## Supplemental Methods

### Detail description of Preprocessing of reference single-cell RNA-seq data

*Processing large inbuilt pool of reference single-cell open-chromatin profile*

We used the logarithmic value of the p-values of gene-enrichment genes and further processed large reference single-cell open chromatin data-sets. The matrices containing the vectors of logarithmic values of gene-enrichment scores for single-cell open chromatin profiles also have high dimensions. Hence, an autoencoder was implemented to reduce the feature dimension for clustering reference scATAC-seq profiles. Here, we used a convolutional autoencoder model, which automates the learning of optimal filter by minimizing the reconstruction error(Masci et al. 2011). The feature extracted in the minimum neuron layer of the autoencoder can then be used as a representation of a data point in a reduced dimension. For each fully connected layer in the architecture, network architecture, we used the sigmoid activation function to learn the non-linearity of the data. The sigmoid function used here is

$$\Phi\,(x) \;=\; \frac{1}{1+e^{-x}} \qquad (1)$$

where the resultant value lies between 0 and 1.
The optimization of the autoencoder was done using a stochastic gradient descent optimizer with a learning rate of 0.01, decay 1e-6, and momentum of 0.9. Stochastic gradient descent replaces the actual gradient by making an estimate derived from a randomly selected subset of the data. Hence, it red,uces the computational burden and achieves faster iterations but compromises for a slower convergence rate. The loss function for optimization used is mean squared error

$$MSE \;=\; \frac{1}{NM}\sum_{i}^{N}\sum_{j}^{M}(I'_{ij} - I_{ij})^2 \qquad (2)$$

Wher $I'_{ij}$ is the reconstructed cell gene-enrichment level of gene j at the output (last layer) of auto-encoder, $I_{ij}$ is input cell gene-enrichment score, and $N, M$ is the number of cells and genes, the respectively. After dimension reduction by feature extraction using an auto-encoder, we used a self-organising map (SOM) to classify reference single-cell ATAC-seq profiles. Our SOM model used Euclidean distance to measure the similarity between sample vectors (features extracted from autoencoder) and nodes. Thus, we clustered the single-cell ATAC-seq profiles (auto-encoder derived features) using SOM and achieved a representative node (cluster) to represent multiple single-cell at a time.

*Processing large inbuilt pool of reference single-cell expression data-sets*

We used cell-type annotation to group single-cell expression profiles instead of auto-encoder and SOM. For multiple human and mouse scRNA-seq profiles, cell-type annotations were already available. We performed annotation for scRNA-seq profiles without cell-type information using matchscore2 (Mereu et al. 2020). After making major clusters of cells using cell-type annotation, KNN based method was used to find sub-clusters of cells in each cluster. The vectors containing the mean values of expression of genes of cells belonging to a sub-cluster were used as their representatives. Thus, for more than 3 million human and mouse scRNA-seq profiles, we made approximately 30,000 representative vectors of expression value. This process helped in two ways; first, the mean vectors were less sparse; second, it increased the speed of searching the matching expression profile for a query using a hierarchical approach (supplemental Fig. S1B). The webserver and standalone versions of scEpiSearch are implemented using Python programming language.

The current version of scEpiSearch consists of a compilation of single-cell expression profiles for 2,239,727 human cells and 2,141,797 mouse cells (total ~4.3 million cells) and scATAC-seq profiles of 742,297 human cells and 81,173 mouse cells (total ~800,000 epigenome profiles) (supplementary Table 1).

**Calculating global accessibility score**

For both species, human and mouse, we have compiled global accessibility peak-list using several published open-chromatin profiles of bulk samples. For this purpose, we used the available peak lists of open chromatin profiles (DNAse-seq and ATAC-seq) of bulk samples, available at the GEO database(Barrett et al. 2005), UCSC genome browser(Lee et al. 2020) and iHEC portal(Bujold et al. 2016). We merged the peaks lying within 1 kb of each other. The number of times a genomic site appeared as a peak in published open chromatin profiles of bulk samples was defined as its global accessibility score(Chawla et al. 2021). Thus we have almost 1 million sites ( width > 1 kbp ) in our global accessibility list for both humans and mice.

**Query preprocessing**

ScEpiSearch first highlights cell-type-specific peaks (mostly enhancers) by dividing the scATAC-seq read-count of every peak in the query peak-list by its global accessibility score (as shown in equation (1)). It finds genes proximal to peaks in the query scATAC-seq profile. In order to find proximal genes quickly, it first uses the pre-existing table of genes proximal to the peaks in the global accessibility list. For every query scATAC-seq profile, scEpiSearch finds an overlap between its peaks and sites in the global accessibility list. With our analysis, we found that most of the time, scEpiSearch achieves overlap with the global accessibility list for 65-80% of the peaks in query. For peaks that overlap with sites in the global accessibility list, ScEpisearch adapts their proximal genes from the pre-existing tables. For peaks that do not overlap with the global accessibility list, it searches proximal genes separately (if the accurate mode is selected).

*Gene enrichment scores of query:* For every cell in the query, scEpiSearch selects genes proximal to peaks with high normalized read-count as foreground while keeping genes near all the peaks of the query cell as background. For every query cell, it uses foreground and background genes to calculate gene-enrichment scores using Fisher's exact test as explained in equation (2).

**Approach of hierarchical Search of scEpiSearch**

The search for matching expression profiles is done hierarchically, first, the MExTEG value is calculated using an expression vector for the cluster of reference scRNA-seq profiles. Notice that for a cluster, the representative expression vector contains the mean of cell-specific expression values of genes from cells belonging to that cluster. Further, the MExTEG for query cells is converted to P-value using the null model of representative expression vectors. Then top N clusters are chosen, which have the lowest p-value for MExTEG. Further, MExTEG for query cells is again computed using a cell-type-specific expression profile of single-cells present in top N-selected clusters. One hundred cells are selected based on a higher MExTEG value. For 100 reference cells with high MExTEG for query, the MExTEG is converted to a p-value score using a null model. Notice that we use p-values based on MExTEG rather than MExTEG score itself to identify the top matching clusters or single-cells to avoid batch effect due to variability in sparseness (or sequencing depth) and bias for certain genes in reference scRNA-seq profiles. If a reference cell has less sparse (of high sequencing depth) expression values, then the MExTEG score would be high for the query as well as null model cells. Hence comparison of the MExTEG score of null models and query cells to calculate significance (or p-value) would reduce the such unseen batch effect.

***Rank-based statistical approach to refine significance of match****:* To reduce bias in the search for matching cells, scEpiSearch further refines or adjusts the rank of matches due to P-values calculated using MExTEG. For this purpose, we keep the precalculated rank of every reference cell for all cells in the null model (false queries explained above). Such rank calculation provides a view of bias in the data and enumerates the number of times a reference cell comes in the top hit for the cells in the null model. Thus, after calculating the P-value of the match and determining the rank of a reference cell for a query cell, we estimate a new P-value. The new P-value of the match between a reference cell and a query cell is calculated as the fraction of cells in the null model for which the same reference cell has a better rank than for the query cell.

**Cross-species Search**

scEpiSearch also allows matching human cell scATAC-seq profiles with mouse reference scRNA-seq dataset. It is based on the fact that cell-type-specific expression of genes in the same cell type from two species is highly likely to be similar, as the same markers are often used to identify different cell-type in both humans and mice. Hence, scEpiSearch uses the approach of highlighting enriched genes proximal to foreground peaks (possibly enhancers) with high cell-type specificity. Therefore theoretically, it is possible to find the correct matching mouse reference expression dataset for query scATAC-seq profiles of human cells using our approach ells. For this purpose, scEpSearch transforms read counts to gene enrichment scores for query scATAC-seq profiles of human cells. For the human cell query, it calculates MExTEG using the reference single-cell expression profile of mouse cells. For calculating P-value, it uses precalculated MExTEG value for a null model made from human scATAC-seq profiles and reference expression dataset of mouse cells. After finding the rank of reference mouse cells for a query human scATAC-seq profile based on MExTEG-based P-value, scEpiSearch calculates a new P-value for the match with a reference mouse expression vector based on its precalculated ranks for the null model made from human cells.

**Word Cloud**

Wordcloud is also shown in the results of scEpiSearch. It displays phenotypic information of top hits in both expression and epigenome across all query cells where the size of each word/match in the word-cloud figure depends upon the frequency of phenotype. We used the”wordcloud” library in python to generate such figures.

**Datasets and parameters used for analysis**

The sources of the single-cell open-chromatin profile dataset used for Figure 2 have been mentioned in supplemental tables S2-3. The sci-ATAC-seq dataset used for unannotated cells from 4 different organs is available in the GEO database (GSE111586). All the single-cell open-chromatin profiles used in the study were unimputed except for the sci-ATAC-seq profile from Cussanovich et al. We used FITs for imputation and recovering signals in unknown cells in 4 organs in datasets by Cussonovich et al. The scATAC-seq profile used to study the lineage for HL60 and K562 cell lines have GEO ids: GSE109828 and GSE65360, respectively.

Notice that we used the read-count matrices for all the case studies without imputation except for the unknown cells from mouse organs published by Cusanovich et al. We used the imputed version of single-cell ATAC-seq profile of cell of mouse organ made available by Sharma et al. at (http://reggen.iiitd.edu.in:1207/FITS/imputed_finaldata/finalDataSets_final/cusanovich_data/FITS_cusanovich_data/)

For embedding case studies, we have used the dataset as such: Human Neuron (GSE97942), cells from Mouse forebrain (GSE1,00033), Human HSC (GSE96769), Mouse HSC (GSE111586), Human Myoblast (GSE109828), Human GM12878 (GSE109828), Mouse B cell (GSE111586) and Human GM12878 (GSE68103), Human Neuron (GSE97942), Mouse Forebrain (GSE1,00033), Human HSC (GSE96769), all mouse HSC from Bone marrow tissue (GSE111586).
A detailed description is also given in the corresponding section below.

**Evaluation based on comparison with the correlation of gene activity, enrichments, and predicted expression values**

While comparing the query scATAC-seq profile with reference cells (Figure 2A), we evaluated the approach of correlating gene scores (gene activity, enrichment score, predicted expression) estimated using single-cell epigenome profiles. For calculating gene-activity scores used Seurat 3.2, and for predicting gene-expression values, we used BABEL (Wu et al. 2021). We used BABEL in default mode after downloading their pre-trained model. We common genes between the BABEL gene list and our list for calculating correlation values. For estimating the gene-enrichment score we used the method described in the Method section. While comparing scATAC-seq profiles to reference expression values, we calculated the correlation between estimated gene-scores of a query with reference gene expression (FPKM, TPM, or UMI-counts).

**Evaluation of co-embedding of scATAC-seq and single-cell expression profiles across species**

Even though Leucken et al. (Luecken et al. 2022) (https://www.nature.com/articles/s41592-021-01336-8) reported that integrative methods for scATAC-seq produced unsatisfactory results, we wanted to be sure, and compare them with scEpiSearch. Hence we used three different methods (Seurat, LIGER and Conos) for comparison. The parameters used for different methods are written below:

Seurat: Seurat 3.2 (Stuart et al. 2019) was used, with details available at https://satijalab.org/seurat/archive/v3.2/atacseq_integration_vignette.html. RNA-seq reference count data and ATAC-seq queries were loaded in R, and Seurat object was created. Further standard analyses were performed on RNA-seq data (normalization, finding variable genes, scaling data and running PCA, t-SNE). ATAC-seq data is analyzed separately, and gene annotation information is added to the Seurat object. Anchors were identified between both modalities. The gene activity scores and scRNA-seq gene expression quantifications are used in canonical correlation analysis with all genes taken which were highly variable in RNA-seq dataset (parameters being reduction = "cca", k.anchor = 5, k.filter = 80). The label transfer was done using option weight.reduction = human.atac[["lsi"]]. Finally, for visualisation, RNA-seq is imputed into the scATAC-seq based on already computed anchors, and then datasets are merged with parameters using default parameters. Notice that single-cell ATAC-seq read-count matrices for mouse were in mm9 format and for human hg19 version was used.

LIGER: Linked Inference of Genomic Experimental Relationships (Liu et al. 2020) was used with details available at (http://htmlpreview.github.io/?https://github.com/welch-lab/liger/blob/master/vignettes/Integrating_scRNA_and_scATAC_data.html) ATAC-seq data was transformed into gene counts so that they could be compared to RNA-seq which were obtained by counting the total number of ATAC-seq reads within gene and promoter region (3 kb upstream) of each gene in each cell. Then after loading read counts in R, the LIGER object was created with the createLiger function. Both datasets are normalised using normalize function. Highly variable genes are identified and combined from both datasets. The parameter datasets. used was set to 2 such that genes could be selected from the RNA-seq dataset in functselect genesenes. Joint matrix factorization (iNMF) was performed on the normalized and scaled RNA and ATAC data using the *optimizeALS* function with the value of k being 20. Finally, to fully integrate datasets quantile normalization was performed through the quantile_norm function with the value of knn_k=5. runUMAP function was used to get coordinates for each cell in integrated visualization with parameters being distance = 'cosine', n_neighbors = 30, min_dist = 0.3. Notice that here we used the hg19 version of our read-count/peakfiles.

Conos : Conos R package (Barkas et al. 2019) was used using the github repository https://github.com/kharchenkolab/conos#basics-of-using-conos. All steps for integration of RNA and ATAC-seq mentioned in the tutorial (http://pklab.med.harvard.edu/peterk/conos/atac_rna/example.html) were followed. The gene activity scores generated from Seurat v3.2 were used for ATAC-seq. For preprocessing step basicP2proc, the pagoda2 package was used. Further buildGraph function was used with parameters k=15, k.self=5, k.self.weigh=0.01, ncomps=30 and n.odgenes=5e3. Embeddings were generated using the embedGraph() function, and coordinates were obtained from largeVis.

ScEpiSearch : The reference dataset was prepared as explained in the methods section. Cross Species search was performed for ATAC-seq query datasets with reference mouse si3 matches were selected and filtereda as per p-value cutoff of 0.05. Coordinates for each query cell was calculated as the average of coordinates of top 5 matches in the reference dataset. Combined visualization of reference and ATAC-seq query is then plotted in R. For most the figures we used t-SNE coordinates of reference cells from mouse cell atlas (MCA) dataset which were provided by Chawla et al. (2021). However, we also performed t-SNE based dimension reduction for reference cells to make sure that our results are replicable. Such as for Figure 2D, we used calculated new t-SNE coordinates for reference cells.

SnapATAC: SnapATAC python package 2.0 was used using the github repository GitHub - kaizhang/SnapATAC2: Single-cell epigenomics analysis tools. All steps for integration of RNA and ATAC-seq mentioned in the tutorial (Multi-modality pipeline: analyzing single-cell multiome data (ATAC + Gene Expression) — SnapATAC2 2.1.2 documentation (kzhang.org)) were followed. scglue.genomics.rna_anchored_prior_graph() was used to find anchors between RNA and ATAC. scglue.models.fit_SCGLUE() was used to train the model on both ATAC, RNA and anchored graphs of both modalities.

***Silhouette coefficient calculation***

Silhouette index or coefficient measures how similar a data-point is to its own predicted-cluster compared to other cluster. For each sample, silhouette coefficent is calculated using the mean intra-cluster distnce (a) and the mea nearest-custer distance b) as:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (3)$$

Where a(i) represents the average distance of $i^{th}$ data-point (sample) to all other data-points in the same cluster, whereas b(i) is the average distance of $i^{th}$ sample with all samples in the closest cluster (other than it a cluster of the ith sample). For the Silhouette coefficient calculation, we used the cluster package in R(https://www.rdocumentation.org/packages/cluster/versions/2.1.2/topics/silhouette), where we considered queries and their similar cell types from a reference in one class and rest other cell types were assigned their corresponding clusters or cell labels. E.g., If query cells were macrophage cells, then query cells and macrophages from the reference were assigned the same cluster, and other cell types were assigned their individual cluster label. For calculating distance among cells, we considered the t-SNE coordinates of cells obtained from each method's t-SNE plot.

***Embedding of multiple scATAC-seq profiles across batches and species***

We tested a few combinations of scATAC-seq read-count matrices with the standalone version of scEpiSearch. We could get satisfactory embedding of scATAC-seq profiles, such as cells of the same type colocalized together while maintaining clear cell-type separability.

Case 1 : We made queries from different batches and species such as Human Neuron (GSE97942), cells from the Mouse forebrain (GSE1,00033), Human HSC (GSE96769), Mouse HSC (GSE111586), Human Myoblast (GSE109828), Human GM (GSE109828), Mouse B cell (GSE111586) and Human

GM12878 (GSE68103) and passed these queries to Embedding module of the standalone version of scEpiSearch. The resultant embedding plot was made using the top ten matching clusters to build an adjacency matrix. NetworkX was used to make graphs and visualize clusters. It can be seen that HSC cells from both species are clustered together, while neuronal cells from both species (Human and Mouse) lie close to each other. The distance between HSC and neuronal cells is evident from the graph (Figure 4A). A separate cluster for B Cells and GM from both species can also be seen while they lie close to HSC cells for being in the class of immune cells. Human myoblast cells do not form a cluster with any group of cells.

Case 2: Here (Figure 4C), we used read-count matrices scATAC-seq profile of 4 cell types, namely: Human lymphoblastoid cell line (GM12878), mouse B cells, Human T cells and mouse T cells. In the 2D embedding plot made by scEpiSearch human lymphoblastoid cells and mouse B cells are colocalized (Figure S10). Similarly, human and mouse T cells are colocalized in scEpiSearch-based embedding. In contrast, the other four tools could not provide such correct embedding.

Case 3: For the third example (Figure 4B), we used four read-count matrices of scATAC-seq profiles of 4 cell types, namely: Human lymphoblastoid cell line (GM12878), mouse B cells, Human embryonic kidney cell line (HEK293T) and cells from the mouse – proximal kidney tubules. For this example, scEpiSearch provided almost correct embedding showing colocalisation of human GM12878 cells and mouse B cells (Figure S9). However, other tools showed wrong co-localizations in the embedding result, such as in results from SCALE mouse B cell and Mouse Proximal tubule were colocalized. SCVI also wrongly colocalized Human GM12878 cells and mouse proximal tubule.

Case 4: To show the consistency of the module, we made queries from different batches and species of a larger number of cells Human Neuron (GSE97942), Mouse Forebrain (GSE1,00033), Human HSC (GSE96769), all mouse HSC from Bone marrow tissue (GSE111586) and passed these queries to Embedding module of the standalone system (supplemental Fig. S12A). The graph has been built by making an adjacency matrix ( using the top twenty matching clusters ), and a clear separability can be seen for both groups.

**Case study of MPAL:**
We performed embedding of scATAC-seq profiles (Figure 5) of cells from two patients with multiple phenotypes acute leukaemia (MPAL)( GEO id: GSE139369 ), peripheral blood mononuclear cells from healthy individuals (GEO id: GSE139369 ) and progenitors of cells in the blood (progenitors of hematopoietic cells) (GEO id: GSE96772), T cells (GEO id:GSE107817) and B cells (GEO id: GSE109828 ). Among progenitors, we used scATAC-seq profile of progenitors of hematopoietic cells in blood, namely MEP (megakaryocytic-erythroid progenitor), CMP (common myeloid progenitor), CLP (common lymphoid progenitor), GMP (granulocyte-monocyte progenitor), MCP (mast cell progenitor).

***Comparison of embedding with other tools***
Other methods meant for the analysis of scATAC-seq profiles do not search reference datasets to find matching cells. However, few researchers have proposed the possibility of embedding scATAC-seq profiles from different sources in a single visualization plot using their tools. We downloaded such tools and compared their performance in embedding scATAC-seq profiles from multiple sources. Here is the description of their sources and the parameters used.

SCALE : SCALE(Xiong et al. 2019) was downloaded from https://github.com/jsxlei/SCALE. It was run with parameters batch_size as 500, seed as 43, min_peaks as 400, lr as 0.0002, and max_iter as 500. After getting latent space representation for each query, we combined them and passed them to t-SNE to get the final plot.

SCVI : SCVI(Lopez et al. 2018) was downloaded from https://github.com/YosefLab/scvi-tools. We passed combined gene enrichment scores of all queries to SCVI, and the model learned the latent representation of query cells. We further passed these latent features to t-SNE to get a scatter plot.

SCANORAMA : SCANORAMA(Hie et al. 2019) was downloaded from https://github.com/brianhie/scanorama. For SCANORAMA also, we combined gene enrichment scores of all queries that were passed to scanorama.correct() function to get integrated form for cells.

MINT : MINT(Rohart et al. 2017) was used by installing mixOmics library in R (http://www.bioconductor.org/packages/release/bioc/html/mixOmics.html). In MINT gene enrichment scores of query-cells were passed to function mint.plsda() and plot was made using their function plotIndiv().

HARMONY **:** HARMONY was used by installing harmonypy library in Python (GitHub - slowkow/harmonypy: Integrate multiple high-dimensional datasets with fuzzy k-means and locally linear adjustments.). run_harmony() was called on combined gene enrichment scores of query cells.

**Calculation of clustering purity**

In order to evaluate the 2D embedding from multiple methods (Figure 4) we performed DBSCAN based clustering of the 2D coordinates of cells.. For eg., if celltypes present were Bcell and Tcell of human and mouse, n_clusters was 2.We used two measures to judge the correctness of embedding. The first measure is called as adjusted Rand index (ARI).

Let, $T = [t_1, \ldots, t_P]$ represents the true p classes consisting of $n_i$ number of observations in class $t_i$ and $V = [v_1, \ldots, v_K]$ be the clustering result with 'k' clusters having $n_j$ number of observations in cluster $v_j$. ARI is calculated as:

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{k}\binom{n_{ij}}{2} - [\sum_{j=1}^{p}\binom{n_i}{2}\sum_{j=1}^{k}\binom{n_j}{2}]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{j=1}^{p}\binom{n_i}{2} + \sum_{j=1}^{k}\binom{n_j}{2}\right] - [\sum_{j=1}^{p}\binom{n_i}{2}\sum_{j=1}^{k}\binom{n_j}{2}]/\binom{n}{2}} \quad (4)$$

Here, $n = \sum_{j=1}^{k} n_j = \sum_{i=1}^{p} n_i$

The second measure we used is called normalized mutual information (NMI), which is calculated as

$$\frac{2I(U,V)}{H(U)+H(V)} \quad (5)$$

Where I(U,V) is mutual information and H(U) and H(V) are the entropies of U and V are cluster labels. For ARI and NMI calculation, we used adjusted_rand_score() and normalized_mutual_info_score() functions from sklearn in python programming language.

**Analysis of the scATAC-seq profile of mESC**

For mESC scATAC-seq profiles, we mapped the reads to mm9 genome using bowtie2. Then we detected regions with non-zero read-count for every cell to call them "peaks". We combined the list of peaks from every cell to get a union list. Notice that we did not merge scATAC-seq profiles of mESC to call peaks, as it often leads to finding peaks only in major populations and loss of information about minor populations. After having a union list of peaks, we estimated the read-count on the peaks for every mESC cell. This led to a large read-count matrix which we used as a query for scEpiSearch.

For selecting the top 1,0000 peaks from each cluster, we used an average of normalized read-counts in cells in a cluster. The normalization of peaks was done using their global accessibility score. We

combined the reads of alignment files (bam files) of cells belonging to a cluster for further analysis. A wiggle track with a bin size of 200 bp was made using the combined read of cells belonging to a cluster. The read-count in bins of 200bp in the wiggle file was normalized according to the sequencing depth (number of reads) in bam file consisting of reads from cells belonging to a cluster, such the total normalized read-counts in bam of clusters is similar. The normalized wiggle files were used for visualization in the UCSC genome browser.

For making boxplots shown in Figure 6D and supplemental Fig. S16, first, the read-counts in 1 kbp up and down of TSS (total 2 Kb/TSS) of gene were estimated one single combined aligned read file (bam) for each cluster. Thus, we achieved a matrix of read-count of size G X 4 where G is the number of genes, and 4 is the number of clusters. We performed quantile normalization of read-counts such that for every cluster, we could have a similar distribution of read-counts on promoters. Then we dropped those genes promoters that did not have at least 5 reads in any cluster after quantile normalization. Thus we used the list of chosen promoters with more than 5 reads in at least one of the clusters, to avoid the effect of noisy read-counts while making boxplots for genes belonging to different gene-sets

## Supplemental Results

### Features implemented in scEpiSearch

**Interactive GUI and Provision of Data privacy:** scEpiSearch provides a very interactive graphical overview of results. The online version offers an interactive result web page that renders a clickable SOM plot for queries and a summary table that updates subsequent tables based on the query row user clicks on. Further, the tool provides an interactive summary plot for both matches. scEpiSearch also gives the convenience of a standalone tool where large queries can be submitted and benefits can be availed on privacy and confidentiality issues that might arise for medical datasets.

### Output and visualization of Results

The output of scEpiSearch is displayed in a scalable manner such that information about search results for multiple cells can be seen at a time. For both webserver and standalone versions a 2D scatter plot is shown, where every dot represents a query scATAC-seq profile. Every dot representing the query scATAC-seq profile is clickable, such that, on a click, it displays the detail about its matching results. Both matching single-cell expression and epigenome profile results are shown in separate panels. For every query cell, top matching results are shown with a p-value. Cell-type annotations, library ID and tissue, is shown for matching reference cell.

Following additional informations can be retrieved from scEpiSearch

**i) Enriched Genes:** scEpiSearch provides information of the top 50 enriched genes for every query cell. The tool also provides particulars if a gene is a marker for any cell type.

**ii) Hierarchical clustering results:** In the results section of scEpiSearch, there is a panel that shows the hierarchical clustering of query cells. The hierarchical clustering is based here on their matching index with reference cells. For this purpose, scEpiSearch first makes a union set of top K-matching reference cells for query scATAC-seq profiles. Then it makes a matrix whose elements represent matching index(P-value) between query scATAC-seq profile with a reference cell in union-set. The matching index matrix is of size N X M, where N is the number of query cells and M is the number of reference cells in union set-set. Using such a matching index matrix, it performs hierarchical clustering.

We also used scEpiSearch to find matched reference single-cell expression profiles for scATAC-seq read-count of peripheral blood mononuclear cells (PBMC) from healthy humans and achieved a result in concordance with the expected distribution of cell-types in the blood (supplemental Fig. S10A.

## Case study of query of unannotated single-cell epigenome profiles in Cusanovich et al. data-set

We applied scEpiSearch to find the closest match for "unknown" cells in the single-cell indexed ATAC-seq (sciATAC-seq) dataset published by Cusanovich et al.(2018) provided very relevant hits. Such as for the sciATAC-seq profile of unknown cells from mouse liver cells, 96% of the top 5 matching scRNA-seq profiles were from hepatocytes (Supplemental Fig. S7) and a majority of unknown cells from the prefrontal cortex had top matching cell-type as neurons. scEpiSearch also provides the list of gene enrichment scores for query cells to help researchers decide on representative genes (or markers) or develop the reliability of results. Such as markers for hepatocytes like *Apob, Tat, Cfhr2, and Mat1a* were present frequently among the top 20 enriched genes for "unknown" query cells from the liver (Supplemental Fig. S7). We have provided some more examples in supplemental Fig. S8 and S9.

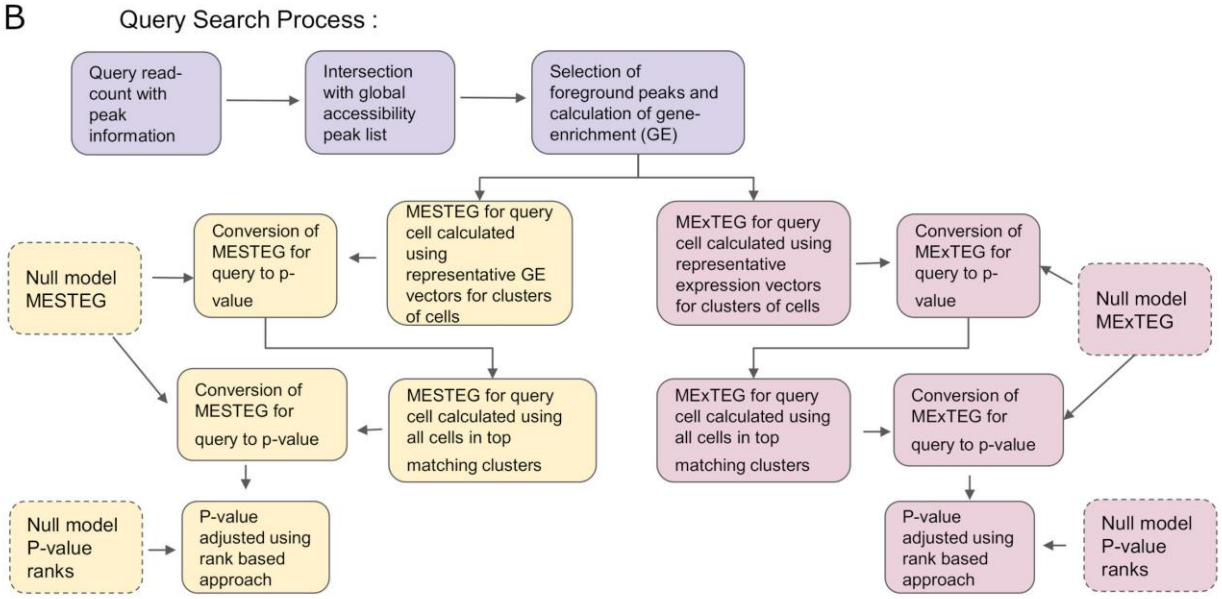
### Case study of query as scATAC-seq profiles from human PBMCs

Application of scEpiSearch for finding matched reference scRNA-seq profiles for scATAC-seq read-count matrixes of PBMCs from healthy humans provided top matches in concordance with the expected distribution of cell types in the blood (Supplemental Fig. S10A) (Kleiveland 2015).

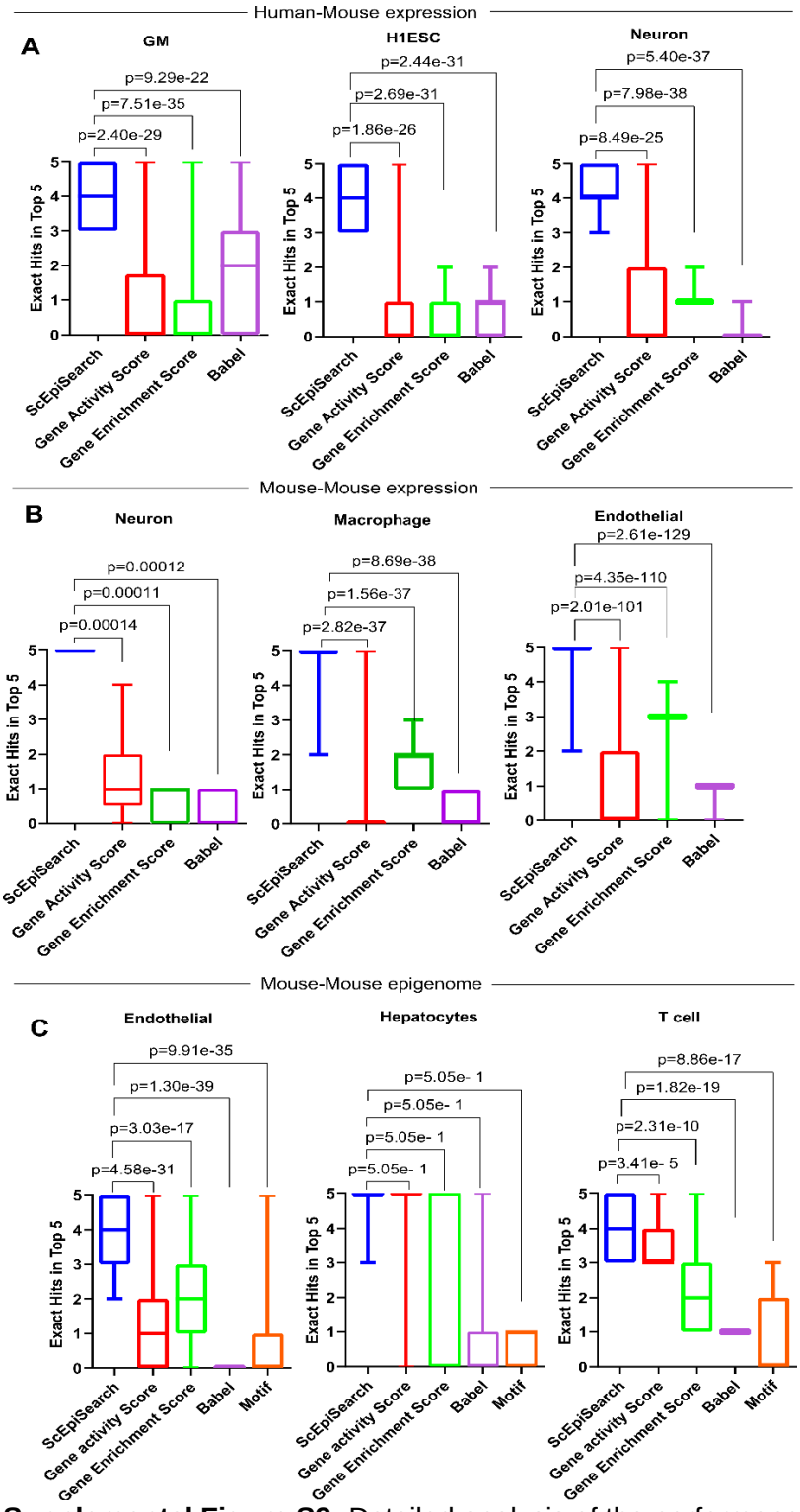
A

| Reference data-set | Number of cells |
|---|---|
| Single Cell Expression (Human) | 2239727 |
| Single Cell Epigenome (Human) | 742297 |
| Single Cell Expression (Mouse) | 2141797 |
| Single Cell Epigenome (Mouse) | 81173 |

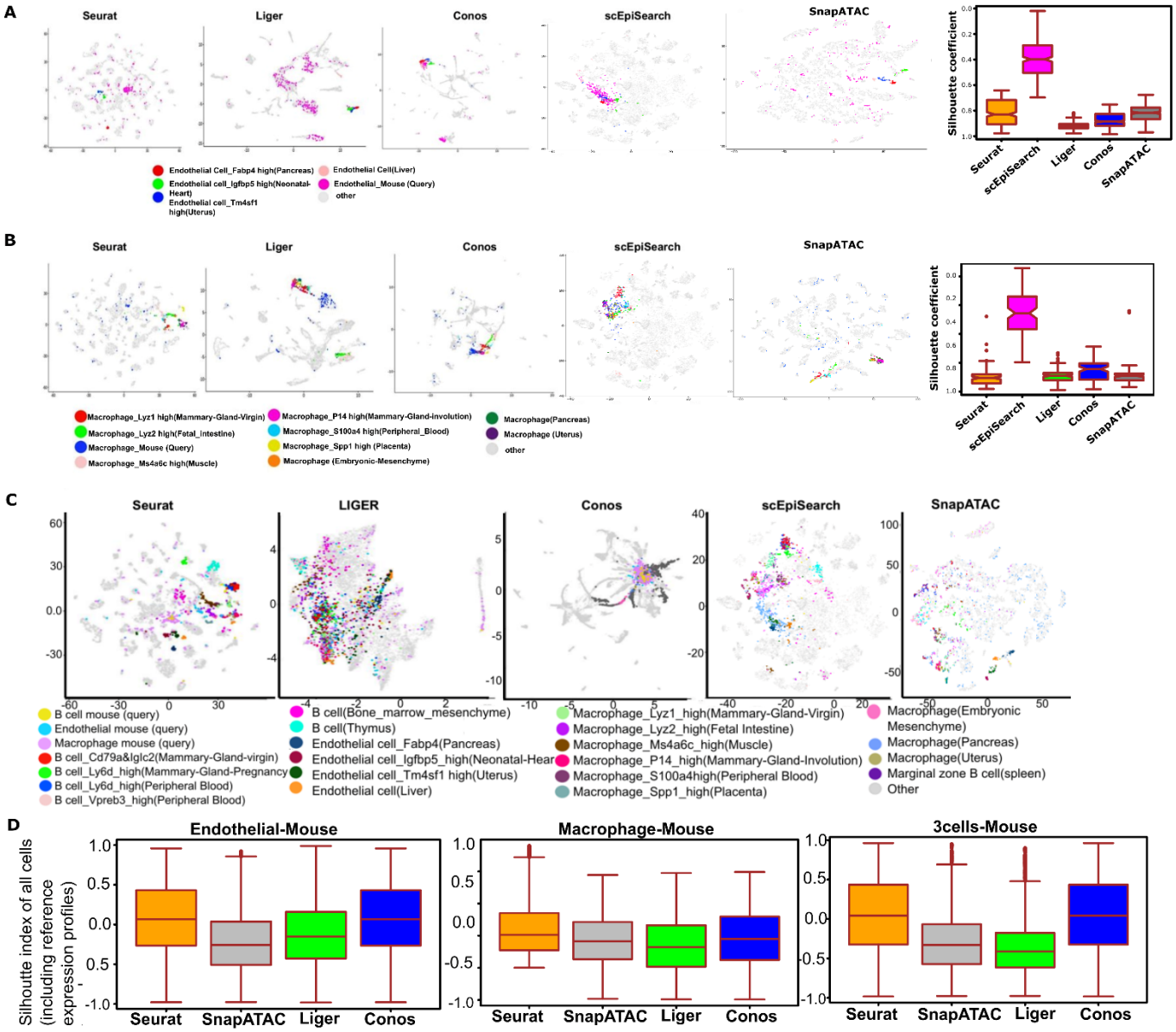
B



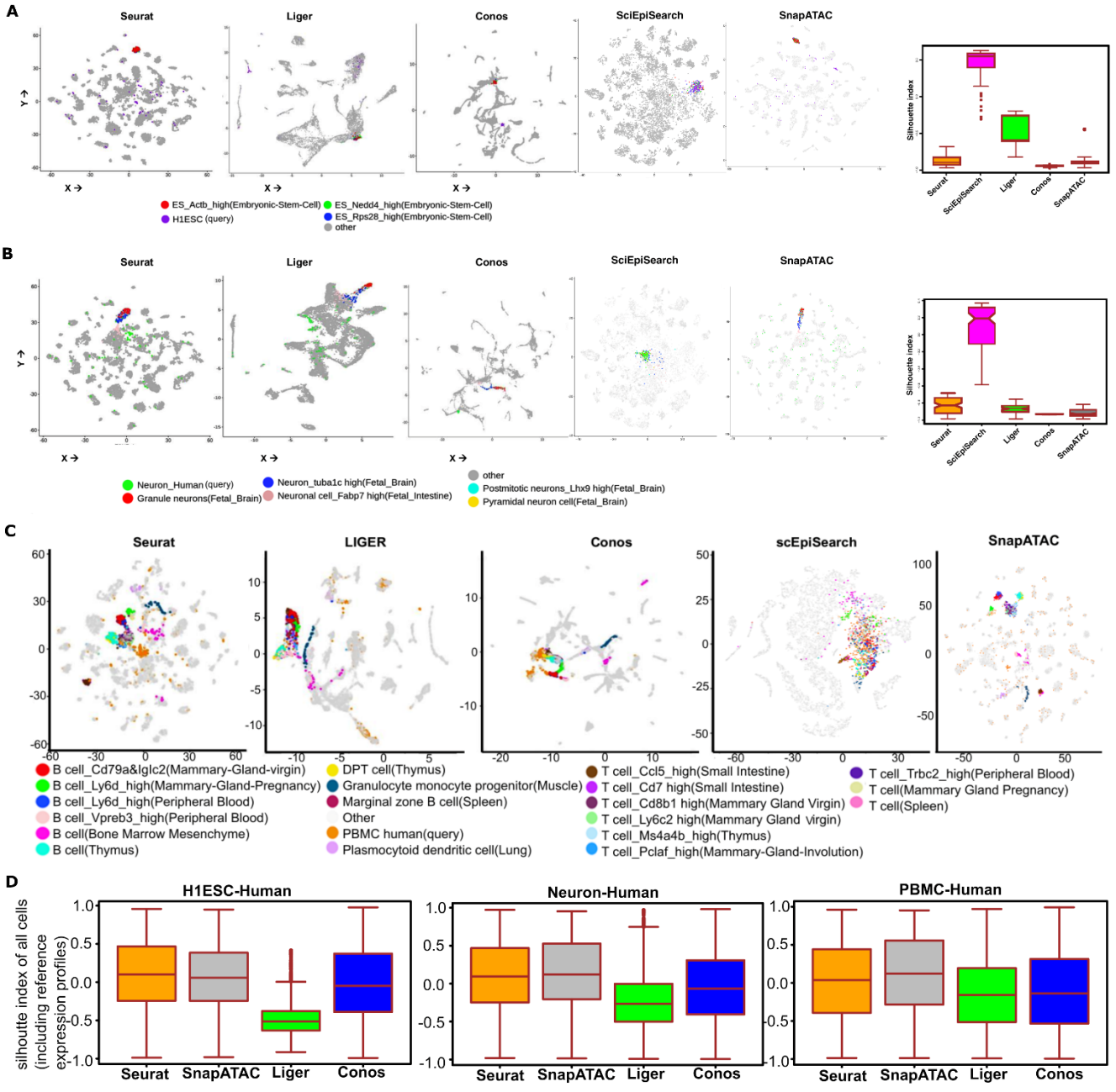

**Supplemental Figure S1: The overview of reference data-set of scEpiSearch and the approach to search it. A)** The current number of cells whose scRNA-seq and scATAC-seq profiles are in the reference dataset. **B)** Flowchart of Query processing by ScEpiSearch. It takes peaks and read-count matrix as input, finds proximal genes, and calculates gene-enrichment (GE) scores for cells. Further, it searches for matching reference epigenome and expression profiles. The part on the right after the step of finding GE scores shows steps for expression match, while the part on the left shows steps for searching similar single-cell open-chromatin profiles.



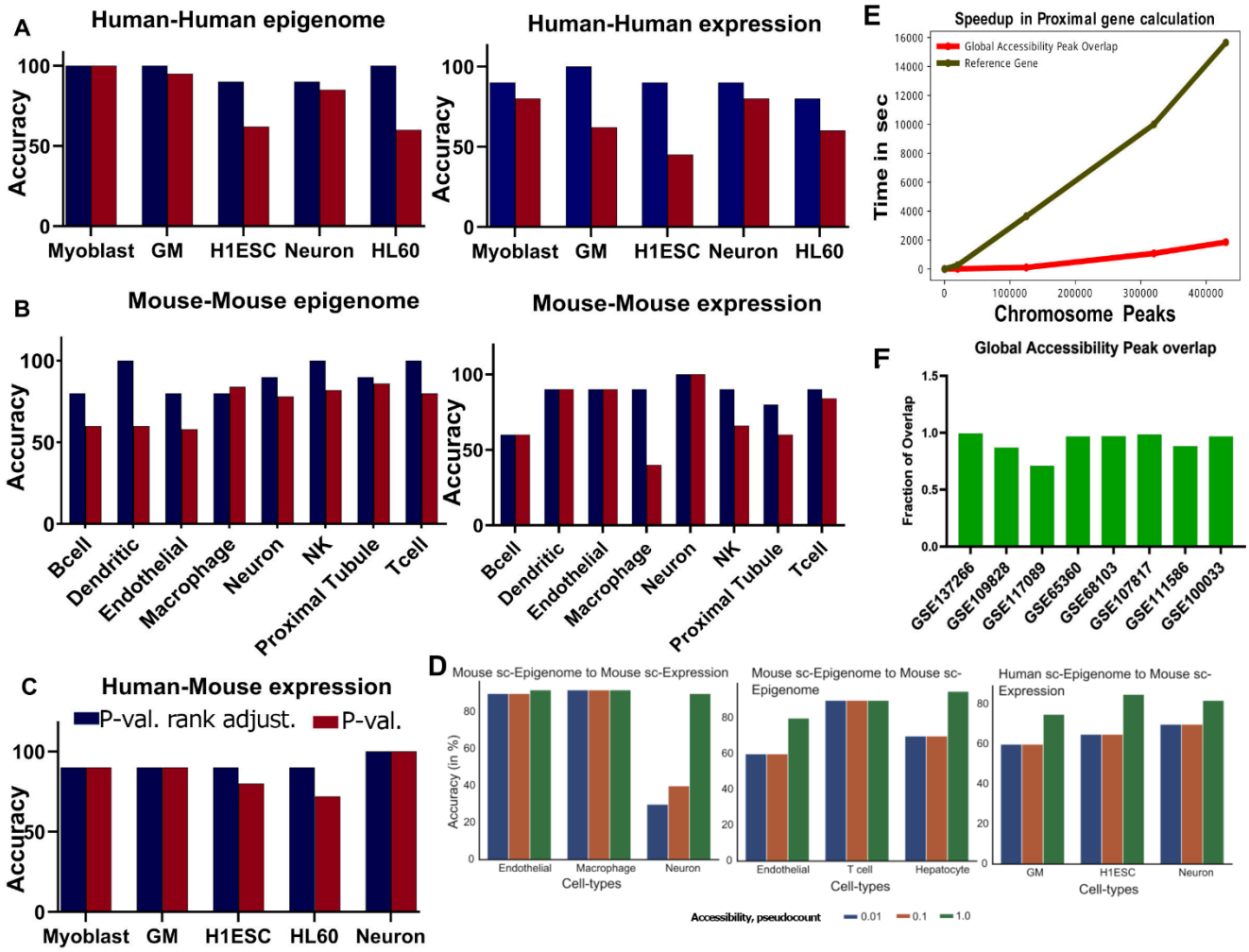

**Supplemental Figure S2:** Detailed analysis of the performance of different approaches to match query single-cell ATAC-seq with reference transcriptome and epigenome. A) Number of times the correct cell type was present in the top 5 matching reference cells are shown as boxplots. Here reference data-set consisted of 10,100 single-cell expression profiles from Mouse expression (MCA) and queries were scATAC-seq profiles of Human GM28178 (GM), H1ESC and Neuron cells B) Boxplots for number of correct result in top-5 its for Mouse-Mouse expression (MCA) for cell types Neuron, Macrophage, Endothelial C) Boxplots for number of correct result in top-5 for matching reference mouse epigenome to mouse epigenome profiles for cell types Endothelial, Hepatocytes and T cell.



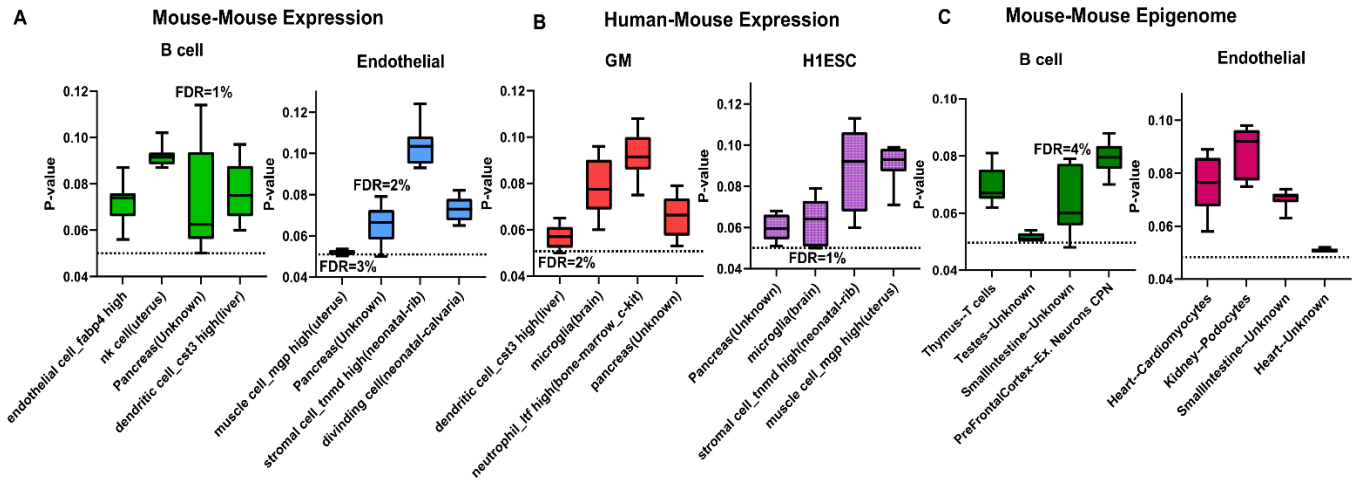

**Supplemental Figure S3: Comparison of scEpiSearch with integrative methods using reference expression and query single-cell epigenome profile from mouse cells.** The reference data-set here consists of single-cell expression profiles of 10,100 cells from mouse cell Atlas (Han et al. 2018). A) The query consisted of a single-cell ATAC-seq profile of mouse endothelial cells (Cusanovich et al., 2018). The silhouette coefficients /indexes of query cells for 4 methods are also shown. B) Here, the query consisted of epigenome profiles of mouse macrophages (Cusanovich et al., 2018). The silhouette coefficients (or indexes) for query celare is also shown. C) simplified version of Fig. 2C where reference cells the of same type the have same colour. D) Incorrect evaluation when silhouette coefficients for all cell, including reference mouse cell atlas cells, are shown in the 2D embedding figures. Corresponding labels on the sub-figures how : mouse endothelial cells (supplemental Fig. S3A) or mouse macrophage (supplemental Fig. S3B) or 3-cell combination (Fig. 2C, supplemental Fig. S3C ).



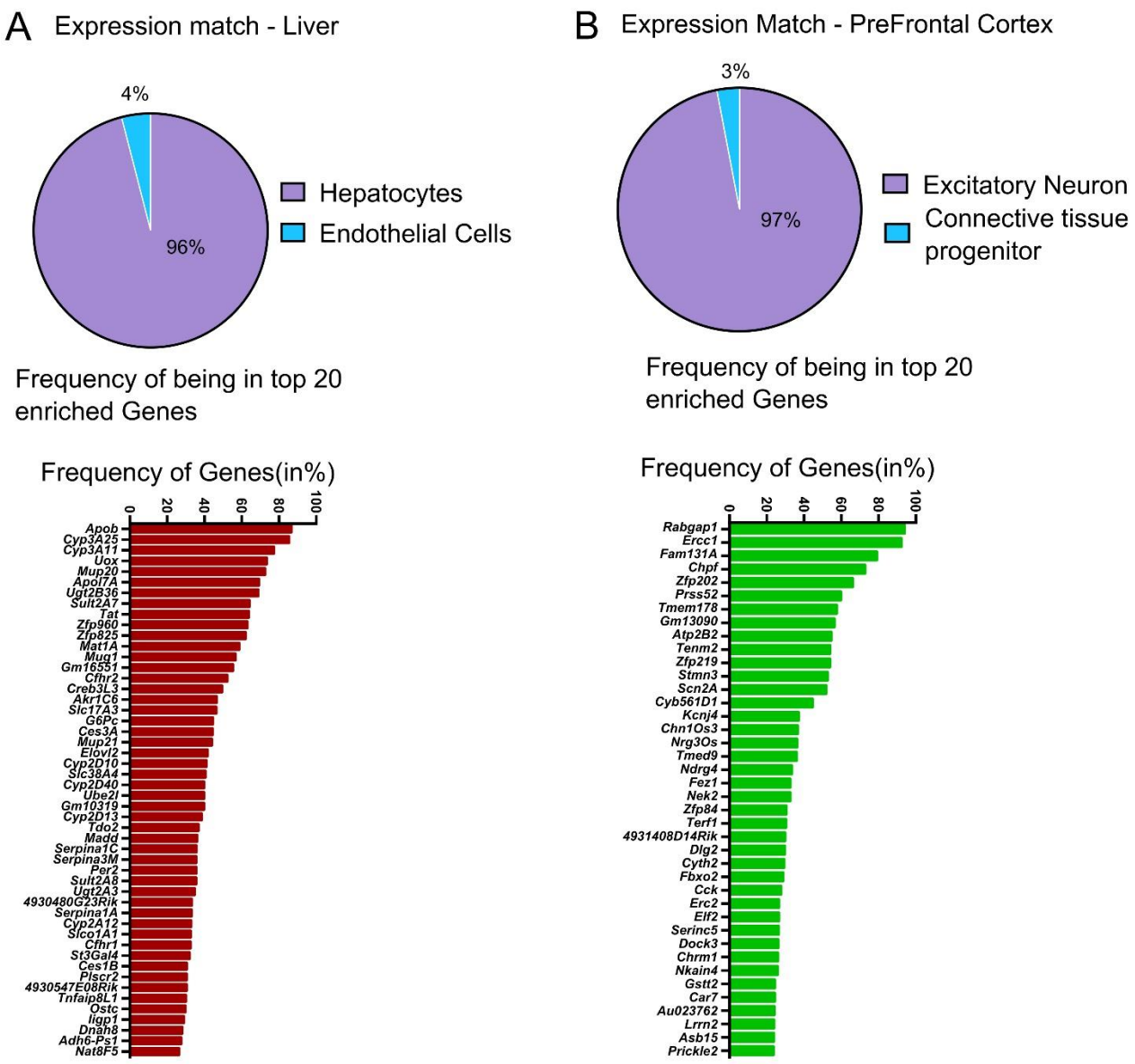

**Supplemental Figure S4: Comparison of scEpiSearch with integrative methods using reference expression of mouse cells and query as the scATAC-seq profile of human cells.** Here, the reference data-set consisted of 10,100 single-cell transcriptomes from mouse cell Atlas (Han et al. 2018). A) The query consisted of a single-cell ATAC-seq profile of Human embryonic stem cells (H1ESC). The silhouette indexes of query cells for 5 methods are also shown. B) Here the query consisted of single-cell epigenome profiles of human neuron cells. The silhouette coefficients for query cells are also shown. C) The simplified version of Fig. 2D where reference cells of the same type have same color. D) The silhouette coefficients (or indexes) for all cells, including reference mouse cell atlas cells in the 2D embedding figures shown. Corresponding labels on sub-figures show: when query single-cell ATAC-seq profiles belonged to H1ESC (Supplemental Fig. S4A) or human neurons (Supplemental Fig, S4B) or human PBMC (Fig. 2D, supplemental Fig. S4C).



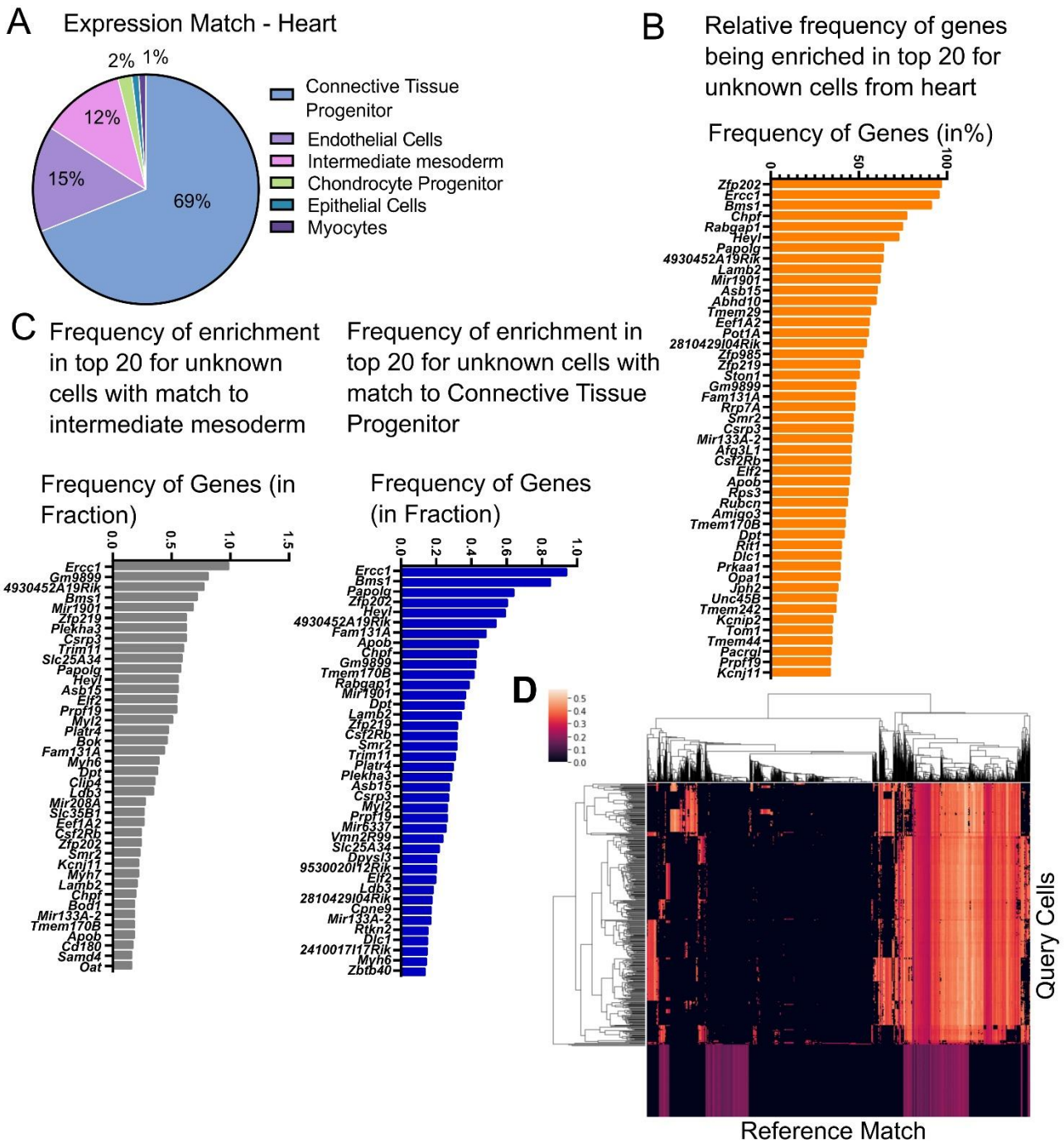

**Supplemental Figure S5**: Evaluation of different different steps of scEpiSearch A) Human-Human epigenome: matching single-cell epigenome profile of human cells to reference human single-cell open-chromatin dataset. The accuracy of results based on MESTEG based p-value and their rank based adjustment is shown. Similar results are shown for comparing query human single-cell open-chromatin profile to reference human single-cell transcriptome data-set (Human-Human expression) B) The perfromance based on p-value vs rank based adjustment is shown for Mouse-Mouse epigenome (comparing mouse single-cell epigenome with reference mouse open-chromatin profiles) and Mouse-Mouse expression (matching mouse single-cell epigenome with reference mouse single-cell transcriptome) C) Human-Mouse epigenome: matching Human single-cell epigenome with reference mouse single-cell data-set. The p-value vs rank based adjustment based results are shown. D) The accuracy of search result, with different values of pseudocount (Ɛ) for normalisation with accessibility score in equation (1). E) Speedup achieved in proximal gene calculation by ScEpiSearch. The green line shows the standard way of finding the proximal gene file (reference Gene) (*accurate*), while the red line shows the method utilizing overlap of query peaks with global accessibility peak-list (*Faster*). F) Fraction of peaks overlapping with pre-compiled global accessibility peak list is shown. X-axis shows the data repositories (GEO id) from where query scATAC-seq profiles were made and Y-axis shows the fraction of overlap of their peaks with peak-list of global accessibility table.



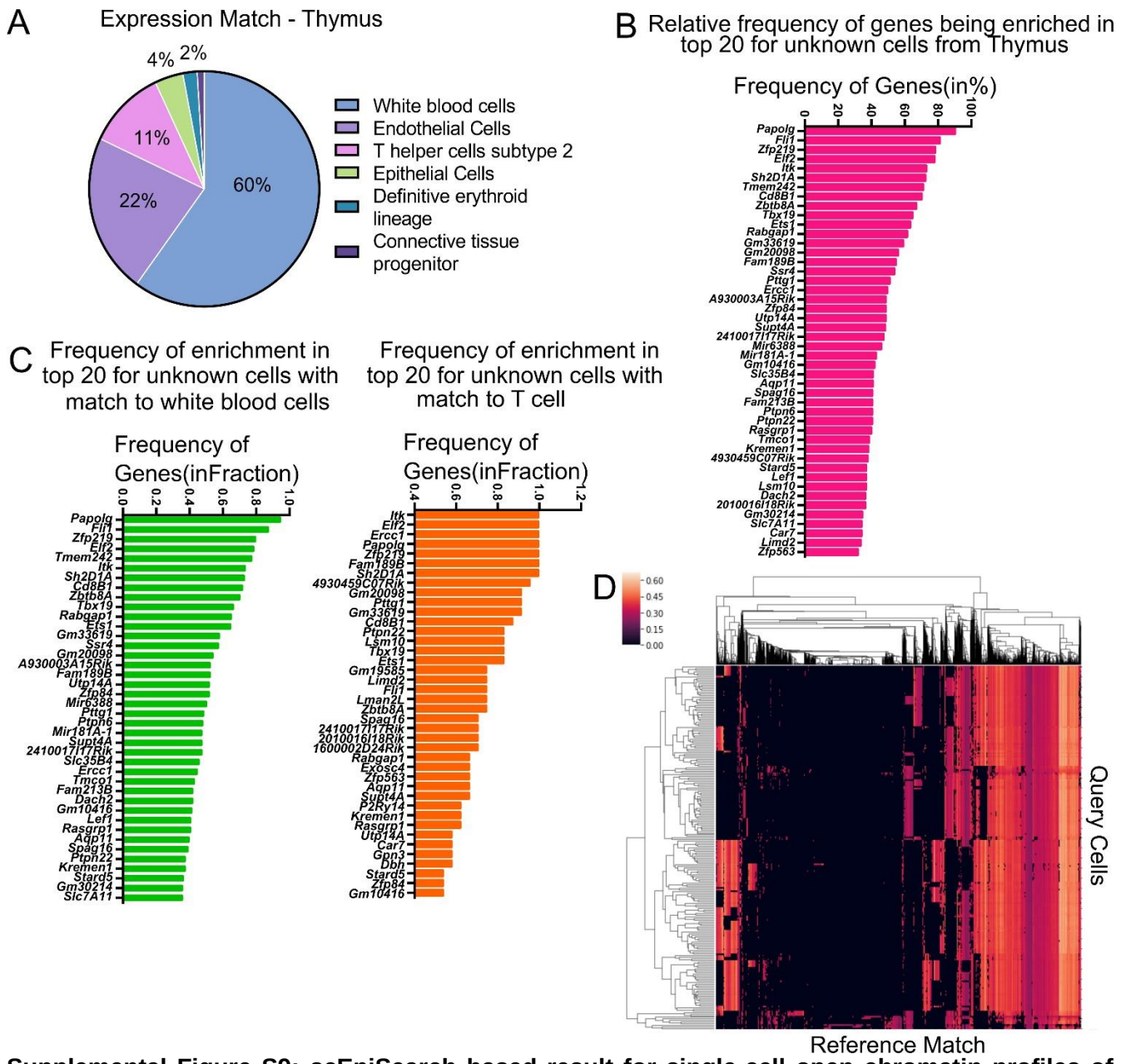

**Supplemental Figure S6:** Results of control experiment where the cell-type relevant to query scATAC-seq profiles were removed from reference data-sets. A) Top hits P-value (rank adjusted) for query scATAC-seq profiles of mouse B cells and Endothelial cells in reference consisting of mouse (MCA) expression where relevant cells (B cells and endothelial cells) were removed. The P-values of match are not significant (i.e. P-value > 0.05). The dotted line shows P-value = 0.05. Whereever the query result had p-value less than 0.05, the fraction of results with P-value less than 0.05 is mentioned as FDR (false detection rate). B) Top hits P-value (rank adjusted) for query scATAC-seq profile of human H1ESC, GM12878(GM) cells in reference mouse (MCA) expression without relevant matching cells. C) Top hits P-value (rank adjusted) for Mouse B cell, Endothelial scATAC-seq profiles in reference data-set of mouse single-cell epigenome where relevant cells were removed. In the absence of relevant cell-types in reference data-set, the matching top profiles not truly similar to query cells have non-signficant p-value (i.e. P-val > 0.05).



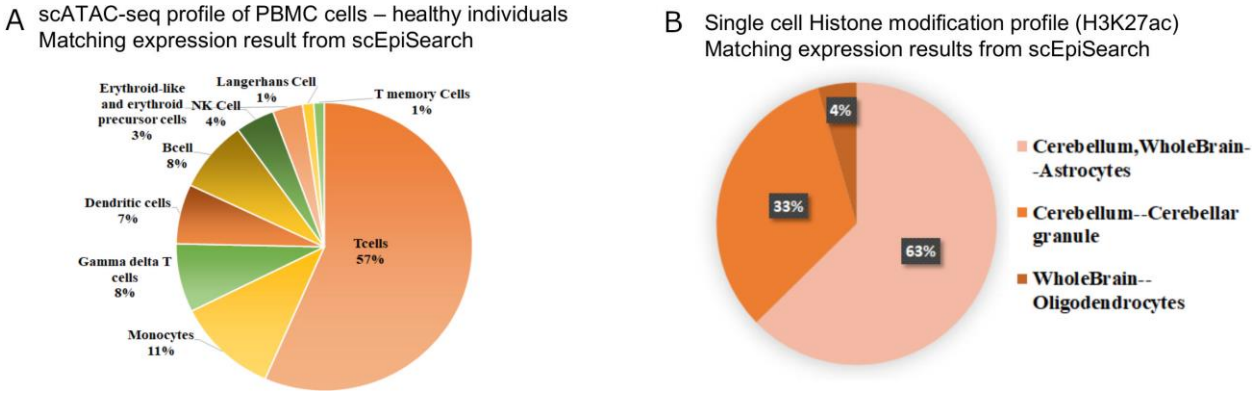

**Supplemental Figure S7: Application of scEpiSearch on scATAC-seq profile of "unknown" cells from mouse organs A)** Pie-chart showing the distribution of top 5 matching single-cell expression profiles to single-cell epigenome profiles of unannotated cells (with "unknown" label) from in vivo samples of adult mouse liver. The bar-plot shows the frequency of appearance in the top 20 enriched genes for query cells (with "unknown label") from the liver. **B)** distribution of top 5 matching single-cell expression profiles to single-cell epigenome profiles of unannotated cells (with "unknown" label) from the prefrontal cortex of adult mice. The bar-plot shows the frequency of being among the top 20 enriched genes for query cells (with "unknown label") from the prefrontal cortex.



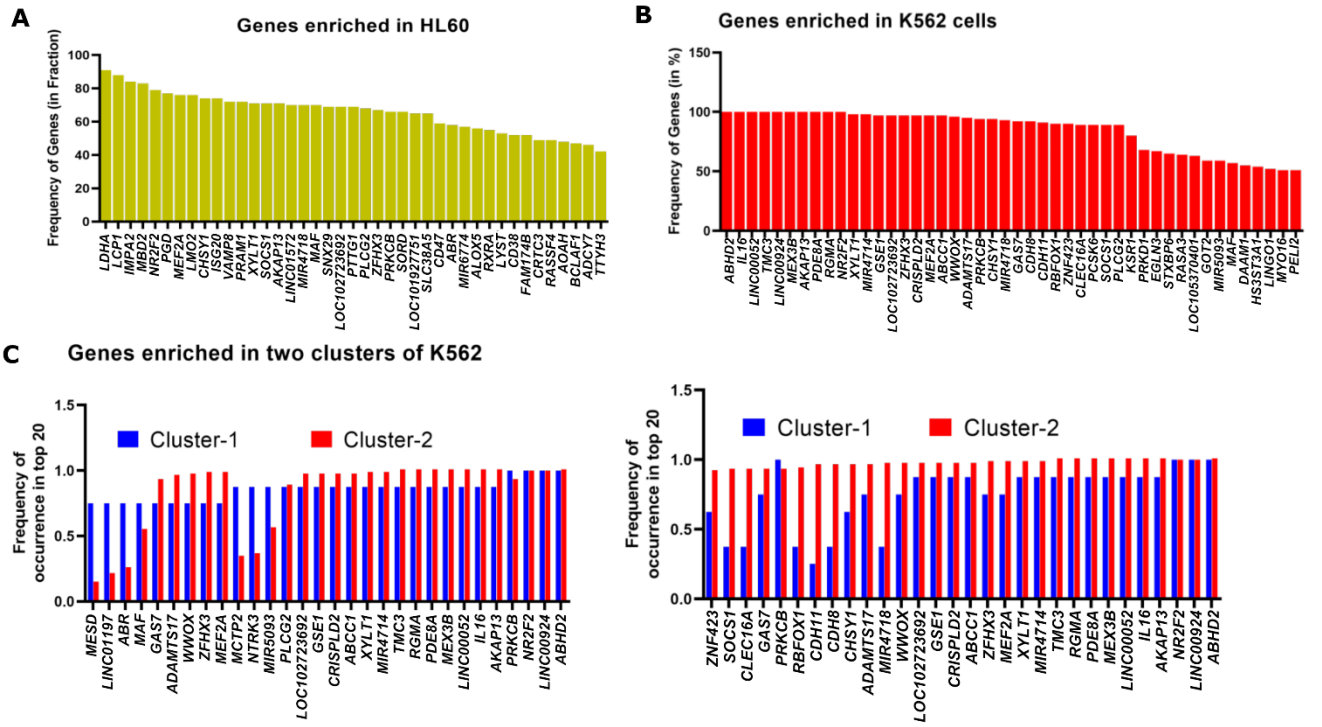

**Supplemental Figure S8: scEpiSearch based result for single-cell open chromatin profiles of unknown cells from mouse Heart tissue (Cusanovich et al., GSE111586)** **A)** Pie chart of the phenotype of top five matches with p-value<0.05 in mouse expression. The connective tissue cells of heart are known as cardiac stromal cells also include mesenchymal stromal cells.(Stadiotti et al. 2020). **B)** Frequencies of appearance in the top twenty enriched genes in all query cells are shown as a bar plot. **C)** Frequencies of appearance in top 20 enriched genes in query cells with three major phenotypes detected based on expression match, namely: Connective tissue progenitor (Blue), endothelial cells (orange), Intermediate mesoderm (grey). **D)** Heatmap made using scores of the match between query and reference epigenome profiles frequently appearing as top hits. The X-axis shows matching epigenome profiles from reference, and Y-axis shows Query cells. Hierarchical clustering of query scATAC-seq and reference cells is also shown.



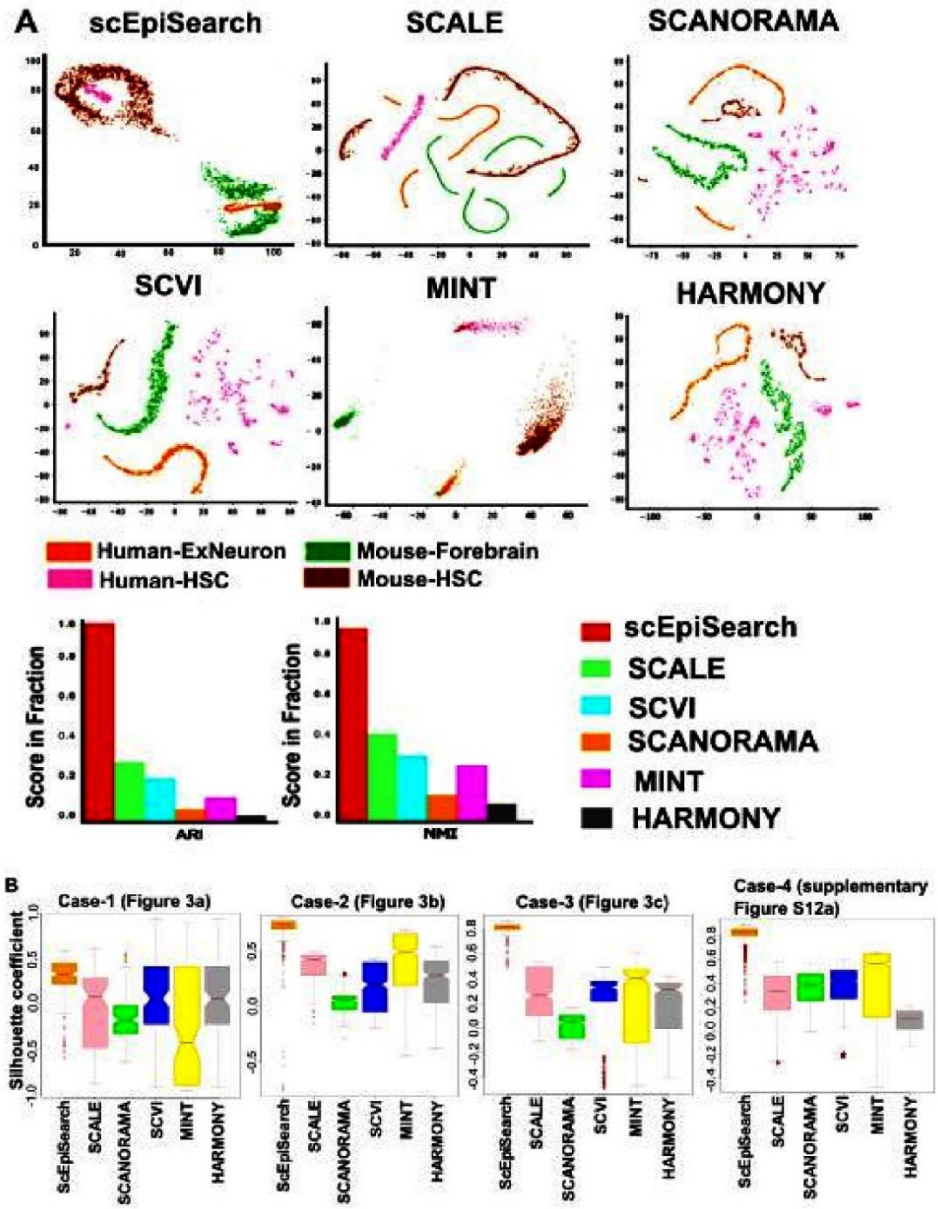

**Supplemental Figure S9: scEpiSearch based result for single-cell open chromatin profiles of single-cells with 'unknown' annotation from mouse (GSE111586) Thymus tissue**. **A)** Pie chart of the phenotype of top five matching reference expression profiles from mouse cell, (with p-value<0.05) for all queries (Unknown cells from thymus). Almost 70% of matching single-cell transcriptome profiles are from white blood cells, including T cells. The thymus is known to be instrumental in the production and maturation of T-lymphocytes (a type of white blood cell) **B)** Frequency of appearance in top twenty enriched genes in all query cells is shown as bar-plot. The X-axis shows genes whose fraction of frequency occurrence is shown on the Y-axis. **C)** Frequency of appearance of genes in the top twenty enriched in all query cells is shown for three major phenotypes in expression match i.e white blood cells (Blue), endothelial cells (orange), T helper cells (grey). **D)** Heatmap of match-scores of epigenome profile matching is shown. Here columns represent top-matching epigenome profiles from reference, and rows represent query cells.



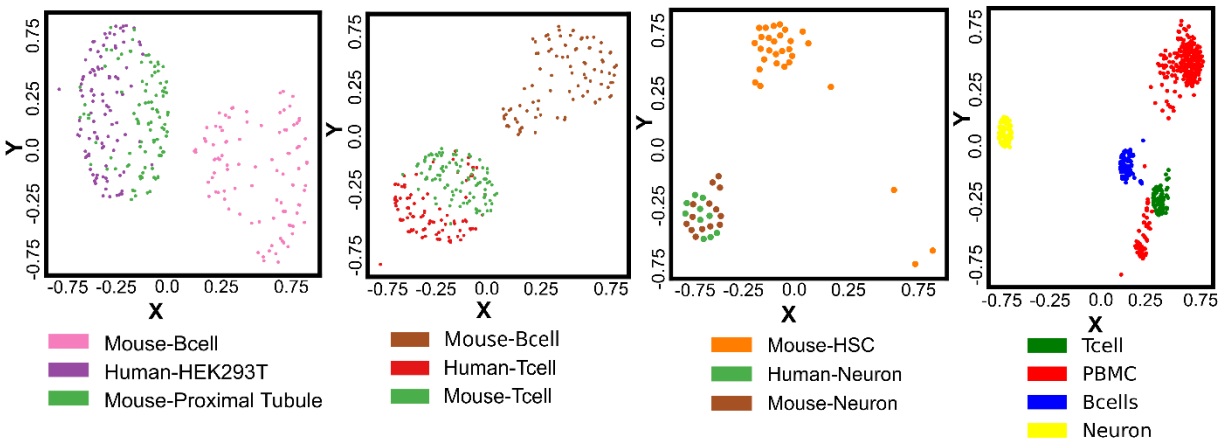

**Supplemental Figure S10:** The pie-chart showing cell types of top matching single-cell expression with reference single-cell epigenomes of human cells. **A)** The result is shown for peripheral blood mononuclear cells (PBMCs) (GEO id: GSE139369) **B)** Result of using scEpiSearch on single-cell histone modification (H3K27ac) profile of Astrocytes cells. (GEO id: GSE157637).



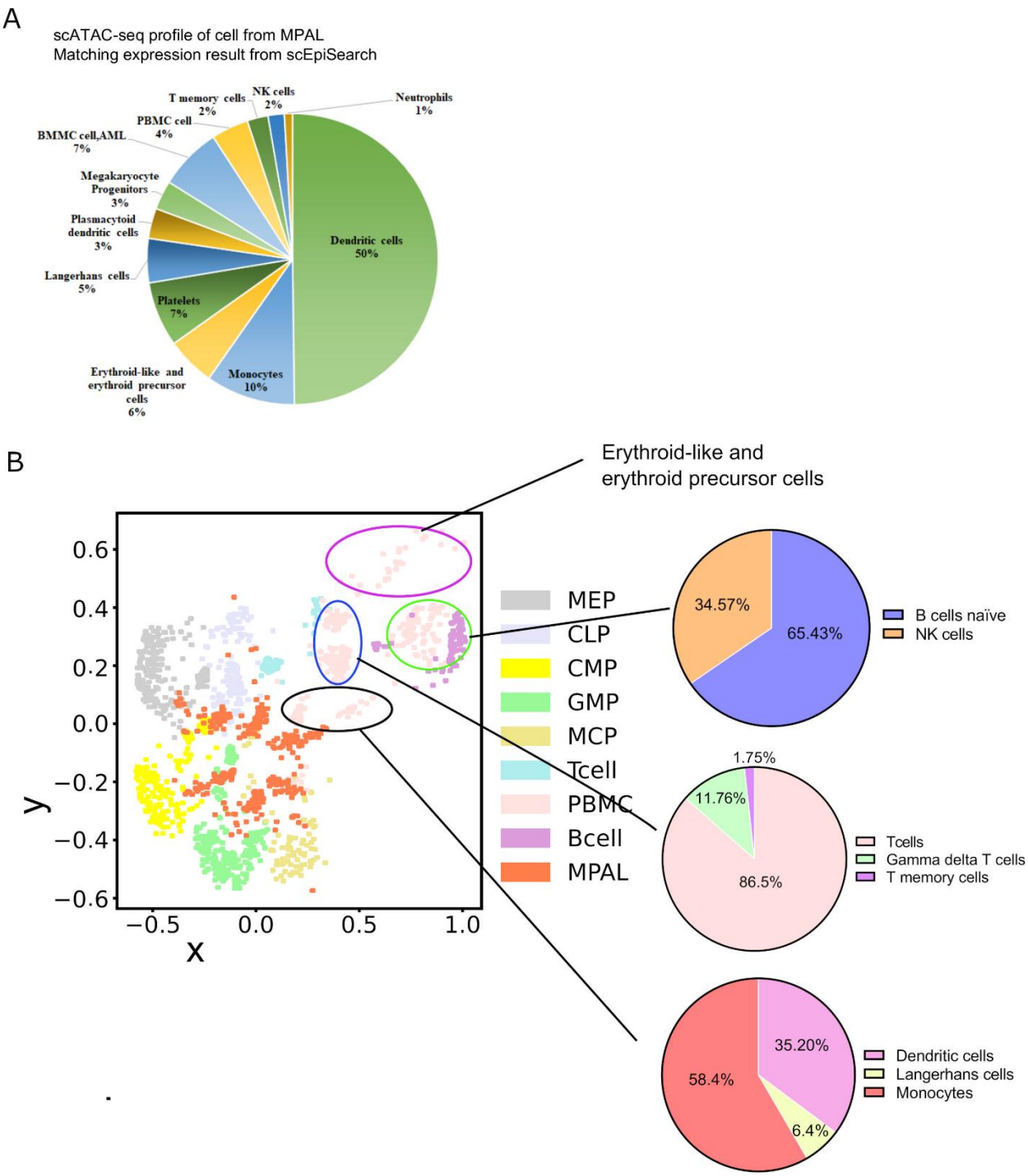

**Supplemental Figure S11: Frequency of being top 50 enriched genes for query scATAC-seq profile of cancer cell lines.** A) for 100 HL60 cells B) for 100 K562 cells C) the frequencies are shown for selected genes for K562 cells belonging to two classes (cluster-1 and cluster-2).



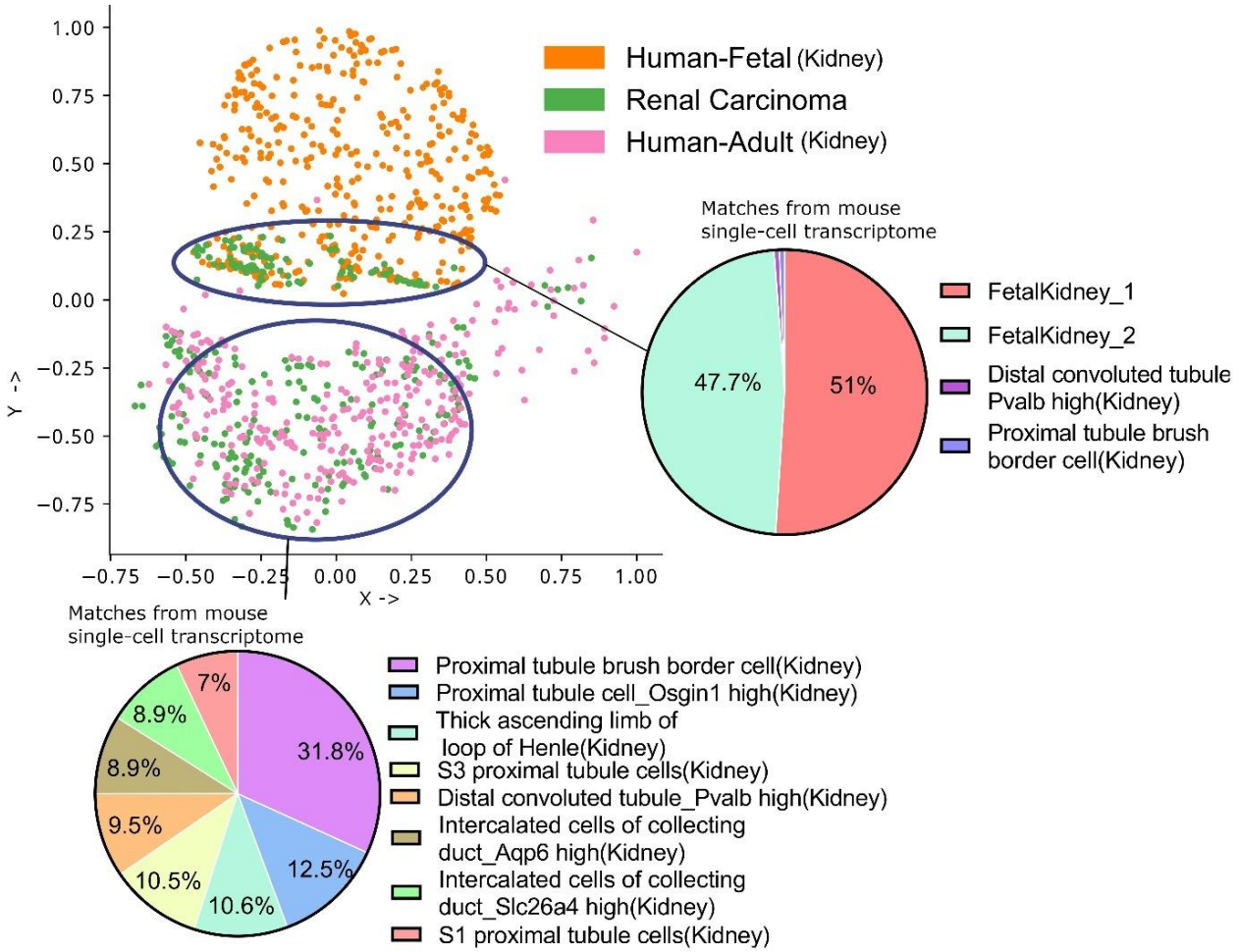

**Supplemental Figure S12:** Evaluation of 2D embedding of query single-cell ATAC-seq profiles from different species and batches. A) Here, queries are made for Human-Neuron, Mouse-Neuron, Human-HSC, Mouse-HSC. None of the other methods evaluated here could provide correct low-dimensional embedding like scEpiSearch. Graph is built using networkx python library. The t-SNE plot of latent representation derived from SCALE is shown. The t-SNE plot of latent representation derived from SCVI is shown. MINT plot is made using their function plotIndiv(). The t-SNE plot is made for SCANORAMA using an integrated representation of query cells given by the package. The right-bottom panel shows clustering-purity in terms of ARI and NMI scores after applying DSCAN on the 2D coordinates (embedding results) using two labels (HSC and forebrain/neruons). B) For evaluation of 2D embedding, box-plots were also made using calculatted silhouette coefficients. The case studies (and corresponding embed plot figure) is also mentioned.



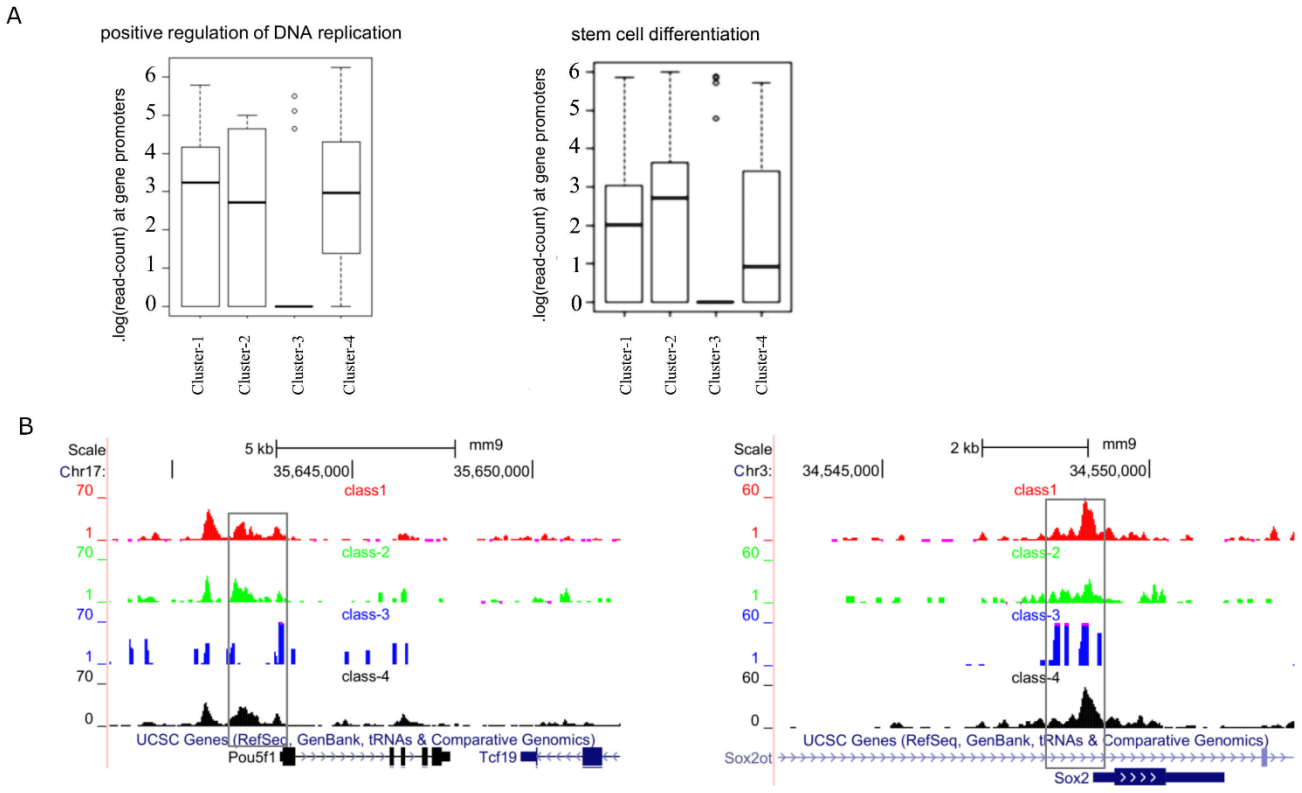

**Supplemental Figure S13: Additional testing of co-embedding function of scEpiSearch.** Here a non-matching cell-type population was used to test if scEpiSearch keeps it separate from other cell-types. Such as B cell should not co-embed with kidney cells and T cells. Neuron cells should be separate from HSC or PBMC and T cells and B cells.




**Supplemental Figure S14: Analysis of scATAC-seq profile of MPAL cells and their coembedding with other cell types. (A)**The pie-chart showing cell-types of top matching single-cell expression for scATAC-seq profile of blood cells from patients with multiple phenotypes acute Leukaemia (MPAL) (GEO id: GSE139369). **(B)** The distribution of cell-types for top-5 matching transcriptional profiles for different groups of PBMC cells. The PBMCs colocalising with B cells have top hit mostly as B cell. While PBMCs near T cells have top matching transcritional profile as T cell.




**Supplemental Figure S15:** A case study of matching single-cell open chromatin profiles of kidney cancer cells to other single cell epigenomes and reference mouse transcriptome by scEpiSearch. The co-embedding of single-cell open chromatin profiles of Papillary renal carcinoma cells(pRCC) from Wang et al. (GSE166547) and fetal Human Kidney (GSE149683 (Domcke et al. 2020)) and adult human kidney (GSE151302 (Muto et al. 2021)) cells. The pie charts show the distribution of the source of matching single-cell profiles for Renal Carcinoma cells overlapping two groups of cells (Fetal kidney and adult kidney), found by scEpiSearch while comparing to all cells from mouse reference transcriptome. The renal kidney cells overlapping with human fetal kidney also had top matching mouse single-cell transcriptome profile from fetal mouse kidney. While the renal kidney cells overlaping with adult human kidney cells had most of the matches from adult muse kidney. The pie chart for renal kidney cells overlapping with adult kidney also shows cell-types from mouse kidney.




**Supplemental Figure S16: The visualization of read-counts at promoters belonging to different types of genes as controls for results in Fig. 6. A**) The read-counts at the promoter of genes belonging to gene-set for two biological functions, namely "positive regulation of DNA replication" and "stem cell differentiation". B) The snapshot of UCSC genome browser shows a similar activity level among the four classes of cells at promoters of *Oct4 (Pou5f1)* and *Sox2* genes.

## Supplemental Tables

**Supplemental Table S1**: Information about sources of reference single-cell expression and open-chromatin datasets. Information is provided for data-sources for both species human and mouse.

**Uploaded as a separate xls file**

**Supplemental Table S2 :** The details of results from scEpiSearch for matching reference cells from Human for query scATAC-seq profile of Human cells.

| Cell-type Query | URL (after http://reggen.iiitd.edu.in:1207/episearch/?job=) | Source GEO ID | Percentage of query with True positive reference-expression in top 5 matches. | Percentage of query with True positive reference-scATAC-seq in top 5 matches. |
|---|---|---|---|---|
| **HL60** | http://reggen.iiitd.edu.in:1207/episearch/?job=-ntlgsuqp4-utvlgmnqx | GSE109828 | Mainly myeloid lineage cells<br>Monocyte-derived macrophages,Langerhans cells ( 80% )<br>Monocyte-derived macrophages,Unknown (70 %)<br>Circulating tumor cells in hepatocellular carcinoma,Dendritic cells (6%)<br>Circulating tumor cells in hepatocellular carcinoma,Monocytes (70 %)<br>Circulating tumor cells in hepatocellular carcinoma,Langerhans cells (2%)<br>Bone marrow,Monocytes (2%) | HL60 ( 90 %) |
| **Myoblast** | http://reggen.iiitd.edu.in:1207/episearch/?job=-ggbqir0do-hlpsn5gn7, http://reggen.iiitd.edu.in:1 | GSE109828 | Myoblast (80%) | Myobast (100%) |

| | | | | |
|---|---|---|---|---|
| | 207/episearch/?job=-ag5z14ze5-v2xf77zlf | | | |
| **GM12878** | http://reggen.iiitd.edu.in:1207/episearch/?job=-vgptvbeci-dlcdcztlk | GSE109828 | Lymphoblastoid cell line,B cells (80 %) Lymphoblastoid cell line,Plasma cells (20%) | GM12878 ( 100 %) |
| **H1ESC** | http://reggen.iiitd.edu.in:1207/episearch/?job=-206f3vcny-33j40evt3 | GSE65360 | Embryoid body,Unknown (87%) | H1ESC (70%) |
| **Neuron** | http://reggen.iiitd.edu.in:1207/episearch/?job=-r9sdpeg5i-pgjxb2nw8 | GSE97942 | Adult,tissue: cortex cell type: neurons (60%) Adult,tissue: cortex cell type: fetal_quiescent (100%) Glioblastoma cell,cell type: Glioblastoma (70%) | Excitatory neuron(80%) Inhibitory Neuron(20%) Cerebrum_Unknown.3 (10%) |

**Supplemental Table S3 :** The details of results from scEpiSearch for matching reference cells from mouse for query scATAC-seq profile of Mouse cells.

| **Celltype Query (mouse)** | **URL http://reggen.iiitd.edu.in:1207/episearch/?job=** | **Source GEO ID:** | Percentage of query with True positive reference-expression in top 5. | Percentage of query with True positive reference-scATAC-seq in top 5. |
|---|---|---|---|---|
| **Neuron** | http://reggen.iiitd.edu.in:1207/episearch/?job=-puamhnftl-zkm98ya79 | GSE111586 | Excitatory neurons (100%) | WholeBrain, Cerebellum --Inhibitory neurons (69.23 %) PreFrontalCortex--Ex. neurons SCPN, CThPN, CPN (84.6 %) |

| | | | | |
|---|---|---|---|---|
| | | | | Cerebellum- - Inhibitory neurons (10.7%%) |
| **Endothelial** | http://reggen.iiitd.edu.in:1207/episearch/?job=-6n9foszqj-jgs0mmstl | GSE111586 | Endothelial cells (90 %) | Heart,Kidney,Heart, WholeBrain,Lung,Liver - - Endothelial I cells (100 %) BoneMarrow--Hematopoietic progenitors ( 35.7 %) |
| **Dendritic** | http://reggen.iiitd.edu.in:1207/episearch/?job=-pp08s9ztk-cp31ehp5c | GSE111586 | White blood cells (90 %) | Lung,BoneMarrow,Spleen --Dendritic cells (100 %) BoneMarrow--Hematopoietic progenitors (50 %) |
| **Macrophage** | http://reggen.iiitd.edu.in:1207/episearch/?job=-leh9jpjui-f7uzo109j | GSE111586 | White blood cells (90%) | Heart,Spleen,Lung,Liver,Lung,Kidney,BoneMarrowLargeIntestine,Thymus--Macrophages ( 92.3 %)<br><br>BoneMarrow--Hematopoietic progenitors (38.46%) |
| **NK cell** | http://reggen.iiitd.edu.in:1207/episearch/?job=-dppduwmaj-9d80zgmvf | GSE111586 | White blood cells (90 %) | Lung, Spleen, BoneMarrow -NK cells (100 %) |
| **T cell** | http://reggen.iiitd.edu.in:1207/episearch/?job=-l2wb6vhcw-pc84v1q1i | GSE111586 | White blood cells (100 %) T helper cells subtype 2 (71.4 %) naive T cells activated (with anti-CD3/CD28) and polarized towards the Th2 subtype with IL4 (2 %) | Spleen,Lung,Thymus,Spleen,Li=ung --T cells ( 100 %) BoneMarrow--Hematopoietic progenitors (3 %) |

| | | | | |
|---|---|---|---|---|
| **HSC-hematopoetic stem cells** | http://reggen.iiitd.edu.in:1207/episearch/?job=-ks5c8qnuc-0q64t7rf1 | GSE111586 | White blood cells (100%) | BoneMarrow--Hematopoietic progenitors (96.6%) |

**Supplemental Table S4:** The details of results from scEpiSearch for matching query scATAC-seq profile of Human cells to mouse cell expression profile.

| **Celltype Query** | **URL http://reggen.iiitd.edu.in:1207/episearch/?job=** | **Source GEO ID:** | Percentage of query with True positive reference-expression in top 5 matches. |
|---|---|---|---|
| Myoblast | http://reggen.iiitd.edu.in:1207/episearch/?job=-km4m8sm1n-by3oshify | GSE109828 | Myocytes( 90%)<br>Osteoblasts ( 33.3%)<br>Intermediate Mesoderm (25%) |
| **GM12878** | http://reggen.iiitd.edu.in:1207/episearch/?job=-uhibbslst-d5c8lzijd | GSE109828 | B cell (90%)<br>White blood cells (90%) |
| **Neuron** | http://reggen.iiitd.edu.in:1207/episearch/?job=-uuo33h9um-g5bgybhol | GSE97942 | Excitatory neurons (100%) |
| **HL60** | http://reggen.iiitd.edu.in:1207/episearch/?job=-ljwz1w0ml-whrhfh401 | GSE109828 | White blood cells (90%) |
| **H1ESC** | http://reggen.iiitd.edu.in:1207/episearch/?job=-p880gctre-yhrfim47c | GSE65360 | Embryonic stem cells (90%) |

**Supplemental Table S5 :** Information about matching human single-cell expression profiles with query scATAC-seq read-counts matrices for K562 and HL60 cells.

**Uploaded as separate xls file**

**Supplemental Table S6 :** The enriched Gene Ontology terms for cluster-specific peaks for different cluster of mESCs

**Uploaded as separate xls file**

**Supplemental Table S7 :** The result of runtime benchmarks (with single CPU core) for scEpiSearch and 4 other integrative methods, using the same reference and query datasets. The reference consisted of 300 single-cells, and was queried against 10,100 single-cells from Mouse Cell Atlas. For preprocessing of reference data-set, scEpiSearch for took 1min39s and 72734.09 MiB memory.

| Method | Time Taken (min: sec) | Memory (in kb) |
|---|---|---|
| ScEpiSearch | 2:48 | 1,379,676 |
| Seurat | 5:33 | 8,781,056 |
| Conos | 2:28 | 4,174,576 |
| LIGER | 4:25 | 14,141,348 |
| SnapATAC | 2:05 | 18,376.77 MiB with GPU |


Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, Kharchenko PV. 2019. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat Methods* **16**: 695-698.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-C, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. 2005. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic acids research* **33**: D562-D566.

Bujold D, de Lima Morais DA, Gauthier C, Côté C, Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T. 2016. The international human epigenome consortium data portal. *Cell systems* **3**: 496-499. e492.

Chawla S, Samydurai S, Kong SL, Wu Z, Wang Z, Tam WL, Sengupta D, Kumar V. 2021. UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic acids research* **49**: e13-e13.

Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH et al. 2020. A human cell atlas of fetal chromatin accessibility. *Science* **370**.

Hie B, Bryson B, Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* **37**: 685-691.

Kleiveland CR. 2015. Peripheral Blood Mononuclear Cells. In *The Impact of Food Bioactives on Health: in vitro and ex vivo models*, doi:10.1007/978-3-319-16104-4_15 (ed. K Verhoeckx, et al.), pp. 161-167, Cham (CH).

Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, Hinrichs AS, Lee BT, Nassar LR, Powell CC. 2020. UCSC Genome Browser enters 20th year. *Nucleic acids research* **48**: D756-D761.

Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. 2020. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature protocols* **15**: 3632-3662.

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**: 1053-1058.

Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**: 41-50.

Masci J, Meier U, Cireşan D, Schmidhuber J. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pp. 52-59. Springer.

Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Alvarez-Varela A, Batlle E, Sagar, Grun D, Lau JK et al. 2020. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* **38**: 747-755.

Muto Y, Wilson PC, Ledru N, Wu H, Dimke H, Waikar SS, Humphreys BD. 2021. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat Commun* **12**: 2190.

Rohart F, Eslami A, Matigian N, Bougeard S, Le Cao KA. 2017. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**: 128.

Stadiotti I, Piacentini L, Vavassori C, Chiesa M, Scopece A, Guarino A, Micheli B, Polvani G, Colombo GI, Pompilio G et al. 2020. Human Cardiac Mesenchymal Stromal Cells From Right and Left Ventricles Display Differences in Number, Function, and Transcriptomic Profile. *Front Physiol* **11**: 604.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902 e1821.

Wu KE, Yost KE, Chang HY, Zou J. 2021. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci U S A* **118**.

Xiong L, Xu K, Tian K, Shao Y, Tang L, Gao G, Zhang M, Jiang T, Zhang QC. 2019. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* **10**: 4576.