BMC Bioinformatics

RESEARCH

Open Access

Orthogonal multimodality integration and clustering in single-cell data

Yufang Liu[1†], Yongkai Chen[1†], Haoran Lu[1], Wenxuan Zhong[1], Guo-Cheng Yuan[2*] and Ping Ma[1*]

[†]Y. Liu, and Y. Chen joint first authors.

*Correspondence: guo-cheng.yuan@mssm.edu; pingma@uga.edu

[1] Department of Statistics, University of Georgia, Athens, GA 30602, USA
[2] Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Abstract

Multimodal integration combines information from different sources or modalities to gain a more comprehensive understanding of a phenomenon. The challenges in multi-omics data analysis lie in the complexity, high dimensionality, and heterogeneity of the data, which demands sophisticated computational tools and visualization methods for proper interpretation and visualization of multi-omics data. In this paper, we propose a novel method, termed Orthogonal Multimodality Integration and Clustering (OMIC), for analyzing CITE-seq. Our approach enables researchers to integrate multiple sources of information while accounting for the dependence among them. We demonstrate the effectiveness of our approach using CITE-seq data sets for cell clustering. Our results show that our approach outperforms existing methods in terms of accuracy, computational efficiency, and interpretability. We conclude that our proposed OMIC method provides a powerful tool for multimodal data analysis that greatly improves the feasibility and reliability of integrated data.

Keywords: Multimodality integration, CITE-seq, Cell clustering

Introduction

Recent advances in single-cell multi-omics have opened up new avenues for delving into the intricacies of cellular diversity and gene expression at the individual cell level [1, 2]. One of the pioneering techniques in this field is Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq), which has emerged as a groundbreaking technology [3, 4]. CITE-seq combines simultaneous measurements of single-cell RNA sequencing (scRNA-seq) [1, 5] with cell surface protein markers detected by antibody-derived tags (ADTs) [6], providing a comprehensive multimodal snapshot of cellular identity and function [7]. Nevertheless, it is challenging to effectively harness and combine data from RNA and cell surface protein marker expression levels. This challenge becomes particularly daunting when dealing with large volumes and high dimensional datasets [8].

To tackle this issue, several methods have been proposed, including weighted nearest neighbor (WNN) [9], multi-omics factor analysis plus (MOFA+) [10] and totalVI [11]. WNN performs clustering analysis by generating the nearest neighbor graph (NNG) [12] for each modality and then constructing a weighted graph that combines these NNGs with weighted connections. As a result, each data point would be assigned

to a cluster based on the weighted contributions of its neighbors. However, the weight associated with each modality is cell-specific, and its value cannot be easily interpreted. In contrast, MOFA+ is a factor analysis model to estimate common factors that capture shared variability among different omics layers. These identified factors are used in downstream analyses, such as feature selection and clustering. Nonetheless, extracting the meaningful factors presents a challenging task, requiring careful consideration to draw valid conclusions from the model. TotalVI processes gene and protein UMI counts as input, establishing the variational autoencoder (VAE) to obtain the latent variables; it then leverages the resulting latent variables for integration, clustering, and visualization purposes. Still, machine learning methods encounter challenges such as elevated computational expenses, the need for parameter tuning, and the interpretation of resulting variables. Consequently, these approaches share a common limitation in terms of interpretability, hindering the extraction of meaningful insights, including the identification of critical predictive features. This limitation leaves two fundamental biological questions inadequately addressed: First, compared to RNA, do ADTs provide an additional significant prediction power in predicting cell type? If so, which ADTs are most needed? Can we quantify this additional prediction power? Second, in each cell cluster and type, which RNAs and ADTs are differentially expressed to provide significant prediction power? In addition to the lack of interoperability, the computational burden of methods such as WNN, MOFA+ and totalVI becomes prohibitively high when analyzing datasets with numerous cells and a large number of features [13].

To address these limitations, we introduce a novel approach called Orthogonal Multimodality Integration and Clustering (OMIC) for the analysis of single-cell multi-omics data. Our method excels at modeling the relationships among multiple variables, facilitating scalable computation, and preserving accuracy in cell clustering compared to existing methods. Most importantly, our approach provides quantitative insights into the contributions of individual features in clustering analysis. To underscore the effectiveness of OMIC methods, we present comprehensive comparisons with the several benchmark methods: WNN, MOFA+, TotalVI, CiteFuse [14] and BREM-SC [15] on the cord blood mononuclear cell (CBMCs) and human bone marrow cell (HBMCs) datasets. Moreover, we perform an additional analysis of OMIC method on the human peripheral blood mononuclear cells (PBMCs) dataset, showing that our method is capable of integrating multiple datasets from multiple batches. To further assess the efficacy of out method in the context of transcriptomic profiling across spatial regions, we perform data integration and clustering utilizing the OMIC approach on a Spatial CITE-seq dataset [16].

Results

Overview of OMIC method

Figure 1 illustrates the OMIC method. To efficiently leverage and combine information from RNA and ADT expression levels, the OMIC method decomposes the ADT expression level into two parts by projecting the ADT expression onto the RNA space, resulting in a decomposition into two orthogonal components, ADT prediction and ADT residual. The predicted ADT represents the portion of the data that can be explained by RNA, while the ADT residual comprises the unexplained portion not attributed to RNA.

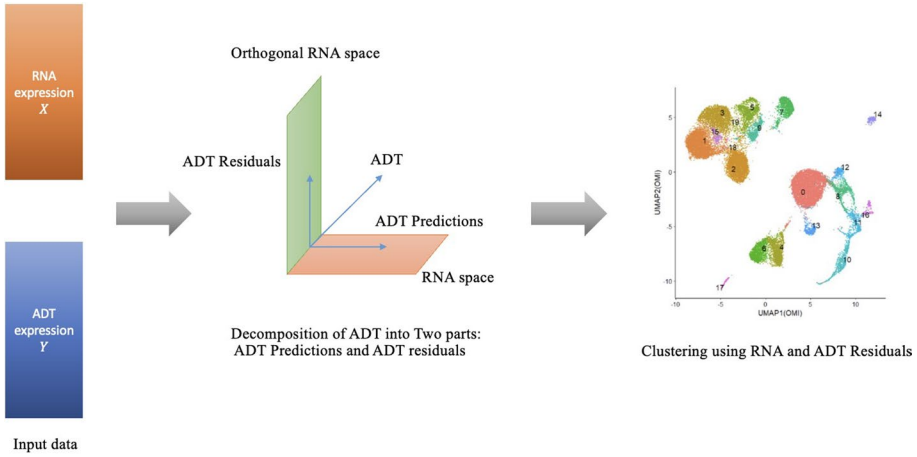

Fig. 1 Outline of the OMIC method. The input is RNA and ADT expression. After performing ADT projections on RNA and Orthogonal RNA space, clustering analysis is conducted based on RNA and ADT residuals. The OMIC method has the capability to identify differentially expressed RNAs and ADTs, thereby offering substantial predictive power of cell types

Consequently, our objective is to integrate the unexplained ADT residual with RNA for the purpose of cell clustering. This methodology eliminates any redundant information between RNA and ADT, thus enhancing precision and efficiency in the clustering process. More importantly, through an examination of how well RNA explains variation in ADT, along with an analysis of the coefficients in the resulting model, we can identify which RNAs and ADTs are differentially expressed, thereby contributing significantly to predictive power.

OMIC method on CITE-seq datasets

Analysis of cord blood mononuclear cells (CBMCs) dataset

We test the performance of the OMIC method on cord blood mononuclear cells (CBMCs) CITE-seq dataset [17]. This dataset contains 8,617 cells. For each cell, 13 cell-surface protein markers are quantified via sequencing their corresponding antibody-derived tags (ADTs), and 20,501 RNA expression levels are measured. There are 15 true cell types in the dataset. We compare RNA only, ADT only, WNN, MOFA+ and totalVI with the OMIC method, each yields 21, 19, 14, 13, 13 and 14 clusters, respectively.

To evaluate the clustering results, we computed the Adjusted Rand Index (ARI) [18], measuring the similarity between true cell type annotations and predicted clusters for each method. An ARI value closer to 1 indicates greater consistency between the clustering results and the ground truth cell type annotations. Figure 2A shows that when leveraging the information of RNA alone, it is challenging to separate the CD14+Monocytes (CD14+Mono) and T/Mono doublets cell groups effectively (ARI = 0.69). Furthermore, Fig. 2B illustrates that using ADT information alone was more problematic, with mouse and human erythroid, DC, and Mk cell groups mixed together. In contrast, by using OMIC (ARI = 0.72, Fig. 2F) to integrate RNA and ADT information, we were able to accurately distinguish between Memory CD4 T and Naive CD4 T groups while effectively separating CD14+Mono and T/Mono doublets cell groups.

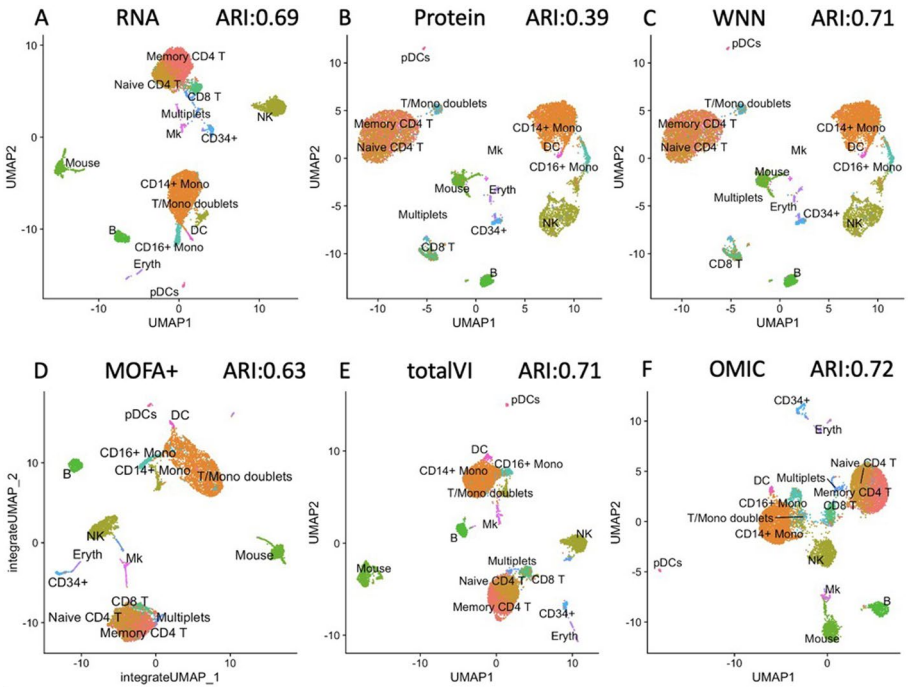

Fig. 2 UMAP visualization of different methods on CBMCs dataset (**A** RNA alone; **B** ADT alone; **C** WNN; **D** MOFA+; **E** totalVI, **F** OMIC)

For comparison, we also applied WNN, MOFA+, and totalVI to analyze this dataset. While WNN (ARI = 0.71) and MOFA+ (ARI = 0.63) methods can also distinguish CD14+Mono and T/Mono doublets, both methods merge Naíve CD4 T and Memory CD4 T cells into a single cluster (Fig. 2C, D). Of note, OMIC does not have this artifact. totalVI has the similar performance (Fig. 2E, ARI = 0.71) to OMIC method.

Furthermore, we conduct analysis with CiteFuse (ARI = 0.63) and BREM-SC (ARI = 0.61) on CBMCs dataset. Our method demonstrates superior performance compared to other methods in the analysis of the CBMCs dataset.

Analysis of human bone marrow cells (HBMCs) dataset

We further analyzed the human bone marrow cells CITE-seq dataset, comprising 30,672 cells [1]. 25 ADTs and 17, 009 genes are profiled for each cell.

It is worth noting that the RNA analysis is more informative than the ADT analysis in identifying progenitor states (the ADT panel contains markers for differentiated cells), while the converse is true of T cell states (where the ADT analysis outperforms RNA) [9]. Thus, integrated information is necessary for cell clustering. We have conducted four analyses using the integrated data of RNA and ADT. There are 27 true cell types in the dataset. We compare WNN, MOFA+, totalVI with OMIC method, each yields 27, 12, 15, and 20 clusters respectively. Of note, CiteFuse and BREM-SC are not feasible for application on this dataset due to the computational constraints of their methods. Our OMIC approach effectively discriminates several significant cell groups, including Naive B cells, Memory B cells, plasmablast cells, and pDC cells, as depicted in Fig. 3D. Notably, our OMIC method performs well with an ARI of 0.89 for this dataset, surpassing

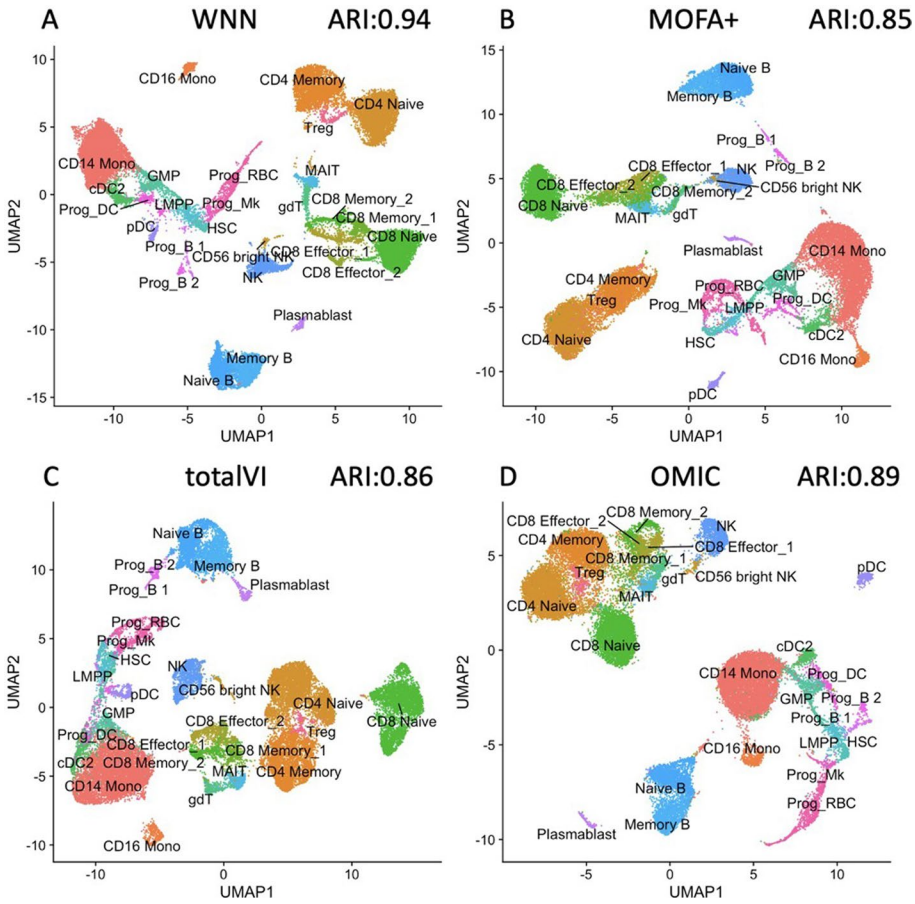

Fig. 3 UMAP visualization of different methods on HBMCs dataset (**A** WNN method; **B** MOFA+ method; **C** totalVI method; **D** OMIC method)

MOFA+ (ARI = 0.85, Fig. 3B) and totalVI (ARI = 0.86, Fig. 3C) but slightly trailing behind WNN (ARI = 0.94, Fig. 3A).

Computational cost analysis

To assess the computational efficiency of OMIC, WNN, MOFA+, totalVI, CiteFuse and BREM-SC, we conducted experiments on the same computer system featuring an Intel Core i7-12700 CPU running at 2.10GHz and 32GB of DDR4 RAM. Our findings indicate that the OMIC method offers the most efficient computational performance for analyzing both CITE-seq datasets.

Specifically, in the case of the HBMCs dataset, the OMIC method completed its computations in a mere 34.99 s, whereas the WNN method, MOFA+, and totalVI method required substantially more time, which are 119.98 s, 378.20 s, and 1247.69 s respectively (Table 1). It is worth noting that neither CiteFuse nor BREM-SC did not work in the HBMCs dataset since CiteFuse method requires at least $O(n^3)$ computational complexity in fusing the similarity matrix whereas the BREM-SC method uses iterative approach such as EM algorithm to solve parameters in the joint likelihood function which is not able to deal with dataset with large number of observations.

Table 1 Time cost of OMIC, WNN, MOFA+, totalVI, CiteFuse, BREM-SC methods, in seconds

Methods	CBMCs dataset	HBMCs dataset
OMIC	**24.61**	**34.99**
WNN	32.76	119.98
MOFA+	303.88	378.20
TotalVI	576.66	1247.69
CiteFuse	> 1 hour	
BREM-SC	> 1 hour	

The minimum time cost is highlighted in bold for each dataset

Table 2 AUCs of identifying true cell groups (CD4 Naíve, CD4 Memory, CD8 Naíve, Treg) using four kinds of information (only RNA information, RNA information and CD4 protein information, RNA information and CD25 protein information, RNA information and CD45RO protein information)

	RNA	RNA+CD4	RNA+CD25	RNA+CD45RO
CD4 Naive	0.684	**0.947**	0.700	0.746
CD4 Memory	0.735	0.767	0.813	**0.899**
CD8 Naive	0.782	**0.981**	0.797	0.813
Treg	0.632	0.752	**0.938**	0.653

The highest AUC is highlighted in bold for each cell type identification

In summary, we conclude that the proposed OMIC method effectively captures valuable biological information from the dataset while demanding significantly less computation time than other methods.

Interpretability

One of the notable strengths of our OMIC model lies in its ability to facilitate straightforward interpretation. Specifically, we can examine how well RNA explains the variance in ADT [19]. This explained variance value serves two key functions: first, it measures how well the model fits the data, with a higher value indicating better fitting. Second, it reflects the level of redundancy between RNA and ADT information, with a high value indicating a large area of overlap. Consequently, it underscores the significance of proteins with lower values, as they contain additional information beyond RNA for cell clustering.

In the HBMCs dataset, we selected three ADTs examples (CD25, CD45RO, and CD4) with relatively low values of the explained variance (0.20, 0.41, and 0.52, respectively) compared to the rest of the other ADTs. We next explored the notable significance of these ADTs in enhancing the clustering outcomes of corresponding four cell groups (CD4 Naive, CD4 Memory, CD8 Naive, and Treg).

For comparison, clustering was conducted using RNA information alone (Fig. 4A). However, this approach led to imperfect clustering (ARI = 0.46) as CD4 and CD8 cells were combined. To address this, we assigned a label to each cell group based on the predominant cell specificity within that group. By doing so, we could determine the accuracy of identifying specific cell types by comparing our assigned label to the ground truth of the cell annotations. In Table 2, we report the AUCs [20] of identifying three cell groups (CD4 Naíve, CD4 Memory, CD8 Naíve, Treg). We find that using RNA

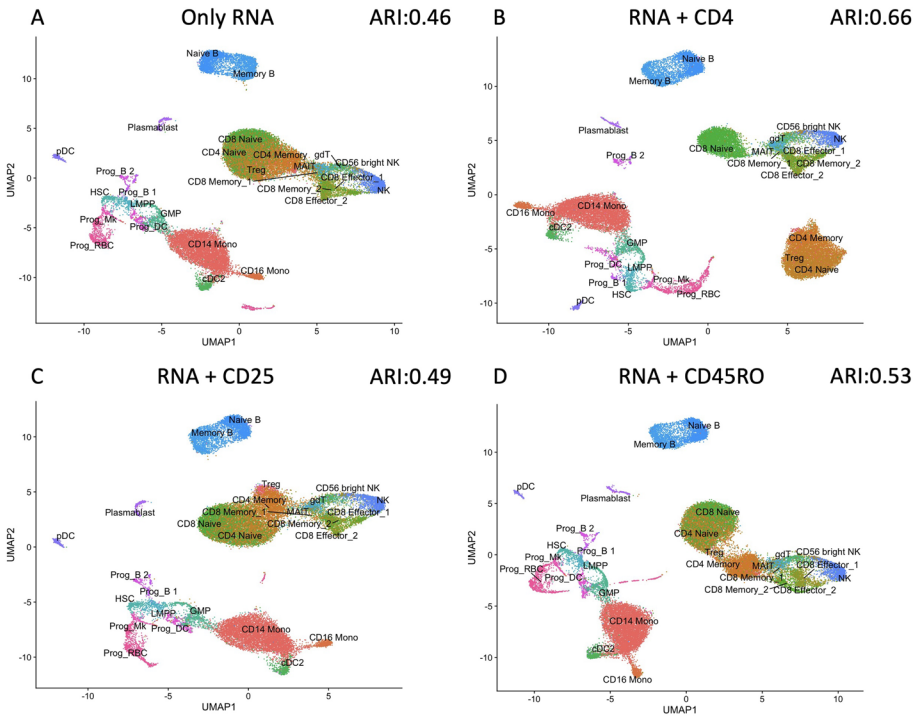

Fig. 4 UMAP visualization of clustering using different information on HBMCs dataset (**A** RNA alone; **B** RNA+CD4; **C** RNA+CD25; **D** RNA+CD45RO)

information alone would cause low AUC values for CD4 Naíve, CD4 Memory, CD8 Naíve, and Treg cell groups, which were 0.684, 0.735, 0.782, and 0.632, respectively.

In contrast, when we added CD4 ADT in the clustering procedure (Fig. 4B), not only did it lead to a complete separation of CD4 and CD8 cells, but it also enabled the identification of subgroups such as CD4 Naíve and CD4 Memory cells (ARI = 0.66). Furthermore, the AUC values for CD4 Naíve, CD4 Memory, and CD8 Naíve cell groups improved to 0.947, 0.767, and 0.981, respectively. Moreover, the accuracy for identifying other cell groups remained largely unchanged. These findings underscore the critical role of CD4 ADT in distinguishing CD4 and CD8 Naíve cells, aligning with existing literature [21].

Moreover, adding CD25 ADT alone in the clustering procedure allowed the detection of Treg group cells. Using RNA alone, the ARI value is only 0.46, but adding CD25 ADT increases the AUC to 0.938 (Fig. 4C). This result is consistent with the CD25 protein serving as a Treg group cell marker [22].

Finally, adding CD45RO ADT information along with all the RNA information in the clustering procedure resulted in better performance than only RNA (Fig. 4D, ARI = 0.53). Combining the results in Fig. 4A–D, we found an interesting fact that CD45RO essentially functions as the primary cell marker for CD4 Memory cells, since the other three pieces of information couldn't distinguish CD4 Memory cell groups as effectively [23]. The AUC for CD4 Memory cell group identification increased to 0.899.

Given the favorable outcomes of OMIC in the clustering analysis, we performed logistic regression independently for each cluster, examining three distinct scenarios within each cluster: one with RNA as the predictor, another with ADT as the

predictor, and a third incorporating the integrated RNA and ADT data as predictors. For RNA, we conducted a Wilcoxon Rank Sum test [24] for each cluster to include the differentially expressed genes (p values ≤ 0.01) in the logistic regression model. We performed a random split of the entire cell dataset into two subsets: one for training (70%) and the other for testing (30%). We repeat this process 100 times for each scenario.

Our focus was directed toward five specific clusters: CD4 Memory, CD4 Naíve, Memory B, Naíve B, and Treg, with the objective of evaluating the contributions of RNA and ADT information. In Memory B and Naíve B clusters, the use of integrated RNA and ADT information as predictors yielded higher AUC compared to using only RNA or ADT information. However, in the Memory B, Naíve B, and Treg cell clusters, the integrated information remained either unchanged or slightly lower than when using only ADT information (Fig. 5). Additionally, when we examined the coefficients in each logistic regression within these five clusters, we discovered that nearly no RNAs were statistically significant for identifying Treg, CD4 Memory, and CD4 Naíve (Fig. 6). This explained why incorporating RNA information did not significantly alter AUC as depicted in Fig. 5. However, in the case of identifying cluster Memory B and Naíve B, the relevance of RNA information became evident upon examining their coefficient values (Fig. 7) [25].

Furthermore, our analysis revealed that for CD4 Memory, ADT CD4 and CD45RO displayed larger positive coefficients, suggesting their significance as cell markers. Similarly, in the CD4 Naíve cluster, ADT CD4 and CD45RA emerged as important cell markers [26]. In the Treg cell group, CD25 exhibited a large positive value, while the coefficient of CD127-IL7Ra is negative, underscoring their utilities in detecting this particular cluster [27].

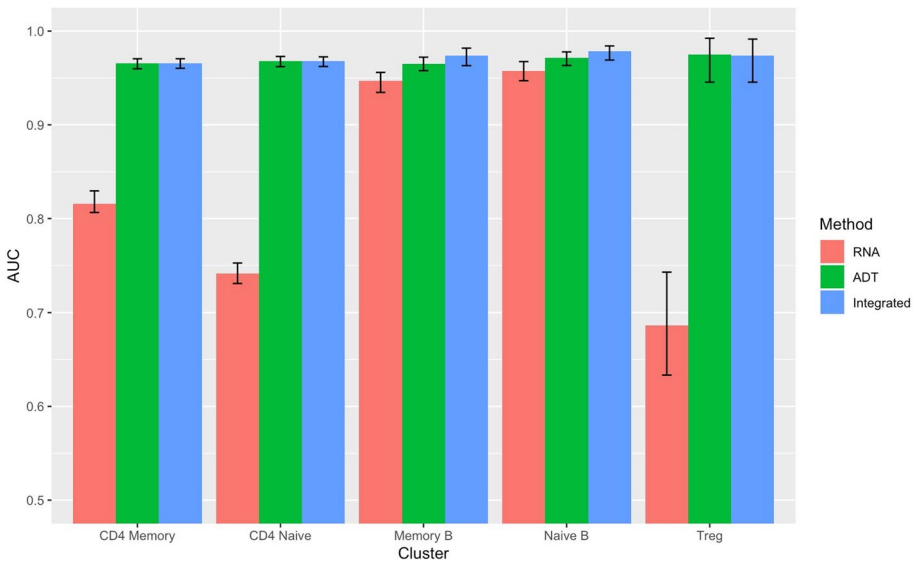

Fig. 5 AUC of classification in the testing set of five clusters under three scenarios: Only RNA, only ADT, and integrated RNA and ADT information as predictors

Fig. 6 Coefficients of logistic regression in the training set of CD4 Memory, CD4 Naíve and Treg using integrated RNA and ADT information as predictors. The size of each dot on the plot corresponds to the absolute value of its respective coefficient, while the color of the dot indicates the sign (positive or negative) of the coefficient

Analysis of the multi-batch CITE-seq data

In this section, we demonstrate the effectiveness of the OMIC method in simultaneously conducting data integration and batch effect correction across multiple batches of CITE-seq data. Our ana lysis focuses on human peripheral blood mononuclear cells (PBMCs), which is a Cite-seq dataset comprising 161,761 cells and measured with 228 antibodies [9]. These samples originate from a cohort of eight volunteers aged between 20 and 49 years participating in an HIV vaccine trial [28, 29]. Treating each of the eight volunteers as individual batches, we conducted batch effect correction and simultaneous integration of RNA and ADT. Without applying batch correction, it becomes evident that the batch effect significantly influences the integration of RNA and ADT data, as well as the clustering process (Fig. 8A). This is evident from the partitioning of several clusters, each associated with different batches. After performing the batch correction, we observed that the cells in different batches are mixed together (Fig. 8C), which implies that the influence of batch effect in clustering has been reduced. Moreover, Fig. 8D shows that several significant cell groups are detected by the OMIC method, including $CD4^+$ T cells, $CD8^+$ T cells, B cells, plasmablast cells, NK cells and so on.

Analysis of the spatial CITE-seq data

With the rapid advancement of spatial omics technologies [16, 30], there arises great interest in validating the effectiveness of the OMIC method when transcriptomics are profiled across spatial regions, particularly in the context of conducting clustering of

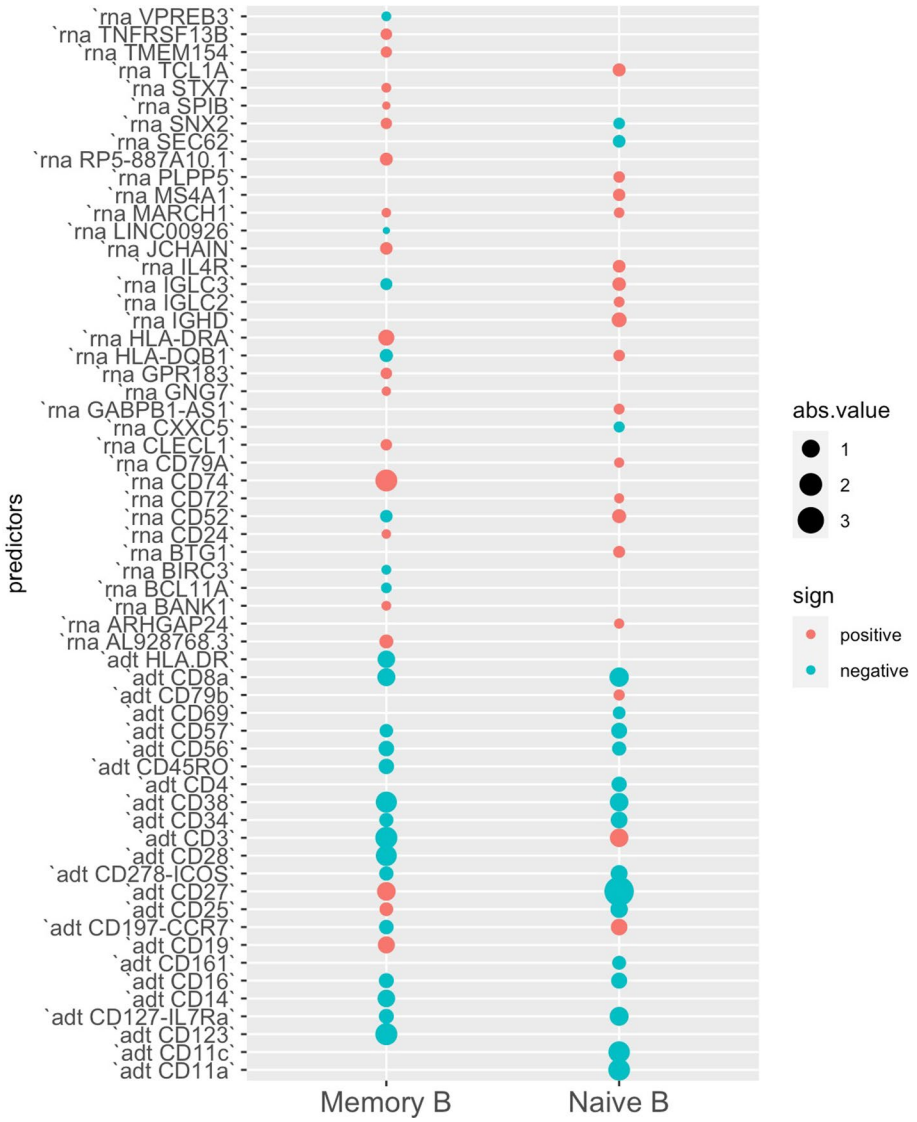

Fig. 7 Coefficients of logistic regression in the training set of Memory B and Naíve B using integrated RNA and ADT information as predictors. The size of each dot on the plot corresponds to the absolute value of its respective coefficient, while the color of the dot indicates the sign (positive or negative) of the coefficient

the spatial regions. To address this problem, we use the OMIC method in conducting data integration and clustering on a Spatial CITE-seq dataset [16]. This dataset comprises profiles of 2, 492 spots on a human tonsil sample. The abundance of 28, 417 genes and 283 ADTs are measured.

In Fig. 9, we provide the clustering results by using RNA alone (Fig. 9A), ADT alone (Fig. 9B), and the integration of RNA and ADT through the OMIC method (Fig. 9C). Clustering using the RNA profiles alone identified seven clusters while clustering using the ADT profiles alone identified five clusters. However, many clusters are mixed together in these two clustering results. By using the OMIC method, we can observe that there are seven resulting clusters,and most of them are well separated.

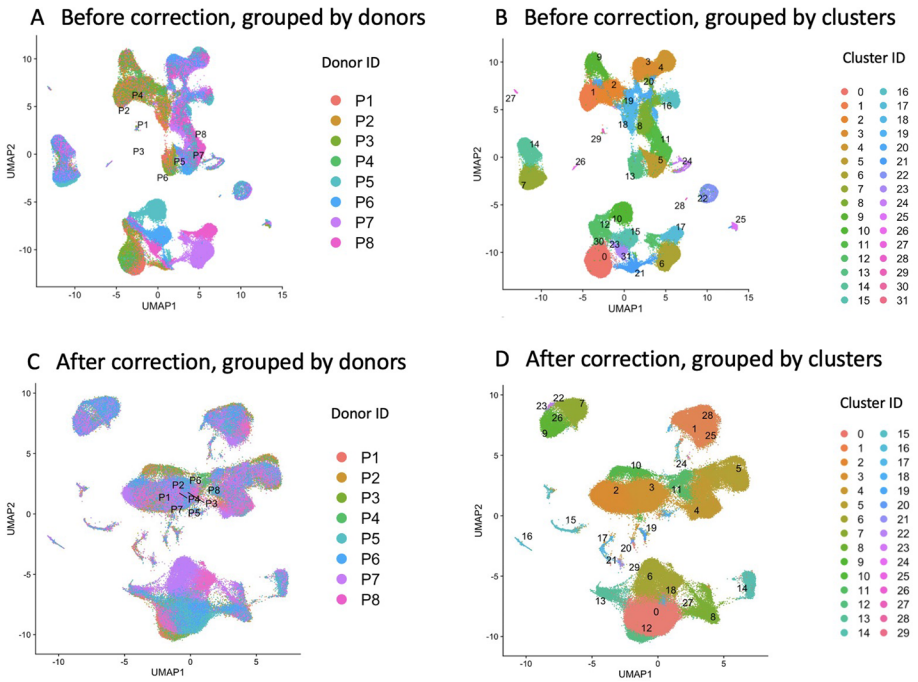

Fig. 8 Batch correction results in PBMCs dataset (**A** Integration of RNA and ADT before batch correction, grouped by donors (batches); **B** Integration of RNA and ADT before batch correction; **C** Integration of RNA and ADT after batch correction, grouped by donors; **D** Integration of RNA and ADT after batch correction)

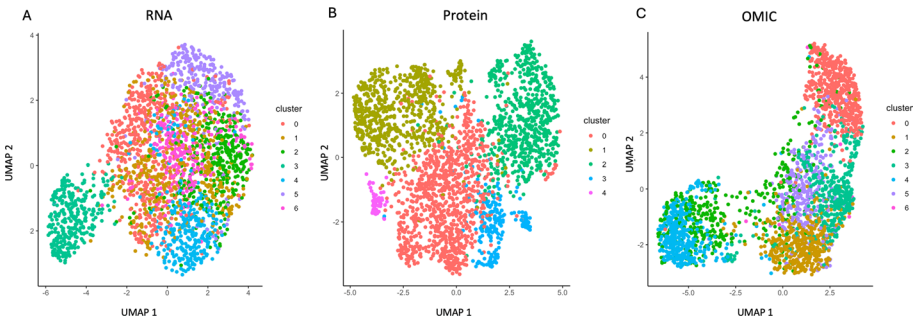

Fig. 9 Clustering results in the spatial CITE-seq dataset (**A** RNA alone; **B** ADT alone; **C** OMIC)

Discussion

The proposed Orthogonal Multimodality Integration and Clustering (OMIC) method represents a significant advancement in the analysis of single-cell multi-omics data integration. While our analysis was only focused on CITE-seq, the same model framework is applicable to other multi-omic data types. In this section, we delve into the key findings, implications, and potential future directions of our work.

Key findings and methodological contributions

Our paper introduces OMIC as a novel approach to address the complexities associated with multimodal single-cell omics data analysis. We have demonstrated its

effectiveness in multiple aspects, emphasizing the following key findings and methodological contributions.

Efficient multimodal data integration: OMIC successfully integrates information from diverse sources, particularly RNA and cell surface protein markers. This integration is pivotal for achieving a more holistic understanding of cellular identity and function.

Improved clustering accuracy: The experimental results presented in this paper showcase OMIC's competitive clustering accuracy compared to existing methods like WNN MOFA+, and totalVI. OMIC excels at distinguishing challenging cell groups, a critical capability for uncovering cellular heterogeneity.

Enhanced interpretability: OMIC's unique feature lies in its interpretability. Researchers can quantitatively assess the contributions of individual features in clustering analysis, fostering a deeper understanding of the biological relevance of integrated data. An investigation into the extent to which RNA can account for variance in ADT, coupled with logistic regression analyses, emphasizes the importance of specific ADTs as crucial cell markers.

Efficiency and scalability: OMIC not only improves accuracy but also offers efficiency gains, particularly with large datasets. It reduces computational burdens, making it a practical choice for researchers dealing with extensive single-cell omics data.

Implications and future directions

The implications of our work are significant, with broad applications in the field of single-cell genomics and cellular biology. While we have demonstrated OMIC's effectiveness on specific datasets, its applicability extends to a wide range of biological contexts. Researchers can explore its utility in various single-cell omics datasets and data types to gain a deeper understanding of cellular processes. Moreover, future studies can leverage OMIC to investigate specific biological questions, such as the identification of key cell markers and the characterization of rare cell populations.

In conclusion, the OMIC method presented in this paper offers a powerful solution to the challenges of multimodal single-cell omics data analysis. Its efficiency, interpretability, and accuracy improvements hold great promise for advancing our understanding of cellular biology at the single-cell level. As researchers continue to explore its applications and refine its methodology, OMIC is poised to have a lasting impact on the field of single-cell genomics.

Method

In this section, we describe our OMIC integration method in detail, while focusing on RNA and ADT data integration.

Data preprocessing

The CBMCs dataset [17] contains 8,617 cells with 20,501 genes and a panel of 10 antibodies. Major cord blood cell types can be discerned by marker gene expression, which has been divided into 17 clusters. The HBMCs dataset [1] consists of 30,672 cells, which contain 17,009 genes and 25 antibodies, where the dataset has been divided into 27 clusters by the cell type marker genes.

Suppose that in this experiment, n cells were sequenced, and two raw count matrices (RNA and ADT) were generated, with each row representing a cell and each column representing a feature. We first perform log-transformation and centered log ratio (CLR) transformation to RNA and ADT raw count matrices, respectively, and then perform standardization to both matrices for these two datasets. The workflow for computing RNA and ADT expressions in CITE-seq data is given as follows: For RNA expressions, we utilize the standard pipelines available in Seurat package V5 [9]. This pipeline includes essential steps such as normalization (using the "NormalizeData" function) and feature scaling (using the "ScaleData" function). In the normalization step, we use "normalization.method = LogNormalize" in the "NormalizeData" function. All other parameters are kept at their default values. For ADT expressions, we use Seurat package V5 and normalize the ADT expression levels within each cell using the centered-log ratio (CLR) transform. Subsequently, we perform feature scaling and centering using the "ScaleData" function. The CLR transform is achieved by using the "NormalizeData" in Seurat by setting "normalization.method = 'CLR' " and "margin = 2". The remaining parameters are set to their default values. Since RNA expression data in these two datasets contains a large number of features, some may not be informative due to uniform or negligible expression across cells, we apply an additional step for these two datasets to reduce the dimensionality of the datasets by screening out such features using Seurat package V5 package [9], which is to use local polynomial to fit the line between the log-variance and log-mean and then calculate the feature variance. This step removes noise and uninformative features, resulting in a selection of p RNA features for analysis. The resulting normalized gene expression measurements are then represented by an $n \times p$ matrix denoted by $\mathbf{X}$, and the normalized ADT measurements are represented by an $n \times q$ matrix denoted by $\mathbf{Y}$.

Orthogonal integration of ADT and RNA

We construct a multivariate linear regression model using the scaled data matrix of RNA as the predictor variables and the scaled data matrix of ADT as the response variables,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}, \tag{1}$$

where $\mathbf{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q]$ is a $p \times q$ matrix of coefficients, $\boldsymbol{\beta}_k = (\beta_{1k}, ...\beta_{pk})'$ is the k-th coefficient vector, $\mathbf{U} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_q]$ is the $n \times q$ residual matrix, and $\boldsymbol{u}_k = (u_{1k}, ..., u_{nk})$ is the k-th residual vector. Note that we assume each row of the residual matrix, denoted by $\boldsymbol{u}^{(i)}, i = 1, ..., n$, is uncorrelated to $\mathbf{X}$ and $\boldsymbol{u}^{(i)} \overset{\text{iid}}{\sim} N_q(\mathbf{0}, \boldsymbol{\Sigma})$. Applying the maximum likelihood estimation, we obtain the estimator of $\mathbf{B}$ and $\boldsymbol{\Sigma}$ [31],

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{Y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}. \tag{3}$$

Further, we obtain the predicted ADT matrix

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}, \tag{4}$$

and the estimated residual matrix

$$\hat{\mathbf{U}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}. \tag{5}$$

Note that $\hat{\mathbf{U}}$ is the projection of ADT information matrix $\mathbf{Y}$ on the orthogonal complement space of the column space of RNA information matrix $\mathbf{X}$. This procedure enables the extraction of additional ADT information that does not overlap with the RNA information.

Combining $\mathbf{X}$ and the residuals, we get the OMIC integrated data $(\mathbf{X}, \hat{\mathbf{U}})$. OMIC integrates RNA and ADT data while removing redundant information. Remarkably, the computational time complexity of our approach is $O(np^2)$.

Clustering

The integrated data $(\mathbf{X}, \hat{\mathbf{U}})$ is log-transformed and standardized using the same method as described in Data preprocessing section. For group cell clustering, a graph-based clustering method is selected. Specifically, a K-nearest neighbor graph is constructed, and the Louvain algorithm [32] is applied to the integrated data. The time complexity for Louvain algorithm is $O(nlog(n))$.

Finally, UMAP (Uniform manifold approximation and projection) [33] visualization is utilized to explore the relationships among cell groups.

We use residuals from ADT data rather than the original ADT data for clustering. Our goal is to incorporate both RNA and ADT information in the clustering process while minimizing redundancy and maximizing computational efficiency. To achieve this, we use the least squares method to project the scaled data information of ADT onto RNA information, removing the redundant overlap. The resulting residuals can be seen as a projection onto the complement space of RNA, which contains only ADT-related information and no RNA-related data.

Through using data $\mathbf{X}$ and $\hat{\mathbf{U}}$, we are actually using the integrated information of RNA and the non-overlapping information ADT in clustering which will be much more time-saving and precise.

We use Adjusted Rand Index (ARI) as the criterion for methods comparison [18]. The ARI is calculated as follows. Given a set $\mathcal{S}$ of n elements and two clustering results of these elements, namely $\mathcal{S}^{(1)} = \{\mathcal{S}_1^{(1)}, ..., \mathcal{S}_r^{(1)}\}$ and $\mathcal{S}^{(2)} = \{\mathcal{S}_1^{(2)}, ..., \mathcal{S}_s^{(2)}\}$, the overlap between $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ can be summarized as $[n_{ij}]$, where n_{ij} denotes the number of objects in common between $\mathcal{S}_i^{(1)}$ and $\mathcal{S}_j^{(2)}$: $n_{ij} = |\mathcal{S}_i^{(1)} \cap \mathcal{S}_j^{(2)}|$. We denote $a_i = \sum_{j=1}^{s} n_{ij}, i = 1, ..., r$ and $b_j = \sum_{i=1}^{r} n_{ij}, j = 1, ..., s$. The ARI is:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}. \tag{6}$$

Classification

Suppose we get s cell clusters in the clustering process. For each cluster j, we define a binary vector $\boldsymbol{z}^{(j)} = (z_1^{(j)}, ..., z_n^{(j)})$, where $z_i^{(j)}$ indicates whether cell i belongs to cluster j with values of either 0 or 1. Here, n represents the total number of cells.

For each cluster j, we create a normalized RNA measurements matrix $\mathbf{X}^{(j)}$ of size $n \times p_j$. The value of p_j is determined by selecting features by identifying differentially expressed genes between cluster j and the remaining cell clusters using the Wilcoxon Rank Sum test [24]. Consequently, we construct an integrated $n \times (q + p_j)$ matrix $(\mathbf{Y}, \mathbf{X}^{(j)}) = (\boldsymbol{x}_1^{(j)}, ..., \boldsymbol{x}_n^{(j)})^T$, where $\boldsymbol{x}_i^{(j)} = (x_{i,1}, ..., x_{i,q}, x_{i,q+1}^{(j)}, ..., x_{i,q+p_j}^{(j)})^T$, to combine ADT and RNA for identifying cluster j.

For each cluster, we build three logistic regression models based on different predictors (RNA alone, ADT alone, and integrated RNA and ADT). Specifically, for the integrated matrix $(\mathbf{Y}, \mathbf{X}^{(j)})$ for cluster j, we have the logistic regression model

$$P(y_i^{(j)} = 1 | \boldsymbol{\beta}^{(j)}, \boldsymbol{x}_i^{(j)}) = p_i^{(j)}(\boldsymbol{\beta}^{(j)}) = \frac{1}{1 + \exp(-(\boldsymbol{x}_i^{(j)})^T \boldsymbol{\beta}^{(j)})}, \tag{7}$$

where $\boldsymbol{\beta}^{(j)} = (\beta_1^{(j)}, ..., \beta_{q+p_j}^{(j)})^T$ is the coefficient vector for cluster j. We estimate $\boldsymbol{\beta}^{(j)}$ via minimizing the negative weighted log-likelihood [34],

$$\begin{aligned} l(\boldsymbol{\beta}^{(j)}) = -\sum_{i=1}^{n} w_i^{(j)} [z_i^{(j)} \log\{p_i^{(j)}(\boldsymbol{\beta}^{(j)})\} \\ + (1 - z_i^{(j)}) \log\{1 - p_i^{(j)}(\boldsymbol{\beta}^{(j)})\}, \end{aligned} \tag{8}$$

where $w_i^{(j)} = 0.5n[z_i^{(j)}/\pi + (1 - z_i^{(j)})/(1 - \pi)]$, with $\pi = \sum_{i=1}^{n} z_i^{(j)}/n$.

The classification criterion is set as follows:

$$\hat{z}_i^{(j)} = \begin{cases} 1 \text{ if } (\boldsymbol{x}_i^{(j)})^T \boldsymbol{\beta}^{(j)} > 0 \\ 0 \text{ if } (\boldsymbol{x}_i^{(j)})^T \boldsymbol{\beta}^{(j)} \leq 0 \end{cases}. \tag{9}$$

Settings of other methods for benchmark

We compared WNN, MOFA+, totalVI, CiteFuse and BREM-SC methods in CBMCs and HBMCs datasets with our OMIC methods in performance. We all followed the recommended settings for these methods.

We utilized the same data preprocessing method in the Data preprocessing section for WNN, MOFA+, and BREM-SC methods. For WNN, we employed the default settings as outlined in the Seurat tutorial [9], followed by clustering with the Louvain algorithm and visualization using UMAP. For MOFA+ method, we utilize z-scored data (also referred to as 'scaled' data) from the two assays view1 and view2, as recommended in the MOFA+ tutorial [10]. All other parameters were set to default values. The Louvain clustering and UMAP visualization were performed by using the learned factors identified through nearest-neighbor analysis. For CiteFuse, we followed the tutorial [14] for data preprocessing, similarity matrix fusion and clustering. For BREM-SC, we take RNA and protein UMI counts as the input and use the function: jointDIMMSC in the tutorial [15] to perform clustering analysis. For totalVI, we followed the tutorial [11] for data preprocessing, model construction, and resulting latent variables extraction for Louvain clustering. In Louvain clustering, we opt for the resolution that maximizes the ARI for each method. For example, when analyzing the CBMCs dataset using the OMIC method, the cluster number is 14, whereas it is 20 when analyzing the HBMCs dataset.

Furthermore, we include a comparison of the OMIC method with other methods, maintaining a fixed cluster number of 14 for CBMCs and 20 for HBMCs. Table 3 demonstrates that our approach exhibits superior accuracy in clustering under these settings.

Batch effect correction

Suppose there are $b = 1, ..., B$ batches of CITE-seq samples. Consider a $n \times B$ binary matrix $\mathbf{Z}$, where its (i, b)th entry z_{ij} indicates that the ith cell belongs to the bth batch if $z_{ij} = 1$. Given the existence of the batch effects, we consider the following ANOVA model of RNA and ADT.

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}_{RNA}^T + \mathbf{X}_0, \tag{10}$$

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma}_{ADT}^T + \mathbf{Y}_0, \tag{11}$$

$$= \mathbf{Z}\boldsymbol{\Gamma}_{ADT}^T + \mathbf{X}_0\mathbf{B} + \mathbf{U}, \tag{12}$$

where $\mathbf{X}_0$, $\mathbf{Y}_0$ represents the main effects of RNA and ADT expression matrices. $\boldsymbol{\Gamma}_{RNA}$ is a $B \times p$ matrix where the bth row represent the batch effect of RNA expression in the bth batch, and $\boldsymbol{\Gamma}_{ADT}$ is a $B \times q$ matrix where the bth row represent the batch effect of ADT expression in the bth batch. Compared to the model (1) where there is no batch effect, our model considered here decomposes the RNA and ADT expression into their batch effect terms and main effect terms in Eq. (10), Eq. (11). To conduct the orthogonal integration, we impose the multivariate linear regression model on their main effect terms $\mathbf{X}_0$ and $\mathbf{Y}_0$.

To estimate the RNA and ADT's batch effects and conduct the orthogonal integration of ADT and RNA, we first estimate $\boldsymbol{\Gamma}_{RNA}$ by taking regression of $\mathbf{X}$ on $\mathbf{Z}$,

$$\hat{\boldsymbol{\Gamma}}_{RNA} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X}, \tag{13}$$

and obtain the RNA expression with the batch effect being corrected as the estimated main effect,

$$\hat{\mathbf{X}}_0 = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T]\mathbf{X}. \tag{14}$$

We next estimate the $\boldsymbol{\Gamma}_{ADT}$ and the coefficient matrix $\mathbf{B}$ by taking regression of $\mathbf{Y}$ on $\mathbf{Z}$ and $\hat{\mathbf{X}}_0$. The detailed formula of the estimates $\hat{\boldsymbol{\Gamma}}_{ADT}$ and $\hat{\mathbf{B}}$ are relegated to the Additional file 1. Then, we could obtain the ADT expression with the batch effect being corrected as the estimated main effect

Table 3 Comparison of the ARI value for different methods when cluster numbers are fixed at 14 in CBMCs and 20 in HBMCs

Dataset / Method	OMIC	WNN	MOFA+	TotalVI	CiteFuse	BREM-SC
CBMCs	0.72	0.71	0.62	0.71	0.63	0.61
HBMCs	0.89	0.89	0.71	0.81	–	–

$$\hat{\mathbf{Y}}_0 = \hat{\mathbf{X}}_0\hat{\mathbf{B}}, \tag{15}$$

and the estimated residual matrix

$$\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}}_{ADT}^{T} - \hat{\mathbf{Y}}_0. \tag{16}$$

Finally, we use the estimated residuals $\hat{\mathbf{U}}$ along with $\hat{\mathbf{X}}_0$ for clustering, which is the same as Sect. .

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05773-y.

Additional file 1. Supplemental Information.

Acknowledgements
This work was partially supported by National Science Foundation grants DMS-1925066, DMS-1903226, DMS-2124493, DMS-2311297, DMS-2319279, and National Institutes of Health grants R01GM152814, RF1MH133703.

Availability of data and materials
The Cord Blood Mononuclear Cells dataset [17] is available at the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) with access no. GSE100866. The Human Bone Marrow Cells dataset [1] is available at the NCBI GEO with access no. GSE128639. The peripheral blood mononuclear cells dataset [9] is available at New York Genome Center (https://atlas.fredhutch.org/nygc/multimodal-pbmc/). The spatial CITE-seq dataset for the human tonsil [16] is available the NCBI GEO with access no. Series GSE213264. Source code for OMIC is made available on https://github.com/lyfhei/OMIC.git.

Declarations

Competing interests
The authors declare that they have no Conflict of interest.

Received: 30 January 2024 Accepted: 10 April 2024
Published online: 25 April 2024

References

1. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177:1888–190221.
2. Cheow LF, Courtois ET, Tan Y, Viswanathan R, Xing Q, Tan RZ, Tan DS, Robson P, Loh Y-H, Quake SR, et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. Nat Methods. 2016;13(10):833–6.
3. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Large-scale simultaneous measurement of epitopes and transcriptomes in single cells. Nat Methods. 2017;14:865–8.
4. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, Wang T, Larsson L, Kaczorowski D, Weisenfeld NI, Uytingco CR, Chew JG, Bent ZW, Chan C-L, Gnanasambandapillai V, Dutertre C-A, Gluch L, Hui MN, Beith J, Parker A, Robbins E, Segara D, Cooper C, Mak C, Chan B, Warrier S, Ginhoux F, Millar E, Powell JE, Williams SR, Liu XS, O'Toole S, Lim E, Lundeberg J, Perou CM, Swarbrick A. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53(9):1334–47.
5. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
6. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, Smibert P, Satija R. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19(1):224.
7. Pombo Antunes AR, Scheyltjens I, Lodi F, et al. Single-cell profiling of myeloid cells in glioblastoma across species and disease stage reveals macrophage competition and specialization. Nat Neurosci. 2021;24:595–610.
8. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.
9. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–358729.
10. Argelaguet R, Arnol D, Bredikhin DEA. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:111.
11. Gayoso A, Steier Z, Lopez R, et al. Joint probabilistic modeling of single-cell multi-omic data with totalvi. Nat Methods. 2021;18:272–82.

12. Eppstein D, Paterson MS, Yao FF. On nearest-neighbor graphs. Discret Comput Geom. 1997;17(3):263–82.
13. Miao Z, Humphreys BD, McMahon AP, Kim J. Multi-omics integration in the age of million single-cell data. Nat Rev Nephrol. 2021;17(11):710–24.
14. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. Citefuse enables multi-modal analysis of cite-seq data. Bioinformatics. 2020;36:4137–43.
15. Wang X, Sun Z, Zhang Y, Xu Z, Xin H, Huang H, Duerr RH, Chen K, Ding Y, Chen W. Brem-sc: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. 2020;48:5814–24.
16. Liu Y, DiStasio M, Su G, Asashima H, Enninful A, Qin X, Deng Y, Nam J, Gao F, Bordignon P, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial cite-seq. Nat Biotechnol. 2023;41(10):1405–9.
17. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865–8.
18. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
19. Lewis-Beck A. Applied regression: an introduction. Thousand Oaks, CA: Sage Publications; 1980.
20. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30(7):1145–59.
21. Sallusto F, Geginat J, Lanzavecchia A. Central memory and effector memory t cell subsets: function, generation, and maintenance. Annu Rev Immunol. 2004;22:745–63.
22. Rodríguez-Perea AL, Arcia ED, Rueda CM, Velilla PA. Phenotypical characterization of regulatory t cells in humans and rodents. Clin Exp Immunol. 2016;185(3):281–91.
23. Wilson NJ, Boniface K, Chan JR, McKenzie BS, Blumenschein WM, Mattson JD, Basham B, Smith K, Chen T, Morel F, Lecron J-C, Kastelein RA, Cua DJ, McClanahan TK, Bowman EP, Waal Malefyt R. Development, cytokine profile and function of human interleukin 17-producing helper t cells. Nat Immunol. 2007;8(9):950–7.
24. Haynes, W.: In: Dubitzky, W., Wolkenhauer, O., Cho, K.H., Yokota, H. (eds.) Wilcoxon Rank Sum Test, pp. 2354–2355. Springer, New York, NY (2013)
25. Kong X-F, Martinez-Barricarte R, Kennedy J, Mele F, Lazarov T, Deenick EK, Ma CS, Breton G, Lucero KB, Langlais D, Bousfiha A, Aytekin C, Markle J, Trouillet C, Jabot-Hanin F, Arlehamn CSL, Rao G, Picard C, Lasseau T, Latorre D, Hambleton S, Deswarte C, Itan Y, Abarca K, Moraes-Vasconcelos D, Ailal F, Ikinciogullari A, Dogu F, Benhsaien I, Sette A, Abel L, Boisson-Dupuis S, Schröder B, Nussenzweig MC, Liu K, Geissmann F, Tangye SG, Gros P, Sallusto F, Bustamante J, Casanova J-L. Disruption of an antimycobacterial circuit between dendritic and helper t cells in human sppl2a deficiency. Nat Immunol. 2018;19(9):973–85.
26. Miyara M, Yoshioka Y, Kitoh A, Shima T, Wing K, Niwa A, Parizot C, Taflin C, Heike T, Valeyre D, Mathian A, Nakahata T, Yamaguchi T, Nomura T, Ono M, Amoura Z, Gorochov G, Sakaguchi S. Functional delineation and differentiation dynamics of human cd4+ t cells expressing the foxp3 transcription factor. Immunity. 2009;30(6):899–911.
27. Carrette F, Surh CD. Il-7 signaling and cd127 receptor regulation in the control of t cell homeostasis. Semin Immunol. 2012;24(3):209–17.
28. Elizaga ML, Li SS, Kochar NK, Wilson GJ, Allen MA, Tieu HVN, Frank I, Sobieszczyk ME, Cohen KW, Sanchez B, Latham TE, Clarke DK, Egan MA, Eldridge JH, Hannaman D, Xu R, Ota-Setlik A, McElrath MJ, Hay CM. NIAID HIV vaccine trials network (HVTN) 087 study team: safety and tolerability of hiv-1 multiantigen pdna vaccine given with il-12 plasmid dna via electroporation, boosted with a recombinant vesicular stomatitis virus hiv gag vaccine in healthy volunteers in a randomized, controlled clinical trial. PLoS ONE. 2018;13(9):0202753.
29. Li SS, Kochar NK, Elizaga M, Hay CM, Wilson GJ, Cohen KW, De Rosa SC, Xu R, Ota-Setlik A, Morris D, Finak G, Allen M, Tieu HV, Frank I, Sobieszczyk ME, Hannaman D, Gottardo R, Gilbert PB, Tomaras GD, Corey L, Clarke DK, Egan MA, Eldridge JH, McElrath MJ, Frahm N. NIAID HIV Vaccine Trials Network: Dna priming increases frequency of t-cell responses to a vesicular stomatitis virus hiv vaccine with specific enhancement of cd8+ t-cell responses by interleukin-12 plasmid dna. Clin Vaccine Immunol. 2017;24(11):00263–17.
30. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, Hoang M, Jung J, Liang Y, McKay-Fleisch J, et al. Multiplex digital spatial profiling of proteins and rna in fixed tissue. Nat Biotechnol. 2020;38(5):586–99.
31. Mardia KV, Kent JTJT, Bibby JMJM. Multivariate Analysis. London: Probability and mathematical statistics. Academic Press; 1979.
32. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):10008.
33. McInnes, L., Healy, J., Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
34. Cox DR. The regression analysis of binary sequences. J Roy Stat Soc Ser B (Methodol). 1958;20(2):215–32.

Publisher's Note