A search-engine for single-cell epigenome profiles for multi-purpose applications

Shreya Mishra[1] , Smriti Chawla[1], Neetesh Pandey[1], Debarka SenGupta[1,2], Vibhor Kumar[1]

[1] Department for Computational Biology, Indraprastha Institute of Information Technology, Delhi 110020, India
[2] Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Australia

Abstract

The true benefits of large data-sets of single-cell epigenome and transcriptome profiles can be availed only when they are searchable to annotate individual unannotated cells. Matching a single-cell epigenome profile to a large pool of reference cells remains as a challenge and largely unexplored. Here, we introduce scEpiSearch, which enables a user to query single-cell open-chromatin read-count matrices for comparison against a large pool of single-cell expression and open-chromatin profiles from human and mouse cells (~ 3.5 million cells). Besides providing accurate search in a short time and scalable visualization of results for multiple query cells, scEpisearch also provides a low-dimensional representation of single-cell open-chromatin profiles. It outperformed many other methods in terms of correct low-dimensional embedding of single-cell open-chromatin profiles originating from different platforms and species. Here we show how scEpiSearch is unique in providing several facilities to assist researchers in the analysis of single-cell open-chromatin profiles to infer cellular state, lineage, potency and representative genes.

Introduction

Owing to the prolific adaptation of single-cell genomics by several research groups, it is now important to handle the challenge of storage, searching, and meta-analysis of single-cell epigenome profiles. Lahnemann et al.[1] have listed the problem of mapping a single-cell to reference atlas as one of the 11 grand challenges in single-cell data science. They have also listed integrating single-cell data across samples and experiments as a grand challenge. For single-cell expression profiles, there have been efforts by our[2] and other groups[3, 4] to tackle such challenges. However, for single-cell epigenome data-sets, there is still a need to resolve several issues associated with such challenges. Therefore, in spite of the availability of large single-cell epigenomic atlases[4-6] and methods for their low-dimensional visualization[7], there is a scarcity of search-engines meant to match a query single-cell epigenome profile to the multi-source collection of reference single-cell profiles.

In comparison to the single-cell transcriptome, single-cell open-chromatin profiles offer new obstacles. The read-count matrices compiled using single-cell open-chromatin tend to have more number of genomic loci (peaks) as features compared to a similar matrix for single-cell expression profile. Hence even conversion of their read-count matrix to gene activity score is time-consuming. Most often, genomic loci (peaks) in read-count matrix of single-cell open chromatin profile compiled by different research groups are not the same. Hence the algorithm

and search methods proposed for single-cell expression profiles cannot be used directly for single-cell open-chromatin profiles. In addition, there are different assays for detection of open chromatin such as DNase-seq (Dnase-I hypersensitive sites sequencing)[8], MNase-seq (Micrococcal-nuclease based hypersensitive sites sequencing)[9] and ATAC-seq (Transposase-Accessible Chromatin using sequencing)[10]. To tackle such issues, we developed scEpiSearch (scepisearch.com), a novel framework to match query single-cell open-chromatin profile with existing reference consisting of single-cell open-chromatin and single-cell expression data-sets. The framework of scEpiSearch resolves the issue of handling non-similar peak-lists of single-cell epigenome profiles from multiple scientific-groups and solves the problem of calculating the statistical significance of the match of the query with single-cell expression and open-chromatin profiles. For this purpose, scEpiSearch uses a gene-enrichment score which is more stringent in representing cell-type specificity than the gene-activity score used by some studies[5] (see supplementary methods).

Results

scEpiSearch has 4 reference data-sets. Two of the 4 reference data-sets consist of more than 3.3 million single-cell expression profiles from human and mouse cells. The other two reference data-sets consist of a single-cell ATAC-seq (scATAC-seq) profile of more than 150000 cells from human and mouse cells (supplementary method, supplementary figures S1-S2). The single-cell expression profiles were classified according to cell-type annotation so that they can be searched in a hierarchical manner. For a query single-cell open-chromatin profile, scEpiSearch first normalizes its read-count with a global accessibility score to highlight enhancers (Figure 1a). For both species, human and mouse, we used the global accessibility peak-list compiled using several published open-chromatin profiles of bulk samples published by different groups and consortiums [11, 12]. Normalization by global accessibility score for peaks removes the bias, which could have come from neighbouring cells in the query. Thus every cell in query is treated independently of each other. Using the predefined peak-list also helps in determining the proximal genes directly. For a query with more 200000 peaks, intersection with predefined peak-list with global accessibility speeds up the process of determining proximal genes by 1000 times (Figure 1b). For most of the query scATAC-seq read-count matrices built from multiple studies, we found that 80-90% of their peaks intersected with our predefined peak-list with global accessibility scores, most of the time. (Supplementary Figure S2c). Hence in the faster version of scEpiSearch, peaks not overlapping with predefined peak-list are dropped.

ScEpiSearch uses peaks with high normalized-counts as foreground and others as the background set to calculate the gene-enrichment score using the Fischer exact test (hypergeometric test) (see supplementary methods). For matching to a reference scRNA-seq profile, its median expression for top 1000 enriched genes (MExTEG) of query cells is calculated. For a query, the MExTEG of a reference cell is converted to P-value using precalculated MExTEG values for cells in the null model (see Supplementary Methods). Since scEpiSearch uses a hierarchical approach, it first matches query with representative expression vectors for clusters of cells. Then references cells from top matching clusters are used to calculate MExTEG values and corresponding P-values. Though cells with the lowest P-value

could be considered as best matches, however, scEpiSearch further calculate a new P-value based on rank of reference cells for a query to reduce bias in the data-set and searching procedure (see supplementary methods and Figure 1a). In other words, scEpiSearch makes a rank adjustment for hits using their precalculated ranks for the null model.

For matching to a reference scATAC-seq profile, its median enrichment score for top enriched genes (MESTEG) of query cells is calculated. The MESTEG value is converted to P-value using precalculated MESTEG scores for cells (vectors of enriched genes) in the null model (supplementary Methods). ScEpiSearch also uses a hierarchical approach to find the most matching scATAC-seq profile in the reference data-set (Figure 1a). After determining the rank of reference cells for a query cell, it calculates a new P-value using the precalculated rank of the same reference for the null model (see supplementary methods).

Even though one can use the webserver version of scEpiSearch, for large read-count matrices, it is advisable to use the stand-alone version (see supplementary Figure S3 and S4). The stand-alone version has its own database of reference expression and gene-enrichment scores of scATAC-seq profiles. Both versions of scEpiSearch have scalable visualization, where results of more than 1000 query scATAC-seq can be visualized interactively (supplementary figure S3). Evaluation using several scATAC-seq profiles with known cell-type annotation revealed that the accuracy of the search of scEpiSearch is about 80-100% in highlighting the correct cell-type among top 5 matches (Figure 1c-d). Using proximal genes enrichment and expression profiles, rather than relying on correlation or distances among cells, helps avoid artefacts due to technique or species specific batch effect. Therefore, scEpiSearch also allows matching query scATAC-seq profiles from human cells to mouse reference single-cell expression profiles with high accuracy (see Figure 1e). There are several possible applications of scEpiSearch, which include i) finding cell types for scATAC-seq profiles from unannotated cells from *in vivo* samples ii) Studying heterogeneity and tracking divergences in state of cells and their potency, for example, one cell-line showing differentiation potential towards several lineages, such as K562 cells[13] iii) Finding matching mouse model for a human cell which is not characterized well. iv) Identifying marker genes representing query cells v) Embedding and clustering multiple scATAC-seq profiles irrespective of their sources, technique and species.

We applied ScEpiSearch for several case studies, including the prediction of lineage and potency of cancer cells. Such as, for scATAC-seq profiles of HL60 (a myeloid leukaemia cell-line), scEpiSearch found top matching expression results from myeloid cell-types (dendritic and Langerhans cells). For 10% of the HL60 cells, scEpiSearch also reported monocytes among the top 3 matching cells (Figure S5). The differentiation potential of HL60 to monocyte is well known. For scATAC-seq profile of K562 cells[10] scEpiSearch reported mostly erythroid and erythroid precursor cells among top 3 matching expression profiles (Figure 2a). Then we used scEpiSearch for another case study of detecting heterogeneity among K562 cells. Clustering of K562 scATAC-seq profiles using their low dimension projections in terms of match scores with reference cells revealed two major clusters (Figure 2c-d). For K562 cells belonging to the smaller cluster, some of the most frequently enriched genes, such as NTRK3 (TRKC)[14] and MAF have higher expression in megakaryocytes than erythrocytes (see cluster-1 in Figure 2c-

d). MAF is known to cause differentiation of hematopoietic progenitor cells towards megakaryocytes and reduction of its transformation towards erythroid lineage[15]. Thus our analysis using scEpiSearch highlighted the bias of underlying chromatin state in some K562 cells towards megakaryocytic lineage.

Further, we used scEpiSearch to annotate cells with “unknown” annotation in the scATAC-seq data-set published by Cusanovich et al. for mouse cells[5]. The top matching reference scRNA-seq profile found by scEpiSearch for scATAC-seq read-counts matrices of unknown cells were very relevant. Such as for the scATAC-seq profile of unknown cells from mouse liver cells, 96% of top 5 matching scRNA-seq profiles were from hepatocytes (Supplementary Figure S6). For the unknown cells from the heart (Figure 3) the top 5 matching expression profiles found by scEpiSearch, were mostly from connective tissue progenitor, endothelial and intermediate mesoderm cells. The presence of connective tissue in the heart as cardiac mesenchymal stromal cells (C-MSC) has been shown previously [16]. For thymus almost 70% of matching scRNA-seq profiles were from white blood cells, including T cells (Supplementary Figure S7). The thymus is known to be instrumental in the production and maturation of T-lymphocytes (a type of white blood cell). We achieved similar coherent results for unknown cells from the pre-frontal cortex (supplementary figure S8).

Last but not least, scEpiSearch can also perform embedding of multiple scATAC-seq profiles with different peak-list and batch-effect irrespective of their species of origin. We compared the performance of scEpiSearch with four methods meant only for embedding (SCANORAMA, MINT, SCVI, SCALE)[17-20] using 4 different collections of scATAC-seq read-count matrices. As shown in Figure 4 and Supplementary Figures S9-S10 , the 2D embedding plot made by scEpiSearch for scATAC-seq profiles has almost correct co-localization of similar cell types irrespective of the species and laboratory of origin. Other available methods of embedding scRNA-seq profiles[17, 19] and scATAC-seq tools[20] could not provide correct results like scEpiSearch (see Figure 4 and Supplementary Figures S9-S10). Overall with different case studies, it becomes quite evident that scEpisearch is useful for the analysis of single-cell open-chromatin profiles in many different ways.

Discussion

The uniqueness of scEpiSearch lies in its collection of reference cells and its statistical approach to reduce bias in the search for matching transcriptome and open-chromatin profiles. It is resourceful not only in terms of collection and search of reference cell profiles but also due to various facilities it provides such as : low-dimensional representation, embedding robust to batch effect, summary word-cloud for overall notion about query scATAC-seq profiles and enrichment scores of genes to highlight possible markers for cell-types. scEpiSearch can also be very useful in the integrated and interactive analysis of scATAC-seq profiles for identifying rare cells. Cross-validation of identification of rare cells using single-cell open chromatin profiles is eminent and imminent for which scEpiSearch can have a significant contribution. Due to its dependence on highlighting enhancers, scEpiSearch can also be used for single-cell histone

modifications profiles[21] showing enhancer activity, such as histone-acetylations, H3K4me1 and H3K4me2[22]. In future, scEpiSearch would be improved further with larger reference data-sets and more functionalities.

Availability: scEpiSearch has been made available as stand-alone tool with a user-friendly graphical user interface available at:
http://reggen.iiitd.edu.in:1207/episearch/index.php?view=download

Alternatively, it can also be assessed freely with some toned-down functionality at scepisearch.com or http://reggen.iiitd.edu.in:1207/episearch/index.php

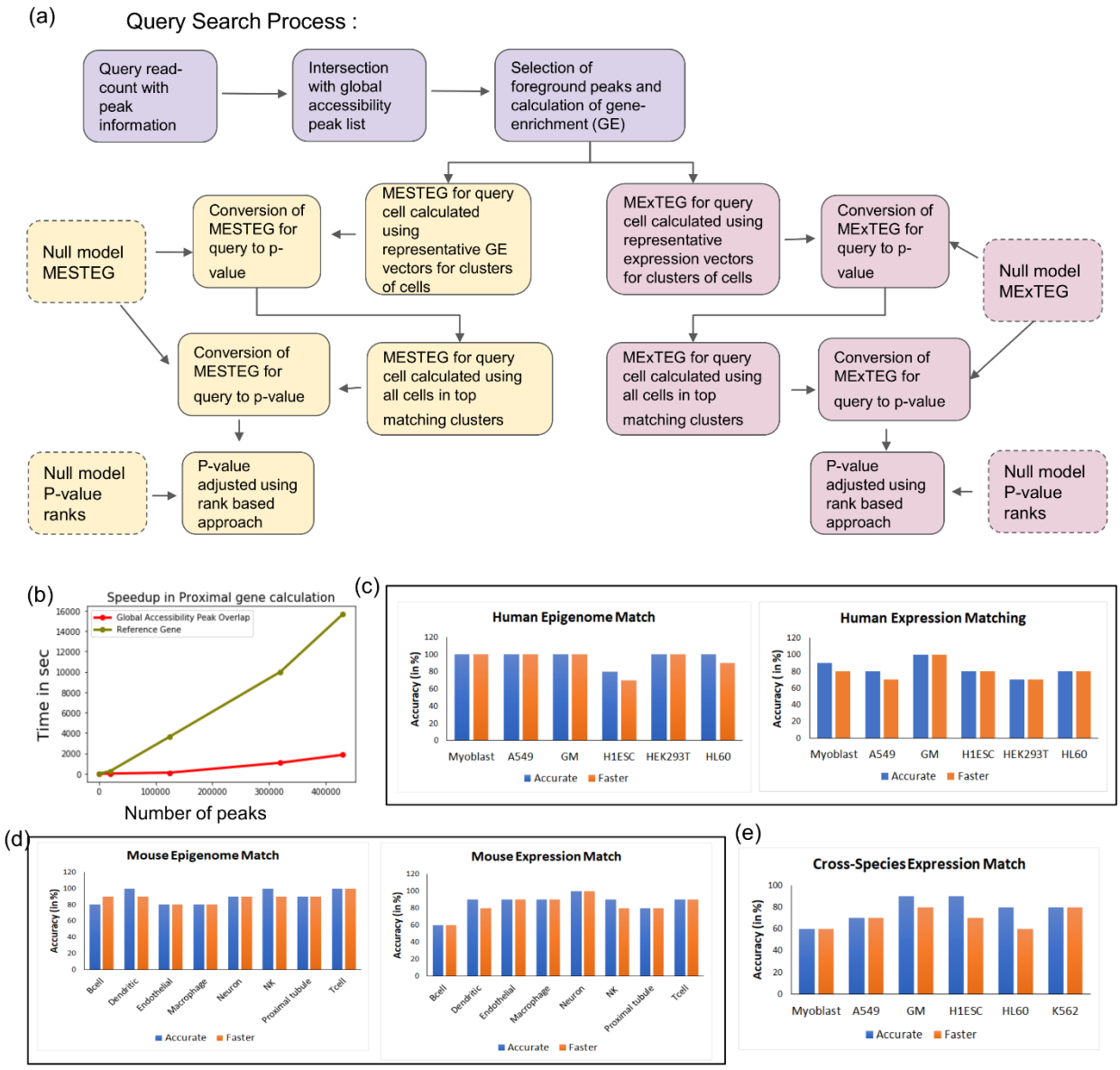


Figure 1: Overview of scEpiSearch and evaluation of its accuracy a) Flowchart of Query processing by ScEpiSearch. It takes peaks and read-count matrix as input and finds proximal genes, and calculates gene-enrichment (GE) scores for cells. Further, it searches for matching reference epigenome and expression 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 profile. **b)** Speedup achieved in proximal gene calculation by ScEpiSearch. Red line shows the standard way of finding the proximal gene file (Reference Gene), while the green lines show the method utilizing overlap of query peaks with global accessibility peak-list. **c)** Accuracy achieved with query scATAC-seq read-count matrices from 5 human cell-types, while searching for matched human single-cell expression and open-chromatin profile. Y-axis shows accuracy in terms of the percentage of query cells for which the top 5 results had correct annotations. Two plots are shown for human expression match and human epigenome match, with each one showing plots for faster and accurate modules of scEpiSearch. Queries shown are made from data

repositories (GEO Id: GSE109828, GSE117089, GSE65360, GSE68103) **d)** Accuracy achieved in Mouse queries (made from data-set GEO id : GSE111586) search for each of the queries mentioned in X-axis. Y-axis shows accuracy achieved in percentage. **e)** Accuracy achieved in the cross-species match, with query scATAC-seq profile from human cells and reference expression values from the mouse. The accuracy is shown as the percentage of query cells that had the same cell-type among the top 5 matches

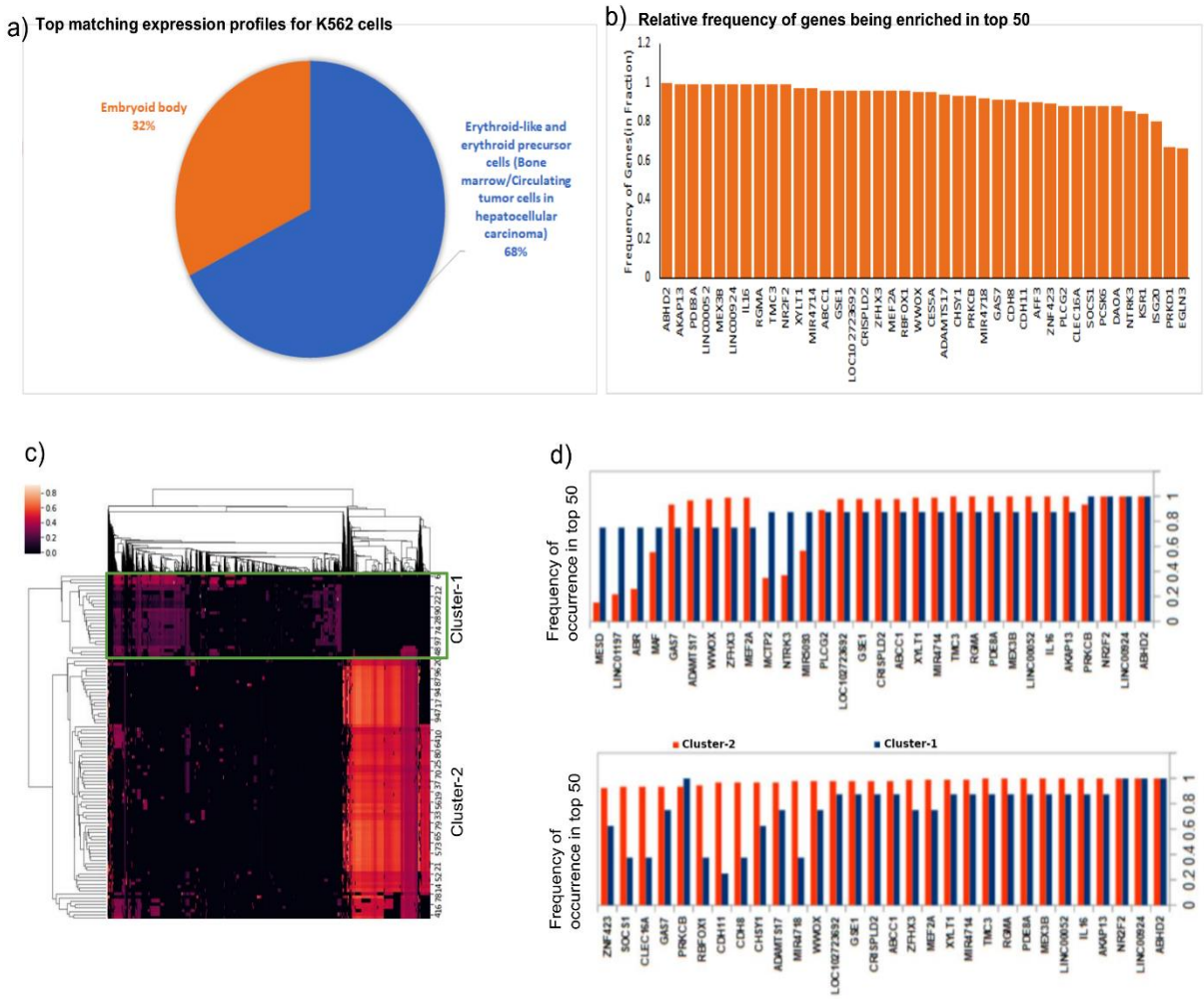


Figure 2: K562 human query with human expression match results are shown. **a)** Pie chart of the phenotype of top three matches in expression with p-value<0.05 in human expression is shown. **b)** Frequency of occurrence in top twenty enriched genes list for all query cells is shown as a bar plot. Here most frequently occurring enriched genes are shown. **c)** Heatmap of epigenome matching profiles is shown. X-axis shows matching epigenome profiles from reference, and Y-axis shows Query cells. Two clusters of query cells are also shown. **d)** Relative frequency of enriched genes in the top twenty is shown for query cells from two clusters. The top panel shows genes sorted according to the frequency in cluster-1. For bottom panel genes are sorted according to the frequency in query cells in cluster-2. Few top frequent genes in cluster-1 and cluster-2 have similar frequencies. However, other genes like ABR, MCTP2, MESO, CHD8, CHD11 show difference in frequency of occurrence in the top 50 enriched genes for query cells in cluster-1 and cluster-2.

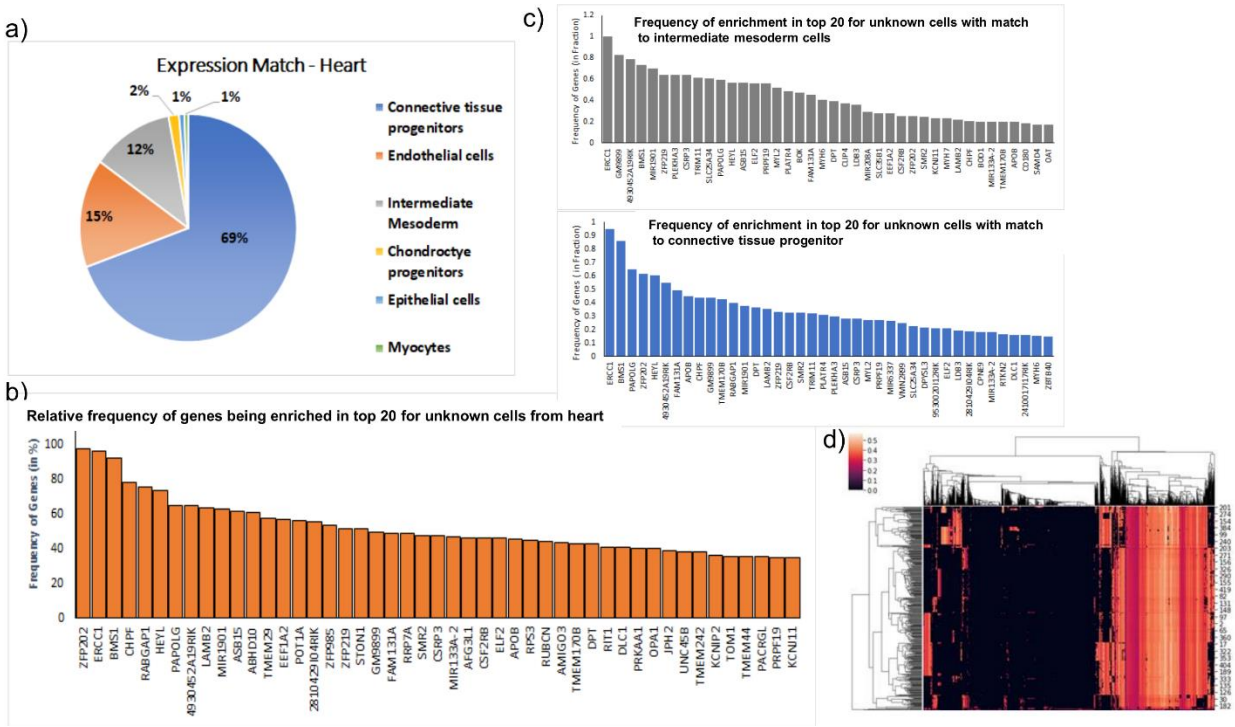


Figure 3: **A** case study of annotating unknown single-cell open chromatin profiles from mouse Heart tissue (Cusanovich et al.) **a)** Pie chart of the phenotype of top five matches with p-value<0.05 in mouse expression. **b)** Frequencies of appearance in the top twenty enriched genes in all query 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.

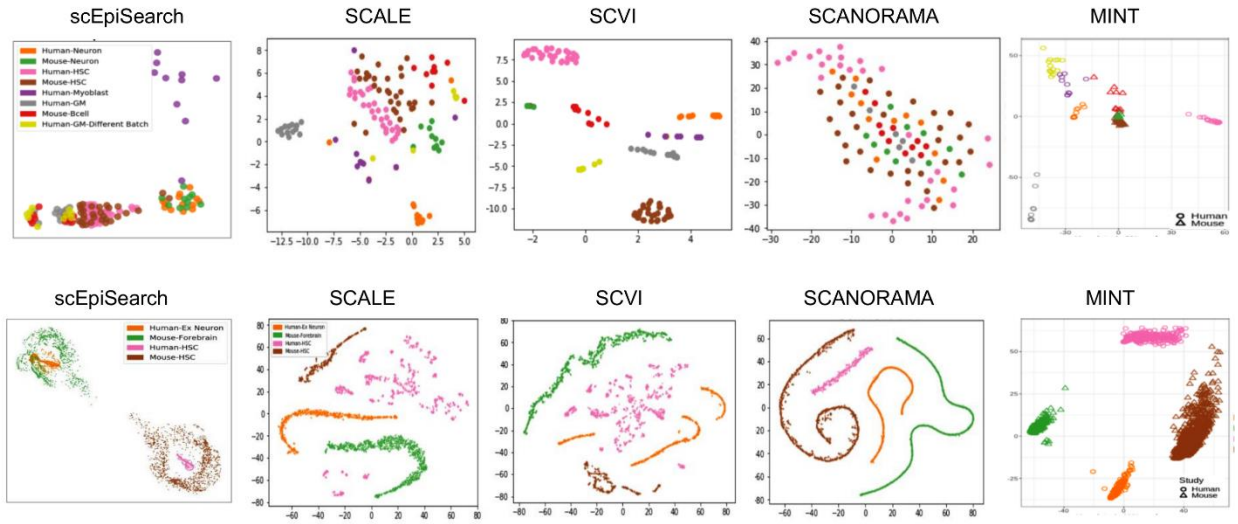


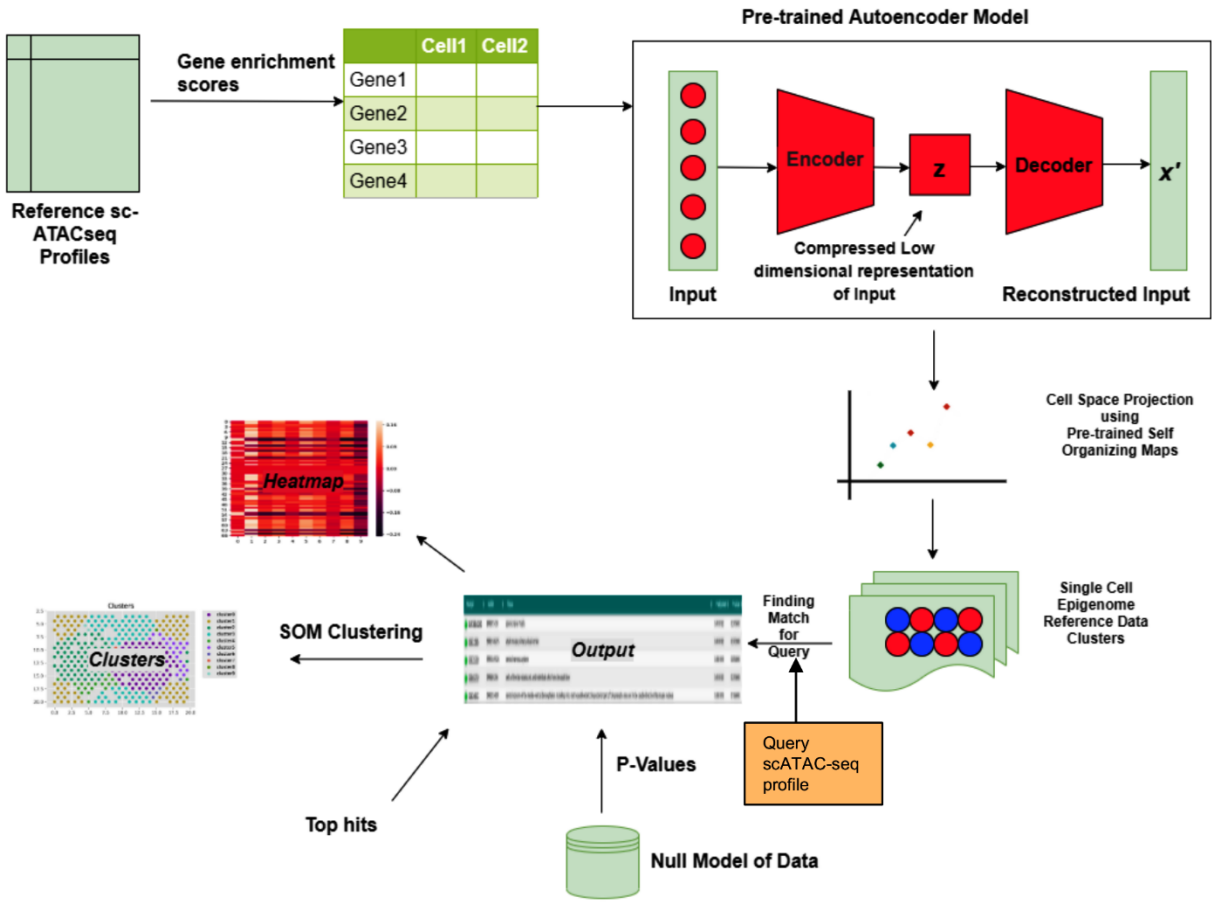
Figure 4: A case-study of embedding multiple query-sets consisting of scATAC-seq from different cell-types, scientific-groups and species. (a) For this case-study, queries consisted of separate read-count matrices for scATAC-seq profiles of human-neuron, mouse-neuron, human-HSC (hematopoietic stem cells), house-HSC, human-Myoblast, human-GM12878 cells from two batches and mouse-B-cells. The peak-list of read-count matrices were different from each other. Embedding plot from 5 different methods are shown here. While SCANORAMA mixed the location of all cells, MINT could not group cells of the same type together like scEpiSearch. (b)Another example of embedding query scATAC-seq profiles from different batches and species. Here, queries are made for Human-Neuron, Mouse-Neuron, Human-HSC, Mouse-HSC. None of the other methods evaluated here could give correct low-dimensional embedding like scEpiSearch.

References:

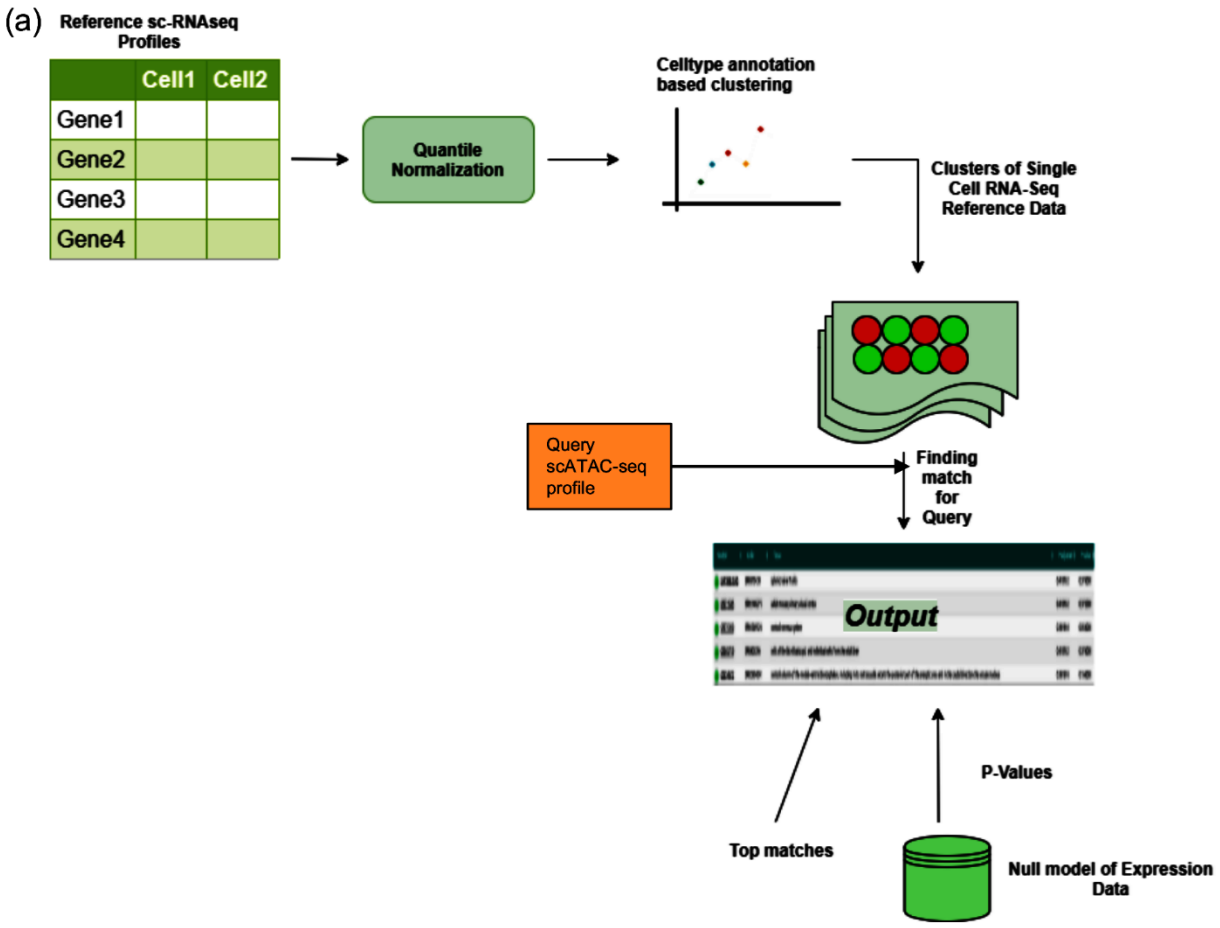
1. Lahnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020).
2. Srivastava, D., Iyer, A., Kumar, V. & Sengupta, D. CellAtlasSearch: a scalable search engine for single cells. *Nucleic Acids Res* **46**, W141-W147 (2018).
3. Cao, Z.J., Wei, L., Lu, S., Yang, D.C. & Gao, G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat Commun* **11**, 3458 (2020).
4. Lake, B.B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).
5. Cusanovich, D.A. et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e1318 (2018).
6. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370** (2020).
7. Ji, Z., Zhou, W., Hou, W. & Ji, H. Single-cell ATAC-seq signal extraction and enhancement with SCATE. *Genome Biol* **21**, 161 (2020).

8. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142-146 (2015).
9. Lai, B. et al. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **562**, 281 (2018).
10. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).
11. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
12. Bujold, D. et al. The International Human Epigenome Consortium Data Portal. *Cell Syst* **3**, 496-499 e492 (2016).
13. Tetteroo, P.A., Massaro, F., Mulder, A., Schreuder-van Gelder, R. & von dem Borne, A.E. Megakaryoblastic differentiation of proerythroblastic K562 cell-line cells. *Leuk Res* **8**, 197-206 (1984).
14. Kizilyer, A. et al. Inhibition of Tropomyosin Receptor Kinase A Signaling Negatively Regulates Megakaryopoiesis and induces Thrombopoiesis. *Sci Rep* **9**, 2781 (2019).
15. Bianchi, E. et al. MYB controls erythroid versus megakaryocyte lineage fate decision through the miR-486-3p-mediated downregulation of MAF. *Cell Death & Differentiation* **22**, 1906-1921 (2015).
16. Pilato, C.A. et al. Isolation and Characterization of Cardiac Mesenchymal Stromal Cells from Endomyocardial Bioptic Samples of Arrhythmogenic Cardiomyopathy Patients. *J Vis Exp* (2018).
17. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* **37**, 685-691 (2019).
18. Rohart, F., Eslami, A., Matigian, N., Bougeard, S. & Le Cao, K.A. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* **18**, 128 (2017).
19. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058 (2018).
20. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* **10**, 4576 (2019).
21. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology* **33**, 1165-1172 (2015).
22. Spicuglia, S. & Vanhille, L. Chromatin signatures of active enhancers. *Nucleus* **3**, 126-131 (2012).

Supplementary Figures



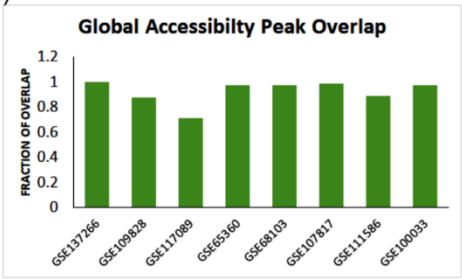
Supplementary Figure S1: Methodology for preparation of Reference Epigenome representatives for scEpiSearch. Our pipeline takes single-cell open-chromatin profiles, calculates their gene enrichment scores, passes them through trained autoencoder to get compressed representation and trained self-organizing map to find clusters. The output for Query match to epigenome profiles generates heatmap, SOM clusters (interactive 2-D plot), top hits, and their p-values. The gene-enrichment scores of query open-chromatin profile are not passed to autoencoder; rather, they are compared to cluster-representative gene-enrichment vectors using a different approach (see supplementary method).



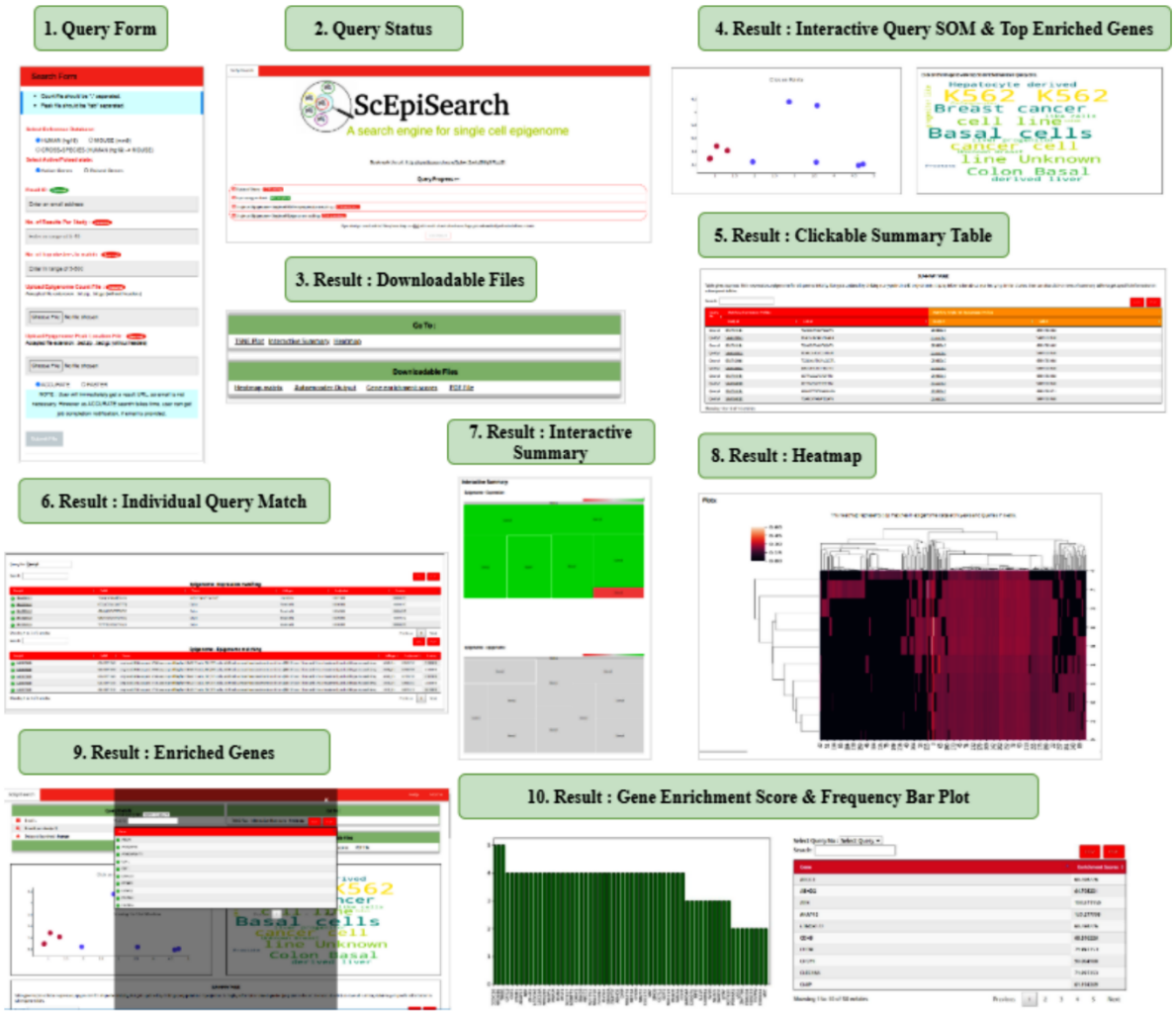
(b)

Reference data-set	Number of cells
Single Cell Expression (Human)	1288653
Single Cell Epigenome (Human)	742297
Single Cell Expression (Mouse)	2141797
Single Cell Epigenome (Mouse)	81173

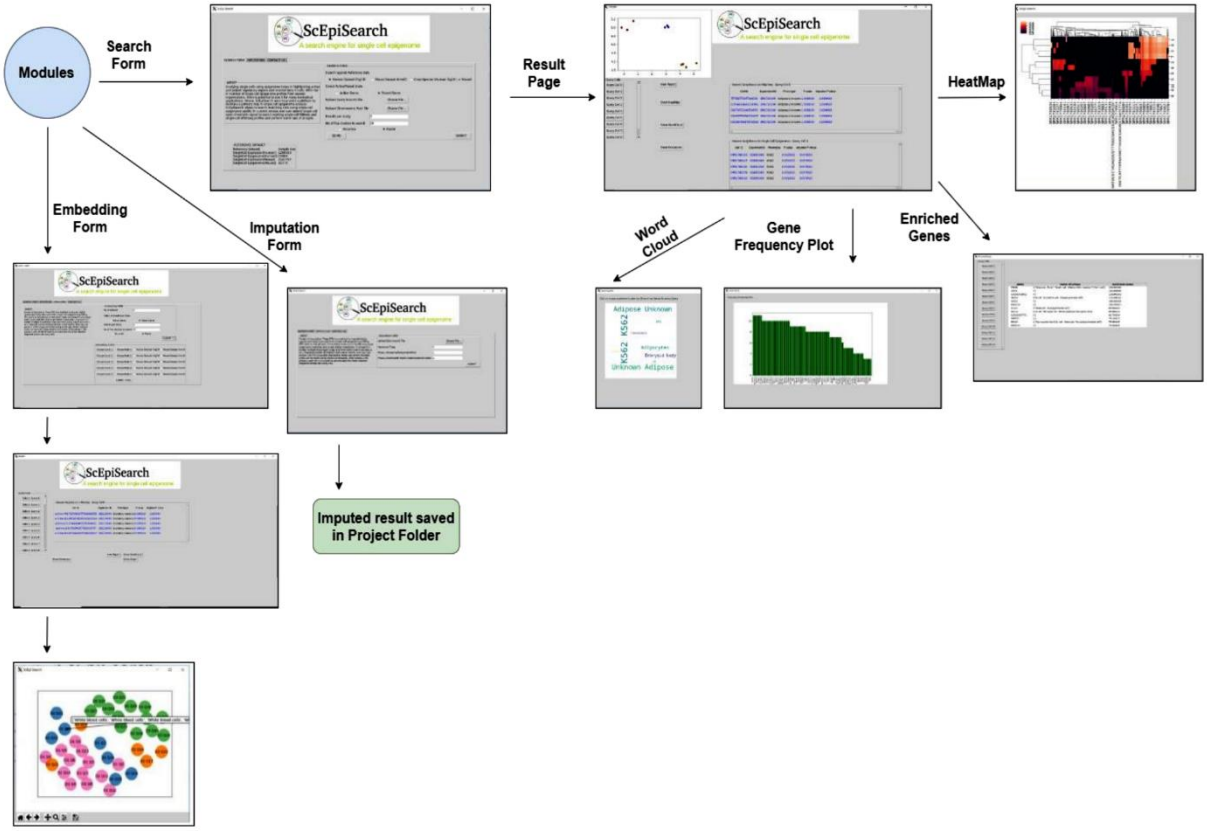
(c)



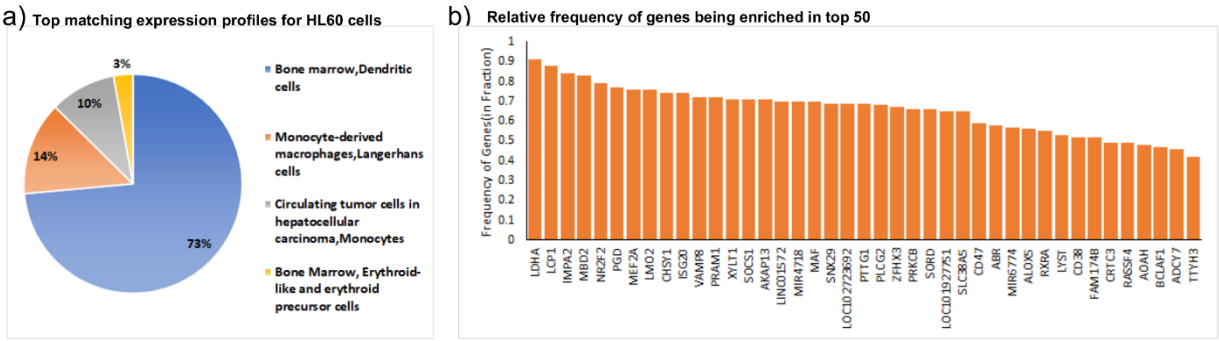
Supplementary Figure S2: Methodology for data preparation from the expression of reference cells. (a) The reference scRNA-seq profiles are clustered according to cell-type annotation. The representative expression vector for clusters of reference-cells is first matched to the query in the first round of search. In the second round of search, the reference cells in top matching clusters are matched to query. The p-value of the match is calculated using the Null model, which could also be considered as simulated false queries. (b) The current number of cells whose scRNA-seq and scATAC-seq profiles are in the reference dataset. (c) The fraction of peaks overlapping with the pre-compiled global accessibility peak list is shown. The X-axis shows the data repositories from where query scATAC-seq profiles were made, and the Y-axis shows the fraction of overlap of their peaks with peak-list of global accessibility table.



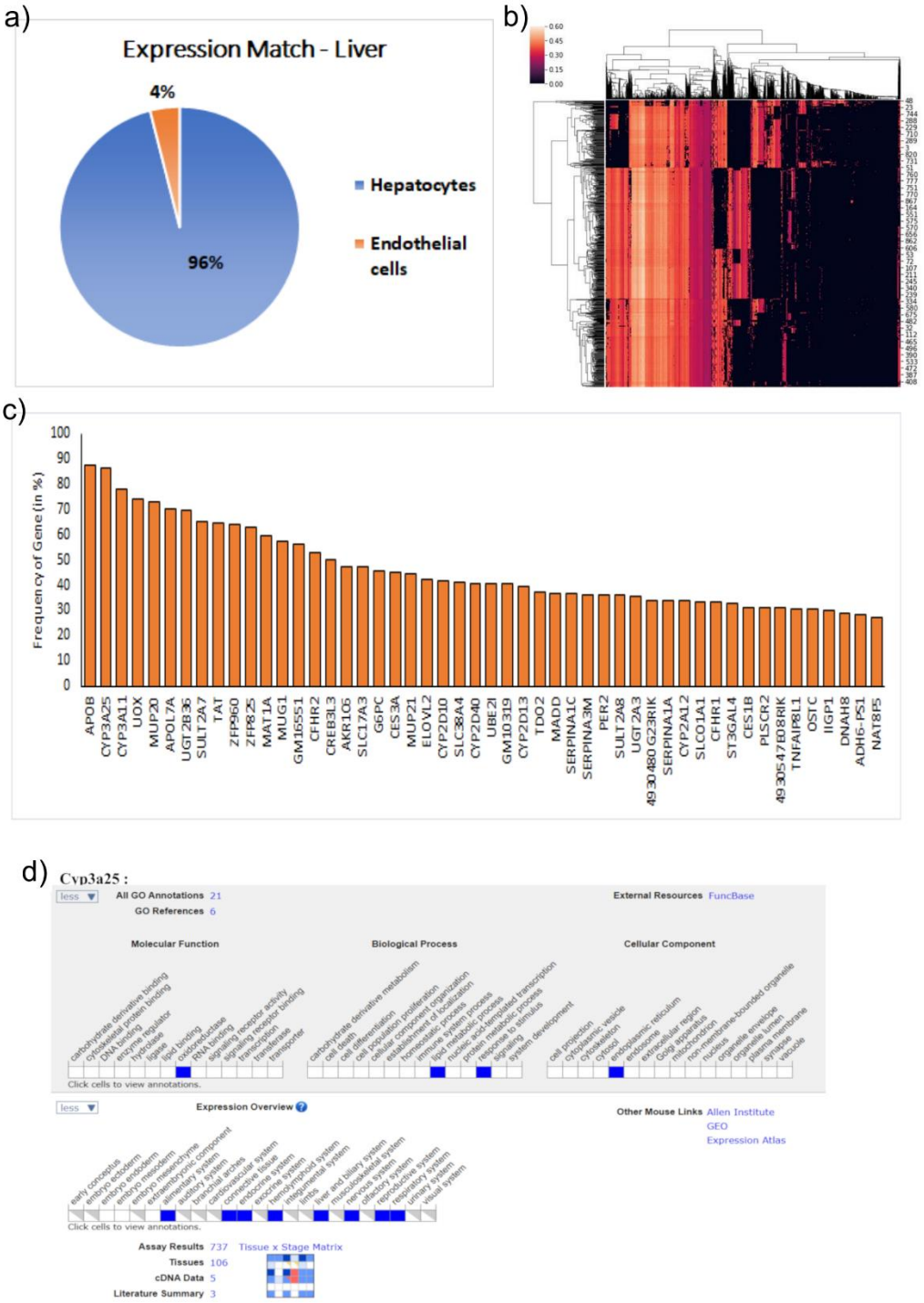
Supplementary Figure S3: The web interface of ScEpiSearch is shown. **1)** Query form to submit peak and read-count file of query and other parameters such as reference database, the number of top matches, no of clusters to search from, accurate/faster search, poised/active stat and annotated or both type of reference cells to search. **2)** Shows job running status for the query. **3)** Panel on the result page to download files available on ScEpiSearch for further analysis for users. **4)** Interactive SOM plot of Queries and a word cloud of all matching phenotypes. **5)** Summary table for Queries with their study Id and cell Id of a topmost hit in both epigenome and expression match. **6)** Individual tables for top N matches with their complete information the user requested for both matches. **7)** Interactive summary plot with the colour and size of each match based on their p-values. **8)** Heatmap from epigenome matches. **9)** When a user clicks on word cloud, a pop-up of enriched genes is shown. 10) The top gene-enrichment scores for query is shown as a bar plot and table in the result page of scEpiSearch.



Supplementary Figure S4: Modules of the standalone tool are shown. It has three modules: **1) Search Form** takes to peak and count file as input and other parameters after execution shows result page and further button clicks leading to the word cloud, enriched genes, heatmap. **2) Imputation form** takes single-cell open chromatin profiles as input and saves the imputed matrix in the project folder. **3) Embedding tool** takes cross-species and cross batch queries. It can take multiple matrics of read-counts and show matching profile from mouse expression reference; an interactive embedding plot is also shown using a spectral clustering approach.

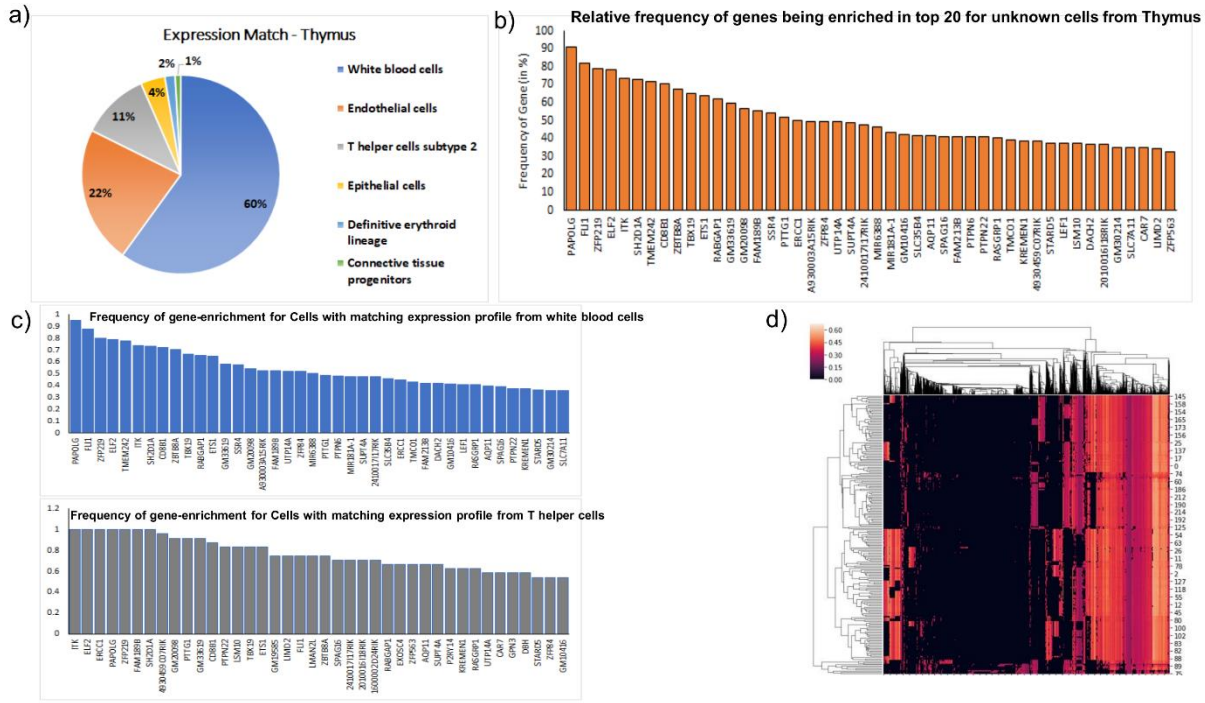


Supplementary Figure S5: HL60 human query (source: GSE109828) with human expression match results are shown. **a)** Pie chart of the phenotype of top three matches in human expression with p-value<0.05. **b)** Frequency of being enriched in the top 50 in all query cells is shown as a bar plot. The X-axis shows genes whose fraction of frequency occurrence is shown on the Y-axis.

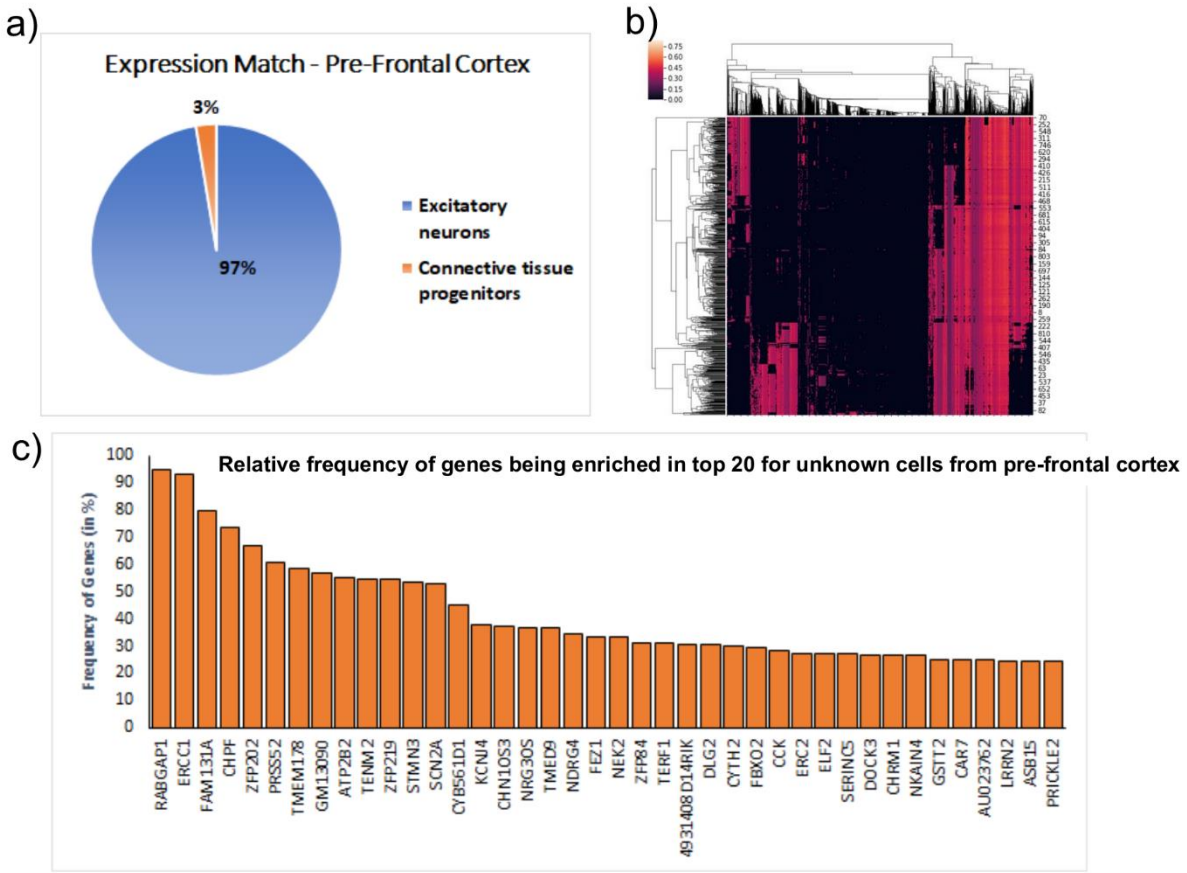


Supplementary Figure S6: A Case study of annotating single-cells with "unknown" annotation from mouse liver tissue published by Cusanovich et al. (GSE111586). **a)** Pie chart of the phenotype of top five matches in mouse expression with p-value<0.05. **b)** Heatmap, showing match score with reference epigenome profiles and hierarchical clustering of query cells. The X-axis shows matching epigenome profiles from reference, and Y-axis shows Query cells. **c)** Frequency of enriched genes in the top twenty in

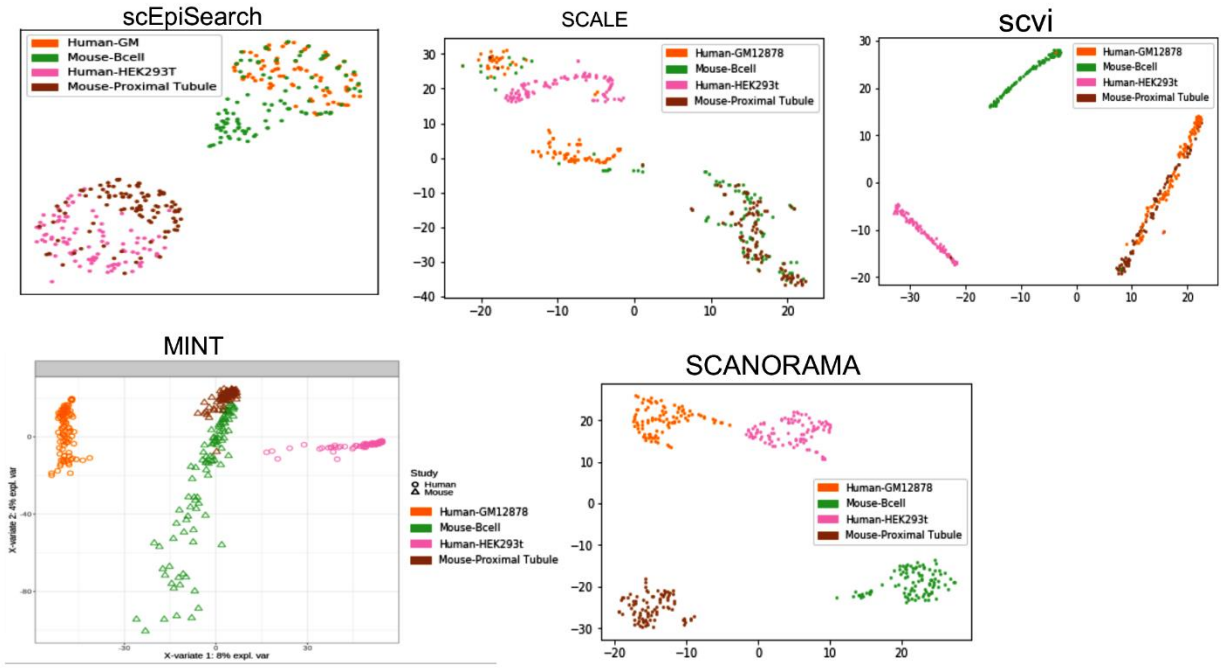
all query cells is shown as a bar-plot. The X-axis shows genes whose fraction of frequency occurrence is shown on the Y-axis. d) Information from the MGI (Mouse Genome Informatics) database about one of the top enriched genes Cyp3a25 found using scATAC-seq profiles of unknown cells from Liver tissue. It is displayed after clicking the relevant gene name on the result page of scEpiSearch.



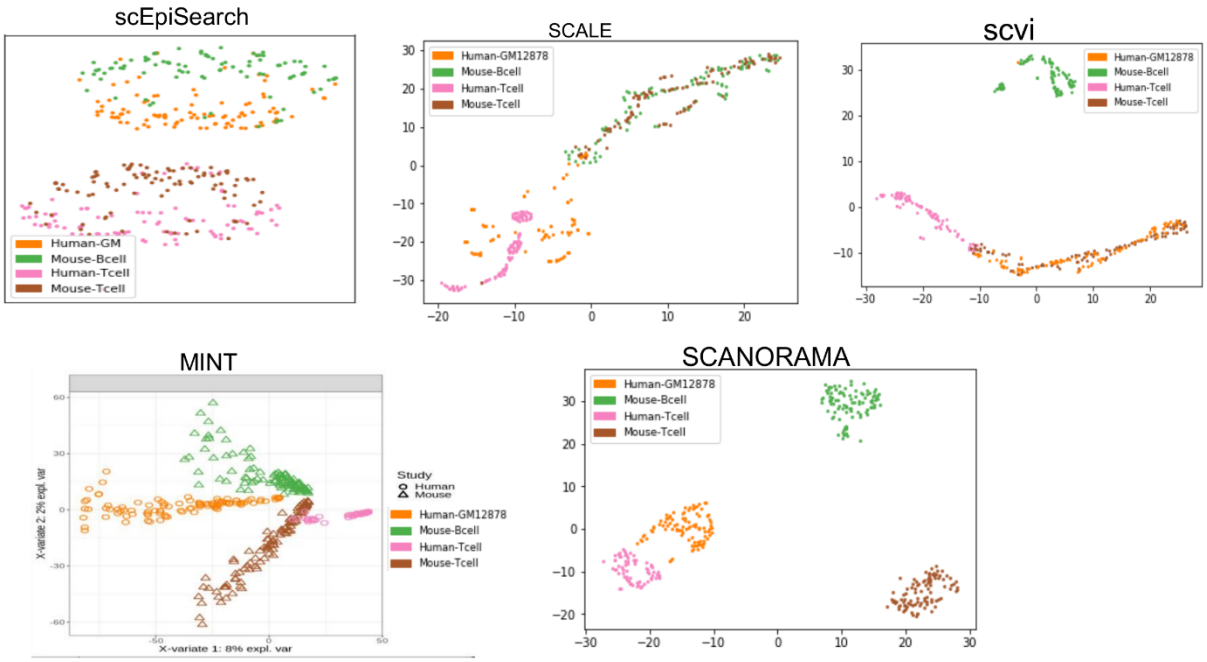
Supplementary Figure S7: A Case study of annotating single-cells with 'unknown' annotation from mouse (GSE111586) Thymus tissue. **a)** Pie chart of the phenotype of top five matching expression profiles from mouse cell (with p-value<0.05) for all queries (Unknown cells from thymus). **b)** Frequency of appearance in the 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. The X-axis shows matching epigenome profiles from reference, and Y-axis shows Query cells.



Supplementary Figure S8: A case study of annotating single-cells with the annotation "unknown" from mouse (GSE111586) pre-frontal cortex tissue is shown. **a)** Pie chart of the phenotype of top five matches with p-value<0.05 in mouse expression is shown. **b)** Heatmap of epigenome matching profiles is shown. The X-axis shows matching epigenome profiles from reference, and Y-axis shows Query cells. **c)** Frequency of enriched genes in the top twenty in all query cells is shown as a bar plot. The X-axis shows genes whose fraction of frequency occurrence is shown on the Y-axis.



Supplementary Figure S9: The plot of embedding results showing the alignment of cells from different species and batches together. Queries are made for Human-GM12878 cell, Mouse B-cell, Human-HEK293T, Mouse-Proximal tubule. Embedding plots from 5 different methods are shown, including scEpiSearch. According to the similarity of tissue or cell-type, none of the other 4 methods evaluated here performed correct low-dimensional embedding of scATAC-seq profiles.



Supplementary Figure S10: The plots show 2D embedding of cells from different species and batches. Queries were made for Human GM12878, Mouse B cell, Human T-cells, Mouse T cell. These 4 cell type queries had their own read-count matrix and peak list. Results of 2D embedding by different methods are shown here. It can be seen that in the embedding plot from scEpiSearch that the mouse and human T cells are co-localized. Similarly, Human GM12878 cells and mouse B cells are localized in the proximity of each other in the result of scEpiSearch. In contrast, none of the other 4 methods showed correct co-localization, according to cell-type.