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, 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 reach 13,627,618. Currently, no other methods can deal with large sparse matrices. The proposed method could not only provide UMAP embedding that coincides with tissue specificity, but also 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

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

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Biomolecules* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

ATAC-seq [1] is a powerful tool to profile chromatin accessibility. In particular, scATAC-seq [2] can profile the accessibility of chromatin throughout the genome within individual cells and several tools have ever been invented [3–8] for scATAC-seq. Although scATAC-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 [9], 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. [10] 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. [11] used a predefined set of genomic regions to interpret the scATAC-seq. Buenrostro et al. [12] 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) [13] 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 [14,15], to our knowledge, there are no studies that applied TD to scATAC-seq.

2. Materials and Methods

Figure 1 shows the analysis flow chart.

2.1. scATAC-seq profiles

The scATAC-seq data set we analyzed in this study is obtained from Gene Expression Omnibus (GEO) [16], the ID is GSE167050 [17]. We used eight samples of GSM5091379

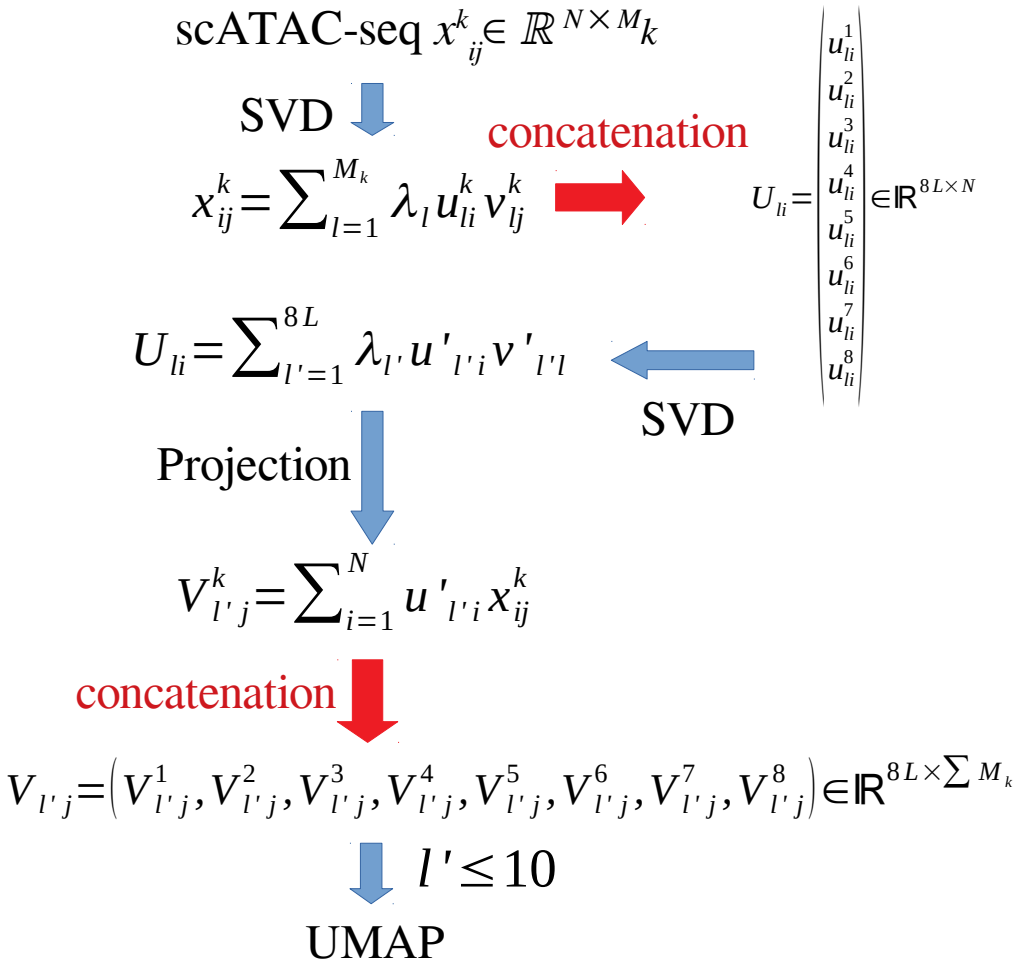


Figure 1. Flow chart of the analysis

to GSM5091386 from four mice tissues (CTX, MGE, CGE and LGE) with two replicates each (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. Nuclei isolation followed the 10X Genomics ATAC nuclei isolation

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	total
number of single cells (M_k)	4,108	6,845	4,013	6,577	4,946	3,465	4,530	4,769	39,253

protocol with several modifications. All steps were performed on ice. snATAC reaction was carried out following 10X Genomics ATAC User Guide (revision C), libraries were prepared following 10X Genomics and Illumina guidelines.

Additional data set for validation is from ID GSE139950 [18] (Table 2). They are mice kidney of unilateral ureter obstruction at Day 0, 2 and 10.. All protocols to generate scATAC-seq data on the 10x Chromium platform, including sample prep, library prep, instrument and sequencing settings.

Table 2. Number of single cells included in the files analyzed in this study.

	Day 0		Day 2		Day 10		
Biological replicates	1	2	1	2	1	2	total
number of single cells (M_k)	10,521	7,499	5,828	9,792	8,103	5,636	47,379

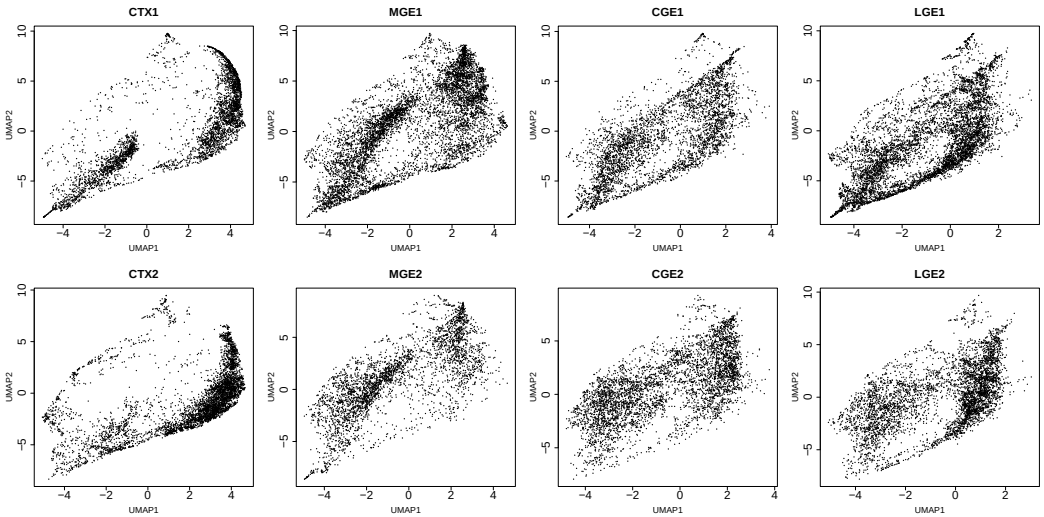


Figure 2. UMAP embedding of eight samples analyzed in this study (see Table 1)

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 the 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 mouse genome bps by the interval length, i.e. 2.5 billion by 200.

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. Please note that x^k_{ij} is standadized as

$$\sum_i x^k_{ij} = 1 \tag{2}$$

(in other word, x^k_{ij} is replaced with $x^k_{ij} / \sum_i x^k_{ij}$) before SVD is applied. 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

$$U_{\ell i} = \sum_{\ell'=1}^{8L} \lambda_{\ell'} u'_{\ell' i} v'_{\ell' \ell} \tag{3}$$

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.

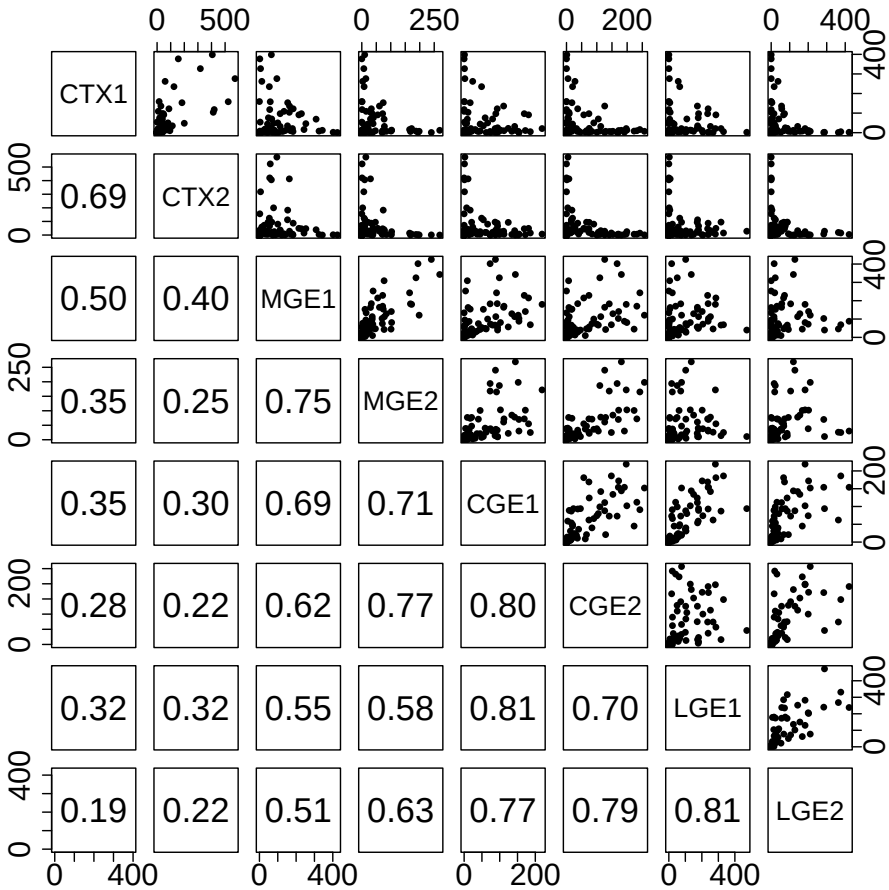


Figure 3. Upper: scatter plot of the number of single cells in Fig. 2, $N_{IJ}, 1 \leq I, J \leq 10$, in one of 10×10 regions, S_{IJ}. Lower: Kendall's's correlation coefficients of N_{IJ} between eight samples.

Low-dimensional embedding can be obtained as

$$V^k_{\ell' j} = \sum_{i=1}^{N} u'_{\ell' i} x^k_{ij} \in \mathbb{R}^{8L \times M_k}. \tag{4}$$

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

The storage of x^k_{ij} 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 [19] in R [20], which is adapted for large sparse matrices.

2.5. UMAP

UMAP was applied to concatenated matrix, $V_{\ell' j}$, $\ell' \leq L$, by umap [21] 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 individual 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

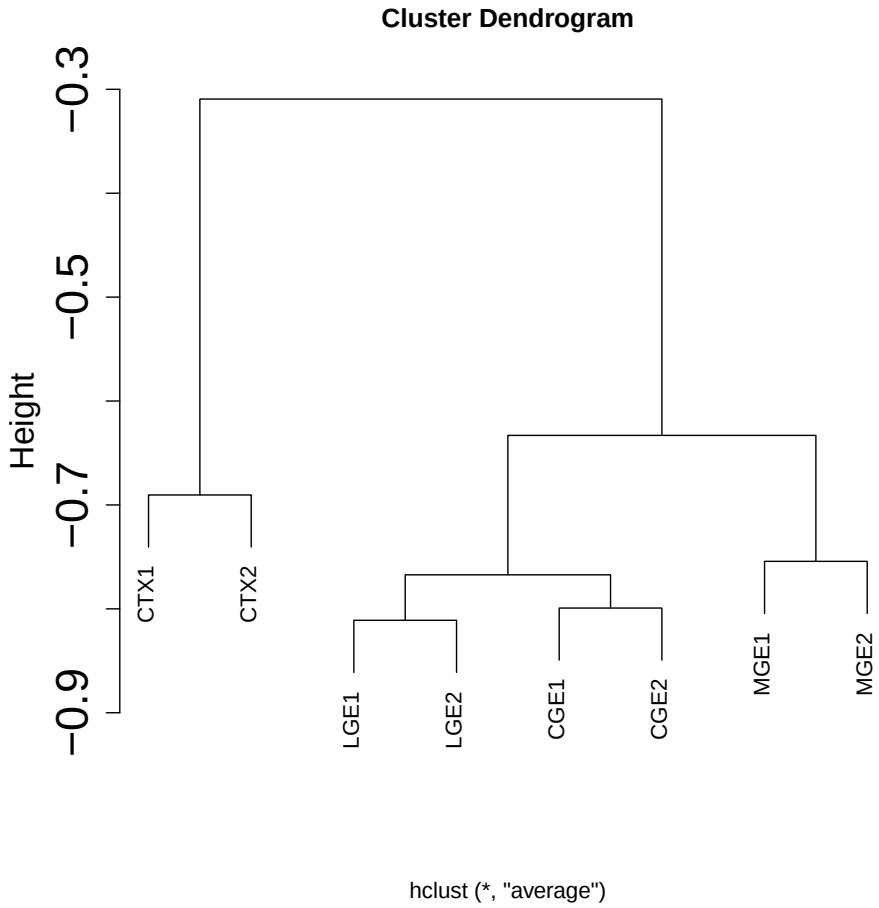


Figure 4. Hierarchical clustering by UPGMA of eight samples (Table 1) using coordinates obtained by UMAP. Distance is represented by negatively signed correlation coefficients in Fig. 3

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 [13], we tried to select genes using u'_{2i}. P-values are attributed to the region i using

$$P_i = P_{\chi^2}\left[> \left(\frac{u'_{2i}}{\sigma_2}\right)^2\right] \tag{5}$$

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 [13] 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 [22] within Biocoundiuctor [23].

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 completely overlapped with each other. Nevertheless, the distribution of individual 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 $u'_{\ell' i}$ calculated to select genes using eq. (5), since the second component was often more associated with biological features in the previous studies [13]. 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 Table S1 using annotatr [22]. 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 1,147 associated gene symbols identified by annotatr to Enrichr [24] (the list of 1,147 gene symbols is available as supplementary material). Then we have found many enrichments as follows (all of full versions of the following tables that list only the top ten are available as supplementary materials). At first, transcription factors (TF) are associated with selected genes of various aspects. Table S2 lists the top 10 TFs in "ENCODE and ChEA consensus TFs from ChIP-X" of Enrichr. As can be seen, the P-values are very small, and thus the results are very significant. Since scATAC-seq is supposed to detect open chromatin to which TFs bind, this is reasonable. Table S3 also lists yet another enrichment of TFs. Not only P-values are as significant as Table S2, but also some TFs, IRF3, SP1, and SP2, are commonly selected. Table S4 also lists additional enrichment with TF. In contrast to Tables S2 and S3 that are based upon experiments, Table S4 is sequence (motif) based. It still has highly significant enrichment of TFs, although its significance steadily decreases. The results listed in Tables S2, S3,and S4 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 S5 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 cells or single cells when we consider tissue specificity.

Although Table S6 also lists the associated brain tissue specificity, some other tissue specificity is also associated. Table S7 is full of transcription activities, and Table S8 is full of DNA binding. It is also coincident with the fact that scATAC-seq detects open chromatin.

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

Although we have made use of scATAC-seq to investigate tissue specificity, if we can discuss about distinction of the cell diversity between individual tissues, it is much better. In our analysis, distinction of the cell diversity between individual tissues should be reflected in that of distribution of single cells embedding in Fig. 2. Thus, we selected

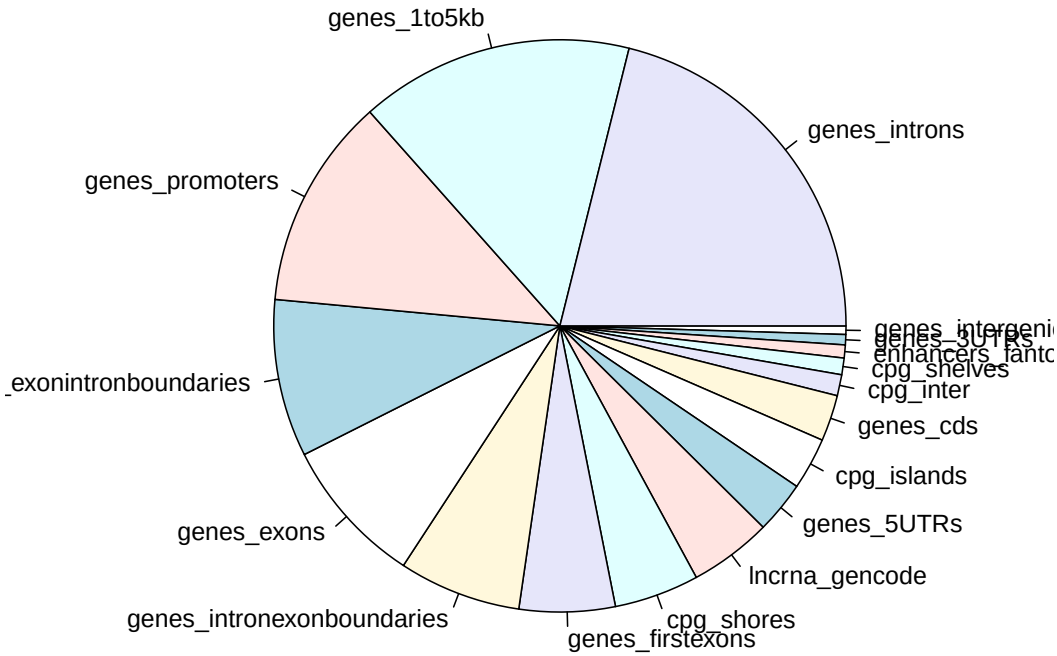


Figure 5. Pie chart of annotations by annotatr [22] about 16,469 regions selected by the proposed method.

top 10 N_{IJ}s whose difference between CTX1 or CTX2 and MGE1 or MGE2 is large. Figure 6 suggest N_{IJ}s selected are coincident among pairs of CTX and MGE. Thus, these likely represents single cell level difference between tissues. To understand biological aspects, we investigated which $V_{\ell' j}$s are most distinct between single cells in (I, J)s associated with top 10 N_{IJ}s. Then we found that $\ell' = 10$ and 7 correspond to the cases CTX $>$ MGE and CTX $<$ MGE, respectively. Then using the corresponding U_{10i} and U_{7i}, we selected 3,138 and 2,510 regions, respectively. Downstream analyses applied to the 16,469 regions in the above were repeatedly applied to these 3138 and 2510 regions, too. Figure 7 shows the annotation pie chart. It is obvious that it differs from Fig. 5. Accessible genomics regions enhanced in Fig. 7 are more enriched with introns and intergenic genomic regions. This suggests that cell type specific genome accessibility is more likely associated with non-exon or non-gene body. Enrichment analyses (see supplementary materials) also shows the clear difference from 16,469 regions. For example, GO BP for CTX $>$ MGE is enriched with synaptic functions. It is reasonable since CTX is more likely associated with synaptic activity than MGE. This means that distinct distribution of single cells in embedding between tissues can kelp us to understand single cell level diversity as well. To our knowledge, no other methods could make use of the difference of single cell distribution in embedding to characterize the single cell level diversity.

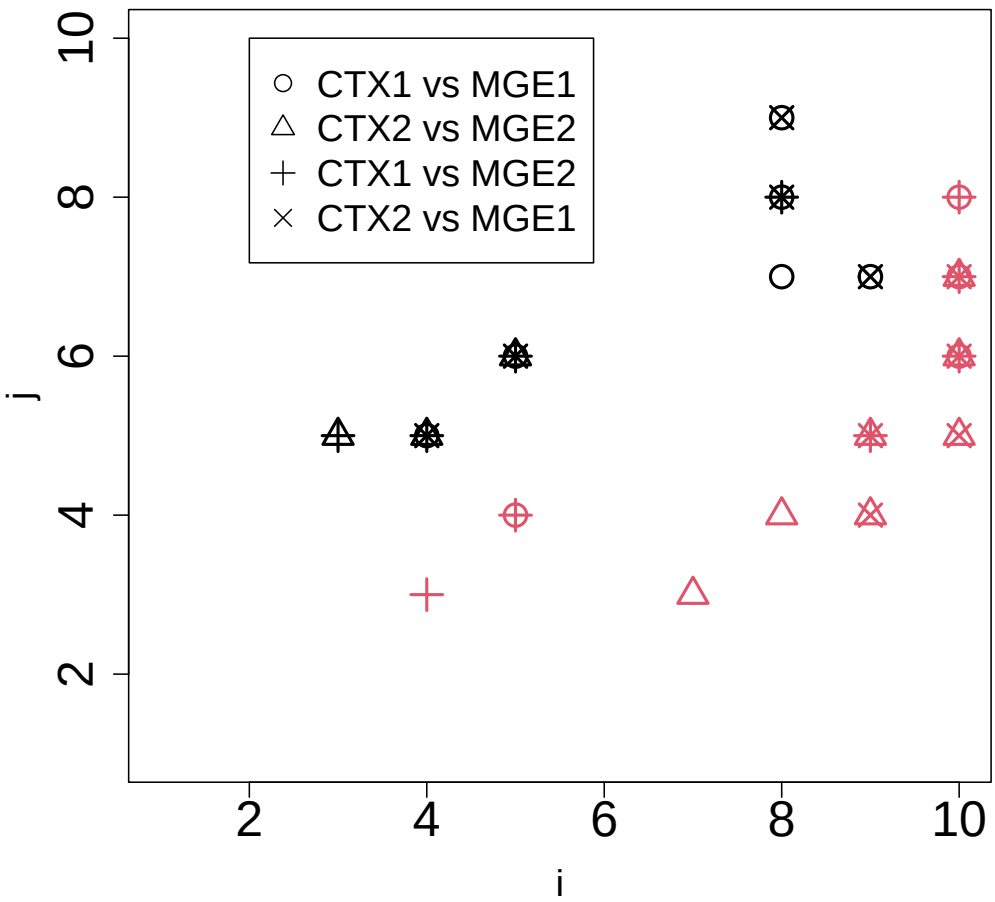


Figure 6. Top 10 N_{IJ}s whose difference between CTX1 or CTX2 and MGE1 or MGE2 is large. Red and black marks correspond to CTX1 or CTX2 larger than MGE1 or MGE2 and CTX1 or CTX2 less than MGE1 or MGE2, respectively. Horizontal and vertical axes represent I and J, respectively.

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 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 [25] 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 [26–29] 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 8 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 do not overlap and are completely separated. In reality, 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. 8 do not look similar. Thus, ours is better than signac in recognizing identities of the sames tissues as well as distinction of the different tissues only using scATAC-seq data set.

One might also wonder if the above excellent performances are accidental and are not reproducible if other data sets are considered. In order to deny this possibility, we consider another data sets and repeated applied the above procedure to it (Figs. 9, 10 and 11). Since it is obvious that embedding is coincident between biological replicates but distinct between not biological replicate samples. Thus it is unlikely that tissue specific embedding is accidentally obtained specific to only one set of samples. Our methods are robust and are expected to be applied to wide range of samples for scATAC-seq.

5. Conclusions

In this paper, we applied TD to an scATAC-seq data set, and the obtained embedding can be used for UMAP, after 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 method. 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.

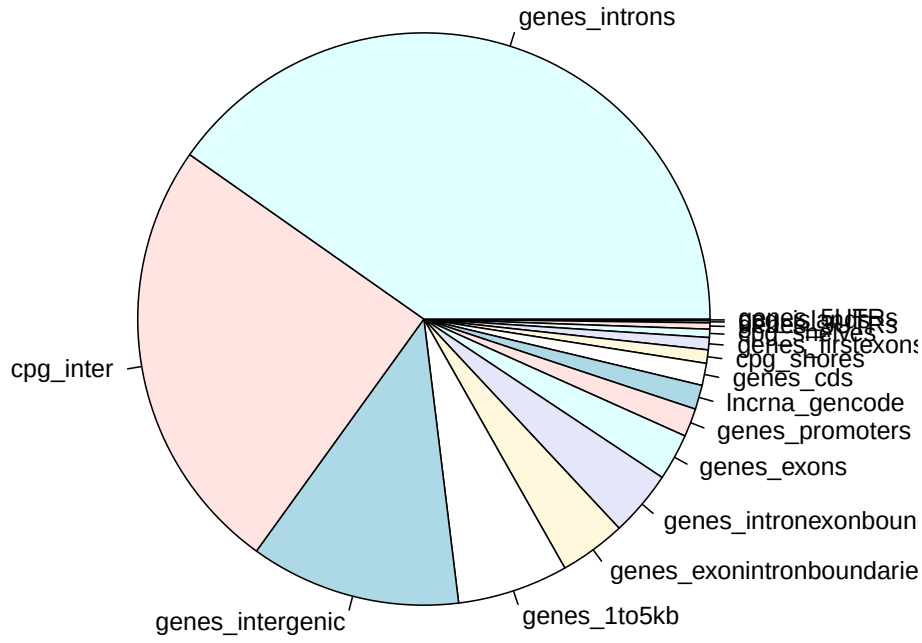
Data Availability Statement: The data used in this study are available in GEO ID GSE167050 and GSE139950. A sample of the source R can be found at https://github.com/tagtag/scATAC-seq (accessed 30 September 2022).

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. Chen, H.; Lareau, C.; Andreani, T.; Vinyard, M.E.; Garcia, S.P.; Clement, K.; Andrade-Navarro, M.A.; Buenrostro, J.D.; Pinello, L. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology* **2019**, *20*. https://doi.org/10.1186/s13059-019-1854-5.
4. Fang, R.; Preissl, S.; Li, Y.; Hou, X.; Lucero, J.; Wang, X.; Motamedi, A.; Shiau, A.K.; Zhou, X.; Xie, F.; et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature Communications* **2021**, *12*. https://doi.org/10.1038/s41467-021-21583-9.
5. Xu, Y.; Begoli, E.; McCord, R.P. sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network. *npj Systems Biology and Applications* **2022**, *8*. https://doi.org/10.1038/s41540-022-00245-6.
6. Guo, H.; Yang, Z.; Jiang, T.; Liu, S.; Wang, Y.; Cui, Z. Evaluation of classification in single cell atac-seq data with machine learning methods. *BMC Bioinformatics* **2022**, *23*. https://doi.org/10.1186/s12859-022-04774-z.
7. Yuan, H.; Kelley, D.R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature Methods* **2022**, *19*, 1088–1096. https://doi.org/10.1038/s41592-022-01562-8.

8. Song, Q.; Zhu, X.; Jin, L.; Chen, M.; Zhang, W.; Su, J. SMGR: a joint statistical method for integrative analysis of single-cell multi-omics data. *NAR Genomics and Bioinformatics* **2022**, *4*. https://doi.org/10.1093/nargab/lqac056.
9. 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.
10. 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.
11. 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.
12. 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.
13. Taguchi, Y.H. *Unsupervised Feature Extraction Applied to Bioinformatics*; Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-22456-1.
14. 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.
15. 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.
16. 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.
17. Lee, D.R.; Rhodes, C.; Mitra, A.; Zhang, Y.; Maric, D.; Dale, R.K.; Petros, T.J. Transcriptional heterogeneity of ventricular zone cells in the ganglionic eminences of the mouse forebrain. *eLife* **2022**, *11*, e71864. https://doi.org/10.7554/eLife.71864.
18. 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. *Nature Communications* **2021**, *12*. https://doi.org/10.1038/s41467-021-26530-2.
19. 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.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.
21. Konopka, T. *umap: Uniform Manifold Approximation and Projection*, 2022. R package version 0.2.8.0.
22. Cavalcante, R.G.; Sartor, M.A. annotatr: genomic regions in context. *Bioinformatics* **2017**, *33*, 2381–2383, [https://academic.oup.com/bioinfor pdf/33/15/2381/25157896/btx183.pdf]. https://doi.org/10.1093/bioinformatics/btx183.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.



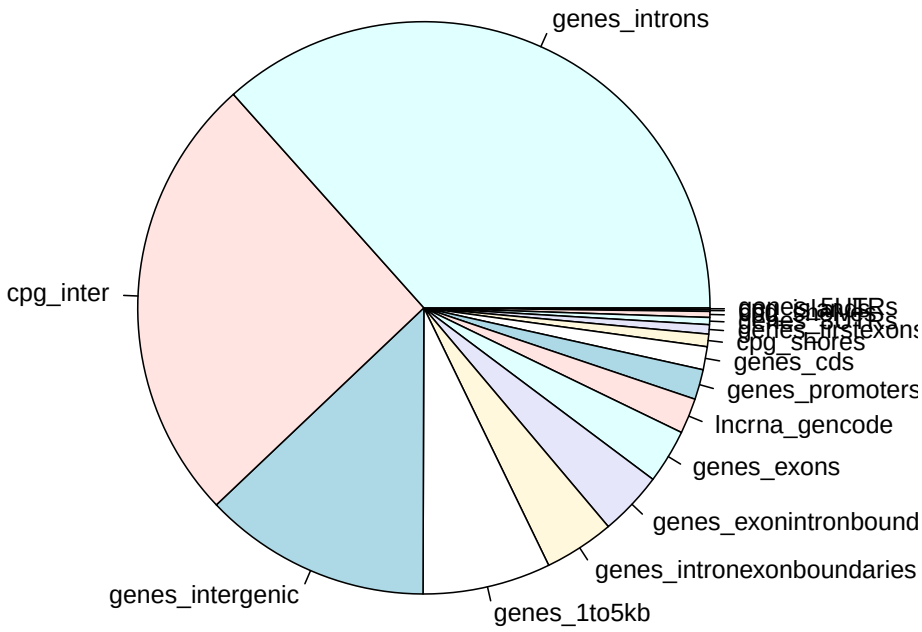


Figure 7. Pie chart of annotations generated as in Fig. 5, but for 3,138 and 2,510 regions associated with N_{IJ} of CTX > MGE (upper) and CTX < MGE (lower), respectively.

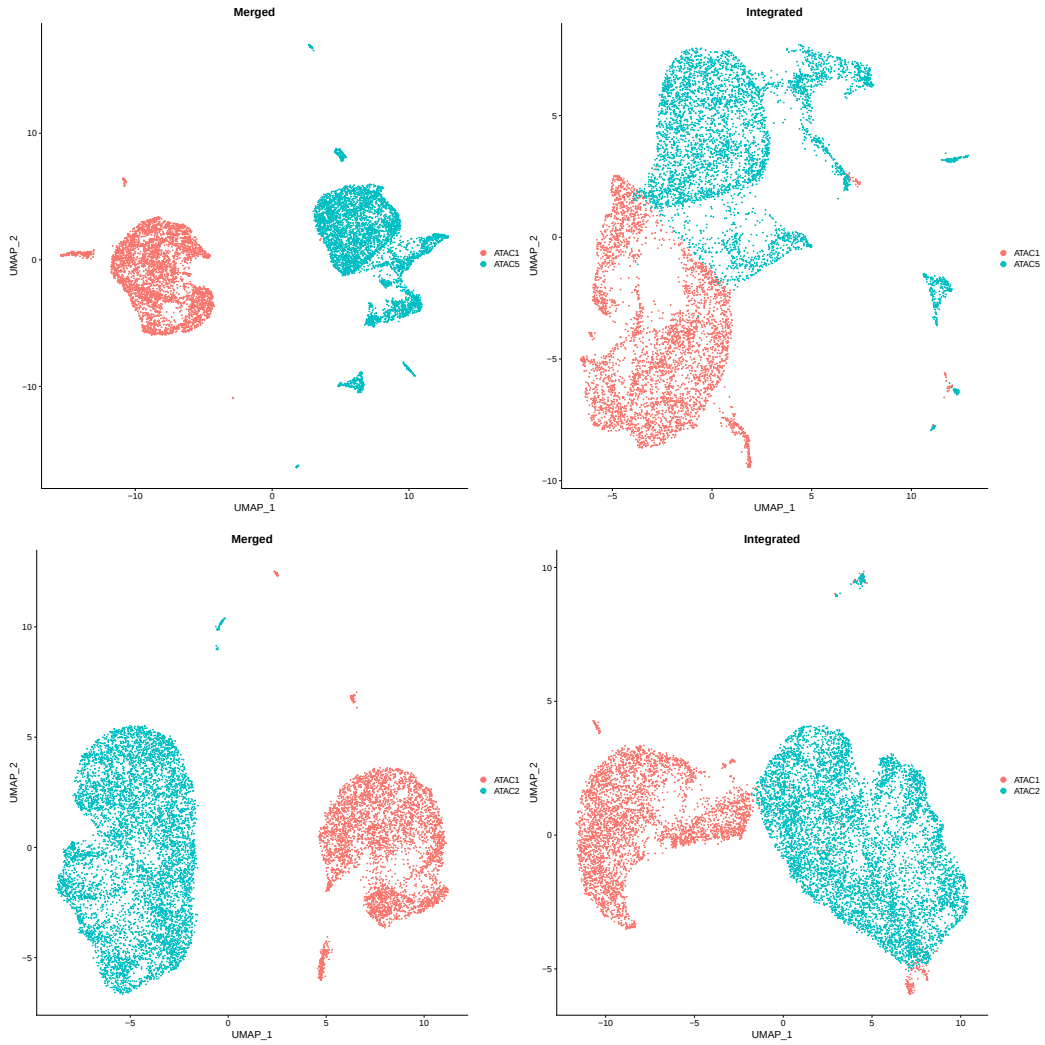


Figure 8. Upper: UMAP representation of signac applied to the pair of CTX1 and CTX2 (ATAC1(orange) and ATAC5(cyan) correspond to CTX1 and CTX2, respectively). Lower: UMAP representation of signac applied to the pair of CTX1 and MGE1 (ATAC1(orange) and ATAC2(cyan) correspond to CTX1 and MGE1, respectively).

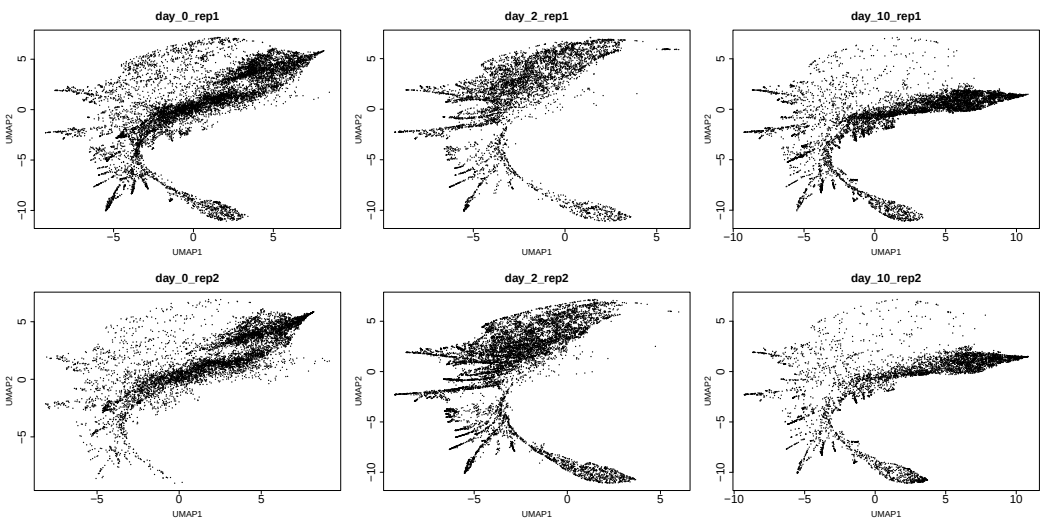


Figure 9. UMAP embedding of six samples analyzed in this study (see Table 2)

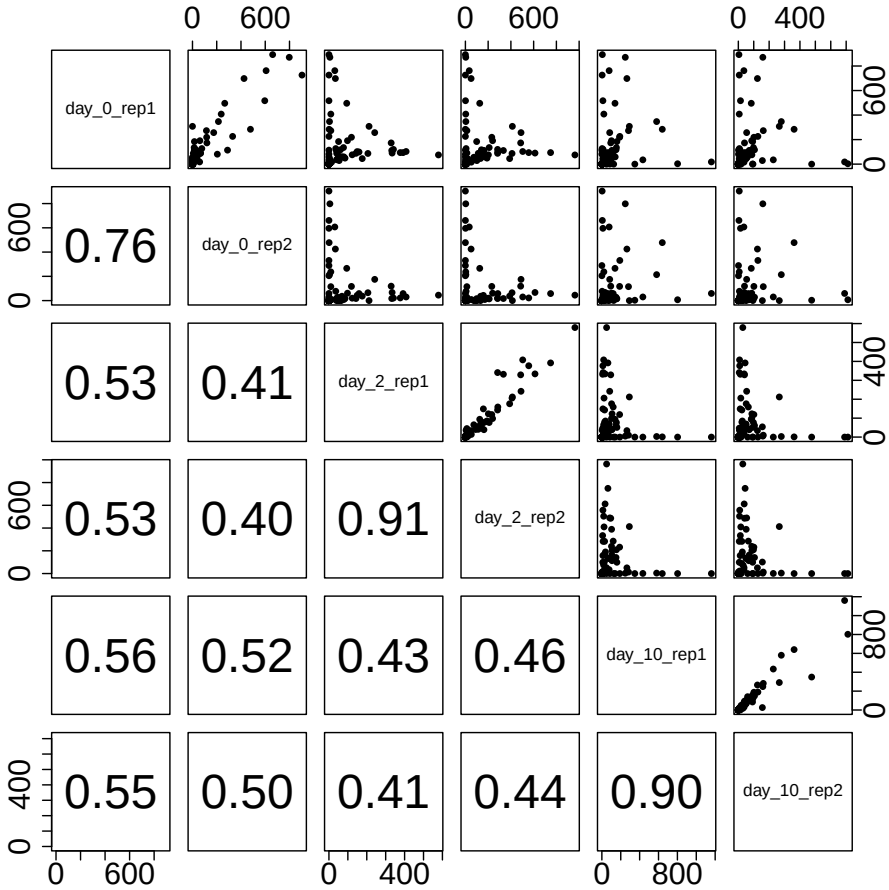


Figure 10. Upper: scatter plot of the number of single cells in Fig. 9, $N_{IJ}, 1 \leq I, J \leq 10$, in one of 10×10 regions, S_{IJ}. Lower: Kendall's's correlation coefficients of N_{IJ} between six samples.

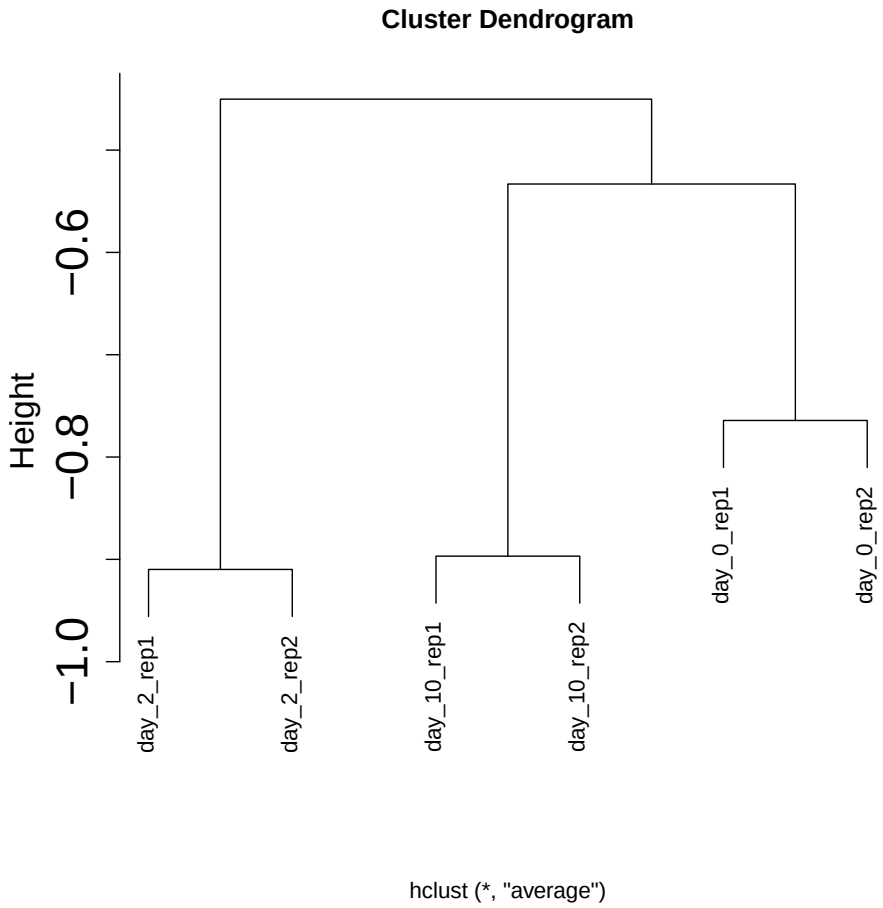


Figure 11. Hierarchical clustering by UPGMA of six samples (Table 2) using coordinates obtained by UMAP. Distance is represented by negatively signed correlation coefficients in Fig. 10