Genome Biology

METHOD Open Access

scNAT: a deep learning method for integrating paired single-cell RNA and T cell receptor sequencing profiles

Biqing Zhu[1,8], Yuge Wang[2], Li-Ting Ku[2], David van Dijk[3,4,8], Le Zhang[5,7,8], David A. Hafler[6,7,8] and Hongyu Zhao[1,2*]

*Correspondence:
hongyu.zhao@yale.edu

[2] Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06511, USA
Full list of author information is available at the end of the article

Abstract

Many deep learning-based methods have been proposed to handle complex single-cell data. Deep learning approaches may also prove useful to jointly analyze single-cell RNA sequencing (scRNA-seq) and single-cell T cell receptor sequencing (scTCR-seq) data for novel discoveries. We developed scNAT, a deep learning method that integrates paired scRNA-seq and scTCR-seq data to represent data in a unified latent space for downstream analysis. We demonstrate that scNAT is capable of removing batch effects, and identifying cell clusters and a T cell migration trajectory from blood to cerebrospinal fluid in multiple sclerosis.

Keywords: scRNA-seq, scTCR-seq, Data integration, Deep learning, Variational autoencoder, Clone expansion

Background

The rapid development of next-generation sequencing (NGS) technologies has enabled researchers to detect complex signals in biological samples, especially at the single-cell resolution through single-cell RNA sequencing (scRNA-seq) and multi-omics [1, 2]. With the generation of massive scRNA-seq data, it is challenging to use traditional linear statistical methods to extract complex patterns in these data. In contrast, deep learning-based methods have been widely employed thanks to their capability of modeling non-linear features [3], ability to conduct transfer learning across different domains [4], and scalability in terms of data size [5], to handle single cell data for different tasks, such as gene imputation [6, 7], batch correction [8–10], cell clustering [11, 12], TCR sequence classification [13], expression program identification [14], and multi-omics data integration [15]. For instance, SAUCIE [9] is an autoencoder designed to perform multiple tasks simultaneously, including data denoising, imputation, visualization, and clustering via various layers of regularizations, including information dimension regularization and maximal mean discrepancy correction. scGNN [6] is a graph neural network (GNN) model which conducts gene imputation and cell clustering iteratively that is regularized by a left-truncated mixture Gaussian model to account for cell-type-specific signals. Moreover, by learning a joint probabilistic distribution of the paired RNA and protein data, totalVI [16] employs a GNN model to jointly analyze the cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) data comprising data integration, joint dimension reduction, and missing protein imputation.

In parallel to scRNA-seq technologies, high-throughput single cell T cell receptor sequencing (scTCR-seq) has been developed to characterize the complex composition of immune repertoires [17, 18]. The TCR of more than 95% of all T cells consists of two chains, TCRα and TCRβ [19]; TCRα is encoded by variable (V) and joining (J) gene segments, and TCRβ by variable (V), diversity (D), and joining (J) segments. The diversity of the TCR repertoires is characterized by the V/(D)/J gene rearrangement and the variability in the complementarity determining region 3 (CDR3) encoded by the V/(D)/J junction [20]. Specifically, the CDR3 is hypervariable for both the TCRα and TCR β chains, and responsible for peptide recognition presented by the cell surface major histocompatibility complex (MHC) protein. scTCR-seq technology has enabled profiling of the diversity of the TCR repertoire, quantify the clonal expansion of T cells in the adaptive immune response [21], and characterize the associations between TCR and T cell subtypes [22].

Analyzing paired scRNA-seq and scTCR-seq data in various biological domains has led to insight into disease pathogenesis. For example, one study found that in multiple sclerosis (MS), genes related to T cell cytotoxic function and activation were upregulated in the clonally expanded T cells from cerebrospinal fluid (CSF) [23]. Similarly, clonal expansion of CD8$^+$ T cells was found in the CSF of patients with Alzheimer's disease [24]. A growing number of methods have been proposed for multi-omics data integration. For example, MOFA+ [25] integrates single-cell multi-modal data by variational inference and reconstructs a low-dimensional data embedding. scAI [26] performs an integrative analysis on transcriptomic and epigenomic data through iterative learning, employing a deep generative probabilistic model. MultiVI [27] integrates different single-cell level data modalities using a deep generative model for probabilistic analysis. However, these methods were specifically designed for numerical data, posing difficulties in their applications to TCR data that contains CDR3 and V/(D)/J gene features, which are considered as text and categorical data, respectively. On the other hand, recent methods have been developed to integrate TCR information or infer the correlation between scRNA-seq and scTCR-seq data. For example, DeepTCR [13] leverages neural networks to integrate V/(D)/J gene and CDR3 data to facilitate downstream analyses such as dimension reduction and clustering. CoNGA [28] is a graph theoretic approach for calculating the correlation between scRNA-seq data and the corresponding TCR sequence data for new cell type discovery. Tessa [29] is a Bayesian model integrating TCRs with T cell gene expression data to better understand the T cell phenotypes. However, DeepTCR only focuses on TCR data and completely neglects RNA modality which contains transcription-level information. CoNGA takes into account RNA data but it still visualizes RNA and TCR data via separate UMAP representations. Furthermore, it collapses all the cells within the same clonotype, resulting in the loss of singe-cell-level resolution. Tessa does not take into consideration the CDR3α and V/(D)/J gene information, and similar to CoNGA, tessa assumes that all the cells having the same clonotype share the same transcriptomic profile, which is unrealistic. Therefore, there is a lack of systematic approaches which jointly consider the two data types as a whole for data integration, although previous studies have revealed that T cells within the same clonotype share similar transcriptome profiles [22, 30]. Pappalardo et al. have also observed this phenomenon in the T cells from the blood of patients with MS [23], with cells within the same clonotype closer to each other in the UMAP plots, indicating similar gene expression patterns. Here, we introduce scNAT, a deep learning method for integrating paired **sc**R**NA**-seq and sc**T**CR-seq profiles, to learn patterns in these two data types through a unified latent space to facilitate downstream analyses. With scNAT, we were able to identify a cluster of T cells that are in transition state as well as a T cell migration trajectory from blood to CSF in an MS dataset.

Results

scNAT model specification

Our method is based on a variational autoencoder (VAE) to integrate both scRNA-seq and paired scTCR-seq data as well as to facilitate scalable unsupervised learning. The input consists of three parts: CDR3 features from scTCR-seq data, V/J gene features from scTCR-seq data, and RNA features from scRNA-seq data. To enable VAE integration, we first transform the α- and β-chain CDR3 variable from a discrete sequence into a continuous numerical space through a one-layer neural network followed by a three-layer convolutional neural network (CNN) (Fig. S1C). Similarly, V/J genes in both chains are provided as categorical variables and first encoded into a one-hot encoding representation and then transformed into a continuous vector by leveraging a trainable embedding layer (Fig. S1B). Additionally, the top 2000 highly variable genes are selected from the RNA data, and the dimension is then reduced by a fully connected layer to be more comparable with the scTCR-seq data (Fig. S1A). Afterwards, the three preprocessed data vectors are concatenated as input to the VAE model to learn the underlying data structure informed by both the RNA and TCR data. The model has three layers, the middle of which is the latent space parametrized by a multivariate normal distribution. Then, these data are reconstructed through deconvolution or fully connected layers (Fig. 1). After the model is trained, the integrated features representing both the scRNA-seq and the scTCR-seq information can be extracted from the latent space to facilitate downstream analyses, such as trajectory analysis or clustering. The details of each step are provided in the "Methods" section.

Application to multiple sclerosis data

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease that affects the central nervous system. MS is genetically mediated involving several immune cells and is primarily driven by T cells [31, 32]. To reveal the details of the immune processes in MS, we applied scNAT to an MS dataset in Pappalardo et al. [23], containing six healthy controls and five patients with MS. For each individual, both blood and cerebrospinal fluid (CSF) were sequenced. In total, there were 48,898 cells, including CD4$^+$ naive T cell, CD4$^+$ memory T cell, CD8$^+$ naive T cell, CD8$^+$ memory T cell, and regulatory T cell (T$_{reg}$) in blood as well as CD4$^+$ memory T cell, CD8$^+$ memory T cell, and T$_{reg}$ in

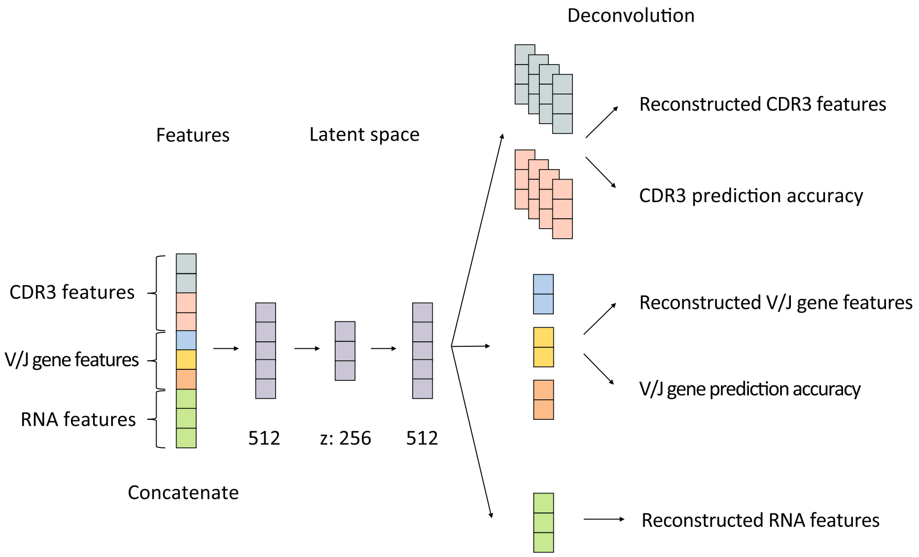

Fig. 1 A variational autoencoder (VAE) model built from the concatenated features. The CDR3, V/J gene, and RNA features were first concatenated and then treated as input to the VAE in order to learn the underlying structure

CSF (Table S1, Fig. S2). To determine the weight parameters for the loss function, we first conducted an extensive grid search on different parameter combinations (Methods, Figs. 2, 3, 4, and S3). Then, these parameters were applied to scNAT as hyperparameters for training and the latent features were extracted to facilitate downstream analysis.

scNAT latent features represent information from both the scRNA-seq as well as the scTCR-seq data

We first compared the UMAP (Methods) calculated by scRNA-seq data only (Fig. 5A–C right), scTCR-seq data only (Fig. S4A), scNAT integrated data using normalized and batched corrected RNA data (Fig. 5A–C left), and scNAT integrated data using RNA raw count data (Fig. S4B). The results showed that the scNAT integrated data with normalized and batch corrected input can preserve tissue and cell type information as the RNA data, but TCR data themselves do not provide direct information in terms of cell types. On the other hand, scNAT with RNA raw count input also struggled with learning the underlying biological structure due to early stopping, which shows the importance of data normalization (Methods). We further investigated the performances of DeepTCR [13], tessa [29], and CoNGA [28] (Methods, Fig. S5). There is no obvious pattern in the tessa TCR network to reflect biological information (Fig. S5A), and tessa was not able to retain much cell type-specific or tissue-specific information through the weighted TCR embedding (Fig. S5B-C). Similar to scNAT with scTCR-seq only data, DeepTCR results also preserve very little cell type or tissue information (Fig. S5D–E). This is expected since the structure of the TCR part of scNAT is similar to DeepTCR. On the other hand, since CoNGA reduced data into one cell per clonotype, the information from UMAP of the scRNA-seq data is very limited compared with the UMAP from the full scRNA-seq data (Figs. S5F-G, 5B-C left). And similar to tessa and DeepTCR, the CoNGA TCR UMAP structure does not reflect cell source information (Fig. S5H–I). The run time and memory information is summarized in Table S2. Taken together, we have further

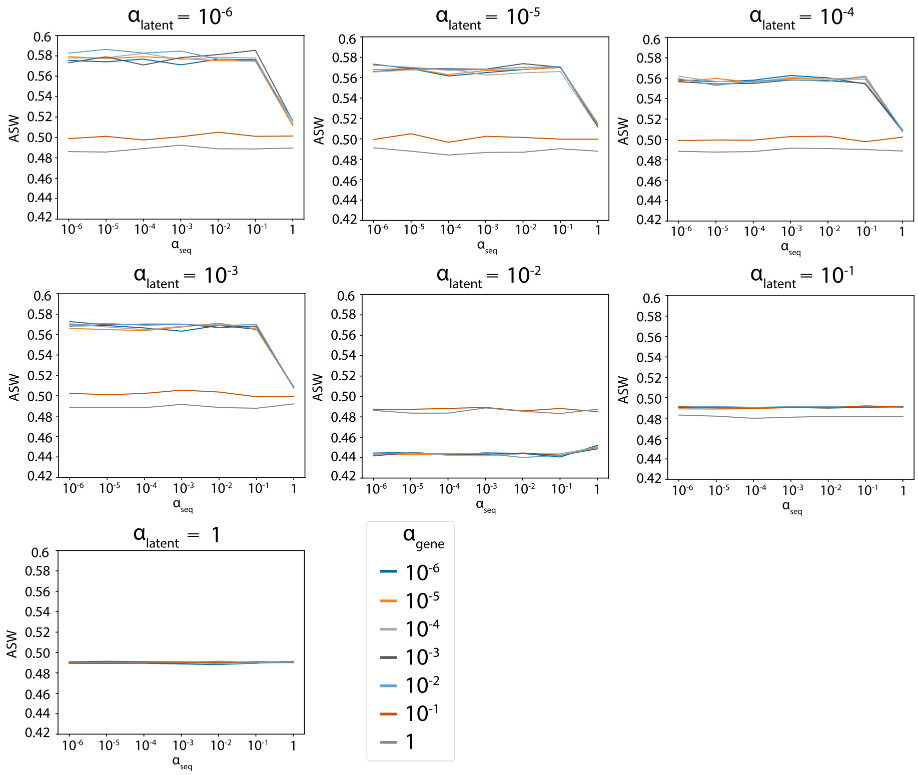

Fig. 2 ASW grid search results on the MS dataset. The ASW under different α_{latent}, α_{gene}, and α_{seq}

demonstrated that by integrating scRNA-seq and scTCR-seq data into a united latent space, scNAT can achieve better performance in terms of cell type and tissue information preservation.

Next, to quantify to which extent scNAT can preserve the information from TCR modality, GLIPH2 [33] was applied to identify T cell clusters by CDR3 similarity (Methods). We selected the six largest GLIPH2 clusters across blood and CSF which had high CD8 percentage (Fig. S6A–C). The results demonstrate that the normalized pair-wise distances within GLIPH2 clusters in the scNAT integrated space are much smaller than those in the RNA space for both tissue types (Fig. 5D, E, Methods), suggesting the clonotype-related information is also incorporated into scNAT integrated data.

scNAT can further remove remaining batch effects

We first demonstrate that scNAT can better remove batch effects than ResPAN. As stated in the previous section, both methods can distinguish blood and CSF (Fig. 5B), and the corresponding cell types in each source are clearly separated in the UMAP plots (Fig. 5C). However, in ResPAN, the batch effects from the CSF in healthy control 3 (HC3) and blood in MS 4 (MS4) were still largely preserved, but in the scNAT processed data, all samples were better mixed (Fig. 5A). To further quantify the batch removal performance, we considered two metrics, 1 − bASW [34] and 1 − kBET [35], where higher scores indicate better batch correction performance (Methods). The results showed that scNAT achieved 16% higher 1 − kBET score than ResPAN (0.387 versus 0.333), and a slightly higher 1 − bASW score (0.566 versus 0.561) (Table 1). We also compared scNAT with popular batch correction approaches designed for scRNA-seq data, including

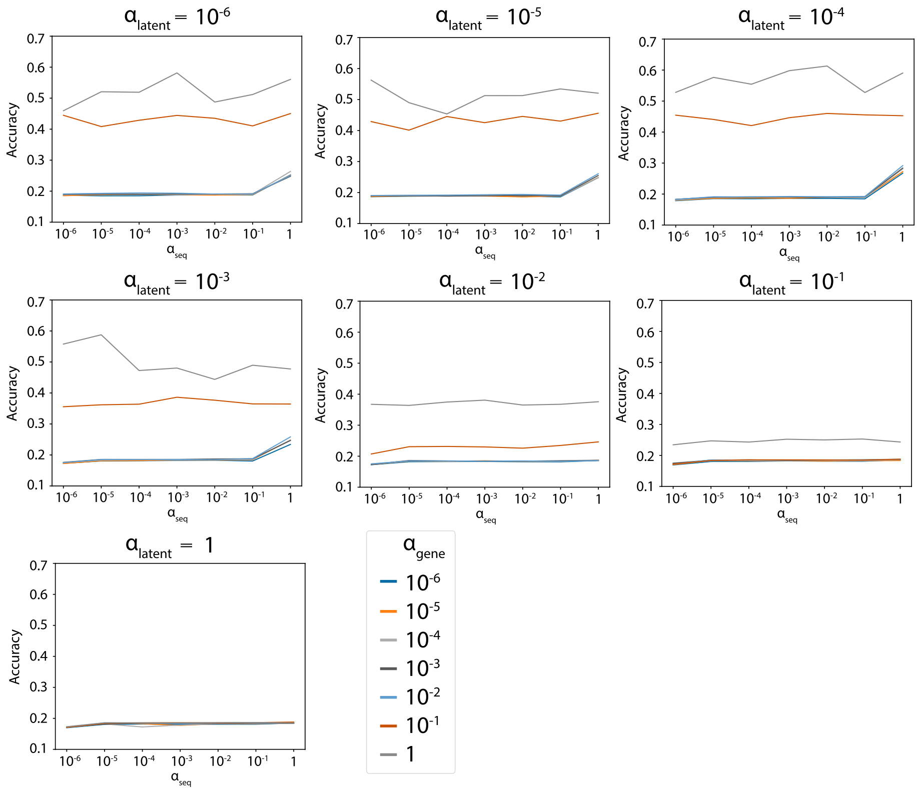

Fig. 3 Accuracy grid search results on the MS dataset. V/J gene and CDR3 prediction accuracy under different α_{latent}, α_{gene}, and α_{seq}

Seurat v4 [36], MNN [37], and BBKNN [38], and found out that scNAT consistently performed the best among these methods (Table 1). These results demonstrate that apart from data integration, scNAT can further remove remaining batch effects.

Characterization of the T cell trafficking trajectory from blood to CSF

Since previous studies have found that T cells can enter the CSF through the choroid plexus for immune surveillance [23, 39], we interrogated if the same observation can be made for the MS dataset to understand the molecular mechanism. To this end, we defined expansion score as in Pappalardo et al. [23], which is the $\log_2$ of the number of cells in a clonal group and corresponds to the clone size, and found that CD8$^+$ memory T cells in both blood and CSF had the highest score, suggesting this cell type was undergoing clonal expansion (Fig. 6A). Further, to examine whether the CD8$^+$ memory T cells in the two sources came from the same clonal group, we selected a few large clones and found that the predominance of the cells overlapped with the cells with higher expansion scores in both blood and CSF (Fig. S7A), consistent with the known observation that CD8$^+$ memory T cells may enter the CSF from blood. To further confirm this observation, we applied Slingshot [40] to infer the cell lineage and pseudotime, and the results showed that there is indeed a trajectory of T cells from blood to CSF with a small cluster of transitioning CD8$^+$ memory T cells in the middle (Fig. 6B–C, F, Methods). The application of Monocle3 [41, 42] resulted in similar patterns (Fig. S8A-B). A trajectory was identified from blood CD8 memory T cells to CSF, with the increase of pseudotime. To

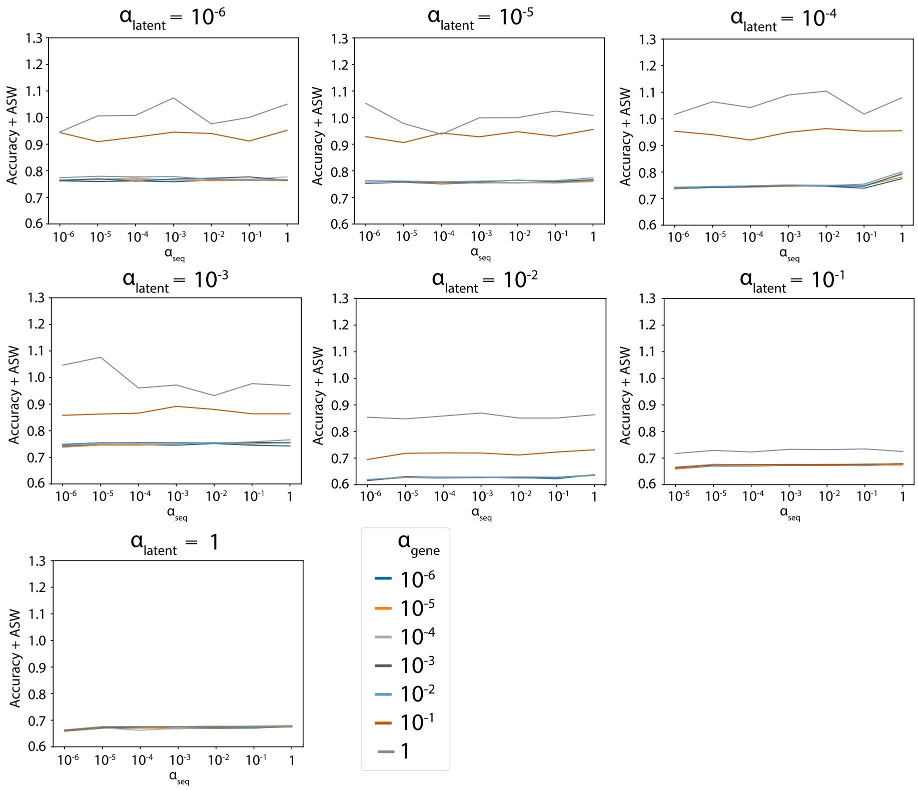

Fig. 4 ASW and accuracy grid search results on the MS dataset. The sum of ASW and accuracy under different α_{latent}, α_{gene}, and α_{seq}

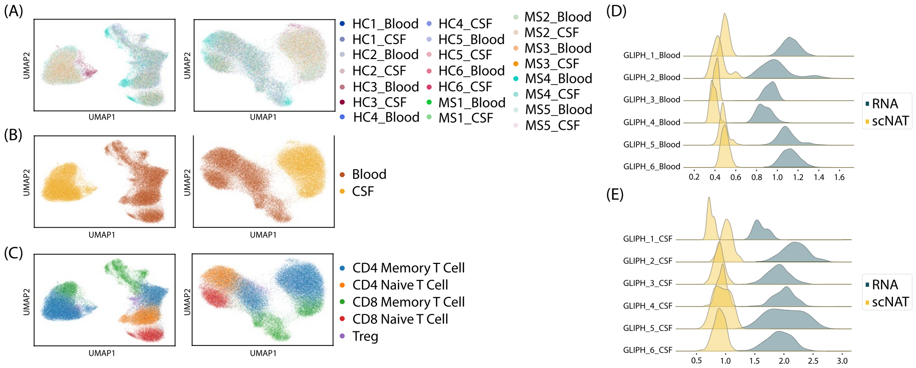

Fig. 5 Latent space and GLIPH2 comparison on the MS dataset. **A**–**C** ResPAN (RNA) UMAP (left) and scNAT UMAP (right) colored by batch (**A**), source (**B**), and cell type (**C**). **D**, **E** Pair-wise normalized Euclidean distance comparison of the GLIPH2 clusters between RNA and scNAT in blood (**D**) and CSF (**E**)

further validate this finding, we applied the same procedure using only the RNA data (Fig. S8C-G), and found that the RNA-only data could partially recover the pattern, with the cells having high expansion score being the CD8 memory T cells towards the high end of the UMAP for both blood and CSF (Fig. S8C-D), and with the cells still migrating from blood to CSF (Fig. S8E-F). However, the RNA data failed to bridge the two tissues, rather the cells that should be in transitioning state were pushed towards the boundaries of blood and CSF (Fig. S8G). Given the trajectory, we then studied which genes are associated with the trajectory to better understand the migration mechanism.

Table 1 Batch correction scores comparing scNAT and other methods

Method	1 − bASW	1 − kBET
scNAT	0.566	0.387
ResPAN	0.561	0.333
Seurat v4	0.550	0.267
MNN	0.561	0.263
BBKNN	0.557	0.224

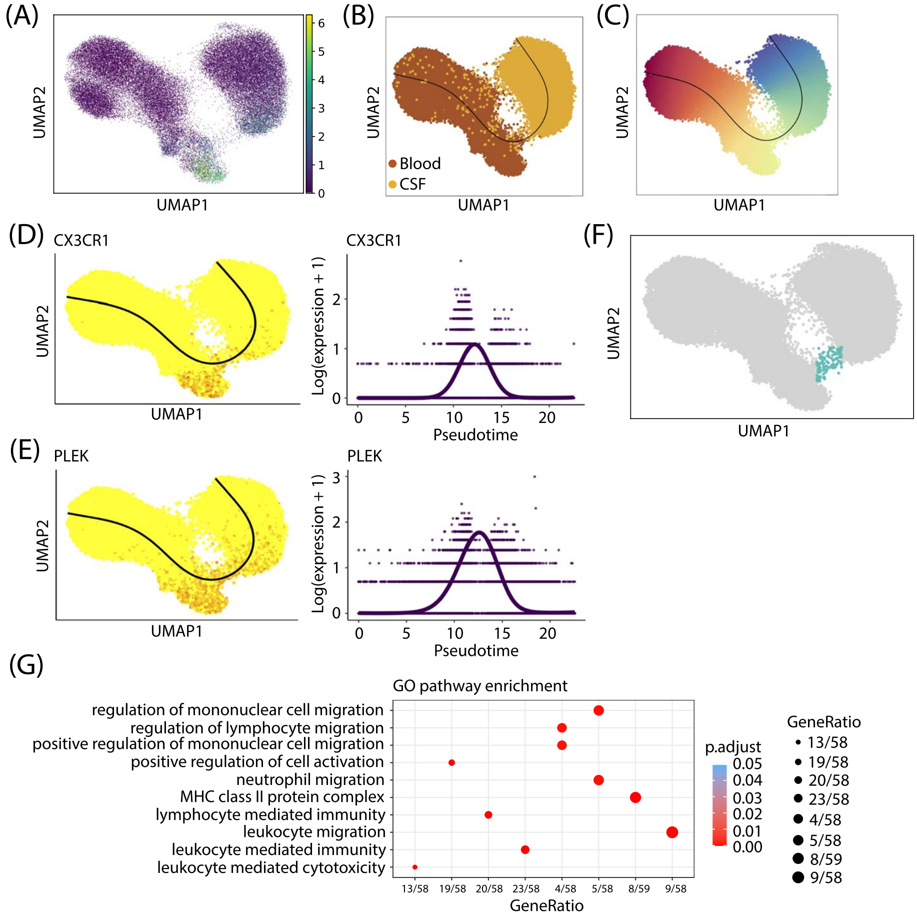

Fig. 6 A T cell trafficking trajectory from blood to CSF identified by scNAT. **A** scNAT UMAP colored by expansion score. **B**, **C** Slingshot inferred trajectory colored by tissue (**B**) and pseudotime (**C**). **D**, **E** Genes that are associated with the trajectory identified by TradeSeq in UMAP (left) and the corresponding smoothers (right). **F** A cluster of T cells that are in transitioning state. **G** Upregulated GO pathways of the transitioning cluster compared with other cells

We applied tradeSeq [43] to infer genes that were differentially expressed along the trajectory (Methods). Overall, we identified 50 DE genes (Figs. 6D-E, S7B-C), and the full gene list can be found in Additional file 2: Table S3. Among these genes, CX3CR1 has been reported to prevent remyelination in a cuprizone model of demyelination [44], underscoring its neuroprotective effects [45] (Fig. 6D). PLEK was identified as an MS risk gene from both RNA [46] and GWAS studies [47] (Fig. 6E). Next, we focused on the small cluster that was in the transitioning state (Fig. 6F). To further corroborate our hypothesis that these cells are migrating and clonally expanded T cells, we examined the distribution of the CDR3β chain sequence in this cluster as well as the distribution of the chain in random cells and found that the CDR3 in the transitioning cluster was more homogeneous than the one in random cells of the same size (Fig. S7D). Finally, the upregulated GO pathways of this cluster with respect to all the other cells (Additional files 3 and 4: Table S4-5) can be broadly divided into two groups, including cell migration and leukocyte cytotoxicity (Fig. 6G), which suggests that the cells were migrating from blood to CSF and were clonally expanded due to activation. Overall, this indicates that T cells in MS can transmigrate from blood to CSF and some of the transitioning cells are under clonal expansion, showing signatures of T cell cytotoxicity.

Discussion

Next-generation sequencing generates vast amounts of molecular data, especially in the field of genomics and transcriptomics, which requires advanced computational and analytical methods to extract meaningful insights and improve our understanding of biological processes. Deep learning has been applied extensively in this field to analyze and interpret these complex data. In this paper, we present scNAT, a deep learning-based model to integrate scRNA-seq and the paired scTCR-seq data for the purpose of novel cell population identification and biological process inference. To the best of our knowledge, our method represents the first approach capable of integrating the three complex data types, including numerical, text, and categorical data in order to perform this task. While previous methods including MOFA+, scAI, and LIGER are capable of integrating multi-omics data, they fail to handle data types other than numerical data. On the other hand, tessa ignores CDR3α as well as V/(D)/J gene information, and assumes all the cells within the same clone have the same RNA profile, which resulted in unstructured data (Fig. S5B–C). In contrast, DeepTCR combines the CDR3 sequences and V/(D)/J gene usage to characterize the TCR data, but it does not take into account RNA information which is a valuable data source for cell type identification and functional characterization of different cell types (Fig. S5D–E). Similarly, although the CoNGA algorithm computes correlations between RNA and TCR data, it does not render a joint representation of the two data modalities, hindering downstream visualization and analysis (Fig. S5F–I).

Through both gird search and real data analysis, we demonstrate that scNAT can balance the RNA and TCR modalities in a data-driven manner, achieved through loss function hyperparameter tuning. While we only explore the integration of the TCR and RNA data, this framework can be extended to include other data modalities such as single-cell BCR-seq [48] and single cell immune profiling with other modalities such as cell surface marker. By utilizing deep learning, it is possible to create a rich feature space that can incorporate multiple data types, providing greater flexibility in data modeling. We demonstrated the performance of scNAT through its application to an MS dataset. We showed that with leveraging information from the TCR data, the resulting UMAP can not only preserve the cell type information but can also incorporate signals from the clonotype data. The application suggests that apart from data integration, scNAT can further remove remaining batch effect from the input data which have been preprocessed by some batch correction method. Moreover, through integrating different data modalities, scNAT can also benefit other downstream analyses, such as trajectory inference.

We note a few limitations in our pipeline. The first is the huge variability in both TCR α and β chains. Due to the randomness in the V/(D)/J combination and the addition or deletion at the CDR3 junction site, it is difficult to perform systematic simulation to evaluate model performance under different scenarios. Furthermore, the black-box nature of the neural network hampers our ability to interpret the results. For example, although there are hyperparameters in the loss function controlling the weight of each modality, understanding which RNA or TCR contributes to the final result of each cell is still challenging.

Conclusions

We demonstrate that the use of deep learning to integrate scRNA-seq and scTCR-seq data allows for a more comprehensive analysis that leverages the strengths of both data types. By combining gene expression information from scRNA-seq with TCR repertoire diversity and clonality information from scTCR-seq, we can better understand the complex immune response and identify novel cell populations that may have been previously overlooked. The resulting integrated data can also be used for trajectory inference, which allows for the study of cellular differentiation and development over time. This approach is the first to handle numerical, text, and categorical data types all together in order to integrate RNA and TCR data, representing a significant advance in the field of single-cell analysis and has the potential to lead to new insights into the immune system and its role in health and disease.

Methods

scTCR-seq data curation

scTCR-seq data in the filtered_contig_annotations.csv files were collected from the two data sources as described in the manuscript. Only T cells with α and β chains were kept and we further chose cells with both RNA and paired TCR data.

scRNA-seq data preprocessing

scRNA-seq data were processed in accordance with the scanpy [49] (v1.7.2, RRID:SCR_018139) workflow for preprocessing. Next, ResPAN [10] (v0.1.0) was applied to remove batch effects. Cell type annotation was achieved via domain knowledge and canonical marker genes.

scNAT architecture

Data transformation

scNAT takes normalized, log-transformed scRNA-seq gene expression with the top 2000 highly variable genes as input. Further, to be more comparable with the dimension of the paired scTCR-seq data, the scRNA-seq input was transformed through a fully connected layer of length 1024 (Fig. S1A). Each V/J gene is considered as a categorical input to the model and was first converted to a one-hot embedding and then transformed into a continuous numeric space by a trainable embedding layer of dimension 48. Finally, all the gene features for both TCRα and TCRβ were concatenated to represent the joint feature of the V/J gene usage for each T cell (Fig. S1B). For the CDR3 sequence, similar to the V/J gene, the amino acid sequence was first transformed to a one-hot embedding and then to a continuous numeric vector. Subsequently, three CNN layers with the feature maps 32, 64, and 128 were added to extract sequences and allow all the sequences to share the same sets of NN parameters. Detailed structure can be found in Sidhom et al. [13]. Finally, the outputs for both TCRα and TCRβ were flattened and concatenated as the CDR3 representation (Fig. S1C). For the TCR data integration (Fig. S4A), only the V/J gene and CDR3 information were input to the model.

VAE structure

The core of the model is a VAE which is able to extract features from the scRNA-seq data, V/J gene usage, and the CDR3 sequences. The three data types were first concatenated to get an overall continuous representation (X) of the multi-omics data, and then a VAE model with layer sizes of 512, 256, and 512 was applied in order to learn the low dimensional distribution of the combined data by a multi-dimensional standard Gaussian distribution parametrization. Finally, the sampled data from the latent space were reconstructed through fully-connected and deconvolutional layers (Fig. 1).

Training VAE

To train the neural network, we implemented the Adam Optimizer with learning rate 0.001 in order to minimize both the variational loss and the reconstruction loss. The variational loss is the Kullback-Leibler (KL) divergence between the encoder's distribution and a unit Gaussian:

$$V_{loss} = D_{KL}(N(\mu(X), \sigma(X)||N(0,1))). \tag{1}$$

The reconstruction loss has three parts. The first component is the RNA reconstructed loss which is defined as the average of the mean squared error between the original RNA input (X_g) and the reconstructed RNA data ($\hat{X_g}$) across all G genes:

$$R_{RNA} = \frac{1}{G}\sum_{g}\left(X_g - \hat{X}_g\right)^2. \tag{2}$$

The second part is the V/J gene cross-entropy loss between the one-hot encoded input V/J gene usage (T_{orig}) and the reconstructed gene (T_{recon}) across all the gene positions (V and J) and the two chains (α and β):

$$R_{gene} = -\left(\sum_{m} T_{orig,m} log(T_{recon,m}) + \sum_{n} T_{orig,n} log(T_{recon,n})\right); m \in \{V_\alpha, J_\alpha\}, \\ n \in \{V_\beta, J_\beta\}. \tag{3}$$

The third piece is the reconstruction loss of the CDR3 sequence represented by the cross-entropy loss between the one-hot encoded CDR3 input (S_{orig}) and the reconstructed sequence (S_{recon}) across the two chains:

$$R_{CDR3} = -(S_{orig,\alpha} log(S_{recon,\alpha}) + S_{orig,\beta} log(S_{recon,\beta})). \tag{4}$$

The total loss is:

$$L = \alpha_{latent} V_{loss} + \alpha_{gene} R_{gene} + \alpha_{seq} R_{CDR3} + R_{RNA}, \tag{5}$$

where α_{latent}, α_{gene}, and α_{seq} are the weights for the variational loss, V/J gene loss, and CDR3 loss, which can be tuned to balance different data modalities as described in the next section. The training with raw RNA counts as input was stopped early after two epochs because there were extreme values in the RNA count matrix, causing infinite values of the weights.

Grid search for parameter tuning

In order to find the optimal combination of the three weights, we conducted grid search to maximize the sum of the normalized mean Silhouette Coefficient of all samples as well as the V/J gene and CDR3 prediction accuracy. We can define a cell type silhouette score for each cell i with $s_{celltype}^{(i)}$, where:

$$s_{celltype}^{(i)} = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{6}$$

here $a(i)$ is the mean distance from cell i to all other cells that belong to the same cell type, $b(i)$ is the lowest average distance from cell i to all cells in the same cell type among all other cell types.

The normalized Silhouette coefficient quantifies how well the integrated data preserve the cell type information and is defined as:

$$ASW = \frac{1}{2}\left(\frac{\sum_{i=1}^{N} s_{celltype}^{(i)}}{N} + 1\right), \tag{7}$$

where N is the total number of cells, and the normalization ensures the ASW is between 0 and 1.

The predicted V/J gene was set to be the one that corresponds to the index with the largest value of the reconstructed data. Similarly, each amino acid in the predicted CDR3 sequence was determined by the index with the largest value of the reconstructed CDR3 sequence in each position. Finally, the accuracy was computed by comparing the predicted data with the original data. The V/J gene and CDR3 prediction accuracy reflects the weight of the TCR modality.

The parameter grid was set to be $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$ for α_{latent}, α_{gene}, and α_{seq}, and during each run, we randomly subsampled 4000 cells to make it computationally efficient. In addition to maximizing the sum of the ASW and accuracy, we (1) put a constraint on ASW such that the resulting value is larger than 0.5 to ensure the performance is better than random, and (2) only selected the parameters whose corresponding loss does not exceed 115% of the minimal loss so that the model fits the data well. The final parameters are $\alpha_{latent} = 10^{-6}$, $\alpha_{gene} = 10^{-3}$, $\alpha_{seq} = 10^{-2}$ for the MS dataset.

Other methods details

Here, we present the full analysis of the comparisons to other methods referenced in the main text. DeepTCR (v2.1.27) was trained under the default parameters within the unsupervised learning setting. For tessa, due to its high computational cost, only 25% of the cells were used to perform the analysis. We first constructed TCR networks that took into account the association with the scRNA-seq data. After that, we calculated UMAP following the steps in the next section. CoNGA analysis was conducted by first computing TCRdist [50] kernel PCs. After that, the scRNA-seq data were reduced to a single cell per TCR clonotype. Finally, the UMAP was computed for RNA and TCR separately.

UMAP calculation

The UMAP from scRNA-seq data only was calculated following the standard scanpy [49] pipeline. For the scNAT integrated data, we first ran PCA based on the latent space representation, and then we further embedded the dimension reduced data into the UMAP space. By setting the model parameter *include_RNA* to *False*, the model only included the TCR data when doing the integration for the scTCR-seq data only version, and the UMAP was calculated the same way as the scNAT integrated data.

GLIPH2 analysis

In order to cluster TCRs based on shared similarity of the CDR3 sequences, GLIPH2 [33] was applied to the MS scTCR-seq data. GLIPH2 clusters TCRs based on global similarity, computed by CDR3 sequences differing by up to one amino acid, and motif-based local similarity, determined by shared enriched CDR3 motifs. The GLIPH2 human CD48 dataset was used for reference. The filtering criteria for GLIPH2 clusters were set to be groups with significant V-gene bias ($P < 0.005$ by GLIPH2) and significant final score ($P < 10^{-5}$ by GLIPH2). The normalized pair-wise distances within each cluster were calculated by dividing the original distance by the mean pair-wise distances across all cells in the RNA PCA or scNAT latent space.

Batch correction metrics

To evaluate the batch correction performance for scNAT and ResPAN, we considered two metrics: bASW (ASW [34] on batch labels) and kBET [35].

bASW

For each tissue, the bASW was computed on batch labels and scaled to ensure it is between 0 and 1 using the equation below:

$$bASW_{tissue} = \frac{ASW_{tissue} + 1}{2}, \tag{8}$$

then we weighted the bASW$_{tissue}$ score from blood and CSF based on their cell number n_{tissue}:

$$bASW = bASW_{blood}\frac{n_{blood}}{n_{blood} + n_{CSF}} + bASW_{CSF}\frac{n_{CSF}}{n_{blood} + n_{CSF}}. \tag{9}$$

Finally, to make sure higher scores correspond to better batch removal performance, the score is scaled by subtracting it from 1:

$$1 - bASW. \tag{10}$$

kBET

To identify if there is any bias in a replicated experiment, kBET determines whether the label distribution of a k nearest neighborhood of a cell is similar to the global label distribution by a χ^2-based test. The test is repeated for random neighborhoods with a fixed size, and the results are summarized by averaging the binary test results to render an overall rejection rate. Since a lower rejection rate indicates well-mixed replicates, similar to the bASW, we use 1−kBET to ensure higher score correspond to better mixing performance.

Trajectory inference

In order to infer the cell trafficking trajectory and underlying pseudotime in the MS dataset, we applied the R (v4.3.2, RRID:SCR 001905) package Slingshot [40] (v3.18, RRID:SCR 017012) and specified the starting cluster and the endpoint. In addition, given the Slingshot trajectory, the method tradeSeq [43] was further implemented to fit a regression model which is able to identify genes that are differentially expressed along this lineage through the R function associationTest. The threshold for DE genes was set to be the Bonferroni corrected p value < 0.05 and the mean log fold change > 3. For Monocle3 [41, 42] (v3.14, RRID:SCR 018685), a principal graph from the reduced dimension space was first learned from reversed graph embedding, and then the cells were assigned a pseudotime based on their projection on the principal graph.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03129-y.

Additional file 1.

Additional file 2: Table S3. Differentially expressed genes identified by tradeSeq in MS dataset.

Additional file 3: Table S4. Differentially expressed genes in transitioning cells in MS dataset.

Additional file 4: Table S5. Gene Ontology canonical pathways in transitioning cells in MS dataset.

Additional file 5. Review history.

Acknowledgements

We thank the colleagues from Zhao Lab for discussion and suggestions.

Review history

The review history is available as Additional file 5.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Methodology, B.Z., Y.W.; software, B.Z.; exploration, L.T.K.; writing-original draft, B.Z.; writing-review and editing, Y.W., D.D., L.Z., D.H., H.Z.; supervision, H.Z. All authors read and approved the final manuscript.

Funding
The study is funded by the joint efforts of the Michael J. Fox Foundation for Parkinson's Research (MJFF) and the Aligning Science Across Parkinson's (ASAP) initiative. MJFF administers the grant [ASAP-000529] on behalf of ASAP and itself. L.Z. receives research funding from NIA R56-AG074015 and NIDA DP2-DA056169. H.Z. receives research funding from NIA R56-AG074015.

Availability of data and materials
Data for all CSF scRNA-seq and single-cell TCR sequencing from the MS study is available through NCBI's dbGaP at accession number phs002222.v1.p1 [51]. Python (v3.6.15, RRID:SCR 008394) source code and scNAT is uploaded on https://github.com/biqing-zhu/scNAT [52] - DOI: 10.5281/zenodo.8341925 [53]. The repository is released under the MIT license.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA. [2]Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06511, USA. [3]Department of Internal Medicine, Yale School of Medicine, New Haven, CT 06511, USA. [4]Department of Computer Science, Yale University, New Haven, CT 06511, USA. [5]Department of Neuroscience, School of Medicine, Yale University, New Haven, CT 06511, USA. [6]Department of Neurology, School of Medicine, Yale University, New Haven, CT 06511, USA. [7]Department of Immunobiology, School of Medicine, Yale University, New Haven, CT 06511, USA. [8]Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, USA, MD 20815.

Received: 30 May 2023 Accepted: 27 November 2023
Published online: 18 December 2023

References

1. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):1–14.
2. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):1–12.
3. Lusch B, Kutz JN, Brunton SL. Deep learning for universal linear embeddings of nonlinear dynamics. Nat Commun. 2018;9(1):4950.
4. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. Proc IEEE. 2020;109(1):43–76.
5. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1(1):18.
6. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021;12(1):1–11.
7. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, et al. Data denoising with transfer learning in single-cell transcriptomics. Nat Methods. 2019;16(9):875–8.
8. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8.
9. Amodio M, Van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods. 2019;16(11):1139–45.
10. Wang Y, Liu T, Zhao H. ResPAN: a powerful batch correction model for scRNA-seq data through residual adversarial networks. Bioinformatics. 2022;38(16):3942–9.
11. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018;9(1):1–13.
12. Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics. 2020;36(16):4415–22.
13. Sidhom JW, Larman HB, Pardoll DM, Baras AS. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. Nat Commun. 2021;12(1):1–12.
14. Wang Y, Zhao H. Non-linear archetypal analysis of single-cell RNA-seq data by deep autoencoders. PLOS Comput Biol. 2022;18(4):e1010025.
15. Lin Y, Wu TY, Wan S, Yang JY, Wong WH, Wang YR. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40(5):703–10.
16. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, Yosef N. Joint probabilistic modeling of single-cell multiomic data with totalvi. Nature methods. 2021;18(3):272–82.

17. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10(1):1–13.
18. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat Biotechnol. 2014;32(7):684–92.
19. Zhao Y, Niu C, Cui J. Gamma-delta ($\gamma\delta$) T cells: friend or foe in cancer development? J Transl Med. 2018;16(1):1–13.
20. Pasetto A, Lu YC. Single-cell TCR and transcriptome analysis: An indispensable tool for studying T-cell biology and cancer immunotherapy. Front Immunol. 2021;12:689091.
21. De Simone M, Rossetti G, Pagani M. Single cell T cell receptor sequencing: techniques and future challenges. Front Immunol. 2018;9:1638.
22. Zemmour D, Zilionis R, Kiner E, Klein AM, Mathis D, Benoist C. Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. Nat Immunol. 2018;19(3):291–301.
23. Pappalardo JL, Zhang L, Pecsok MK, Perlman K, Zografou C, Raddassi K, et al. Transcriptomic and clonal characterization of T cells in the human central nervous system. Sci Immunol. 2020;5(51):eabb8786.
24. Gate D, Saligrama N, Leventhal O, Yang AC, Unger MS, Middeldorp J, et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. Nature. 2020;577(7790):399–404.
25. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):1–17.
26. Jin S, Zhang L, Nie Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol. 2020;21:1–19.
27. Ashuach T, Gabitto MI, Koodli RV, Saldi GA, Jordan MI, Yosef N. MultiVI: deep generative model for the integration of multimodal data. Nat Methods. 2023;20:1–10.
28. Schattgen SA, Guion K, Crawford JC, Souquette A, Barrio AM, Stubbington MJ, et al. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). Nat Biotechnol. 2022;40(1):54–63.
29. Zhang Z, Xiong D, Wang X, Liu H, Wang T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat Methods. 2021;18(1):92–9.
30. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell. 2018;174(5):1293–308.
31. Høglund RA, Maghazachi AA. Multiple sclerosis and the role of immune cells. World J Exp Med. 2014;4(3):27.
32. Van Langelaar J, Rijvers L, Smolders J, Van Luijn MM. B and T cells driving multiple sclerosis: identity, mechanisms and potential triggers. Front Immunol. 2020;11:760.
33. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol. 2020;38(10):1194–202.
34. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
35. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9.
36. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
37. Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7.
38. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36(3):964–5.
39. Strazielle N, Creidy R, Malcus C, Boucraut J, Ghersi-Egea JF. T-lymphocytes traffic into the brain across the blood-CSF barrier: evidence using a reconstituted choroid plexus epithelium. PLoS ONE. 2016;11(3):e0150945.
40. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19:1–16.
41. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979–82.
42. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14(3):309–15.
43. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun. 2020;11(1):1201.
44. Lampron A, Larochelle A, Laflamme N, Préfontaine P, Plante MM, Sánchez MG, et al. Inefficient clearance of myelin debris by microglia impairs remyelinating processes. J Exp Med. 2015;212(4):481–95.
45. Mai W, Liu X, Wang J, Zheng J, Wang X, Zhou W. Protective effects of CX3CR1 on autoimmune inflammation in a chronic EAE model for MS through modulation of antigen-presenting cell-related molecular MHC-II and its regulators. Neurol Sci. 2019;40:779–91.
46. Hoppmann N, Graetz C, Paterka M, Poisa-Beiro L, Larochelle C, Hasan M, et al. New candidates for CD4 T cell pathogenicity in experimental neuroinflammation and multiple sclerosis. Brain. 2015;138(4):902–17.
47. James T, Lindén M, Morikawa H, Fernandes SJ, Ruhrmann S, Huss M, et al. Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients. Hum Mol Genet. 2018;27(5):912–28.
48. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from single cells. Bioinformatics. 2017;33(3):425–7.
49. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):1–5.
50. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017;547(7661):89–93.
51. Pappalardo JL, Zhang L, Pecsok MK, Perlman K, Zografou C, Raddassi K, et al. Transcriptomic and Clonal Characterization of T Cells in the Human Central Nervous System. Datasets. Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002222.v1.p1. Accessed 22 Oct 2020.

52. Zhu B. scNAT: A deep learning method for integrating paired single cell RNA and T cell receptor sequencing profiles. GitHub. 2023. https://doi.org/10.5281/zenodo.8341925.
53. biqing-zhu. biqing-zhu/scNAT: v0. Zendo. 2023. https://doi.org/10.5281/zenodo.8341925.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.