure 4F. scATAcat achieves the highest scores in almost all metrics, in particular showing a perfect score in cluster accuracy. Label-transfer approach demonstrates a good performance across all the metrics except the median precision score where Cellcano performs equally well with scATAcat and median recall score where Cellcano outperforms label-transfer approach.

The results of second evaluation strategy are shown in Supplementary Figure S4N. scATAcat shows high scores across metrics. Label-transfer and Cellcano show moderate performance, with label-transfer excelling in median F1 score and Cellcano in median recall score. This evaluation includes EpiAnno as well, as the common cell-types are defined in a pairwise manner, but it does not perform well.

We examined the potential impact of the clustering parameter choice on performance metrics. As presented in Supplementary Figure S4O and Supplementary Table S11, most of the metrics show rather robust performance across varying clustering parameters while median recall and macro *F*1 show the highest variance.

Performance of scATAcat in annotating Granja PBMC scATAC-seq data

The scATAC-seq data from (44) serves as the next dataset we use to test and compare our scATAcat cell-type annotation method. This dataset has been annotated by the authors which we take as the ground-truth labels. We use the same prototypes and the differentially accessible regions as those used in the 10X PBMC dataset with scATAcat. We follow the standard scATAcat pipeline to preprocess and cluster the cells. Supplementary Figure S5A shows the result of the clustering. We then annotate the 11 pseudobulks by the prototype labels using scATAcat. The 3D projection plot depicting the co-embedding of the pseudobulks with prototypes is shown in Supplementary Figure S5B. We also calculate the high-dimensional Euclidean distances between the prototypes and pseudobulks, assigning each cluster to its nearest prototype cell type. As shown in Supplementary Figure S5C, similar to the observations in the 10X PBMC sc-multiome dataset, some pseudobulks, such as cluster 2, show mixed signal. This may indicate a poor clustering performance. Additionally, despite projecting rather far from B cells in the 3D PCA plot, cluster 8 is annotated as a B cell. Given that the projection plot reflects only the first three PCs, it may not not be sufficient to capture the real relationships in data. This once again highlights the importance of considering higher dimensions when determining the distances.

We tried marker based annotation after obtaining gene activity scores. With the exception of the B cell marker MS4A1, which displays a weak but distinct signal around cluster 4 (Supplementary Figure S5D), all other marker genes failed to highlight particular clusters (Supplementary Figure S6B–L). We therefore excluded marker based annotation from our evaluation again.

To annotate the cells via label-transfer, we follow the same procedures and the reference dataset as previously used for annotating 10X PBMC sc-multiome with the same approach. Supplementary Figure S5E shows the UMAP embedding of the resulting annotations. Although B cells and NK cells shows a coherent signal, neither the CD8 and CD4 T cells nor the monocytes and dendritic cells form distinct clusters, indicating poor annotation performance.

For Cellcano, we made use of Satpathy PBMC sc-multiome data and the cell-type labels as the training data. The gene activity scores of the query data is then annotated with the trained model.

Similarly, for EpiAnno, we use the ArchR derived peaks of Satpathy PBMC sc-multiome data as the reference dataset. We next obtain the peaks of query Granja PBMC scATAC-seq data with ArchR and unify these peak regions to prepare the inputs for EpiAnno. We train Epianno with the Satpathy data and determine the cell-type labels of the query data, leveraging the trained model.

Performance assessment

The reference datasets utilized in different annotation methods contains annotations of varying detail, and the used nomenclature may differ. In order to include as many cells as possible, we adjust the annotations to a coarser annotation scheme. Supplementary Table S5 presents both the original and adjusted annotations. Figure 5A shows the results of the first evaluation strategy. scATAcat does well across all the methods and metrics except for cluster accuracy, where label-transfer also reaches a perfect score, and median precision where EpiAnno performs equally well as scATAcat. The performance ranking of label-transfer, Cellcano and EpiAnno varies depending on the metric considered, indicating that each method has unique advantages that may be leveraged for specific question.

The second evaluation strategy also resulted in similar results with scATAcat being the best performing method. Both scATAcat and label-transfer reach a perfect score for Cohen's kappa score as shown in Supplementary Figure S5F.

The results of testing the effect of clustering on evaluation metrics are presented in Supplementary Table S12. Again, the metrics perform quite stably across parameter choices. Only cluster accuracy displays higher variation, possibly reflecting clustering granularity.

Performance of scATAcat in annotating Granja BMMC scATAC-seq data

Another datatset we leverage from (44) is the BMMC scATAC-seq data. We applied scATAcat, marker based annotation, label-transfer, Cellcano and EpiAnno to this dataset, following the same strategies and the reference datasets introduced in the NeurIPS bone marrow sc-multiome data. The results are shown in Supplementary Figures S7 and S8.

Performance assessment

We compared the annotations obtained from these methods to the ground-truth annotations provided by the authors. We again adjusted the annotations across different reference datasets and the original and used annotations are presented in Supplementary Table S6. Figure 5B shows the performance metrics across methods when calculated with the first evaluation strategy. EpiAnno performed poorly similar to its application to NeurIPS bone marrow data. To successfully calculate the metrics, we excluded this method

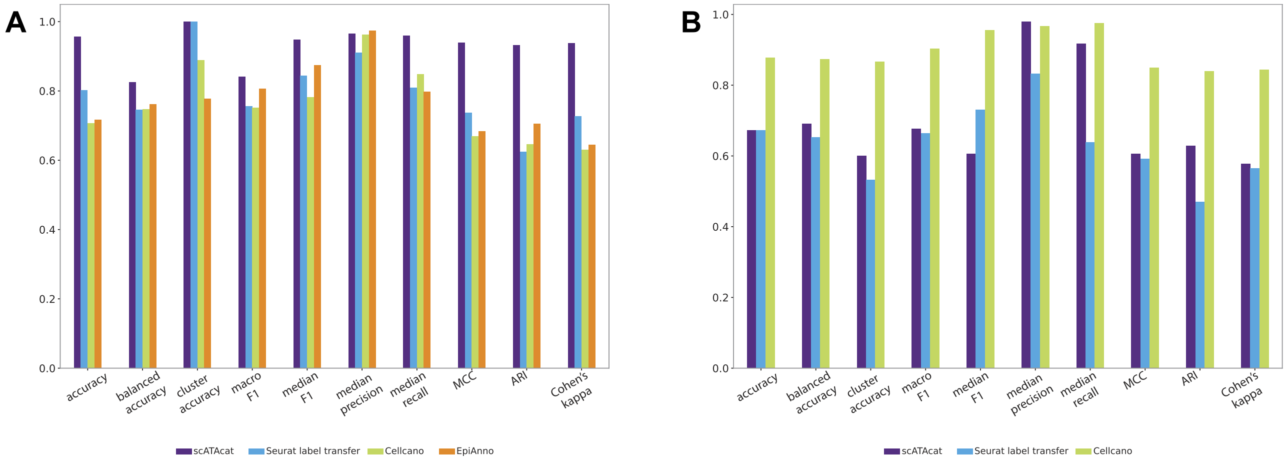


Figure 5. Performance evaluation of methods in annotating Granja data. Evaluation of cell-type annotations performances for (**A**) PBMC and (**B**) BMMC scATAC-seq data across methods.

from the evaluation. Cellcano showed its best performance on this data, achieving the highest scores across metrics, except for median precision, where scATAcat performed better. Similar result are observed for the second evaluation strategy, shown in Supplementary Figure S7F. Cellcano performed the best across all methods, except for the median F1 score, where label-transfer achieved an almost perfect score.

Performance of scATAcat in annotating Corces brain scATAC-seq data

We then proceeded to evaluate the performance of scATAcat in a different tissue by applying all the methods to scATAC-seq data from the human brain. To apply scATAcat, we leverage the sorted bulk ATAC-seq data of brain cell-types from BOCA2 (53) as the prototypes. We define the differentially accessible regions. Supplementary Figure S12D shows the number of differential regions across cell-types. Apparently, in these brain cell types their number is larger than in blood cell and to make use of these regions we combine the top 5000 regions per comparison to obtain the differential regions for this analysis. The scATAC-seq data is processed as introduced earlier and the clustering of the cells resulted in 10 clusters (Figure 6A). The coembedding of the pseudobulks of these clusters along with the bulk prototypes is shown in Figure 6B. Most of the clusters project close to oligodendrocytes and none of the clusters are too close to GABAergic or glutamatergic neurons. Nevertheless, when the distances are calculated in higher dimensions, the similarities between clusters and prototypes become more apparent (Figure 6C) and we annotate each cluster by the closest prototype. For this dataset, the marker gene based annotation showed promising results for some of the marker genes like OLIG2 and GAD1, as can be seen from Supplementary Figure S9D–L. However, in the given clustering granularity, it is still challenging to make cell-type assignments with this approach.

We used Allen Brain Map primary motor cortex snRNA-seq as the reference data (shown in Supplementary Figure S9B) for the label-transfer approach. Supplementary Figure S9C shows the UMAP embedding of the data with cells colored by the label-transfer-based cell-type labels. In this, most of the cells are annotated as excitatory (glutamatergic) neurons and the cluster boundaries are fuzzy indicating rather poor annotation.

For both Cellcano and EpiAnno we used the scATAC-seq part of the human cerebral cortex sc-multiome data from (48) as the training dataset. We follow the same strategies introduced in the earlier sections to prepare the input data and train the models to obtain predicted cell-type annotations of the Corces scATAC-seq data.

Performance assessment

We next compared and evaluated the annotations provided by each method taking the author annotations as the ground-truth. The both the ground-truth annotations and those derived from each method were adjusted to a coarser annotation scheme to enable evaluation of most number of cells, as presented in Supplementary Table S7. Figure 6D shows the results of the first evaluation strategy. Although label-transfer achieves perfect sore for median precision, and Cellcano does better for ARI, scATAcat shows the best performance for all the other metrics. In the second evaluation strategy, methods exhibit varied strengths (Supplementary Figure S9N): EpiAnno is high in accuracy score, whereas Cellcano showcase the highest adjusted rand index.

Supplementary Figure S9O and Supplementary Table S14 show the effect of clustering granularity on the evaluation metrics. The lowest clustering parameters leading to a very coarse clustering resulted in the highest scores, possibly because they did not account for all cell types. Subsequently, this dataset showed improved scores with increasing clustering granularity in all metrics apart from balanced accuracy and median recall.

Performance of scATAcat in annotating Morabito brain cortex scATAC-seq

The Morabito brain cortex scATAC-seq is also annotated with all five annotation methods following the same instructions to the annotation of Corces brain scATAC-seq data. The results are shown and detailed in Supplementary Figures S10 and S11, and Supplementary Tables S8 and S15.

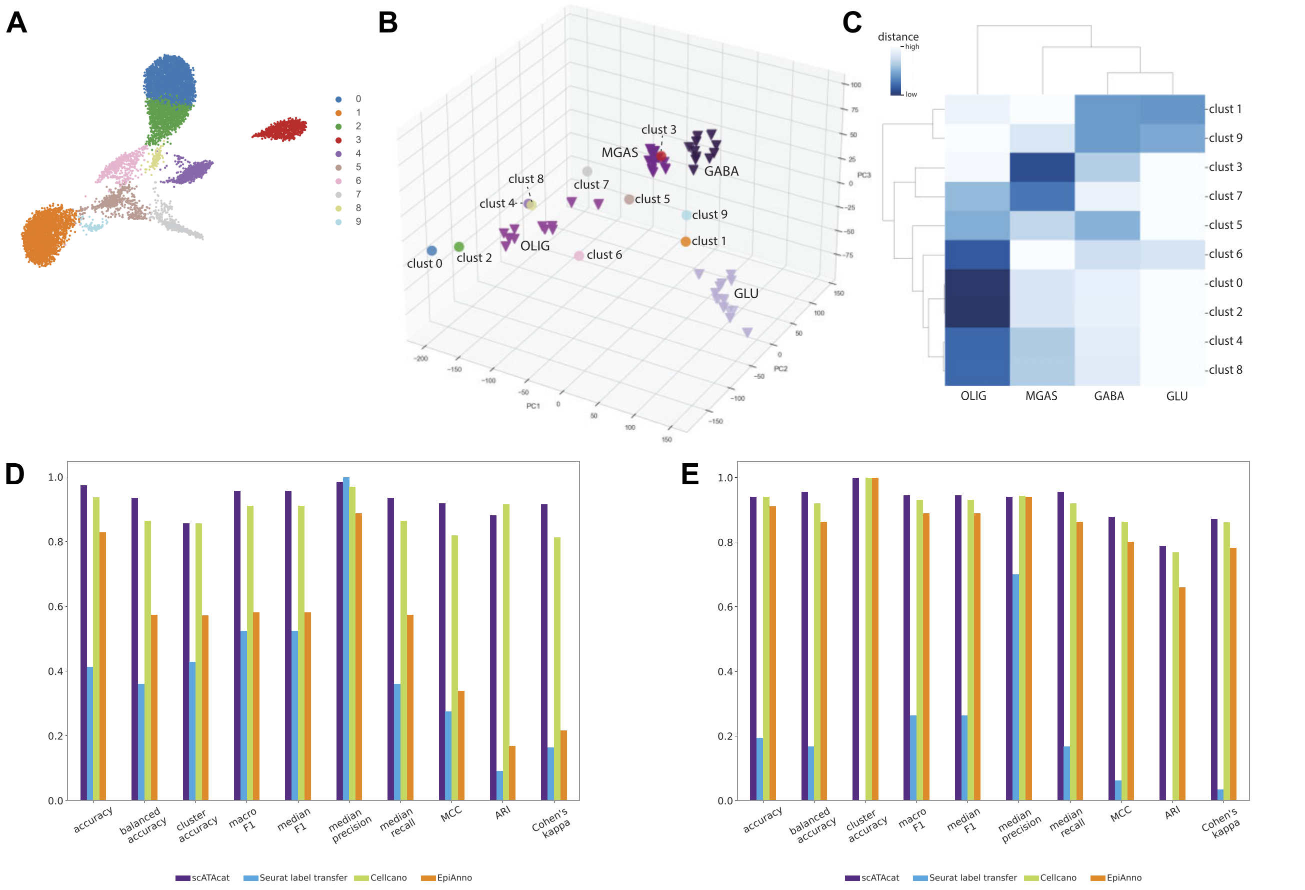


Figure 6. Performance of scATAcat in annotating Corces brain and Morabito brain cortex scATAC-seq data. (**A**) UMAP embedding of Corces brain scATAC-seq data colored by the clustering. (**B**) 3D PCA projection of the Corces brain scATAC-seq pseudobulks together with prototypes. Triangles represents bulk samples while circles represent pseudobulk-clusters. (**C**) Heatmap of the high-dimensional Euclidean distances between the pseudobulks and prototypes shown in (B). Microglia and astrocytes (MGAS), GABAergic neurons (GABA), Glutamatergic neurons (GLU), Oligodendrocytes (OLIG). (D, E) Evaluation of cell-type annotations performances (**D**) for Corces brain data and (**E**) for Morabito brain cortex data across methods.

Performance assessment

The performances of the methods in the conservative evaluation strategy is presented in Figure 6E. All methods, except for label-transfer, demonstrated good performance. However, scATAcat was the best across many metrics, although by a small margin compared to Cellcano. The second evaluation strategy showed similar results (Supplementary Figure S10F). scATAcat again shows the best scores across metrics except for median *F*1 score where both Cellcano and EpiAno achieved higher scores.

In this dataset, increasing clustering granularity positively affects evaluation metrics, peaking at a granularity level of 0.7 with the highest scores across several metrics, as shown in Supplementary Figure S11N and Supplementary Table S15. However, beyond this peak, the improvement in metrics does not consistently continue, indicating a limit to the benefits of increased granularity.

Discussion

Current single-cell methods like scATAC-seq hold the promise of unraveling the gene regulation at a more precise resolution. Studying the accessibility profiles of cells enables understanding of the regulatory landscape governing each cell type. However, the resulting data presents a number of challenges which need to be tackled to fully capitalize on the valuable of information they offer.

One of the foremost challenges in scATAC-seq data analysis is to annotate cell-types. Typically, the methods developed for cell-type annotation in scRNA-seq data are borrowed to annotate the cells in scATAC-seq data. This is enabled by transforming the scATAC-seq data into scRNA-seq-like-data using the accessibility around the genes as a proxy for expression. In this study, we have demonstrated that this approach may not yield optimal results, and does not fully exploit the potential of scATAC-seq, which inherently offers more cell-type specificity. The other two methods used in this study, Cellcano and EpiAnno, are specifically developed for cell-type annotation in scATAC-seq data. These approaches use existing annotated scATAC-seq datasets to train a machine learning model, which is then applied to new datasets for cell-type annotation. However, this strategy faces challenges similar to those of the label-transfer method. Most of the scATAC-seq datasets used for training are annotated using scRNA-seq-based methods by relying on the predicted expression levels of marker genes, which we showed to be problematic. We introduced scATAcat to address this issue of circular reliance on scRNA-seq-based annotations. We also argue that annotation within the same modality would improve the cell-type annotations. scATAcat enables cell-type annotation of clusters in scATAC-seq data by leveraging prototype accessibility profiles typically obtained from bulk ATAC-seq data.

Another challenge in scATAC-seq data is the sparsity. Limited information in individual cells makes it impractical to perform annotation at this level. As a remedy, we propose the pseudobulk. Given that the clustering is performed sufficiently fine-grained to not merge different cell-types, pseudobulk reflects the accessibility of the cell-type appropriately. Consequently, all the cells in a pseudobulk share the same cell-type annotation.

One of the key decisions in scATAC-seq data analysis is to determine an appropriate reference frame. Label-transfer and Cellcano use genes, offering stability but only leveraging a limited portion of the ATAC-seq data. On the other hand, EpiAnno relies on the ATAC-peak regions which vary across datasets. This poses challenges when performing supervised annotation. More specifically, it requires choosing between using the peaks from either the training or test dataset and adjusting the other accordingly. In their original paper, authors base the analysis on the test peaks and adapt the training data to this. This approach is problematic in a supervised learning context, where the model should be blind to test data and it's features. Also, this requires retraining the model for each new dataset. When designing scATAcat, we replace the ATAC-peaks by ENCODE cCREs. As ENCODE cCREs are derived through the integration of various experiments across cell types, the advantage here is the stability of the reference frame which is a prerequisite for data integration across different experiments.

In addition to cell-type annotation, our program scATAcat provides useful visualizations for the interpretation of the these annotations. In one, we provide the heatmap of the Euclidean distances which illustrates the similarity between all the clusters and prototypes. Additionally, we provide a bipartite heatmap showing the similarities of pseudobulks to each prototype cell-type. This heatmap serves as a quantitative indicator, providing a measure of confidence in the annotations. In this way, one can, e.g. identify potential doublet clusters showing similarity to multiple cell-types, as well as a new cell-type with poor annotation confidence to all the prototypes.

As reported in (59), the definition of ground-truth data is generally a challenge in single-cell studies. We mostly rely on the annotations provided in the publications of the datasets derived, except for the first dataset where the cell-type annotations are derived experimentally. Consequently, our evaluations (as well as any other one) may be influenced by the method through which the ground-truth is established.

In our study, the observed performance of scATAcat was superior to other methods except for two datasets. One of these datasets is the 10X PBMC sc-multiome dataset where the label-transfer approach showed the best performance. However, in the absence of an external annotation for it, we annotated the scRNA-seq part of the multiome data with the label-transfer approach and used these annotations as the ground-truth. This may, of course, introduce a bias in favor of this method.

Interestingly, on the Granja BMMC dataset Cellcano performs better than scATAcat, in contrast to the NeurIPS bone marrow data where it is the other way around. Besides the cluster granularity as one possible reason, it is interesting to consider the dynamic character of hematopoietic development. Bone marrow consists of the progenitors and the terminal cell-types derived from these progenitors. The bone marrow progenitor cells follow a differentiation trajectory and are expected to show a dynamic and heterogeneous accessibility profile (40) whereas terminal cell states, such as PBMCs, mostly consist of naïve or resting cells (56), suggesting more stable chromatin accessibility profiles. This may make it harder to capture the true chromatin accessibility signatures of progenitor cells, consequently making their cell-type annotation harder. The Granja dataset includes more progenitor cells than NeurIPS bone marrow data which may explain scATAcat's poorer performance. Additionally, the better performance of Cellcano in this case indicates that here the neural network indeed contributes in capturing this complex relationship.

While our method has provided valuable insights, it also comes with limitations. First of all, scATAcat requires prototype cells as input, consequently, as in any reference based tool, can only annotate the cell-types for provided prototypes. Secondly, we build on the assumption that the pseudobulk clusters are sufficiently pure. However, this assumption may not hold true especially in the case of complex samples including similar cell-types. In general, we recommend opting for a higher number of clusters to ensure more homogeneous clusters. That being said, we also provided evidence that our method is fairly robust to clustering resolution (Supplementary Tables S9–S15).

A potential further extension of the method would be the use of single-cell-pseudobulk prototypes instead of bulk prototypes. This approach would leverage the increasingly available and more homogenous scATAC-seq atlases. Additionally, conceptually our method is not restricted to scATAC-seq data. We are currently working on applying it to single-cell epigenetic data like single-cell DNA methylation or single-cell ChIP-seq.

Data availability

scATAcat is available as a python package at https://github.com/aybugealtay/scATAcat and https://zenodo.org/records/12586074.

The scripts used to generate the figures from this manuscript along with the figures/tables from the Supplementary materials and their respective outputs can be accessed at https://github.com/aybugealtay/scATAcat_paper and https://zenodo.org/records/13381495.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors would like to thank Zhicheng Ji for his valuable inputs regarding feature selection step. The authors thank Ektha Shah for discussions on co-embedding choices. The authors also gratefully acknowledge the IT group of the Max Planck Institute for Molecular Genetics for their support and flawless maintenance of the computational resources.

Funding

A.A. gratefully acknowledges funding by Deutsche Forschungsgemeinschaft in the International Research Training Group 'Dissecting and Reengineering the Regulatory Genome' [IRTG2403].

Conflict of interest statement

None declared.

References

1. Poirier,M.G., Bussiek,M., Langowski,J. and Widom,J. (2008) Spontaneous access to DNA target sites in folded chromatin fibers. *J. Mol. Biol.*, **379**, 772–786.
2. Corces,M.R., Buenrostro,J.D., Wu,B., Greenside,P.G., Chan,S.M., Koenig,J.L., Snyder,M.P., Pritchard,J.K., Kundaje,A., Greenleaf,W.J., *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48**, 1193–1203.
3. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
4. Eberwine,J., Sul,J.Y., Bartfai,T. and Kim,J. (2013) The promise of single-cell sequencing. *Nat. Methods*, **11**, 25–27.
5. Zeng,H. (2022) What is a cell type and how to define it?. *Cell*, **185**, 2739–2755.
6. Trapnell,C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.
7. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M., *et al.* (2017) The human cell atlas. *eLife*, **6**, e27041.
8. Stuart,T., Srivastava,A., Madad,S., Lareau,C.A. and Satija,R. (2021) Single-cell chromatin state analysis with Signac. *Nat. Methods*, **18**, 1333–1341.
9. Granja,J.M., Corces,M.R., Pierce,S.E., Bagdatli,S.T., Choudhry,H., Chang,H.Y. and Greenleaf,W.J. (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, **53**, 403–411.
10. Gralinska,E., Kohl,C., Fadakar,B.S. and Vingron,M. (2022) Visualizing cluster-specific genes from single-cell transcriptomics data using association plots. *J. Mol. Biol.*, **434**, 167525.
11. Gralinska,E. and Vingron,M. (2023) Association Plots: visualizing cluster-specific associations in high-dimensional correspondence analysis biplots. *J. Roy. Stat. Soc. Ser. C: Appl. Stat.*, **72**, 1023–1040.
12. Kiselev,V.Y., Yiu,A. and Hemberg,M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, **15**, 359–362.
13. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
14. Aran,D., Looney,A.P., Liu,L., Wu,E., Fong,V., Hsu,A., Chak,S., Naikawadi,R.P., Wolters,P.J., Abate,A.R., *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172.
15. de Kanter,J.K., Lijnzaad,P., Candelli,T., Margaritis,T. and Holstege,F.C. (2019) CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.*, **47**, e95.
16. Cortal,A., Martignetti,L., Six,E. and Rausell,A. (2021) Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat. Biotechnol.*, **39**, 1095–1102.
17. Alquicira-Hernandez,J., Sathe,A., Ji,H.P., Nguyen,Q. and Powell,J.E. (2019) ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.*, **20**, 264.
18. Hu,J., Li,X., Hu,G., Lyu,Y., Susztak,K. and Li,M. (2020) Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat. Mach. Intell.*, **2**, 607–618.
19. Ma,F. and Pellegrini,M. (2020) ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics*, **36**, 533–538.
20. Song,Q., Su,J. and Zhang,W. (2021) scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat. Commun.*, **12**, 1–11.
21. Xu,C., Lopez,R., Mehlman,E., Regier,J., Jordan,M.I. and Yosef,N. (2021) Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, **17**, e9620.
22. Abdelaal,T., Michielsen,L., Cats,D., Hoogduin,D., Mei,H., Reinders,M.J. and Mahfouz,A. (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 194.
23. Ma,W., Lu,J. and Wu,H. (2023) Cellcano: supervised cell type identification for single cell ATAC-seq data. *Nat. Commun.*, **14**, 1864.
24. Chen,X., Chen,S., Song,S., Gao,Z., Hou,L., Zhang,X., Lv,H. and Jiang,R. (2022) Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nat. Mach. Intell.*, **4**, 116–126.
25. Chen,S., Yan,G., Zhang,W., Li,J., Jiang,R. and Lin,Z. (2021) RA3 is a reference-guided approach for epigenetic characterization of single cells. *Nat. Commun.*, **12**, 2177.
26. Abascal,F., Acosta,R., Addleman,N.J., Adrian,J., Afzal,V., Aken,B., Akiyama,J.A., Jammal,O.A., Amrhein,H., Anderson,S.M., *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
27. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R. and et,al. (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.
28. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
29. Thibodeau,A., Eroglu,A., McGinnis,C.S., Lawlor,N., Nehar-Belaid,D., Kursawe,R., Marches,R., Conrad,D.N., Kuchel,G.A., Gartner,Z.J., *et al.* (2021) AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol.*, **22**, 252.
30. Zandigohar,M. and Dai,Y. (2022) Information retrieval in single cell chromatin analysis using TF-IDF transformation methods. arXiv doi: https://arxiv.org/abs/2212.05184, 10 December 2022, preprint: not peer reviewed.
31. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
32. McInnes,L., Healy,J. and Melville,J. (2018) Umap: uniform manifold approximation and projection for dimension reduction. arXiv doi: https://arxiv.org/abs/1802.03426, 09 February 2018, preprint: not peer reviewed.
33. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 5233.
34. Virshup,I., Rybakov,S., Theis,F.J., Angerer,P. and Wolf,F.A. (2021) anndata: annotated data. bioRxiv doi: https://doi.org/10.1101/2021.12.16.473007, 19 December 2021, preprint: not peer reviewed.
35. Hu,C., Li,T., Xu,Y., Zhang,X., Li,F., Bai,J., Chen,J., Jiang,W., Yang,K., Ou,Q., *et al.* (2023) CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.*, **51**, D870–D876.
36. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
37. Levine,J.H., Simonds,E.F., Bendall,S.C., Davis,K.L., Amir,E.A.D., Tadmor,M.D., Litvin,O., Fienberg,H.G., Jager,A., Zunder,E.R., *et al.* (2015) Data-driven phenotypic dissection of AML reveals

progenitor-like cells that correlate with prognosis. *Cell*, **162**, 184–197.

38. Zhang,Y., Zhang,F., Wang,Z., Wu,S. and Tian,W. (2022) scMAGIC: accurately annotating single cells using two rounds of reference-based classification. *Nucleic Acids Res.*, **50**, e43.
39. Gorodkin,J. (2004) Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.*, **28**, 367–374.
40. Buenrostro,J.D., Corces,M.R., Lareau,C.A., Wu,B., Schep,A.N., Aryee,M.J., Majeti,R., Chang,H.Y. and Greenleaf,W.J. (2018) Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, **173**, 1535–1548.
41. Human PBMC sc-multiome dataset, 10X Genomics, https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0.
42. Luecken,M., Burkhardt,D., Cannoodt,R., Lance,C., Agrawal,A., Aliee,H., Chen,A., Deconinck,L., Detweiler,A., Granados,A., *et al.* (2021) A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: Vanschoren,J. and Yeung,S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. Vol. 1*. Curran.
43. Lance,C., Luecken,M.D., Burkhardt,D.B., Cannoodt,R., Rautenstrauch,P., Laddach,A., Ubingazhibov,A., Cao,Z.-J., Deng,K., Khan,S., *et al.* (2022) Multimodal single cell data integration challenge: results and lessons learned. bioRxiv doi: https://doi.org/10.1101/2022.04.11.487796, 12 April 2022, preprint: not peer reviewed.
44. Granja,J.M., Klemm,S., McGinnis,L.M., Kathiria,A.S., Mezger,A., Corces,M.R., Parks,B., Gars,E., Liedtke,M., Zheng,G.X., *et al.* (2019) Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.*, **37**, 1458–1465.
45. Corces,M.R., Shcherbina,A., Kundu,S., Gloudemans,M.J., Frésard,L., Granja,J.M., Louie,B.H., Eulalio,T., Shams,S., Bagdatli,S.T., *et al.* (2020) Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.*, **52**, 1158–1168.
46. Morabito,S., Miyoshi,E., Michael,N., Shahin,S., Martini,A.C., Head,E., Silva,J., Leavy,K., Perez-Rosendahl,M. and Swarup,V. (2021) Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.*, **53**, 1143–1155.
47. Satpathy,A.T., Granja,J.M., Yost,K.E., Qi,Y., Meschi,F., McDermott,G.P., Olsen,B.N., Mumbach,M.R., Pierce,S.E., Corces,M.R., *et al.* (2019) Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.*, **37**, 925–936.
48. Zhu,K., Bendl,J., Rahman,S., Vicari,J.M., Coleman,C., Clarence,T., Latouche,O., Tsankova,N.M., Li,A., Brennand,K.J., *et al.* (2023) Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Sci. Adv.*, **9**, eadg3754.
49. Allen Brain Map primary motor cortex snRNA-seq, human M1, 10x Genomics, https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x.
50. Bakken,T.E., Jorstad,N.L., Hu,Q., Lake,B.B., Tian,W., Kalmbach,B.E., Crow,M., Hodge,R.D., Krienen,F.M., Sorensen,S.A., *et al.* (2021) Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*, **598**, 111–119.
51. Calderon,D., Nguyen,M.L., Mezger,A., Kathiria,A., Müller,F., Nguyen,V., Lescano,N., Wu,B., Trombetta,J., Ribado,J.V., *et al.* (2019) Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.*, **51**, 1494–1505.
52. Hitz,B.C., Lee,J.-W., Jolanki,O., Kagda,M.S., Graham,K., Sud,P., Gabdank,I., Strattan,J.S., Sloan,C.A., Dreszer,T., *et al.* (2023) The ENCODE uniform analysis pipelines. bioRxiv doi: https://doi.org/10.1101/2023.04.04.535623, 06 April 2023, preprint: not peer reviewed..
53. Hauberg,M.E., Creus-Muncunill,J., Bendl,J., Kozlenkov,A., Zeng,B., Corwin,C., Chowdhury,S., Kranz,H., Hurd,Y.L., Wegner,M., *et al.* (2020) Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.*, **11**, 5581.
54. Roy,A., Wang,G., Iskander,D., O'Byrne,S., Elliott,N., O'Sullivan,J., Buck,G., Heuston,E.F., Wen,W.X., Meira,A.R., *et al.* (2021) Transitions in lineage specification and gene regulatory networks in hematopoietic stem/progenitor cells over human development. *Cell Rep.*, **36**, 11.
55. Pasquini,G., Arias,J. E.R., Schäfer,P. and Busskamp,V. (2021) Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.*, **19**, 961–969.
56. Kleiveland,C. (2015) Peripheral blood mononuclear cells. In: *The Impact of Food Bioactives on Health: In Vitro and Ex Vivo Models*. pp. 161–167.
57. Jensen,K.C., Higgins,J.P., Montgomery,K., Kaygusuz,G., Rijn,M.V.D. and Natkunam,Y. (2007) The utility of PAX5 immunohistochemistry in the diagnosis of undifferentiated malignant neoplasms. *Mod. Pathol.*, **20**, 871–877.
58. Sender,R., Fuchs,S. and Milo,R. (2016) Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.*, **14**, e1002533.
59. Lähnemann,D., Köster,J., Szczurek,E., McCarthy,D.J., Hicks,S.C., Robinson,M.D., Vallejos,C.A., Campbell,K.R., Beerenwinkel,N., Mahfouz,A., *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.

Received: January 19, 2024. **Revised:** September 11, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 23, 2024