Efficient Integration of Bulk and Single-Cell Multi-omics Data Using Latent Factor Models

Qianxing Mo†
Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida 33612, USA
Email: qianxing.mo@moffitt.org

Adam B. Olshen
Department of Epidemiology and Biostatistics, Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA 94143, USA

Integrative clustering (iCluster) methods are designed for joint clustering analysis of bulk multi-omics data measured on the same samples and have been successfully applied to identify integrative molecular subtypes for a variety of cancers. The iCluster methods are based on latent factor models that require intensive computation for feature and model selection. The computational cost has limited their application to single cell sequencing studies, which typically generate multi-omics data for tens of thousands of cells in a single experiment. Here, we present a new workflow and a new parameter-tuning algorithm for the iCluster methods that substantially reduces the computational burden, thereby enabling the application of iCluster methods to the integrative analysis of single-cell multi-omics data. We use one bulk multi-omics dataset and two single cell multi-omics datasets to demonstrate the updated iCluster methods. Our results show that the updated iCluster methods compare favorably to the Weighted Nearest Neighbor (WNN) method and the Multi-Omics Factor Analysis (MOFA2) method for integrative analysis of single cell multi-omics data. The new algorithm has been implemented in R functions iClusterPlus2 and tune.iCluster2b, which are available at https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html.

Keywords: Integrative clustering; iCluster; Latent factor; Multi-omics; iClusterPlus; iClusterPlus2.

1. Introduction

Over the past 30 years, high-throughput technologies, such as microarrays and next-generation sequencing, have enabled the generation of massive amounts of molecular data, including copy number, gene expression, DNA methylation, somatic mutations, and other molecular features of cancer. The promise of these technologies has led to large-scale collaborative initiatives, such as The Cancer Genome Atlas (TCGA)[1, 2] and the International Cancer Genome Consortium (ICGC)[3]. In these projects, as well as in similar but smaller-scale studies conducted by individual laboratories, the same biological samples were often subjected to multiple molecular assays.

† To whom correspondence should be addressed.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

In the 2010s, single-cell sequencing emerged as a prominent technology. While initially focused primarily on gene expression, single-cell sequencing has been expanded to enable the measurement of many other molecular features, including DNA mutations, DNA methylation, and DNA–protein interactions. Common examples of multimodal single-cell experiments include cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)[4], RNA sequencing combined with assay for transposase-accessible chromatin using sequencing (ATAC-seq)[5], and single-cell nucleosome, transcriptome and methylation sequencing (scNMT-seq)[6].

Although each type of multi-omics data could be analyzed separately, integrative analysis is biologically and scientifically appealing because different molecular measurements can provide complementary information about the underlying biology of cancer and other diseases. However, integrative analysis is challenging because multi-omics data are typically high-dimensional and exhibit different noise characteristics, scales and distributions. The iCluster method[7, 8] was one of the first approaches designed for integrative analysis of continuous multi-omics data. It was subsequently extended to iClusterPlus[9], which can jointly model multiple types of omics data, including continuous, binary, multi-categories and count data. In addition, a Bayesian extension of iClusterPlus, iClusterBayes[10], was developed to incorporate Bayesian variable selection[11]. To transform high-dimensional omics data into a low-dimensional space while retaining the most important patterns, the iCluster methods use low dimensional latent factors to model the high-dimensional multi-omics data. In addition, they employ lasso[12] or Bayesian variable selection approaches to identify informative features. Other integrative methods, including Cobolt[13] and MOFA[14] also use latent factor models for dimensionality reduction. A major distinction among these methods is the approach to estimate the latent factors. For example, the iCluster methods use expectation maximization (EM) algorithm and classical Bayesian methods, while Cobolt uses multimodal variational autoencoder (MVAE) and MOFA2 uses a variational Bayesian approach.

Although the iCluster methods have been successfully applied to bulk multi-omics data to identify integrative cancer subtypes across a variety of cancers, they have not been applied to analysis of single cell sequencing experiments, which typically generate multi-omics data for tens of thousands of cells. We realize that a major limitation of the iCluster methods is their intensive computational requirements for feature and model selection, which has limited their applications to large-scale multi-omics datasets. Therefore, in this manuscript, we introduce a new workflow and a novel parameter-tuning algorithm designed to substantially reduce the computational burden of the iCluster methods, thereby enabling their application to large single-cell multi-omics datasets.

The manuscript is organized as follows. In the Methods section, we briefly introduce the iCluster methods, the new workflow and the lasso parameter-tuning algorithm. In the Application section, we apply the new methods to one bulk multi-omics dataset and two single cell multi-omics datasets, and compare their performance with that of the popular Seurat WNN[15, 16] method and MOFA2. Finally, in the Discussion, we summarize and discuss the proposed methods and their potential applications.

2. Methods

2.1. *A Latent factor model for joint dimension reduction of continuous multi-omics data*

Suppose there are n samples assayed by T high-throughput techniques generating T omics datasets. These multi-omics data can be represented as data matrices for integrative analyses. Let X_t be the tth data matrix with dimensions $p_t \times n$, where p_t is the number of features (rows) and n is the number of samples (columns). Shen et al. [7, 8] proposed a Gaussian latent factor model for joint decomposition of X_t ($t = 1, 2, \ldots, \mathrm{T}$), which can be expressed as

$$X_t = W_t Z + E_t, \quad t = 1, 2, \ldots, \mathrm{T}. \tag{1}$$

In model (1), W_t denotes the weight (loading) matrix with dimensions $p_t \times k$, corresponding to p_t features and k latent factors; Z denotes the latent factor matrix with dimensions $k \times n$, corresponding to k latent factors and n samples; and E_t denotes the random error matrix with dimensions $p_t \times n$, where the column vector $e_{ti} (i = 1, \ldots, n)$ represent independent random errors that follow a multivariate normal distribution with mean 0 and a diagonal covariance matrix $\Psi_t = (\sigma_1^2, \ldots, \sigma_{p_t}^2)$. The model (1) can be viewed as a high-rank matrix X_t decomposed into two low-rank matrices W_t and Z ($k \ll p_t, n$) plus a random error matrix E_t. Sample i ($i = 1, 2, \ldots, \mathrm{n}$) is associated with the ith column vector $z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})'$ of matrix Z that captures the correlation structure of multi-omics data. As a result, the latent factor vectors can be used for sample clustering. The jth row vector $w_{tj} = (w_{tj1}, \ldots, w_{tjk})$ of W_t indicate how strongly feature j is associated with each factor. For feature j in the tth data matrix, the larger the absolute value of $|w_{tjk}|$ the stronger the association with factor k. Therefore, the weight matrix can be used to identify the important multi-omics features that contribute to the joint sample clustering. Shen, Wang and Mo[8] used a penalized EM algorithm to estimate the optimal matrices W_t and Z in which a lasso-type penalty term is imposed on W_t to introduce sparsity. When an appropriate lasso parameter is chosen, some rows of W_t become all zeros, indicating their corresponding features are non-informative, while the informative features are not all zeros. Therefore, the latent factor model (1) can be used to perform both sample clustering and feature selection, which are both important for integrative clustering analysis of multi-omics data.

2.2. *Latent factor models for joint dimension reduction of heterogeneous multi-omics data*

The original iCluster method is designed to integrate continuous multi-omics data and assumes that each data type follows a normal distribution. Although this assumption is generally appropriate for some normalized molecular measurements, it is not suitable for many commonly encountered data types. In practice, multi-omics datasets are heterogeneous, which include continuous, count, binary, and categorical variables. For example, gene expression in RNA-seq is usually measured as read counts, DNA copy number alterations are often classified as gain, loss, or normal, and somatic mutation is typically represented by a binary variable denoting the presence or absence of a mutation. To accommodate this heterogeneity, Mo et al. developed the iClusterPlus[9] method that extends the original latent factor framework by modeling each data type using an appropriate member of the generalized linear model (GLM) family. This capability represents a substantial advancement over the original iCluster method by broadening its

applicability to diverse multi-omics datasets. In the following, we briefly describe the models implemented in iClusterPlus.

In the iClusterPlus model, x_{ijt} denote a variable for the jth ($j = 1, \cdots, p_t$) omics feature of the ith ($i = 1, \cdots, n$) sample in the tth ($t = 1, \cdots, T$) dataset. Assume sample i is associated with a latent factor $z_i = \left(z_{i1,}\ z_{i2,} \ldots,\ z_{ik}\right)'$ with k elements. When x_{ijt} is a continuous variable, it is assumed to follow a Gaussian distribution, with its mean modeled as a linear function of z_i:

$$x_{ijt} = \alpha_{jt} + \beta_{jt} z_i + \varepsilon_{ijt}, \varepsilon_{ijt} \sim N(0,\ \sigma_{jt}^2), \quad (2)$$

where α_{jt} is the intercept, $\beta_{jt} = (\beta_{1jt}, \ldots, \beta_{kjt})$ are the slope coefficients, and ε_{ijt} is the random error assumed to be independent, identically and normally distributed with mean 0 and variance σ_{jt}^2. When $x_{ijt}$is a count variable, it is assumed to follow a Poisson distribution, with its mean modeled as a function of z_i through a Poisson regression with a log link function:

$$\log(\lambda(x_{ijt}|\ z_i)) = \alpha_{jt} + \beta_{jt} z_i, \quad (3)$$

where $\lambda(x_{ijt}|\ z_i)$ is the conditional mean of the count given z_i. When $x_{ijt}$is a binary variable with realized value 0 or 1, it is assumed to follow a Bernoulli distribution, with the probability of being 1 modeled as a function of z_i through a logistic regression model:

$$\log\left(\frac{P(x_{ijt}=1|z_i)}{1-P(x_{ijt}=1|z_i)}\right) = \alpha_{jt} + \beta_{jt} z_i, \quad (4)$$

where $P(x_{ijt} = 1|z_i)$ is the probability of $x_{ijt} = 1$ given z_i. Finally, when x_{ijt} is a multi-categorical variable, it is assumed to follow a multinomial distribution, with category probabilities modeled as functions of z_i through a multinomial logistic regression model:

$$P(x_{ijt}=\mathrm{c}|z_i) = \frac{\exp(\alpha_{jct}+\beta_{jct} z_i)}{\sum_{l=1}^{C}\exp(\alpha_{jlt}+\beta_{jlt} z_i)},\ \mathrm{c} = 1, 2, \ldots, \mathrm{C}, \quad (5)$$

where $P(x_{ijt}=\mathrm{c}|z_i)$ denotes the probability of x_{ijt} being the *c* category.

In the iClusterPlus model, all the omics data are linked through a common latent factor representing the underlying biological structure shared across multi-omics platforms. Parameter estimation is performed like in iCluster using a penalized EM algorithm with sparsity-inducing penalties to identify informative features while simultaneously estimating the latent factors and model parameters.

2.3. *A dynamic algorithm for searching optimal lasso parameters in T-dimensional space*

In the joint modeling of multi-omics data, iCluster imposes a lasso penalty on parameter estimation to facilitate identifying the omics features that contribute to sample clustering. When an appropriate lasso parameter is chosen, the coefficients β_{jt} corresponding to non-informative

features are shrunk to zero, whereas those associated with informative features retain non-zero values. Sample clusters are determined based on the estimated latent factors z_i. It is a computationally intensive process to search for optimal lasso parameters in a T-dimensional space, especially when $T \geq 2$. In order to efficiently search for the optimal combination of penalty parameters, iClusterPlus uses the uniform design to generate good lattice points from the search domain. For example, the recommended lattice points in a 4-dimensional space are 307, 562, 701, 1019, 2129, 3001, 4001, 5003 and 6007. Usually, the more lattice points that are searched, the more likely an optimal model is identified. Originally, iCluster was developed to identify cancer subtypes. In practice, that is typically a small number of subtypes. Therefore, it is only necessary to test a small number of latent factors (e.g., $k = 1 - 6$). Suppose there were four omics datasets and we chose $k = 1 - 6$, we would need to evaluate at least 307×6 models. The optimal *k* is selected at the elbow point of the plot of the deviance ratio or the Bayesian information criterion (BIC) versus *k*, that is where the deviance ratio or BIC levels off. Using this empirical rule, iCluster divides n samples into $k + 1$ clusters.

Although iCluster models work well for joint clustering analysis of bulk cancer multi-omics datasets, they face challenges in the analysis of larger single cell multi-omics data, which can consist of dozens of cell types among thousands of cells. To reduce computational burden, we propose the following workflow and parameter-tuning algorithm to optimize the iCluster methods.

(1) Determine an optimal number of factors by performing principal component analysis (PCA) of concatenated multi-omics data. An optimal number of factors can be identified from elbow of the PCA variance plot.
(2) For each of the multi-omics dataset, set the search domain for the lasso parameter and calculate the change of lasso parameter for each the iteration of the EM algorithm. Let λ_{t0} and λ_{t1} denote the minimum and maximum lasso parameters of dataset t, respectively. The change of lasso parameter ($\lambda_{t\Delta}$) for each iteration of the EM algorithm is calculated as
$$\lambda_{t\Delta} = (\lambda_{t1} - \lambda_{t0})/N,$$
where N is the maximum number of EM iterations.
(3) For dataset t, increase or decrease lasso parameter λ_t based on the BIC change. The following is the pseudo code for updating λ_t at each EM iteration.

```
Initialize BIC_min_t = BIC of the initial model for dataset t
Initialize BIC_rate_cutoff_t = 0.05
for i = 1 to N do
  for t = 1 to T do
     BIC_current_t = BIC of current model for dataset t
     BIC_rate_t = (BIC_current_t - BIC_min_t)/BIC_min
     if BIC_rate_t > BIC_rate_cutoff_t then
       λ_t = λ_t + λ_tΔ
     else if BIC_rate_t < −BIC_rate_t then
       BIC_min_t = BIC_current_t
       λ_t = λ_t − λ_tΔ
  end for
end for
```

(4) After the iterations, perform uniform manifold approximation and projection (UMAP) on the latent factor matrix to obtain a two-dimensional embedding and visualize the sample/cell clusters in the resulting UMAP space.
(5) Perform clustering analysis for the samples or cells using partition-based methods (e.g. k-means) or graph-based methods (e.g., Louvain[18] method), depending on the clustering structure observed in the UMAP embedding.

In the above algorithm, we set a cutoff for the BIC rate of change between 2 consecutive iterations. If the rate of change is less than the cutoff (e.g., 0.05), λ_t will remain the same. Suppose the number of iterations is N. The possible lattice points in a T-dimensional space will be N^T, which is quickly prohibitive with the old algorithm. We use three real examples to illustrate the new workflow.

2.4. *Comparisons with other methods*

We ran WNN using default parameters. For iCluster and MOFA2, we chose the same number of latent factors. We ran MOFA2 using the same overall workflow as iCluster except in the step used to estimate the latent factors.

3. Application

3.1. Integrative clustering analysis of uveal melanoma (UM) multi-omics data.

We first applied the new algorithm to the bulk uveal melanoma (UM) multi-omics data generated by TCGA[2]. The UM dataset consists of somatic mutation, DNA copy number variation, DNA methylation and RNA-seq gene expression data for 80 UM primary samples. The level-3 UM multi-omics data available at http://firebrowse.org were processed to form 4 data matrices in which the columns represent the samples, and the rows represent the omics features. The data procession was detailed in Mo et al.[17]. Briefly, somatic mutation was coded as 1 and wild-type was coded as 0. Genes with a somatic mutation rate $\geq 2\%$ in the samples were used for iCluster analysis. We logit transformed the methylation beta values and then selected the top 25% most variable genes. We log2 transformed the mRNA-seq normalized counts plus 1 and selected the top 25% most variable genes. For copy number data, we condensed the chromosome segments to 2632 chromosome regions using the CNregions function in the iClusterPlus package. Finally, we standardized omics feature values across samples for the methylation, mRNA and copy number data.

To determine the number of factors used for the latent factor model, we concatenated the standardized methylation, mRNA and copy number data matrix and then performed PCA on the concatenated matrix. Figure 1 shows the proportion of the total variance explained by the first 30 principal components (PCs). The explained variance decreases rapidly for the first few PCs, after which the rate of decline slows markedly beginning around the fifth PC. The optimal number of latent factors for iCluster modeling can be selected near this inflection point. We chose $k = 7$ as the number of latent factors for iCluster integration. The resulting sample clustering was highly consistent for values of $k \geq 7$, indicating a stable solution.

It took about a minute to run the new algorithm on a Mac Studio with a M4 Max chip, which was implemented in a new function named iClusterPlus2. In contrast, the first version of iClusterPlus took about 12 hours to evaluate 701 combinations of lasso parameters for each of the latent factors ($k = 1 - 5$). Both versions of iClusterPlus generated the same clustering results.

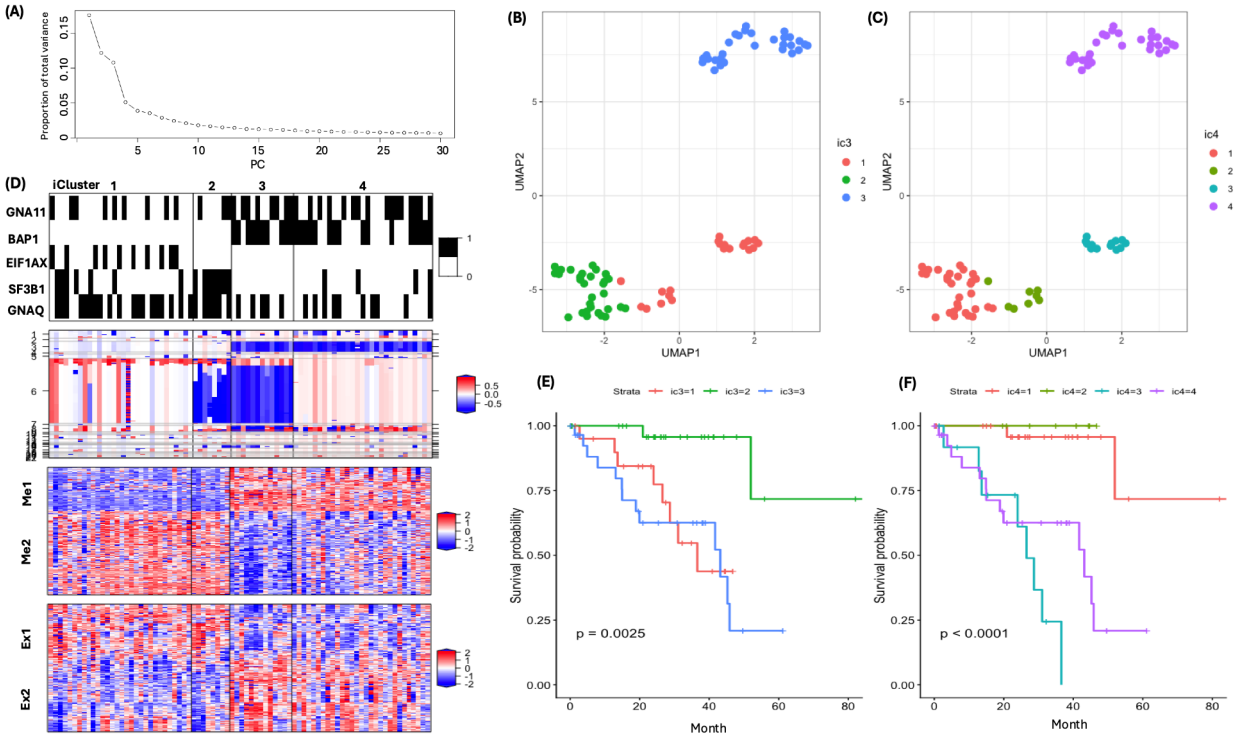


Figure 1. iCluster analysis of UM multi-omics data. (A) PCA variance plot: proportion of total variance vs. PC. (B) UM samples distributed on UMAP 2-dimensional space; UM samples are grouped into 3 clusters by K-means clustering. (C) UM samples distributed on UMAP 2-dimensional space; UM samples are grouped into 4 clusters by K-means clustering. (D) Heatmap of the top driver features of the multi-omics data. Somatic mutation (the top panel) is indicated by black bar; DNA copy number (the 2nd panel from the top): copy number loss, normal and gain are indicated by blue, white and red, respectively; Methylation (the 3rd panel from the top): hypomethylation and hypermethylation are indicated by blue and red, respectively. mRNA (the bottom panel): low and high expression are indicated by blue and red, respectively. (E) Overall survival of the 3 iClusters. (F) Overall survival of the 4 iClusters.

To visualize the sample clusters in the latent factor space, we performed UMAP on the latent factor matrix. Figure 1B and 1C show the 3 and 4 major clusters on the UMAP 2-dimensional space, respectively. It can be seen that cluster 1 of Figure 1B is further divided into 2 sub-clusters (clusters 2 and 3) of Figure 1C. The iCluster analysis identified the somatic mutations *GNAQ*, *SF3B1*, *EIF1AX*, *BAP1* and *GNA11* were drivers of UM subtyping (Figure 1D, top panel). The driver copy number alterations included chr3 loss, chr6 gain and loss, and chr8 gain (Figure 1D, the 2nd panel from the top). The driver methylated genes formed 2 major gene clusters labeled as Me1 and Me2, which had opposite methylation patterns from each other (Figure 1D, the 3rd panel from the top). In terms of gene expression, the driver genes also formed 2 major gene clusters labeled as Ex1 and Ex2 that had opposite expression patterns (Figure 1D, the bottom panel). Overall, it appears that the multi-omics patterns drive the samples clustering into two major clusters 1+2 and 3+4, which coincides with the chr3 status (chr3 normal in clusters 1+2, chr3 loss

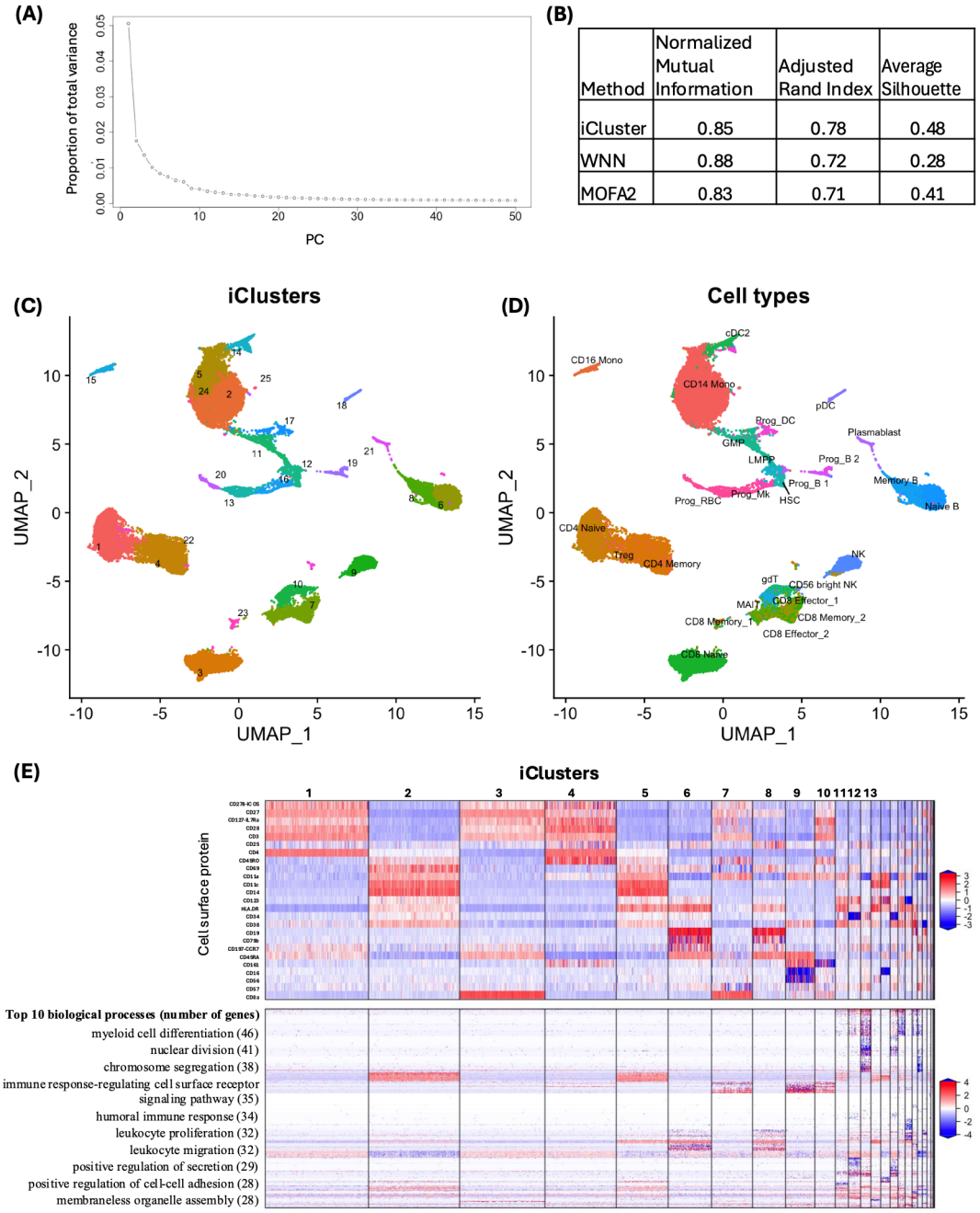
in clusters 3+4). The loss in chr6q arm is the major visible omics pattern that divides clusters 1+2 and 3+4 further into two smaller clusters, where clusters 2 and 3 are associated with chr6q loss, while clusters 1 and 4 are not. It is well known that the most common chromosomal aberration in this disease is the loss of one copy of chr3 (monosomy 3). Patients with monosomy 3 are associated with a worse overall survival than those with normal chr3, while there is no significant difference in overall survival between clusters 1 and 2, nor between clusters 3 and 4 (Figure 1F). In the 3-cluster solution, both clusters 2 and 3 include patients with chr3 loss, which may explain why there is not significant survival difference between the two clusters (Figure 1E). Statistically, it is challenging to determine which clustering result (the 3-cluster solution or the 4-cluster solution) is better. Clinically, it is clear that the 4-cluster solution is better because it is more consistent with the clinical observation.

3.2. Integrative clustering analysis of CITE-seq data.

Our second application is to single cell multi-omics data collected using the CITE-seq method[9]. The dataset consists of single cell RNA-seq (scRNA-seq) and antibody-derived tag (ADT) data from the same 30,672 cells from human bone marrow cells first introduced by Stoecklus et al.[9] with the experiment performed by Stuart et al.[16]. The data can be obtained by installing the SeuratData R package (https://github.com/satijalab/seurat-data). The cells include 17009 RNA features and 25 ADT features. We processed the CITE-Seq data following the Seurat Weighted Neighbor Analysis (WNN) vignette (https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis), which included normalization, scaling and feature selection. As a result, the 2000 most highly variable RNA-seq genes and all the 25 ADT features were used for iCluster and WNN analyses.

To estimate the ideal number of factors to use for iCluster modeling, we performed PCA of the concatenated matrix of ADT and RNA-seq. We found that the explained variances level off around the 10th PC (Figure 2A). For iCluster modeling of single cell multi-omics data, we suggest selecting the PC after which the explained variance levels off. As an example, we selected $k = 20$ as the number of latent factors for iCluster modeling. Actually, we found that the clustering results were generally consistent for a range of factors (e.g., $k = 20 - 30$). Because features had already been selected, we utilized a small value for the lasso parameters (λ=0.1) for both the RNA and ADT data. The model ran in only 17 seconds on the M4 Mac Studio mentioned earlier.

We utilized the Louvain[18] method (a graph-based clustering method implemented in the RphenoGraph function in R) to identify cell clusters in the latent factor space. As a result, we identified 25 clusters ranging in size from 4736 to 29 cells. These cell clusters are visualized on the 2-dimensional UMAP of the latent factors (Figure 2C). The annotated cell clusters are shown on Figure 3D. The most abundant cells are CD4 naïve, CD14 mono, CD8 naïve, CD4 memory, CD14 mono and naïve B cells (Figure 3D). The ADT and RNA expression patterns are shown on Figure 2E. We can see that the iClusters 1 (CD4 naïve), 3 (CD8 naïve) and 4 (CD4 memory) are characterized by high expression of CD4, CD8a, CD4 and CD45RO, respectively (Figure 2E, top panel). In addition, the three iClusters are also characterized by high expression of CD278-ICOS, CD27, CD127-IL7Ra, CD28 and CD3 (Figure 2E, top panel).



Method	Normalized Mutual Information	Adjusted Rand Index	Average Silhouette
iCluster	0.85	0.78	0.48
WNN	0.88	0.72	0.28
MOFA2	0.83	0.71	0.41

Figure 2. iCluster analysis of CITE-seq data. (A) The PCA variance plot of the CITE-seq data. (B) Comparison of iCluster, WNN and MOFA2 in terms of normalized mutual information, adjusted rand index and average silhouette. (B) Cell clusters identified by iCluster. (D) Annotated cell clusters. (E) Heatmaps of cell surface markers of the ADT data (top panel) and the top 500 genes of the scRNA-seq data (bottom panel).

To identify the most important genes that drove the cell clustering, we ranked the RNA-seq genes by the sum of absolute values of the model parameters (βs) from the largest to the smallest.
Figure 2E shows the top 500 genes of the RNA-seq data. Immune related biological processes (gene ontology terms) are enriched in the top 500 RNA-seq genes, which include myeloid cell differentiation (46 genes), immune response-regulating cell surface receptor signaling pathway (35

genes), humoral immune response (34 genes), leukocyte proliferation (32 genes) and migration (32 genes) (Figure 2E, bottom panel).

To evaluate the performance of iCluster, WNN and MOFA2, we calculated three performance metrics for which higher values indicate better performance: normalized mutual information (NMI), adjusted rand index (ARI) and average silhouette (AS). These three metrics were calculated by comparing the cell clusters identified by each method with the annotated cell types. iCluster achieved the highest ARI and AS values, and the second-highest NMI value (Figure 2B). Overall, iCluster demonstrated the best performance.

3.3. Integrative clustering analysis of scRNA-seq and scATAC-seq data.

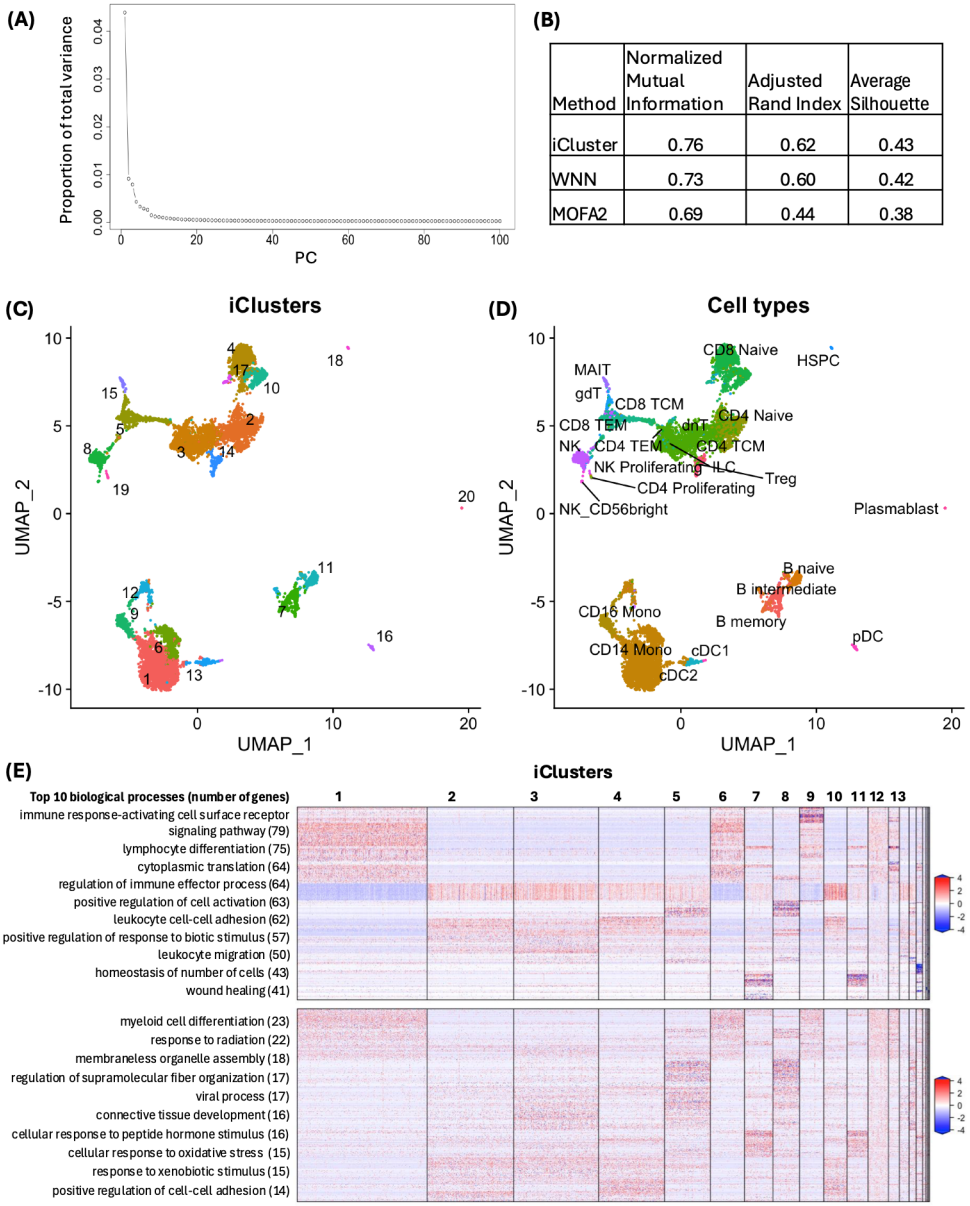
Our third example is another common type of single cell experiment where scRNA-seq and scATAC-seq[5] are applied to the same set of cells, in this case 11,909 PBMCs, and can be found at the 10x Genomics website. ATAC-seq is a technique that is used to measure chromosome accessibility. ATAC-seq data are correlated within regions, with high regional counts implying open chromatin regions. Integrating scRNA-seq and scATAC-seq should allow better identification of cell types.

We used the Seurat package to process the scRNA-seq and ATAC-seq data, which included filtering out cells with low quality data, normalization and scaling of the sequence counts and feature selection. Specifically, we filtered out cells with percent of RNA mapped to mitochondrial genes $\geq 20\%$, total ATAC-seq counts $\leq 5{,}000$ or $\geq 70{,}000$, and total RNA-seq counts $\leq 1{,}000$ or $\geq 25{,}000$. As a result, 10,412 cells remained for iCluster analysis. We selected the 3000 most variable genes for the RNA-seq data and top 20% most variable regions for the ATAC-seq data. In addition, we merged neighboring ATAC regions into the same regions, which led to 15,441 regions for iCluster analysis.

To determine the number of factors used for iCluster modeling, we performed PCA of the concatenated matrix of RNA-seq and ATAC-seq. We found that the variances level off around the 10th PC (Figure 3A). However, since the first PC was dominated by sequencing depth, Seurat only used the 2nd to the 50th PCs for the WNN integrative analysis (see Seurat WNN tutorial). In order to have a fair comparison, we chose $k = 50$ latent factors for iCluster modeling. Then, we identified the cell clusters in the latent factor space using the Louvain method and visualized the cell clusters on a 2-dimensional UMAP of the latent factors. We annotated the cells using Azimuth software (https://azimuth.hubmapconsortium.org/) with PBMC as the reference.

We identified 20 cell clusters with sizes ranging from 2138 to 18 cells (Figure 3C). The most abundant cell types are CD14 Mono, CD4 Naïve, CD4 TCM, CD8 Naïve and CD8 TEM (Figure 3D). HSPC, Plasmablast and pDC are among the least abundant cells and they are well separated from the other mature immune cells on the UMAP plot, which are expected (Figure 3D). To identify the most important features that drove the cell clustering, we ranked the RNA-seq genes and ATAC-seq regions by the sum of absolute values of the model parameters (βs) from the largest to the smallest. Figure 3E shows the top 750 genes of the RNA-seq data and 1544 regions of the ATAC-seq data. The top biological processes (gene ontology terms) involved in the top 750 RNA-seq genes are immune related, which include the immune response-activating cell

surface receptor signaling pathway (79 genes), lymphocyte differentiation (75 genes), regulation of immune effector process (64 genes), positive regulation of cell activation (63 genes), leukocyte cell-cell adhesion (62 genes) and positive regulation of response to biotic stimulus (57 genes). A lot of genes falling in the top ATAC-seq regions are also involved in immune-related processes, which include myeloid cell differentiation (23 genes), response to radiation (22 genes) and viral process (17 genes); the others are involved in cellular signal pathways such as cellular response to peptide hormone stimulus, cellular response to oxidative stress (15 genes) and positive regulation of cell-cell adhesion (14 genes).



Method	Normalized Mutual Information	Adjusted Rand Index	Average Silhouette
iCluster	0.76	0.62	0.43
WNN	0.73	0.60	0.42
MOFA2	0.69	0.44	0.38

Figure 3. iCluster of RNA-seq and ATAC-seq data. (A) The PCA variance plot of the combined RNA-seq and ATAC-seq matrices. (B) Comparison of iCluster, WNN and MOFA2 in terms of normalized mutual information, adjusted rand index and average silhouette. (C) Cell clusters identified by iCluster. (D) Annotated cell clusters. (E) Heatmap of the top features in the RNA-seq data (top panel) and the ATAC-seq data (bottom panel).

Using the cell types predicted by Azimuth as the reference, we calculated the NMI, ARI and AS for the cell clusters identified by iCluster, WNN and MOFA2. Compared to WNN and MOFA2, iCluster achieved the highest values for all the three metrics, indicating the best overall performance (Figure 3B).

Discussion

In this manuscript, we introduce a new workflow and parameter-tuning algorithm for the iCluster methods to integrate multi-omics generated from the same samples or cells. In the new approach, we first determine the optimal number of latent factors for iCluster modeling using principal component analysis of the concatenated data matrix. Then, we use a dynamic algorithm to search for optimal lasso penalty parameters during iCluster modeling. Next, we visualize sample or cell clusters on a two-dimensional UMAP embedding of the latent factor matrix. Finally, we determine cluster memberships utilizing either partition-based or graph-based clustering of the latent factors depending on the clustering structure observed in the UMAP embedding.

This new approach overcomes a major limitation of the previous iCluster workflow, which requires an extensive grid search over different lasso penalty parameters and numbers of latent factors. As a result, the new approach substantially reduces the computational burden of the iCluster methods while maintaining their ability to integrate multi-omics data. Using TCGA uveal melanoma and two single-cell multi-omics datasets, we demonstrate that the new approach can be used not only to identify cancer molecular subtypes but also to identify cell clusters in single-cell multi-omics datasets comprising tens of thousands of cells. These results demonstrate the broad applicability and computational efficiency of the proposed approach for both bulk and single-cell multi-omics data analyses.

We compared the updated iCluster methods to the popular Seurat WNN method and the MOFA2 method. Overall, the updated iCluster methods demonstrated overall superior performance compared with the WNN and MOFA2 methods across three evaluation metrics: NMI, ARI, and AS score. In addition, because the iCluster methods are model-based, the model parameters can be used to rank multi-omics features to identify the most important features that drive the cell clustering. This feature-selection capability is an important advantage of model-based methods. In contrast, the WNN method does not provide a direct approach for identifying and ranking multi-omics features that contribute most to cell clustering in single-cell multi-omics data analysis.

One limitation of our study is that we compared the iCluster methods only with the WNN and MOFA2 methods. Several other methods have been developed for integrative analysis of multi-omics data and could also be considered for comparison. In future work, we plan to conduct more extensive comparisons with additional methods.

In summary, we developed a new computationally efficient approach for the iCluster methods that substantially improves their applicability to both bulk and single-cell multi-omics data. We believe that this approach will be a useful computational tool for multi-omics data analysis in biological research.

References

1.Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–25.
2.Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. Cancer Cell. 2018;33(1):151.
3.International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8.
4.Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865–8.
5.Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5.
6.Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018;9(1):781.
7.Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12.
8.Shen R, Wang S, Mo Q. Sparse Integrative Clustering of Multiple Omics Data Sets. Ann Appl Stat. 2013;7(1):269–94.
9.Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013;110(11):4245–50.
10.Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2018;19(1):71–86.
11.Savitsky T, Vannucci M, Sha N. Variable Selection for Nonparametric Gaussian Process Priors: Models and Computational Strategies. Stat Sci. 2011;26(1):130–49.
12.Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 1996;58(1):267–88.
13.Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 2021;22(1):351.
14.Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020; 21(1):111.
15.Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888–902 e21.
16.Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87 e29.
17.Mo Q, Wan L, Schell MJ, Jim H, Tworoger SS, Peng G. Integrative Analysis Identifies Multi-Omics Signatures That Drive Molecular Classification of Uveal Melanoma. Cancers (Basel). 2021;13(24).
18.Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el AD, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015;162(1):184–97.