NAR Genomics and Bioinformatics, 2024, **6**, lqae141
https://doi.org/10.1093/nargab/lqae141
Advance access publication date: 16 October 2024
Standard Article

MoNETA: MultiOmics Network Embedding for SubType Analysis

Giovanni Scala[1,†], Luigi Ferraro[2,†], Aurora Brandi[1], Yan Guo[2], Barbara Majello[1] and Michele Ceccarelli[2,*]

[1]Department of Biology, University of Naples 'Federico II', 80128 Naples, Italy
[2]Sylvester Comprehensive Cancer Center, University of Miami, 33136, Miami, USA

[*]To whom correspondence should be addressed. Tel: +1 305 243 3016; Fax: +1 305 243 3016; Email: mxc2982@miami.edu
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Cells are complex systems whose behavior emerges from a huge number of reactions taking place within and among different molecular districts. The availability of bulk and single-cell omics data fueled the creation of multi-omics systems biology models capturing the dynamics within and between omics layers. Powerful modeling strategies are needed to cope with the increased amount of data to be interrogated and the relative research questions. Here, we present MultiOmics Network Embedding for SubType Analysis (MoNETA) for fast and scalable identification of relevant multi-omics relationships between biological entities at the bulk and single-cells level. We apply MoNETA to show how glioma subtypes previously described naturally emerge with our approach. We also show how MoNETA can be used to identify cell types in five multi-omic single-cell datasets.

Introduction

The advancements in molecular biology and biotechnology have fostered the development of systems biology as an interdisciplinary field that aims to characterize the complexity of biological systems by integrating information from various levels of organization, ranging from molecular to cellular components. Understanding the intricate regulatory mechanisms governing biological processes and how they are affected by human disease can facilitate advancements in medical research, biotechnology, and personalized medicine (1). The development of novel analytic approaches is facilitated by the availability of large-scale datasets covering several human diseases, such as cancer, within international consorti such The Cancer Genome Atlas (TCGA) (2), the Human Genome Project (3) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (4). These community efforts provided researchers with an extensive array of cell profiles (5), greatly expanding our understanding of characteristics underlying cellular activities. The increasing accessibility of this data has encouraged a growing interest in exploring the intricate interplay of omics data to decipher emerging phenotypes, offering promising prospects across various scientific domains (6–9). The plethora of data derived from these technologies enables a more refined exploration of the critical factors influencing disease subtyping and targeted therapies (10).

Statistical machine learning models can be used for the integration of multi-omics data into a lower-dimensional space as a means to gain a comprehensive visualization and understanding of the data itself (6–9). Current methods for integrating multi-omics data demand significant computational resources. This is because they involve complex calculations and storage of large intermediate structures, like distance matrices and networks, that capture similarities between different entities. These networks, originally termed Patient Similarity Networks (PSN) (11), are a common approach. However, we extend this concept to form an Entities Similarity Network (ESN), recognizing its capacity to describe not only patients but also cells, genes, proteins and other molecular entities. Furthermore, multi-omics data can overcome several approaches, such as monoplex networks (12), focused on a single molecule layer by exploiting communications and interactions across various biological layers.

Here, we introduce a novel general-purpose method for multi-omics integration called MoNETA (Multi-Omics Network Embedding for SubType Analysis). This innovative approach facilitates identifying significant multi-omics relationships in biological samples and individual cells, offering speed and scalability. Our choice lies in adopting a biologically informed multiplex network (13) for MoNETA, driven by three key reasons: (i) networks are well-suited for describing entities as nodes and expressing their similarity through edges; (ii) including information related to biological relationships within the multi-omics context improves the overall model's efficacy, merging all monoplex networks into a singular, comprehensive multi-omics network; iii) heterogeneous networks allow for the integration of entities even when they lack information across all modalities.

Received: March 25, 2024. **Revised:** July 19, 2024. **Editorial Decision:** September 27, 2024. **Accepted:** October 4, 2024

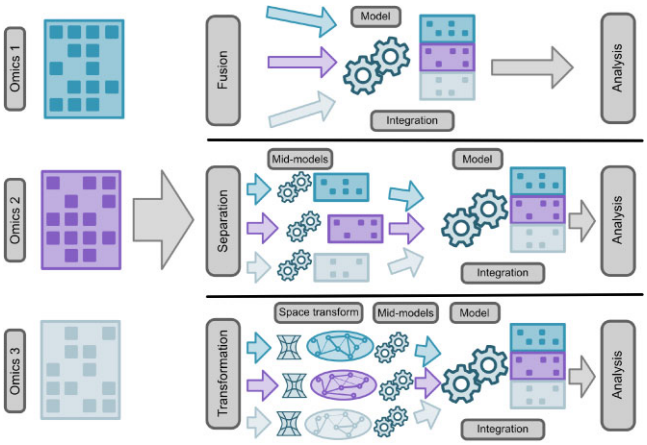

Figure 1. Representation of multi-omics data integration techniques. Fusion-based approaches unite diverse omics data into a singular model. Separation-based approaches build intermediate models for each omics, merging insights for a comprehensive analysis. Transformation-based approaches apply graph or kernel-dependent algorithms before integration, enhancing entity description.

Related works

Multi-omics integration is an active research area, and many methods have been proposed (see (7) for a review on PSN). In general, we can classify these contributions according to the phase of the process where the merging of the feature takes place (Figure 1): *Fusion-based*, *Separation-based* and *Transformation-based*. *Fusion-based* integration is executed by concatenating different omics data sets. *Separation-based* methods isolate distinct omics data sets for independent analyses before subsequent integration, allowing focused exploration of each data type's unique characteristics before merging them. *Transformation-based* approaches apply transformations to individual omics data sets before their integration, enhancing the compatibility and meaningful convergence of diverse data sources.

Fusion-based approaches involve constructing an unified model for analyzing combined omics data. This integration can occur through concatenation and blending methods. Concatenation (14,15) is a straightforward juxtaposition of multi-omics data. Essentially, the data from different molecular profiling platforms are combined side by side, creating larger matrices that encapsulate the entirety of the information with the drawback of not providing any dimensionality reduction. Blending methods (16–19) utilize dimensionality reduction techniques to merge multi-omics information. This process retains the most crucial information for describing an entity while significantly reducing the overall dimension of the data. Although more intricate in implementation, blending methods are favored for their efficiency in managing data. Fusion methods encompass a range of techniques, and one notable approach is Multi-Omics Factor Analysis (MOFA/MOFA+) (20,21). This method tackles a challenge by deducing interpretable (hidden) factors that capture biological and technical variations. Specifically, MOFA utilizes Factor Analysis to reduce the dimensionality of the data. MOFA also aims to untangle whether each factor is unique to a single type of data or is evident in multiple types, thereby uncovering shared patterns of variation across various omics layers.

The separation-based approach (22–27) involves constructing an intermediate model for each omics dataset and subsequently merging the outcomes for analysis using a joint model. This method proves highly effective in elucidating relationships within individual omics datasets. By doing so, the approach enables a more reliable feature selection process prior to the final integration. The results from these intermediate models are combined during the last integration step. The final phase of the analysis is then executed, utilizing various techniques such as majority voting (23). MDNNMD (Multimodal Deep Neural Network for Multi-dimensional Data integration) (25) is a separation-based approach that combines deep learning methodologies with a thorough analysis of molecular characteristics to enhance the diagnosis, treatment and prevention of breast cancer. This method consists of three autonomous models, each responsible for encoding specific omics data types. The autonomy of these models enables a detailed exploration of the complex molecular landscape associated with breast cancer. In the final stage, MDNNMD merges the predictive scores generated by each independent model.

Raw omics data remain unaltered in the previously discussed methods, preserving their original structure. Contrastingly, transformation-based approaches (28–31) introduce the application of graph or kernel-dependent algorithms before integration. These structural transformations are well-suited for describing entities and identifying potential patterns within the data. The investigation into optimal distance metrics (11,32,33) or kernel functions (34–37) stands out as the most crucial step in constructing these methods. While kernel functions demonstrate superior performance compared to network-based methods (7), the latter are more accessible to interpret and involve less time. When integrating multi-omics datasets, this becomes a significant consideration, especially in extensive analyses like pan-cancer or single-cell integration, where matrices can become considerably large. The efficiency and interpretability of network-based methods, even with potential performance trade-offs, make them a practical choice. Following this, a separation- or fusion-based approach could be employed to achieve the intended task. An example of transformation-based method is the weighted-nearest neighbor (WNN) analysis (38), available in the Seurat package (39) for single-cell analysis. The key steps involve independently calculating nearest neighbors for each data modality to form k-nearest neighbor (KNN) graphs. Predictions for each cell molecular profile are made based on its neighbors in each modality, with accuracy comparisons determining modality-specific weights. These weights, reflecting the relative utility of each data type, are used to construct an integrated WNN graph for downstream analysis.

Materials and methods

MoNETA (Multi-omics Network Embedding for SubType Analysis) adopts a transformation-based methodology by defining networks for individual modalities and subsequently merging them into a comprehensive network. The output of this integration process is an embedding matrix that depends on all the omics layers. Our implementation of MoNETA prioritizes scalability, can face the complex nature of multi-omics data and can manage non-overlapping entity sets among omics-layers, i.e. does not require that a given entity is represented in all layers. MoNETA is implemented as an R package available at the following repository: https://github.com/BioinfoUninaScala/MoNETA.

Datasets

In order to evaluate our model, we use one tri-modal bulk dataset, two bi-modal, two tri-modal, and one hepta-modal single-cell datasets S1.

Pan-glioma data

A multi-omic cohort of glioma samples (40) was downloaded using TCGABiolink (41). The integrative supervised analysis reported in (40) has classified glioma, both lower grade and high grade (Glioblastoma), into seven classes according to the methylation, expression, mutational, and copy number profiles. Here, we considered 788 TCGA cases based on RNA sequencing, copy number variation and methylation availability. In particular, for each sample, we retrieved CNV profiles for 24 776 genes, DNA methylation levels for 1300 glioma-specific CpG sites, and mRNA levels of 12 985 genes of the combined gene expression matrix (40).

Single-cell multi-omics data

Integrating multi-omics single-cell datasets aims to leverage information from diverse omics layers, enabling the comprehensive characterization of distinct cell types and functional states. To test the ability of MoNETA to handle multi-omic datasets, we considered five omics datasets (Supplementary Table S1).

The first bi-modal dataset, generated by (42) using the CITE-seq protocol on human PBMC and lung cells, included 10,470 cells with measurements of 33 514 RNA levels and 52 surface proteins. (43) developed another bi-modal CITE-seq dataset focusing on human bone marrow (BM) mononuclear cells. This comprehensive dataset comprises over 30 672 cells, annotated into 27 cell types, and includes 17 009 transcript levels and 25 surface proteins. A tri-modal dataset of human PBMCs, created by (44) using the TEA-seq protocol, comprised 25 517 cells and featured 36 601 RNA levels, 128 853 accessible chromatin ATAC-seq peaks, and 47 surface proteins. (45) provided another tri-modal dataset, DOGMA-seq, on human PBMCs. This dataset encompasses 13 763 cells, distributed across 27 cell types, with 36 495 RNA levels, 68 963 accessible chromatin ATAC-seq peaks and 210 surface protein abundances. The hepta-modal dataset, generated by (46) using Paired-TAG and Paired-seq protocols, includes 52 781 genes and 2 965 565 1000-length bins distributed among DNA matrices as follows: 2 443 832 ATAC bins, 500 634 H3K4me1 bins, 1 144 963 H3K4me3 bins, 452 748 H3K27me3 bins, 471 509 H3K27ac bins and 519 186 H3K9me3 bins. Due to the large number of cells, we performed subsampling, selecting four out of the initial 22 cell lines. Each cell line represents a different macro-class: non-neuronal cells (NonNeu), cortical (FC), inhibitory (InNeu) and hippocampal (HC) neurons. The filtered dataset comprises 7556 cells, annotated as follows: 2553 cells as BR.InNeu.CGE, 1463 cells as BR.NonNeu.Microglia, 2046 cells as FC.ExNeu.PT and 1495 cells as HC.ExNeu.Subiculum. Each cell includes transcriptomic data, with DNA-level data distributed as follows: 1484 cells for ATAC, 1374 cells for H3K27ac, 835 cells for H3K27me3, 1513 cells for H3K4me1, 807 cells for H3K4me3 and 1544 cells for H3K9me3. To enable the application of other integration methods, such as MOFA (20,21) and WNN (38), we had to manipulate the omics matrices in order that all matrices contain the same cells, eventually with empty values. This manipulation is not necessary for the MoNETA pipeline.

Integrating multi-omics data through network embedding

The MoNETA workflow, depicted in Figure 2, begins with a collection of L omic matrices denoted as $M = \{M_\alpha | \alpha = 1, \ldots, L\}$. Each matrix $M_\alpha \in \mathbb{R}^{f_\alpha \times n_\alpha}$ contains measurements of f_α features for a subset n_α of entities, which could be bulk samples or single-cells from a total cohort of n entities. The fundamental goal of the MoNETA approach is to derive a matrix $EM \in \mathbb{R}^{n \times d}$, where $d << n$, encapsulating a latent space representation of the multi-omics molecular profiles embedded within the input dataset. To construct this model, MoNETA follows a sequence of four essential steps. It begins with the retrieval the input data, basic preprocessing include filtering for expression and scaling. Following this, individual omics networks are computed. These networks are then integrated to create a biologically informed multi-omics network. The final step involves executing a Random Walker with Restart procedure (47) to achieve the multi-omics embedding.

Single omics networks computation

For each omics matrix M_α, MoNETA builds an Entity Similarity Network, $ESN_\alpha = (V_\alpha, E_\alpha)$. Here, V_α denotes a set of nodes corresponding to entities, and E_α represents a set of edges connecting every entity with its nearest neighbors within that specific omics layer. Selecting the nearest-neighbor entities is a computationally expensive task that involves exhaustively scanning all data points for each entity. To address this, we utilize the VP-tree (Vantage Point tree) method (48), which avoids the need to compute the full $n_\alpha \times n_\alpha$ distance matrix. This approach allows for the fast retrieval of a set of closest neighbors for each entity and provides the flexibility to choose distance metrics based on the nature of the omics data. MoNETA uses the *buildIndex* function from the BiocNeighbors package to construct a VP-tree. Subsequently, the *queryKNN* function is utilized to extract a set of neighbors for each entity. Two approaches can be employed to select the neighborhood for each entity: the static kNN method, where each entity is linked to its k closest neighbors, and the dynamic neighborhood k^*nn approach. In the latter, a variable number of neighbors k^*nn, ranging from 1 to a user-defined maximum max_k, is selected for each node by applying the k^*nn algorithm as defined in (49) that employs a greedy approach to determine the optimal set of neighbors for each node based on the distances from its closest max_k neighbor nodes.

Multi-omics network integration

Subsequently, the networks ESN_α created for each omics are merged into a multiplex graph $G_M = (V_M, E_M)$. For each entity i in each omics α, there is a corresponding node $v_{\alpha i}$ in V_M, such that $V_M = \{v_{\alpha i}, i = 1 \ldots n, \alpha = 1 \ldots L\}$. The edge set E_M is the union of three distinct sets: E_O, E_S and E_N. The set E_O consists of all edges within each ESN_α:

$$E_O = \cup_\alpha ENS_\alpha$$

The set E_S contains edges connecting the same entity across different omics:

$$E_S = \{(v_{\alpha i}, v_{\beta i}), i = 1 \ldots n, \alpha, \beta = 1 \ldots L, \alpha \neq \beta\}$$

The set E_N includes edges connecting an entity in one omics to the neighbors of the same entity in another omics:

$$E_N = \{(v_{\alpha i}, v_{\beta j}), i, j = 1 \ldots n, \alpha, \beta = 1 \ldots L, \alpha \neq \beta, (v_i, v_j) \in E_\alpha\}$$

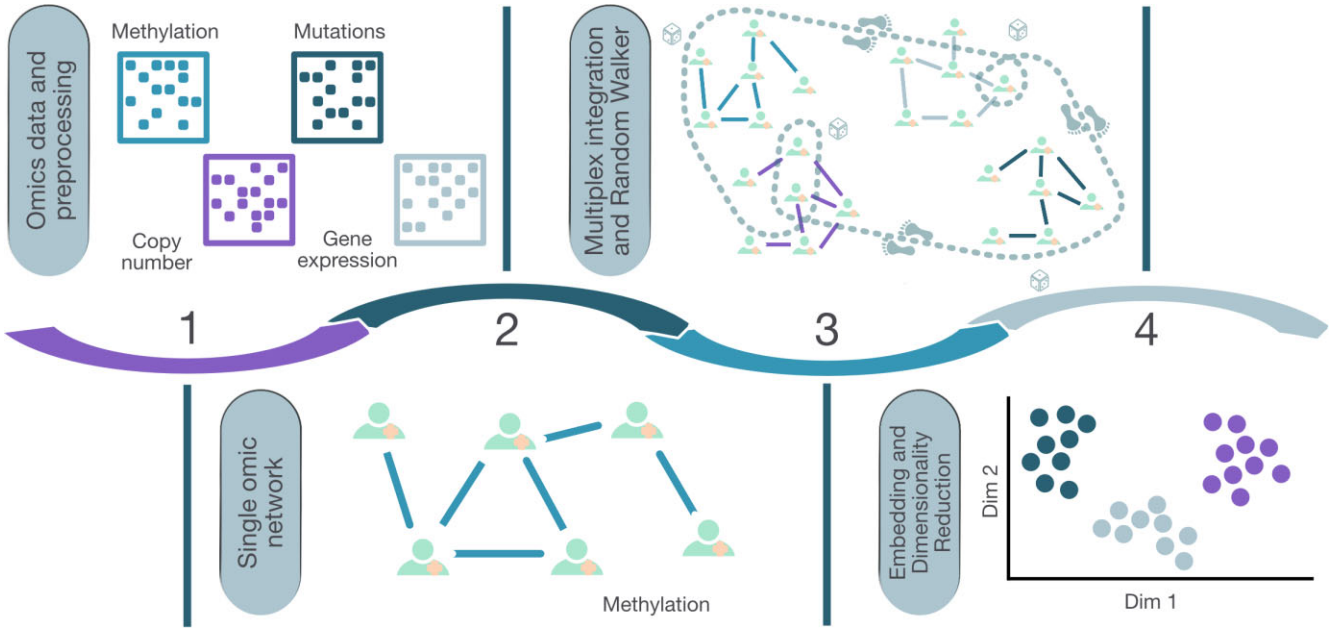

Figure 2. MoNETA workflow. MoNETA workflow for the analysis and visualization of multi-omics networks comprises four steps. The initial step involves data acquisition and pre-processing. The second step involves computing single omics networks to understand the underlying relationships between each omics data. This step is followed by multi-omics network integration, where individual networks are combined into a multiplex network. Random Walker is then applied to the multiplex network to identify nodes with similar attributes. The final step in the workflow involves embedding and dimensionality reduction of the data, which is necessary for visualizing high-dimensional data that highlights the underlying patterns and relationships in the data.

Thus, the final edge set E_M of the multiplex network is:

$$E_M = E_O \cup E_S \cup E_N$$

These inclusions enrich the network with comprehensive inter-omic relationships, offering a holistic view of the biological system. This approach defines the structural components of the multiplex graph, elucidating the intricate connections between nodes that represent the same biological entity across diverse omics layers.

Dimensionality reduction

The multiplex graph serves as a model representing the omics neighborhood of each entity across various omics. From this structure, the multi-omics neighborhood of each entity is derived by computing the probability distribution of reaching other nodes in the multi-omics network through a random walk process originating from the entity node. MoNETA employs a modified version of the Random Walk with Restart procedure defined in (47). This procedure involves a random walker associated with each entity, starting from the entity-associated node (seed node) in a randomly chosen layer. In each step, the walker randomly: (i) moves to a neighbor node on the same layer; (ii) shifts to the same node but on a different layer with a probability driven by an omics transition probability matrix $\Delta \in \mathbb{R}^{L \times L}$; (iii) restarts from the seed with probability r, randomly choosing among the different L layers, based on a vector of user-provided probabilities $\tau = [\tau_1, \ldots, \tau_L]$. Upon completion, the Random Walk with Restart algorithm computes a matrix $RW \in \mathbb{R}^{n \times n}$ containing the stationary probabilities distribution of visiting other nodes in the multiplex network starting from each entity associated node. MoNETA uses the $L \times L$ omics transition probability matrix Δ as input, guiding the random walk process towards biologically meaningful inter-layer passages and favoring associations between nodes in layers exhibiting stronger molecular interdependence (e.g. methylation-transcription or snv-cnv). Each element $\Delta_{i,j}$ represents a random walker's probability of moving from layer i to layer j. MoNETA allows a data-driven creation of the matrix Δ. The underlying concept for the automatic construction of Δ is to favor transitions between layers where nodes share similar neighbors compared to layers where the same node has vastly different neighbors between two layers. $\Delta_{i,j}$ is proportional to the Jaccard index between the sets of edges of layer i and layer j.

The RW matrix is then fed into a dimensionality reduction algorithm to yield the ultimate embedded matrix, $EM \in \mathbb{R}^{n \times d}$, where $d << n$, indicating the latent space dimensionality. Various dimensionality reduction algorithms are implemented, including Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), t-distributed Stochastic Neighbor Embedding (t-SNE) or the multiVERSE algorithm (13,50). The multiVERSE algorithm is a rapid and scalable method for learning node embeddings, applying the VERSE algorithm (50) to multiplex and multiplex-heterogeneous networks. Specifically, it tackles an optimization problem aimed at minimizing the Kullback–Leibler divergence between the graph similarity distribution and the embedding similarity distribution while updating the embeddings. If the multiVERSE algorithm is chosen as dimensionality reduction method, the resulting low dimensional matrix must be fed into one of the other three algorithms to visualize the nodes neighborhood on a two dimensional space.

MoNETA R package

MoNETA is distributed as an R package available on GitHub at https://github.com/BioinfoUninaScala/MoNETA. MoNETA provides a set of functions: (i) to import, normalize and filter omics data; (ii) to perform all the integration

and dimensionality reduction steps as defined above exploiting multi-core execution; (iii) to produce interactive visualizations of omics/integrated data and similarity networks amended with user-provided entity annotations. Finally, the package can be used throughout a Shiny-based web application providing an interactive GUI (Supplementary Figure S6) that executes the main steps of loading, integration, visualization, and clustering of multi-omics data. Additionally, it is comprised within a Docker container, accessible at https://hub.docker.com/r/bioinfouninascala/moneta.

Benchmarking

We compare the performance of MoNETA with MOFA (Multi-Omics Factor Analysis) and Seurat v4 (39) WNN (Weighted-nearest neighbor) method (38), using two key metrics: Normalized Mutual Information (NMI) score (51) and Accuracy score. The NMI score quantifies the similarity between predicted and actual labels while considering the data distribution. It normalizes the Mutual Information score, providing a robust measure of clustering accuracy independent of the dataset scale. A higher NMI score indicates a more accurate and consistent clustering. The predicted labels were assigned through the k-means clustering algorithm (52), with '*k*'—the number of clusters - equal to the number of actual labels. To the sake of repeatability, we fixed the seed, and the computation of the NMI score was repeated 100 times. The Accuracy score was computed by using the *k*-nearest neighbor (knn) classifier (53), included in the 'class' R package (54). This method assigns to each sample/cell the most representative label among its k-nearest neighbors. The size of the neighborhood was set equal to 20.

The benchmarking evaluations were performed on MOFA matrix, on MoNETA VERSE embedding, and on the PCA embedding of the weighted shared nearest neighbors (wsnn) graph. This choice was made because the metrics need to be computed in Euclidean space, which UMAP does not provide.

Results

Glioma sub-typing

We applied MoNETA to characterize glioma subtypes across various omics layers. Gliomas, the most aggressive form of brain tumors, are traditionally categorized based on molecular, histological and clinical attributes (55). Previous integrative supervised analyses (40) classified gliomas into seven classes, considering methylation, expression, mutational, and copy number profiles. IDH-mutant gliomas segregate into Codel, G-CIMP-high and G-CIMP-low classes. The G-CIMP-low subtype was identified through supervised gene expression and DNA methylation analysis. Similarly, IDH-wildtype gliomas are divided into four groups, including the PA-like subgroup discovered based on clinical grade and copy number profiles (40).

Our study analyzed copy number, gene expression profiling, and DNA methylation of 788 TCGA cases (Figure 3A). MoNETA was applied to obtain an integrated multi-omics similarity matrix, revealing the natural segregation of molecular subtypes. Samples were separated based on IDH status, and within the mutant samples, two main clusters emerged: Codel and non-Codel (G-CIMP-high and G-CIMP-low), evident in t-SNE, PCA, UMAP and VERSE spaces (Figure 3B). G-CIMP-low, associated with poor survival among IDH-mutant cases, clustered together, bordering the IDH-mutant cluster toward the wild-type, often linked to recurrent Glioblastoma (56) and subject to epigenetic reprogramming (57). Similarly, the PA-like group, exhibiting the best survival among gliomas (40), was positioned at the border of the IDH-wildtype group toward mutants or even clustered together in the VERSE and UMAP spaces. This analysis further confirms that these two groups genuinely embody glioma subtypes with 'outlier' characteristics, both from the clinical and molecular points of view. At the same time, an unique omics platform cannot characterize them. Overall, the biologically informed integrative analysis performed by MoNETA can uncover complex and non-trivial relationships between molecular layers. WNN in UMAP space successfully distinguished the major groups and subgroups, similar to MoNETA. However, it incorrectly clustered the LGm6-GBM group with the IDH mutant cluster (Figure 3C), potentially leading to inaccuracies in subsequent analyses. MOFA did not clearly separate these outlier subtypes. For example, the t-SNE reduction partially merged the IDH-wildtype group with the C-CIMP subgroup, resulting in erroneous clusterization (Figure 3D).

We evaluated the algorithms based on structural preservation using both NMI and accuracy scores over single-omics, MOFA, WNN and MoNETA embeddings. MOFA and WNN exhibited lower NMI scores, while MoNETA, demonstrated high accuracy and the highest NMI scores. This suggests that our model effectively integrates multi-omics data, extracting key features. Although methylation omics performed exceptionally well with the highest accuracy score, it did not yield the highest NMI, as the classification presented by (40) is predominantly driven by methylation. It is worth noticing the G-CIMP-low and the PA-like naturally emerge from the unbiased analysis with MONETA. Finally, we compared the execution times of the three considered algorithms, MoNETA outperforms MOFA by an order of magnitude, whereas WNN was the fastest method, since it was specifically developed for single-cells large datasets (Supplementary Table S2).

Single-cell data integration

We also assessed MoNETA on the task of integrating multi-omics single-cell data. Here, the information from diverse omics layers can be used to characterize distinct cell types and functional states better. The integrated low-dimensional latent space can serve this purpose. This scenario is characterized by a substantial volume of observations in the order of thousands, low measurement quality, and a significant number of missing variables. Network-based integration of such data poses computational challenges, particularly when contrasted with typical bulk sequencing experiments where the number of observations is usually hundreds. To evaluate MoNETA's performance in integrating multi-omic datasets of this nature, we used five distinct datasets: PBMC cells from various tissues assayed with bi-modal CITE-seq and tri-modal single-cell sequencing protocols TEA-seq and DOGMA-seq, CITE-seq dataset of human bine marrow (BM) mononuclear cells, and cells from the frontal cortex and hippocampus extracted from Mus musculus composing a hepta-modal dataset (Supplementary Table S1). The first dataset of from PBMC and lung cells (42) uses the CITE-seq protocol (58). MoNETA integration (Figure 4A, Supplementary Figure S1A), compared to single omics views, consolidates observable patterns in each layer, such as pDC, fibroblasts and epithelial cells from the

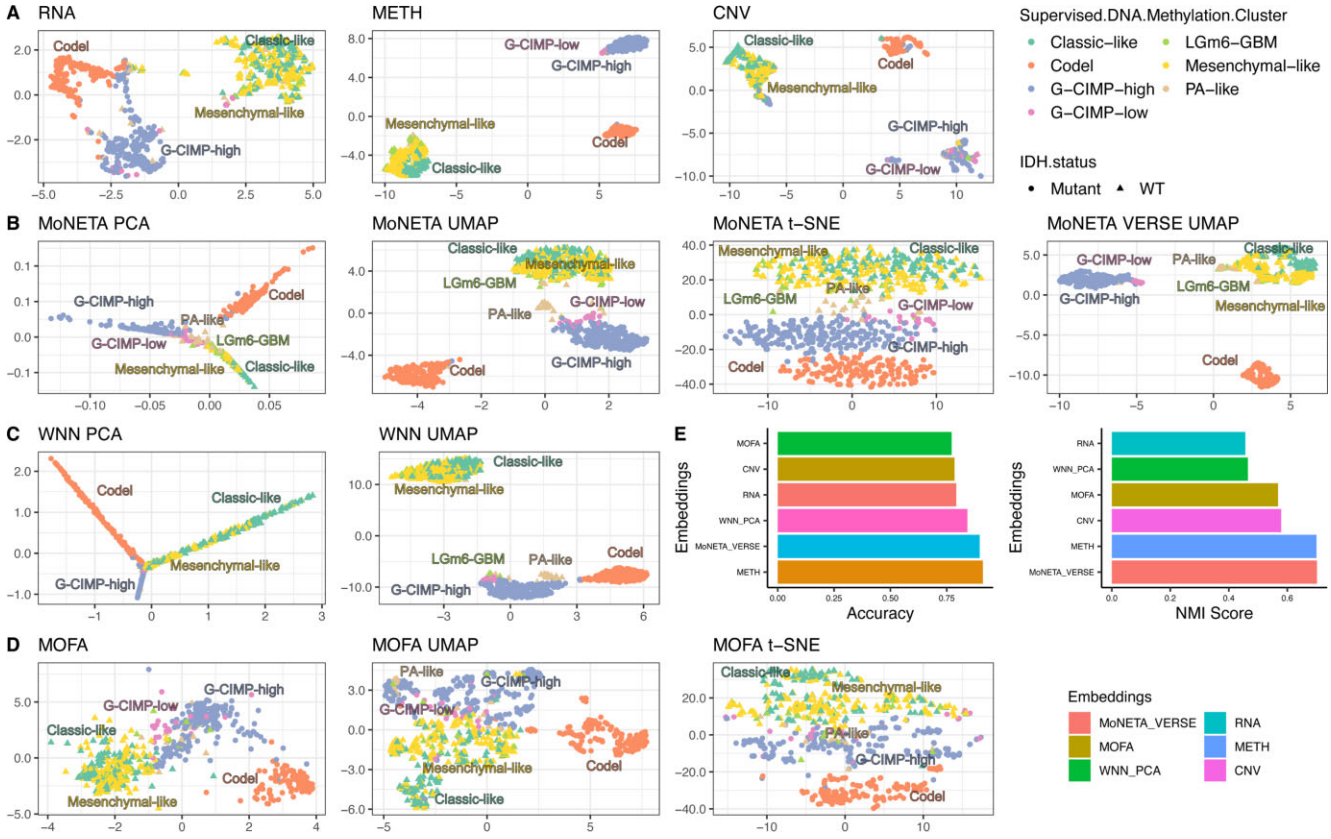

Figure 3. Glioma sub-typing. Embedding of glioma data. Colors show glioma subtypes identified in (40) while shapes are associated with IDH sample status. (**A**) Single-omics embeddings. (**B**) MoNETA embeddings. (**C**) WNN embeddings. (**D**) MOFA embeddings. (**E**) Comparison of MOFA, WNN and MoNETA embeddings using accuracy and NMI scores.

ADT layer. It also reveals patterns not evident in either layer, such as the separation of T Treg and T CD EM from TCD4 cluster. This separation is not evaluable in MOFA projections (Supplementary Figure S1B), where T Treg, T CD EM and T Prolif mix together, as do T CD4 and T CD8. In contrast, the WNN UMAP shows this differentiation more clearly, though the cell types are still very close in space, which could lead to errors in clustering analysis (Supplementary Figure S1C).

The second dataset is a bi-modal focusing on human bone marrow (BM) mononuclear cells categorized into 27 cell types (43). All the integration methods compared were able to distinguish most of the cell types into distinct clusters, primarily relying on ADT layer projection (Supplementary Figure S2). Notably, MoNETA VERSE UMAP and WNN UMAP more accurately identified CD8 Effectors from CD8 Memory, whereas MOFA tended to merge these two groups. In general the clusters are more distinct and compact, with a better differentiation and less overlap between different cell types.

The third dataset, from (44), involves PBMC cells from human blood using the tri-modal TEA-seq protocol. The integrated map reveals cell groups that are only observable in specific layers, such as the distinct separation of natural killer cells from other groups. While naive and activated B cells are clustered together in the integrated map, they show significant distinctions compared to their distribution in individual layers (Figure 4B). These distinctions are also observable in MOFA and WNN projections (Supplementary Figure S3). However, in WNN, T CD8 and T CD4 macro groups are clustered together, whereas in other projections, they are well separated. Additionally, the MoNETA VERSE UMAP space divides T CD8 Naive and Mono CD14 into two smaller groups, suggesting that further analysis could be conducted to distinguish these cell types more specifically.

The fourth dataset, another tri-modal on human PBMCs, was processed using DOGMA-seq (45). This dataset includes 13 763 cells categorized into 27 cell types, with data on 36 495 RNA levels, 68 963 accessible chromatin ATAC-seq peaks, and 210 surface protein abundances. All projections successfully identified clusters of NK and B cells, clearly visible in the ADT layer and slightly in the RNA layer (Supplementary Figure S4). All the clusters were homogeneously separated by all methods. MoNETA VERSE UMAP space further divided two clusters, one containing CD4 Naive, CD4 TCM, Eryth and Treg, and another containing CD8 TEM and CD8 Naive, into two subclusters each. This division likely reflects differences in the functional collaboration and cell communication strategies within these subgroups.

The Paired-Tag/seq datasets (46,59) posed the greatest challenge as they comprise 7556 cells with seven omics layers. MOFA and WNN can handle it after adding missing value columns to the DNA omics matrices. MoNETA, on the other hand, can natively manage these complexities. Some layers, such as H3K27me3, H3K4me3, H3K9me3 and ATAC, do not show a clear separation of the four cell types (Figure 4C). In contrast, data from H3K27ac, H3K4me1 and RNA can distinguish these types effectively. This distinction is also clearly visible in MoNETA VERSE UMAP space, which further divides these groups into smaller subgroups,

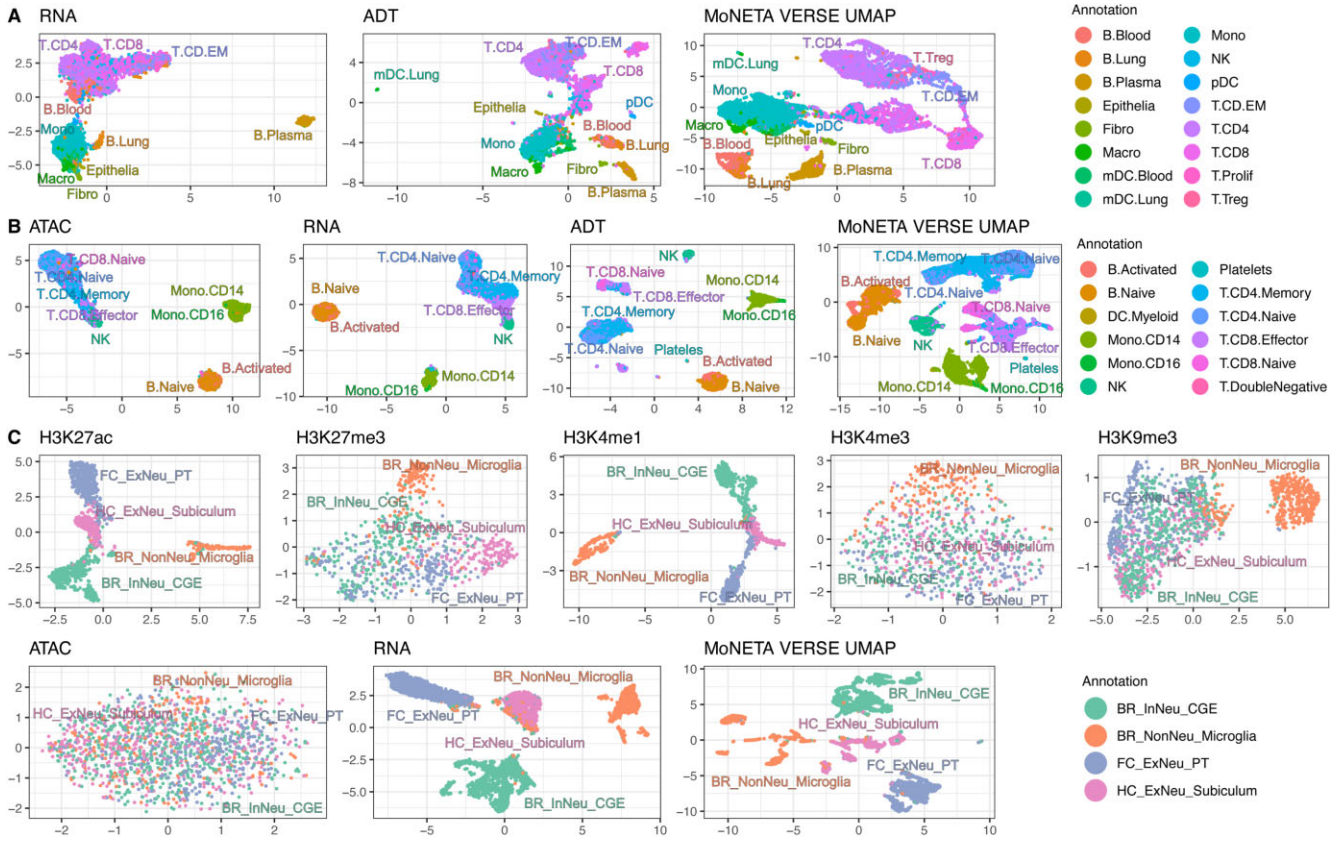

Figure 4. Single-cell LUNG-CITE, PBMC-TEA and paired clustering. Plots comparing the projections of cells obtained using data from individual omics layers with the MoNETA multi-omics integration for: (**A**) lung derived single-cell PBMC bi-modal assay; (**B**) blood-derived PBMC tri-modal assay; (**C**) adult mouse-brain derived hepta-modal single-cell Paired assay. Colors show cell types as identified in their annotation.

suggesting internal divisions that could correspond to different subtypes. Although MOFA and WNN successfully separated the four main groups, they could not identify these finer subgroups, highlighting MoNETA superior ability to discover subtypes (Supplementary Figure S5A–C). MoNETA outperformed both methods in terms of accuracy and NMI score (Supplementary Figure S5D).

Discussion

In the era of expanding multi-omics data, the need for integrative models capable of deciphering complex relationships among diverse biological layers has never been more pressing. MoNETA, an integrative multi-omics model, exploits a multiplex network representation of the similarity between entities at single layers. The data is then embedded using neighboring information with a random walk procedure. We have evaluated its performance on two different tasks: the stratification of cancer using multiple molecular layers such as copy number, gene expression, and DNA methylation. Compared to state-of-the-art approaches such as MOFA (20,21) and WNN (38), our model exhibited better performances regarding the compactness of the clusters measured by the accuracy score and similarity with previously established semi-supervised classification. We have shown that the glioma subtypes described in (40) naturally emerge with our unbiased approach. This is also true for the two classes, G-CIMP-low and PA-like, which were originally derived in a supervised procedure accounting for variation of gene expression and methylation (G-CIMP-low) and methylation and clinical grade and copy number (PA-like). Moreover, MoNETA is scalable enough to be applied to single-cell datasets. We showcased how using multiple molecular layers in a single-cell can be employed to characterize immune cell types better. We have seen that NK and B cells are better clustered in the integrative embedding rather than in single layers. Missing data often affect single-cell datasets, requiring state-of-the-art methods to account for the possibility that not all entities have corresponding data in every omics layer. We demonstrated that MoNETA can natively overcome this issue, successfully identifying granular subtypes in a hepta-modal dataset. Finally, it is essential to point out the limitations of our method. One limitation of MoNETA, which is shared with similar approaches, is the limited explainability due to the loss of information about the single features driving the clustering in the embedding space. This can be overcome by post-hoc differential analysis between the clusters in the combined space. Another limitation is that the network building is fixed on a similarity network; more efforts will be dedicated in the future to include prior network-encoded knowledge such as pathways, gene regulatory networks, or protein-protein interaction networks.

Data availability

MoNETA is accessible as an R package on GitHub at https://github.com/BioinfoUninaScala/MoNETA. Within this package, the TCGA glioma dataset, integral to our analyses, is readily available. A stable release was deposited on

Zenodo at https://zenodo.org/records/13864141. Our package offers a user-friendly Shiny-based web application featuring an interactive GUI. The application is encapsulated within a Docker container, accessible at https://hub.docker.com/r/bioinfouninascala/moneta.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

Research reported in this publication was performed in part at the Biostatistics and Bioinformatics Shared Resource of the Sylvester Comprehensive Cancer Center at the University of Miami, RRID: SCR_022890, which is supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under award number P30CA240139. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Funding

EU funding within the NextGeneration EU: MUR PNRR National Center for Gene Therapy and Drugs based on RNA Technology [CN_00000041]. EU funding within the Next Generation EU, Mission 4 Component 1, CUP E53D23004630001.

Conflict of interest statement

None declared.

References

1. Civelek,M. and Lusis,A.J. (2014) Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.*, **15**, 34–48.
2. Aaltonen,L.A., Abascal,F., Abeshouse,A., Aburatani,H., Adams,D.J., Agrawal,N., Ahn,K.S., Ahn,S.M., Aikata,H., Akbani,R., *et al.* (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.
3. Es,L. (2001) Initial sequencing and analysis of the human genome. *nature*, **409**, 860–921.
4. Edwards,N.J., Oberti,M., Thangudu,R.R., Cai,S., McGarvey,P.B., Jacob,S., Madhavan,S. and Ketchum,K.A. (2015) The CPTAC data portal: a resource for cancer proteomics research. *J. Proteome Res.*, **14**, 2707–2713.
5. O'Donnell,S.T., Ross,R.P. and Stanton,C. (2020) The progress of multi-omics technologies: determining function in lactic acid bacteria using a systems level approach. *Front. Microbiol.*, **10**, 3084.
6. Cai,Z., Poulos,R.C., Liu,J. and Zhong,Q. (2022) Machine learning for multi-omics data integration in cancer. *Iscience*, **25**, 103798.
7. Gliozzo,J., Mesiti,M., Notaro,M., Petrini,A., Patak,A., Puertas-Gallardo,A., Paccanaro,A., Valentini,G. and Casiraghi,E. (2022) Heterogeneous data integration methods for patient similarity networks. *Brief. Bioinform.*, **23**, bbac207.
8. Subramanian,I., Verma,S., Kumar,S., Jere,A. and Anamika,K. (2020) Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insight*, **14**, 1177932219899051.
9. Reel,P.S., Reel,S., Pearson,E., Trucco,E. and Jefferson,E. (2021) Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.*, **49**, 107739.
10. Zhang,B., Wang,J., Wang,X., Zhu,J., Liu,Q., Shi,Z., Chambers,M.C., Zimmerman,L.J., Shaddox,K.F., Kim,S., *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.
11. Navaz,A.N., El-Kassabi,H.T., Serhani,M.A., Oulhaj,A. and Khalil,K. (2022) A novel patient similarity network (PSN) framework based on multi-model deep learning for precision medicine. *J. Pers. Med.*, **12**, 768.
12. Grover,A. and Leskovec,J. (2016) node2vec: Scalable Feature Learning for Networks. *KDD*, **2016**, 855–864.
13. Pio-Lopez,L., Valdeolivas,A., Tichit,L., Remy,É. and Baudot,A. (2021) MultiVERSE: a multiplex and multiplex-heterogeneous network embedding approach. *Sci. Rep.*, **11**, 8794.
14. Stetson,L.C., Pearl,T., Chen,Y. and Barnholtz-Sloan,J.S. (2014) Computational identification of multi-omic correlates of anticancer therapeutic response. *BMC Genomics*, **15**, S2.
15. Auslander,N., Yizhak,K., Weinstock,A., Budhu,A., Tang,W., Wang,X.W., Ambs,S. and Ruppin,E. (2016) A joint analysis of transcriptomic and metabolomic data uncovers enhanced enzyme-metabolite coupling in breast cancer. *Sci. Rep.*, **6**, 29662.
16. Yuan,Y., Savage,R.S. and Markowetz,F. (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.
17. Zhang,S., Liu,C.-C., Li,W., Shen,H., Laird,P.W. and Zhou,X.J. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.
18. Yang,Z. and Michailidis,G. (2016) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, **32**, 1–8.
19. Gao,Y.-L., Hou,M.-X., Liu,J.-X. and Kong,X.-Z. (2019) An integrated graph regularized non-negative matrix factorization model for gene co-expression network analysis. *IEEE Access*, **7**, 126594–126602.
20. Argelaguet,R., Velten,B., Arnol,D., Dietrich,S., Zenz,T., Marioni,J.C., Buettner,F., Huber,W. and Stegle,O. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.
21. Argelaguet,R., Arnol,D., Bredikhin,D., Deloro,Y., Velten,B., Marioni,J.C. and Stegle,O. (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biol.*, **21**, 111.
22. Ciaramella,A., Nardone,D. and Staiano,A. (2020) Data integration by fuzzy similarity-based hierarchical clustering. *BMC Bioinformatics*, **21**(Suppl. 10), 350.
23. Drăghici,S. and Potter,R.B. (2003) Predicting HIV drug resistance with neural networks. *Bioinformatics*, **19**, 98–107.
24. Hoadley,K.A., Yau,C., Wolf,D.M., Cherniack,A.D., Tamborero,D., Ng,S., Leiserson,M.D., Niu,B., McLellan,M.D., Uzunangelov,V., *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**, 929–944.
25. Sun,D., Wang,M. and Li,A. (2018) A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Computat. Biol. Bioinform.*, **16**, 841–850.
26. Phan,J.H., Hoffman,R., Kothari,S., Wu,P.-Y. and Wang,M.D. (2016) In: *Integration of Multi-modal Biomedical Data to Predict Cancer Grade and Patient Survival*. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 577–580.
27. De Tayrac,M., Lê,S., Aubry,M., Mosser,J. and Husson,F. (2009) Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, **10**, 32.
28. Nguyen,H., Shrestha,S., Draghici,S. and Nguyen,T. (2019) PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, **35**, 2843–2846.
29. Sienkiewicz,K., Chen,J., Chatrath,A., Lawson,J.T., Sheffield,N.C., Zhang,L. and Ratan,A. (2022) Detecting molecular subtypes from multi-omics datasets using SUMO. *Cell Rep. Methods*, **2**, 100152.

30. Shin,H., Lisewski,A.M. and Lichtarge,O. (2007) Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics*, **23**, 3217–3224.
31. Tsuda,K., Shin,H. and Schölkopf,B. (2005) Fast protein classification with multiple networks. *Bioinformatics*, **21**(Suppl. 2), ii59–ii65.
32. Chen,S., Ma,B. and Zhang,K. (2009) On the similarity metric and the distance metric. *Theor. Comput. Sci.*, **410**, 2365–2376.
33. Rappoport,N. and Shamir,R. (2019) NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, **35**, 3348–3356.
34. Yan,K.K., Zhao,H. and Pang,H. (2017) A comparison of graph-and kernel-based–omics data integration algorithms for classifying complex traits. *BMC Bioinformatics*, **18**, 539.
35. Lanckriet,G.R., De Bie,T., Cristianini,N., Jordan,M.I. and Noble,W.S. (2004) A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2626–2635.
36. Seoane,J.A., Day,I.N., Gaunt,T.R. and Campbell,C. (2014) A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, **30**, 838–845.
37. Wu,C.-C., Asgharzadeh,S., Triche,T.J. and D'Argenio,D.Z. (2010) Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. *Bioinformatics*, **26**, 807–813.
38. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
39. Hao,Y., Stuart,T., Kowalski,M.H., Choudhary,S., Hoffman,P., Hartman,A., Srivastava,A., Molla,G., Madad,S., Fernandez-Granda,C., *et al.* (2024) Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.*, **42**, 293–304.
40. Ceccarelli,M., Barthel,F.P., Malta,T.M., Sabedot,T.S., Salama,S.R., Murray,B.A., Morozova,O., Newton,Y., Radenbaugh,A., Pagnotta,S.M. and et,al. (2016) Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, **164**, 550–563.
41. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I., *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
42. Buus,T.B., Herrera,A., Ivanova,E., Mimitou,E., Cheng,A., Herati,R.S., Papagiannakopoulos,T., Smibert,P., Odum,N. and Koralov,S.B. (2021) Improving oligo-conjugated antibody signal in multimodal single-cell analysis. *Elife*, **10**, e61973.
43. Stuart,T., Srivastava,A., Madad,S., Lareau,C.A. and Satija,R. (2021) Single-cell chromatin state analysis with Signac. *Nat. Methods*, **18**, 1333–1341.
44. Swanson,E., Lord,C., Reading,J., Heubeck,A.T., Genge,P.C., Thomson,Z., Weiss,M.D., Li,X.-j., Savage,A.K., Green,R.R., *et al.* (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife*, **10**, e63632.
45. Mimitou,E.P., Lareau,C.A., Chen,K.Y., Zorzetto-Fernandes,A.L., Hao,Y., Takeshima,Y., Luo,W., Huang,T.-S., Yeung,B.Z., Papalexi,E., *et al.* (2021) Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.*, **39**, 1246–1258.
46. Zhu,C., Zhang,Y., Li,Y.E., Lucero,J., Behrens,M.M. and Ren,B. (2021) Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods*, **18**, 283–292.
47. Valdeolivas,A., Tichit,L., Navarro,C., Perrin,S., Odelin,G., Levy,N., Cau,P., Remy,E. and Baudot,A. (2019) Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, **35**, 497–505.
48. Yianilos,P.N. (1993) Data structures and algorithms for nearest neighbor search in general metric spaces, Vol. 93. In: *Soda*. Society for Industrial and Applied Mathematics, pp. 311–321.
49. Anava,O. and Levy,K.Y. (2017) k*-Nearest neighbors: from global to local. arXiv doi: https://arxiv.org/abs/1701.07266, 25 January 2017, preprint: not peer reviewed.
50. Tsitsulin,A., Mottin,D., Karras,P. and Müller,E. (2018) Verse: Versatile graph embeddings from similarity measures. In: *Proceedings of the 2018 World Wide Web Conference*. pp. 539–548.
51. Strehl,A. and Ghosh,J. (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, **3**, 583–617.
52. Wu,J. (2012) In: *Advances in K-means Clustering: A Data Mining Thinking*. Springer Science & Business Media.
53. Mucherino,A., Papajorgji,P.J. and Pardalos,P.M. (2009) In: *K-nearest Neighbor Classification*. Springer, pp. 83–106.
54. Venables,W.N. and Ripley,B.D. (2002) Modern Applied Statistics with S. 4th edn., Springer, NY.
55. Louis,D.N., Perry,A., Wesseling,P., Brat,D.J., Cree,I.A., Figarella-Branger,D., Hawkins,C., Ng,H., Pfister,S.M., Reifenberger,G., *et al.* (2021) The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology*, **23**, 1231–1251.
56. Malta,T.M., de Souza,C.F., Sabedot,T.S., Silva,T.C., Mosella,M.S., Kalkanis,S.N., Snyder,J., Castro,A. V.B. and Noushmehr,H. (2018) Glioma CpG island methylator phenotype (G-CIMP): biological and clinical implications. *Neuro-oncology*, **20**, 608–620.
57. Mazor,T., Chesnelong,C., Pankov,A., Jalbert,L.E., Hong,C., Hayes,J., Smirnov,I.V., Marshall,R., Souza,C.F., Shen,Y., *et al.* (2017) Clonal expansion and epigenetic reprogramming following deletion or amplification of mutant IDH1. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 10743–10748.
58. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.
59. Zhu,C., Yu,M., Huang,H., Juric,I., Abnousi,A., Hu,R., Lucero,J., Behrens,M.M., Hu,M. and Ren,B. (2019) An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.*, **26**, 1063–1070.

Received: March 25, 2024. **Revised:** July 19, 2024. **Editorial Decision:** September 27, 2024. **Accepted:** October 4, 2024