ADVANCED SCIENCE Open Access
www.advancedscience.com

A Cell Cycle-Aware Network for Data Integration and Label Transferring of Single-Cell RNA-Seq and ATAC-Seq

*Jiajia Liu, Jian Ma, Jianguo Wen, and Xiaobo Zhou**

In recent years, the integration of single-cell multi-omics data has provided a more comprehensive understanding of cell functions and internal regulatory mechanisms from a non-single omics perspective, but it still suffers many challenges, such as omics-variance, sparsity, cell heterogeneity, and confounding factors. As it is known, the cell cycle is regarded as a confounder when analyzing other factors in single-cell RNA-seq data, but it is not clear how it will work on the integrated single-cell multi-omics data. Here, a cell cycle-aware network (CCAN) is developed to remove cell cycle effects from the integrated single-cell multi-omics data while keeping the cell type-specific variations. This is the first computational model to study the cell-cycle effects in the integration of single-cell multi-omics data. Validations on several benchmark datasets show the outstanding performance of CCAN in a variety of downstream analyses and applications, including removing cell cycle effects and batch effects of scRNA-seq datasets from different protocols, integrating paired and unpaired scRNA-seq and scATAC-seq data, accurately transferring cell type labels from scRNA-seq to scATAC-seq data, and characterizing the differentiation process from hematopoietic stem cells to different lineages in the integration of differentiation data.

1. Introduction

In recent years, advances in single-cell RNA sequencing (scRNA-seq) technology have enabled us to generate high-throughput gene expression data through different sequencing methods at single-cell resolution.[1] The evolution of these technologies has significantly expanded the adoption of single-cell RNA sequencing across diverse fields, furnishing a more comprehensive and profound perspective on the comprehension of cell heterogeneity and functions.[1c,2] There exists a multitude of methods for generating scRNA-seq data, with over a dozen most commonly used scRNA-seq protocols accessible.[3] These technologies generate scRNA-seq data derived from distinct experiments, encompassing variations in capture timing, handlers, reagent batches, equipment, and even technological platforms.[3,4] These inherent dissimilarities engender batch effects within scRNA-seq data,[2b,5] which become the priority and grand challenges in single-cell RNA-seq data analysis.[6]

In addition, the emergence of single-cell multi-omics technologies has enabled insights into complex cellular microenvironments and biological processes, offering many exciting biological opportunities from perspectives other than transcriptomics, such as genomics, epigenomics, proteomics, metabolomics, spatial transcriptomics, etc.[7] Particularly, single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) is an epigenomic profiling technology for studying the chromatin accessibility of individual cells.[8] It provides the ability to examine the openness of chromatin regions in the nucleus at the single-cell level, which is unavailable in single-cell RNA-sequencing data.[8a] This enhances our comprehension of the epigenetic state, cell-type heterogeneity, and cell state.[9] However, scATAC-seq data has more extreme sparsity than scRNA-seq data, which also increases the difficulty of analysis based on scATAC-seq data.[8b,10] Meanwhile, cell type annotation for scATAC-seq data is challenging due to lack of specifically designed tools and use of unintuitive cis- and trans-regulatory elements in single-cell ATAC-seq data.[11] In recent years, advanced technologies have made it possible to simultaneously characterize gene expression and chromatin accessibility in the same cell, which we call the generated data paired data.[12] These techniques provide tools for the integrated analysis of scRNA-seq and scATAC-seq data, which can apply the information obtained from

J. Liu, J. Wen, X. Zhou
Center for Computational Systems Medicine
McWilliams School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX 77030, USA
E-mail: Xiaobo.Zhou@uth.tmc.edu

J. Ma
Department of Electronic Information and Computer Engineering
The Engineering & Technical College of Chengdu University of Technology
Leshan, Sichuan 614000, China

X. Zhou
McGovern Medical School
The University of Texas Health Science Center at Houston
Houston, TX 77030, USA

X. Zhou
School of Dentistry
The University of Texas Health Science Center at Houston
Houston, TX 77030, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202401815

DOI: 10.1002/advs.202401815

the large amount of annotated scRNA-seq data for the cell type annotation of scATAC-seq data.[13] However, single-omics data are more readily available than paired scRNA-seq and scATAC-seq data. That is, we can generate scRNA-seq and scATAC-seq data of different single-cell experimental samples separately, but they are from the same organ or tissue, which we refer to as unpaired data in this study.[1b] Therefore, in cases where paired data are not abundant, integrating these unpaired data is a better option for researchers to conduct broader studies. However, unpaired data frequently introduce complexity to subsequent analyses due to differences in features and sparsity level, so it is important to develop novel data integration methods that can be applied to unpaired scRNA-seq and scATAC-seq data.

In summary, the current integration of scRNA-seq and scATAC-seq data can be divided into three types: [14] 1) intra-modality integration, the integration of the same omics data (scRNA-seq data) measured from different cells and different experiments; 2) paired inter-modality integration, that is, the integration of scRNA-seq and scATAC-seq data measured from the same cell; and 3) unpaired inter-modality integration, that is, the integration of scRNA-seq and scATAC-seq data generated from different cells, samples or experiments. Corresponding integration methods have also been developed for different integration types: 1) methods for intra-modality integration employ dimensionality reduction algorithms to reduce the complexity of the data and identify the common biological signal across datasets to align cells or cell populations to integrate scRNA-seq datasets from different sources or experiments, but most of them struggle with excessive data scale, run time or resource requirements.[15] 2) methods for paired inter-modality integration apply matrix factorization,[16] weighted nearest neighbor algorithm,[17] and neural networks [18] to integrate scRNA-seq and scATAC-seq data measured within a cell and to obtain a joint profile of cellular state. These methods are specially designed for paired data, making their application to other unpaired data challenging. 3) unpaired data not only have different features but also frequently exhibit significant variations in cell count, thus giving rise to a distinct category of integration methods for unpaired data. Methods for unpaired inter-modality integration focus on finding solutions for the manifold learning and cell alignment in the embedding space using neural networks.[19] There is also a non-neural network approach that uses the non-negative matrix factorization approach and online learning algorithm to incorporate new data without recalculating from scratch.[20] However, despite the aforementioned approaches, most of them are specialized to address specific challenges within particular integration type of single-cell data. Currently, there is a lack of a comprehensive approach capable of simultaneously addressing all three types of integration issues outlined above. The single-cell data integration method to be developed needs to consider the sparsity, data scale, feature difference, high dimensionality, and other inherent disparities of scRNA-seq and scATAC-seq data. In addition, the cell cycle is often considered a confounding factor in the study of cell population and cell heterogeneity based on single-cell RNA-sequencing data.[21] How it will work in the integrated analysis of scRNA-seq and scATAC-seq data is still unknown.

To address such concerns, we developed an advanced Cell Cycle-Aware Network (CCAN) with the aim of extracting intrinsic biological signals masked by context-specific patterns (i.e., cell type-specific heterogeneity) and confounding factors (i.e., cell cycle effects, batch effects, and noise). Notably, CCAN can integrate single-cell multi-omics data and remove cell cycle effects from the integrated data while maintaining heterogeneity between cell types. CCAN is based on a domain separation network, adding a periodic activation function to the private decoder to simulate the dynamic process of the cell cycle, and projecting single-cell data from different platforms or modalities into a common low-dimensional space through shared projection. The distribution constraint function and the class alignment loss function are added to the shared embedding space to make the distribution of different data as similar as possible and the difference between different types of data to be maximized. In addition to single-cell data integration, CCAN enables cell type prediction of scATAC-seq data via transferring the cell type annotation information of scRNA-seq data to scATAC-seq data. Validations based on multiple sets of data prove that CCAN can not only eliminate the batch effect between scRNA-seq data from different platforms, but also integrate paired and unpaired scRNA-seq data and scATAC-seq data well in the embedding space. Integration of unpaired data enables accurate cell type prediction for scATAC-seq data. Furthermore, CCAN can maintain cell differentiation trajectories when integrating single-cell differentiation data.

2. Results

2.1. Overview of CCAN Approach

As illustrated in **Figure 1**, CCAN is a self-supervised approach using the labeled transcriptomic profile of scRNA-seq data (source domain) and unlabeled profile from same/different omics data (target domain), such as gene expression of scRNA-seq data and chromatin accessibility of scATAC-seq data. CCAN uses a domain separation network (DSN) to integrate data from source and target domains and transfer the annotations from source domain to target domain. Shared encoders and private encoders in DSN are three-layer perceptrons to learn noncircular and circular embeddings of both domains, separately. The shared embedding function projects a high-dimensional profile of each cell to a low-dimensional vector, which distinguishes biological meaningful signals from circular confounding factors (private embedding) and transforms the embeddings of cells from different domains into a similar distribution. In the decoder, we use sine and cosine as the activation functions specific for private embeddings, followed by a two-layer perceptron performing noncircular transformations mapping the embedded data to the original space. The training of CCAN has four main steps: 1) pretraining of the cell cycle-aware domain separation network; 2) label transferring from source domain to target domain; 3) refining CCAN by introducing a cluster alignment loss and 4) finalization and applications of CCAN model. We assessed CCAN using several real single-cell datasets, including two scRNA-seq datasets from different protocols,[15c] paired and unpaired scRNA-seq and scATAC-seq datasets,[12c,22] and single-cell differentiated datasets from different modalities,[23] etc. Evaluation results indicate that CCAN is a versatile method that can be used for multi-tasks, including single-cell multi-omics data integration, batch effect removal, cell cycle effects removal, label transferring, and cell type prediction (Figure 1). CCAN is an effective method and

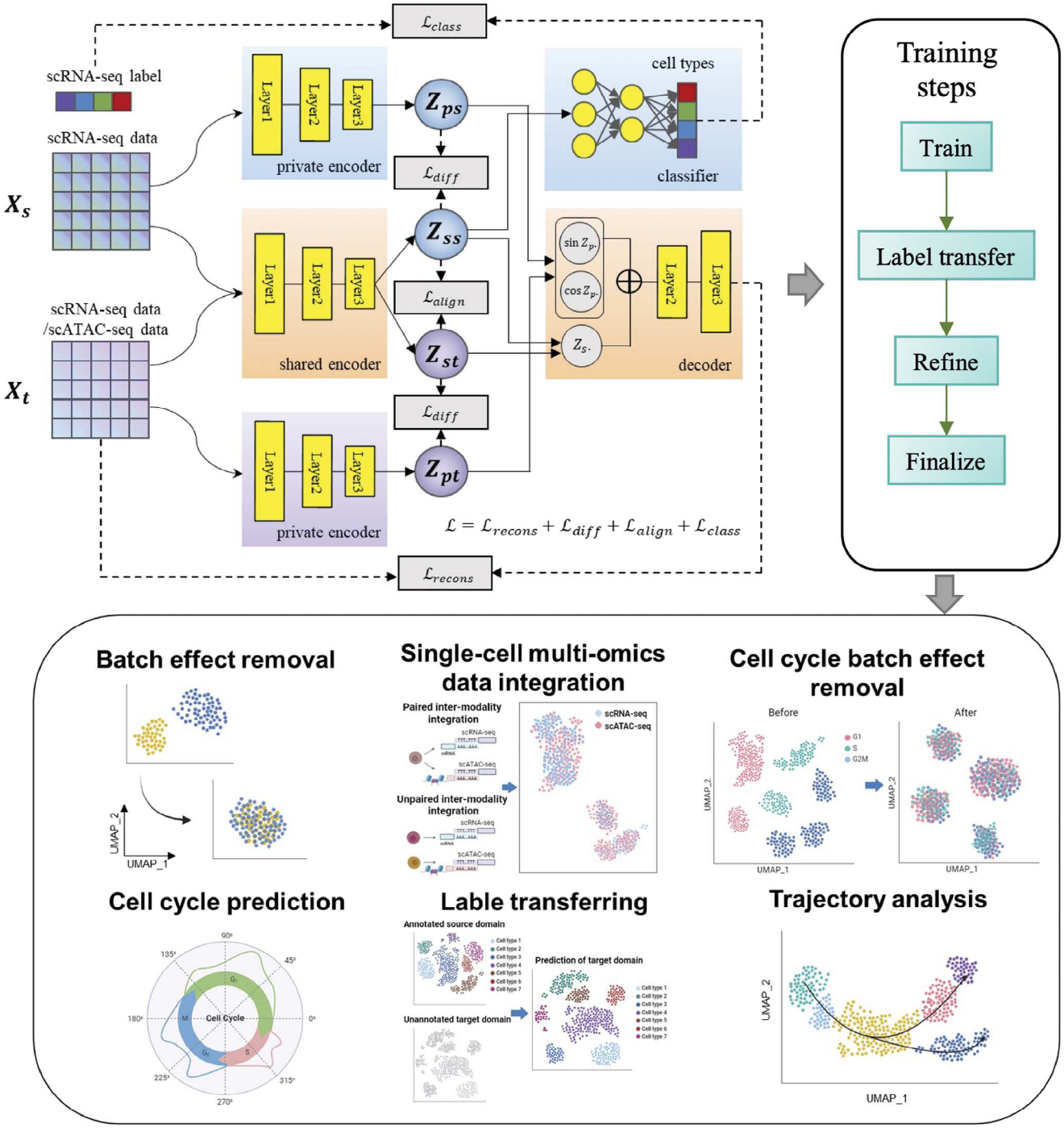


Figure 1. Overview of CCAN approach. CCAN is based on a domain separation network, taking labeled scRNA-seq data as source domain and unlabeled scRNA-seq or scATAC-seq data as target domain. Both source and target domain are fed into a shared encoder and two private encoders. The shared encoder learns the common information from both domains and employs an alignment loss to constrain the distribution of source and target domain to be similar in the shared embedding space. Private encoders are specific to source or target domain to extract periodic information of the cell cycle effect. An orthogonal difference loss between the shared embedding and private embedding enforces features learned by the shared encoder and the private encoder to be as different as possible in the low-dimensional space. Both source and target domain have the same structure of decoder by contacting the shared embedding and private embedding to reconstruct the original data. A supervised classifier is trained on the labeled scRNA-seq data and can be used to predict the label of target domain. CCAN has four main training steps and can be applied to multi-tasks, including batch effect removal, single-cell multi-omics data integration, cell cycle effect removal, cell cycle prediction, label transferring and trajectory analysis, etc.

competitive in various applications compared with other existing methods.

2.2. Batch Effect Removal of scRNA-Seq Datasets from Different Protocols

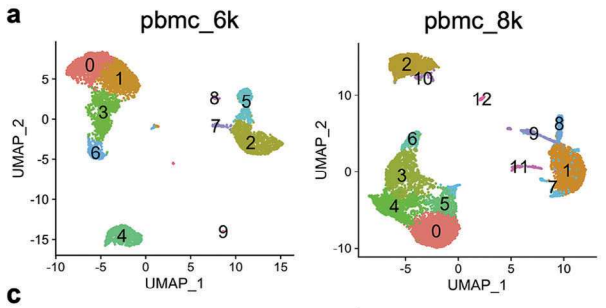
The emergence of advanced technologies enabled comprehensive transcriptional characterization of cell-type heterogeneity across a variety of biological and clinical conditions, integrating these scRNA-seq datasets from different protocols while maintaining cell-type heterogeneity has become very challenging. To benchmark the performance of CCAN against other existing methods for scRNA-seq data integration, we applied two scRNA-seq datasets of human Peripheral Blood Mononuclear Cells (PBMC), each assayed on the Chromium 10X platform but prepared with different protocols: 3′ end v1 and 3′ end v2 chemistries.[15c] We denoted them as pbmc_6k and pbmc_8k respectively in this study. We first clustered the two scRNA-seq datasets separately using the Louvain clustering algorithm,[24] resulting 10 clusters of the pbmc_6k data and 13 clusters of the pbmc_8k data (**Figure 2a**). Then we used canonical cell type marker genes to annotate PBMC clusters before integration (Figure 2b; Figures S1 and S2, Supporting Information). The annotation resulted six major cell populations (Figure 2c): monocytes (CD14+/FCGR3A+/MS4A7+/LYZ+), dendritic cells (FCER1A+), B cells (MS4A1+), T cells (CD3D+), megakaryocytes (PPBP+), and natural killer (NK) cells (CD3D-/GNLY+).[17,25] These marker genes were significantly expressed in the corresponding annotated cell types (Figure 2d). We compared CCAN with six existing methods for scRNA-seq integration, including Harmony,[15c] Seurat V4,[17] online iNMF,[20] Conos,[15a] Scanorama,[15b] and BBKNN.[15d] As illustrated in Figure 2e, the two PBMC data were completely separated before integration, which showed the confident existence of the batch effect. These batch effects significantly impeded the accurate classification of cell types. Evidently, the same cell types were subdivided into two distinct categories prior to integration, including B cells, T cells, and monocytes. CCAN outperformed other methods by accomplishing a perfect integration, effectively eliminating dataset-specific variations and mixing all cells within each cluster in the projected space. The projection distribution on the two coordinate axes of UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) [26] visualization also illustrated that the distributions of pbmc_6k and pbmc_8k exhibited near-complete overlap, further affirming the successful integration of PBMCs using CCAN. To quantitatively measure the performance of CCAN, we calculated kBET (k-nearest neighbor Batch Effect Test),[27] a metric used to assess the batch effects in single-cell RNA sequencing data by comparing the distribution of nearest neighbors between cells from different batches. Compared to other methods, CCAN had the lowest kBET score when choosing different numbers of nearest neighbors (Figure 2f), indicating that cells from different batches had more similar nearest neighbors after integration using CCAN. What's more, a perfect integration not only required no significant variability between the two scRNA-seq data, but also needed to enhance the differences between cell types, which was valuable for subsequent downstream analyses. We applied k-means clustering to the integrated data, and compared the clustering results with the annotated cell types. By calculating the three clustering evaluation metrics of Rand Index (RI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI), we observed that the results of CCAN were slightly inferior to those of BBKNN and Seurat V4, but always higher than Harmony, online iNMF, Conos and Scanorama (Figure 2g). However, considering the performance of both data mixing and cell type separation, CCAN was still a better choice for integration that had less impact by batch effects and was more reliable for downstream analyses. Details in the calculation of three clustering evaluation metrics are presented in Text S1 (Supporting Information).

2.3. Cell Cycle Identification and Cell Cycle Effect Removal of the Integrated Data

To evaluate the CCAN's performance in cell cycle identification, we utilized a mouse embryonic dataset [28] with known ground truth cell cycle information as one batch. Gaussian noise was added to simulate another batch.[4c,29] Subsequently, CCAN was applied to integrate two batches and estimate cell cycle pseudotime from cell cycle-specific variations. We leveraged a Gaussian Mixture Model (GMM) with three components (Figure S3, Supporting Information) to discretize the continuous pseudotime generated by CCAN into discrete cell cycle stages. In our comparison, CCAN was benchmarked against two marker gene-based methods: Seurat [4a] and reCAT [30] and one deep learning-based method: Cyclum.[21c] Seurat employs 43 S phase marker genes and 54 G2M phase marker genes to delineate cell cycle stages,[4a] while reCAT utilizes 378 cell cycle genes from Cyclebase3 [31] to estimate cell cycle stages.[30] To evaluate the performance of these models, we employed four classification metrics: Accuracy, Rand Index (RI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). As illustrated in **Figure 3a**, CCAN demonstrated outstanding performance in identifying cell cycle stages compared with marker gene-based methods.

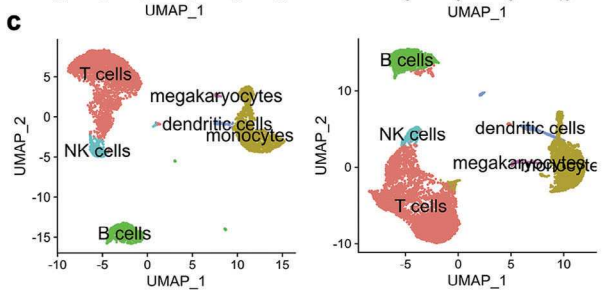
Cell cycle dysregulation presents as changes in cell distribution across different stages of the cell cycle and variations in the expression of cell cycle regulatory genes.[32] To further assess the CCAN's performance in identifying cell cycles from different conditions. We used scRNA-seq tumor data of 176644 cells from breast cancer patients.[33] CCAN was utilized to integrat data with two conditions (letrozole alone and intermittent high-dose ribociclib) and predict cancer cell cycle transitions. We calculated the proportion of cancer cells in the mitotic (S/G2) phase in biopsies from each patient (Figure S4a, Supporting Information). During the combination therapy, there was an increase in the proportion of cancer cells in the mitotic and growth (S/G2) phases. We studied the expression fluctuations of CDK6 and CDKN2A genes throughout the cell cycle under combination therapy, a reduction in CDK inhibitor 2A and an increase in CDK6 expression from G1 to S/G2 phase were observed (Figure S4b, Supporting Information). In conclusion, CCAN accurately estimates cancer cell cycle transitions.

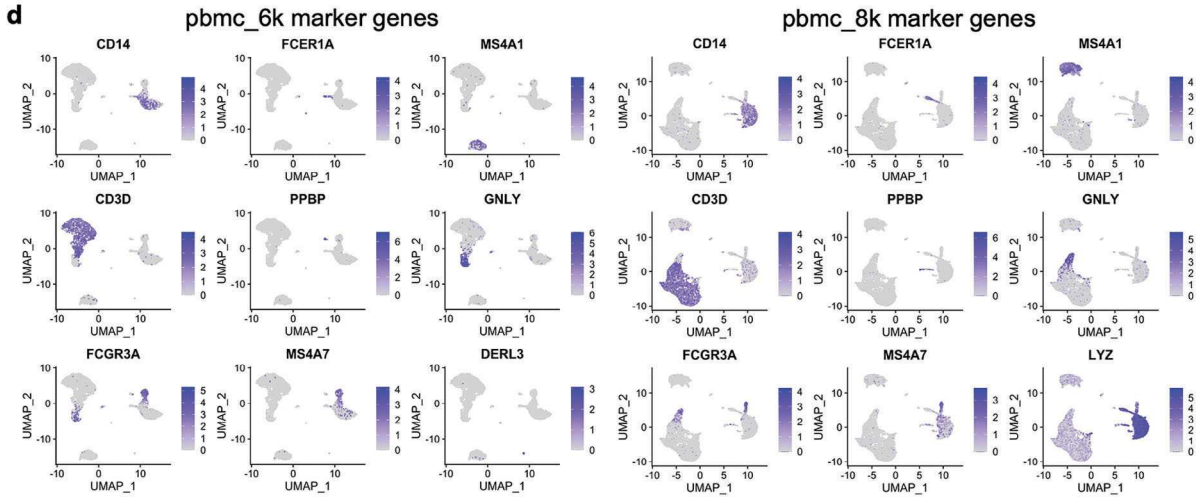
In addition to the impact of data batches on cell type heterogeneity, cell cycle is often seen as a confounding factor when studying differences between cell types. Due to the lack of real data with both cell cycle and cell type labels as ground truth,

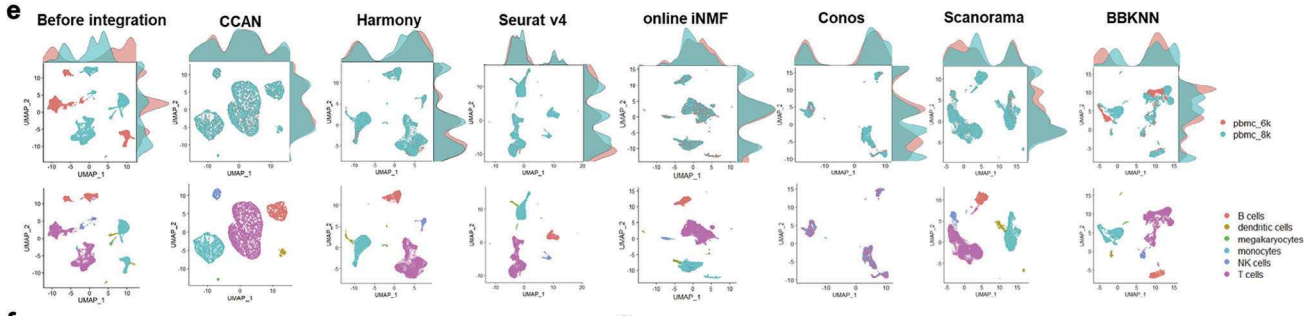


b

Cell types	Markers	Pbmc_6k clusters	Pbmc_8k clusters
monocytes	CD14+ FCGR3A+ MS4A7+ LYZ+	2,5	1,7,8
dendritic cells	FCER1A+	7	9,12
B cells	MS4A1+	4,9	2
T cells	CD3D+	0,1,3	0,3,4,5,10
megakaryocytes	PPBP+	8	11
natural killer (NK) cells	CD3D- GNLY+	6	6







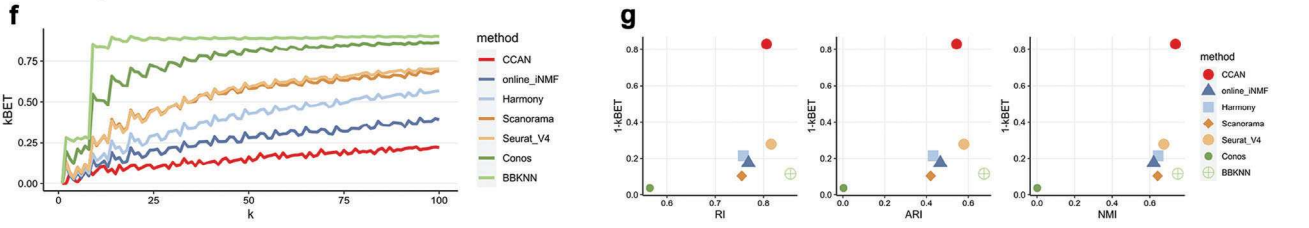


Figure 2. Annotation of PBMCs and batch effect removal of scRNA-seq datasets from different protocols. a) UMAP visualization of two single-cell PBMC data from different protocols. Different colors represent different clusters using Louvain clustering method. b) Annotation of different clusters of two PBMC data. Different cell clusters were annotated into six major cell types using marker genes. c) UMAP visualization of two single-cell PBMC data

we use the simulated data in Cyclum [21c] to characterize the cell cycle effect in virtual samples. The simulated dataset comprises two subclones. The first subclone, designated as subclone 1, consists of samples from the mESC data,[28] which includes experimental cell cycle labels serving as ground truth. The second subclone is generated by doubling the expression levels of a randomly selected set of genes, encompassing varying numbers of known cell-cycle and non-cell-cycle genes. Cells from these two subclones are amalgamated into a virtual tumor sample. To simulate batch effects in our data, we randomly partitioned all merged samples into two groups and introduced Gaussian noise to one of them.[4c,29] This process was repeated 10 times, generating 10 sets of simulated data. We then assessed the effectiveness of CCAN alongside Cyclum, Seurat, and ccRemover in removing the cell cycle effect. Figure 3b presents UMAP visualizations of a representative sampled dataset. CCAN successfully segregates two subclones accurately, with a noticeable distinction between phenotypes compared to the data before removing the cell cycle effect. Cyclum separates cells into two groups, but they do not correspond to the expected subclones. Seurat and ccRemover fail to distinguish between the two subclones. Figures S5–S9 (Supporting Information) illustrate the UMAP visualizations of 10 simulated datasets before and after removing the cell cycle effect using four methods. To quantify the effectiveness of cell cycle effect removal, we employ the metric of separability to measure the degree to which the data points in different classes can be correctly separated or discriminated by the provided subclone clusters. A higher separability value indicates better discrimination between subclones. Figure 3c highlights CCAN's superior separability, underscoring its prowess in removing the cell cycle effect from integrated data.

For the PBMC data before integration, we meticulously selected a subset of variable genes from the PBMCs and proceeded to conduct Principal Component Analysis (PCA) based on these selected genes. Notably, some genes associated with the cell cycle prominently featured among the top 10 principal components. For instance, in the pbmc_6k dataset, genes such as *TYMS*, *RRM2*, *BIRC5*, *PCNA*, and *HMGB2* displayed significant associations with principal components 7, 8, 9, and 10 (PC_7, PC_8, PC_9, and PC_10). Similarly, within the pbmc_8k dataset, genes *TYMS*, *BIRC5*, and *MKI67* were prominent in principal component 10 (PC_10) (Figure 3d). Despite their minimal or low expression in the majority of cells, several cell cycle-related genes (such as *PCNA* and *HMGB2* in pbmc_6k and *BIRC5* in pbmc_8k) exhibited normal expression patterns, thereby corroborating the presence of cell cycle effects. Inspired by Cyclum,[21c] we developed a cell-cycle aware module in CCAN to remove cell cycle effects in the integration of scRNA-seq data. CCAN employed a distinct sinusoidal component in the private autoencoder to effectively capture the circular trajectory in the high-dimensional gene expression space. The private embedding space was formed by single cells sampled at various stages of a periodic process. In essence, the private encoder in CCAN was dedicated to pinpointing an optimal cell embedding within this circular space, which we denoted as cell cycle pseudotime. We compared the performance of CCAN with other existing cell cycle effect removal methods, including Seurat, Cyclum, and ccRemover. Before integration, we used the cell cycle marker genes in Seurat [17] to identify the cell cycle phases of cells in the two scRNA-seq datasets, and used the identified cell cycle as a benchmark label to evaluate the effectiveness of cell cycle effects removal. UMAP visualization of the integrated data after cell cycle effects removal showed that integration using CCAN was unable to clearly distinguish the three annotated cell cycle phases, as did Seurat, Cyclum, and ccRemover. However, the advantage of CCAN over other methods was that only CCAN did not introduce extra noise to the downstream analysis based on cell types after removing the cell cycle effects, which was reflected in the fact that CCAN can accurately distinguish six different cell types after removing the cell cycle effects (Figure 3e). We calculated the metric of separability to quantitatively measure the discrimination between cell types after removing the cell cycle effect using CCAN, Seurat, Cyclum, and ccRemover (Figure S10, Supporting Information). This also demonstrated the effectiveness and reliability of CCAN in removing cell cycle effects.

2.4. Integration of Joint Profiling of scRNA-Seq Data and scATAC-Seq Data

Advanced technologies make it possible to simultaneously measure gene expression and chromatin accessibility in the same cell, such as SNARE-seq.[12c] We applied CCAN to the paired single-cell dataset, including joint profiling of scRNA-seq data and scATAC-seq data from adult mouse brain using SNARE-seq technology.[12c] We compared the performance of CCAN with eight existing integration methods, containing scMVP,[18a] scAI,[16] scMVAE,[18b] Seurat v4,[17] scJoint,[19b] GLUE,[19a] SCALEX,[19c] and online_iNMF.[20] Among these methods, scMVP, scAI, and scMVAE are specifically designed only for integrating paired data, so their generation of the integrated data is based on concatenation of features in the latent dimension (Figure S11, Supporting Information). While Seurat v4, scJoint, GlUE, SCALEX, and online_iNMF have the capability to integrate both paired and unpaired datasets. Their integration strategy for paired datasets in the comparison is to treat different modalities as two datasets from different experiments and integrate them together, so their integration is based on concatenation of cells (Figure S11, Supporting Information). The paired datasets are measured in the same cells, which provides ground truth labels that allow CCAN to integrate without predicting the pseudo label of the scATAC-seq data first (see Experimental

from different protocols with annotated cell types. d) UMAP visualization of marker expression in pbmc_6k and pbmc_8k data, respectively. e) UMAP visualization of scRNA-seq datasets before integration and after integration using CCAN, Harmony, Seurat V4, online iNMF, Conos, Scanorama and BBKNN, respectively. The upper panel was colored by different data batches and the bottom panel colored by annotated cell types. f) Line plot of kBET scores of integrated data using different methods when choosing different numbers of nearest neighbors. k is neighborhood size. g) scatter plot of kBET score and three clustering metrics including RI, ARI, and NMI of different methods. Different methods were represented using different colors and different shapes. RI, Rand Index; ARI, Adjusted Rand Index; NMI, Normalized Mutual Information.

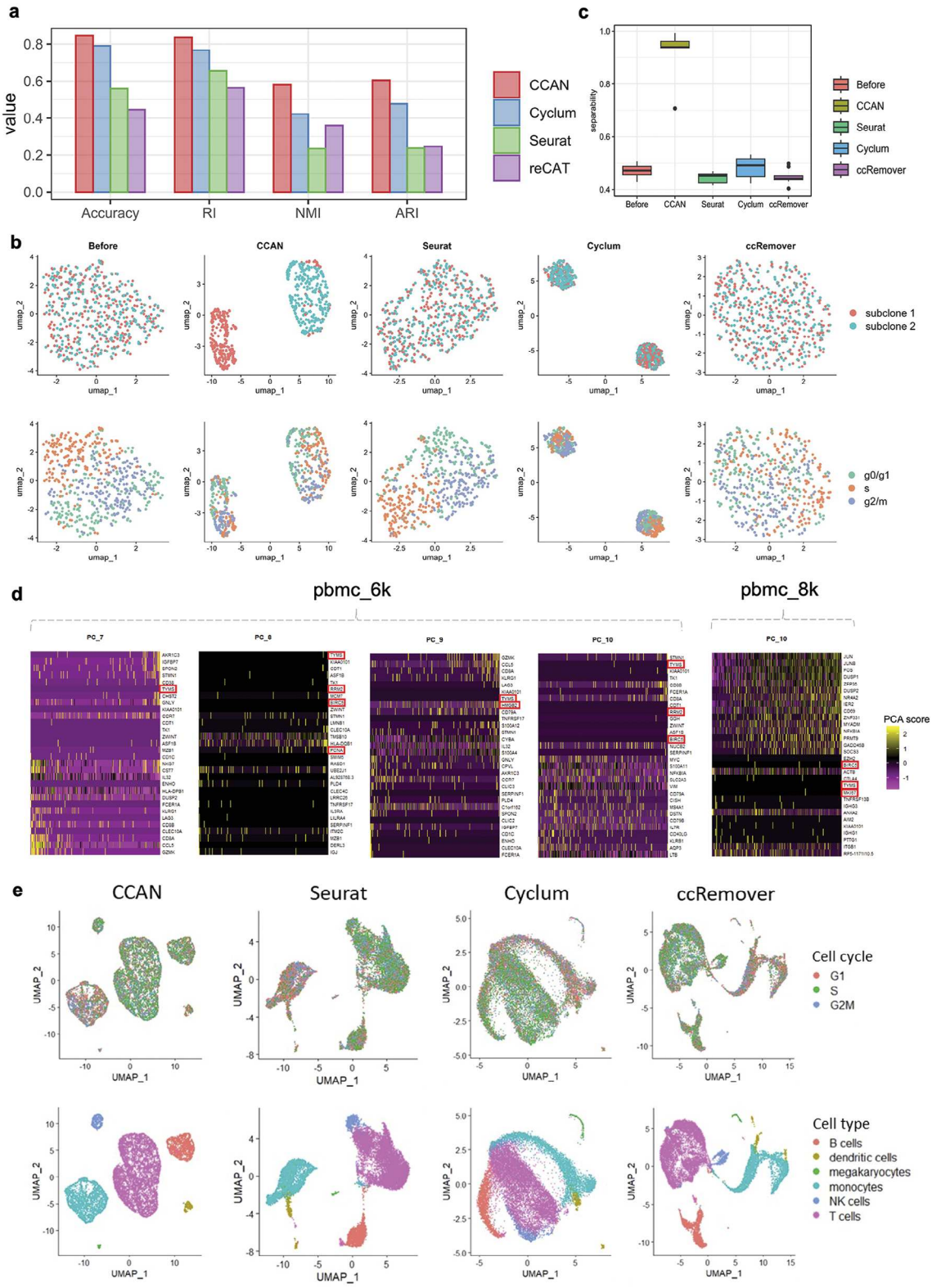


Figure 3. Cell cycle identification and cell cycle effect removal of virtual mESC data and single-cell PBMC data. a) Bar plot of four clustering metrics including Accuracy RI, ARI and NMI of different methods in evaluating the performance of cell cycle identification. Different methods were represented using different colors. RI, Rand Index; ARI, Adjusted Rand Index; NMI, Normalized Mutual Information. b) UMAP visualization of integrated mESC data before and after removing cell cycle effects using CCAN, Seurat, Cyclum and ccRemover. Top panels are colored by subclones and bottom panels are colored by cell cycle groundtruth. c) The separability of mESC data before and after removing cell cycle effect using CCAN, Seurat, Cyclum and

Section). In addition, CCAN's integration for paired data is also the concatenation of learned features of scRNA-seq and scATAC-seq data in the latent space. UMAP visualization of the integrated data shows the excellent integration ability of CCAN, which enables the integrated data to distinguish 13 different cell types better than other methods. Following CCAN, GLUE exhibits the second-best performance. Notably, CCAN's integrated data showcases a tighter aggregation of cells of the same type, with cells of different types are more dispersed. Although scAI, scMVAE, Seurat V4, and SCALEX can also clearly distinguish the large clusters such as L2 /3T, L4, L5CT, and L6IT, different cell types are not far away from each other (**Figure** 4a). The results from three clustering evaluation metrics, including Rand Index (RI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI), computed based on k-means clusters and ground truth, further corroborate that CCAN outperforms other methods in both integrating cells and accurately separating different cell types (Figure 4b).

2.5. Integration of Unpaired Datasets and Label Transferring from scRNA-Seq Data to scATAC-Seq Data

Although more and more technologies are able to measure gene expression and chromatin accessibility in the same cell, there are still unpaired datasets from the same tissue. There is no correspondence between cells in scRNA-seq data and cells in scATAC-seq data. The integration of unpaired datasets maps both modalities to a common space, which allows tools and analyses designed for scRNA-seq data to have the potential to be applied to scATAC-seq data. Variations between different modalities are different from batch effects between scRNA-seq datasets from different protocols, since scATAC-seq data characterize the chromatin accessibility instead of transcriptome profile and have more sparsity than scRNA-seq data. Even though the integration of unpaired datasets from different modalities has many challenges, CCAN still has excellent performance compared with other existing methods (**Figure** 5a). We assessed the performance of CCAN using scRNA-seq data from CITE-seq and scATAC-seq data from ASAP-seq. The CITE-seq generates filtered scRNA-seq data based on the condition of mitochondrial reads greater than 10%, a number of expressed genes fewer than 500, and total number of UMI fewer than 1000. For the scATAC-seq data from ASAP-seq, we filtered out cells with a number of peaks more than 1 00 000 and calculated the gene activity matrices for scATAC-seq data using Signac.[34] After that, 17668 overlapped genes are selected as the input features of CCAN. In comparison to the state before integration, CCAN effectively integrates data from two distinct modalities, leading to the clear identification of seven cell types in the integrated dataset. CCAN is comparable to Seurat v4, GLUE, SCALEX, and online_iNMF (Figure 5a). Although the integration performance of CCAN is slightly inferior to scJoint, its label transferring exhibits higher accuracy in identifying cell types in scATAC-seq data when compared to scJoint and Seurat v4. This is evident in the superior accuracy, F1 score, precision, and recall values of CCAN (Figure 5b; Text S1, Supporting Information). In terms of predicted cell types for scATAC-seq data, CCAN's results closely align with the golden standard, indicating a high level of accuracy (Figure 5c). This suggests that CCAN achieves a well-balanced integration of data and label transferring, allowing it to accurately predict cell types for scATAC-seq data without compromising the performance of data integration for both modalities.

2.6. Trajectory Inference and Pseudotime Analysis on the Integrated scRNA-Seq Data and scATAC-Seq Data

We used the human hematopoiesis dataset [23] to evaluate the performance of the integration of differentiated datasets. The human hematopoiesis dataset profiles the chromatin accessibility and gene-expression data of single cells that undergo a differentiation path from hematopoietic stem cells (HSC) dividing into branches. One branch differentiates into plasmacytoid dendritic cells (pDC), the other goes through common myeloid progenitor (CMP) and differentiates into megakaryocyte erythroid progenitor (MEP) and granulocyte–monocyte progenitors (GMP). Granulocyte–monocyte progenitors (GMP) can further differentiate into Monocyte (mono) cells (**Figure** 6a). The integration of CCAN effectively merged the majority of cells from both scRNA-seq and scATAC-seq data within the embedding space. However, a small subset of cells from the scATAC-seq data remained distinct. This observation shows CCAN's capability to mitigate modality-specific variations (Figure 6b). Figure 6c shows the visualization of inferred trajectories of the hematopoiesis stem cell differentiation process based on the joint embedding of scRNA-seq and scATAC-seq data. Cells are colored with ground-truth cell types. Lineage 1 and Lineage 2 were inferred and smoothed using Slingshot.[35] In the inferred Lineage 1 of the hematopoiesis dataset, a clear trajectory emerges, with hematopoietic stem cells (HSC) transitioning through the common myeloid progenitor (CMP) stage and further differentiating into megakaryocyte erythroid progenitor (MEP) cells. Along Lineage 2, HSCs follow a similar path through the CMP stage, eventually leading to a mixed cluster comprising granulocyte–monocyte progenitors (GMP) and Monocyte (mono) cells. Notably, toward the end of Lineage 2, we observe a trend wherein GMP cells transition toward plasmacytoid dendritic cell (pDC) differentiation in the integrated data. The literature [36] provides evidence suggesting that GMP cells undergo a series of differentiation steps, including differentiation into earlier progenitor cells, followed by further differentiation into various types of immune cells such as dendritic cells. This finding potentially elucidates the observed trajectory from GMP cells to pDC cells. Based on the cell type annotations on the integrated HSC dataset, we can distinctly observe a lineage progression from hematopoietic stem cells (HSC) to plasmacytoid dendritic cells (pDC), a trajectory that is not readily inferred by Slingshot. For a more intuitive representation, we manually added this lineage (referred to as Lineage 3) to highlight this

ccRemover. d) Heatmap of variable genes in principle components of PBMCs in two scRNA-seq data. Only the principal components related to cell cycle genes among the top ten principal components are shown. Cell-cycle genes are marked in red boxes. e) UMAP visualization of integrated PBMC data after removal of cell cycle effects using CCAN, Seurat, Cyclum, and ccRemover. Top panels are colored by annotated cell cycle and bottom panels are colored by annotated cell types.

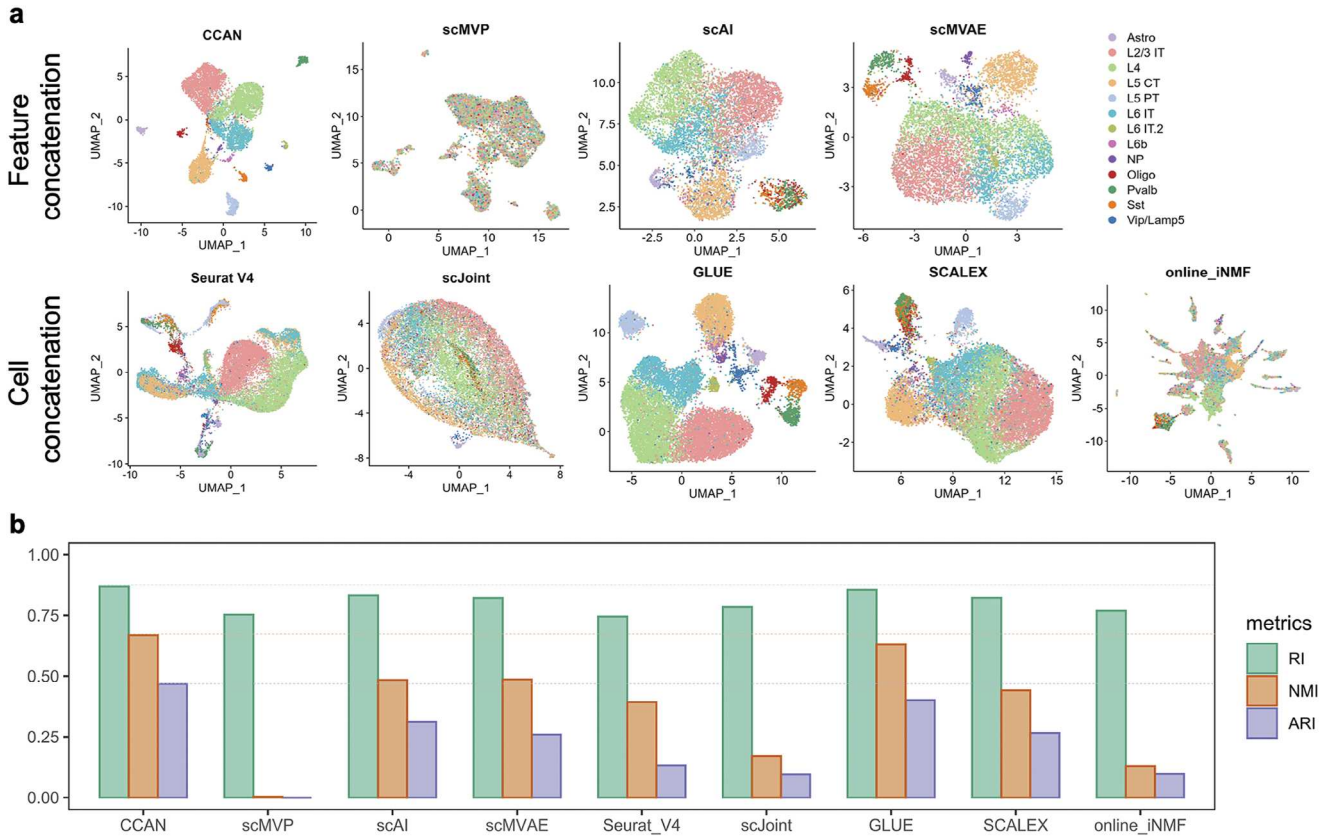


Figure 4. Data integration of paired datasets of scRNA-seq data and scATAC-seq data from adult mouse brain using SNARE-seq technology. a) UMAP visualization of integrated SNARE-seq data using CCAN, scMVP, scAI, scMVAE, Seurat v4, scJoint, GLUE, SCALEX, and online_iNMF, respectively. Different colors represent different cell types annotated by Seurat. b) Bar plot of three clustering metrics including RI, ARI, and NMI of different methods. Different metrics were represented using different colors. RI, Rand Index; ARI, Adjusted Rand Index; NMI, Normalized Mutual Information.

developmental pathway. Pseudotime analysis [37] based on the inferred trajectories confirms the ability of our model to integrate differentiated scRNA-seq data and scATAC-seq data (Figure 6d).

Differential gene analysis can identify differences in gene expression between different cell types during the differentiation process. By calculating the Pearson correlation between gene expression and the inferred cell differentiation pseudotime, we found the nine genes most related to pseudotime, are *KLF1, IRF8, SPIB, IRF7, IRF4, ZNF683, STAT2, PRDM1*, and *BLC11A* (Figure 6e). These genes are called transition genes.[38] They are highly expressed in leaf node cells on the cell differentiation trajectory (Figure S12, Supporting Information). According to some published literature studies, the expression of KLF1 is limited to erythrocytes and megakaryocytic-erythroid progenitor cells MEP;[39] *IRF8, IRF7, IRF4, STAT2, PRDM1*, and *BLC11A* are all marker genes for plasmacytoid dendritic cells and are highly expressed in pDCs.[40] In addition, we also used DESeq2 [41] to identify differentially expressed genes between different cell types. As shown in Figure S13 (Supporting Information), *HOXA6, PRDM16, IRF8, GATA1, CEBPD*, and *CEBPE* are differentially expressed genes in hematopoietic stem cells, common myeloid progenitor cells, common myeloid progenitors, plasmacytoid dendritic cells, megakaryocyte-erythroid progenitors, monocytes, and granulocyte-monocyte progenitors, respectively. They are marker genes of different cell types.[40e,42] Enrichment analysis based on differential genes across all cell types showed that these differential genes were concentrated in pathways related to hematopoietic stem cell differentiation (Figure 6h). Differential gene analysis further confirmed the effectiveness and accuracy of CCAN in integrating differentiation data with multi-modalities.

3. Discussion

Data integration of single-cell multi-omics has enhanced our investigation of cell functions and internal regulatory mechanisms beyond single omics viewpoints. However, single-cell multi-omics integration has numerous challenges, including issues such as omics-variance, sparsity, cell heterogeneity, and confounding factors. The cell cycle confounders in scRNA-seq data inspired us to think about the cell cycle effects in the integration of multi-omics data of cells, especially the integration of scRNA-seq and scATAC-seq data. In this study, we developed CCAN, a cell cycle-aware network for data integration of scRNA-seq and scATAC-seq data and label transferring from scRNA-seq to scATAC-seq. CCAN is based on a domain separation network, which includes periodic activation functions (sine and cosine) in the private autoencoder to simulate and remove cell cycle effects from the integrated single-cell multi-omics data, and projects single-cell data from different platforms or different omics into

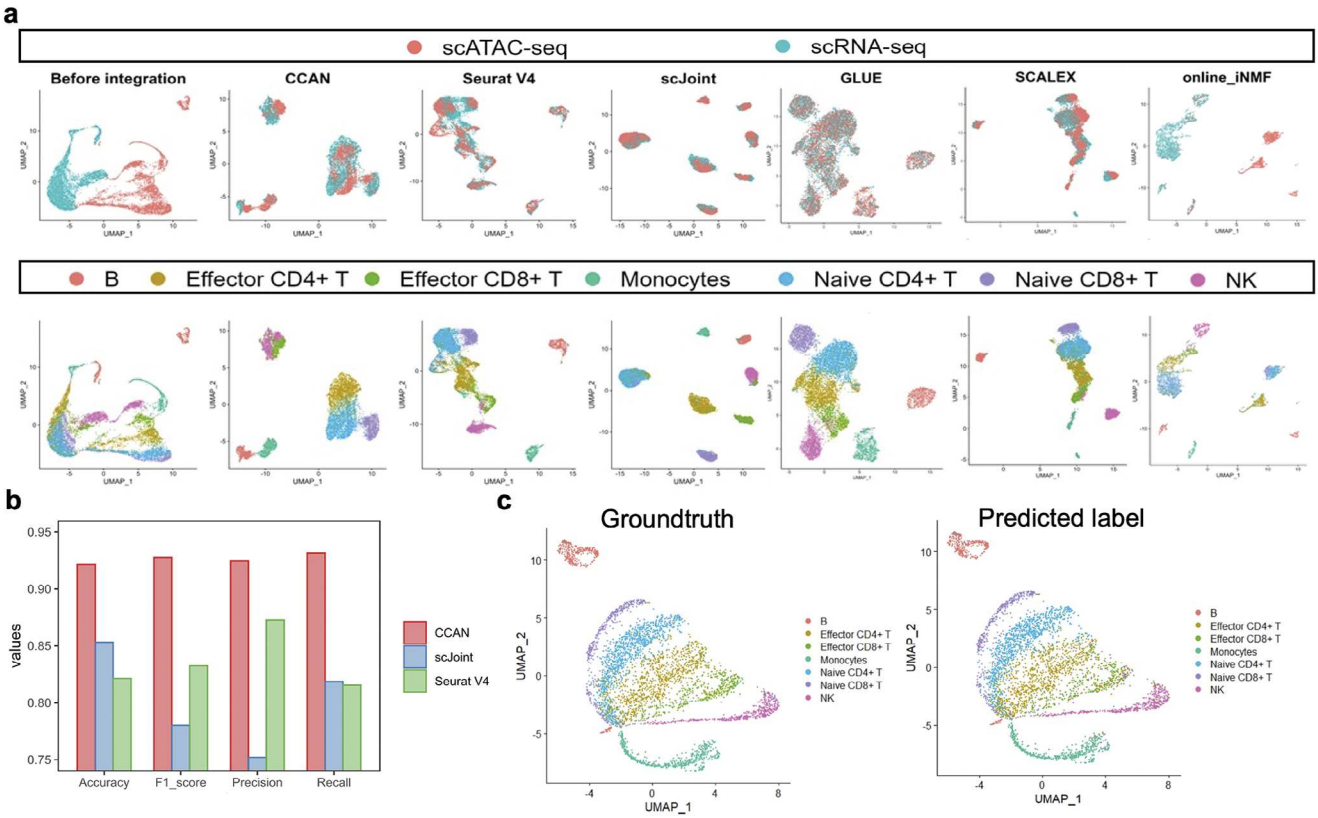


Figure 5. Integration of unpaired datasets and label transferring from scRNA-seq data to scATAC-seq data. a) UMAP visualization of unpaired scRNA-seq data and scATAC-seq data before and after integration using CCAN, Seurat V4, scJoint, GLUE, SCALEX, and online_iNMF, respectively. Different colors in the upper panel represent different modalities and different colors in the lower panel represent different cell types. b) Label prediction comparison of CCAN with scJoint and Seurat V4 using four clustering metrics including Accuracy, F1 score, Precision and Recall. In multi-class prediction evaluation, Precision, Recall and F1 score represent Macro-Precision, Macro-Recall and Macro-F1 score, separately. Different methods are represented using different colors. c) UMAP visualization of scATAC-seq data labeled with ground-truth cell types and CCAN-predicted cell types, respectively. Different colors represent different cell types.

a common low-dimensional space through shared autoencoder to integrate while maintaining heterogeneity between cell types. The domain adaptive network solves the problem of inconsistent distribution between single-cell data from different platforms or different omics. The class alignment loss is added to the hidden layer of the domain adaptive network to enhance the differences between different cell types in the integrated data. At the same time, by introducing sine and cosine activation functions into the network, the impact of the cell cycle on cell type heterogeneity can be eliminated while effectively integrating single-cell multi-omics data, further improving the performance of data integration.

The design of the cell cycle-aware module within CCAN was inspired by Cyclum. The cell cycle-aware module within CCAN shares similarities with Cyclum's circular component. Both modules leverage nonlinear periodic functions to emulate the circular trajectory of the cell cycle. However, a key distinction lies in CCAN's integration of an orthogonal loss, absent in Cyclum. This additional loss function facilitates the separation of features between the private and shared embeddings, a crucial factor contributing to CCAN's superior performance in cell cycle effect removal. Furthermore, CCAN integrates the cell cycle module into the private decoder of a domain separation network, enhancing its adaptability to multi-omics data integration tasks. Conversely, Cyclum embeds the circular component within an autoencoder framework, thereby limiting its use in the integration of single-cell multi-omics data.

Through comprehensive downstream analyses across diverse data integration scenarios, it has been demonstrated that CCAN (Cell Cycle-Aware Network) possesses the capability to not only mitigate batch effects and cell cycle effects in single-cell RNA sequencing (scRNA-seq) data originating from different platforms but also seamlessly integrate both paired and unpaired scRNA-seq data and single-cell ATAC-seq (scATAC-seq) data. The integration of unpaired scRNA-seq data and scATAC-seq data is particularly noteworthy, as CCAN facilitates accurate cell type prediction for scATAC-seq data by leveraging the transformation of annotation information gleaned from scRNA-seq data. This unique feature enhances the utility of CCAN in deciphering cellular heterogeneity and functional states across diverse data modalities. Furthermore, CCAN exhibits remarkable versatility by not only integrating differentiated data from various modalities but also capturing the intricacies of cell differentiation trajectories. This is exemplified in its ability to characterize the differentiation process from hematopoietic stem cells (HSC) to different branches, providing a holistic understanding of cell fate determination in

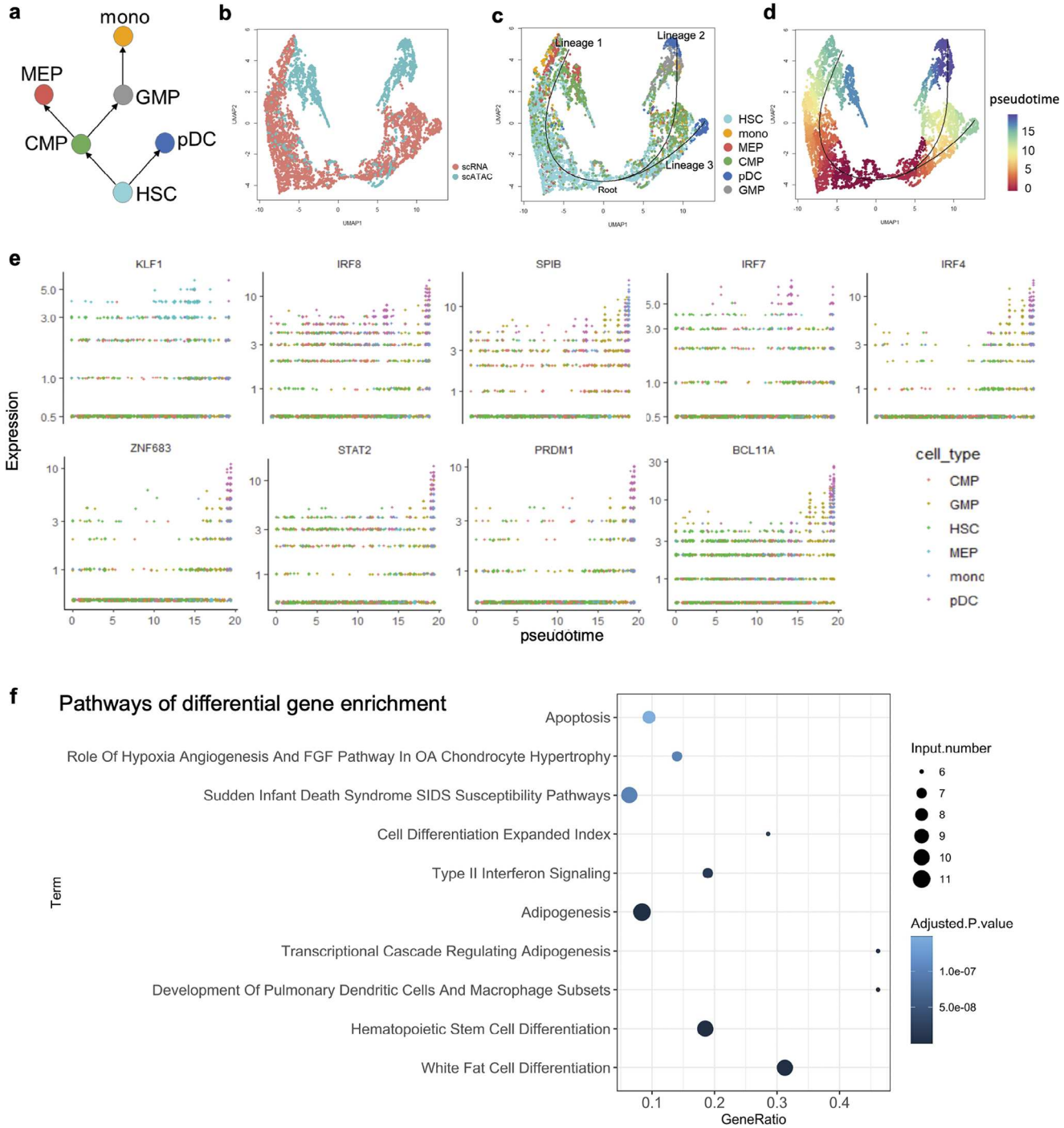


Figure 6. Trajectory inference and pseudotime analysis on the integrated scRNA-seq data and scATAC-seq data. a) Reference of human hematopoietic differentiation process from literature. hematopoietic stem cells (HSC) devided into branches, one branch differentiates in to plasmacytoid dendritic cells (pDC), the other goes through common myeloid progenitor (CMP) and differentiates into two different cell types, including megakaryocyte erythroid progenitor (MEP) and granulocyte–monocyte progenitors (GMP). Granulocyte–monocyte progenitors (GMP) can further differentiate into Monocyte (mono) cells. b) UMAP visualization of integrated scRNA-seq data and scATAC-seq data colored by different modalities. c) UMAP visualization of integrated cells along the inferred trajectory, cells are colored by different cell types. Lineage 1 and Lineage 2 were inferred and smoothed using Slingshot, Lineage 3 was manually added for a more intuitive representation. d) UMAP visualization of integrated scRNA-seq data and scATAC-seq data colored by differentiation pseudotime. e) scatter plot of gene expression along the pseodotime of nine highly-correlated genes. f) bubble plot of enriched pathways of differentially expressed genes.

the context of integrated differentiation data. In essence, CCAN emerges as a powerful tool for unraveling the complexities of cellular dynamics across heterogeneous datasets.

In CCAN, we use a domain separation network that requires the source domain data and the target domain data to have the same number of features. The label transferring in the domain separation network limits CCAN to be used only when the data in the source domain and the target domain have the same cell types. This limitation is not prone to over-correction, which is likely to occur especially when integrating collections of datasets with considerable differences in cellular composition. Currently, there are many methods for analyzing and removing cell cycle effects based on scRNA-seq data, but there are few methods for cell cycle analysis based on scATAC-seq data. Therefore, in this manuscript, the cell cycle effects analysis was focused on the integration of scRNA-seq data from different platforms. When integrating scRNA-seq and scATAC-seq data, we can regard the private-encoder embedding as a confounding factor that affects data integration and cell classification. Besides, we only used single-cell datasets with two modalities, including gene expression profiling from scRNA-seq data and chromatin accessibility from scATAC-seq data. However, with the ongoing advancements in high-throughput single-cell sequencing technologies, the accessibility to analyze various molecular components such as DNA, mRNA, and proteins at a single-cell resolution has expanded significantly. Recognizing the potential for a more nuanced understanding through the integration of diverse omics data types, we envision extending the capabilities of CCAN in future development. The objective is to broaden CCAN's scope to encompass the integration and comprehensive analysis of a more extensive array of single-cell omics data, thereby facilitating diverse analytical objectives.

4. Experimental Section

Basic Structure of Domain Separation Network in CCAN: Domain separation network (DSN) is a specific type of neural network architecture used for domain adaptation, it was designed based on the labeled source-domain data and unlabeled target-domain data.[43] In CCAN, source-domain data is gene expression profiles of scRNA-seq data and target-domain data can be same/different single-cell omics data, such as gene expression of scRNA-seq data from different protocols, chromatin accessibility of scATAC-seq data, etc. In the basic structure of DSN, it contains one shared encoder and two domain-specific private encoders to extract the common and private representations from input data. The common and private components of the same domain should be totally split to make sure the independence of these parts. A shared decoder reconstructs the input domain by cascading the shared and private embeddings. Given a labeled dataset in a source domain and an unlabeled dataset in a target domain, cell type classification was mainly used as the cross-domain task, training on the shared embedding from the source domain that generalizes to the target domain, which requires the high invariance between shared embeddings of source and target domains. To achieve this goal, alignment was considered in the embedding space to eliminate differences between domains. Objectively, DSN is a model that produces a shared representation that is similar for both domains and a private representation that is different and transfers the classification label from source domain to target domain.

Cell Cycle-Aware Module in CCAN: The cell cycle has been recognized as a confounding factor in the analysis of cell type-dependent processes. In CCAN, to achieve better cell type transfer from the source domain to the target domain, it is necessary not only to account for differences in the shared embeddings between different domains, but also to remove cell cycle effects from the shared representations. Considering the mutual independence between shared and private components in DSN, a cell cycle-aware module was introduced in the private part of DSN to characterize the dynamic process of cell cycle. The high degree of differences between shared and private components can make shared components out of the influence of private components, that is, the prediction of cell types based on shared components can eliminate the effect of cell cycle modeling based on private components. Taking the source data as X_s and the target data as X_t, the objective of the cell cycle-aware module is to infer the cell cycle pseudotime Z_p for cells from their corresponding profiles X_s or X_t. DSN is an autoencoder-based network, in the encoder, a standard multi-layer perceptron with hyperbolic tangent activation functions was used in the private part (also as circular part) and selu activation functions in the shared part (also as acyclic part). The encoder of source data X_s is as below

$$Z_s = \begin{bmatrix} Z_{sp}^{(circular)} \\ Z_{ss}^{(acyclic)} \end{bmatrix} = \begin{bmatrix} W_3^{(circular)} \tanh\left(W_2^{(circular)} \tanh\left(W_1^{(circular)} X_s + b_1\right) + b_2\right) \\ W_3^{(acyclic)} selu\left(W_2^{(acyclic)} \text{selu}\left(W_1^{(acyclic)} X_s + b_1\right) + b_2\right) \end{bmatrix} \quad (1)$$

where W's and b's are the weight matrices and bias vectors of the encoder. Z_{sp} and Z_{ss} are private embedding and shared embedding of the source domain. Z_s represents the cascade of Z_{sp} and Z_{ss} in the hidden layer. In the decoder, cosine and sine were used as the activation functions in the first layer, followed by two layers performing linear transformations. The reconstruction of source data $\widehat{X_S}$ can be represented mathematically as

$$\hat{X}_s = \left[V^{(circular)} V^{(acyclic)}\right] \begin{bmatrix} \sin Z_{sp}^{(circular)} \\ \cos Z_{sp}^{(circular)} \\ Z_{ss}^{(acyclic)} \end{bmatrix} = VZ_s \quad (2)$$

where V's are the weight matrices of the decoder and $\widehat{X_S}$ is the reconstructed matrix of source domain generated by the decoder. The target domain has the same autoencoder structure of cell cycle-aware module in the private part as follows.

$$Z_t = \begin{bmatrix} Z_{tp}^{(circular)} \\ Z_{ts}^{(acyclic)} \end{bmatrix} = \begin{bmatrix} W_6^{(circular)} \tanh\left(W_5^{(circular)} \tanh\left(W_4^{(circular)} X_t + b_1\right) + b_2\right) \\ W_3^{(acyclic)} selu\left(W_2^{(acyclic)} \text{selu}\left(W_1^{(acyclic)} X_t + b_1\right) + b_2\right) \end{bmatrix} \quad (3)$$

$$\hat{X}_t = \left[V^{(circular)} V^{(acyclic)}\right] \begin{bmatrix} \sin Z_{tp}^{(circular)} \\ \cos Z_{tp}^{(circular)} \\ Z_{ts}^{(acyclic)} \end{bmatrix} = VZ_t \quad (4)$$

where Z_{tp} and Z_{ts} are private embedding and shared embedding of the target domain. Z_t represents the cascade of Z_{tp} and Z_{ts} in the hidden layer. $\hat{X}_t$ is the reconstructed matrix of target domain.

Algorithm Design and Training of CCAN: CCAN uses a domain separation network (DSN) to integrate data from source and target domains and transfer the annotations from source domain to target domain. Shared encoder and private encoders in DSN are three-layer perceptrons to learn circular and acyclic embeddings of both modalities, separately. The shared

embedding function projects a high-dimensional profile of each cell to a low-dimensional vector, which distinguishes biological meaningful signals from circular confounding factors (private embedding) and transforms the embeddings of cells from different domains into a similar distribution. In the decoder, sine and cosine were used as the activation functions specific for private embeddings, followed by a two-layer perceptron performing noncircular transformations mapping the embedded data to the original space. CCAN used the labeled transcriptomic profile of scRNA-seq data (source domain) and unlabeled profile from same/different omics data (target domain) as input. Taking scRNA-seq data as source domain and scATAC-seq data as target domain as an example, the training of CCAN has four main steps:

Step 1: Pretraining of the cell cycle-aware domain separation network. A cell-cycle aware domain separation network was used to perform joint embedding and modality alignment in a common embedding space through a multi-objective loss $\mathcal{L}$. First, CCAN trains the source and target autoencoders by minimizing the data reconstruction error.

$$\mathcal{L}_{recons} = \frac{1}{N_s}\sum_{i=1}^{N_s}\left\|X_s^{(i)} - \widehat{X_s^{(i)}}\right\|^2 + \frac{1}{N_t}\sum_{i=1}^{N_t}\left\|X_t^{(i)} - \widehat{X_t^{(i)}}\right\|^2 \quad (5)$$

where N_s and N_t are number of cells in source and target data, X_s and X_t are input source and target data, $\widehat{X}_s$ and $\widehat{X}_t$ are the decoder-reconstructed data. Second, it learns shared signals between the scRNA-seq data and scATAC-seq data as well as private signals that are unique to the scRNA-seq data and scATAC-seq data. CCAN applies an orthogonal constrain $\mathcal{L}_{diff}$ to push the features of the shared and the private embeddings apart from each other

$$\mathcal{L}_{diff} = \left\|Z_{ss}^T Z_{sp}\right\|^2 + \left\|Z_{ts}^T Z_{tp}\right\|^2 \quad (6)$$

where Z_{ss} and Z_{ts} are the embedded source and target data from shared encoder, Z_{sp} and Z_{tp} are the embedded data from domain-specific private encoders. The rationale is to disentangle biological signals specific for cell type heterogeneity from cell cycle confounders. Third, CCAN regularizes the embeddings of scRNA-seq data and scATAC-seq data to make their distributions to be similar. Aligning the distributions across cells from different domains in the shared embedding space can alleviate the out-of-distribution problem between different modalities. The Maximum Mean Discrepancy loss $\mathcal{L}_{MMD}$ was used as the alignment loss $\mathcal{L}_{align}$ to align the distribution of the scRNA-seq data and scATAC-seq data

$$\mathcal{L}_{align} = \mathcal{L}_{MMD} = \left\|\frac{1}{N_s}\sum_{i=1}^{N_s} k\left(Z_{ss}^{(i)}\right) - \frac{1}{N_t}\sum_{i=1}^{N_t} k\left(Z_{ts}^{(i)}\right)\right\|^2 \quad (7)$$

where k represents the kernel function. Finally, after the unsupervised pretraining of the autoencoder, a supervised cell type classification model can be trained from the aligned shared embedding using the labeled scRNA-seq data, refer as a cross-entropy classification loss $\mathcal{L}_{class}$

$$\mathcal{L}_{class} = \sum_{i=1}^{K} y_i \log\left(p_i\right) \quad (8)$$

where y_i is ground-truth cell type label of scRNA-seq data and p_i is the predicted probability generated by the cell type classifier. In summary, the total loss in the pretraining of CCAN is

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{recons} + \mathcal{L}_{diff} + \mathcal{L}_{align} \quad (9)$$

Multiple losses will be given different weights during training.

Step 2: Label transferring from source domain to target domain. After pretraining of CCAN, the trained cell type classification network is applied to the unlabeled scATAC-seq data, transferring the cell type information of scRNA-seq data to annotate the scATAC-seq data, regarded as label transferring. CCAN makes it possible to remove circular confounders while performing accurate annotations for scATAC-seq data. After the label transferring in step 2, a pseudo-label of the scATAC-seq data was obtained.

Step 3: Refining CCAN by introducing a cluster alignment loss. The joint embedding and label prediction performance was improved using the ground truth of scRNA-seq data and the pseudo-label of the scATAC-seq data. A cluster alignment loss $\mathcal{L}_{ca}$ was added to refine the neural networks in CCAN.

$$\mathcal{L}_{ca} = \frac{1}{N_s N_t}\sum_{i=1}^{N_s}\sum_{j=1}^{N_t}\left[\delta d\left(z_{ss}^i, z_{ts}^j\right) + (1-\delta)\max\left(0, m - d\left(z_{ss}^i, z_{ts}^j\right)\right)\right] + \frac{1}{K}\sum_{k=1}^{K}\left\|\lambda_s - \lambda_t\right\|^2 \quad (10)$$

where m is the distance threshold, λ_s and λ_t are centroids of embedded source and target data, K is the number of cell types. $\mathcal{L}_{ca}$ is a class-conditional loss that forces the features from the same class to concentrate together and the features from different classes to be separated. In addition, it introduces a conditional feature matching loss to improve the alignment between two domains that aligns the clusters which correspond to the same class but come from different domains. Thus, the updated alignment loss is

$$\mathcal{L}_{align} = \mathcal{L}_{MMD} + \mathcal{L}_{ca} \quad (11)$$

Step 4: Finalization of CCAN model. The last step generates the joint profile of two modalities and finalizes the annotation of scATAC-seq data. Besides, more downstream analyses can be operated on the joint embedding profile of scRNA-seq and scATAC-seq data when the shared embedding of scRNA-seq and scATAC-seq was cascaded.

Balanced Mini-Batch Training for Cluster Alignment: In the refining step, a cluster alignment loss in CCAN was introduced, which challenges the choice of batch size in neural network training. The batch size selection in the pretraining step is no longer suitable for the refining step to calculate the cluster alignment loss, because it cannot guarantee that each batch size of data can contain all cell types, and may cause extreme imbalance of batch-size data, which will affect the performance of class alignment. To overcome this problem, a balanced mini-batch training in the optimization of the cluster alignment loss was applied that can virtually balance the class ratio of training samples in CCAN. The numbers of samples for each class in a mini-batch are restricted to be the same. This method does not modify or discard cells in the input data, so it can avoid oversampling and under-sampling problems while improving cluster alignment performance, achieving a better integration of single-cell multi-omics data.

CCAN Parameters: By default, the following parameters were set for CCAN: batch size: 64, the hidden-layer dimensions in the encoder: [512, 256], the hidden-layer dimensions in the classifier: [32, 16], the dimension of latent space: 64, learning rate: 1e-3, number of training epochs: 1000. Parameters are optimized via grid search and may vary based on input data.

Datasets: CCAN is an effective tool for multitasking. In order to verify the performance of CCAN in different application scenarios, several real single-cell datasets were used (Table S1, Supporting Information), including scRNA-seq data of peripheral blood mononuclear cells from different protocols,[15c] breast cancer single-cell dataset,[33] paired scRNA-seq and scATAC-seq data generated from SNARE-seq technology,[12c] unpaired scRNA-seq and scATAC-seq data of different cells from the same tissue [22] and single-cell multi-omics data of human hematopoietic differentiation.[23,44] Virtual datasets were also generated based on the real mESC data, as shown below [28]:

1) scRNA-seq data of PBMCs: The scRNA-seq data of PBMCs consisted of two scRNA-seq data, each assayed on the Chromium 10X platform, but using different library construction protocols: 3′ end v1 (3pV1) and 3′ end v2 (3pV2) chemistries. This dataset was pre-processed following the vignette in Seurat (https://satijalab.org/seurat/articles/pbmc3k_tutorial). The 3pV1 scRNA-seq data has 5356 cells after data pre-processing and the 3pV2 scRNA-seq data has 8806 cells after data pre-processing.

2) mESCs data: This scRNA-seq dataset of mouse embryonic stem cells were downloaded from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2805/. The cells were stained with Hoechst and sorted using FACS for respective cell-cycle fractions (G1, S and G2/M phase). Two hundred eighty-eight mouse embryonic stem cells were sequenced using HighSeq 2000 sequencing system.
3) Breast cancer single-cell dataset: This data comprised patients with ER+ breast cancer undergoing neoadjuvant endocrine therapy (letrozole) with or without a CDK4/6 inhibitor (ribociclib), sampled at the start of treatment, after 14 days, and after 180 days of treatment, using 10x technology for single-nucleus RNA sequencing. This dataset is available from Gene Expression Omnibus (GEO) with GEO Series ID GSE158724.
4) SNARE-seq data: Single-nucleus chromatin accessibility and mRNA expression sequencing (SNARE-seq) is a droplet-based method to simultaneously profile transcriptome and chromatin accessibility in single nucleus. This dataset is regarded as paired single-cell RNA-seq and ATAC-seq data for adult mouse cerebral cortex. SNARE-seq data of adult mouse brain was pre-processed following the vignette in Signac (https://stuartlab.org/signac/1.2.0/articles/snareseq.html). 8055 cells were used in CCAN after data pre-processing. This dataset is available from Gene Expression Omnibus (GEO) with GEO Series ID GSE126074.
5) Unpaired scRNA-seq and scATAC-seq data: This dataset includes scRNA-seq data from CITE-seq and scATAC-seq data from ASAP-seq. The CITE-seq generates filtered scRNA-seq data based on the condition of mitochondrial reads greater than 10%, number of expressed genes fewer than 500 and total number of UMI fewer than 1000. For the scATAC-seq data from ASAP-seq, cells with a number of peaks more than 100 000 were filtered out and calculated the gene activity matrices for scATAC-seq data using Signac.[34] After that, 17 668 overlapped genes are selected as the input features of CCAN. This dataset is available from Gene Expression Omnibus (GEO) with GEO Series ID GSE156478.
6) Single-cell differentiated data: This dataset includes scRNA-seq and scATAC-seq of human hematopoietic differentiation. The scATAC-seq and scRNA-seq were performed separately on different cells and there is no paired relationship between cells from the scRNA-seq data and cells from the scATAC-seq data. Six cell types shared by scRNA-seq and scATAC-seq were selected and used in CCAN, they are hematopoietic stem cells (HSC), plasmacytoid dendritic cells (pDC), common myeloid progenitor (CMP), megakaryocyte erythroid progenitor (MEP), Monocyte (mono) and granulocyte–monocyte progenitors (GMP).

Input Format for Compared Methods: CCAN was compared with several existing integration methods for different integration situations (Table S2, Supporting Information). These methods are developed for different integration types, so they have different format requirements for input data.

1) Harmony, Conos, Scanorama and BBKNN: These methods are designed specific for scRNA-seq data integration. Their required input is scRNA-seq data with gene expression matrix.
2) Seurat V4 and online_iNMF: The scRNA-seq data is gene expression matrix with genes as rows and cells as columns. The scATAC-seq data is gene activity matrix pro-pressed using Signac.
3) scMVP: scMVP takes raw count of scRNA-seq and term frequency–inverse document frequency (TF-IDF) transformed scATAC-seq as input.
4) scAI: The data of scRNA-seq is a matrix with genes as rows and cells as columns. The data of scATAC-seq is a sparse/binary epigenomic profile with regions as rows and cells as columns.
5) scMVAE: The raw count data of scRNA-seq and scATAC data (gene activity format). Row indicates variable (genes and loci), and column indicates sample (cell).
6) scJoint: The input of scJoint consists of gene activity score matrix, calculated from the accessibility peak matrix of scATAC-seq, and gene expression matrix including cell-type labels from scRNA-seq experiments.
7) GLUE: GLUE requires gene expression profile and peak matrix with *AnnData* format as input. In the pre-processing step of scRNA-seq data, highly variable genes were selected using Seurat v3. Then, the data is normalized and scaled, and PCA dimensionality reduction is performed, using 100 principal components by default. For scATAC-seq, GLUE uses LSI dimensionality reduction with a default dimensionality of 100.
8) SCALEX: SCALEX requires sparse matrix of gene expression and gene activity as two batches. Although it is designed for unpaired data, it can be applied to paired data as well via changing the cell names of one batch.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This study was supported by the National Institutes of Health [R01GM123037, U01AR069395, R01DE027027 and R01CA241930 to X.Z.] the National Science Foundation [NSF2217515 and NSF2326879 to X.Z.]; The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Funding for open access charge: Dr & Mrs Carl V. Vartian Chair Professorship Funds to Dr. Zhou from the University of Texas Health Science Center at Houston.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

J.L. and X.Z. conceived the study. J.L. developed the model and performed all analyses. J.M. implemented balanced mini-batch sampling in the model and helped test the methods. J.W. helped with cell cycle identification and cell cycle dysregulation analysis. J.L. wrote the manuscript and all authors revised it. All authors read and approved the final version of the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

batch effect, cell cycle effect, domain separation network, single-cell multi-omics integration

Received: February 20, 2024
Revised: April 22, 2024
Published online: June 17, 2024

[1] a) M. S. Kowalczyk, I. Tirosh, D. Heckl, T. N. Rao, A. Dixit, B. J. Haas, R. K. Schneider, A. J. Wagers, B. L. Ebert, A. Regev, *Genome Res.* **2015**, *25*, 1860; b) I. C. Macaulay, C. P. Ponting, T. Voet, *Trends Genet.* **2017**, *33*, 155; c) E. Papalexi, R. Satija, *Nat. Rev. Immunol.* **2018**, *18*, 35; d) A. Tanay, A. Regev, *Nature.* **2017**, *541*, 331.

[2] a) L. Gonzalez-Silva, L. Quevedo, I. Varela, *Trends Cancer.* **2020**, *6*, 13; b) D. Lahnemann, J. Koster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, et al., *Genome Biol.* **2020**, *21*, 31; c) F. Wu, J. Fan, Y. He, A. Xiong, J. Yu, Y. Li, Y. Zhang, W. Zhao, F. Zhou, W. Li, J. Zhang, X. Zhang, M. Qiao, G. Gao, S. Chen, X. Chen, X. Li, L. Hou, C. Wu, C. Su, S. Ren, M. Odenthal, R. Buettner, N. Fang, C. Zhou, *Nat. Commun.* **2021**, *12*, 2540;

[3] E. Mereu, A. Lafzi, C. Moutinho, C. Ziegenhain, D. J. McCarthy, A. Alvarez-Varela, E. Batlle, D. G. Sagar, J. K. Lau, S. C. Boutet, C. Sanada, A. Ooi, R. C. Jones, K. Kaihara, C. Brampton, Y. Talaga, Y. Sasagawa, K. Tanaka, T. Hayashi, C. Braeuning, C. Fischer, S. Sauer, T. Trefzer, C. Conrad, X. Adiconis, L. T. Nguyen, A. Regev, J. Z. Levin, S. Parekh, A. Janjic, et al., *Nat. Biotechnol.* **2020**, *38*, 747.

[4] a) A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, *Nat. Biotechnol.* **2018**, *36*, 411; b) J. Gehring, J. Hwee Park, S. Chen, M. Thomson, L. Pachter, *Nat. Biotechnol.* **2020**, *38*, 35; c) L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, *Nat. Biotechnol.* **2018**, *36*, 421.

[5] M. Eisenstein, *Nat. Biotechnol.* **2020**, *38*, 254.

[6] M. D. Luecken, M. Buttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colome-Tatche, F. J. Theis, *Nat. Methods.* **2022**, *19*, 41.

[7] a) M. Efremova, S. A. Teichmann, *Nat. Methods.* **2020**, *17*, 14; b) H. Gao, B. Zhang, L. Liu, S. Li, X. Gao, B. Yu, *Brief Bioinform.* **2023**, *24*, bbad081; c) L. Heumos, A. C. Schaar, C. Lance, A. Litinetskaya, F. Drost, L. Zappia, M. D. Lucken, D. C. Strobl, J. Henao, F. Curion, C. H. B. Schiller, F. J. Theis, *Nat. Rev. Genet.* **2023**, *24*, 550.

[8] a) J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, *Nat. Methods.* **2013**, *10*, 1213; b) H. Chen, C. Lareau, T. Andreani, M. E. Vinyard, S. P. Garcia, K. Clement, M. A. Andrade-Navarro, J. D. Buenrostro, L. Pinello, *Genome Biol.* **2019**, *20*, 241; c) S. Sinha, A. T. Satpathy, W. Zhou, H. Ji, J. A. Stratton, A. Jaffer, N. Bahlis, S. Morrissy, J. A. Biernaskie, *Genomics Proteomics Bioinformatics* **2021**, *19*, 172.

[9] a) J. M. Granja, S. Klemm, L. M. McGinnis, A. S. Kathiria, A. Mezger, M. R. Corces, B. Parks, E. Gars, M. Liedtke, G. X. Y. Zheng, H. Y. Chang, R. Majeti, W. J. Greenleaf, *Nat. Biotechnol.* **2019**, *37*, 1458; b) R. K. Kawaguchi, Z. Tang, S. Fischer, C. Rajesh, R. Tripathy, P. K. Koo, J. Gillis, *Brief Bioinform.* **2023**, *24*, bbac541.

[10] a) R. Fang, S. Preissl, Y. Li, X. Hou, J. Lucero, X. Wang, A. Motamedi, A. K. Shiau, X. Zhou, F. Xie, E. A. Mukamel, K. Zhang, Y. Zhang, M. M. Behrens, J. R. Ecker, B. Ren, *Nat. Commun.* **2021**, *12*, 1337; b) Z. Ji, W. Zhou, W. Hou, H. Ji, *Genome Biol.* **2020**, *21*, 161.

[11] 10x Genomics, Cell Type Annotation Strategies for Single Cell ATAC-Seq Data, Technical Note, Document Number CG000234, 10xGenomics, Pleasanton, CA **2020**.

[12] a) S. Ma, B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, T. Law, C. Lareau, Y. C. Hsu, A. Regev, J. D. Buenrostro, *Cell.* **2020**, *183*, 1103; b) J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, F. J. Steemers, A. C. Adey, C. Trapnell, J. Shendure, *Science* **2018**, *361*, 1380; c) S. Chen, B. B. Lake, K. Zhang, *Nat. Biotechnol.* **2019**, *37*, 1452; d) C. Zhu, M. Yu, H. Huang, I. Juric, A. Abnousi, R. Hu, J. Lucero, M. M. Behrens, M. Hu, B. Ren, *Nat. Struct. Mol. Biol.* **2019**, *26*, 1063.

[13] A. Ma, G. Xin, Q. Ma, *Nat. Commun.* **2022**, *13*, 2728.

[14] R. Argelaguet, A. S. E. Cuomo, O. Stegle, J. C. Marioni, *Nat. Biotechnol.* **2021**, *39*, 1202.

[15] a) N. Barkas, V. Petukhov, D. Nikolaeva, Y. Lozinsky, S. Demharter, K. Khodosevich, P. V. Kharchenko, *Nat. Methods.* **2019**, *16*, 695; b) B. Hie, B. Bryson, B. Berger, *Nat. Biotechnol.* **2019**, *37*, 685; c) I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, S. Raychaudhuri, *Nat. Methods.* **2019**, *16*, 1289; d) K. Polanski, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, J. E. Park, *Bioinformatics.* **2020**, *36*, 964.

[16] S. Jin, L. Zhang, Q. Nie, *Genome Biol.* **2020**, *21*, 25.

[17] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck 3rd, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, *Cell.* **2021**, *184*, 3573.

[18] a) G. Li, S. Fu, S. Wang, C. Zhu, B. Duan, C. Tang, X. Chen, G. Chuai, P. Wang, Q. Liu, *Genome Biol.* **2022**, *23*, 20; b) C. Zuo, L. Chen, *Brief Bioinform.* **2021**, *22*, bbaa287.

[19] a) Z. J. Cao, G. Gao, *Nat. Biotechnol.* **2022**, *40*, 1458; b) Y. Lin, T. Y. Wu, S. Wan, J. Y. H. Yang, W. H. Wong, Y. X. R. Wang, *Nat. Biotechnol.* **2022**, *40*, 703; c) L. Xiong, K. Tian, Y. Li, W. Ning, X. Gao, Q. C. Zhang, *Nat. Commun.* **2022**, *13*, 6118.

[20] C. Gao, J. Liu, A. R. Kriebel, S. Preissl, C. Luo, R. Castanon, J. Sandoval, A. Rivkin, J. R. Nery, M. M. Behrens, J. R. Ecker, B. Ren, J. D. Welch, *Nat. Biotechnol.* **2021**, *39*, 1000.

[21] a) M. Barron, J. Li, *Sci. Rep.* **2016**, *6*, 33892; b) M. Chen, X. Zhou, *Sci. Rep.* **2017**, *7*, 13587; c) S. Liang, F. Wang, J. Han, K. Chen, *Nat. Commun.* **2020**, *11*, 1441; d) J. Liu, M. Yang, W. Zhao, X. Zhou, *Nucleic Acids Res.* **2022**, *50*, 704; e) S. C. Zheng, G. Stein-O'Brien, J. J. Augustin, J. Slosberg, G. A. Carosso, B. Winer, G. Shin, H. T. Bjornsson, L. A. Goff, K. D. Hansen, *Genome Biol.* **2022**, *23*, 41.

[22] E. P. Mimitou, C. A. Lareau, K. Y. Chen, A. L. Zorzetto-Fernandes, Y. Hao, Y. Takeshima, W. Luo, T. S. Huang, B. Z. Yeung, E. Papalexi, P. I. Thakore, T. Kibayashi, J. B. Wing, M. Hata, R. Satija, K. L. Nazor, S. Sakaguchi, L. S. Ludwig, V. G. Sankaran, A. Regev, P. Smibert, *Nat. Biotechnol.* **2021**, *39*, 1246.

[23] J. D. Buenrostro, M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang, W. J. Greenleaf, *Cell.* **2018**, *173*, 1535.

[24] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.: Theory Exp.* **2008**, *2008*, P10008.

[25] H. Guo, J. Li, *Genome Biol.* **2021**, *22*, 69.

[26] L. McInnes, J. Healy, J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction **2018**.

[27] M. Buttner, Z. Miao, F. A. Wolf, S. A. Teichmann, F. J. Theis, *Nat. Methods.* **2019**, *16*, 43.

[28] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, O. Stegle, *Nat. Biotechnol.* **2015**, *33*, 155.

[29] T. Wang, T. S. Johnson, W. Shao, Z. Lu, B. R. Helm, J. Zhang, K. Huang, *Genome Biol.* **2019**, *20*, 165.

[30] Z. Liu, H. Lou, K. Xie, H. Wang, N. Chen, O. M. Aparicio, M. Q. Zhang, R. Jiang, T. Chen, *Nat. Commun.* **2017**, *8*, 22.

[31] A. Santos, R. Wernersson, L. J. Jensen, *Nucleic Acids Res.* **2015**, *43*, D1140.

[32] H. K. Matthews, C. Bertoli, R. A. M. de Bruin, *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 74.

[33] J. I. Griffiths, J. Chen, P. A. Cosgrove, A. O'Dea, P. Sharma, C. Ma, M. Trivedi, K. Kalinsky, K. B. Wisinski, R. O'Regan, I. Makhoul, L. M. Spring, A. Bardia, F. R. Adler, A. L. Cohen, J. T. Chang, Q. J. Khan, A. H. Bild, *Nat Cancer* **2021**, *2*, 658.

[34] T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, R. Satija, *Nat. Methods.* **2021**, *18*, 1333.

[35] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, S. Dudoit, *BMC Genomics.* **2018**, *19*, 477.

[36] A. T. Satpathy, J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, P. Shah, J. C. Bell, D. Jhutty, C. M. Nemec, J. Wang, L. Wang, Y. Yin, P. G. Giresi, A. L. S. Chang, G. X. Y. Zheng, W. J. Greenleaf, H. Y. Chang, *Nat. Biotechnol.* **2019**, *37*, 925.

[37] J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, *Nature*. **2019**, *566*, 496.
[38] H. Chen, L. Albergante, J. Y. Hsu, C. A. Lareau, G. Lo Bosco, J. Guan, S. Zhou, A. N. Gorban, D. E. Bauer, M. J. Aryee, D. M. Langenau, A. Zinovyev, J. D. Buenrostro, G. C. Yuan, L. Pinello, *Nat. Commun.* **2019**, *10*, 1903.
[39] M. R. Tallack, G. W. Magor, B. Dartigues, L. Sun, S. Huang, J. M. Fittock, S. V. Fry, E. A. Glazov, T. L. Bailey, A. C. Perkins, *Genome Res.* **2012**, *22*, 2385.
[40] a) Y. A. Ko, Y. H. Chan, C. H. Liu, J. J. Liang, T. H. Chuang, Y. P. Hsueh, Y. L. Lin, K. I. Lin, *Front Immunol* **2018**, *9*, 1828; b) G. C. Ippolito, J. D. Dekker, Y. H. Wang, B. K. Lee, A. L. Shaffer 3rd, J. Lin, J. K. Wall, B. S. Lee, L. M. Staudt, Y. J. Liu, V. R. Iyer, H. O. Tucker, *Proc Natl Acad Sci U S A.* **2014**, *111*, E998; c) T. Tamura, P. Tailor, K. Yamaoka, H. J. Kong, H. Tsujimura, J. J. O'Shea, H. Singh, K. Ozato, *J. Immunol.* **2005**, *174*, 2573; d) S. Ning, J. S. Pagano, G. N. Barber, *Genes Immun.* **2011**, *12*, 399; e) D. Sichien, C. L. Scott, L. Martens, M. Vanderkerken, S. Van Gassen, M. Plantinga, T. Joeris, S. De Prijck, L. Vanhoutte, M. Vanheerswynghels, G. Van Isterdael, W. Toussaint, F. B. Madeira, K. Vergote, W. W. Agace, B. E. Clausen, H. Hammad, M. Dalod, Y. Saeys, B. N. Lambrecht, M. Guilliams, *Immunity.* **2016**, *45*, 626; f) J. Dai, N. J. Megjugorac, S. B. Amrute, P. Fitzgerald-Bocarsly, *J. Immunol.* **2004**, *173*, 1535; g) B. Cisse, M. L. Caton, M. Lehner, T. Maeda, S. Scheu, R. Locksley, D. Holmberg, C. Zweier, N. S. den Hollander, S. G. Kant, W. Holter, A. Rauch, Y. Zhuang, B. Reizis, *Cell.* **2008**, *135*, 37.
[41] M. I. Love, W. Huber, S. Anders, *Genome Biol.* **2014**, *15*, 550.
[42] a) C. A. Spek, H. L. Aberson, J. M. Butler, A. F. de Vos, J. Duitman, *Cells* **2021**, *10*, 2233; b) P. Shyamsunder, M. Shanmugasundaram, A. Mayakonda, P. Dakle, W. W. Teoh, L. Han, D. Kanojia, M. C. Lim, M. Fullwood, O. An, H. Yang, J. Shi, M. Z. Hossain, V. Madan, H. P. Koeffler, *Blood.* **2019**, *133*, 2507; c) J. Y. Noh, S. Gandre-Babbe, Y. Wang, V. Hayes, Y. Yao, P. Gadue, S. K. Sullivan, S. T. Chou, K. R. Machlus, J. E. Italiano jr., M. Kyba, D. Finkelstein, J. C. Ulirsch, V. G. Sankaran, D. L. French, M. Poncz, M. J. Weiss, *J. Clin. Invest.* **2015**, *125*, 2369; d) F. Aguilo, S. Avagyan, A. Labar, A. Sevilla, D. F. Lee, P. Kumar, I. R. Lemischka, B. Y. Zhou, H. W. Snoeck, *Blood.* **2011**, *117*, 5057; e) M. Abuhantash, E. M. Collins, A. Thompson, *Biochem. Soc. Trans.* **2021**, *49*, 1817.
[43] a) K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, *Adv Neural Inf Process Syst.* **2016**, *29*; b) D. He, Q. Liu, Y. Wu, L. Xie, *Nature Machine Intelligence.* **2022**, *4*, 879.
[44] M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, D. Pe'er, *Nat. Biotechnol.* **2019**, *37*, 451.