DeepMAPS: Single-cell biological network inference using heterogeneous graph transformer

Anjun Ma[1,2,*], Xiaoying Wang[3,*], Cankun Wang[1], Jingxian Li[3], Tong Xiao[2], Juexing Wang[4, 5], Yang Li[1], Yuntao Liu[3], Yuzhou Chang[1,2], Duolin Wang[4,5], Yuexu Jiang[4,5], Jinpu Li[4,5], Li Su[5,6], Shaopeng Gu[1], Gang Xin[2], Zihai Li[2], Bingqiang Liu[3,$], Dong Xu[4,5,$], Qin Ma[1,2,$]

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA,

[2] Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center, Columbus, OH, 43210, USA,

[3] School of Mathematics, Shandong University, Jinan, Shandong, 250100, China,

[4] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

[5] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA.

[6] MU Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

* These authors contributed equally to the paper as first authors

[$] To whom correspondence should be addressed

Abstract

We present DeepMAPS (Deep learning-based Multi-omics Analysis Platform for Single-cell data) for biological network inference from single-cell multi-omics (scMulti-omics). DeepMAPS includes both cells and genes in a heterogeneous graph to simultaneously infer cell-cell, cell-gene, and gene-gene relations. The multi-head attention mechanism in a graph transformer considers the heterogeneous relation among cells and genes within both local and global context, making DeepMAPS robust to data noise and scale. We benchmarked DeepMAPS on 18 scMulti-omics datasets for cell clustering and biological network inference, and the results showed that our method outperformed various existing tools. We further applied DeepMAPS on lung tumor leukocyte CITE-seq data and matched diffuse small lymphocytic lymphoma scRNA-seq and scATAC-seq data. In both cases, DeepMAPS showed competitive performance in cell clustering and predicted biologically meaningful cell-cell communication pathways based on the inferred gene networks. Note that we deployed a webserver using DeepMAPS implementation equipped with multiple functions and visualizations to improve the feasibility and reproducibility of scMulti-omics data analysis. Overall, DeepMAPS represents a heterogeneous graph transformer for single-cell study and may benefit the use of scMulti-omics data in various biological systems.

Keywords: Single-cell multi-omics, graph neural network, heterogeneous graph transformer, multi-head attention mechanism, biological network, webserver

Main

Single-cell sequencing, such as single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq), reshapes the investigation of cellular heterogeneity and yields novel insights in neuroscience, cancer biology, immuno-oncology, and therapeutic responsiveness[1, 2]. However, an individual single-cell modality only reflects a snapshot of genetic features and partially depicts the peculiarity of cells, leading to characterization biases in complex biological systems[2, 3]. Single-cell multi-omics (scMulti-omics) allows for the generation and quantification of multiple modalities simultaneously to fully capture the intricacy of complex molecular mechanisms and cellular heterogeneity. Such analyses advance various biological studies when paired with robust computational analysis methods[4].

The existing tools for integrative analyses of scMulti-omics data, e.g., Seurat[5], MOFA+[6], Harmony[7], and totalVI[8], reliably predict cell types and states, remove batch effects, and reveal relationships or alignments among various modalities. However, most existing methods do not explicitly consider in their models the crosstalk between cells or different molecular modalities. Hence, these models cannot infer the active biological networks of diverse cell types and have limited power to elucidate the response of these complex networks to external stimuli in specific cell types.

Recently, graph neural networks (GNN) have shown a unique strength in learning low-dimensional representations of individual cells by propagating neighbor cell features and constructing cell-cell relations in a global cell graph[9-12]. For example, our in-house tool scGNN, a novel GNN model, has demonstrated superior cell clustering and gene imputation performance based on large-scale scRNA-seq data[13]. Furthermore, a heterogeneous graph with different types of nodes and edges has been widely used to model a multi-relational knowledge graph[14]. It provides a natural representation framework for integrating scMulti-omics data and learning the underlying cell-type-specific biological networks. Moreover, a recent development in the attention mechanism for modeling and integrating heterogeneous relationships has made deep learning models explainable and enabled the inference of cell-type-specific biological networks[14, 15].

To this end, we developed **DeepMAPS** (**Deep** learning-based **M**ulti-omics **A**nalysis **P**latform for **S**ingle-cell data), a heterogeneous graph transformer framework for cell-type-specific biological network inference from scMulti-omics data. DeepMAPS formulates high-level representations of relations between cells and genes in a heterogeneous graph, with cells and genes as the two disjoint node sets in this graph. Projecting the features of genes and cells into the same latent space is an effective way to harmonize the imbalance of different batches and create a solid foundation for cell clustering (i.e., node clustering) as well as the prediction of cell-gene and gene-gene relations in a specific cell cluster (i.e., link prediction). Most importantly, the attention mechanism in this transformer model

enhances biological interpretability and enables the identification of active gene modules in each cell cluster. Overall, DeepMAPS is an end-to-end and hypotheses-free framework that provides the first deep learning tool to infer cell-type-specific biological networks from scMulti-omics data. We deployed DeepMAPS as a code-free web portal along with Docker to ensure the reproducibility of scMulti-omics data analysis and lessen the programming burden for biologists who lack sufficient computational skills or resources (**Fig. 1**).

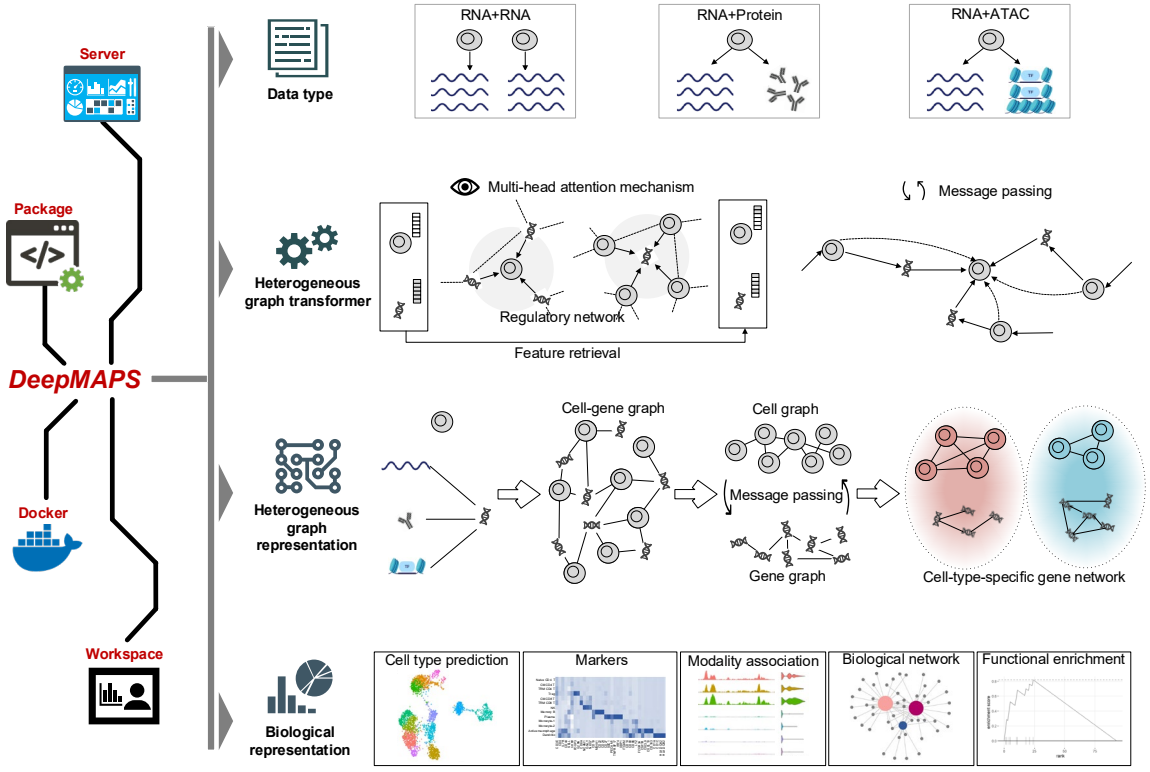

Fig. 1. DeepMAPS is a Deep learning-based Multi-omics Analysis Platform for Single-cell data. It allows for the joint analysis of multiple scRNA-seq, CITE-seq (matched RNA and protein profilings), and matched single-cell RNA and ATAC-seq data sets. The core method includes the representation of cell-gene relations via a heterogeneous graph and a transformer with a graph attention mechanism. DeepMAPS provides interactive and interpretable graphical representations to deliver cell clusters and various cell-type-specific biological networks based on modality types. DeepMAPS is provided as a web portal to ensure robustness and reproducibility, along with a docker container. Workspace is provided for job saving and retrieval. DeepMAPS also supports diverse interpretations, including but not limited to joint cell clustering, marker identification, modality associations, cell-type-specific biological network inference, and functional enrichment.

Results

Overview of DeepMAPS

There are five steps in the DeepMAPS pipeline to accomplish a joint analysis of scMulti-omics data (**Fig. 2** and **Methods**). (*i*) Data are preprocessed by removing low-quality cells

and modalities,= and then different normalization methods are applied according to the specific data types. (*ii*) An integrated cell-gene matrix is generated to represent the combined activity of each gene in each cell. Different data integration methods are applied for different scMulti-omics data types. (*iii*) A heterogeneous graph transformer (**HGT**) model is built to jointly learn the low-dimensional embedding for cells and genes and generate an attention score to indicate the importance of a gene to a cell. (*iv*) Cell clustering and functional gene modules are predicted based on HGT-learned embeddings and attention scores. (*v*) Diverse biological networks, e.g., gene regulatory networks and gene association networks, are inferred for each cell type.

To learn joint representations of cells and genes, we first generate a cell-gene matrix integrating all heterogeneous information of the input scMulti-omics data. A heterogeneous graph with cell nodes and gene nodes is then constructed, wherein a cell-gene edge represents the integrated gene activity score in the matrix, and the initial embedding of each node is learned from the gene-cell integrated matrix via a two-layer GNN graph autoencoder. The entire heterogeneous graph is then sent to a graph autoencoder to learn relations between the cells and genes and update the embedding of each node. Here, DeepMAPS adopts a heterogeneous multi-head attention mechanism to model the overall topological information (global relationships) and neighbor message passing (local relationships) on the heterogeneous graph.

In each layer, each node (either a cell or a gene) is considered a target and its 1-hop neighbors as sources. DeepMAPS evaluates the importance of its neighbor nodes and the amount of information that can be passed to the target based on the synergy of node embedding. As a result, cells and genes with high positively correlated embeddings are more likely to pass messages within each other, thus maximizing the similarity and disagreement of the embeddings. To make the unsupervised training process feasible on a large heterogeneous graph, DeepMAPS is first performed on 50 subgraphs sampled from the heterogeneous graph, covering a minimum of 50% of all nodes to train for the shared parameters between different nodes, information which is later used for testing of the whole graph. As an important training outcome, an attention score is given to represent the importance of a gene to a cell. A high attention score for a gene to a cell implies that the gene is relatively important for defining cell identity and characterizing cell heterogeneity. This discrimination allows for the construction of reliable gene association networks in each cell cluster as the final output of DeepMAPS. We then build a Steiner Forest Problem (SFP) model[16] using to identify genes with higher attention scores and similar embedding features unique to a cell cluster. The gene-gene and gene-cell relations in the optimized solution of the SFP model mirror the co-expression relations between genes and the specificity of gene attention to a cell type. A gene network established from SFP contains genes that are of the most important in characterizing the identity of that cell

cluster based on their attention scores and embedding similarities, and these genes are considered to be cell-type-active.

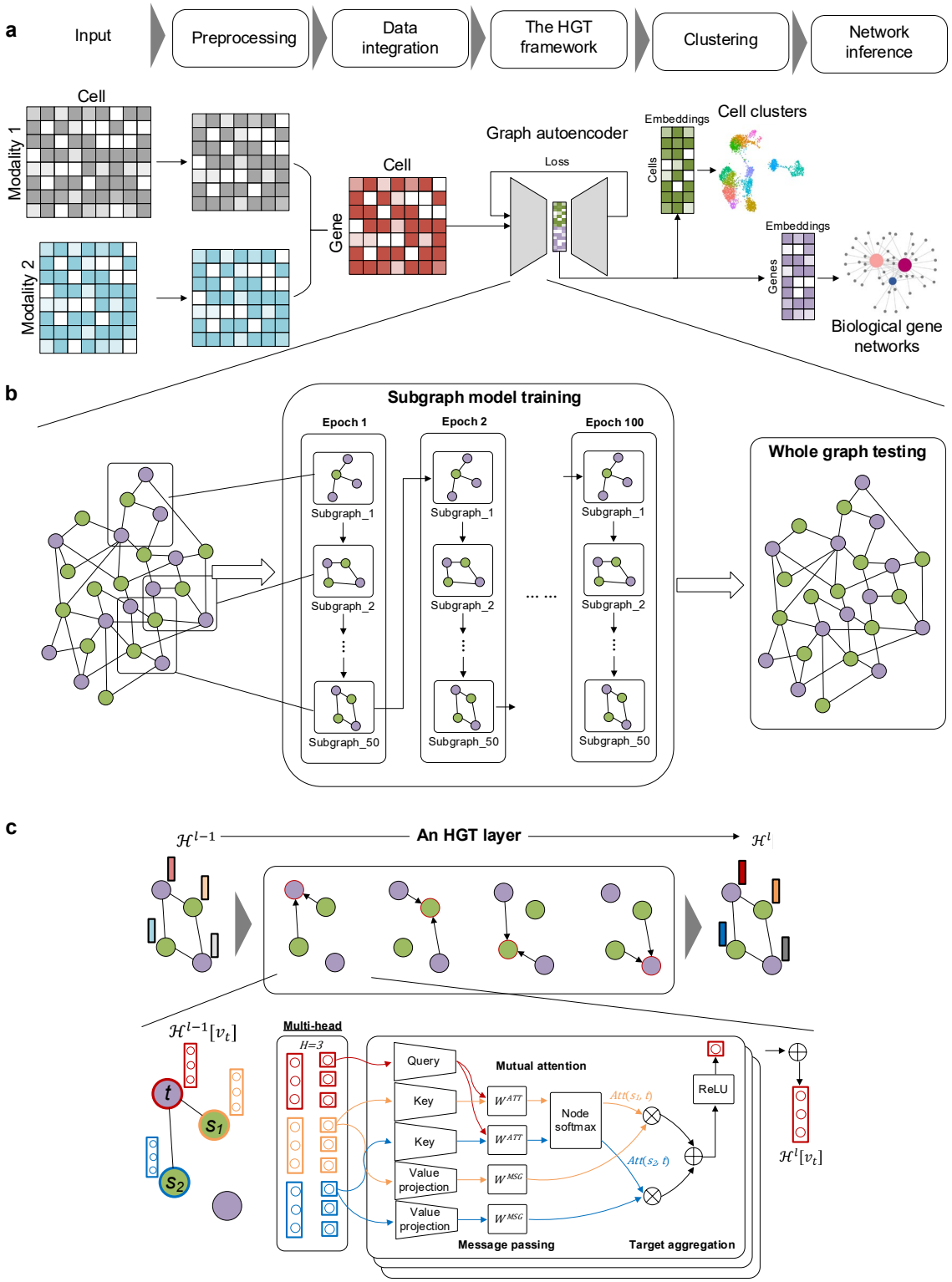

Fig. 2. DeepMAPS framework. (**a**) The five major steps and graphical illustration of DeepMAPS.

(**b**) Relations in the heterogeneous graph of cells and genes are learned in an HGT-based graph autoencoder. The hyperparameters are first trained in 50 subgraphs and 100 epochs and then applied to test the whole graph. (**c**) For each subgraph training and whole graph testing, multiple HGT layers are applied. An example with two cells (purple) and two genes (green) is shown for one HGT layer. The embeddings of the target cell (red) and two neighbor genes (orange and blue) are separated evenly into three heads. For each head, the HGT layer calculates the attention score of neighbor genes to the target cell and updates the target's embedding of this head. The entire target cell embedding is updated by concatenating all three heads to complete one HGT layer (Method).

DeepMAPS achieves superior performances in joint cell clustering and biological network inference from scMulti-omics data

We evaluated DeepMAPS on 18 scMulti-omics datasets, including two multiple scRNA-seq data sets (Data 1-2), eight CITE-seq datasets (Data 3-10), and eight matched scRNA-seq and scATAC-seq (scRNA-ATAC-seq) datasets measured from the same cell (Data 11-18) (**Supplementary Data 1**). Specifically, Data 1–2 and 17–18 have benchmark annotations provided in the original manuscripts. These data cover a number of cells ranging from 549 to 30,672; an average read depth (considering scRNA-seq data only) ranging from 2,933 to 645,526; and a zero-expression rate (considering scRNA-seq data only) from 71% to 97% (**Fig. 3a**).

We compared DeepMAPS with four benchmarking tools (Seurat, MOFA+, TotalVI, and Harmony (**Methods**)) using the default settings in terms of the Average Silhouette Width (ASW) (**Fig. 3a**) and Adjusted Rand Index (ARI) (**Fig. 3b**) to evaluate the performance of joint cell clustering for all three scMulti-omics data types. DeepMAPS was trained by each scMulti-omics data type and each dataset in an unsupervised way. One set of parameters was chosen as the default for all datasets within the same data type based on the grid optimization of hyper-parameter combinations (**Supplementary Data 2-4**). The results showed that, in all three scMulti-omics data scenarios, DeepMAPS distinctly outperformed the other tools in most cases. Note that for Data 2 and 18, though DeepMAPS did not achieve the best ASW compared to the other tools, its performance regarding ARI comparison to the benchmark label was the highest.

We show the UMAP results for the cell clustering of Data 17, a cancer cell line (n=549) with benchmarked cell labels for scRNA-ATAC-seq data (**Fig. 3c**). Compared with the original cell line labels, DeepMAPS was the only tool that accurately separated each cell type with minimum mismatches (ARI=0.97). In contrast, Seurat (ARI=0.88) and MOFA+ (ARI=0.79) mistakenly divided the PDX1 or PDX2 population into two clusters and included more mismatches. DeepMAPS also produced better UMAP visualization than other tools for datasets without benchmark labels, without mixing clusters or separating random cells (Fig. 3d).

In order to evaluate the robustness of DeepMAPS, we performed a leave-out test on benchmark datasets (**Supplementary Data 5**). For data with benchmark labels, we first filtered cells by removing a cluster of cells based on benchmark labels and then performed analysis using DeepMAPS and other tools; for other datasets, we removed cells based on clusters identified by each tool. The results showed that DeepMAPS achieved better performance compared to all other benchmarking methods (**Fig. 3e)**. Another test was performed on a series of data simulated for different read depth rates (**Fig. 3f**, **Supplementary Data 6**). For each testing dataset, the clustering results of DeepMAPS were consistent with high AWS or ARI scores, indicating that the message passing and attention mechanism used by DeepMAPS helped maintain cell-cell relations and tolerance to varying read depths.

We further evaluated the two kinds of biological networks that DeepMAPS can produce (**Supplementary Data 7**). For the gene association network for all scMulti-omics data types, we used the centrality scores and enriched pathways to compare DeepMAPS with IRIS3[11]. IRIS3 is an in-house tool for identifying cell-type-specific regulons and constructing gene regulatory networks from scRNA-seq data. It has superior performance over other public tools, such as SCENIC[10]. We also compared our results to the co-expression network constructed from the whole gene list with significance cutoffs, which is widely used in single-cell studies. Both the average closeness centrality and eigenvector centrality scores of networks constructed by DeepMAPS showed significantly higher scores than the other two methods when using all 18 benchmark datasets (**Fig. 3g-h**). Moreover, for the gene regulatory network constructed from scRNA-ATAC-seq data, we evaluated the number of significantly enriched pathways in a TF-regulon and a cell-type-specific regulon (**Fig. 3i-j**). The results indicated that DeepMAPS could construct more compatible and biologically reasonable gene networks for each cell type and outperform the other methods.

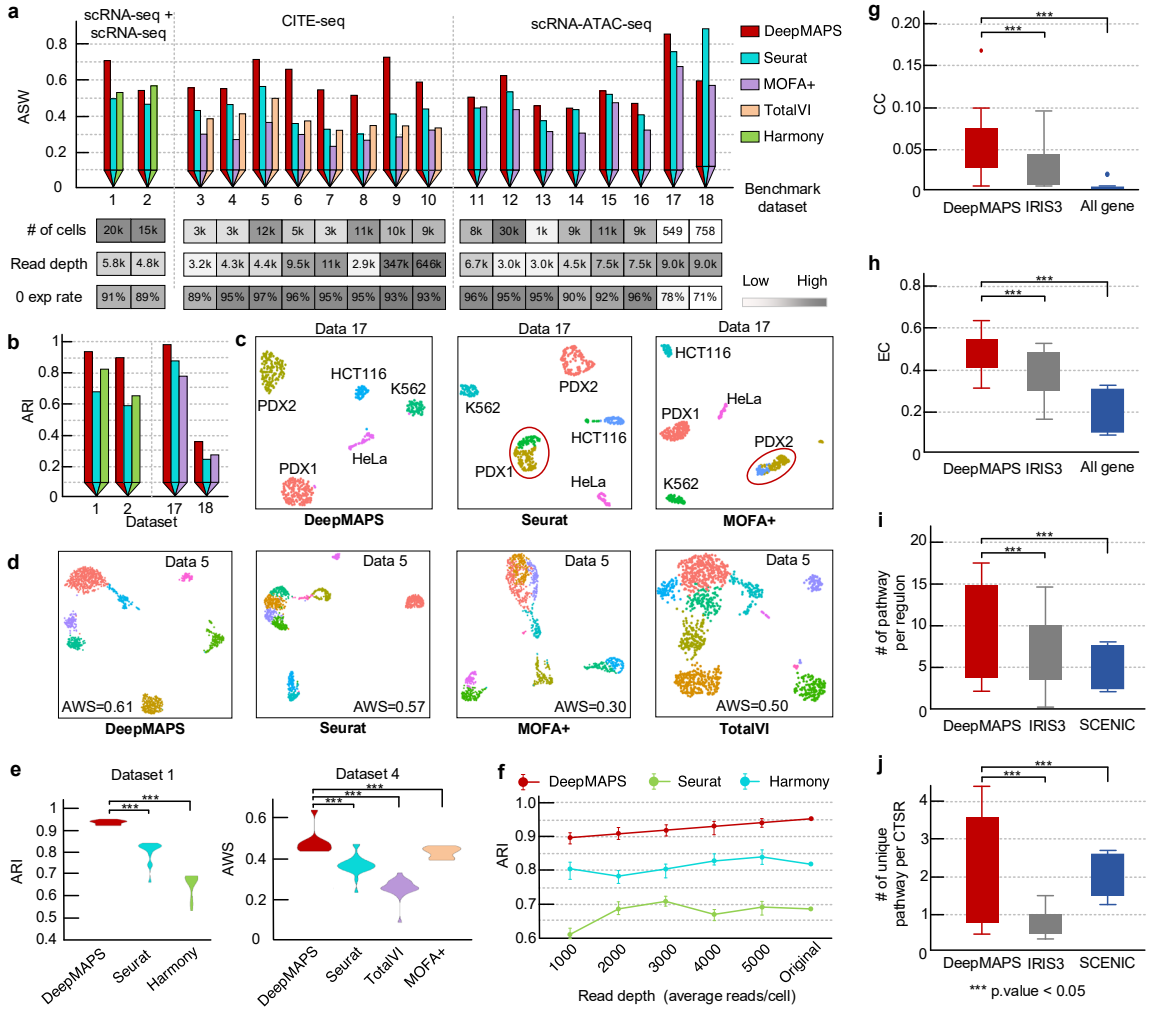

Fig. 3. Benchmarking of DeepMAPS in cell clustering and biological network inference. (**a**) Benchmark cell clustering results of all 18 datasets in terms of ASW without using benchmark cell labels. Different benchmarking tools were selected for comparison based on the capability of the tool. The heatmaps indicate the number of total cells, average gene expression read depth per cell, and average RNA zero-expression rate in each data. (**b**) Results comparison on four datasets with benchmarking cell labels in terms of adjusted rand index. (**c**) UMAP comparison of Data 17 (with benchmark labels) between DeepMAPS and other tools. Cluster labels were annotated based on cell correspondence to the original cell label. (**d**) UMAP comparison of Data 5 (without benchmark labels) between DeepMAPS and other tools. (**e**) Robustness test of DeepMAPS using the cell cluster leave-out method for Data 1 (with benchmark labels) and Data 4 (without benchmark labels). Details can be found in the Method section. (**f**) Robustness test of DeepMAPS to different read depths with Data 1. (**g** and **h**) Evaluation and comparison of gene association network inference of DeepMAPS and other methods. Closeness centrality and betweenness centrality were used to indicate the compactness and connectivity of networks inferred from different methods. (**i** and **j**) Evaluation and comparison of gene regulatory networks (regulons) identified by DeepMAPS, IRIS3,

and SCENIC, based on the number of functional pathways enriched in a regulon or cell-type-specific regulon.

DeepMAPS accurately identify cell types and infer cell-cell communication in PBMC and lung tumor immune CITE-seq data

We present a case study that applies DeepMAPS to a published mixed peripheral blood mononuclear cells (PBMC) and lung tumor leukocytes CITE-seq dataset to demonstrate capacity in modeling scMulti-omics in characterizing cell identities. The dataset includes RNAs and proteins measured on 3,485 cells. DeepMAPS identified 13 cell clusters, including four $CD4^+$ T cell groups (naïve, central memory (CM), tissue-resident memory (TRM), and regulatory (Treg)), two $CD8^+$ T cell groups (CM and TRM), a natural killer cell group, a memory B cell group, a plasma cell group, two monocyte groups, one tumor-associated macrophage (TAM) group, and a dendritic cell (DC) group. We annotated each cluster by visualizing the expression levels of curated maker genes and proteins (**Fig. 4a-b**, **Supplementary Data 8**). Compared to cell types identified using only proteins or RNA, we isolated or accurately annotated cell populations that could not be characterized using the individual modality analysis. In summary, by combining signals captured from both RNA and proteins, DeepMAPS successfully identified biologically reasonable and meaningful cell types in the CITE-seq data.

We then compared the modality correlation between two cell types. We used the top differentially expressed genes and proteins between memory B cells and plasma cells and performed hierarchical clustering of the correlation matrix. The result clearly stratified these features into two anticorrelated modules: one associated with memory B cells and the other with plasma cells (**Fig. 4c**). Furthermore, we found that the features in the two modules significantly correlated with the axis of maturation captured by our HGT embeddings (**Supplementary Fig. 1**). For example, we observed that one HGT embedding (the 51st) showed distinctive differences between plasma cells and memory B cells (**Fig. 4d-e**). Similar findings were also observed when comparing EM $CD8^+$ T cells and TRM $CD8^+$ T cells and showed a much closer relationship for expression correlations (**Fig. 4f**). Nevertheless, it is possible to identify a representative HGT embedding (56th) that maintains embedding signals for a defined separation of the two groups (**Fig. 4g-h**). These results point to any two cell clusters consisting of coordinated activation and repression of multiple genes and proteins, leading to a gradual transition in cell state that can be captured by a specific dimension of the DeepMAPS latent HGT space. On the other hand, we generated the gene-associated networks with genes showing high attention scores for EM $CD8^+$ T cells, TRM $CD8^+$ T cells, memory B cells, and plasma cells and observed diverse patterns (**Supplementary Fig. 2**).

Based on the cell types and raw data of gene and protein expressions, we inferred cell-cell communication using CellChat[17]. We constructed communication networks among

different cell types within multiple signaling pathways (**Fig. 4i)** and further applied CellChat[17] to detect any ligand-receptor interactions that have been validated between any cell types we identified. For example, we observed a CD6-ALCAM signaling pathway existing between DC (source) and TRM $CD4^+$ T cells (target) in the lung cancer tumor microenvironment (TME). Previous studies have shown that ALCAM on antigen-presenting DCs interact with CD6 on the T cell surface and contribute to T cell activation and proliferation[18-20]. As another example, we identified the involvement of the NECTIN-TIGIT signaling pathway during the interaction between the TAM (source) and TRM $CD8^+$ T cells (target), which is supported with a previous report that NECTIN (CD155) expressed on TAM could be immunosuppressive when interacting with surface receptors, TIGIT, on $CD8^+$ T cells in the lung cancer TME[21, 22].

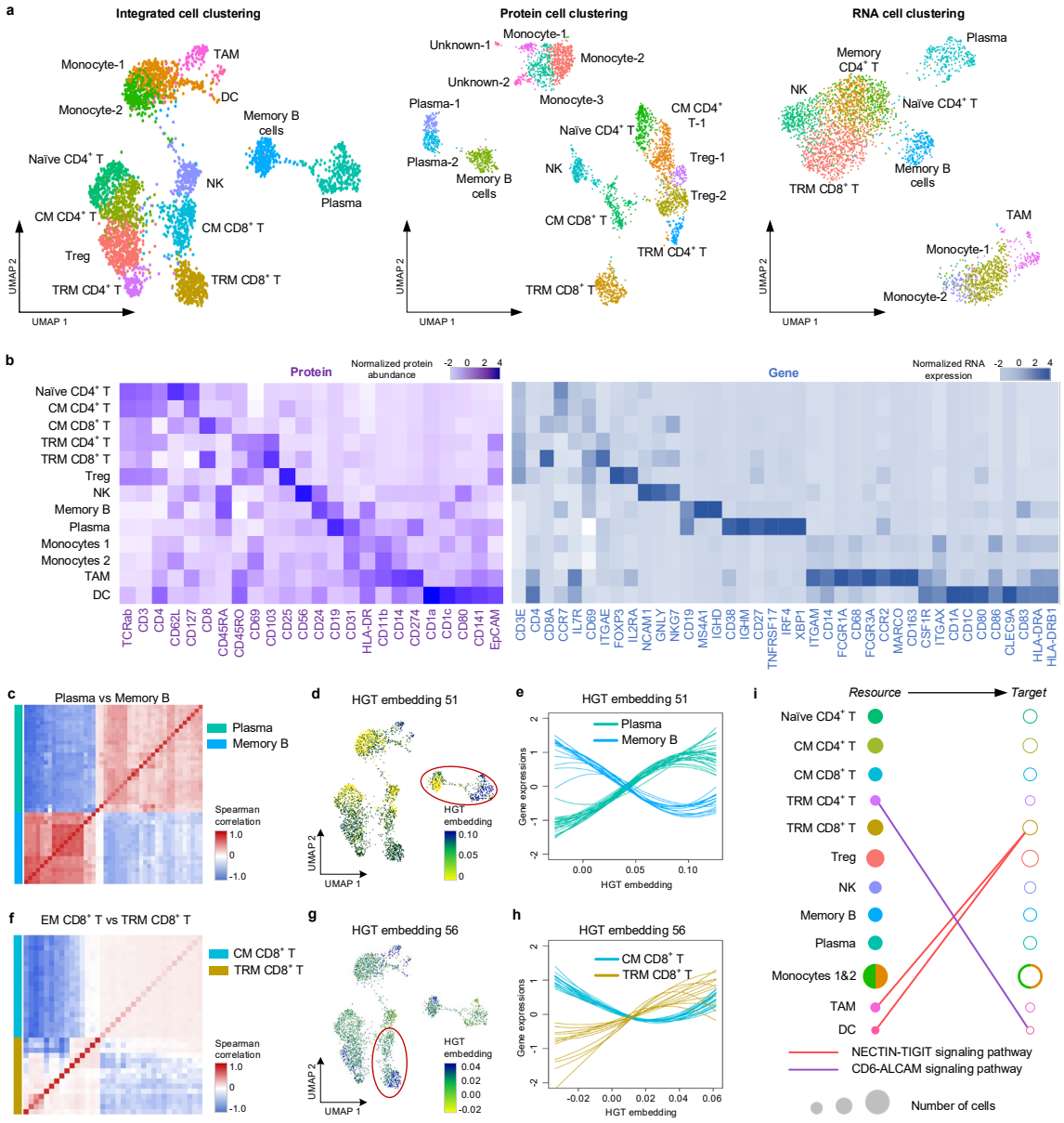

Fig. 4. DeepMAPS identification of heterogeneity in CITE-seq data of PBMC and lung tumor leukocytes. (**a**) UMAPs for DeepMAPS cell clustering results from integrated RNA and protein data, protein data only, and RNA data only. Cell clusters were annotated based on curated marker proteins and genes. (**b**) Heatmap of curated marker proteins and genes that determine the cell clustering and annotation. (**c**) Heatmap of the correlation comparison of top differentially expressed genes and proteins in plasma cells and memory B cells. (**d**) UMAP is colored by the 51st embedding, indicating distinct embedding representations in plasma cells and memory B cells. (**e**) Expression of top differentially expressed genes and proteins in ***c*** as a function of the 51st embedding to observe the pattern relations between plasma cells and memory B cells. Each line represents a gene/protein, colored by cell types. For each gene, a line was drawn using a loess smoothing function based on the corresponding embedding and scaled gene expression in a cell. (**f-h**) Similar

visualization was conducted to compare EM $CD8^+$ T cells and TRM $CD8^+$ T cells (***c-e)***. (**i**) Two signaling pathways, NECTIN and ALCAM, are shown to indicate the predicted cell-cell communications between two cell clusters. A link between a filled circle (resource cluster with highly expressed ligand coding genes) and an unfilled circle (target cluster with highly expressed receptor coding genes) indicates the potential cell-cell communication of a signaling pathway. Circle colors represent different cell clusters, and the size represents the number of cells. The two monocyte groups were merged.

DeepMAPS identifies specific gene regulatory networks in diffuse small lymphocytic lymphoma scRNA-seq and scATAC-seq data

To further extend the power of DeepMAPS to gene regulatory network inference, we used a single-cell multiome ATAC+Gene expression dataset available on the 10X Genomics website (10X Genomics online resource, **Supplementary Data 1**). The raw data is derived from 14,566 cells of flash-frozen intra-abdominal lymph node tumor from a patient diagnosed with diffuse small lymphocytic lymphoma (DSLL) of the lymph node lymph. We developed a new method to achieve the integration of gene expression and chromatin accessibility by balancing the weight of each modality of a gene in a cell, based on RNA velocity (**Fig. 5a** and **Method**). We demonstrate that using velocity-weighted integration methods can achieve better and rigorous cell clustering results than the Seurat weighted nearest neighbor method (**Supplementary Fig. 3**). To build TF-gene linkages, we considered the gene expression, gene accessibility, TF-motif binding affinity, peak to gene distance, and TF-coding gene expression. Genes found to be regulated by the same TF in a cell cluster are grouped as a regulon. We considered regulons with higher centrality scores to have greater influences on the characterization of the cell cluster. Regulons regulated by the same TF across different cell clusters are compared for differential regulon activities. Those with significantly higher regulon activity scores (RAS) are considered to be the cell-type-specific regulons in the cell cluster.

DeepMAPS identified 11 cell clusters in the DSLL data. All clusters were manually annotated based on curated gene markers (**Fig. 5b**, **Supplementary Data 9**). Two DSLL-like cell clusters (DSLL state-1 and state-2) were observed. The RNA velocity-based pseudotime analysis performed on the three B cell clusters (normal B cell and two DSLL states) assumed that the two DSLL states were derived from normal B cells, and state-1 is derived earlier than state-2, although the two states seemed to be partially mixed (**Fig. 5c**). We further selected the top 20 TFs with the highest regulon centrality scores in each of the three cell clusters (**Fig. 5d**, **Supplementary Data 10**). Interestingly, these TFs showed distinctions between the normal and the two DSLL states and inferred variant regulatory patterns within the two DSLL states. For regulons shared by all three B cell clusters, EGR1, MEF2B, and FOS were transcriptionally active in both normal B and DSLL cells and responsible for regulating B cell development, proliferation, and germinal center formation[23-26]. E2F6, ELF3, and KLF16 were identified as shared only in the two DSLL

states, with reported roles in tumorigenesis[27-32]. Further, JUN, MAFK, and MAFG, which encode the compartments of the activating protein-1 (AP-1),[26, 33, 34] were found to be active in DSLL state-1 while NFKB1, coding for a subunit of the NF-κB protein complex[35, 36], was found to be active in DSLL state-2.

We constructed a gene regulatory network consisting of the four cell-type-specific regulons (JUN, KLF16, GATA1, and FOS) (**Fig. 5e**, **Supplementary Fig. 4**) in DSLL state-1 with RAS that are significantly higher than normal B cells and DSLL state-2 (**Fig. 5f**). KLF16 reportedly promotes the proliferation of both prostate[31] and gastric cancer cells[32]. FOS and JUN are transcription factors in the AP-1 family, regulating the oncogenesis of multiple types of lymphomas[26, 33, 34, 37], and GATA1 is essential for hematopoiesis, the dysregulation of which is implicated in multiple hematologic disorders and malignancies[38, 39]. Distinct regulatory patterns were also observed when we zoomed in to a single regulon (**Fig. 5g**, **Supplementary Fig. 5**). As the most active regulon in DSLL state-1, JUN was found to regulate five unique downstream genes and 12 genes shared with DSLL state-2. Downstream genes, including CDK6[25, 26], IGF2R[40], and RUNX1[41], are critical for cell proliferation, survival, and development functions in DSLL.

Moreover, we further built connections between upstream cell-cell communication signaling pathways and downstream regulatory mechanisms in DSLL cells. We identified a cell-cell communication between macrophage and the two DSLL states via the B cell activation factor (BAFF) signaling pathway, based on the integrated GAS matrix using CellChat[17], which includes BAFF as the ligand on macrophage cells and TACI (transmembrane activator and calcium-modulator and cyclophilin ligand interactor) as the receptor on DSLL cells (**Fig. 5h**). BAFF signaling is critical to the survival and maturation of normal B cells[42, 43], while aberrations contribute to the resistance of malignant B cells to apoptosis[44, 45]. We observed that the expression of the TACI coding gene, TNFRSF13B, was explicitly higher in the two DSLL states, while the corresponding chromatin accessibility maintained high peaks in state-1 (**Fig. 5i**). Upon engagement with its ligand, TACI has been reported to transduce the signal and eventually activate the AP-1[46, 47] and NF-κB[48, 49] transcriptional complexes for downstream signaling in B cells. JUN (a subunit of AP-1) was identified as the most specific and key regulator in state-1 responsible for cell proliferation and regulating downstream oncogenes, such as CDK6 that has been reported to promote the proliferation of cancer cells in multiple types of DSLLs as well as other hematological malignancies[50-52]. It is now clear that BAFF signaling first appears in DSLL state-1 and triggers the activation of the JUN regulatory mechanism, leading to a high regulon activity of JUN. The JUN regulon specifically accelerates the proliferation and oncogenesis in DSLL, leading to a more terminal differentiatial stage of DSLL (state-2). As a result, state-1 includes cells undergoing rapid cell proliferation and differentiation, transitioning from normal B cells to matured DSLL. In short, DeepMAPS can construct gene regulatory networks and identify cell-type-specific regulatory patterns to offer a better

understanding of cell states and developmental orders in diseased subpopulations.

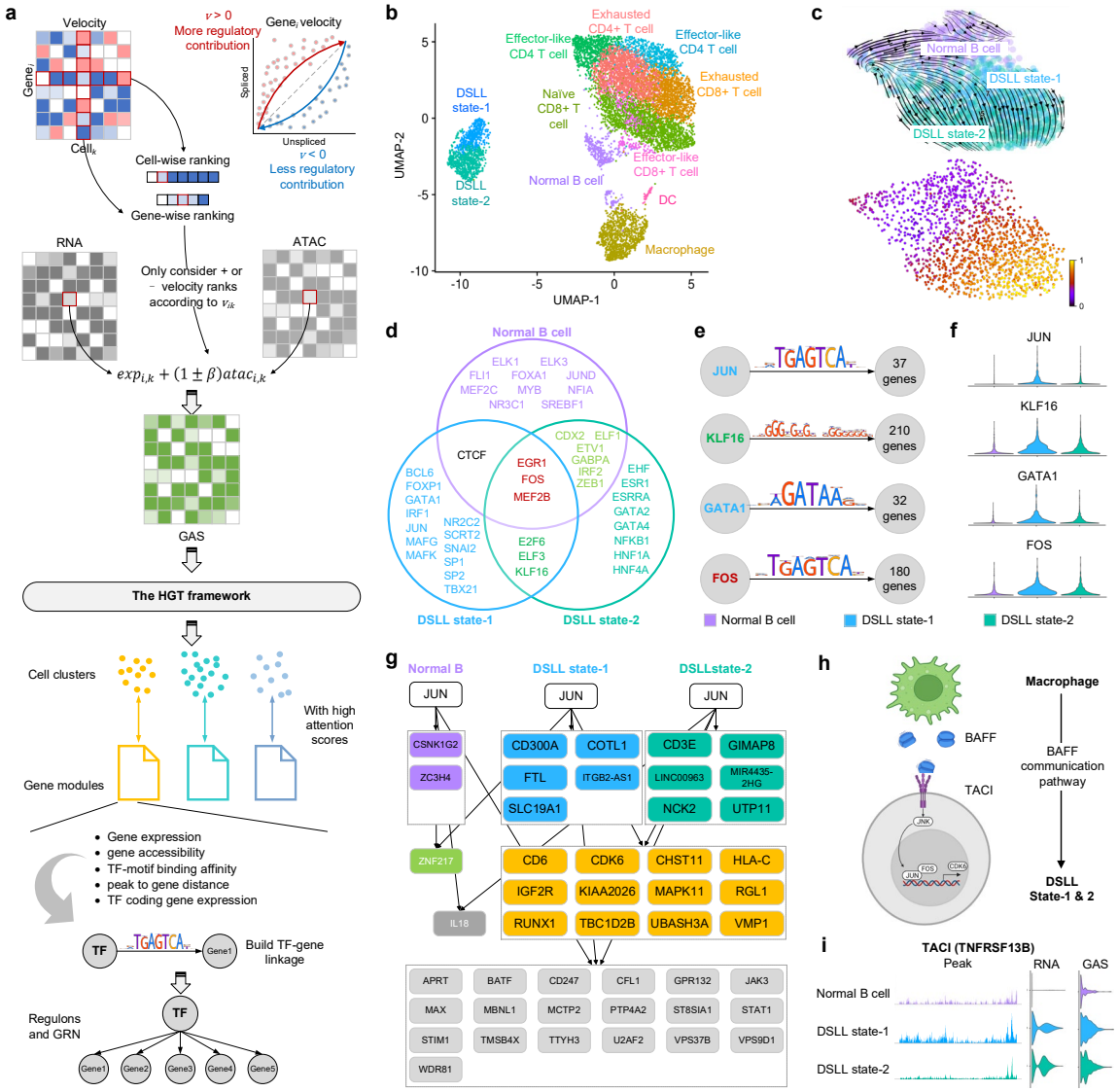

Fig. 5. DeepMAPS identifies specific gene regulatory networks in DSLL subnetworks. (a) Conceptual illustration of DeepMAPS analysis of scRNA-ATAC-seq data. Modalities are first integrated based on a velocity-weighted balance. The integrated gene activity matrix (GAS) was then used to build a heterogeneous graph as input into the HGT framework. The cell cluster and gene modules with high attention scores were then used for building TF-gene linkages and determining regulons in each cell cluster. (b) The UMAP shows the clustering results of DeepMAPS. Cell clusters were manually annotated based on curated marker genes. (c) The observed and the extrapolated future states (arrows) based on the RNA velocity of the normal B cell and the two DSLL states are shown (top panel). Velocity-based trajectory analysis shows the pseudotime from the top to the bottom right (bottom panel). (d) Selected 20 TF in each of the three clusters, representing the top 20 regulons with the highest centrality scores. Colors represent regulons uniquely identified in each cluster or shared between different clusters. (e) Regulons in DSLL state-1 showed a significant difference in regulon

activity compared to the other clusters. Motif shape and number of regulated genes are also shown. (f) Violin plots of regulon activities of the four regulons compared between the three clusters. (g) The downstream-regulated genes of JUN (the most differentially active regulon in DSLL state-1) in the three clusters. (h) An illustration of the BAFF signaling pathway identified from GAS-based cell-cell communication prediction using CellChat. The BAFF signaling pathway was found to exist between macrophage and both DSLL states. It further activates the JUN regulon and enables the transcription of genes like CDK6. (i) The ATAC peak, RNA expression, and GAS level of TNFRSF13B (the coding gene of TACI, the receptor in the BAFF signaling pathway).

DeepMAPS provides a multi-functional and user-friendly web portal for analyzing scMulti-omics data

Researchers who lack sufficient computational skills prefer to use webservers or dockers to lessen the programming burden of data analysis. Hence, a code-free and interactive platform for single-cell sequencing data analysis is urgently needed in the public domain. Because of the complexity of single-cell sequencing data, more and more webservers and dockers have been developed in the past three years[53-65] (**Supplementary Data 11**). However, most of these tools only provide minimal functions such as cell clustering and differential gene analysis. They do not support the joint analysis of scMulti-omics data and especially lack sufficient support for biological network inference. To this end, we developed DeepMAPS as a first-of-its-kind web portal to support online and code-free computational analysis for scMulti-omics data (**Fig. 6a**). The webserver supports the analysis of multiple RNA-seq data, CITE-seq data, and scRNA-ATAC-seq data using DeepMAPS (**Fig. 6b**). Some other methods, e.g., Seurat, are also incorporated as an alternative use for the users' convenience. Three major steps—data preprocessing, cell clustering and annotation, and network construction—are included in the server. In addition, the DeepMAPS server supports real-time computing and interactive graph representations. Users may register for an account to have their own workspace to store and share analytical results.

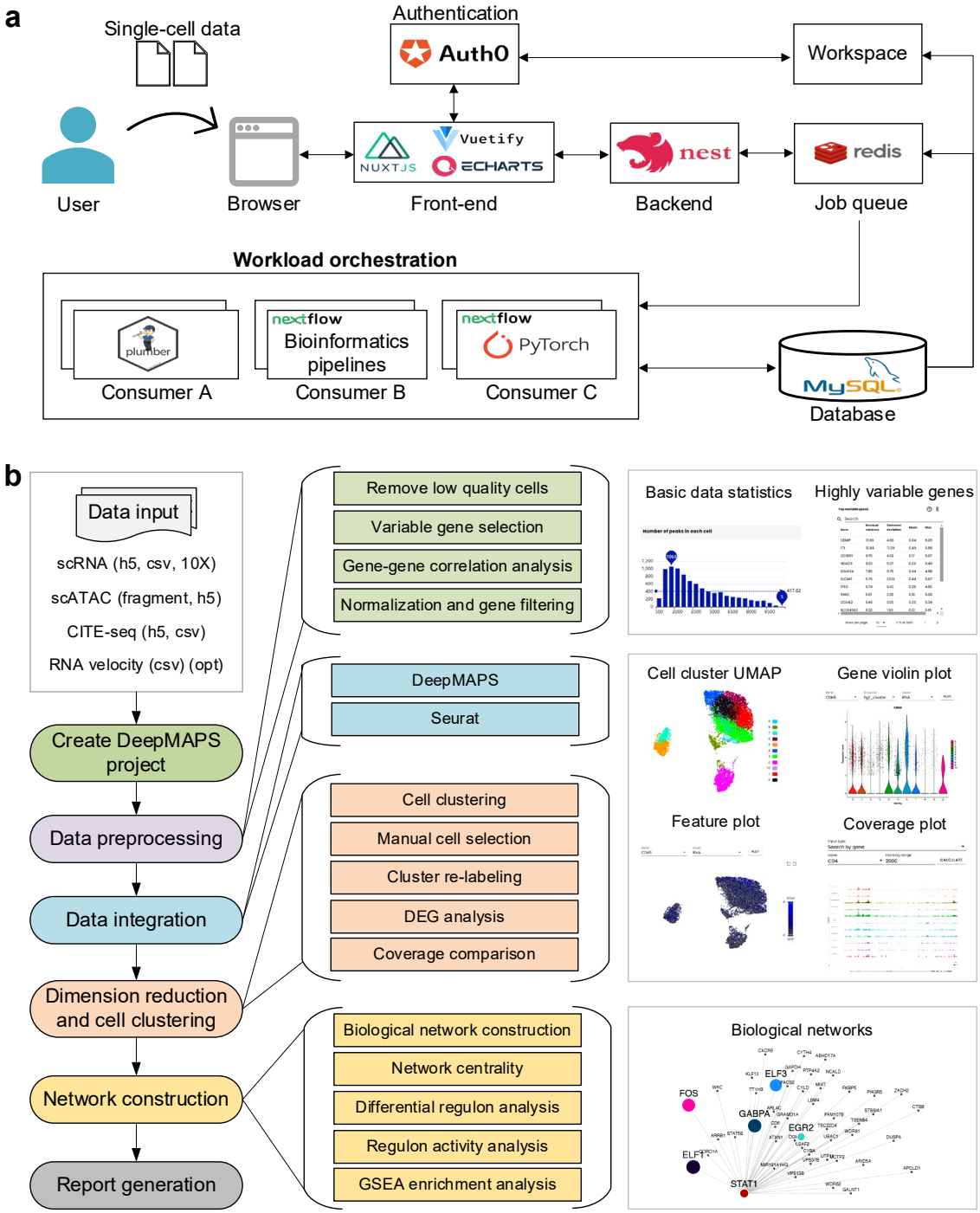

Fig. 6. The organization of the DeepMAPS web portal. (a) The software engineering of DeepMAPS and an overview of the framework. (b) Pipeline illustration of the server, including major steps (left), detailed analyses (middle), and featured figures and tables (right).

Conclusion and discussion

We have highlighted DeepMAPS as the first deep-learning framework that implements heterogeneous graph representation learning and a graph transformer in the study of

scMulti-omics data. By building a heterogeneous graph containing both cells and genes, DeepMAPS identifies the joint embedding of them simultaneously and enables the inference of cell-type-specific biological networks along with cell types in an intact framework. Furthermore, the application of heterogeneous graph transformer models the cell-gene relation in an interpretable uniform multi-relations. In such a way, the training and learning process in a graph can be largely shortened to consider cell impacts from a further distance.

By jointly analyzing gene expression and protein abundance, DeepMAPS accurately identified and annotated 13 cell types in a mixed CITE-seq data of PBMC and lung tumor leukocytes based on curated markers that cannot be fully elucidated using a single modality. We also proved that the embedding features identified in DeepMAPS capture statistically significant signals and amplify them when the original signals are noisy. Additionally, we identified biologically meaningful cell-cell communication pathways between DC and TRM $CD4^+$ T cells based on the gene network inferred in the two clusters. For scRNA-ATAC-seq, we employed an RNA velocity-based method to dynamically integrate gene expressions and chromatin accessibility that enhanced the prediction of cell clusters. Using such a method, we identified distinct gene regulatory patterns among normal B cells and two DSLL development states. We further elucidated the deep biological connections between cell-cell communications and the downstream gene regulatory networks, which helped characterize and define DSLL states. The identified TFs and genes can be potential markers for further validation and immuno-therapeutical targets in DSLL treatment.

While there are advantages and improved performances for analyzing scMulti-omics data, there is still room to improve the power of DeepMAPS further. First of all, the computational efficiency for super-large datasets (e.g., more than 1 million cells) might be a practical issue considering the complexity of the heterogeneous graph representation (may contain billions of edges). Moreover, DeepMAPS is recommended to be run on GPUs, which leads to a potential problem of reproducibility. Different GPU models have different floating-point numbers that may influence the precision of loss functions during the training process. That is to say, for different GPU models, DeepMAPS may generate slightly different cell clustering and network results, which is the main reason for the development of the webserver. Lastly, the current version of DeepMAPS is based on a bipartite heterogeneous graph with genes and cells. Separate preprocessing and integration steps are required to transfer different modalities to genes for integration into a unique cell-gene matrix. To fully achieve an end-to-end framework for scMulti-omics analysis, the bipartite graph can be extended to a multipartite graph, where each modality can be included as a node type. Such a multipartite heterogeneous graph can also include knowledge-based biological information, such as known molecular regulations and more than two modalities in one graph. However, by including more node types, the computational burden will be increased

geometrically, which requires a dedicated discovery of model optimization in the future.

In summary, we evaluated DeepMAPS as a pioneer study for the joint analysis of scMulti-omics data and cell-type-specific biological network inference. It will likely provide new visions of deep learning deployment in single-cell biology. With the development and maintenance of the DeepMAPS webserver, our long-term goal is to create a deep learning-based eco-community for archiving, analyzing, visualizing, and disseminating AI-ready scMulti-omics data.

Methods

Methods

Data preprocessing and data integration

Different data preprocessing and integration methods are designed for multiple scRNA-seq data, CITE-seq data, and scRNA-ATAC-seq. The method for each type of multi-omic data are shown as follows.

Multiple scRNA-seq data

DeepMAPS takes raw scRNA-seq gene expression profiles as inputs. Genes expressed in less than 0.1% of total cells or cells with a maximum of 0.1% of genes expressed are removed. To integrate multiple scRNA-seq datasets, we first reduced the dimension of multiple scRNA-seq data using canonical correlation analysis (CCA) and subsequently searched for mutual nearest neighbors (MNNs) in the shared low-dimensional space. The program then calculates vectors of each cell and corrects gene expressions in each dataset. The output would be an integrated matrix with I genes and J cells, aggregated from all datasets. Finally, the shared genes were normalized and re-scaled as in Seurat v3. The expression level of gene i in cell j was denoted as x_{ij}.

CITE-seq data

Low-quality genes and cells are removed as described above. Log normalization is then applied on the RNA matrix $X^{R\prime}$ with I_1 genes and J cells as follows:

$$x_{ij}^{R} = \log\left(1 + \frac{x_{ij}^{R\prime} \times 10{,}000}{\sum_i x_{ij}^{R\prime}}\right) \tag{1}$$

where $x_{ij}^{R\prime}$ represents the expression of gene i in cell j, $i = 1, 2, \ldots, I_1,$, $j = 1, \ldots, J$, and 10,000 is the scaling factor. Then, the top 2,000 (or I_1, whichever is smaller) highly variable genes are selected from matrix X^R. We still use X^R to represent the submatrix only containing these top-ranked genes.

Subsequently, RNA matrix X^R were spliced by rows with the protein matrix including I_2 proteins to obtain matrix X'. Then, a joint centered log-ratio (CLR) normalization was perfomed to obtain the processed matrix X as follows:

$$x_{ij} = \log\left(1 + \frac{x'_{ij}}{L_j}\right) \tag{2}$$

$$L_j = \exp\left(\frac{\sum_{i \in K_j} \log\left(1 + x'_{ij}\right)}{n(K_j)}\right) \tag{3}$$

$$K_j = \{\, i \mid x'_{ij} > 0 \} \tag{4}$$

where x'_{ij} represents the expression for gene/protein i in cell j, $i = 1, \ldots, min(2000, I_1) + I_2$, $j = 1, \ldots, J$ and $n(K_j)$ represents the number of genes/proteins with non-zero expression

values in cell j.

Matched scRNA-seq and scATAC-seq data (scRNA-ATAC-seq)

Data filtering and quality control are performed to obtain the RNA matrix and binarized ATAC matrix, X^R *and* X^C, respectively, following a standard protocol[5]. We first annotated peak regions in the scATAC-seq based on the method described in MAESTRO[66]. The regulatory potential R_{ik} for peak k and gene i is calculated independently by the exponential weight decay with the distance from the peak to the transcription start site (TSS):

$$r_{ik} = \begin{cases} 0, d_{ik} > 150\text{kb or peak p located in any nearby genes} \\ \frac{1}{L}, \text{peak } k \text{ located at the exon regions of the gene } i \\ 2^{-\frac{d_{ik}}{d_0}}, \text{else} \end{cases} \quad (4)$$

where d_{ik} is the distance between the center of peak k and the TSS of gene i and L refers to the length of the exon region. To ensure computational efficiency, we set the R_{ik} as 0 if d_{ik} was over 150 kb. The default value of d_0 is set to 10kb.

If a peak k was located at the exon regions of the gene i, the regulatory potential R_{ik} would be divided by L. The gene regulatory activity matrix X^A is defined in which the activity of gene i in cell j is:

$$x_{ij}^A = \sum_i r_{ik} x_{kj}^C . \quad (5)$$

We assume that the activity of a gene to a cell is determined by both gene expression activity and gene regulatory activity with different contributions. Unlike the contribution weights determined directly based on the expression and chromatin accessibility values in Seurat v4 (weighted nearest neighbor)[5], we hypothesize that the relative contribution of the expression and chromatin accessibility of a gene to a cell is dynamic rather than static and not accurately determined with a snapshot of the cell. RNA velocity is determined by the abundance of unspliced and spliced mRNA in a cell. The amount of unspliced mRNA is determined by gene regulation and gene transcription rate, and the amount of spliced mRNA is determined by the difference between unsliced mRNA and degraded mRNA. We reasoned that for genes with positive RNA velocities, there are higher potentials to drive genes to be transcribed. Thus, its regulatory activity related to chromatin accessibility has a greater influence than the gene expression in defining the overall gene activity in a cell of the current snapshot. For genes with negative velocities, the transcription rate tends to be decelerated, and regulatory activity has less influence on the cell than gene expression activity. We defined a gene activity score (GAS) matrix X, of which the score of gene i in cell j is defined which integrated RNA and ATAC information as follows,

$$x_{ij} = \begin{cases} x_{ij}^{R}.\sum_{i} x_{ij}^{R} + (1+b^{+}) \cdot x_{ij}^{A} \cdot \left((1+b^{+}) \cdot \sum_{i} x_{ij}^{A} \right), \text{for } o_{ij} > 0 \\ x_{ij}^{R}.\sum_{i} x_{ij}^{R} + (1-b^{-}) \cdot x_{ij}^{A} \cdot \left((1+b^{+}) \sum_{i} x_{ij}^{A} \right), for\ o_{ij} < 0 \\ x_{ij}^{R}.\sum_{i} x_{ij}^{R} + x_{ij}^{A}.\sum_{i} x_{ij}^{A}, \text{for } o_{ij} = 0 \end{cases} \quad (6)$$

where x_{ij} and x_{ij}^{A} represent the gene expression and gene regulatory activity for gene i in cell j, respectively, which are devided by the sum of genes, and o_{ij} denotes the RNA velocity for gene i in cell j, calculated using CellRank[67]. The weight was defined as:

$$b^{+} = \frac{\sqrt{\left(\left|\{i \mid o_{ij} > 0\}\right| - rank(o_{ij} \cdot \left|\{i \mid o_{ij} > 0\}\right|)\right)^{2} + \left(\left|\{j \mid o_{ij} > 0\}\right| - rank(o_{ij} \cdot \left|\{j \mid o_{ij} > 0\}\right|\right)^{2}}}{\sqrt{\left(\left|\{i \mid o_{ij} > 0\}\right| - 1\right)^{2} + \left(\left|\{j \mid o_{ij} > 0\}\right| - 1\right)^{2})}} \quad (7)$$

$$b^{-} = \frac{\sqrt{\left(\left|\{i \mid o_{ij} < 0\}\right| - rank\left(o_{ij} \cdot \left|\{i \mid o_{ij} < 0\}\right|\right)\right)^{2} + \left(\left|\{j \mid o_{ij} < 0\}\right| - rank(o_{ij} \cdot \left|\{j \mid o_{ij} < 0\}\right|\right)^{2}}}{\sqrt{\left(\left|\{i \mid o_{ij} < 0\}\right| - 1\right)^{2} + \left(\left|\{j \mid o_{ij} < 0\}\right| - 1\right)^{2})}} \quad (8)$$

Construction of gene-cell heterogeneous graph

After the data preprocessing and integrations, we obtaind a matrix, $X = \left(x_{ij}\right), i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, for any types of scMulti-omics data that integrated information from both modalities, with J columns for cells and I rows for genes. Values in the integrated matrix, x_{ij}, represent either normalized gene expressions (for multiple scRNA-seq and CITE-seq) or GAS (for scRNA-ATAC-seq). Given an integrated matrix $X \in \mathbb{R}^{I \times J}$ with I genes and J cells generated in the last step, a heterogeneous graph, denoted as $G = (V^{C}, V^{G}, E)$, where $V^{C} = \left\{v_{1}^{C}, v_{2}^{C}, \ldots, v_{|J|}^{C}\right\}$ and $V^{G} = \left\{v_{1}^{G}, v_{2}^{G}, \ldots, v_{|I|}^{G}\right\}$ represent the set of cells and genes in the integrated data and $E = \left\{e_{ij}\right\}, i \in V^{G}, j \in V^{C}$ represent the edges between gene i and cell j with the weight corresponding to the expression level of gene i in cell j. To initialize embedding of each node for gene i and cell j in the matrix X , a feature selection using autoencoder with two layers in both the encoder and decoder was used with the objective function of minimizing the mean squared error (MSE) between X and X'. The initial embedding of cells and genes are separately trained using the same feature autoencoder architecture with different loss functions. The loss function for gene embeddings is

$$loss(j) = \sum_{i} (X - X_{i})^{2} \quad (9)$$

and the loss function for cell embeddings is

$$loss(i) = \sum_{j} \left(X - X_j\right)^2 \tag{10}$$

Learning joint embeddings via a heterogeneous graph transformer

We proposed an unsupervised HGT framework[14, 15] to learn joint embeddings of gene i and cell j. Given a heterogeneous bipartite graph $G = (V^C, V^G, E)$, DeepMAPS would extracts all linked gene-cell node pairs, which are denoted as (s, t), where t is the target node and s is the neighbor node of t. The training processes consist of four steps: (*i*) calculating multi-head attention; (ii) passing heterogeneous message; (*iii*) aggregating neighbors' information; and (*iv*) calculating loss function. To learn joint embeddings of gene i and cell j, we aggregated information from s to get a contextualized representation for t and to simultaneously learn the representations of V_C and V_G. It is noteworthy that, to handle the heterogeneous relations in the graph, attention will be calculated via multiple heads where each node type (either cell node or gene node) has a unique head in attention and the nodes are linearly projected to a low dimensional space to model the distribution differences maximally

1) Calculating multi-head attention.

We extracted all linked node pairs, where a target node $v_t \in V^C \cup V^G$ was directly linked to its source node (neighbor node) $v_s = V^C \cup V^G$ through edge e. Let $\mathcal{H}^l$ denote the embedding of the lth HGT layer where $l \geq 1$, the contextualized representation of v_t on the lth layer is denoted as $\mathcal{H}^l[v_t]$, which can be learned by its own presentation layer $\mathcal{H}^{l-1}[v_t]$ and its neighbor layer $\mathcal{H}^{l-1}[v_s]$ from the $(l-1)^{th}$layer. The overall model is formulated as:

$$\mathcal{H}^l[v_t] \leftarrow \underset{\forall s,t \in V, \forall e \in E}{Aggregate}(\mathcal{H}^{l-1}[Attention(v_s, v_t, e) \cdot Message(v_s)]) \tag{11}$$

where *Attention* estimates the importance of each neighbor; *Message* extracts the information passed from the neighbors; *Aggregate* is the final step to aggregate the neighborhood message by the attention weight.

The multi-head attention is proposed in the attention level to calculate h-head attention for each edge $e = (v_s, v_t)$. Each target node v_t in the h-th head, $h = 1, \ldots, H$, was mapped into a target-node vector $T_{v_t}^h$ in each HGT layer via linear projection $T_{\tau(t)}^h$: $\mathbb{R}^d \rightarrow \times\ \mathbb{R}^{\frac{D}{H}}$, where D is the dimension of initial node feature, and $\frac{D}{H}$ is the feature dimension per head. Similarly, each neighbor node v_s in the head h was mapped into a key vector $K^h(v_s)$ with a linear projection $K_{\tau(s)}^h$: $\mathbb{R}^D \rightarrow \times\ \mathbb{R}^{\frac{D}{H}}$:

$$T^h(v_t) = Target_{\tau(t)}^h\left(\mathcal{H}^{(l-1)}[v_t]\right) \tag{12}$$

$$K^h(v_s) = Key_{\tau(s)}^h\left(\mathcal{H}^{(l-1)}[v_s]\right). \tag{13}$$

The similarity between the queries and keys was measured (e.g., scaled dot product operator) as attention. Then we calculated the multi-head attention for source node s to target node t by the dot product. To maximize parameter sharing while still maintaining the specific characteristics of different relations, we parameterize weight matrices $W_{\phi(e)}^{ATT}$ of the interaction operators. The hth head attention can be defined as:

$$ATThead^h(v_s, v_t, e) = \left(K^h(v_s) W_{\phi(e)}^{ATT} Q^h(v_t)^T\right) \cdot \frac{\mu\langle \tau(v_s), \tau(v_t)\rangle}{\sqrt{d}} \tag{14}$$

where $(\cdot)^T$ is the transposal function and μ is a prior tensor of the significance for each edge e, serving as an adaptive scaling to the attention. The attention score in the h-th head in the lth layer was defined as:

$$Att(v_s, v_t, e) = \underset{\forall s \in V(t)}{\mathrm{Softmax}}\left(\underset{H}{||} ATThead^h(v_s, v_t, e)\right) \tag{15}$$

where $\underset{H}{||} ATThead^h(\cdot)$ is the concatenation of embeddings in all H heads.

2) Passing heterogeneous message

To alleviate the distribution differences of different types of nodes and edges, we incorporated the types of edges into the message passing. The hth head message for each edge (v_s, v_t) was defined as:

$$M^h(v_s, v_t, e) = Message_{\tau(s)}^h\left(\mathcal{H}^{l-1}[v_s]\right) W_{\phi(e)}^{MSG} \tag{16}$$

Which means each source node s in the head h was mapped into a message vector by a linear projection $Message_{type(s)}^h : \mathbb{R}^d \rightarrow\times\ \mathbb{R}^{\frac{D}{H}}$, where $W_{\phi(e)}^{MSG} \in \mathbb{R}^{\frac{D}{H}\times\frac{D}{H}}$ is a distinct edge-based matrix for each edge, and $\phi(e)$ is the edge type of the heterogeneous graph. After multi-head aggregation, the degree of message passing was calculated as

$$Msg(v_s, v_t, e) = \underset{H}{||} head^h(v_s, v_t) \tag{17}$$

3) Aggregating neighbors' information

To obtain the representations of each node, we aggregated multi-head attentions and messages. The attention vectors was regarded as the weights for message representation. The representation of target nodes $\mathcal{H}^l[v_t]$ was updated as:

$$\mathcal{H}^l[v_t] = Linear_{v_t}\left(\sigma\left(\underset{\forall v_s \in N(v_t)}{\bigoplus}\left(Att(v_s, v_t, e) \cdot Msg(v_s, v_t, e)\right)\right)\right) + \mathcal{H}^{l-1}[v_t] \tag{18}$$

Where $Linear_{v_t}$ is a linear projection.

4) Calculating loss function.

The original application of HGT was to solve node classification problems. To utilize HGT

for training cell and gene embeddings without supervised classification labels, we used a graph autoencoder (GAE) framework. The whole HGT structure was deployed as an encoder in the GAE, while two embedding matrices M^C and M^G are used to record the trained embeddings of cells and genes from the HGT encoder where the columns were the same embedding dimensions and rows were cells and genes. A decoder was used to reconstruct the heterogeneous graph by the inner product of $M^C$and M^G. The loss function of GAE was calculated as:

$$loss = KL(X, < M^G, M^C >) \tag{19}$$

where $KL(\cdot)$ represents KL-divergence, and <·> denotes the inner product.

HGT training on subgraphs

To improve the efficiency and capability HGT model on such a giant heterogeneous graph (tens of thousands of nodes and millions of edges), we performed model training on subgraphs and multiple mini-batches based on the premise of HGSampling[14]. For a sampled node, we added all its direct neighbors into the corresponding budget and added its normalized degree to these neighbors, which were then used to calculate the sampling probability. Such normalization is equivalent to accumulating the probability of random walk of each sampled node to its neighborhood, avoiding the sampling being dominated by high-degree nodes. The sampler constructed a number of small subgraphs from the given giant graph, and these subgraphs were fed into the HGT model in batches with multiple GPUs. Finally, these distributed training results were collected to build the whole graph with representations on the head node.

The core of HGSampling is to sample heterogeneous subgraphs with similar proportions in different types of nodes to avoid sampling highly imbalanced subgraphs in the training process. We sampled a number of subgraphs for each epoch training, and the one-hot of each node was put into the trained model to obtain all node embedding. Given a node v_t which had been sampled and a dictionary $Dict[\tau]$ for each node type τ, we added all the first-neighbor nodes of ***t*** into the corresponding $Dict[\tau]$ and added the normalized degree of v_t to these neighbors to calculate the sampling probability:

$$Prob(\tau, v_s) = \frac{Dict[\tau][v_s]^2}{\left|\left|Dict[\tau]\right|\right|_2^2} \tag{20}$$

where $Prob(\tau, s)$ is the sampling probability for each source node s of type τ, $Dict[\tau]$ is all node for type τ with the normalized degree, $||.||_2$is the L2-norm, and $Dict[\tau][s]$ is the normalizer degree for the source node s of node type τ. Then we sampled all nodes types according to the probability in $Dict[\tau]$, and moved them out of $Dict[\tau]$. We repeated this sampling 50 times to obtain 50 subgraphs that maximize the coverage of the whole heterogeneous graph, and each subgraph was trained with 100 epochs.

Cell clustering and cell-type-active gene association network prediction

Cell clustering.

We apply the Louvain clustering method (igraph v1.2.7, R package) to predict m cell clusters $\{C_k, k = 1, 2, \cdots, m\}$ on cell-embedding matrix M^C.

Attention-based gene module detection.

To infer the connection of genes and cell clusters, we extracted the attention value of gene i to cell j in the h-th head through the step of multi-head attention calculation. We defined the importance function Imp of gene i to cell j as:

$$Imp(i,j) = \sqrt{\sum_{H} ATThead^{h}(i,j)^{\,2}} \tag{21}$$

We assigned genes to each cell with a threshold of $mean_g\big(Imp(i,j)\big) + sd_g(Imp(i,j))$. A gene is considered to be one of the active genes in a cell, if Imp is higher than the threshold.

The Steiner Forest Problem (SFP) model

We build an SFP model on a heterogeneous graph to extract the most critical gene-gene and gene-cell relations contributing to the gene module specificity in a cell cluster. The input of this model includes three parts:

1. gene-gene relations ($\tilde{E}^1$) defined by the embedding (resulted from HGT) Pearson's correlation between genes,
2. gene-cell relations are defined by the attention score of a gene to a cell ($\tilde{E}^2$).
3. a set of cell clusters, $\{\mathrm{C_k}, \mathrm{k} = 1, 2, \cdots, \mathrm{m}\}$, predicted by the HGT model.

We define a weighted heterogeneous graph $\tilde{G} = \big(\tilde{V}^1 \cup \tilde{V}^2, \tilde{E}^1 \cup \tilde{E}^2\big)$ in which nodes represent genes ($\tilde{V}^1$) and cells ($\tilde{V}^2$), and edges represent both gene-gene ($\tilde{E}^1$) and gene-cell ($\tilde{E}^2$) relations. We formulate this problem using a combinatorial optimization model defined as below

$$\min_{\tilde{E}' \subseteq \tilde{E}^1 \cup \tilde{E}^2} \sum_{e \in \tilde{E}'} w(e) \tag{22}$$

s.t.

$$r_{\tilde{E}'}(u,v) = 1, \forall\, u, v \in \mathrm{C_k}, k = 1, 2, \cdots, m \tag{23}$$

where $r_{\tilde{E}'}(u,v)$ is a binary indicator function representing whether two nodes, u and v, are connected (1) or not (0) in the subgraph induced by $\tilde{E}'$in $\tilde{G}$. We aim to identify the minimum weighted edge set, $\tilde{E}'$, from the heterogeneous graph $\tilde{G}$, so that cells in the same cell type could be connected to each other via edges in $\tilde{E}'$.

First, given the huge size of $\tilde{G}$, containing $|\tilde{V}^1 \cup \tilde{V}^2|$ nodes and $|\tilde{E}^1 \cup \tilde{E}^2|$ edges, we convert $\tilde{G}$ into a sparser graph, $\tilde{G}'$, by iteratively finding a global alignment between genes

and cells based on the gene-cell edges, using the maximum matching theory[68]. In graph theory, a matching or independent edge set in an undirected graph is a set of edges without common nodes, and a maximum matching in a weighted graph is a matching $\widetilde{M}$ that yields the maximum sum of edge weights. To fulfill this task, we build a weighted bipartite graph, $\tilde{G}^B = (\tilde{V}^1 \cup \tilde{V}^2, \widetilde{E}^2)$, by only retaining the gene-cell edges. The objective is to identify an edge subset, $\widetilde{M}$, to satisfy

$$\max_{\widetilde{M} \subseteq \tilde{E}^2} \sum_{e \in \widetilde{M}} w(e) \tag{24}$$

We calculate $\widetilde{M}$ of $\tilde{G}^B$ using the igraph R package[69]. Then, we remove the cell nodes incident by edges in M. The prediction of maximum matching and deletion of cell nodes incident to edges in previously identified maximum matchings is repeated until there is no cell node in the remaining graph. We compute the union of all the matchings as $\widetilde{E}^3$, and then construct $\tilde{G}' = (\tilde{V}^1 \cup \tilde{V}^2, \tilde{E}^1 \cup \tilde{E}^3)$. Finally, the weights of gene-cell edges, $\widetilde{E}^3$, and gene-gene edges, $\tilde{E}^1$, are normalized by the following two functions, respectively.

$$\frac{\max\left(w(\tilde{E}^3)\right) - w(\tilde{E}^3)}{\max\left(w(\tilde{E}^3)\right) - \min\left(w(\tilde{E}^3)\right)} \tag{25}$$

$$\frac{\max\left(w(\tilde{E}^1)\right) - w(\tilde{E}^1)}{\max\left(w(\tilde{E}^1)\right) - \min\left(w(\tilde{E}^1)\right)} \tag{26}$$

Second, we find the edge set, $\tilde{E}'$, of the SFP, $\tilde{T}$, as follows. To begin, we utilize the igraph R package to calculate a minimum spanning forest (MSF)[68], $\tilde{T}$, of $\tilde{G}'$. An MSF means that each pair of nodes in the same connected component could be connected to each other. Herein, we only need the edges to connect cell nodes belonging to the same cell type. Therefore, we iteratively remove the gene nodes with degree one and their unique neighbor is not a cell node from $\tilde{T}$, until no such gene node exists in $\tilde{T}$. Finally, we output the edge set of $\tilde{T}$, i.e., $\tilde{E}'$, as the solution to the SFP model.

For each connected component of the SFP, the gene-gene edges denote the co-expression relations among genes in the same module, while the set of gene-cell edges represents the cell type specificity of this gene module, and this module is a cell-type-active gene module.

Construct gene regulatory network from scRNA-ATAC-seq data

Infer master TFs and GRNs in each cell type.

To quantify the intensity of genes regulated by TFs, we design a regulatory intensive (RI) score, which can be decomposed into two components: (i) the regulatory potential (R_{ip}) of peaks calculated in the preprocessing step, and (ii) the binding affinity (BA) score of TFs to the peaks. The TF binding profiles were obtained from the JASPAR database. To reduce false positives of binding sites, we selected significant binding sites with *p*-values for TF binding profile matches less than 0.0001. The BA score was the transformed relative score

obtained from TF binding profiles, and the RI score of TF q to the gene i in the cell j was then defined as:

$$r_{i,j,q}^{Int} = \sum_{k} b_{qk}^{A} \cdot r_{ik} \tag{27}$$

Master TFs are genes at the top of a gene regulation hierarchy, particularly in regulatory pathways related to cell fate and differentiation. To infer cell type master TFs, we constructed cell-type-specific GRN with an RI score as edge weight and calculate centrality, which reflected the importance of each node in the network to rank the TFs in each cell type. TFs with high ranks were regarded as master TFs. Considering the RI score of TFs to genes, eigenvector centrality which assigned relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contributed more to the score of the node in question than equal connections to low-scoring nodes is applied to infer master TFs. The eigenvector centrality (EC_v) of a node v in GRN was defined as:

$$EC_v = \alpha_{max}(v) \tag{28}$$

where α_{max} is the eigenvector corresponding to the largest eigenvalue of the weighted adjacency matrix of a GRN.

Identity differential regulon (CTSRs)

We identify CTSRs by logFC and Wilcoxon rank-sum test to detect regulons associated with disease states. For a cell type active regulon, we defined a regulon activity score (RAS) as:

$$RAS(q, C_k) = \frac{\sum_{i \in I_q} \sum_{j \in C_k} x_{i,j} \cdot r_{i,j,q}^{Int}}{I \cdot J} \tag{29}$$

where I_q denotes genes regulated by TF q in cell type C_k. The significance of the difference is calculated using the Wilcoxon rank-sum test. If the BH-adjusted *p*-value is less than 0.05 between different cell clusters and the log fold change larger than 0.10, we consider the regulon to be differentially active in this cluster, and it is defined as a CTSR.

Benchmarking quantification and statistics

Adjusted rand index (ARI)

ARI is used to compute similarities by considering all pairs of the samples assigned in clusters in the current and previous clustering adjusted by random permutation. A contingency table is built to summarize the overlaps between the two cell label lists with b elements (cells) to calculate the ARI. Each entry denotes the number of objects in common between the two label lists. The ARI score can be calculated as:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{30}$$

Where $E[.]$ is the expectation, RI is the unadjusted rand index, which is defined as,

$$RI = \frac{J_a + J_b}{C_n^2} \tag{31}$$

where J_a is the number of cells that are assigned to the same cell cluster as benchmark

labels, and J_b is the number of cells that are assigned to different cell clusters as benchmark labels.

Average Silhouette Width (ASW)

Unlike ARI, which requires known ground truth labels, silhouette weight refers to a method of interpretation and validation of consistency within clusters of data. The silhouette weight indicates how similar an object is to its cluster (cohesion) compared to other clusters (separation). The silhouette width ranges from −1 to +1, where a high value indicates that the object is well matched to its cluster. The silhouette score $sil(j)$ can be calculated by:

$$sil(j) = \frac{n(j) - m(j)}{\max\{m(j), n(j)\}} = \begin{cases} 1 - \frac{m(j)}{n(j)}, & if\ m(j) < n(j) \\ 0, & if\ m(j) = n(j) \\ \frac{n(j)}{m(j)} - 1, & if\ m(j) > n(j) \end{cases} \tag{32}$$

where $m(j)$ is the average distance between a cell i and all other cells in the same cluster, and $n(j)$ is the average distance of i to all cells in the nearest cluster to which j does not belong. We take the average silhouette of all cells in a cluster as the average silhouette weight (ASW) to represent the cell distance of the cell cluster.

Closeness centrality (CC)

The closeness centrality (CC)[70] of a vertex u is defined by the inverse of the sum length of the shortest paths to all the other vertices v in the undirected weighted graph. The formulation is defined as:

$$CC(u) = \frac{1}{\sum_{u \neq v} d_w(u, v)} \tag{33}$$

where $d_w(u, v)$ is the shortest weighted path between u and v. If there is no path between vertex u and v, the total number of vertices is used in the formula instead of the path length. The CC is calculated using igraph R package with function igraph::betweenness.

Eigenvector centrality (EC)

Eigenvector centrality (EC)[71] scores correspond to the values of the first eigenvector of the graph adjacency matrix. The EC score of vertex u is defined as:

$$EC(u) = x_u = c \sum_{v \in g} a_{uv} x_v \tag{34}$$

where c is inverse of the eigenvalues of eigenvector $x = [x_1, x_2, \dots, x_n]$, a_{uv} is the adjacent weighted matrix of undirect graph g. The EC is calculated using igraph R package with function igraph::evcent.

Pathway enrichment test

To evaluate the function of the regulatory network, we use pathway enrichment analysis[72] to identify pathways that are significantly represented in each cell cluster active regulon

and count the number of regulon-enriched pathways. The pathway enrichment analysis is performed using enrichR R package [73].

Robustness evaluation

Cell cluster leave-out test

For a benchmark dataset with a real cell type label, we removed all cells in one cell type and ran DeepMAPS. Then, we traverse all cell types (one at a time) to evaluate the robustness with ARI. We removed cells in predicted cell clusters from DeepMAPS and other benchmark tools for data without benchmark labels.

Read depth simulation test

We performed a down-sampling simulation for gene expressions to test the robustness of DeepMAPS to read depth. Let matrix C be the $N \times M$ expression count matrix, where N is the number of cells and M is the number of genes. Define the cell sequencing depths $c_i = \sum_{j=1}^{M} C_{ij}$, i.e., the column sums of C. Thus, the average sequencing depth of the experiment is $\bar{c} = \frac{\sum_{i=1}^{n} c_i}{N}$. Let $t < \bar{c}$ be our target downsampled sequencing depth and let C^* be the $N \times M$ downsampled matrix. We perform the downsampling as follows:

For each spot $i = 1, \ldots, N$:

1) Define the total counts to be sampled in the cell i as $t_i = \frac{t \times c_i}{\bar{c}}$.
2) Construct the character vector of genes to be sampled as $G_i = \{\underbrace{1,\ldots,1}_{C_{i1}}, \underbrace{2,\ldots,2}_{C_{i2}}, \ldots, \underbrace{M,\ldots,M}_{C_{iM}}\}$.
3) Sample t_i elements from G_i without replacement and define N_j as the number of times gene j was sampled from G_i for $j = 1, \ldots, M$.
4) Let $C_{ij}^* = N_j$.

Using this method, the average downsampled sequencing depth is:

$$\overline{C}^* = \frac{\frac{t}{\bar{c}}c_1 + \frac{t}{\bar{c}}c_2 + \cdots + \frac{t}{\bar{c}}c_n}{N} = \frac{\frac{t}{\bar{c}}\sum_{i=1}^{N} c_i}{N} = \frac{t}{\bar{c}} \times \bar{c} = t \tag{35}$$

Note that, this method preserves the relative total counts of each cell, i.e., cells that had higher sequencing depths in the original matrix have proportionally higher depths in the downsampled matrix.

Comparisons with existing tools

In order to assess the performance of DeepMAPS alongside other proposed scMulti-omics benchmark tools, we compared DeepMAPS with Seurat (v 3.2.3 and v 4.0.0, https://github.com/satijalab/seurat), MOFA+ (v 1.0.0, https://github.com/bioFAM/MOFA2), Harmony (v 0.1, https://github.com/immunogenomics/harmony), and TotalVI (v 0.10.0, https://github.com/YosefLab/scvi-tools). Because of the integration capability, DeepMAPS

was compared with Seurat v 3.2.3 and Harmony on multiple scRNA-seq data, with Seurat v4.0.0, MOFA+, and TotalVI on CITE-seq data, and with Seurat v4.0.0 and MOFA+ on scRNA-ATAC-seq data. All benchmark tools used the default settings. We also evaluated the performance of gene association network inference with IRIS3[11] and a normal gene co-expression network inference method. Specifically, in IRIS3, cell-gene biclusters were first identified using QUBIC2[74]. Cell-type-active gene modules were identified in each cell cluster (using the same cell label predicted in DeepMAPS to ensure comparability) by performing a cell-wise hypergeometric enrichment.

On the other hand, all genes are selected to calculate a gene expression correlation score (Pearson's correlation) between any pair of two genes using cells in one cell cluster. Gene pair expression correlations with a BH-adjusted p-value smaller than 0.05 are retained and used to build the overall co-expression network in one cell cluster. Co-expressed sub gene modules are inferred by performing Louvain clustering on the co-expression network. For scRNA-ATAC-seq data, we compared regulon and cell-type-specific regulon inferred from DeepMAPS with IRIS3 in terms of enriched biological pathways. Noted that IRIS3 only supports regulon inference from scRNA-seq data based on *de novo* motif findings; thus, we used the GAS matrix generated in DeepMAPS as an input of IRIS3.

DeepMAPS server construction

DeepMAPS runs on an HPE XL675d RHEL system with 2×128-core AMD EPYC 7H12 CPU, 64GB RAM, and 2×NVIDIA A100 40GB GPU. The backend is written in TypeScript using the NestJs framework. Auth0 is used as an independent module to provide user authentication and authorization services. Redis houses a queue of all pending analysis jobs. There are two types of jobs in DeepMAPS: The stateful jobs are handled by the Plumber R package to provide real-time interactive analysis; the stateless jobs, such as CPU-bound bioinformatics pipelines and GPU training tasks that could take a very long time, are constructed using Nextflow (**Supplementary Data 12**). All running jobs are orchestrated using Nomad, allowing each job to be assigned with proper cores and storage and keeping jobs scalable based on the server load. The job results are deposited to a MySQL database. The frontend is built with NUXT, Vuetify as the UI library, Apache ECharts, and Cytoscape.js for data visualization. The frontend server and backend server are communicated using REST API.

Data availability

All data used for benchmarking and case study are collected from the public domain and can be retrieved using links or accession numbers provided in **Supplementary Data 1.**

Code availability

The source code of DeepMAPS Docker is freely available at (https://github.com/OSU-BMBL/deepmaps). The DeepMAPS webserver is available at https://bmblx.bmi.osumc.edu/).

Acknowledgments

This work was supported by awards R01-GM131399, R35-GM126985, and U54AG075931 from the National Institute of General Medical Sciences of the National Institutes of Health. The work was also supported by award NSF1945971 from the National Science Foundation. In addition, we thank Dr. Fei He from the Northeast Normal University for his valued suggestions in framework construction and data testing.

Contributions

Q.M., B.L., and D.X. conceived the basic idea and designed the framework. X.W. wrote the backbone of DeepMAPS. C.W. built the backend and frontend server. S.G. designed the interactive figures on the server. Y. Liu carried out RNA velocity calculations. Y. Li designed the SFP model for gene module predictions. A.M, X.W., and J.L. carried out benchmark experiments. X.W., Y.C., and B.L. performed robustness tests. A.M., J.L., X.W., X.G., and T.X. carried out the case study. J.W., D.W., Y.J., J.L., and L.S. performed tool optimizations. A.M. led the figure design and manuscript writing. All authors participated in the interpretation and writing of the manuscript.

Corresponding author

Correspondence to Dr. Qin Ma: Email: qin.ma@osumc.edu; Dr. Dong Xu: xudong@missouri.edu; Dr. Bingqiang Liu: bingqiang@sdu.edu.cn.

Ethics declarations

Competing interests

The authors declare no competing interests.

Reference

1. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257-272 (2019).
2. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends in Biotechnology* (2020).
3. Argelaguet, R., Cuomo, A.S.E., Stegle, O. & Marioni, J.C. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* (2021).
4. S Teichmann, M.E. Method of the Year 2019: Single-cell multimodal omics. *Nat Methods* **17**, 1 (2020).
5. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e3529 (2021).
6. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 111 (2020).
7. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289-1296 (2019).
8. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods* **18**, 272-282 (2021).
9. Li, Y. et al. Elucidation of Biological Networks across Complex Diseases Using Single-Cell Omics. *Trends Genet* (2020).
10. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
11. Ma, A. et al. IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. *Nucleic Acids Res* **48**, W275-W286 (2020).
12. Han, P., Gopalakrishnan, C., Yu, H. & Wang, E. Gene Regulatory Network Rewiring in the Immune Cells Associated with Cancer. *Genes (Basel)* **8**, 308 (2017).
13. Wang, J. et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nature Communications* **12**, 1882 (2021).
14. Hu, Z., Dong, Y., Wang, K. & Sun, Y. in Proceedings of The Web Conference 2020 2704–2710 (Association for Computing Machinery, Taipei, Taiwan; 2020).
15. Wang, X. et al. in The World Wide Web Conference 2022–2032 (Association for Computing Machinery, San Francisco, CA, USA; 2019).
16. Schäfer, G. in Encyclopedia of Algorithms. (ed. M.-Y. Kao) 2099-2102 (Springer New York, New York, NY; 2016).
17. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nature Communications* **12**, 1088 (2021).
18. Ampudia, J. et al. CD6-ALCAM signaling regulates multiple effector/memory T cell functions. *The Journal of Immunology* **204**, 150.113-150.113 (2020).
19. Skonier, J.E. et al. Mutational analysis of the CD6 ligand binding domain. *Protein Engineering, Design and Selection* **10**, 943-947 (1997).
20. Gimferrer, I. et al. Relevance of CD6-mediated interactions in T cell activation and proliferation. *Journal of immunology (Baltimore, Md. : 1950)* **173**, 2262-2270 (2004).
21. Johnston, R.J., Lee, P.S., Strop, P. & Smyth, M.J. Cancer Immunotherapy and the Nectin

976 Family. *Annual Review of Cancer Biology* **5**, 203-219 (2021).
977 22. Li, X.-Y. et al. CD155 loss enhances tumor suppression via combined host and tumor-
978 intrinsic mechanisms. *J Clin Invest* **128**, 2613-2625 (2018).
979 23. Gururajan, M. et al. Early growth response genes regulate B cell development, proliferation,
980 and immune response. *Journal of immunology (Baltimore, Md. : 1950)* **181**, 4590-4602
981 (2008).
982 24. Oh, Y.-K., Jang, E., Paik, D.-J. & Youn, J. Early Growth Response-1 Plays a Non-redundant
983 Role in the Differentiation of B Cells into Plasma Cells. *Immune Netw* **15**, 161-166 (2015).
984 25. Brescia, P. et al. MEF2B Instructs Germinal Center Development and Acts as an Oncogene in
985 B Cell Lymphomagenesis. *Cancer Cell* **34**, 453-465.e459 (2018).
986 26. Trøen, G. et al. Constitutive expression of the AP-1 transcription factors c-jun, junD, junB,
987 and c-fos and the marginal zone B-cell transcription factor Notch2 in splenic marginal zone
988 lymphoma. *J Mol Diagn* **6**, 297-307 (2004).
989 27. Sánchez-Beato, M. et al. Abnormal PcG protein expression in Hodgkin's lymphoma. Relation
990 with E2F6 and NFkappaB transcription factors. *J Pathol* **204**, 528-537 (2004).
991 28. Saha, A., Robertson, E.S. & Goodrum, F. Mechanisms of B-Cell Oncogenesis Induced by
992 Epstein-Barr Virus. *J Virol* **93**, e00238-00219 (2019).
993 29. Yachida, S. et al. Genomic Sequencing Identifies ELF3 as a Driver of Ampullary Carcinoma.
994 *Cancer Cell* **29**, 229-240 (2016).
995 30. Wang, H. et al. Overexpression of ELF3 facilitates cell growth and metastasis through
996 PI3K/Akt and ERK signaling pathways in non-small cell lung cancer. *Int J Biochem Cell Biol*
997 **94**, 98-106 (2018).
998 31. Zhang, J. et al. KLF16 Affects the MYC Signature and Tumor Growth in Prostate Cancer.
999 *Onco Targets Ther* **13**, 1303-1310 (2020).
1000 32. Ma, P. et al. KLF16 promotes proliferation in gastric cancer cells via regulating p21 and
1001 CDK4. *Am J Transl Res* **9**, 3027-3036 (2017).
1002 33. Mathas, S. et al. Aberrantly expressed c-Jun and JunB are a hallmark of Hodgkin lymphoma
1003 cells, stimulate proliferation and synergize with NF-κB. *The EMBO Journal* **21**, 4104-4113
1004 (2002).
1005 34. Eferl, R. & Wagner, E.F. AP-1: a double-edged sword in tumorigenesis. *Nature Reviews*
1006 *Cancer* **3**, 859-868 (2003).
1007 35. Nagel, D., Vincendeau, M., Eitelhuber, A.C. & Krappmann, D. Mechanisms and consequences
1008 of constitutive NF-κB activation in B-cell lymphoid malignancies. *Oncogene* **33**, 5655-5665
1009 (2014).
1010 36. Jost, P.J. & Ruland, J.r. Aberrant NF-κB signaling in lymphoma: mechanisms, consequences,
1011 and therapeutic implications. *Blood* **109**, 2700-2707 (2006).
1012 37. Garces de Los Fayos Alonso, I. et al. The Role of Activator Protein-1 (AP-1) Family Members
1013 in CD30-Positive Lymphomas. *Cancers (Basel)* **10**, 93 (2018).
1014 38. Crispino, J.D. & Horwitz, M.S. GATA factor mutations in hematologic disease. *Blood* **129**,
1015 2103-2110 (2017).
1016 39. Shimizu, R., Engel, J.D. & Yamamoto, M. GATA1-related leukaemias. *Nat Rev Cancer* **8**, 279-

1017 287 (2008).
1018 40. Mosquera Orgueira, A. et al. Detection of new drivers of frequent B-cell lymphoid
1019 neoplasms using an integrated analysis of whole genomes. *PLOS ONE* **16**, e0248886 (2021).
1020 41. Blyth, K. et al. Runx1 promotes B-cell survival and lymphoma development. *Blood Cells Mol*
1021 *Dis* **43**, 12-19 (2009).
1022 42. Mackay, F., Schneider, P., Rennert, P. & Browning, J. BAFF AND APRIL: a tutorial on B cell
1023 survival. *Annu Rev Immunol* **21**, 231-264 (2003).
1024 43. Smulski, C.R. & Eibel, H. BAFF and BAFF-Receptor in B Cell Selection and Survival. *Frontiers*
1025 *in Immunology* **9** (2018).
1026 44. Yang, S., Li, J.Y. & Xu, W. Role of BAFF/BAFF-R axis in B-cell non-Hodgkin lymphoma. *Crit*
1027 *Rev Oncol Hematol* **91**, 113-122 (2014).
1028 45. He, B. et al. Lymphoma B cells evade apoptosis through the TNF family members BAFF/BLyS
1029 and APRIL. *Journal of immunology (Baltimore, Md. : 1950)* **172**, 3268-3279 (2004).
1030 46. Xia, X.Z. et al. TACI is a TRAF-interacting receptor for TALL-1, a tumor necrosis factor family
1031 member involved in B cell regulation. *J Exp Med* **192**, 137-143 (2000).
1032 47. Laâbi, Y., Egle, A. & Strasser, A. TNF cytokine family: more BAFF-ling complexities. *Curr Biol*
1033 **11**, R1013-1016 (2001).
1034 48. Mackay, F. & Schneider, P. TACI, an enigmatic BAFF/APRIL receptor, with new unappreciated
1035 biochemical and biological properties. *Cytokine Growth Factor Rev* **19**, 263-276 (2008).
1036 49. Rihacek, M. et al. B-Cell Activating Factor as a Cancer Biomarker and Its Implications in
1037 Cancer-Related Cachexia. *Biomed Res Int* **2015**, 792187-792187 (2015).
1038 50. Su, H., Chang, J., Xu, M., Sun, R. & Wang, J. CDK6 overexpression resulted from microRNA-
1039 320d downregulation promotes cell proliferation in diffuse large B-cell lymphoma. *Oncol*
1040 *Rep* **42**, 321-327 (2019).
1041 51. Lee, C., Huang, X., Di Liberto, M., Martin, P. & Chen-Kiang, S. Targeting CDK4/6 in mantle
1042 cell lymphoma. *Ann Lymphoma* **4**, 1 (2020).
1043 52. Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nature*
1044 *Reviews Cancer* **17**, 93-115 (2017).
1045 53. Li, K. et al. cellxgene VIP unleashes full power of interactive visualization, plotting and
1046 analysis of scRNA-seq data in the scale of millions of cells. *bioRxiv*, 2020.2008.2028.270652
1047 (2020).
1048 54. Pereira, W. et al. Asc-Seurat – Analytical single-cell Seurat-based web application. *bioRxiv*,
1049 2021.2003.2019.436196 (2021).
1050 55. Gardeux, V., David, F.P.A., Shajkofci, A., Schwalie, P.C. & Deplancke, B. ASAP: a web-based
1051 platform for the analysis and interactive visualization of single-cell RNA-seq data.
1052 *Bioinformatics* **33**, 3123-3125 (2017).
1053 56. Li, B. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and
1054 single-nucleus RNA-seq. *Nature Methods* **17**, 793-798 (2020).
1055 57. Hillje, R., Pelicci, P.G. & Luzi, L. Cerebro: interactive visualization of scRNA-seq data.
1056 *Bioinformatics* **36**, 2311-2313 (2019).
1057 58. Prompsy, P. et al. Interactive analysis of single-cell epigenomic landscapes with

1058 ChromSCape. *Nature Communications* **11**, 5702 (2020).
1059 59. Bolisetty, M.T., Stitzel, M.L. & Robson, P. CellView: Interactive exploration of high
1060 dimensional single cell RNA-seq data. *bioRxiv*, 123810 (2017).
1061 60. Mohanraj, S. et al. CReSCENT: CanceR Single Cell ExpressioN Toolkit. *Nucleic Acids Research*
1062 **48**, W372-W379 (2020).
1063 61. Patel, M.V. iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA
1064 sequencing data. *Bioinformatics* **34**, 4305-4306 (2018).
1065 62. Yousif, A., Drou, N., Rowe, J., Khalfan, M. & Gunsalus, K.C. NASQAR: a web-based platform
1066 for high-throughput sequencing data analysis and visualization. *BMC Bioinformatics* **21**, 267
1067 (2020).
1068 63. Zhu, Q. et al. PIVOT: platform for interactive analysis and visualization of transcriptomics
1069 data. *BMC Bioinformatics* **19**, 6 (2018).
1070 64. Innes, B. & Bader, G. scClustViz Single-cell RNAseq cluster assessment and visualization
1071 [version 2; peer review: 2 approved]. *F1000Research* **7** (2019).
1072 65. Granja, J.M. et al. ArchR is a scalable software package for integrative single-cell chromatin
1073 accessibility analysis. *Nat Genet* **53**, 403-411 (2021).
1074 66. Wang, C. et al. Integrative analyses of single-cell transcriptome and regulome using
1075 MAESTRO. *Genome Biology* **21**, 198 (2020).
1076 67. Lange, M. et al. CellRank for directed single-cell fate mapping. *bioRxiv*,
1077 2020.2010.2019.345983 (2020).
1078 68. Diestel, R., Schrijver, A. & Seymour, P.D. in MATHEMATISCHES FORSCHUNGSINSTITUT
1079 OBERWOLFACH REPORT NO. 16/2007 (Citeseer, 2007).
1080 69. Csardi, G. & Nepusz, T. The igraph software package for complex network research.
1081 *InterJournal, complex systems* **1695**, 1-9 (2006).
1082 70. Li, G.S., Li, M., Wang, J.X., Li, Y.H. & Pan, Y. United Neighborhood Closeness Centrality and
1083 Orthology for Predicting Essential Proteins. *Ieee Acm T Comput Bi* **17**, 1451-1458 (2020).
1084 71. Parisutham, N. & Rethnasamy, N. Eigenvector centrality based algorithm for finding a
1085 maximal common connected vertex induced molecular substructure of two chemical graphs.
1086 *J Mol Struct* **1244** (2021).
1087 72. Reimand, J. et al. Pathway enrichment analysis and visualization of omics data using
1088 g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* **14**, 482-517 (2019).
1089 73. Kuleshov, M.V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016
1090 update. *Nucleic Acids Res* **44**, W90-97 (2016).
1091 74. Xie, J. et al. QUBIC2: a novel and robust biclustering algorithm for analyses and
1092 interpretation of large-scale RNA-Seq data. *Bioinformatics* **36**, 1143-1149 (2020).
1093