Briefings in Bioinformatics, 2025, **26(1)**, bbae615
https://doi.org/10.1093/bib/bbae615
Problem Solving Protocol

TriTan: an efficient triple nonnegative matrix factorization method for integrative analysis of single-cell multiomics data

Xin Ma[1,*], Lijing Lin[2], Qian Zhao[1], Mudassar Iqbal[1,*]

[1]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Rd, Manchester, M13 9PL, United Kingdom
[2]Centre for Health Informatics, Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Rd, Manchester, M13 9PL, United Kingdom

*Corresponding authors: Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Rd, Manchester, M13 9PL, United Kingdom. E-mail: maxine_c137@qq.com, mudassar.iqbal@manchester.ac.uk

Abstract

Single-cell multiomics have opened up tremendous opportunities for understanding gene regulatory networks underlying cell states by simultaneously profiling transcriptomes, epigenomes, and proteomes of the same cell. However, existing computational methods for integrative analysis of these high-dimensional multiomics data are either computationally expensive or limited in interpretation. These limitations pose challenges in the implementation of these methods in large-scale studies and hinder a more in-depth understanding of the underlying regulatory mechanisms. Here, we propose TriTan (Triple inTegrative fast non-negative matrix factorization), an efficient joint factorization method for single-cell multiomics data. TriTan implements a highly efficient factorization algorithm, greatly improving its computational performance. Three matrix factorization produced by TriTan helps in clustering cells, identifying signature features for each cell type, and uncovering feature associations across omics, which facilitates the identification of domains of regulatory chromatin and the prediction of cell-type-specific regulatory networks. We applied TriTan to the single-cell multiomics data obtained from different technologies and benchmarked it against the state-of-the-art methods where it shows highly competitive performance. Furthermore, we showed a range of downstream analyses conducted utilizing TriTan outputs, highlighting its capacity to facilitate interpretation in biological discovery.

Keywords: multi-omics; single cell; machine learning; gene regulation

Introduction

Single-cell multiomics technologies can simultaneously profile multiple omics within the same cell, including transcriptomics, epigenomics, and proteomics [1]. These approaches provide a more comprehensive view of cellular function as they capture multiple layers of regulatory machinery of the cell. Integration of multiomics information can reveal new insights into the underlying mechanisms that regulate gene expression and how these may alter in response to different stimuli. With the rapid advancement in these technologies, we face a complex data integration challenge. The unique characteristics of data from each modality, such as sparsity, dimension, and scale, make it difficult to synthesize a coherent joint representation. In order to effectively combine information from different omics, it is necessary to design bespoke methodologies that are efficient in handling the increasing scale of data and flexible enough to incorporate more omics as they become available.

Several recent studies have offered innovative ways of integrating single-cell multiomics data. One category of methods focuses on finding relationships among data from different omics and then performing the integration. For example, Seurat v4 [2] utilizes the weighted nearest neighbors (WNNs) approach to assign cell-specific weights, allowing it to effectively integrate multiple modalities by learning the contribution of each modality to the overall cellular profile. Similarly, Schema [3] employs a metric learning strategy to identify and extract the informative features from each modality, followed by an integration step that combines these features, enabling a more precise representation of cellular states across different omics. Another approach consists of projecting the feature spaces from different omics into a common latent space. In this category, MOFA+ [4] performs low-dimensional reconstructions for each modality and integrates them. It models variations across multiple modalities using factor modeling with variational inference. MOJITOO [5] efficiently infers multiple modalities' shared representations through canonical correlation analysis (CCA). Integrative nonnegative matrix factorization (iNMF) [6] extends NMF to multiomics for joint cell clusters identification. Mowgli [7] combines iNMF with optimal transport to improve data integration and capture relationships across multiomics. Besides, deep generative models have been employed to perform single-cell multiomics data integration. Among these, totalVI [8] uses a variational autoencoder (VAE) model to jointly analyze CITE-seq data (RNA and

Received: June 12, 2024. **Revised:** October 15, 2024. **Accepted:** November 19, 2024

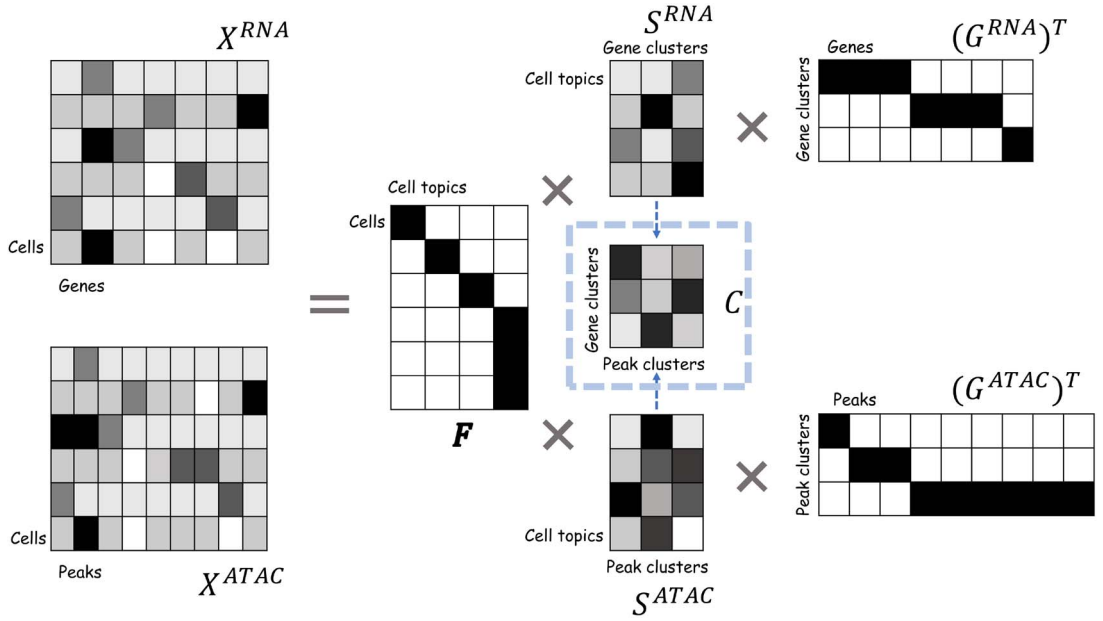

Figure 1. Schematic of TriTan. It receives two (or more) paired single-cell matrices whose rows represent the same cells, where each matrix represents a particular omic. In this example, scRNA-seq and scATAC-seq are being considered. Each input matrix is decomposed into three matrices where F is a cell cluster matrix shared across all modalities. Each column of F represents a distinct cell cluster and each row contains a single nonzero element indicating the cluster assignment of the corresponding cell. G^{RNA} and G^{ATAC} are feature cluster matrices whose columns represent different feature clusters and rows contain a single nonzero element indicating the cluster assignment of the corresponding feature. Middle matrices S^{RNA} and S^{ATAC} (association matrices) are condensed representations of the input matrices, which can be considered as associations between cell clusters and feature clusters. After the factorization, we calculated the matrix C using S^{RNA} and S^{ATAC} to show feature correlations across modalities.

cell surface proteins from the same cell), representing data as a composite of biological and technical factors. CLUE [9] considers cross-encoders in a multi-modal VAE and constructs a joint latent cell representation. Although deep learning methods have shown great potential in clustering cells, they inherently lack biological interpretability thus limiting their utility for downstream analysis.

The majority of the methods discussed above have a primary focus on the estimating embedding at the cell level. Among them, matrix factorization methods have become popular as they provide a low-dimensional representation of the data as well as feature contributions for the discovered topics. However, standard two-matrix factorization methods suffer from several limitations, including slow computational speed on larger datasets, arbitrariness in determining the number of cell topics, and no systematic way to find signatures for inferred cell topics or the associations across omics. The critical need to address above limitations and find the functional roles of features in individual omics, as well as the associations between them, motivates us to look beyond the existing two-matrix factorization methods. We have designed TriTan, an efficient and interpretable method that decomposes the input single-cell multiomics matrices into low-dimensional matrices (see cartoon in Fig. 1). TriTan does not rely on predefined parameters, such as the rank of factor matrices and uses an automated approach for identifying the number of cell and feature clusters, making it a more flexible and adaptive tool for analyzing complex datasets.

The rest of the manuscript is organized as follows. In the Materials and methods section, we present a detailed, step-by-step derivation and explanation of TriTan. In the Results section, we benchmarked TriTan against state-of-the-art methods, Seurat v4 (WNN), MOFA+, and MOJITOO using three public multiomics datasets. Furthermore, through an exemplar analysis using PBMC-10k data, we show TriTan's utility in identifying cell-type-specific signature feature sets, and uncovering cross-omics associations.

Materials and methods

TriTan

TriTan is designed to address the following problem: A dataset $\{X^n \in \mathbb{R}^{m\times D_n}\}_{n=1}^N$ includes N single-cell data from different omics, where m represents the number of cells shared across modalities, and D_n represents the number of features in modality n. We aim to discover K_1 cell clusters that are shared across all modalities, and simultaneously, find K_2^n feature clusters for the nth modality. Note that K_1 and K_2^n are not predetermined but rather learned by TriTan. Additionally, TriTan discovers modality-specific association matrices, representing the weights of feature clusters across all cell clusters.

Overall, TriTan learns a representation for each modality denoted as $X^n \approx FS^n(G^n)^T$. In this representation, F is a binary matrix shared across all modalities, where each row has a single nonzero element indicating the cluster assignment for the corresponding cell. G^n is a binary feature cluster matrix where each row has a single nonzero element indicating the cluster assignment for the corresponding feature. S^n represents the modality-specific association matrix. In the paper, $\|\cdot\|$ denotes the norm $\|A\|^2 = \mathrm{Tr}(AA^T)$, and $A_{i\cdot}$ and $A_{\cdot j}$ represent the ith row and jth of column of A, respectively.

An efficient multiplicative update algorithm for nonnegative matrix tri-factorization

Traditional nonnegative matrix tri-factorization (NMTF) methods [10], which infer three nonnegative latent matrices based on multiplicative update (MU) [11], suffer from slow computational speed

and considerable memory consumption. To accelerate NMTF on large-scale data, we design a novel optimization algorithm to speed up the factorization. Firstly, we introduce our factorization approach for the scenario of a single matrix X. We aim to minimize the loss function

$$L = \left\| X - FSG^T \right\|^2.$$

L can be minimized over F, G, and S in an alternating fashion. Since F and G are binary cluster assignment matrices, the computation of F and G can be simplified. With fixed S and G, letting $U = SG^T$, in order to minimize $L = \sum_{i=1}^{m} \|X_{i\cdot} - F_{i\cdot}U\|^2$, it suffices to minimize $\|X_{i\cdot} - F_{i\cdot}U\|^2$ for each row i of X. Since $F_{i\cdot}$ has only one nonzero element, this is equivalent to $\mathbf{min}_k \|X_{i\cdot} - U_{k\cdot}\|^2$. Therefore, for the elements of $F_{i\cdot}$:

$$F_{ij} = \begin{cases} 1 & j = \arg\min_k \|X_{i\cdot} - U_{k\cdot}\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

The computation of G, given fixed F and S, follows the same strategy as the above. With $V = FS$, for each column j of X, we aim to minimize $\|X_{\cdot j} - V(G^T)_{\cdot j}\|^2 = \|X_{\cdot j} - VG_{j\cdot}\|^2$, which is equivalent to $\mathbf{min}_k \|X_{\cdot j} - V_{\cdot k}\|^2$. Therefore, for the elements of $G_{i\cdot}$:

$$G_{ij} = \begin{cases} 1 & j = \arg\min_k \|X_{\cdot i} - V_{\cdot k}\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

The computation of S, given fixed F and G, uses the following procedure:

$$\frac{\partial L}{\partial S} = -2F^TXG + 2F^TFSG^TG = 0,$$
$$S = (F^TF)^{-1}F^TXG(G^TG)^{-1}.$$

Reducing the dimensions during MUs

To enhance the efficiency of the factorization method described above, we propose to use singular value decomposition (SVD) to reduce dimensions during MUs. This is mainly designed to accelerate the factorization while minimizing the loss of accuracy. It also helps to reduce the imbalance in the number of features between different modalities (for instance, the number of features can vary significantly between scRNA-seq data and scATAC-seq data).

For a given modality n, input matrix $X^n \in \mathbb{R}^{m \times D_n}$ is firstly decomposed using SVD: $X^n = P^n\mathbf{\Sigma}^n(Q^n)^T$, where $(P^n)^TP^n = I$, $(Q^n)^TQ^n = I$ and $\mathbf{\Sigma}^n$ is a rectangular diagonal matrix with nonnegative numbers (singular values) on the diagonal. We approximate X^n by a truncated SVD, i.e. $X^n \approx P_l^n\mathbf{\Sigma}_l^n(Q_l^n)^T$ using the first l SVD-components where l can be chosen using a scree plot of the singular values in $\mathbf{\Sigma}^n$ [12]. In the methodology description, we use the notation l consistently. However, in practice, different numbers of components are employed for the cell and feature spaces, taking into account the specific datasets based on the corresponding singular value scree plots (see details in supplementary Table. S4). Also our software package allows for user-defined number of components for reducing the feature space through SVD decomposition for each modality. To reduce computational time and memory usage, F^n and G^n in the previous section are updated with X^n projected to the lower dimension spaces $X^nQ_l^n$ and $(P_l^n)^TX^n$, respectively (as shown below in the equations (2) and (4)).

Joint MU for single-cell multiomics data

In previous sections, we described our efficient implementation for NMTF for a single modality. Building on that, here we present TriTan's joint MUs scheme for single-cell multiomics data.

Considering the variation in data attributes across different omics, we have introduced an omic-specific weights matrix, $\tilde{w}^n = \mathrm{diag}(\tilde{w}_i^n)$, a diagonal matrix with entry $\tilde{w}_i^n$ representing the weight for cell i in modality n.

$$\tilde{L} = \sum_{n=1}^{N} \left\| \tilde{w}^n X^n - FS^n(G^n)^T \right\|^2.$$

Instead of assigning weights to each modality based on prior knowledge, we adopt an alternative strategy, approximating the omic-specific cell weights $\tilde{w}^n$ by w^n within the updates, as shown below in the equations (3).

The iterative process is outlined as follows. In the $(t+1)$th iteration, ${}^{(t+1)}F^\star \in \mathbb{R}^{m \times K_1'}$ (K_1' is a hyperparameter representing the initial estimate for the number of cell clusters, $K_1' < K_1$), an intermediate weighted cell cluster matrix, is constructed from two components: ${}^{(t+1)}w^n$ and ${}^{(t+1)}F^n$, as per the formula

$${}^{(t+1)}F_{i\cdot}^\star = \sum_{n=1}^{N} {}^{(t+1)}w_i^n \cdot {}^{(t+1)}F_{i\cdot}^n. \tag{1}$$

In this equation, ${}^{(t+1)}F^n$ represents the inference of cell types from individual omics data and is updated with ${}^{(t)}U^n = {}^{(t)}S^n \cdot {}^{(t)}G^{nT}$:

$${}^{(t+1)}F_{ij}^n = \begin{cases} 1 & j = \arg\min_k \|(X^nQ_l^n)_{i\cdot} - ({}^{(t)}U^nQ_l^n)_{k\cdot}\|^2, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

And ${}^{(t+1)}w^n$, which reflects the contribution of each omic to the cell, is calculated based on the prior state ${}^{(t)}F^\star$:

$$E^n = \left| X^nQ_l^n - {}^{(t)}F^\star \cdot {}^{(t)}S^n \cdot {}^{(t)}G^{nT}Q_l^n \right|,$$
$$e_i^n = \frac{1}{\|E_{i\cdot}^n\|/\mu(E_{i\cdot}^n)}, \quad {}^{(t+1)}w_i^n = \frac{e_i^n}{\sum_{n=1}^{N} e_i^n}. \tag{3}$$

This w^n accounts for both the reconstruction loss and inter-modality biases (e.g. scRNA-seq data usually have much more average counts than scATAC-seq data), aiding in the inference of a shared low-dimensional cell representation shown in the next section and the integration of different omics.

${}^{(t+1)}G^n$ is then computed for each modality using ${}^{(t+1)}F^\star$ and ${}^{(t+1)}V^n = {}^{(t+1)}F^\star \cdot {}^{(t)}S^n$:

$${}^{(t+1)}G_{ij}^n = \begin{cases} 1 & j = \arg\min_k \left\| (P_l^{nT}X^n)_{\cdot i} - (P_l^{nT} \cdot {}^{(t+1)}V^n)_{\cdot k} \right\|^2, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We then update ${}^{(t+1)}S^n$ for each modality:

$$A = {}^{(t+1)}F^\star, \quad B = {}^{(t+1)}G^n,$$
$${}^{(t+1)}S^n = (A^TA)^{-1}(A^TX^nB)(B^TB)^{-1}. \tag{5}$$

This iterative process allows us to progressively infer and integrate the data structure from individual omics.

Model selection strategy

After $t^\star$ iterations described in the previous section (also see Algorithm 1), we have obtained an initial weighted shared matrix $F^\star \in \mathbb{R}^{m \times K_1'}$, offering preliminary insights into the cells. We then use it to obtain initial cell cluster assignments c for each cell i:

$$c_i = \arg\max_k F^\star_{ik} \tag{6}$$

Subsequently, by employing w^n, a shared cell embedding is constructed, effectively integrating the information from each omic and presenting the unique characteristics to describe each cell:

$$Z = \left[w^1(X^1 Q_l^1), \ldots, w^N(X^N Q_l^N)\right]. \tag{7}$$

Then, we rearrange rows of Z into a row block matrix, denoted as $\tilde{Z}$, where the kth block $\tilde{Z}_k$ corresponds to cells being assigned to the kth cluster (based on equations (6)), i.e.

$$\tilde{Z} = \begin{bmatrix} \tilde{Z}_1 \\ \vdots \\ \tilde{Z}_{K_1'} \end{bmatrix}, \qquad \tilde{Z}_k = \left[Z_{i\cdot}\Big|_{c_i=k}\right]. \tag{8}$$

We next input $\tilde{Z}$ to HDBSCAN [13] for a local structure search. For each $\tilde{Z}_i$, HDBSCAN applies a density-based transformation to the data space, allowing it to identify stable dense points. The final shared cell cluster assignment matrix F in TriTan's output is then updated:

$$F_{ij} = \begin{cases} 1 & j = \arg\min_k \|Z_{i\cdot} - U^c{}_{k\cdot}\|^2, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Here, $U^c \in \mathbb{R}^{K_1 \times N_l}$ represents the matrix formed by concatenating the centroids of each cluster of dense data points, where K_1 represents the number of centroids.

Features space is typically more complex compared with cells. Therefore, we employ an adaptive approach to infer feature clusters. For all omics, G^n are initialized in the space of $\mathbb{R}^{K_2 \times D_n}$, where K_2, a hyperparameter, is set to 100 by default, which exceeds the expected number of feature clusters. During the main factorization iterations, the matrices F, S^n, and G^n are alternately updated, with the objective of minimizing the overall objective function. Initially, G^n may contain many columns filled exclusively with zeros, indicating that no features have yet been assigned to those clusters, as per the aforementioned equation. As the iterations progress, the number of all-zero columns in these matrices gradually decreases, allowing the inferred number of feature clusters to approach the desired count and contributing to the reduction of the objective function. When the value of the objective function reaches a predetermined threshold, the count of nonzero columns in G^n corresponds to the final number of feature clusters K_2^n we aim to identify.

This iterative process allows us to flexibly adapt to the complexity and structure of the data, ensuring accurate representation and meaningful interpretation of the feature space.

Single-cell multiomics datasets

We have used multiple publicly available single-cell multiomics (two modalities—scRNA and scATAC) datasets of varying sizes to comprehensively benchmark TriTan against other methods and leverage the outputs from TriTan for extensive downstream analyses. These include data generated using 10X Multiome from humanperipheralbloodmononuclearcells and bone marrow mononuclear cells (NeurIPS 2021, GSE194122) as well as mouse skin cells (GSM4156597) using SHARE-seq protocol [14]. These data are summarized in Table 1.

The NeurIPS 2021 dataset [15], which is currently the largest dataset we used, was created as part of the NeurIPS competition and is utilized as a benchmark for the computational complexity of integration methods.

Algorithm 1 TriTan Algorithm

Input: $\{X^n\}_{n=1}^N$, convergence threshold $t^\star$, and tolerance ε
Initialization: $\{G^n\}_{n=1}^N$ and $\{S^n\}_{n=1}^N$ are initialized from Uniform distribution $U(0,1)$, weights $\{w_n = diag(1/N)\}_{n=1}^N$
Set iteration counter $t = 0$
repeat
  $t \leftarrow t + 1$
  for $n = 1$ **to** N **do**
    Update F^n using Equation (2)
  end for
  Update $F^\star$ using Equation (1)
  for $n = 1$ **to** N **do**
    Update matrices G^n, S^n, and weights w^n using Equations(4), (5), and (3)
  end for
until $t \geq t^\star$
Update clustering assignments c using Equation (6)
Define Z and $\tilde{Z}$ using Equations (7) and (8)
Update F using Equation (9)
repeat
  $t \leftarrow t + 1$
  for $n = 1$ **to** N **do**
    Update matrices G^n and S^n using Equations (4) and (5)
  end for
until $L = \sum_{n=1}^N \left\|X^n - FS^n(G^n)^T\right\|^2 < \epsilon$
Output: $F, Z, \{G^n\}_{n=1}^N, \{S^n\}_{n=1}^N$

Data preprocessing

We performed uniform preprocessing across all datasets using Scanpy package for each method. For scRNA-seq matrices, we performed `sc.pp.filter_genes()`, `sc.pp.normalize_total()`, `sc.pp.log1p()`, and `sc.pp.highly_variable_genes()` to filter low quality genes. For scATAC-seq modality, we first ran TF-IDF transformation and then used `sc.pp.log1p()` and `sc.pp.highly_variable_genes()` to filter low-quality peaks. We used different parameters in regard to different datasets and different omics, respectively, for the above functions. The corresponding parameter values are shown in Supplementary Table S3.

Implementation of competing methods

We selected three most popular and representative methods in single-cell multiomics data integration area. Below we describe these methods.

Seurat v4 (WNN):

The pipeline of Seurat v4 [2] uses WNN approach for single-cell multiomics datasets. For each modality, it reduces the data dimension, learns cell-specific weights, and constructs an integrated WNN graph. We followed the recommended tutorial

Table 1. Multiomics datasets

Dataset	Protocol	Species	No. of cells	No. of cell types
PBMC-10K	10X Multiome	Human	11 787	13
Skin-SHARE	SHARE-seq	Mouse	34 774	23
NeurIPS 2021	10X Multiome	Human	69 249	22

from Seurat v4 website (https://satijalab.org/seurat/articles/multi-modal_vignette.html) and executed Seurat v4 (WNN) in R for all datasets reported in this manuscript.

MOFA+

MOFA+ [4] is a popular method based on factor analysis that provides a general framework for the integration of multiomics datasets in an unsupervised manner. We followed the recommended tutorial from MOFA+ website (https://biofam.github.io/MOFA2/tutorials.html) and ran MOFA+ in Python with default parameters for all datasets.

MOJITOO

MOJITOO [5] is a highly efficient method published recently, which uses CCA to detect a shared representation for single-cell multiomics data. MOJITOO takes *SeuratObject* as the input and implements CCA for the identification of a shared representation of cells. We ran MOJITOO in R for all the datasets, following the tutorial (https://github.com/CostaLab/MOJITOO/blob/main/vignettes/SeuratObject_integration.Rmd).

Results

Benchmarking of single-cell multi-modal integration methods

We evaluated TriTan and existing methods described above using three publicly available multiomics datasets (Table 1).

Assessing the computational requirements of processing single-cell multiomics data with an ever-increasing number of cells is a highly important metric in evaluating the effectiveness of methods. To do this, we randomly sampled cells from NeurIPS 2021 data and generated six datasets with an increasing number of cells, from 3000 to 69 249. We evaluated TriTan and three methods using these seven datasets and analyzed the time and memory requirements in our benchmark by running them on a parallel 16-core job with a 512GB RAM/core node. For all methods, we used their default settings. Computational running time and memory comparisons are shown in Fig. 2A where TriTan shows a very competitive performance overall. It is the fastest among all the methods, except in 60 000 cells dataset with MOJITOO is slightly faster. However, TriTan is not the most efficient method regarding memory requirements. WNN and MOJITOO, which are quite similar, have the lowest memory requirements across all seven datasets and TriTan ranks third.

Next, we used ground truth labels to evaluate the ability of TriTan and other methods to accurately recover cell clusters. We employed the Adjusted Rand Index (ARI) [16], which considers all pairs of samples and counts pairs that are assigned in the same or different clusters in the predicted and ground truth labels, and Normalized Mutual Information (NMI), which measures how much information the clustering result shares with the ground truth. TriTan achieves the highest ARI scores in both the PBMC-10K and Skin-SHARE datasets, while ranking close second in the NeurIPS 2021 dataset. Regarding NMI scores, TriTan holds the second-highest scores in the PBMC-10K and Skin-SHARE datasets, but slightly worse than MOJITOO and WNN in the NeurIPS 2021 dataset, as shown in Fig. 2B.

Besides, we employed silhouette scores, using ground truth labels and inferred shared cell embedding from each method, to measure how compact cells of the same type are in the joint embedding. As shown in Fig. 2B, TriTan obtains the second rank in all datasets, while WNN is the best in all datasets. Additionally, the structure preservation score is an important metric used to evaluate how well a method maintains the structure of individual modalities (e.g. RNA and ATAC) after data integration. As seen in Fig. 2B, TriTan consistently demonstrates competitive and balanced performance in preserving the structures of both RNA and ATAC modalities. In PBMC-10K dataset, it has the highest ATAC structure preservation score, while it performs the best in RNA structure preservation in NeurIPS 2021 dataset. In Skin-SHARE dataset, TriTan ranks first for both RNA and ATAC preservation scores. MOJITOO also captures information from all individual modalities uniformly and performs the best in ATAC structure preservation in the NeurIPS 2021 dataset. WNN shows better performance than TriTan in RNA structure preservation in the PBMC-10K dataset but less in ATAC preservation.

Overall, this benchmarking analysis shows that TriTan is highly competitive against state-of-the-art methods in terms of the accuracy of clustering recovery, the inference of the informative latent embedding. Importantly, TriTan distinguishes itself by offering advantages in terms of speed and interpretability, which enhances downstream analysis and facilitates hypothesis generation, as discussed in the following sections.

TriTan finds signature feature sets for each cell topic using association matrices

In this section, we explored the utility of association matrices produced by TriTan for different modalities, highlighting the additional information they provide that is lacking in the typically used two-matrix NMF methods.

Association matrices show the associations of multiple feature clusters with cell topics, as illustrated in Fig. 3A. The two heatmaps on the top display normalized association matrices derived from scATAC-seq and scRNA-seq data. The bottom two heatmaps present binarized association matrices using a fixed threshold (we have chosen a strict threshold of 0.9 in this instance), highlighting those feature sets that are most significant in defining and differentiating cell types.

Through binarized association matrices, we have identified some feature clusters that are unique to certain cell types. For example, geneset-38, geneset-0, and geneset-28, along with peakset-13, peakset-18, and peakset-24, exhibit a unique association with plasmacytoid dendritic cells (pDCs), geneset-9 and peakset-6 are double negative T (DNT) cells specific, while peakset-26 and geneset-46 and geneset-48 exclusively identifies natural killer (NK) cells. These signature gene sets and peak sets are collectively associated with a specific cell type and potentially domains of regulatory chromatin (DORCs).

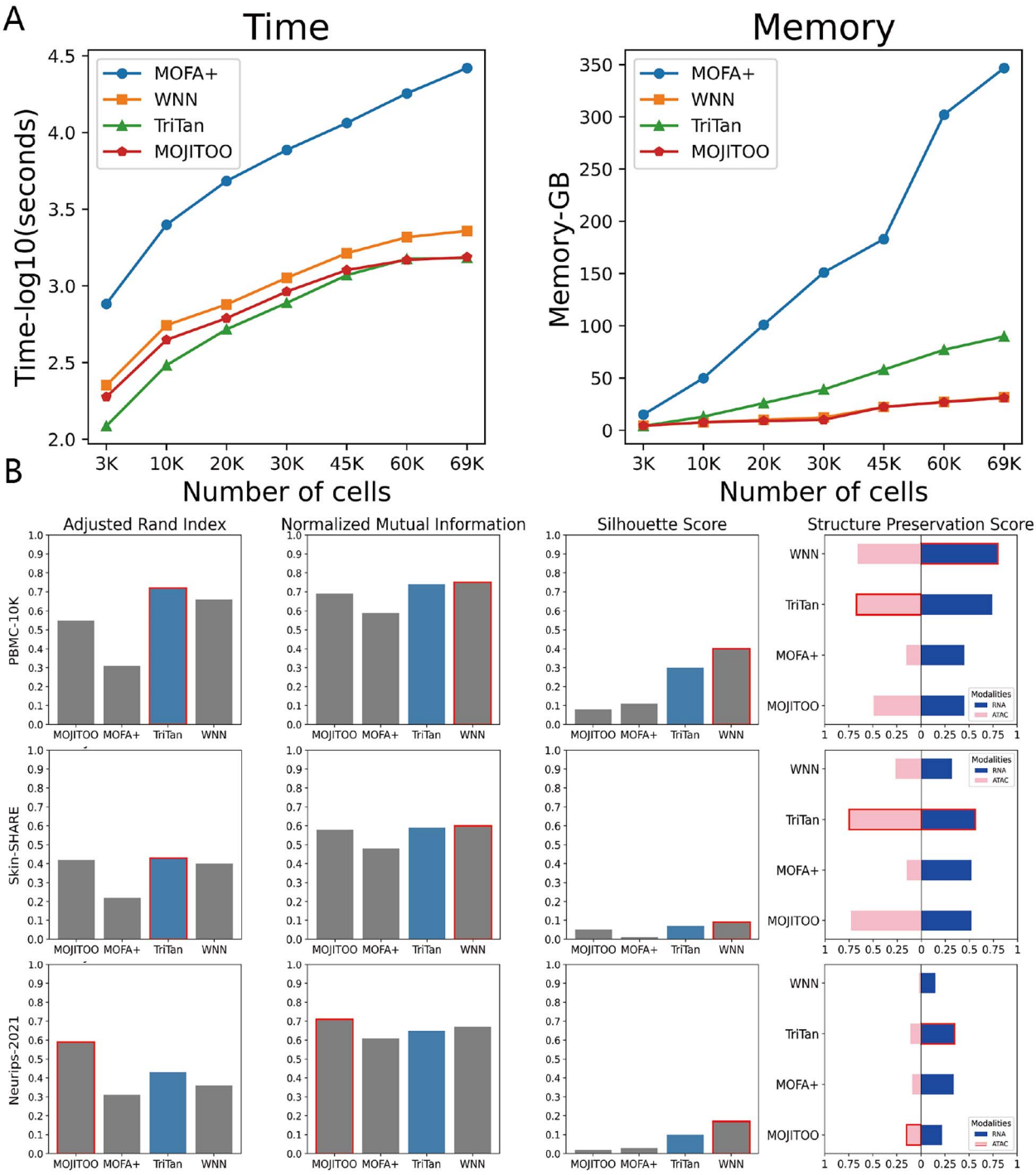

Figure 2. (**A**) Time and memory usage for TriTan and three competing methods against different number of cells (x-axis) randomly sampled from the NeurIPS 2021 data. For each method, the left subplot shows elapsed time (log10 of seconds) and the right subplot shows peak memory (Gigabytes) required by each method. (**B**) Comparisons of methods in terms of ARI, NMI, silhouette scores, and structure preservation scores for three different datasets. The top-ranked method is outlined with red boxes.

Besides, effective integration of multiomics data can enhance cell type identification compared with mono-omic analyses. As observed in Supplementary Fig. S1, scRNA-seq alone could not effectively distinguish between CD-8 effector and NK cells. However, the incorporation of scATAC-seq significantly improves the separation between these two cell types. Peakset-26, peakset-4, and peakset-23 are uniquely associated with NK cells, while peakset-37, peakset-0, and peakset-43 demonstrate marked differences in CD-8 effector cells compared with other cell types. In addition to identifying signature gene sets and peak sets with known cell types, TriTan can also identify potential subtypes within the ground truth. For example, cell clusters 8, 10, and 18 detected by TriTan were all labeled as pro-B cells according to the ground truth; however, their gene expression levels and peak counts varied, suggesting potential differentiation or distinct stages within the pro-B cell population.

To validate that the signature gene sets we identified are truly cell-type-specific and can serve as, or include, markers for specific cell types, we performed cell marker enrichment analysis using clusterProfiler 4.0 [17]. We use CellMarker 2.0 [18] human

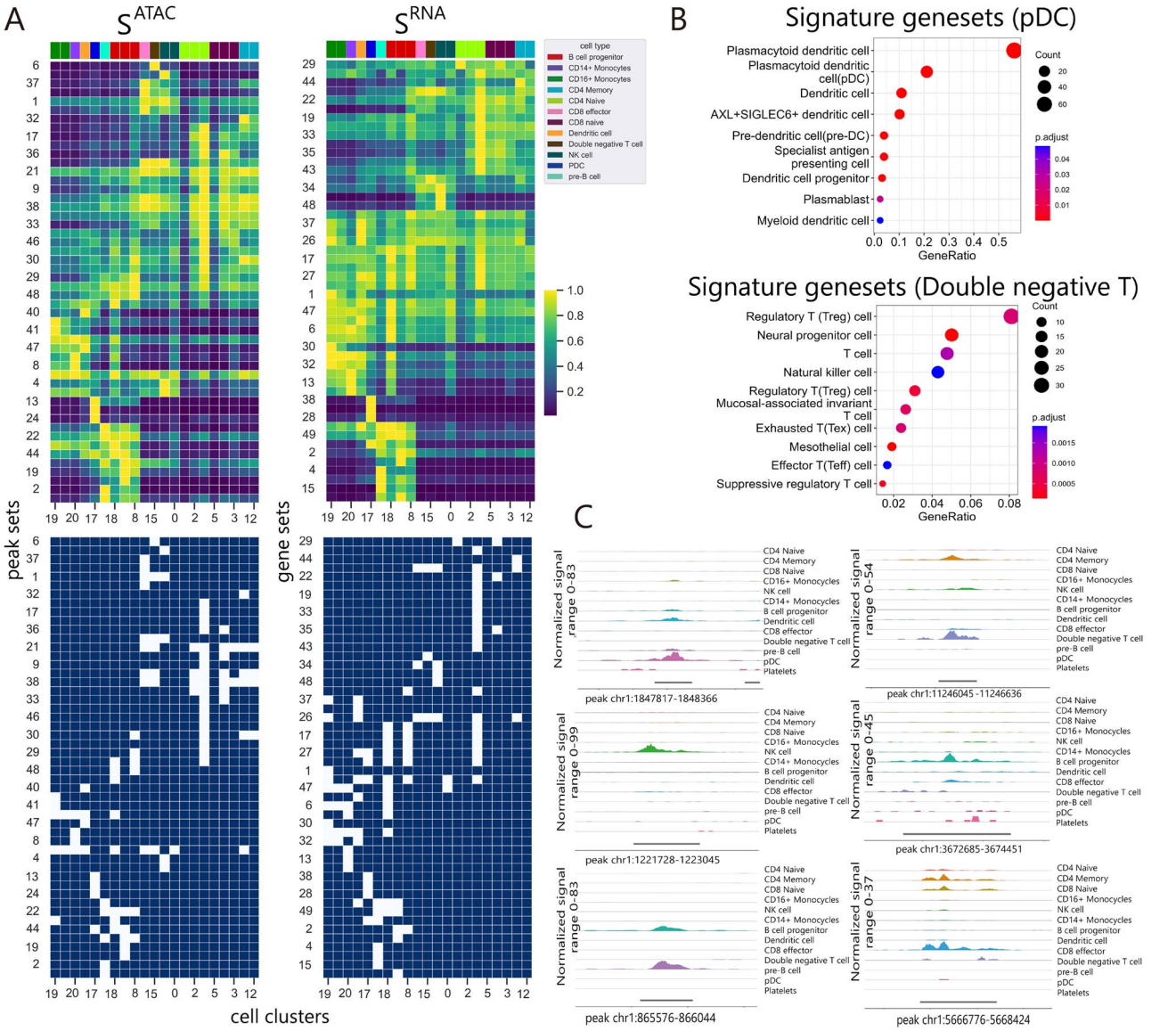

Figure 3. (**A**) Clustered heatmaps (top row) visualize normalized association matrices from each modality. Normalization was done row-wise, dividing by the maximum value. For heatmaps, each row represents a feature cluster (Y-axis) and each column represents a cell cluster (X-axis). We also show two binarized association matrices with a threshold of 0.9 at the bottom, allowing us to define signature feature sets for each cell type. Each cell is color-coded based on its ground truth from the literature, as indicated at the color legend. Please note that the heatmaps may not display the full labels on the X-axis and Y-axis. Complete heatmaps with full labels are shown in the Supplementary Fig. S1. (**B**) Cell Marker over-representation analysis is performed with clusterProfiler 4.0 [17] using CellMarker 2.0 human dataset [18]. Here we present the dotplots for signature gene sets of pDC cells (geneset-38, geneset-0, and geneset-28) and DNT cells (geneset-9). More examples are shown in Supplementary Fig. S4. (**C**) DNA accessibility visualization using the `CoveragePlot()` function in Signac [19] for the signature peak sets of pDCs (peakset-13, peakset-18, and peakset-24), DNT cells (peakset-6), NK cells (peakset-26, peakset-4, peakset-23), B cells progenitor (peakset-19, peakset-25, and peakset-45), pre-B cells (peakset-2, peakset-11, peakset-12), and CD8 effector cells (peakset-37, peakset-0, and peakset-43).

database, which provides a manually curated collection of experimentally supported markers of various cell types in different human tissues. As shown in Fig. 3B, we presented the analysis on pDC-specific gene sets (as described above) and DNT-specific gene sets, showing highly relevant terms enriched. This analysis validates that these gene sets not only differentiate cell types but are also linked to significant biological functions. Additionally, we selected one peak from the signature peak set for each cell topic and visualized DNA accessibility information using the `CoveragePlot()` function in Signac [19], which computes the averaged frequency of sequenced DNA fragments for different groups of cells within a given genomic region. They all exhibit high signal in the corresponding cell types, as seen in Fig. 3C. More examples are shown in Supplementary Fig. S3.

These results demonstrate TriTan's ability to capture both modality-unique and modality-shared information and show that TriTan's outputs, especially the middle association matrices, provide a unique advantage over existing methods for characterizing cell topics based on distinct combinations of gene sets and peak sets, enabling a deeper understanding of regulatory mechanisms underlying cell states.

TriTan identifies DORCs

In this section, we further explored gene-peak associations by utilizing association matrices from RNA and ATAC modalities. For each gene set, Pearson correlations were computed against peak sets using their respective weights in association matrices across all cell topics.

We visualized the correlation matrix with absolute Pearson correlations in Fig. 4A. Correlation matrix without the application of absolute function can be found in Supplementary Fig. S5A. We identified certain gene sets and peak sets that are strongly correlated and have the potential to be functionally linked as DORCs. To validate this hypothesis, we used the Peak-set Enrichment of Gene-Sets (PEGS) tool [20] to test the mutual enrichment of identified gene and peak sets in the genome and contrasted the output against the correlation matrix obtained from scRNA and scATAC association matrices. Here, using PEGS, input peaks are extended to ± 2kb (promoter-proximal peaks) and ± 10kb (promoter-proximal peaks&enhancer-distal peaks). The enrichment of the input gene set is calculated among the genes whose TSSs overlap with the extended peaks. We constructed the enrichment matrix that contains the *P*-values for each gene set's enrichment for each peak set as shown in Fig. 4B and Supplementary Fig. S5. The color of each cell in the heatmap represents the *P*-value of enrichment, and the x-axis and y-axis correspond to the gene sets and peak sets, respectively, in the same order as in the correlation matrix in Fig. 4A. Remarkably, the correlation matrix and the enrichment matrix exhibit a high degree of similarity, as shown in Fig. 4A and B, showing gene and peak sets with a high degree of correlation and enrichment. So we can use DORCs to represent these high correlated and significant enriched gene sets and peak sets.

In the clustered heatmap of the correlation matrix (Fig. 4A), the rectangles drawn using different colors represent these DORCs. Given our modality-specific association matrices, we observed that they correspond to the cell-type groups, such as pro-B and pre-B cells group, CD14+ mono and CD16+ mono group, among others. Thus, we selected these DORCs and examined their enrichment again using PEGS with varying peak extending distances (up to ± 150kb). As shown in Fig. 4C, the combined signature gene sets of these clusters exhibited strong mutual enrichment to the combined signature peak sets.

Based on our analysis, we have demonstrated that the highly related feature sets identified by TriTan across different omics exhibit regulatory interactions. Building on this, we will further predict cell-type-specific regulons (TFs and their target genes), as described in the next section.

TriTan predicts cell-type-specific gene regulatory network and delineates B cell trajectories at single-cell resolution

To further demonstrate that the outputs of TriTan can improve the understanding of gene regulation, we use PBMC-10K dataset to perform regulatory networks analysis for B cell populations. B lymphopoiesis, such as differentiation from pro-B cell (B-lymphoid progenitors) to pre-B cell (B cell precursors), is a genetically and epigenetically highly regulated process [24]. Therefore, it is suitable for study on the transcription factor (TF) regulatory networks.

Firstly, we used signature peak sets identified by TriTan for B cells to enrich motifs by Multiple EM for Motif Elicitation (MEME). Then, using the Tomtom tool within MEME suite [25], we compared the identified motifs with JASPAR database and retained only the ChIP-verified human TFs. The top five enriched motifs were identified and they correspond to five functional domains, respectively (see examples in Fig. 4H).

These motif and domain pairs suggest various interactions between promoters and TFs during B lymphopoiesis. SPIB, corresponding to the ETS domain and PU-box motif pair, inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression [26], while the double mutation of its orthologs in mice develops pre-B cell acute lymphoblastic leukemia [27]. MEF2C, a transcriptional activator of DNA repair, regulated the genome integrity and cell survival of pro-B, and its deficiency leads to a reduced chromatin accessible state in target regions in pre-B [28]. TEAD1/2, a member of the Hippo signaling pathway, regulated the growth and self-renewal of nonlymphoid cells, was upregulated in leukemia, and was inhibited by IKAROS, which regulates pre-B development [29]. EGR1 is also related to the early differentiation of B cells [30]. PAX5 is a master regulator in the development of B cells [31]. After losing IL7 signaling, pro-B cells lacking Pax5 can differentiate into other hematopoietic cell types, instead of entering B-cell lineage [32–34]. As Pax5 is activated, it can enter the pre-B stage, while differentiation from pre-B into plasma cells reduces the expression of Pax5 [31, 35]. Furthermore, conditional deletion of Pax5 in pre-B resulted in the retrodifferentiation of cells into pro-B [31, 36].

Next, we demonstrated that the gene sets containing these transcription factors exhibit differential expression corresponding to changes in B cell development. For example, the relative expression level of geneset-15 containing PAX5 is very low in non-B lymphocytes, but it is highly expressed in pro-B and pre-B, and the expression in pre-B is higher than that in pro-B (Fig. 4E), which is consistent with the above. EBF1 is also a key regulatory factor for the early development of B cells, which specifically binds to the promoter of Cd79a [37, 38]. Geneset-42 (Fig. 4E) contains EBF1, CD79a, and CD79b, and its relative expression levels in pre-B and pro-B are much higher than those in other cell types, which well reflects the relationship and characteristics of the three genes in the early development of B cells. IL7R, the cell surface receptor, exists in geneset-35 (Fig. 4E), and its expression is higher in pro-B cells than in pre-B cells. It is consistent with its function. Under the IL7 environment, IL7R activates the downstream JAK1/3 and STAT5a/b pathways, promotes the proliferation of pro-B, and prevents apoptosis and cell movement [39, 40]. The pre-BCR signaling reduces the adhesion of pre-B by increasing CXCR4 and reducing FAK, which indirectly leads to reduced exposure of IL7 in pre-B cells and attenuates IL7R signaling [39, 40]. Besides, the high expression of IL7R in T cells is also consistent with its previously reported role in T cell development [39, 41]. In this gene set, Bcl-2, which can be activated by IL7, and Foxo1, which acts as an enhancer to regulate the expression of IL7R, were also found [39, 42, 43]. This indicates that our method enables the enrichment of gene sets with biological function correlation.

Following the previous analysis, we pick the transcription factor PAX5, construct PAX5 regulon, and detected the difference in the PAX5 regulons between pro-B and pre-B (seen in Supplementary Tables S1 and S2) to show how gene expression is regulated by TFs. To construct the PAX5 regulons specific to pro-B and pre-B cells, we used the cell-type-specific signature gene sets identified by association matrices outputted by TriTan. For pre-B cells, geneset-4, geneset-39, and geneset-15 were used, while for pro-B cells, geneset-2 and geneset-45 were used. For each cell type, we first used the Ensembl tool [44] to locate the TSS of each gene and then use MEME suite to select those genes with PAX5 motifs enriched within ± 2kb of their TSS as the target genes for the PAX5 regulons. We performed a comparative enrichment analysis (Seen in Fig. 4F and G). The main enriched GO terms in pro-B are related to B cell proliferation, cell receptor signal transduction, and B cell activation, which may be related to IL7-mediated signal transduction [39, 41]. The GO terms of MHCII and SYNTAXIN are specifically enriched in pre-B, which may be related to the production process of pre-BCR receptor in

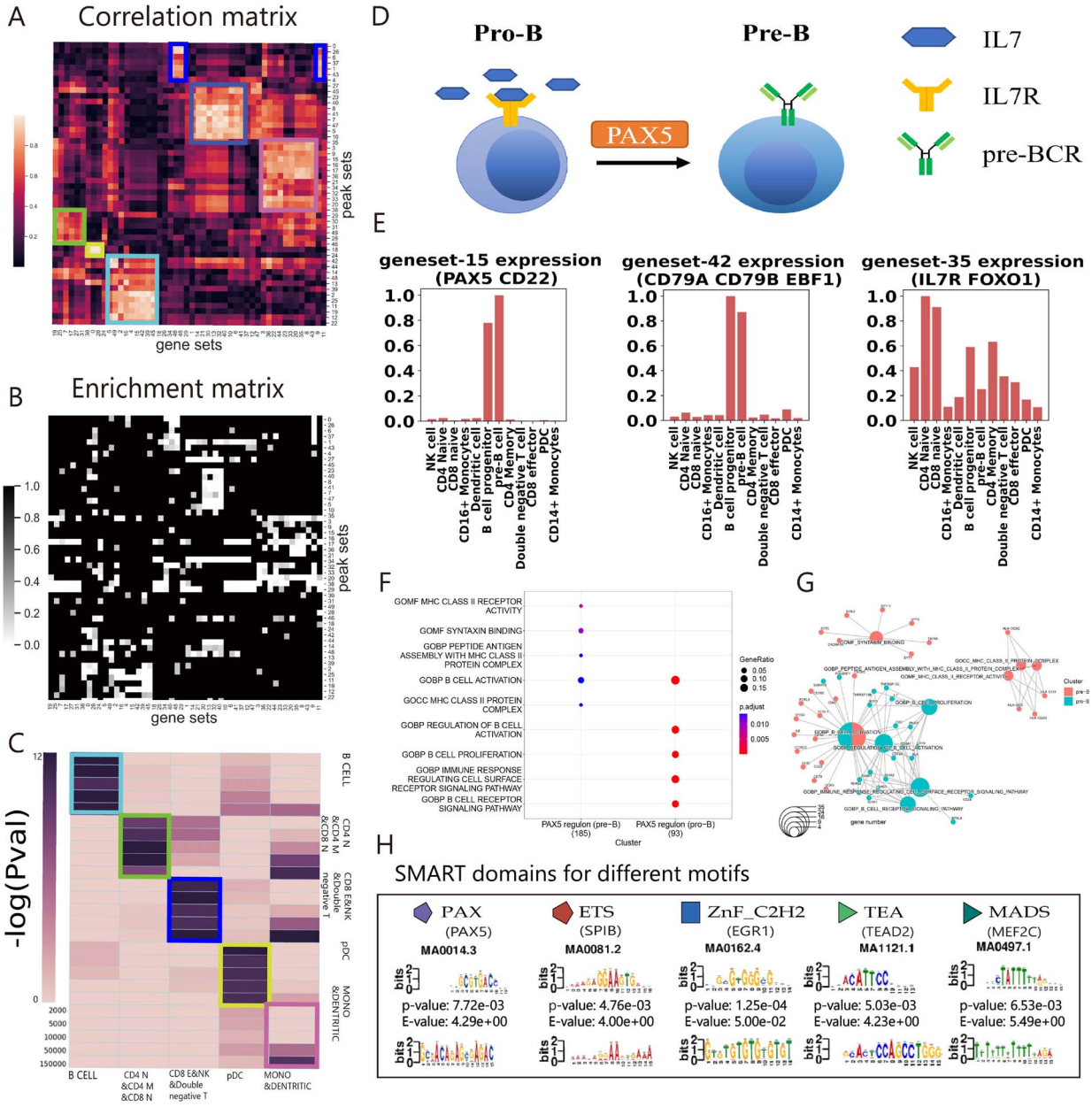

Figure 4. Downstream analysis of TriTan's output. (**A**) The clustered heatmap of the correlation matrix, obtained from RNA and ATAC association matrices, highlighting clusters of gene sets and their potentially regulating peak sets (DORCs), which can be traced back to specific cell-type groups. The x-axis shows the gene sets, and the y-axis shows peak sets, and the color represents the absolute value of their Pearson correlations. (**B**) Clustered heatmap of the enrichment matrix from PEGS analysis, x-axis shows the gene sets, and y-axis shows peak sets (expanded to ± 2kb) with the same order as of correlation matrix. (**C**) PEGS analysis for enrichment of signature gene sets (x-axis) and peak sets (y-axis, expanded to different genomic distances (2000, 5000, 10 000, 50 000, 150 000)) derived from rectangles in (A). Axis labels show their corresponding cell type groups. The cell color represents -log10(p-value) of enrichment. (**D**) Cartoon diagram illustrating the transition of B cells from the pro-B stage to the pre-B stage, which is regulated by PAX5 and the interaction between IL7 and IL7R. (**E**) The barplots showing some important gene sets' expressions reflected by their weights in association matrix across each cell type. (**F**) Comparative enrichment analysis was performed on PAX5 regulon (pre-B) and PAX5 regulon (pro-B). The regulons of PAX5 in pre-B and pro-B were analyzed by GO enrichment, and the GO database used was from Molecular Signature Database (MSigDB) C5 [21, 22]. The enrichment results were compared by the `compareCluster()` function of the clusterProfiler package, and visualized by `dotplot()` function. The size of the circle represents the gene ratio, which is obtained by the number of genes belonging to the corresponding GO term compared with the total number of genes. The color represents the *P*-value. (**G**) Gene-Concept Network Plot of PAX5 regulon (pre-B) and PAX5 regulon (pro-B). The regulons of PAX5 in pre-B and pro-B were analyzed by GO enrichment, and the GO database used was from Molecular Signature Database (MSigDB) C5 [21, 22]. The enrichment results were compared by the `compareCluster()` function of the clusterProfiler package, and visualized by the `centplot()` function. Centplot shows the relationship between genes and GO terms. Red represents genes from pre-B and corresponding enriched GO terms, and blue represents genes from pro-B and corresponding enriched GO terms. The size of the GO circle represents the number of enriched genes. (**H**) Five motifs correspond to five functional domains, respectively. Motifs were enriched by MEME based on the set of sequences identified by TriTan from B cells located ± 2kb of the TSS. The top five motifs were selected to be compared with the JASPAR database through Tomtom, retaining only those motif and TF pairs that have been validated by ChIP-seq data from human sources. The obtained TFs were put into the SMART [23] database for domain identification. Motif 1: mainly enriched to proteins with EST (E twenty-six) domains, such as SPIB, ELF1, and ETV1; motif 2: PAX (paired box) domain, such as PAX5; motif 3: TEA domain, such as TEAD1, TEAD2, and TEAD4; motif 4: MADS domain, such as MEF2A and MEF2C; motif 5: ZnF_C2H2 (Zinc finger C2H2-type) domain, such as EGR1 and KLF9. The bottom row in the figure represents the motifs enriched by our analysis, the middle row displays the best-matched motifs detected in the database by Tomtom, and the top row corresponds to the associated protein domains.

pre-B and the corresponding membrane fusion [45, 46]. Through this analysis, we showed that signature feature sets enriched by TriTan reflect changes during B cell development. Additionally, the regulons identified using TriTan appear to be functionally relevant, highlighting its potential for studying the regulatory mechanisms underlying cell types.

Conclusions

Here, we have presented TriTan, a triple nonnegative matrix factorization algorithm for integrative analysis of single-cell multiomics. TriTan is highly competitive against three state-of-the-art methods in term of the accuracy of cell clustering recovery, inferring the informative latent embedding, and computational efficiency within our benchmark study using three single-cell multiomics datasets from two different experimental technologies. It is important to note that clustering evaluation metrics (e.g. ARI scores) requires ground truth labels, which may be biased toward the methods used to derive those labels, such as WNN.

Importantly, we want to highlight the biological interpretation TriTan can provide to support downstream analysis. TriTan provides both middle association matrices and correlation matrices. Association matrices represent the weights of each feature set across different cell types, enabling users to explore how feature sets from different omics identify or differentiate a cell type. Correlation matrices, showing the association between feature sets across omics, can identify gene-peak linkages and DORCs. Finally, by scanning the motifs over the specific DORCs for each cell type, we enable the prediction of cell-type-specific regulons and the inference of gene regulatory networks. This powerful feature of TriTan will enable researchers to gain deeper insights into the complex interactions across different molecular features. While TriTan's factorization output can currently be used to generate hypotheses in bespoke analyses, a valuable future direction would be to extend the implementation with enhanced downstream analysis capabilities. This would help create a more user-friendly and integrated platform to make it widely accessible to researchers.

TriTan is available as a Python package, with extensive documentation and Jupyter notebook based tutorials at https://tritan-tutorial.readthedocs.io/en/latest/. TriTan is highly versatile and simple to apply across different paired multiomics datasets, including those with more than two modalities. A detailed tutorial demonstrating its application to a CITE-seq dataset (available at https://openproblems.bio/events/2021-09_neurips/documentation/data/dataset) is provided at https://tritan-tutorial.readthedocs.io/en/latest/notebooks/Cite-seq.html. We have also expanded its use case to a tri-modal PBMC-DOGMA dataset (RNA/ATAC/protein), with an accompanying tutorial available at https://tritan-tutorial.readthedocs.io/en/latest/notebooks/PBMC-DOGMA.html.

Key Points

- TriTan is a highly efficient and interpretable triple nonnegative matrix factorization method for multiomics data integration, showing good performance in clustering cells, identifying signature features for each cell type, and uncovering feature associations across different omics layers, which facilitates the identification of Domains Of Regulatory Chromatin (DORCs) and the prediction of cell-type-specific regulatory networks.
- TriTan uses a novel joint multiplicative update algorithm to significantly enhance its processing speed and employ an automatic model selection procedure for both cells and features clustering.
- We showed a range of downstream analyses conducted utilizing the outputs from TriTan, showing its capacity to facilitate interpretation in biological discovery.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Competing interests: No competing interest is declared.

Funding

M.I. was funded by Medical Research Council (MR/X014088/1).

Data availability

The data underlying this article are available at https://doi.org/10.48420/23283998.v1 and https://doi.org/10.48420/23289797.v1.

References

1. Mirjana E, Sarah AT. Computational methods for single-cell omics across modalities. *Nat Methods* 2020; **17**:14–7. https://doi.org/10.1038/s41592-019-0692-4.
2. Yuhan H, Stephanie H, Erica A-N. *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
3. Rohit S, Brian LH, Ashwin N. *et al.* Schema: Metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome Biol* 2021;**22**:1–24.
4. Ricard A, Damien A, Danila B. *et al.* Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:1–17.
5. Mingbo C, Zhijian L, Ivan GC. MOJITOO: a fast and universal method for integration of multimodal single-cell data. *Bioinformatics* 2022;**38**:i282–9. https://doi.org/10.1093/bioinformatics/btac220.
6. Daniel DL, Sebastian H, S. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;**401**:788–91. https://doi.org/10.1038/44565.
7. Geert-Jan H, Ina MD, Gabriel P. *et al.* Paired single-cell multi-omics data integration with mowgli. *Nat Commun* 2023;**14**:7711. https://doi.org/10.1038/s41467-023-43019-2.
8. Adam G, Zoë S, Romain L. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nat Methods* 2021;**18**:272–82.
9. Xinming T, Zhi-Jie C, Xia C. *et al.* Cross-Linked Unified Embedding for cross-modality representation learning. *Advances in Neural Information Processing Systems* 2022;**35**:15942–55. https://proceedings.neurips.cc/paper_files/paper/2022/file/662b1774ba8845fc1fa3d1fc0177ceeb-Paper-Conference.pdf.
10. Chris D, Tao L, Wei P. *et al.* Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and*

data mining (KDD '06). Association for Computing Machinery, New York, NY, USA, 2006, 126–35. https://doi.org/10.1145/1150402.1150420.
11. Daniel L, Sebastian H, S. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems 13. Proceedings of the 2000 Conference, NIPS 2000. Neural Information Processing Systems Foundation. 14th Annual Neural Information Processing Systems Conference, Denver, CO, November 27*. 2000, pp. 535–41.
12. Antonella F. A review on the selection criteria for the truncated svd in data science applications. *Journal of Computational Mathematics and Data Science* 2022;**5**:100064. https://doi.org/10.1016/j.jcmds.2022.100064.
13. Leland M, John H, Steve A. hdbscan: Hierarchical density based clustering. In: *Journal of Open Source Software, The Open Journal*, 2017; **2**. https://doi.org/10.21105/joss.00205.
14. Sai M, Bing Z, Lindsay ML. *et al*. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* 2020;**183**:1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056.
15. Malte L, Daniel B, Robrecht C. *et al*. A sandbox for prediction and integration of dna, rna, and proteins in single cells. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*. 2021.
16. Lawrence H, Phipps A. Comparing partitions. *Journal of classification* 1985;**2**:193–218. https://doi.org/10.1007/BF01908075.
17. Tianzhi W, Erqiang H, Shuangbin X. *et al*. Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The. Innovation* 2021;**2**:100141.
18. Congxue H, Tengyue L, Yingqi X. *et al*. CellMarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2022;**51**:D870–6. https://doi.org/10.1093/nar/gkac947.
19. Tim S, Avi S, Shaista M. *et al*. Single-cell chromatin state analysis with signac. *Nat Methods* 2021;**18**:1333–41.
20. Briggs P, Hunter AL, Yang S-H. *et al*. Pegs: An efficient tool for gene set enrichment within defined sets of genomic intervals. *F1000Research* 2021;**10**:570. https://doi.org/10.12688/f1000research.53926.2.
21. Arthur L, Aravind S, Reid P. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40. https://doi.org/10.1093/bioinformatics/btr260.
22. Aravind S, Pablo T, Vamsi KM. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**:15545–50. https://doi.org/10.1073/pnas.0506580102.
23. Ivica L, Supriya K, Peer B. SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Res* 2020;**49**:D458–60. https://doi.org/10.1093/nar/gkaa937.
24. Sören B, Rudolf G. The regulatory network of B-cell differentiation: A focused view of early B-cell factor 1 function. *Immunol Rev* 2014;**261**:102–15.
25. Timothy LB, Mikael B, Fabian AB. *et al*. MEME suite: Tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**:W202–8. https://doi.org/10.1093/nar/gkp335.
26. Heike S, Sean AD, Maho N. *et al*. Spi-B inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression. *Blood* 2008;**112**:1804–12. https://doi.org/10.1182/blood-2008-01-136440.
27. Kristen MS, Stephen KHL, Ian W. *et al*. Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood* 2011;**118**:2801–8. https://doi.org/10.1182/blood-2011-02-335539.
28. Wenyuan W, Tonis O, Amélie M. *et al*. MEF2C protects bone marrow B-lymphoid progenitors during stress haematopoiesis. *Nat Commun* 2016;**7**:12376.
29. Yeguang H, Zhihong Z, Mariko K. *et al*. Superenhancer reprogramming drives a B-cell– Epithelial transition and high-risk leukemia. *Genes Dev* 2016;**30**:1971–90. https://doi.org/10.1101/gad.283762.116.
30. Dinkel A, Warnatz K, Ledermann B. *et al*. The transcription factor early growth response 1 (Egr-1) advances differentiation of pre-B and immature B cells. *J Exp Med* 1998;**188**:2215–24. https://doi.org/10.1084/jem.188.12.2215.
31. César C, Alexandra S, Alessio D. *et al*. Pax5: The guardian of B cell identity and function. *Nat Immunol* 2007;**8**:463–70.
32. SL, B, Rolink AG. *et al*. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* 1999;**401**:556–62. https://doi.org/10.1038/44076.
33. Rolink AG, Nutt SL, Melchers F. *et al*. Long-term in vivo reconstitution of T-cell development by Pax5-deficient B-cell progenitors. *Nature* 1999;**401**:603–6. https://doi.org/10.1038/44164.
34. Christoph S, Ludovica B, Fritz M. *et al*. Multiple hematopoietic cell lineages develop in vivo from transplanted Pax5-deficient pre-B I-cell clones. *Blood* 2002;**99**:472–8. https://doi.org/10.1182/blood.V99.2.472.
35. Meinrad B. Transcriptional control of early B cell development. *Annu Rev Immunol* 2004;**22**:55–79. https://doi.org/10.1146/annurev.immunol.22.012703.104807.
36. Alessio D, Alexandra S, Qiong S. *et al*. Gene repression by Pax5 in B cells is essential for blood cell homeostasis and is reversed in plasma cells. *Immunity* 2006;**24**:269–81. https://doi.org/10.1016/j.immuni.2006.01.012.
37. Sören B, Rui L, Rudolf G. Defining B cell chromatin: Lessons from EBF1. *Trends Genet* 2018;**34**:257–69. https://doi.org/10.1016/j.tig.2017.12.014.
38. Rui L, Pierre C, Senthilkumar R. *et al*. Dynamic EBF1 occupancy directs sequential epigenetic and transcriptional events in B-cell programming. *Genes Dev* 2018;**32**:96–111. https://doi.org/10.1101/gad.309583.117.
39. João TB, Scott KD, Benedict S. Flip the coin: IL-7 and IL-7R in health and disease. *Nat Immunol* 2019;**20**:1584–93. https://doi.org/10.1038/s41590-019-0479-x.
40. Chris F, Sandra Z, Jeffrey JB. *et al*. Cell circuits between B cell progenitors and IL-7+ mesenchymal progenitor cells control B cell development. *J Exp Med* 2018;**215**:2586–99. https://doi.org/10.1084/jem.20180778.
41. Eva B, Anne M, Mauro T. *et al*. Do CD8 effector cells need IL-7R expression to become resting memory cells? *Blood* 2006;**108**:1949–56.
42. Yann MK, Daniel RB, Roberto T. *et al*. Foxo1 links homing and survival of naive T cells by regulating L-selectin, CCR7 and interleukin 7 receptor. *Nat Immunol* 2009;**10**:176–84. https://doi.org/10.1038/ni.1689.
43. Daniel R, Alice M, Ruben V. *et al*. STAT5 is essential for IL-7-mediated viability, growth, and proliferation of T-cell acute lymphoblastic leukemia cells. *Blood. Advances* 2018;**2**:2199–213. https://doi.org/10.1182/bloodadvances.2018021063.
44. Fiona C, James EA, Jamie A. *et al*. Ensembl 2022. *Nucleic Acids Res* 2021;**50**:D988–95.
45. Mark KB, JoséE G, Lisa AE. *et al*. The syntaxin family of vesicular transport receptors. *Cell* 1993;**74**:863–73. https://doi.org/10.1016/0092-8674(93)90466-4.
46. Julia M, Urszula E, Luca D. *et al*. MHC class II cell-autonomously regulates self-renewal and differentiation of normal and malignant B cells. *Blood* 2019;**133**:1108–18.

Briefings in Bioinformatics, 2025, **26(1)**, bbae615
https://doi.org/10.1093/bib/bbae615
Problem Solving Protocol