MDPI

Article

Tensor Decomposition Discriminates Tissues Using scATAC-seq

Y-H. Taguchi [1,*] and Turki Turki [2]

[1] Department of Physics, Chuo University, Tokyo 112-8551, Japan
[2] Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; tturki@kau.edu.sa
* Correspondence:tag@granular.com; Tel.: +81-3-3817-1791

Abstract: ATAC-seq is a powerful tool for measuring the landscape structure of a chromosome. scATAC-seq is a recently updated version of ATAC-seq performed in a single cell. The problem of scATAC-seq is data sparsity and most of the genomic sites are inaccessible. Here, tensor decomposition (TD) was used to fill in missing values. In this study, TD was applied to massive scATAC-seq data and generated averaging within 200 bp intervals, and this number can be as high as 13,627,618. Currently, no other methods can deal with large sparse matrices. The proposed method not only could provide UMAP embedding coincident with tissue specificity but also could select genes associated with various biological enrichment terms and transcription factor targeting. This suggests that TD is a useful tool to process large sparse matrix generated from scATAC-seq.

Keywords: scATAC-seq; tensor decomposition; large sparse matrices; single-cell applications

1. Introduction

ATAC-seq [1] is a powerful tool to profile chromatin accessibility. In particular, scATAC-seq [2] can profile the accessibility of chromatin across the genome within individual cells. Although scRNA-seq is an effective tool, data need additional information to interpret its results. scATAC-seq data can be analyzed in two ways: 1) the scATAC-seq data coupled to scRNA-seq data [3], or 2) tracing for potential factors that bind to identified open chromatin regions from scATAC-seq data. This prevents us from understanding scATAC-seq as it is.

For example, Satpathy et al. [4] tried to find cell type-specific cis- and trans-regulatory elements, mapping of disease-associated enhancer activity, and reconstruction of trajectories of differentiation from progenitors to diverse and rare immune cell types. Giansanti et al. [5] used a predefined set of genomic regions to interpret the scATAC-seq. Buenrostro et al. [6] also tried to find an association with transfactors and cis elements. Although these are only a few examples, it is obvious that they need massive external information to interpret the scATAC-seq results.

In this paper, we applied tensor decomposition (TD) [7] to the scATAC-seq data set and found that the low-dimensional embedding obtained by UMAP applied to that obtained by TD is highly tissue-specific. This can open numerous avenues of research to make use of scATAC-seq data without additional biological information.

Although there are some studies that applied TD to scRNA-seq [8,9], to our knowledge, there are no studies that applied TD to scATAC-seq.

2. Materials and Methods

Figure 1 shows the flow chart of the analysis.

2.1. scATAC-seq profiles

The scATAC-seq data set we analyzed in this study is obtained from Gene Expression Omnibus (GEO) [10], the ID is GSE167050. We used eight samples of GSM5091379 to GSM5091386 from four mice tissues (CTX, MGE, CGE and LGE) with two replicates each

Citation: Taguchi, Y.-H.; Turki, T. Title. *Biomolecules* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

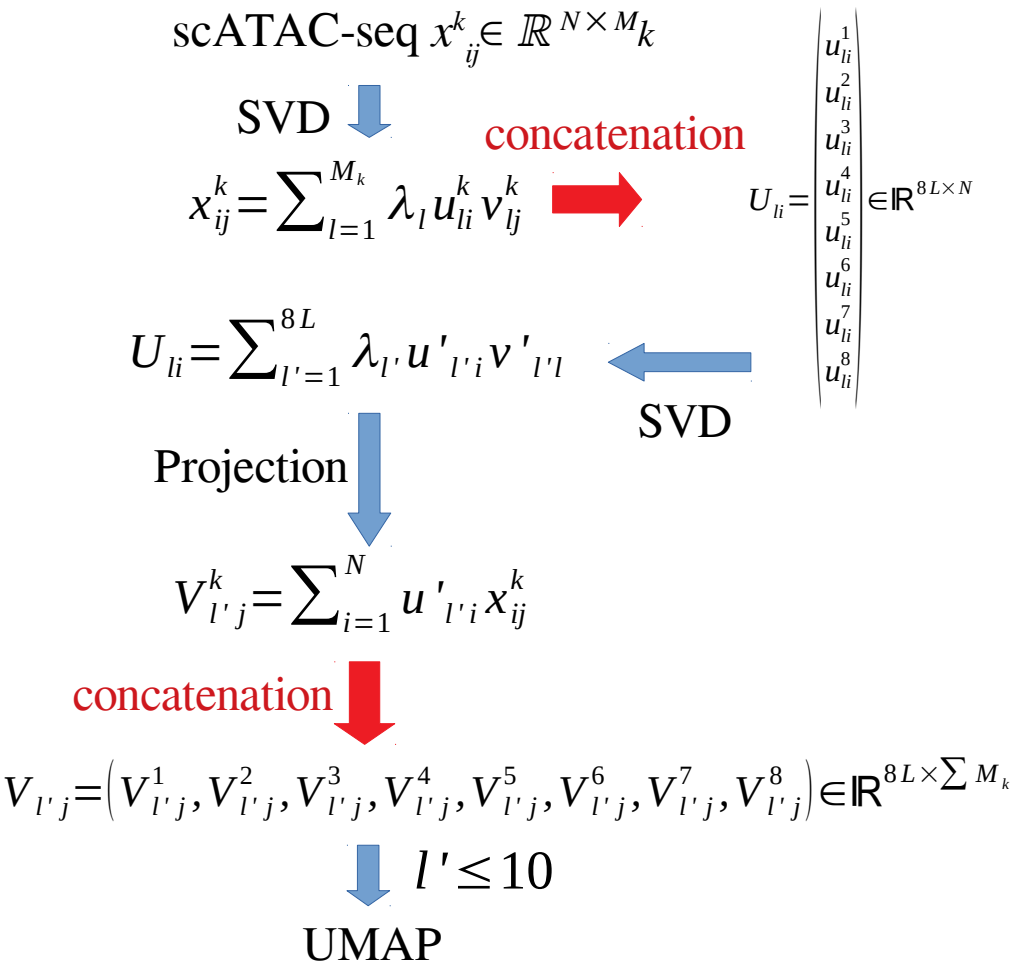

Figure 1. Flow chart of the analysis

(Table 1). Each of the eight samples is associated with three files, barcodes, features, and matrices, which correspond to single cells, genomic coordinates, and scATAC-seq values, respectively.

Table 1. Number of single cells included in the files analyzed in this study. CTX: Cortex, MGE: Medial ganglionic eminence, CGE: Caudal ganglionic eminence, LGE: Lateral ganglionic eminence

Tissues	CTX1	MGE1	CGE1	LGE1	CTX2	MGE2	CGE2	LGE2
number of single cells (M_k)	4108	6845	4013	6577	4946	3465	4530	4769

2.2. Pre-processing scATAC-seq profile

The values stored in the matrix files are averaged over 200 bp intervals, which are supposed to correspond to the length of one wrap of chromatin and linker. Since this results in a sparse matrix, it is stored in a sparse matrix format having columns and rows equivalent to the number of single cells and total number of intervals, which is up to 13,627,618. This value is approximately equal to 12.5 million, which is calculated by dividing the total number of mice genome bps by the interval length, i.e. 2.5 billion by 200.

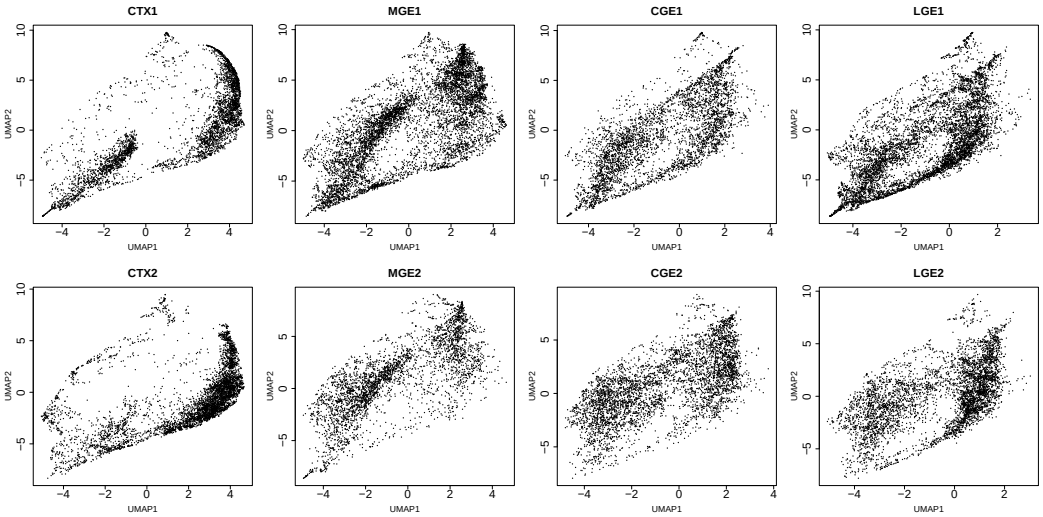

Figure 2. UMAP embedding of eight samples analyzed in this study (see Table 1)

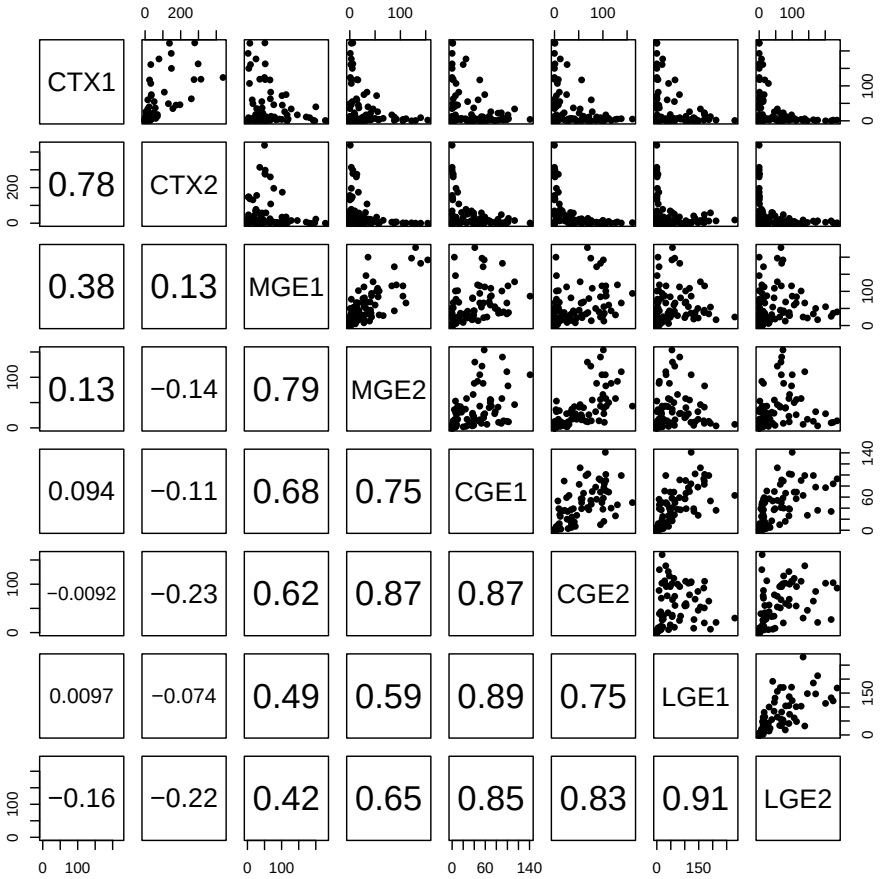

Figure 3. Upper: scatter plot of the number of single cells, $N_{IJ}, 1 \leq I, J \leq 10$, in one of 10×10 regions, S_{IJ}. Lower: Spearman's correlation coefficients of N_{IJ} between eight samples.

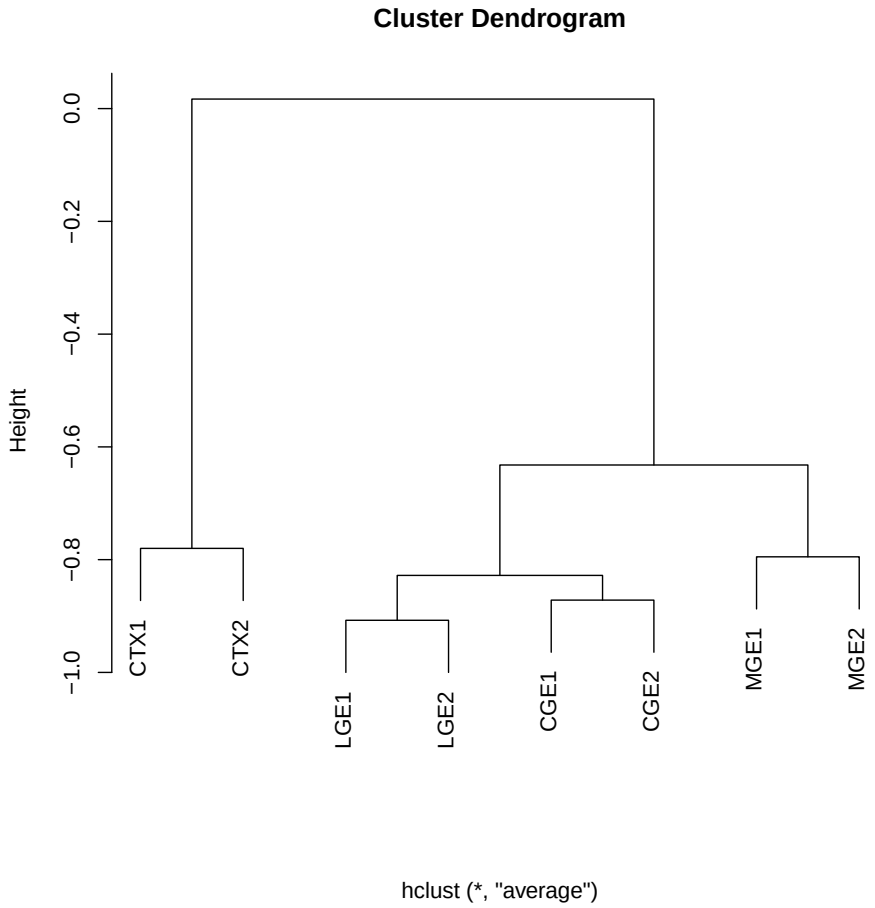

Figure 4. Hierarchical clustering by UPGMA of eight samples (Table 1) using coordinates obtained by UMAP. Distance is represented by negatively signed correlation coefficients in Fig. 3

2.3. Application of singular value decomposition

To obtain a low-dimensional embedding of the individual samples k that will be reformatted as a tensor, we applied the singular value decomposition (SVD) to matrix, $x^k_{ij} \in \mathbb{R}^{N\times M_k}$ (N = 13,627,618 and M_ks are in Table 1) as

$$x^k_{ij} = \sum_{\ell=1}^{M_k} \lambda_\ell u^k_{\ell i} v^k_{\ell j} \tag{1}$$

where $u^k_{\ell i} \in \mathbb{R}^{M_k\times N}$ and $v^k_{\ell j} \in \mathbb{R}^{M_k\times M_k}$ are orthogonal singular value matrices. Then we get concatenated matrix $U_{\ell i} \in \mathbb{R}^{8L\times N}$ of $u_{\ell i}$ where $L < \min(M_k)$ is the used number of principal components. SVD is applied to $U_{\ell i}$ and we get

$$U_{\ell i} = \sum_{\ell'=1}^{8L} \lambda_{\ell'} u'_{\ell' i} v'_{\ell' \ell} \tag{2}$$

where $u'_{\ell' i} \in \mathbb{R}^{8L\times N}$ and $v_{\ell' \ell} \in \mathbb{R}^{8L\times L}$ are orthogonal singular value matrices.

Low-dimensional embedding can be obtained as

$$V^k_{\ell' j} = \sum_{i=1}^{N} u'_{\ell' i} x^k_{ij} \in \mathbb{R}^{8L\times M_k}. \tag{3}$$

and we get concatenated matrix $V_{\ell' j} \in \mathbb{R}^{8L \times \sum_{k=1}^{8} M_k}$

2.4. Sparse matrix format and singular value decomposition

Storing x_{ij}^{k} and $U_{\ell i}$ in a sparse matrix format allows us to deal with a matrix that has a dimension N greater than 10 million. It also enables us to apply SVD to these using the irlba package [11] in R [12], which is adapted for large sparse matrices.

2.5. UMAP

UMAP was applied to concatenated matrix, $V_{\ell' j}$, $\ell' \leq L$, by umap [13] function in R with the option `umap.defaults$n_neighbors <-30` (other options remain as default).

2.6. Estimation of distribution of single cells in UMAP

To quantify the distribution of single cells in UMAP, we divided the entire region into 10×10 regions. The region $S_{IJ}, 1 \leq I, J \leq 10$ is $\{(x,y)|(I-1)\Delta x \leq x \leq I\Delta x, (J-1)\Delta y \leq y \leq J\Delta y\}$ where $\Delta x = \frac{\max(x_j)-\min(x_j)}{10}$ and $\Delta y = \frac{\max(y_j)-\min(y_j)}{10}$ and x_j and y_j are coordinates of jth cell in UMAP embedding. The number of single cells, N_{IJ}, in the region S_{IJ} is equal to the number of $(x_j, y_j) \in S_{IJ}$.

2.7. UPGMA

UPGMA was performed with `hclust` function in R using option method='average'. Negative signed correlation coefficient of N_{IJ} between pairs of samples were used as distance.

2.8. Gene selection

As described in the previous study [7], we tried to select genes using u'_{2i}. P-values are attributed to the region i using

$$P_i = P_{\chi^2}\left[> \left(\frac{u'_{2i}}{\sigma_2} \right)^2 \right] \tag{4}$$

with excluding is with $u'_{2i} = 0$. $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution where the argument is larger than x and σ_2 is the standard deviation. P-values are corrected by BH criterion [7] and is associated with adjusted P-values less than 0.01 are selected.

2.9. Genome region annotation

The selected genomic regions are evaluated by `annotate_regions` function in annotatr package [14] within Biocoundiuctor [15].

3. Results

After obtaining concatenated matrix $V_{\ell' j}$ as described in the Methods section, UMAP was applied to $V_{\ell' j} \in \mathbb{R}^{L \times \sum_{k=1}^{8} M_k}$, i.e., the first L dimensions of $V_{\ell' j}$ (in this case, $L = 10$). Figure 2 shows the two-dimensional embedding of eight samples in Table 1. As observed in the investigated coordinates, the distribution of single cells in eight samples fully overlapped with each other. Nevertheless, the distribution of single cells seems to be somewhat tissue-specific. To quantify the similarity of distributions, we divided the whole region into 10×10 regions and counted the number of cells in the individual regions. Figure 3 shows the scatter plots and correlation coefficients of N_{IJ} between eight samples. The correlation coefficients between similar tissues were generally high. Whereas in few cases, correlation coefficients between different tissues were also high. To see if correlation coefficients were useful for classifying samples, we applied UPGMA (unweighted pair group method with arithmetic mean) to negatively signed correlation coefficients (Fig. 4). It is obvious that similar tissues were paired in the clustering. In addition to this, two CTX

samples were clustered together, apart from six ganglionic eminence samples (LGE, CGE, and MGE); which is also biologically reasonable.

Next, we tried to select genes using the results of proposed method and biologically evaluate selected genes. We used u'_{2i} among the ten $u'_{\ell' i}$ computed to select genes using eq. (4), since the second component was often more associated with biological features in the previous studies [7]. As a result, we selected 16,469 regions. This is only 0.1 % of all 13,627,618 regions. We have biologically evaluated selected regions as in Fig. 5 and

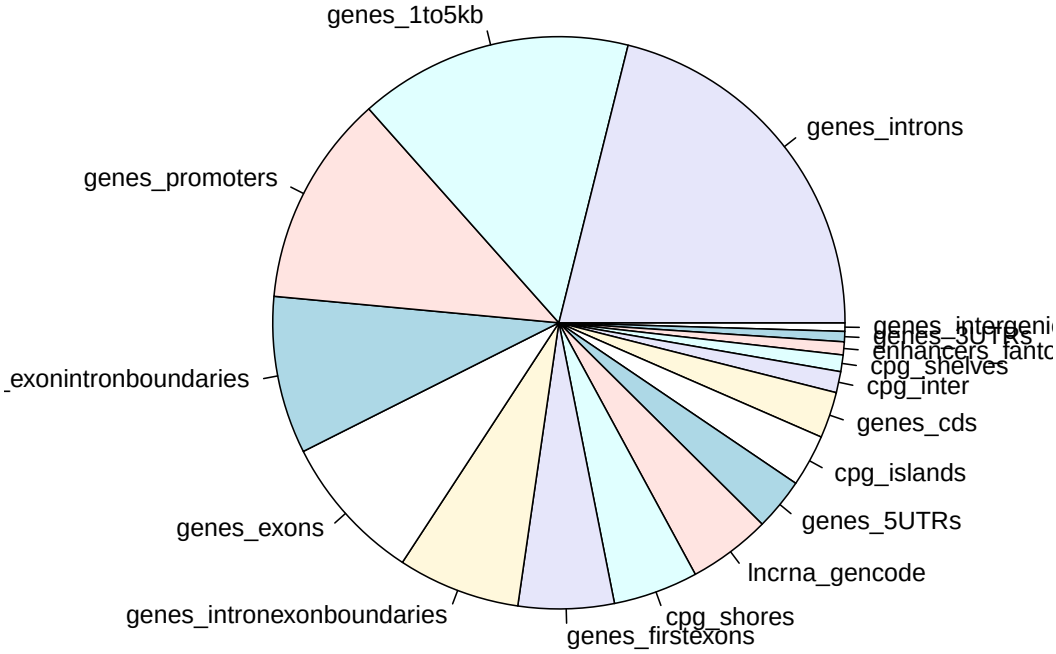

Figure 5. Pie chart of annotations by annotatr [14] about 16,469 regions selected by the proposed method.

Table 2 using annotatr [14]. It is obvious that selected regions are associated with numerous functional sites in spite of the very small number of selected regions compared with human genome (less than 0.01 % as mentioned above).

To further evaluate selected regions, we uploaded associated 1,147 gene symbols identified by annotatr to Enrichr [16] (the list of the 1,147 gene symbols is available as supplementary material). Then we have found many enrichment as follows (All of full versions of the following tables that list only top ten are available as supplementary materials). At first, transcription factors (TFs) are associated with selected genes from various aspects. Table 3 lists the top 10 TFs in "ENCODE and ChEA consensus TFs from ChIP-X" of Enrichr. As can be seen, *P*-values are very small, thus the results are very significant. Since scATAC-seq is supposed to detect open chromatin to which TFs bind, this is reasonable. Table 4 also lists yet another enrichment of TFs. Not only *P*-values are as

Table 2. Details of number of annotations in pie chart shown in Fig. 5

genes_intergenic	genes_3UTRs
909	1097
enhancers_fantom	cpg_shelves
1388	1772
cpg_inter	genes_cds
2250	4925
cpg_islands	genes_5UTRs
5555	5565
lncrna_gencode	cpg_shores
8806	9054
genes_firstexons	genes_intronexonboundaries
10218	13052
genes_exons	genes_exonintronboundaries
15803	16799
genes_promoters	genes_1to5kb
22589	29173
genes_introns	
39852	

Table 3. Top 10 TF in "ENCODE and ChEA Consensus TFs from ChIP-X" of Enrichr

Term	Overlap	P-value	Adjusted P-value
PBX3 ENCODE	238/1269	3.42×10^{-64}	3.55×10^{-62}
IRF3 ENCODE	158/663	3.81×10^{-56}	1.98×10^{-54}
NFYB ENCODE	429/3715	3.78×10^{-54}	1.31×10^{-52}
NFYA ENCODE	312/2250	7.23×10^{-54}	1.88×10^{-52}
FOS ENCODE	148/637	5.10×10^{-51}	1.06×10^{-49}
SP1 ENCODE	153/707	1.53×10^{-48}	2.65×10^{-47}
SP2 ENCODE	168/994	2.36×10^{-38}	3.51×10^{-37}
CREB1 CHEA	205/1444	1.74×10^{-35}	2.26×10^{-34}
RFX5 ENCODE	104/559	3.46×10^{-27}	4.00×10^{-26}
UBTF ENCODE	183/1631	2.20×10^{-19}	2.28×10^{-18}

Table 4. Top 10 TF in "ENCODE TF ChIP-seq 2015" of Enrichr

Term	Overlap	P-value	Adjusted P-value
IRF3 HepG2 hg19	162/755	6.11×10^{-51}	4.98×10^{-48}
IRF3 HeLa-S3 hg19	264/1809	6.24×10^{-49}	2.55×10^{-46}
FOS GM12878 hg19	210/1244	2.51×10^{-48}	6.83×10^{-46}
SP1 K562 hg19	203/1249	5.26×10^{-44}	1.07×10^{-41}
CHD2 MEL cell line mm9	255/1826	1.52×10^{-43}	2.10×10^{-41}
FOS K562 hg19	270/2000	1.54×10^{-43}	2.10×10^{-41}
SP2 H1-hESC hg19	185/1184	1.64×10^{-37}	1.91×10^{-35}
SP1 HCT116 hg19	219/1577	1.45×10^{-36}	1.47×10^{-34}
SP2 HepG2 hg19	211/1507	1.11×10^{-35}	1.01×10^{-33}
FOS HeLa-S3 hg19	146/860	1.85×10^{-33}	1.51×10^{-31}

significant as Table 3, but also some TFs, IRF3, SP1, and SP2, are commonly selected. Table 5 also lists additional TF enrichment. In contrast to Tables 3 and 4 that are based upon experiments, Table 5 is sequence (motif) based. It still has highly significant enrichment of TFs although significance steadily decreases. The results listed in Tables 3, 4, and 5 coincide with the fact that the scATAC-seq detects open chromatin to which TFs bind.

Next we consider tissue specificity. Even if selected genes are associated with TF target genes, if it is not related with tissues where HTS was performed, it is not trustable. Table 6

Table 5. Top 10 TF in "TRANSFAC and JASPAR PWMs" of Enrichr

Term	Overlap	P-value	Adjusted P-value
SP1 (mouse)	242/2360	1.56×10^{-20}	4.84×10^{-18}
PCBP1 (human)	142/1360	1.26×10^{-12}	1.96×10^{-10}
SP1 (human)	141/1406	2.95×10^{-11}	3.05×10^{-9}
TEAD2 (mouse)	154/1591	4.74×10^{-11}	3.67×10^{-9}
TEAD4 (human)	127/1354	1.93×10^{-8}	1.20×10^{-6}
TCFAP2A (human)	126/1367	6.01×10^{-8}	2.92×10^{-6}
SMAD4 (mouse)	141/1580	6.60×10^{-8}	2.92×10^{-6}
SP3 (human)	121/1332	2.47×10^{-7}	8.98×10^{-6}
EGR1 (mouse)	141/1617	2.61×10^{-7}	8.98×10^{-6}
E2F6 (human)	114/1358	2.26×10^{-5}	6.99×10^{-4}

Table 6. Top 10 experiments in "Allen Brain Atlas 10x scRNA 2021" of Enrichr

Term	Overlap	P-value	Adjusted P-value
Mouse 359 OPC down	96/841	6.78×10^{-11}	4.30×10^{-8}
Human Endo L2-5 NOSTRIN SRGN down	83/703	2.76×10^{-10}	8.75×10^{-8}
Mouse 372 SMC down	62/466	5.24×10^{-10}	8.77×10^{-8}
Mouse 375 VLMC down	68/535	5.52×10^{-10}	8.77×10^{-8}
Mouse 357 Astro down	77/651	1.15×10^{-9}	1.46×10^{-7}
Human Astro L1-6 FGFR3 PLCG1 down	95/880	1.67×10^{-9}	1.77×10^{-7}
Mouse 356 Astro down	73/617	3.10×10^{-9}	2.67×10^{-7}
Human VLMC L1-5 PDGFRA COLEC12 down	83/742	3.68×10^{-9}	2.67×10^{-7}
Mouse 374 VLMC down	64/513	3.78×10^{-9}	2.67×10^{-7}
Mouse 377 Micro down	50/365	9.54×10^{-9}	6.06×10^{-7}

lists top 10 experiments in "Allen Brain Atlas 10x scRNA 2021" of Enrichr. It is obvious that selected genes are also associated with tissue specificity. Interesting, Allen Brain Atlas, which does not consider single cell is not coincident with the selected genes (not shown here). This suggests that we have to take into account whether it is taken from bulk or single cells when we consider tissue specificity.

Table 7. Top 10 cells in "CellMarker Augmented 2021" of Enrichr

Term	Overlap	P-value	Adjusted P-value
Radial Glial cell:Undefined	6/11	1.26×10^{-5}	8.04×10^{-3}
Neural Stem cell:Brain	8/25	5.14×10^{-5}	1.46×10^{-2}
Neural Stem cell:Undefined	12/58	9.15×10^{-5}	1.46×10^{-2}
Purkinje cell:Brain	16/96	1.06×10^{-4}	1.46×10^{-2}
Natural Killer T (NKT) cell:Fetal Kidney	312/4543	1.41×10^{-4}	1.46×10^{-2}
Mesoderm cell:Undefined	16/99	1.55×10^{-4}	1.46×10^{-2}
Astrocyte:Embryonic Prefrontal Cortex	37/342	1.60×10^{-4}	1.46×10^{-2}
Cancer Stem cell:Brain	8/30	2.15×10^{-4}	1.71×10^{-2}
Pancreatic Polypeptide cell:Pancreas	6/18	3.59×10^{-4}	2.53×10^{-2}
Neural Progenitor cell:Embryonic Prefrontal Cortex	21/166	5.48×10^{-4}	2.74×10^{-2}

Although Table 7 also lists the associated brain tissue specificity, some other tissue specificity is also associated. Table 8 is full of transcription activities and Table 9 is full of DNA binding. It is also coincident with that scATAC-seq detects open chromatin.

All of these analyses suggest that the selected genes are biologically reasonable.

4. Discussion

Although TD can generate the feature that can cluster samples properly (Fig. 4), if other methods cannot do this, the proposed method is more efficient and unique in terms

Table 8. Top 10 terms in "GO Biological Process 2021" of Enrichr

Term	Overlap	P-value	Adjusted P-value
negative regulation of transcription, DNA-templated (GO:0045892)	117/948	1.90×10^{-15}	6.51×10^{-12}
positive regulation of transcription, DNA-templated (GO:0045893)	123/1183	6.35×10^{-11}	1.09×10^{-7}
negative regulation of transcription by RNA polymerase II (GO:0000122)	82/684	1.68×10^{-10}	1.92×10^{-7}
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	69/547	5.72×10^{-10}	4.90×10^{-7}
positive regulation of transcription by RNA polymerase II (GO:0045944)	97/908	1.93×10^{-9}	1.33×10^{-6}
negative regulation of nucleic acid-templated transcription (GO:1903507)	59/464	7.67×10^{-9}	4.38×10^{-6}
regulation of transcription by RNA polymerase II (GO:0006357)	188/2206	1.04×10^{-8}	5.08×10^{-6}
positive regulation of nucleic acid-templated transcription (GO:1903508)	62/511	1.92×10^{-8}	8.22×10^{-6}
regulation of transcription, DNA-templated (GO:0006355)	184/2244	2.44×10^{-7}	9.30×10^{-5}
negative regulation of neuron differentiation (GO:0045665)	10/24	3.47×10^{-7}	1.19×10^{-4}

Table 9. Top 10 terms in "GO Molecular Function 2021" of Enrichr

Term	Overlap	P-value	Adjusted P-value
sequence-specific double-stranded DNA binding (GO:1990837)	81/712	2.63×10^{-9}	1.61×10^{-6}
sequence-specific DNA binding (GO:0043565)	78/707	2.02×10^{-8}	6.17×10^{-6}
double-stranded DNA binding (GO:0003690)	71/651	1.39×10^{-7}	2.83×10^{-5}
DNA-binding transcription factor binding (GO:0140297)	29/208	8.46×10^{-6}	1.29×10^{-3}
transcription regulatory region nucleic acid binding (GO:0001067)	29/212	1.23×10^{-5}	1.50×10^{-3}
DNA binding (GO:0003677)	74/811	5.08×10^{-5}	5.18×10^{-3}
acetylation-dependent protein binding (GO:0140033)	7/21	1.14×10^{-4}	8.75×10^{-3}
lysine-acetylated histone binding (GO:0070577)	7/21	$.14 \times 10^{-4}$	8.75×10^{-3}
dihydropyrimidinase activity (GO:0004157)	4/6	1.47×10^{-4}	1.00×10^{-2}
mRNA binding (GO:0003729)	30/263	2.62×10^{-4}	1.61×10^{-2}

of tensor representation. Currently, there are very few tools to process scATAC-seq data with only matrix data. For example, although the extended data in Fig. 1 of the past study [17] summarizes ten *de facto* standard methods that can deal with a scATAC-seq data set, no methods can process the scATAC-seq data set with only matrix data. We also tried some methods [18–21] not included in the above list. No tools could efficiently process x_{ij}^k efficiently, it is possible due to the large N. Thus, the proposed method is the only one that can deal with a data set of this size. One of the reasons why the proposed method can handle this large N is that it stores the data set in a sparse matrix format. SVD performs using the tool adapted for the sparse matrix format, and we do not need to process dense format. Thus, the proposed method can deal with huge data sets. Of course, it is not a only reason because one of the most popular tools among these tools, signac, that can accept sparse matrix format cannot process this large data set as a whole at all (because of not enough memory) as mentioned above, either.

To further confirm the inferiority of signac to our method, we applied signac to two pairs of samples, that is, CTX1 and CTX2, as well as CTX1 and MGE1, since signac was unable to process the eight samples at once as mentioned above, although signac could accept a sparse matrix format in contrast to other methods specific to the scATAC sequence. Figure 6 shows the results when two signac-implemented strategies, merge and integration, are applied to the two pairs of samples, respectively. It is obvious that signac did not recognize that CTX1 and CTX2 are the same tissue, since CTX1 and CTX2 are not overlapped at all and are completely separated. In actual, the separation between CTX1 and CTX1 is similar to that between CTX1 and MGE1 that are not the same tissues. In addition to this, in contrast to CTX1 and CTX2 in Fig. 2 where they look similar, those in Fig. 6 do not look similar at all, either. Thus, ours are better than signac in recognizing identities of the sames tissues as well as distinction of the different tissues only using scATAC-seq data set.

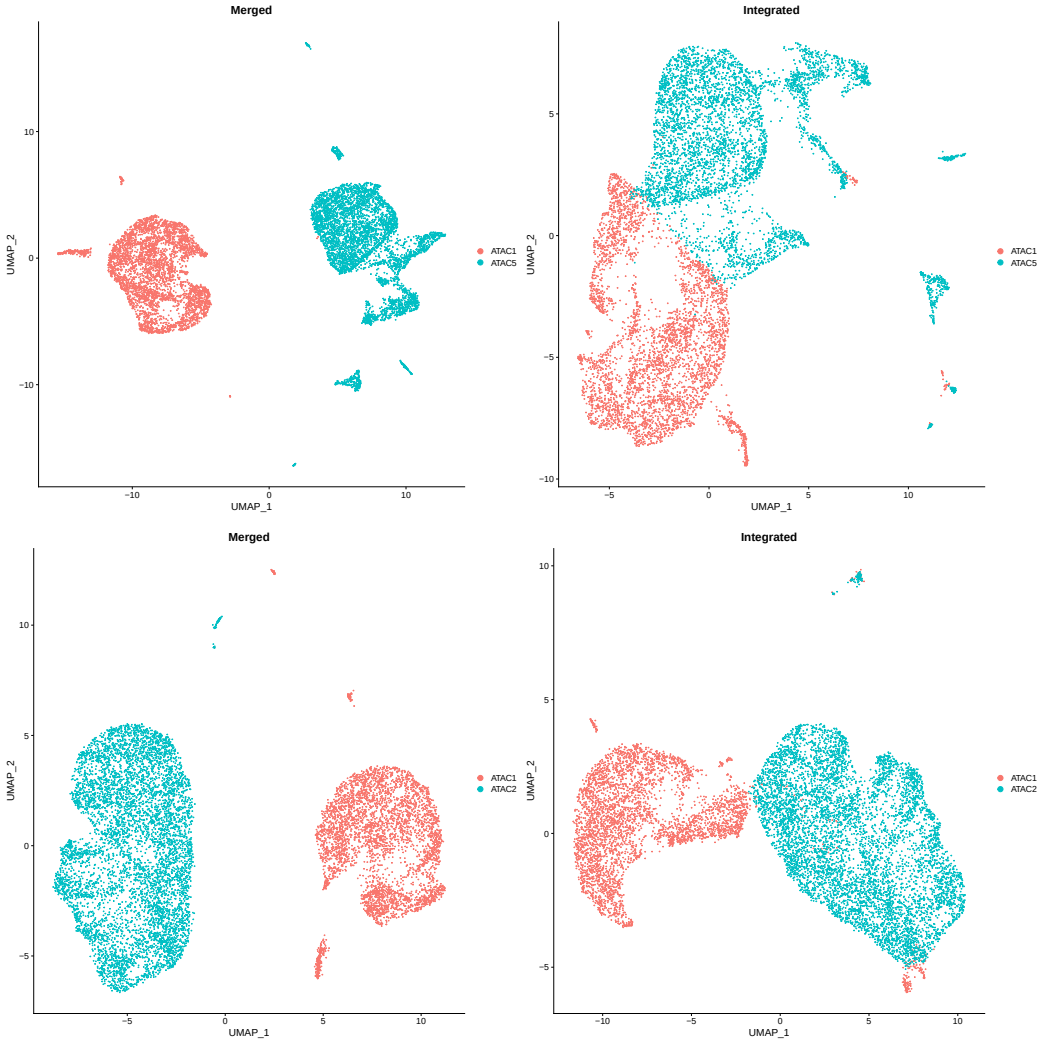

Figure 6. Upper: UMAP representation of signac applied to the pair of CTX1 and CTX2 (ATAC1(orange) and ATAC5(cyan) correspond to CTX1 and CTX2, respectively). Lower: UMAP representation of signac applied to the pair of CTX1 and MGE1 (ATAC1(orange) and ATAC2(cyan) correspond to CTX1 and MGE1, respectively).

5. Conclusions

In this paper, we applied TD to an scATAC-seq data set and the obtained embedding can be used for UMAP, following which the embedded material obtained by UMAP can differentiate tissues from which the scATAC sequence was retrieved. TD can deal with large sparse data sets generated by approximately 200 bp intervals, as these can be stored in a sparse matrix format. The large size of these data sets cannot be processed by any other methods. The proposed method is the only method that can deal with high-resolution native scATAC-seq data sets.

Author Contributions: Y.-H.T. planned the research and performed the analyses. Y.-H.T. and T.T. evaluated the results, discussions, and outcomes and wrote and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by KAKENHI (Grant Numbers 20K12067) to Y.-H.T.

Data Availability Statement: The data used in this study are available in GEO ID GSE167050. A sample of the R source can be found at https://github.com/tagtag/scATAC-seq (Accessed 30th September 2022).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

References

1. Grandi, F.C.; Modi, H.; Kampman, L.; Corces, M.R. Chromatin accessibility profiling by ATAC-seq. *Nature Protocols* **2022**, *17*, 1518–1552. https://doi.org/10.1038/s41596-022-00692-9.
2. Baek, S.; Lee, I. Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Computational and Structural Biotechnology Journal* **2020**, *18*, 1429–1439. https://doi.org/https://doi.org/10.1016/j.csbj.2020.06.012.
3. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **2019**, *177*, 1888–1902.e21. https://doi.org/https://doi.org/10.1016/j.cell.2019.05.031.
4. Satpathy, A.T.; Granja, J.M.; Yost, K.E.; Qi, Y.; Meschi, F.; McDermott, G.P.; Olsen, B.N.; Mumbach, M.R.; Pierce, S.E.; Corces, M.R.; et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **2019**, *37*, 925–936.
5. Giansanti, V.; Tang, M.; Cittaro, D. Fast analysis of scATAC-seq data using a predefined set of genomic regions [version 2; peer review: 2 approved]. *F1000Research* **2020**, *9*. https://doi.org/10.12688/f1000research.22731.2.
6. Buenrostro, J.D.; Wu, B.; Litzenburger, U.M.; Ruff, D.; Gonzales, M.L.; Snyder, M.P.; Chang, H.Y.; Greenleaf, W.J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **2015**, *523*, 486–490.
7. Taguchi, Y.H. *Unsupervised Feature Extraction Applied to Bioinformatics*; Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-22456-1.
8. Pan, X.; Li, Z.; Qin, S.; Yu, M.; Hu, H. ScLRTC: imputation for single-cell RNA-seq data via low-rank tensor completion. *BMC Genomics* **2021**, *22*. https://doi.org/10.1186/s12864-021-08101-3.
9. Mitchel, J.; Gordon, M.G.; Perez, R.K.; Biederstedt, E.; Bueno, R.; Ye, C.J.; Kharchenko, P.V. Tensor decomposition reveals coordinated multicellular patterns of transcriptional variation that distinguish and stratify disease individuals. *bioRxiv* **2022**, [https://www.biorxiv.org/content/early/2022/02/18/2022.02.16.480703.full.pdf]. https://doi.org/10.1101/2022.02.16.480703.
10. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Research* **2012**, *41*, D991–D995, [https://academic.oup.com/nar/article-pdf/41/D1/D991/3678141/gks1193.pdf]. https://doi.org/10.1093/nar/gks1193.
11. Baglama, J.; Reichel, L.; Lewis, B.W. *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*, 2021. R package version 2.3.5.
12. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.
13. Konopka, T. *umap: Uniform Manifold Approximation and Projection*, 2022. R package version 0.2.8.0.
14. Cavalcante, R.G.; Sartor, M.A. annotatr: genomic regions in context. *Bioinformatics* **2017**, *33*, 2381–2383, [https://academic.oup.com/bioinformatics/article-pdf/33/15/2381/25157896/btx183.pdf]. https://doi.org/10.1093/bioinformatics/btx183.
15. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **2015**, *12*, 115–121. https://doi.org/10.1038/nmeth.3252.
16. Kuleshov, M.V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.; Jagodnik, K.M.; Lachmann, A.; et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **2016**, *44*, W90–W97, [https://academic.oup.com/nar/article-pdf/44/W1/W90/18788036/gkw377.pdf]. https://doi.org/10.1093/nar/gkw377.
17. Granja, J.M.; Corces, M.R.; Pierce, S.E.; Bagdatli, S.T.; Choudhry, H.; Chang, H.Y.; Greenleaf, W.J. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **2021**, *53*, 403–411.
18. Xiong, L.; Xu, K.; Tian, K.; Shao, Y.; Tang, L.; Gao, G.; Zhang, M.; Jiang, T.; Zhang, Q.C. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* **2019**, *10*, 4576.
19. Li, Z.; Kuppe, C.; Ziegler, S.; Cheng, M.; Kabgani, N.; Menzel, S.; Zenke, M.; Kramann, R.; Costa, I.G. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat Commun* **2021**, *12*, 6386.
20. Kopp, W.; Akalin, A.; Ohler, U. Simultaneous dimensionality reduction and integration for single-cell ATAC-seq data using deep learning. *Nat Mach Intell* **2022**, *4*, 162–168.
21. Stuart, T.; Srivastava, A.; Madad, S.; Lareau, C.; Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods* **2021**. https://doi.org/10.1038/s41592-021-01282-5.