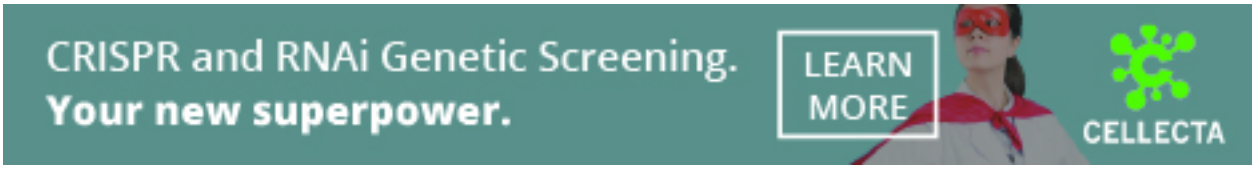
A gene regulatory network–aware graph learning method for cell identity annotation in single-cell RNA-seq data

Mengyuan Zhao, Jiawei Li, Xiaoyi Liu, et al.

Genome Res. 2024 34: 1036-1051 originally published online August 12, 2024
Access the most recent version at doi:10.1101/gr.278439.123



Method

A gene regulatory network–aware graph learning method for cell identity annotation in single-cell RNA-seq data

Mengyuan Zhao,[1,2] Jiawei Li,[3] Xiaoyi Liu,[4] Ke Ma,[5] Jijun Tang,[1] and Fei Guo[6]

[1]College of Computer Science and Control Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; [2]University of Chinese Academy of Sciences, Beijing 100190, China; [3]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China; [4]Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29208, USA; [5]College of Engineering, Southern University of Science and Technology, Shenzhen 518055, China; [6]School of Computer Science and Engineering, Central South University, Changsha 410083, China

Cell identity annotation for single-cell transcriptome data is a crucial process for constructing cell atlases, unraveling pathogenesis, and inspiring therapeutic approaches. Currently, the efficacy of existing methodologies is contingent upon specific data sets. Nevertheless, such data are often sourced from various batches, sequencing technologies, tissues, and even species. Notably, the gene regulatory relationship remains unaffected by the aforementioned factors, highlighting the extensive gene interactions within organisms. Therefore, we propose scHGR, an automated annotation tool designed to leverage gene regulatory relationships in constructing gene-mediated cell communication graphs for single-cell transcriptome data. This strategy helps reduce noise from diverse data sources while establishing distant cellular connections, yielding valuable biological insights. Experiments involving 22 scenarios demonstrate that scHGR precisely and consistently annotates cell identities, benchmarked against state-of-the-art methods. Crucially, scHGR uncovers novel subtypes within peripheral blood mononuclear cells, specifically from $CD4^+$ T cells and cytotoxic T cells. Furthermore, by characterizing a cell atlas comprising 56 cell types for COVID-19 patients, scHGR identifies vital factors like IL1 and calcium ions, offering insights for targeted therapeutic interventions.

[Supplemental material is available for this article.]

In single-cell transcriptome sequencing, high-resolution quantification of gene expression profiles provides insights into cellular heterogeneity and the molecular underpinnings of tissue phenotype variations (Kalucka et al. 2020; Argelaguet et al. 2021). Analyzing cell type composition using scRNA-seq offers an unprecedented chance to unveil cell subpopulations, reconstruct differentiation trajectories, and create comprehensive organismal cell atlases. Meanwhile, the extensive patient-specific cell atlas has notably expedited the investigation of biomarkers, pathogenic mechanisms, and prognostic implications related to outbreaks and intricate disorders (Puram et al. 2017; Semrau et al. 2017; Wang et al. 2023). Numerous extensive cell atlases are accessible owing to the substantial accumulation of single-cell omics data (Han et al. 2020; Travaglini et al. 2020; La Manno et al. 2021). Nonetheless, annotating cells within such data remains a formidable challenge. One aspect is that scRNA-seq data exhibit high noise and sparsity owing to limited mRNA capture. Conversely, newly generated data sets typically stem from diverse batches, technologies, tissues, and even species, resulting in batch effects and biological heterogeneity.

To address these issues, the typical approach is dividing extensive cells into biologically meaningful clusters according to expression patterns. Then, specific types are assigned to each cluster using prior knowledge, such as marker genes (Lopez et al. 2018; Pliner et al. 2019; Zhang et al. 2019; Cheng and Ma 2022; Yu et al. 2022; Zhai et al. 2023). Additionally, some researchers are using a well-annotated cell atlas as a reference and constructing cell communication networks by assessing expression similarities. This helps transfer labels to newly sequenced data sets (Michielsen et al. 2021; Shao et al. 2021; Song et al. 2021). However, the efficacy of the above approaches demonstrates notable variability across diverse data sets. This variance curtails their potential across a broad spectrum of application scenarios. Moreover, available approaches only rely on transcriptomics expression profile but neglect genetic interactions that are not disturbed by the above factors.

In contrast to scRNA-seq data sets, gene regulatory networks are independent of transcriptome sequencing technologies, which reflects the intrinsic regulatory mechanism in gene expression processes and plays a vital role in cell differentiation as well as subtype determination (Zhao et al. 2021, 2022). Studies have indicated that incorporating gene regulatory relationships enhances cell characterization and supports downstream tasks like cell clustering and lineage reconstruction (Aibar et al. 2017; Elyanow et al. 2020; Kamimoto et al. 2023; So et al. 2023). Gene regulatory networks potentially dominate gene expression profiles, resulting in genes colocated within analogous topological structures, often displaying correlated expression patterns. Genes involved in gene regulatory networks serve as intermediaries for coexpressing cells, thus enhancing the ability to identify an extended array of cell–cell

Corresponding authors: guofei@csu.edu.cn, jj.tang@siat.ac.cn
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278439.123.

34:1036–1051 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/24; www.genome.org

interactions. The exchange of information and expression patterns between cells serves as important markers for delineating cell types (Miller et al. 2021; Armingol et al. 2022).

To this end, we propose an automated annotation tool for single-cell transcriptome data, named single-cell hybrid graph neural network with gene regulations (scHGR). Leveraging large-scale experimentally validated gene regulatory relationships from databases, scHGR employs a graph representation algorithm to assemble correlations of gene network topologies. The tool builds a multilayer hybrid graph neural network (HGNN), incorporating nodes that represent both cells and genes, and implements a weighted-feature aggregation algorithm to enhance its predictive capabilities. The overarching goal of this study is to establish scHGR as a powerful computational instrument for efficient and accurate cell annotation. Moreover, we illustrate the versatility of scHGR in supporting downstream analyses, such as differential gene expression (DEG) analysis, subtype revealing, Gene Ontology (GO) enrichment, and cell communication inference.

Results

Overview of scHGR

To summarize, we use a deep learning framework based on (HGNNs) to transfer labels from reference data sets to newly sequenced cells based on gene regulatory relationships, which are prevalent within the same species. As inputs, scRNA-seq expression profiles and cell-level labels are provided for reference data sets, as well as scRNA-seq expression profiles for query data sets (Fig. 1A). There is also an optional input in which additional gene relationships can be uploaded if needed. In the absence of that, scHGR employs gene regulatory relationships from the self-equipped gene regulation repository (Fig. 1B). scHGR integrates gene relationships that are prevalent in human and mouse from the GREDB, BioGRID, TRRUST, and RegNetwork databases. In contrast to cell type–specific and tissue-specific gene regulatory networks, such gene relationships do not vary with cell type and tissue. Transcriptional Regulatory Relationships Unraveled by Sentence-based Text Mining (TRRUST) is a manually annotated database that efficiently extracts regulatory relationships from more than 20 million published articles (Han et al. 2018). RegNetwork is developed based on 25 databases and integrates in-library regulations and potential relationships (Liu et al. 2015). BioGRID is a biomedical interactive repository that includes focused low-throughput studies and large high-throughput data sets (Oughtred et al. 2019). GREDB is a publicly available, manually managed biological database that preserves human gene expression regulatory relationships (Huang et al. 2019). By means of structure deep network embedding (SDNE) (Wang et al. 2016), a nonlinear graph

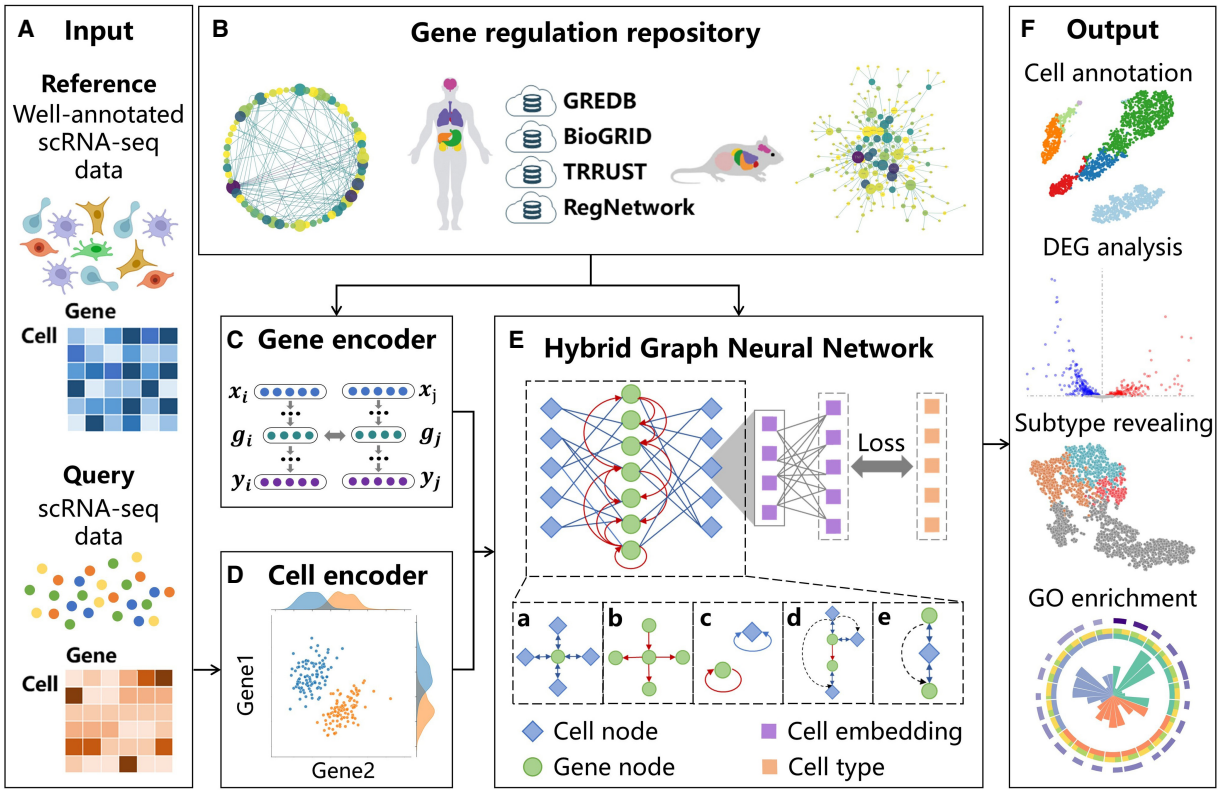


Figure 1. Overview of the scHGR. (*A*) The input of scHGR includes reference data sets equipped with single-cell resolution gene expression profiles as well as type information for each cell, and query data sets equipped with scRNA-seq data. (*B*) The underlying gene regulatory relationships to train scHGR are mainly from GREDB, BioGRID, TRRUST, and RegNetwork, which contain more than 2 million records pertaining to humans and mice. (*C*) Gene encoder generates the initial embedding of gene nodes via graph embedding algorithm. (*D*) Cell encoder encodes high-dimensional expression matrix into low-dimensional embedding of cell nodes. (*E*) A hybrid graph neural network (HGNN) factors in six types of edges. It optimizes parameters by minimizing the gap between prediction and reference cell types. The blue diamond and green circle represent the cell nodes and gene nodes in HGNN; the purple square represents cell embedding; and the orange square represents the cell types provided by reference data set. (*F*) The output of scHGR includes automatic cell annotation and differential expression gene (DEG) analysis based on propagation weight, revealing novel cell subtypes with biological significance and Gene Ontology (GO) enrichment.

representation algorithm, gene encoder integrates local topology and global gene network structure to generate 400-length vectors of each gene, referred to as the initial embedding of gene nodes (Fig. 1C). The cell encoder filters genes based on the dispersion of expression profiles. Using the dimensionality reduction algorithm, each cell is converted into a 400-length vector called the initial embedding of a cell node (Fig. 1D). An HGNN is defined as a directed weighted graph with nodes that are either cells or genes, whose initial embedding is determined by the corresponding encoder. Edges can be derived from gene expression profiles, gene regulations, and self-feedback loops (Methods). scHGR consists of three layers of HGNN (Fig. 1E). Based on the construction rules of HGNN topology (Methods), the edges of a one-layer HGNN contain only two types: gene–cell edges and gene–gene edges. In a three-layer HGNN, the feature aggregation pathways include gene–gene–cell, gene–cell–gene, gene–gene–gene, cell–gene–cell, and cell–gene–gene. In the last layer, the types of cell nodes are assigned by a multiclass classifier. To update HGNN's parameters and node embeddings until convergence or early stopping, the loss between predicted labels and true labels provided by reference is applied. scHGR supports functions involving cell annotation and subtype identification based on cell embeddings, and it also supports downstream tasks such as DEG analysis and GO enrichment thanks to gene propagation scores (Methods) (Fig. 1F). Supplemental Table S1 provides a comprehensive overview of all data sets involved in this study.

scHGR improves accuracy of cell annotations on intra–data sets

To assess the performance of scHGR, we compare it to a variety of cell annotation tools, including a decision-tree (DT)-based method (scHPL) (Michielsen et al. 2021), a neural network-based method (ACTINN) (Ma and Pellegrini 2020), a graph attentional architecture method (scGAC) (Cheng and Ma 2022), a graph neural network model (scDeepSort) (Shao et al. 2021), and a deep generative model (scVI) (Lopez et al. 2018). SCENIC utilizes gene relationships to infer cell types, so it also serves as a benchmarking approach (Aibar et al. 2017). In addition, the benchmark method also includes CellAssign (Zhang et al. 2019) and Garnett (Supplemental Note S1; Pliner et al. 2019). An analysis of large-scale single-cell atlases from different species is conducted to determine the effectiveness of cell annotation within data sets. Both the PBMC-FACS and Allen Mouse Brain (AMB) data sets are a large-scale single-cell transcriptome atlas with more than 10,000 cells. The PBMC-FACS possesses balanced cell types, whereas AMB from the mouse represents the scenarios in which the size of each type differs. The fivefold cross-validation strategy is applied for each data set to quantitatively measure the performance of annotation tools.

As shown in Figure 2, A and B, multiple metrics are considered to quantify the effectiveness of different annotation technologies comprehensively. For PBMC-FACS, all metrics lead to the same conclusion, that scHGR outperforms other state-of-the-art approaches. Particularly, the *F1*-score is 5% higher than the second place, indicating scHGR's strength in obviating false-positive samples. scVI follows scHGR, which is supposed to alleviate the inherent limitations of raw sequencing data owing to high sparsity. In AMB, scHGR's predictions are practically reliable with an average accuracy and Matthew's correlation coefficient (MCC) of 99%. The gene relationships in SCENIC are inferred from GENIE3 (Huynh-Thu et al. 2010), which affects its performance in cell classification tasks. Although scHGR exhibits minor improvements in accuracy and MCC over scVI, scHGR has more advantages in precision, recall, and *F1*-score. In AMB, the precision, recall, and *F1*-score of scHGR are 24%, 23%, and 23% higher, respectively, than those of scVI (Supplemental Fig. S1). Quantitatively, scHGR achieves a 6% improvement in precision, recall, and *F1*-score over competing methods, suggesting its effectiveness in annotating large-scale data sets. To validate the importance of integrating gene regulatory networks into HGNN in improving the accuracy of scHGR, we include scHGRS in the comparison methods. As a variant of scHGR, scHGRS omits gene relationships, and scHGRS relies only on gene expression profiles to construct an HGNN, which omits edges between genes; other architectures are consistent with scHGR. The results in PBMC-FACS and AMB demonstrate that scHGR improves over scHGRS in all metrics, especially recall and *F1*-score, suggesting that the integration of gene regulatory networks reduces false-negative rate.

Additionally, we investigate the reliability of assignment on specific cell populations. In the case of PBMC-FACS (Fig. 2C; Supplemental Fig. S2), scHPL incorrectly determines partial cells as root nodes of DT and fails to infer their type. scGAC mixes up $CD4^+/CD45RA^+/CD25^-$ naive T, $CD4^+/CD45RO^+$ memory, and $CD8^+/CD45RA^+$ naive cytotoxic, accounting for its low *F1*-score in Figure 2A. As subpopulations of T cells, all methods cannot discriminate well among $CD4^+$ T helper2, $CD4^+/CD25^-$ T reg, and $CD4^+/CD45RA^+/CD25^-$ naive T. Still, taking $CD8^+$ cytotoxic T and $CD8^+/CD45RA^+$ naive cytotoxic into account, scHGR is superior to scDeepSort as well as scVI. They mispredict more $CD4^+$ T helper2 cells as $CD4^+/CD45RA^+/CD25^-$ naive T compared with scHGR. Our approach is capable of accurately capturing significant differences between similar subpopulations of cells. For AMB (Fig. 2D; Supplemental Fig. S3), scHPL prefers to annotate rare cells belonging to large populations, such as L6 IT, whereas scGAC appears to mistakenly assign other cells to rare populations, such as Sncg. As a result, such tools would be inhibited by imbalanced cell types. scVI incorrectly annotates most Sncg cells as Vip cells and has minimal efficacy on Serpinf1, Astro, and VLMC cells. In contrast, the annotations of scHGR almost exactly match with reference labels, proving its accuracy and reliability.

scHGR enhances annotation robustness and reveals novel subtypes on inter–data sets

As single-cell sequencing technology has advanced, increasing amounts of data are being generated from a variety of experimental platforms and sequencing technologies. To demonstrate the annotation performance of scHGR for data sets with batch effect, we follow the same strategy as that in highly cited papers (Abdelaal et al. 2019; Lin et al. 2020; Song et al. 2021; Zeng et al. 2022). Therefore, we collect PBMC data sets from different resources (Ding et al. 2020), consisting of 10x chromium version 2 (PBMC1-10X2), 10x chromium version 3 (PBMC1-10X3), Smart-seq2 (PBMC1-SM2), CEL-seq2 (PBMC1-CS2), Drop-seq (PBMC1-DS), inDrops (PBMC1-D), and Seq-Well (PBMC1-SW) from human1. The PBMC data set from human2 based on 10x chromium version 2 (PBMC2-10X2) is also included here to assess the methods' effectiveness across samples. Because of the large number of cells and similar sequencing approaches for PBMC1-10X2 and PBMC1-10X3, they are adopted as reference data sets, and the others are treated as query data sets. Specifically, $X^{m\times n} = \{X_1^{m_1\times n}, X_2^{m_2\times n}\}$ denotes the reference expression matrix including PBMC1-10X2 and PBMC1-10X3, where n is the number of filtered genes, and m_1 and m_2 represent the cell number included in PBMC1-10X2 and

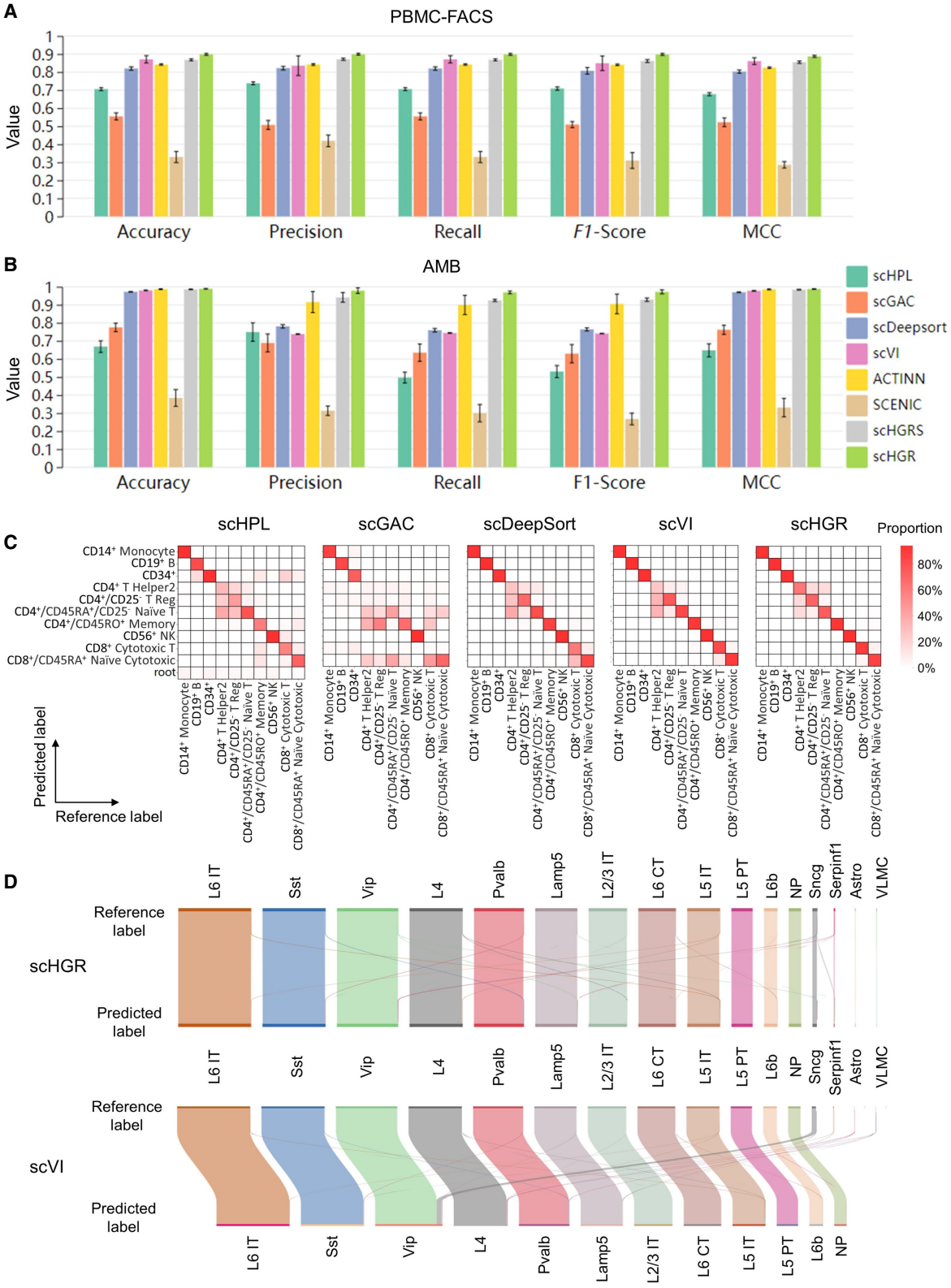


Figure 2. Performance comparison of scHGR and state-of-the-art methods on intra–data sets. Involving four state-of-the-art techniques, the experiments are conducted separately on PBMC-FACS (*A*) of human and AMB (*B*) of mouse and take the mean value of fivefold cross-validation to generate the histogram. Each group represents an evaluation metric, and each color indicates a method. The segment at the *top* of each bar represents the error segment of the corresponding method. (*C*) Heatmap of the confusion matrix of PBMC-FACS, with rows denoting predicted labels by the corresponding method and columns denoting labels provided by reference data sets. Notably, scHPL assigns some cells as root results from its tree architecture. (*D*) For AMB, the Sankey diagram compares cell type assignments between reference data sets and predicted labels from scHGR and scVI.

PBMC1-10X3, respectively. Meanwhile, their labels are concatenated as reference labels. Consistent with previous studies, we performed routine data preprocessing (Abdelaal et al. 2019) without including a batch correction step. To ensure fairness, we ran the benchmark methods with the same preprocessing steps as scHGR.

Gene regulatory relationships assist in automatic annotation under a cross-technology scenario

The experiments on cross-technology data sets reveal varying results (Fig. 3A): Almost perfect annotations on PBMC1-SM2 (scHGR: *F1*-score = 94%) illustrate high consistency between 10x and Smart-seq2, but the generally poor outcome when querying PBMC1-SW (*F1*-score < 60% for all methods except scHGR) implies the conflict from 10x data to Seq-Well data. Different from other data sets, all methods show outstanding *F1*-score and accuracy on PBMC2-10X2. This is because the query data set and reference data set are obtained by the same sequencing technology, avoiding the data deviation caused by technical differences. Apart from scHPL, the accuracies of all the methods differ slightly, exceeding 70%. Table 1 records the *F1*-score for all methods on cross-technology data sets. Accordingly, the average *F1*-score of scHGR is higher than those of other methods, proving the effectiveness of HGNN across technologies. The scRNA-seq data generated from different technologies, platforms, and operators exhibit intrinsic differences owing to measurement bias, technical settings, and other factors during sequencing. In such cases, cells belonging to the same population may exhibit distinct expression distribution characteristics. Instead, gene regulatory relationships avoid the interference introduced by the sequencing process. Relying on them constitutes the key component in determining cell identity and explains the high robustness achieved by scHGR.

Other than quantitative evaluations, expression profiles and cell representations from different approaches are visualized by the t-distributed stochastic neighbor embedding (t-SNE) algorithm (van der Maaten and Hinton 2008). Owing to the diversity of core architectures, some models do not provide cell representations, which are ignored here. In Figure 3B, the expression profile displays all cells overlapping so that it is impossible to distinguish different types. scGAC incorrectly mixes cytotoxic T cell and $CD4^+$ T cell in a mass. scVI is able to roughly separate each type, but in contrast, scHGR separates cytotoxic T cell and $CD4^+$ T cell more neatly, as do $CD14^+$ monocyte and B cell. Such a phenomenon verifies that the cell embedding provided by scHGR can not only effectively identify cell types with expression characteristics but also distinguish similar cell types.

scHGR identifies novel subtypes based on scRNA-seq of PBMC

scHGR not only automatically assigns cell identities but also reveals novel subpopulations by taking abundantly annotated atlases as references. By transferring the labels of reference data sets to query data sets, scHGR may reveal the cell types that are not discovered by previous research on query data sets. Combining PBMC1-10X2 and PBMC1-10X3 as reference data sets, scHGR uncovers new subtypes that are not contained in their benchmark annotations in PBMC1-D, PBMC1-SW, and PBMC1-SM2 (Fig. 3C; Supplemental Fig. S4). In Figure 3C, the cell type "Nature killer cell" assigned by scHGR is derived from reference labels. Furthermore, to verify whether the novel population Nature killer cells isolated from $CD4^+$ T cell and cytotoxic T cell is biologically meaningful, the DEGs covering 89 upregulated and 414 downregulated genes are filtered by edgeR (Fig. 3D; Robinson et al. 2010). Enrichment analysis of such DEGs reveals that they are functionally associated with host–virus interaction, cytokine receptor activity, etc. (*P*-value < 0.05), which is in accordance with the biological meanings of subtypes assigned by scHGR (Fig. 3E; Lee et al. 2001; Popko et al. 2013; Ngo et al. 2023).

scHGR supports cell type transfer in cross-species scenarios

With the aid of well-studied species information, cross-species automatic cell annotation can not only accelerate research progress on other species but also deepen our insights into evolutionary and developmental biology (Hodge et al. 2019; Shafer 2019; Wang et al. 2021). To assess the performance of all tools in cross-species scenarios, we employ human and mouse data sets that are thoroughly studied and precisely labeled, and they serve as background information and references for subsequent studies of other species (Baron et al. 2016). To maintain the consistency of genes in reference and query, the R package "homologene" is employed to extract homologous genes from them, which are utilized to calculate dispersion and constitute the hybrid graph skeleton (R Core Team 2022).

Together with "human–mouse" and "mouse–human" (Fig. 4A), scDeepSort and scVI show unstable performance with accuracy drop-offs of >20%, indicating that their effectiveness largely depends on the quality and scale of reference data sets. In contrast, scHGR remains stable and effective owing to its HGNN forming cross-species information flows through homologous genes. For human–mouse, scHPL is at least 17% lower in accuracy than other methods. scHGR achieves the highest accuracy, which is 2% higher than the second place. However, in mouse–human scenario, the predictions are generally unsatisfactory, caused by the insufficient number of cells in reference. Only scGAC and scHGR remain reliable under such inadequate training. There may be essential distinctions in the gene expression patterns of different species, making it more difficult to extract cell identity information from transcriptome features solely. In scHGR, transcriptomically distant cells are bridged by homologous genes and their relationships. Attaining homotypic cell communication across species facilitates improving the reliability of cross-species annotation.

Annotation performance on diabetes mellitus overcomes variations in cell proportions

Despite a considerable amount of research on healthy tissues and organs, there is still a need for more exploration and emphasis on diseases. Although there is a vast literature on diseases such as cancer, autoimmune disorders, and neurodegenerative conditions, it is important to continue expanding our understanding of various diseases through extensive research. Emerging diseases and epidemics have unclear mechanisms, and their expression patterns differ from healthy samples. Therefore, even experts may have difficulty labeling them. If an effective classifier is available by studying healthy cell atlases, automatic annotation of disease samples can undoubtedly accelerate the research process and uncover special gene expression patterns from a statistical perspective. Subsequently, we test the efficiency of annotation tools for such cross-disease scenarios, taking type 2 diabetes mellitus cells as an example (the human data sets serve as references) (Baron et al. 2016). The accuracy of both scVI and scHGR is as high as 94%, indicating their reliability under such conditions. Figure 4B compares the distribution of pancreas cells from healthy humans, patients with diabetes mellitus, and the prediction of scHGR. Intuitively, the proportions of some types vary obviously without

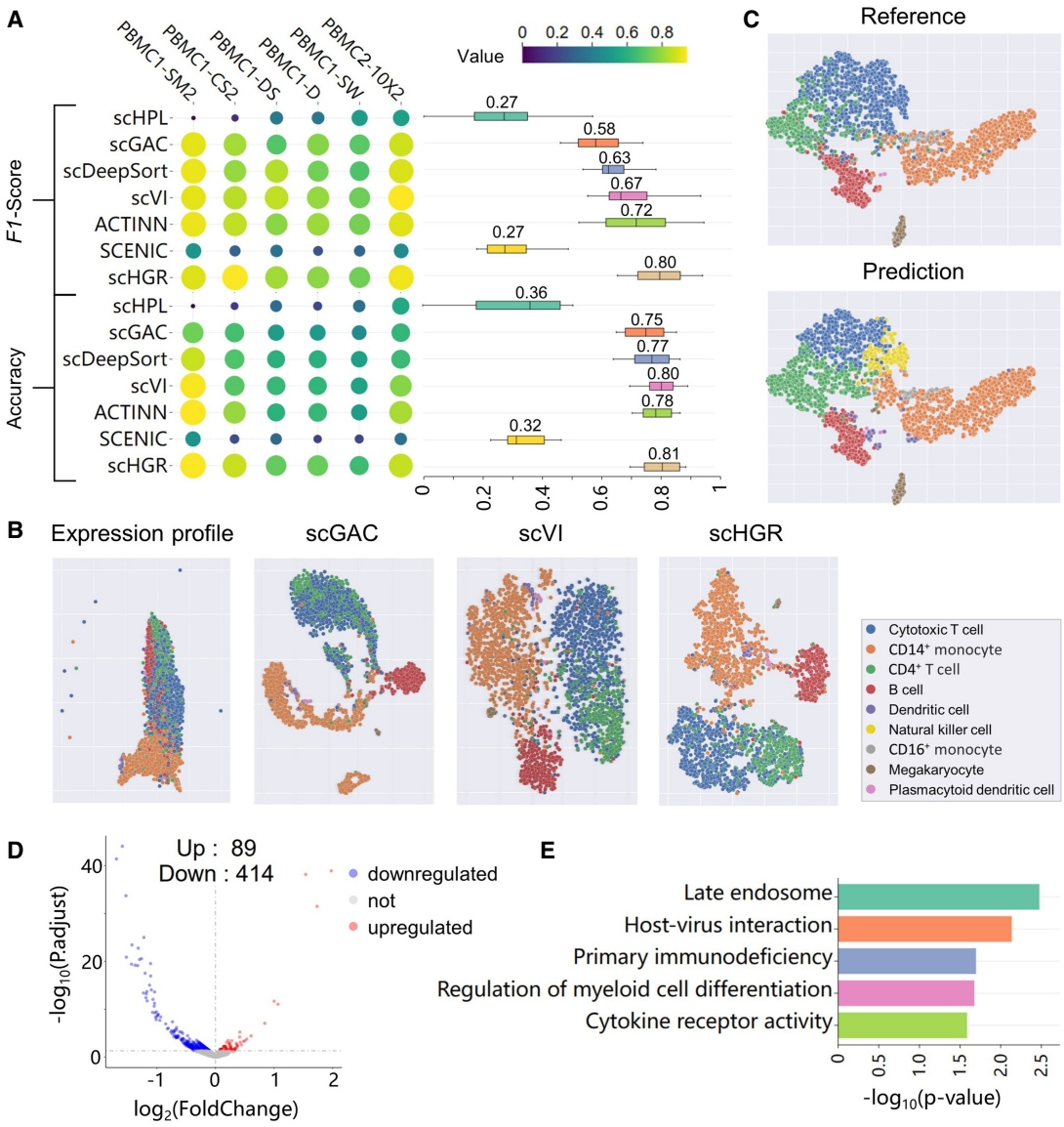


Figure 3. Performance evaluation of different tools across sequencing technologies and novel subtype revealing by scHGR. (*A*) Heatmap of the *F1*-score and accuracy resulting from cross-technology experiments, where rows represent methods and columns represent data sets. Box plots summarize the maximal, minimal, median, and quantile accuracy, as well as *F1*-scores for different methods. The median value is labeled *above* the corresponding box. (*B*) Colored according to reference data set labels, t-SNE plots the raw expression profiles of PBMC1-SW and cell embeddings generated from scGAC, scVI, and scHGR. scHPL and scDeepSort are not contained because cell embeddings are not available from their output. (*C*) t-SNE visualizes cell embeddings in PBMC1-D derived from scHGR. (*Top*) The figure is colored referring to the cell type provided by reference data sets. (*Bottom*) The figure is colored referring to the cell type predicted by scHGR. (*D*) Volcano plots of DEGs detected from the expression profiles in Nature killer cells. The number of upregulated and downregulated genes is marked on the graph. (*E*) The GO enrichment result of DEGs.

a consistent trend in diabetes mellitus. Beta and ductal rise by 10%, whereas alpha and acinar decline by >6%. Although taking human as reference, scHGR yields almost exactly correct results in diabetes mellitus, demonstrating that it is not misled by cell distribution in the reference atlas when making decisions. It can accurately extract cell type–specific expression patterns, which can reflect the disease mechanism as well.

scHGR maintains superiority in scenarios with diverse CDIs

As defined in Methods, the classification difficulty index (CDI) is used to measure the annotation difficulty of different reference-query scenarios (Tran et al. 2020; Zhang et al. 2022). Consistent with our expectations (Supplemental Table S2), cell annotation tasks have the highest difficulty in cross-species scenarios (average CDI = 0.61), followed by inter–data sets (average CDI = 0.54) and, finally, intra–data sets (average CDI = 0.13). It is because of the inherent noise introduced by different sequencing technologies and the considerable biological heterogeneity between different species increasing the challenges of cell type mapping. The performance of all methods in a total of 22 experimental scenarios is summarized in Figure 4C. The variation in accuracy indicates it is less difficult to perform label transfer on intra–data sets, with all methods remaining highly effective and stable. scGAC and

Table 1. The *F1*-score of different tools on cross-technology data sets

Data set	PBMC1-SM2	PBMC1-CS2	PBMC1-DS	PBMC1-D	PBMC1-SW	PBMC2-10X2	Average
scHPL	0.0000	0.1582	0.3336	0.2100	0.3568	0.5718	0.2717
scGAC	0.7419	0.6670	0.5317	0.5201	0.4620	0.6323	0.5925
scDeepSort	0.8707	0.6881	0.6056	0.6046	0.5388	0.6451	0.6588
scVI	0.9376	0.6965	0.6373	0.6246	0.5548	0.7750	0.7043
ACTINN	**0.9482**	0.8021	0.6104	0.6362	0.5247	0.8232	0.7241
SCENIC	0.4892	0.2298	0.3189	0.1793	0.2098	0.3564	0.2972
scHGR	0.9436	**0.8683**	**0.7225**	**0.7309**	**0.6553**	**0.8664**	**0.7978**

Bold indicates the maximum value in that column.

scHPL are inferior to scDeepSort, scVI, and scHGR, whose accuracy sustains at approximately one in mouse data sets. scHGR performs better on the AMB data set, which is cell type abundance imbalanced, compared with the cell type abundance balanced PBMC data set. The average CDI of PBMC-FACS and AMB is 0.2665 and 0.0023, respectively. Therefore, although AMB is a class-imbalanced data set, the classification difficulty of AMB is lower than that of PBMC-FACS from the perspective of batch effect and cell type accessibility. The following are inter–data set comparisons, in which most methods drop accuracy significantly. Using a Wilcoxon rank-sum test, the statistical significance *P*-values of scHGR, scDeepSort, scHPL, and scVI are 0.0014, 0.0312, 0.0003, and 0.0030, respectively. Performance fluctuates in cross-species data sets, with scDeepSort providing a 56% variance. Taking into account all data sets, scHGR has the highest accuracy, with an average accuracy of 0.88. The second place is occupied by scVI, with an average accuracy of 0.84. By evaluating the performance stability of each method in different scenarios with interquartile range (IQR), scHGR achieves first place and outperforms the second place by 2% (the IQRs of scHPL, scGAC, scDeepSort, scVI, ACTINN,

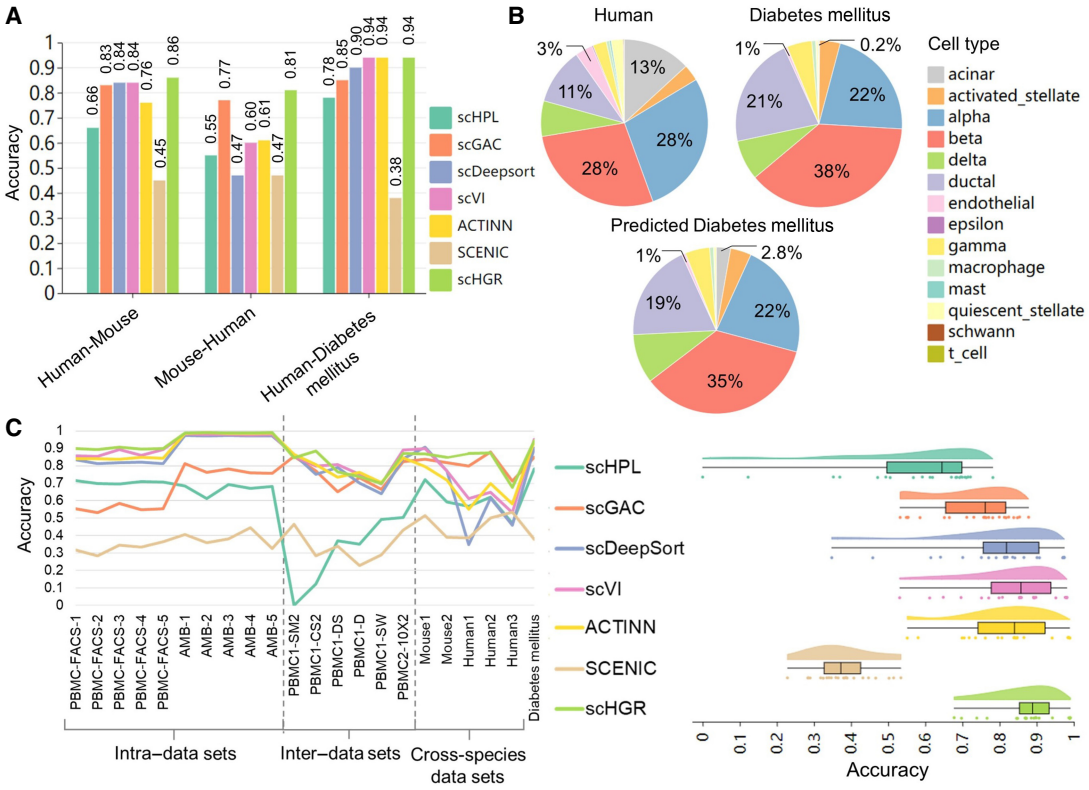


Figure 4. Performance comparison on multiple scenarios and summary of the overall experiments. (*A*) The bar chart exhibits the accuracy of all the approaches in three scenarios, which include annotating mouse cells by referring to human data sets (human–mouse) and vice versa (mouse–human) and annotating diabetes mellitus human cells by referring to human data sets (human–diabetes mellitus). The specific values are marked *above* the corresponding bars. (*B*) For human and diabetes mellitus, pie charts are produced based on their a priori labels and the percentage of each type. In contrast, the predicted diabetes mellitus is drawn based on labels annotated by scHGR. Comparing human and diabetes mellitus, five types with the highest ratio variation are marked with specific values at corresponding parts. (*C*) A total of 22 scenarios are divided into three groups: intra–data sets, inter–data sets, and cross-species data sets. The line chart reflects the variation in accuracy. Note that scHPL identifies all cells in PBMC1-SM2 as root, so it fails to annotate their types. The *right* chart contains an accuracy-based scatter plot (each point corresponds to a scene); a box plot that summarizes the maximal, minimal, and median scores, etc.; and a cloud plot that illustrates the distribution of accuracy values across various data sets.

SCENIC, and scHGR are 20%, 16%, 15%, 16%, 18%, 10%, and 8%, respectively).

The scHGR-guided cell atlas provides novel insights for COVID-19

The previous section verified the reliability of scHGR in inferring cell identities regarding disease. Based on the well-labeled healthy human lung atlas, we utilize scHGR to automatically annotate cells from COVID-19 patients. This is for further exploration of potential molecular mechanisms. The full molecular cell atlas of histologically normal lung tissues in humans is derived from Travaglini et al. (2020), which is set as the reference data set. The COVID-19 data sets are from Liao et al. (2020) and include six severe patients and three mild patients.

scHGR achieves cell annotation and marker gene identification for COVID-19

scHGR characterizes the COVID-19 data sets into 56 cell types as well as subtypes, covering a comprehensive range of categories, including immune cells, epithelial cells, etc. Comparing healthy, mild, and severe samples, the overall proportion of immune cells increases along with COVID-19 progression (Fig. 5A; Supplemental Tables S3–S5). We visualize different types of cells in latent space using t-SNE, UMAP, and PCA (Supplemental Fig. S5). Figure 5B demonstrates five increased cell types, in which the proportions of EREG$^+$ dendritic and myeloid dendritic type 1 double, intermediate monocyte, OLR1$^+$ classical monocyte, and proliferating macrophage are roughly three times higher in severe than in healthy samples. Meanwhile, scHGR also detects that some immune cells are reduced in severe than in mild, for example, CD4$^+$ memory/effector T, CD8$^+$ naive T, CD8$^+$ memory/effector T, and B (Supplemental Fig. S6), which is consistent with previous studies (Samprathi and Jayashree 2021). The fluctuations of such populations are significant biomarkers for diagnosing COVID-19 progression stages and assessing therapeutic responses.

scHGR not only has the ability to identify cells but also possesses the function of gene analysis. Owing to the gene-mediated cell communication network, scHGR distributes propagation scores to each gene while automatically annotating cells. The

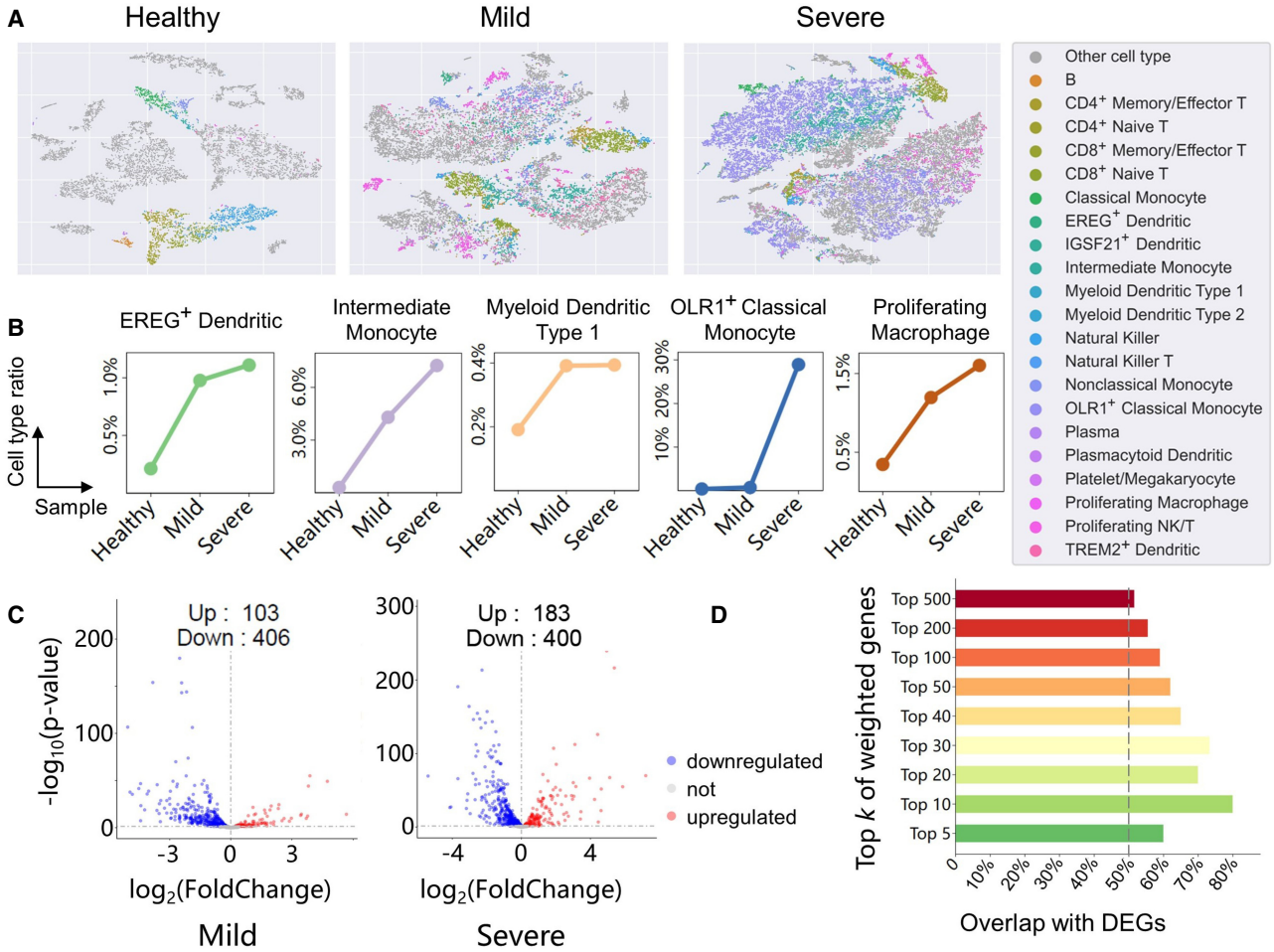


Figure 5. Cell annotation and DEG analysis for COVID-19. (*A*) t-SNE projection of COVID-19 data sets including healthy, mild, and severe samples. Colored points represent immune cells except for macrophage (for the clarity in figure, macrophage has been removed owing to its excessive number of cells). Each color represents one cell type, and gray points represent cells not belonging to immune types. (*B*) Because the number of cells in each data set differs, the line charts exhibit the ratio of cell populations instead of cell numbers. These populations consist of EREG$^+$ dendritic, intermediate monocyte, myeloid dendritic type 1, OLR1$^+$ classical monocyte, and proliferating macrophage, which correlates with the progression of COVID-19. The *x*-axis is samples, and the *y*-axis is ratios of cell population. (*C*) Volcano plots of DEGs detected from scRNA-seq data of mild and severe patients. The number of upregulated and downregulated genes is marked on the graph. (*D*) Bar chart compares the overlap of top *k* genes selected by scHGR with DEGs in severe patients. Supplemental Figure S7 shows the overlap corresponding to mild patients.

propagation score–based rank implies the importance of each gene in the current assignment. Specifically, *IL1B* ranks 70 places higher in severe than in mild, corresponding to the significant increase of such proinflammatory factors in patients with severe disease from the pathology. As for the validated vital factors for the pathogenesis of COVID-19 (Liao et al. 2020), including *IL1B*, *IL17C*, *CXCL9*, *CXCL11*, etc., they were targeted by scHGR at the top of the ranking. All of them are situated within the top 200 in mild and even within the top 100 in severe, indicating their functional expression is positively correlated with disease course (Supplemental Tables S6, S7).

Cell atlas and gene ranking provide insights into COVID-19 pathogenesis

A wealth of studies are available for COVID-19 either in terms of cell distributions or of gene functions, but the same gene may have distinct expression patterns in different populations, such as marker genes. Accordingly, we exploit cell type–specific expression to explore the crucial drivers that accelerate the progression of patients from mild to severe. Targeting cell types whose proportions follow a strict monotonic increase from healthy to mild to severe, we select genes with differential expression. Figure 5C illustrates that more DEGs are detected by edgeR from severe than mild in the above cell types, especially upregulated genes. The DEGs in severe patients tend to have a lower *P*-value (higher significance), indicating that their associated biological function is vital for the patient's condition. Intuitively, DEGs are not perfectly matched with top genes ranked by scHGR, which is caused by their different filtering criteria. Nevertheless, the overlap between top *k* genes ranked by scHGR and DEGs remains >50% in severe (Fig. 5D), and in mild, the overlap becomes lower with the increase of *k*-value (Supplemental Fig. S7).

To keep the overlap >50% in all samples, we separately collected the top 50 genes in mild and severe for GO enrichment analysis (Fig. 6). The GO terms and corresponding *P*-values can be found in Supplemental Table S8. Overall, the GO terms in mild and severe are partially identical and include functions relevant to apoptosis and immune response. In mild patients (Fig. 6A), the GO terms related to cell division, antigen binding, and inflammatory response predominate. Among them, "cellular response to interleukin 1" highlights the critical role of interleukin 1 (IL1) in the mild phase of COVID-19 because it performs a crucial effect in inducing cytokine storm in COVID-19 infection, and studies have validated that IL1 blockers are optimistic in reducing mortality (Mardi et al. 2021; Dimosiari et al. 2023). In severe patients (Fig. 6B), the enrichment of apoptosis and immune response is higher than that in mild patients, which is in line with what Figure 5C reveals. "Positive regulation of release of sequestered calcium ion into cytosol" rises the concentration of calcium ions and exacerbates the patient's olfactory dysfunction (Abdelazim et al. 2022). Therefore, giving medications that inhibit calcium ion increase may be an effective treatment to alleviate olfactory disorders. "Cellular response to interferon-gamma" reflects the fluctuation of interferon gamma (IFNG). As IFNG and tumor necrosis factor alpha (TNF) combine to induce inflammatory cell death (Karki et al. 2021), adopting neutralizing antibodies against IFNG and TNF is beneficial to patients' treatment.

Cell communication analysis reveals signaling pathways for mild and severe patients

After scHGR annotation of scRNA-seq data from COVID-19 patients, we combine CellChat (Jin et al. 2021) to investigate cell communication among cell types. Comparing the cell communication networks of mild and severe indicates the pathways in severe are more numerous and intricately connected than in mild, which is verified quantitatively by the number of interactions (Fig. 7A; Supplemental Fig. S8A,B). For concision and clarity, a total of 56 cell types are categorized into four groups (epithelial, immune, endothelial, and stromal populations) according to the tissue compartments (Travaglini et al. 2020). Figure 7, B and C, compares the cell communication among tissue compartments. Autocrine signaling between immune cells is predominant in mild; however in severe, immune interaction with the endothelial and stromal cells becomes primary signaling pathways. This is consistent with immune cell–epithelial cell interactions increasing the infectivity of epithelial cells as revealed by Chua et al. (2020). The important role of immune populations in both mild and severe determines the subsequent studies focusing on immune cell types (Supplemental Fig. S8C,D). The integrated immune cell communication network covering mild and severe indicates that a majority of cell communication converges on several cell types (Fig. 7D). From the aspect of network structure, OLR1$^+$ classical monocyte, CD8$^+$ naive T, CD4$^+$ memory/effector T, and intermediate monocyte may have main influence in the COVID-19 patients (Supplemental Fig. S9). From the aspect of pathways, RESISTIN is first secreted by classical monocyte and then drives CD8$^+$ naive T, classical monocyte, intermediate monocyte, macrophage, and OLR1$^+$ classical monocyte (Fig. 7G,H). These targeted types match the cell types with increased size as revealed in the subsection "scHGR achieves cell annotation and marker gene identification for COVID-19." The B, CD4$^+$ memory/effector T, and CD8$^+$ naive T cells coordinate the production of cyclophilin A (PPIA, also known as CYPA) signaling and further dominate plasma cells.

Based on the cell types filtered by scHGR, Figure 7, E and F, and Supplemental Figure S10 present the interaction of EREG$^+$ dendritic, intermediate monocyte, myeloid dendritic type 1, OLR1$^+$ classical monocyte, and proliferating macrophage with other cell types. The signaling pathways of all types increase with the progression of COVID-19, especially proliferating macrophage, intermediate monocyte, and OLR1$^+$ classical monocyte. It is a reflection of monocytes exhibiting increased cell communication in COVID-19 patients (Wahiduzzaman et al. 2022). Comparing the number of interactions involved in each signaling pathway, MIF and CYPA exceeded the others (Supplemental Fig. S11A). Notably, MIF consistently ranks first in both mild and severe experiments. Such a phenomenon indicates MIF serum levels and COVID-19 severity are highly correlated, which has also been concluded by previous studies (Armingol et al. 2021; Dheir et al. 2021; Wang et al. 2022). The communication network of MIF among immune cell types demonstrates that MIF signaling interacts primarily among natural killer, B, CD4$^+$ memory/effector T, CD8$^+$ naive T, and intermediate monocyte cells (Supplemental Fig. S11B). By applying the network centrality analysis, MIF signaling is primarily autocrine by B and CD8$^+$ naive T cells in mild. However, severe MIF signaling is mainly secreted by plasma cells and dominates CD8$^+$ naive T cells and proliferating NK/T cells (Fig. 7I,J; Supplemental Fig. S11C).

Discussion

Single-cell transcriptomics provides biologists with valuable insights into the cell atlas of tissues and organs. It helps them research the cellular heterogeneity that contributes to inter-individual variation, species evolution, and disease (Han et al.

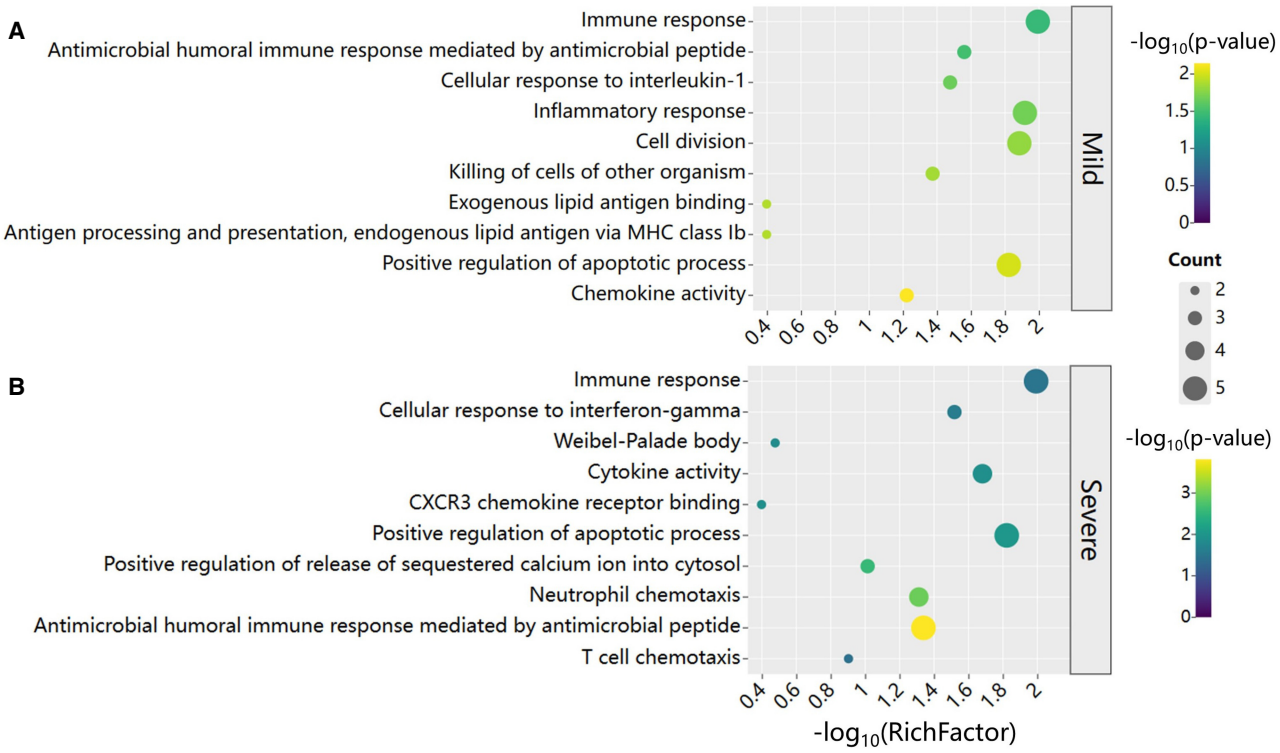


Figure 6. GO enrichment analysis for COVID-19 by scHGR. Based on the top 50 genes selected by scHGR, GO enrichment is performed separately in mild (*A*) and severe (*B*) patients. Rows represent GO terms, and columns represent $-\log_{10}$(RichFactor), where RichFactor = *n*/*N*, *n* is the number of genes belonging to GO term in gene list, and *N* is the number of genes belonging to GO term in genome. The color of bubbles represents the significance of GO term, and the size represents the number of genes enriched in GO item.

2020; Liu et al. 2023; Suoangbaji et al. 2023). Profiling cells is a critical step in processing single-cell data. Exact cell identity annotation is the basis for exploring cell subtypes, inferring differentiation trajectories, and other downstream analyses. We introduce an HGNN-based automatic annotation tool, scHGR, which integrates large-scale gene regulatory relationships with transcriptome data (Elyanow et al. 2020; Shao et al. 2021).

Conventional cell annotation techniques rely on gene expression profiles exclusively, without mentioning the essential role of gene regulatory networks in cell division and differentiation. Introducing data set–independent prior knowledge leads to such assignments based not only on transcriptomic expression patterns but also on gene regulatory relationships. Cells of the same type are thought to possess similar expression patterns in certain gene combinations; however, these genes may not follow easily recognized modes such as high expression but rather serve as communication mediators for cells that coexpress them. By integrating gene information flows with the cell–cell communication networks, scHGR is capable of emulating molecular mechanisms in which genes act as "bridges" in biological systems while achieving distantly related cellular communications.

Thus, scHGR is a single-cell resolution annotation tool that integrates gene regulatory networks into cell identity inference. For scRNA-seq data sets from various technologies, tissues, and diseases, scHGR targets representative genes based on distribution divergences between a reference and query. It embeds the high-dimensional raw data into low-dimensional nodes via cell encoders. Because gene relationships are intrinsically a network structure, a graph-embedding technique is applied to transform it nonlinearly into gene nodes. These nodes are ensured to be within the same potential space as cell nodes. By combining gene regulations with cell nodes to compose HGNN with additional constraints (node embeddings, edge weights, etc.), multiple perspectives have been addressed to mine cell type–specific expression patterns. Last but not least, the gene and cell embeddings are updated simultaneously during training, demonstrating that scHGR can not only reveal cell subtypes but also competently handle DEG analysis, GO enrichment, and other downstream tasks.

A series of experiments have shown that scHGR outperforms the latest tools, regardless of annotation reliability or stability. scHGR improves annotation robustness in data sets with batch discrepancies. In the case of type 2 diabetes mellitus, although its cellular components differ significantly from the reference (*P*-value = 0.0001), scHGR has an annotation accuracy of 94%, indicating the cell embedding it uncovers is cell state perceptible. On this basis, scHGR provides exhaustive annotation of cells from COVID-19 patients covering 56 types as well as subtypes. Benefiting from the introduction of gene regulatory networks, it emphasizes the dominance of positive regulation on calcium ion concentration in severe patients and inspires the treatment for olfactory dysfunction.

scHGR provides a low-dimensional representation for each cell, which is the cell node feature from the third layer of HGNN. Therefore, scHGR also facilitates cell-relevant analyses not covered in this study, such as differentiation trajectory inference and pseudotime assignment. In addition, scHGR provides representation and propagation score for each gene. This means we can smoothly conduct gene-associated downstream tasks, such as cell type–

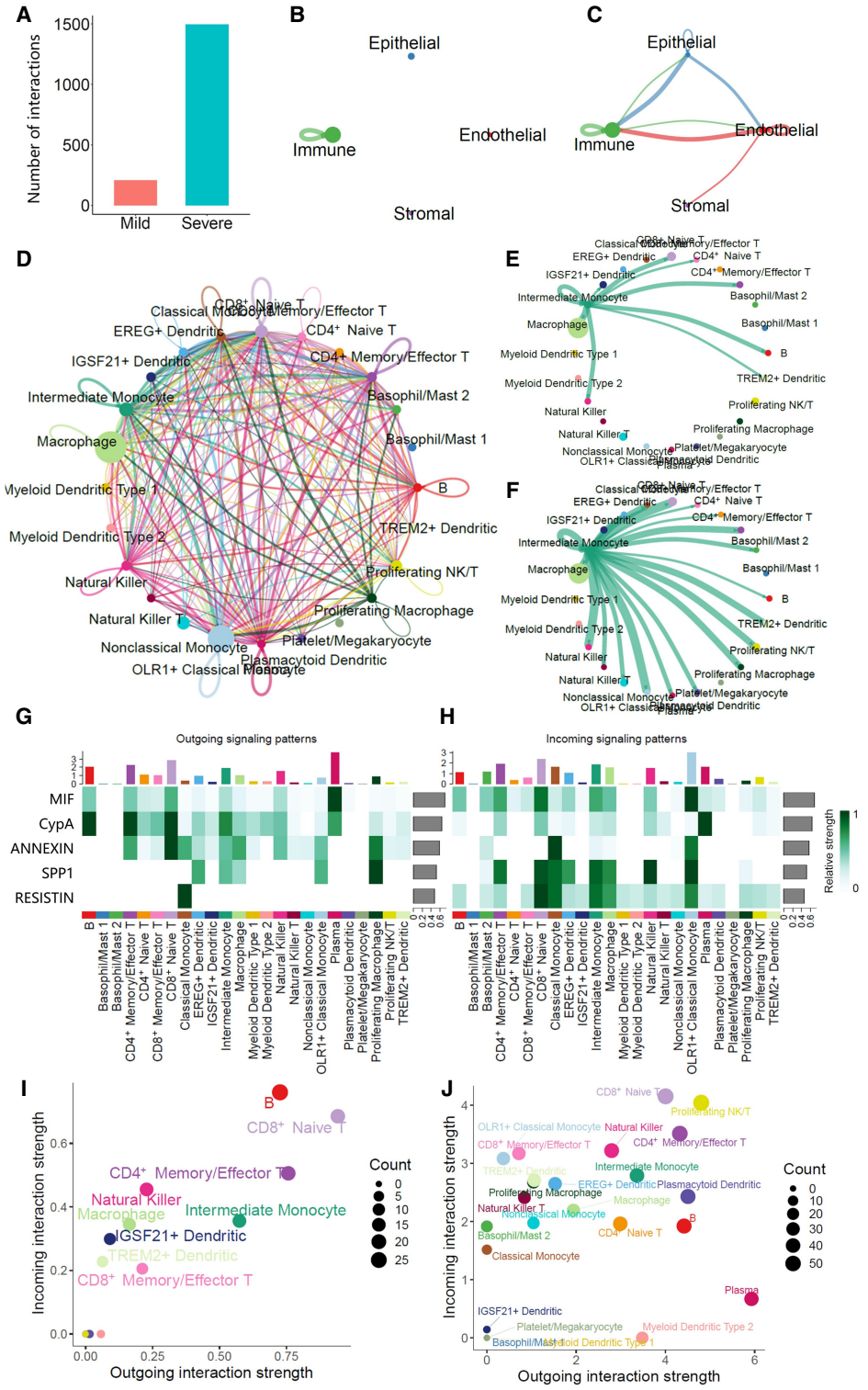


Figure 7. Comprehensive analysis of the cell communication network for COVID-19. (*A*) The number of interactions of the cell communication networks. (*B*,*C*) Cell communication networks at tissue compartment level for mild and severe patients, respectively. (*D*) Cell communication network of immune cell types based on integrated data. (*E*,*F*) Visualization of signaling pathways sending from intermediate monocyte cells for mild (*E*) and severe (*F*). Circle sizes are proportional to the number of cells in each cell type, and edge width represents the communication number. Edge colors are consistent with the source cell type. (*G*,*H*) The outgoing or incoming heatmap of each cell type to the inferred signaling pathways. The color bar represents relative signaling strength of a signaling pathway across cell types. The *top* colored bar plot shows the total signaling strength of a cell type by summarizing all signaling pathways displayed in the heatmap. The *right* gray bar plot shows the total signaling strength of a signaling pathway by summarizing all cell types displayed in the heatmap. (*I*,*J*) Total incoming and outgoing signaling strengths in different cell types for mild (*I*) and severe (*J*). The size of points matches the count of interactions.

specific expression pattern analysis, GO enrichment, etc., which are not supported by other cell annotation tools. With minor changes to encoder modules, this framework can be extended to multi-omics annotation and data integration, benefiting from the gene-mediated cell communication network architecture's scalability. Although equipped with a large-scale gene regulation repository, scHGR also offers an interface for additional lists of gene relationships. In conclusion, we believe that scHGR is a powerful tool for annotating single-cell data, especially in gene-combined tasks.

Methods

Gene regulation repository

scHGR is equipped with a variety of validated gene regulatory relationships. In contrast to cell type–specific and tissue-specific gene regulatory networks, the extracted gene relationships are prevalent within the same species and do not vary with cell type and tissue. Gene regulations play a major role in the encoding of gene nodes and the construction of hybrid graph networks. The gene regulatory relationships are collected from public databases and platforms, including TRRUST, RegNetwork, BioGRID, and GREDB. We assess the feasibility of using different genetic relationships for each specific query data set in Supplemental Note S2.

TRRUST is a manually annotated database that efficiently extracts regulatory relationships from more than 20 million published articles, involving human and mouse data (Han et al. 2018). Developed based on 25 databases, RegNetwork integrates in-library regulatory relations and potential relationships inferred based on transcription factor binding sites (Liu et al. 2015). Only experimentally validated gene regulations were adopted in scHGR. BioGRID is a biomedical interactive repository that includes focused low-throughput studies and large high-throughput data sets involving 80,848 papers and multiple species (Oughtred et al. 2019). GREDB is a publicly available, manually managed biological database that preserves human gene expression regulatory relationships (Huang et al. 2019). scHGR collected a number of gene regulatory relationships that are prevalent in human and mouse from the above sources to constitute the gene regulation repository (Fig. 1B), which comprises 26,842 genes and 918,880 relationships for human and 29,670 genes and 2,004,761 relationships for mouse.

Gene encoder

Because gene expression pattern is highly correlated with cell differentiation and biological function, scHGR introduces the key element that influences gene expression levels, the gene regulatory network, into the task of cell identity annotation.

To integrate the intricate relationships among genes into gene nodes, we apply a classical semisupervised graph embedding strategy termed structure deep network embedding (SDNE). SDNE is a deep learning–based graph representation algorithm for mapping graph structural information to continuous vector space. Given the graph adjacency matrix, SDNE employs an autoencoder containing two hidden layers to make the representations of neighboring nodes similar. It enables connected nodes to form similar embeddings to capture the local graph topology, corresponding to the loss function below:

$$L_{1st} = \sum_{i,j=1}^{n} S_{i,j} \|g_i - g_j\|_2^2,$$

where S is the graph adjacency matrix, and n is the number of nodes in the graph. g_i and g_j denote the embedding of nodes i and j, respectively. As a 400-dimensional vector, g is the output of gene encoder and initializes gene node features in the HGNN.

In addition, SDNE causes nodes with similar neighbor structures to possess similar embeddings to preserve global graph structure. The loss is defined as

$$L_{2nd} = \sum_{i=1}^{n} \|y_i - x_i\|_2^2,$$

where $x_i = s_i$, which means the ith row of adjacency matrix is used as the initial feature of node i to integrate the neighbor information. y_i is the reconstructed feature of node i after decoding. As a result, the final loss function of the gene encoder is

$$Loss = L_{2nd} + L_{1st} + L_{reg},$$

where L_{reg} denotes a regularization term to avoid overfitting.

Cell encoder

For each data set, we remove cell populations containing fewer than 10 cells to ensure the deep neural network can train adequately. In terms of gene filtering, different strategies were adopted depending on the data distribution. Above all, all genes are ranked by dispersion, which is written as

$$Dispersion(g) = \frac{Variance(g)}{Mean(g)}.$$

If the reference and query data follow a close distribution, that is, their expression profiles differ within fivefold (Fig. 1D), the top 1000 genes are selected, and original reads are taken as the input of scHGR. Otherwise, the top 5000 genes are kept, and gene reads require log-normalization at first. For each cell, the principal component analysis (PCA) algorithm is employed to transform the high-dimensional raw sequencing data into 400-dimensional vectors as the initial feature of cell nodes in HGNN.

Hybrid graph neural network

As the basis of cell type transfer, a critical step in scHGR is constructing the HGNN. This integrates reference and query data into a unified latent space. Let $X^{m\times n} = \{X_r^{m_r\times n}, X_q^{m_q\times n}\}$ denote the joint expression matrix including reference and query scRNA-seq data, where n is the number of filtered genes (1000 or 5000 depending on data distribution), and m_r and m_q represent the cell number included in the reference and query data sets, respectively.

Both genes and cells are nodes in this hybrid graph. The adjacency matrix $E^{(m+n)\times(m+n)}$ is determined according to the following rules: First, if the expression level of gene a in cell b is w ($w>0$), then there is a bidirectional edge with weight w between nodes a and b. Second, based on the gene regulation repository module, if gene a regulates gene b, there is an edge from node a to b with a weight equal to the maximum value in reference expression matrix. It can be formulated as

$$e_{a,b} = \mathrm{Max}(X_r^{m_r\times n}),$$

where $e_{a,b} \in E^{(m+n)\times(m+n)}$ denotes the weight of edge from a to b. Third, each node has a self-loop to reflect the cell's own biological signal and gene self-regulation. The output of cell encoder and gene encoder constitutes the initial feature of nodes.

We customize HGNN that takes into account inter-cell and inter-gene information propagation. The embedding of node a in the $(k+1)th$ layer can be formulated as

$$h_a^{k+1} = \sigma(W^k \times MEAN(h_a^k) + b^k),$$

where W^k and b^k mean the learnable shared weight matrix and bias of layer k, and σ means the rectified linear unit (ReLU) activation function to explore the nonlinear correlation. $MEAN(h_a^k)$ represents an improved node embedding update approach that integrates neighbor node information with weight. The formula is as follows:

$$MEAN(h_a^k) = \frac{h_a^k + \sum_{b \in Neighbor(a)} \alpha_b \times h_b^k \times e_{a,b}}{1 + |Neighbor(a)|},$$

where $|Neighbor(a)|$ denotes the number of neighbors corresponding to node a, and α_b denotes the propagation score of node b. As a learnable parameter, α_b is used to tune the importance of each gene or cell in node embedding updating. HGNN accounts for cell and gene heterogeneity. On one hand, the HGNN takes into account the heterogeneity of gene and cell nodes in its topology. Despite using a similar method for integrating neighbor information, the aggregation process for gene nodes and cell nodes is different owing to their distinct neighbor compositions. On the other hand, α in the formula $MEAN(h_a^k)$ represents the propagation score for each node and serves as a node-specific trainable parameter. Therefore, the use of different α values for genes and cells addresses the issue of them potentially not belonging to the same latent space. The cell embedding derived from the third layer of HGNN is followed by a linear classifier layer, and the cross-entropy loss is applied to optimize parameters:

$$y_a = softmax(Wh_a^{last} + b),$$
$$L_{CE} = -\sum_{a=1}^{m_r} l_a \log y_a,$$

where h_a^{last} represents the embedding of node a in the last layer, and l_a represents the label of cell a provided by reference data set.

Experiment setting

Notably, because query data cells are ignored by the mask, they do not participate in loss minimization. In addition, we apply a mini-batch strategy to avoid excessive memory occupation and an early stop strategy to avoid overfitting in the model training stage. Specifically, HGNN trains only one batch at a time, feeds in the adjacency matrix and node features associated with this batch, and updates the whole graph by training the local models in turn. Instead of specifying the epoch number, model training is terminated once the accuracy of training set achieves 100% or the accuracy of validation set does not improve during 50 consecutive epochs. Upon completion of training, scHGR automatically annotates query cells within seconds. We evaluate different numbers of hidden layers and different dimensions of node embedding. By using the same experimental strategy as in the section "scHGR enhances annotation robustness and reveals novel subtypes on inter–data sets," Supplemental Figure S12 demonstrates that the model achieves optimal performance when the hidden layer is two and the node dimension is 400. In this study, we train a three-layer HGNN with an input unit of 400 and hidden unit of 200. As a default, we set a maximum of 1000 epochs and use the Adam optimizer with a learning rate of 0.05.

Evaluation metrics

To comprehensively evaluate the performance of scHGR, we apply a variety of statistical indicators to measure the similarity between annotation from methods and cell labels from data sets. Let $L = \{L_1, L_2, \ldots, L_n\}$ represent cell labels containing n cells, and let $P = \{P_1, P_2, \ldots, P_n\}$ represent predicted annotation of n cells. The calculation of accuracy depends on whether the prediction of each sample strictly matches its label, and it is formulated as

$$Accuracy(L, P) = \frac{1}{n}\sum_{i=1}^{n} 1(L_i = P_i).$$

In a study involving multiple cell types, it is common practice to calculate the statistical metrics separately for each cell type and then compute the average across all types as the final evaluation result. For any cell type j, considering a binary classification task of all cells, the following indicators can be obtained:

- True positive (TP_j)—cells with label j and predicted result j;
- False positive (FP_j)—cells with label not j but predicted result j; and
- False negative (FN_j)—cells with label j but predicted result not j.

Based on the above statistics, the precision, recall rate and *F1*-score of j can be directly calculated:

$$Precision_j = \frac{TP_j}{TP_j + FP_j},$$
$$Recall_j = \frac{TP_j}{TP_j + FN_j},$$
$$F1_j = \frac{2 \times Precision_j \times Recall_j}{Precision_j + Recall_j}.$$

Finally, the average value of all types is taken as the final metric, which is formulated as follows:

$$Precision = \frac{1}{m}\sum_{j=1}^{m} Precision_j,$$
$$Recall = \frac{1}{m}\sum_{j=1}^{m} Recall_j,$$
$$F1 = \frac{1}{m}\sum_{j=1}^{m} F1_j,$$

where m denotes the total number of cell types annotated by algorithm. The MCC is calculated according to the formula defined in scikit-learn (Pedregosa et al. 2011):

$$MCC = \frac{c \times n - \sum_{j=1}^{m} p_j \times t_j}{\sqrt{\left(n^2 - \sum_{j=1}^{m} p_j^2\right) \times \left(n^2 - \sum_{j=1}^{m} t_j^2\right)}},$$

where c denotes the number of cells correctly predicted, t_j denotes the number of cells with label j, and p_j denotes the number of cells with prediction j.

Referring to the method of Zhang et al. (2022), we apply the CDI to assess the difficulty of cell annotation in different scenarios. CDI is calculated based on batch effect and cell type accessible ratio, which are the major factors affecting cell type transfer from reference to query. The formula for CDI is as follows:

$$CDI = \sqrt[2]{\frac{(1 - F1_{ASW})^2 + (1 - P)^2}{2}},$$

where P refers to the proportion of cells whose types are covered by reference data sets in the query data sets and $F1_{ASW} = \frac{2(1 - ASW_{batch\_norm})(ASW_{celltype\_norm})}{1 - ASW_{batch\_norm} + ASW_{celltype\_norm}}$ (Tran et al. 2020). Higher CDI means more difficult cell annotation tasks. The IQR is applied to compare the stability of results from different

methods. It is defined as follows:

$$IQR = Q_3 - Q_1,$$

where Q_3 is the third quartile, and Q_1 is the first quartile. In this study, statistical significance is defined using the Wilcoxon signed-rank test by default. The Wilcoxon rank-sum test is applied to variables with different sample sizes.

scRNA-seq data sets

All data used in this study is publicly available, their statistics information is presented in Supplemental Table S1. The raw data of PBMC-FACS, AMB, PBMC1-10X2, PBMC1-10X3, PBMC1-SM2, PBMC1-CS2, PBMC1-DS, PBMC1-D, PBMC1-SW, and PBMC2-10X2 can be downloaded from Zenodo (https://doi.org/10.5281/zenodo.3357167). The scRNA-seq data of human lung atlas are available on Synapse (https://www.synapse.org/#!Synapse:syn21041850). The human, mouse and diabetes mellitus data sets are available in the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE84133. The COVID-19 data sets involving mild patients and severe patients can be downloaded in GEO under the accession number GSE145926.

PBMC-FACS

The fluorescence-activated cell sorting-based peripheral blood mononuclear cells (PBMC-FACS) is from Zheng et al. (2017), who profiled 68k cells and generated well-labeled reference transcriptomes. We followed the same preprocessing steps taken in a prior analysis of this data set (Abdelaal et al. 2019), which covered cell filter and gene deletion. After preprocessing, the final data set contains 20,000 cells and 21,952 genes with balanced cell types.

AMB

The Allen Mouse Brain (AMB) data set is from Tasic et al. (2018), who reported 23,822 single-cell transcriptomes by Smart-seq v4 kit with cluster-assigned identity. Following the preprocessing steps by Abdelaal et al. (2019), the final data set contains 12,832 cells and 42,625 genes and represents an unbalanced annotation task involving rare cell types of 11 cells and large populations of up to 1848 cells.

Cross-technology PBMC data sets

We collect a series of PBMC scRNA-seq data sets from Ding et al. (2020), who compared seven scRNA-seq technologies on human peripheral blood mononuclear cells with two experiments (PBMC1 and PBMC2). Following the preprocessing steps by Abdelaal et al. (2019), the final data sets cover seven scRNA-seq technologies. The number of genes in each data set is consistently 33,694, but the cell numbers differ largely. More detailed information can be found in Supplemental Table S1.

Cross-species data sets

Both the mouse and human data sets are from Baron et al. (2016), who conducted a droplet-based scRNA-seq method to profile the transcriptomes of more than 12,000 individual pancreatic cells. After preprocessing, the mouse data sets consist of two healthy pancreas samples (Mouse1, Mouse2), involving 14,861 genes and 1886 cells, labeled as 13 unbalanced cell populations. The human data sets consist of three healthy pancreas samples (Human1, Human2, Human3), involving 17,499 genes and 7266 cells, labeled as 14 unbalanced cell populations. In addition, a pancreas cell atlas is reported for a type 2 diabetes mellitus (T2D) patient, which involves 17,499 genes and 1303 cells.

The cell atlas of lung tissues

The full molecular cell atlas of histologically normal lung tissues in humans is derived from Travaglini et al. (2020), who applied droplet- and plate-based scRNA-seq of approximately 75,000 human cells across all lung tissue compartments and created an extensive cell atlas of the human lung. After preprocessing, the final data set covers 65,662 cells and 26,484 genes with 57 transcriptionally distinct cell populations.

COVID-19 data sets

Liao et al. (2020) profiled bronchoalveolar lavage fluid cells from patients with varying severity of COVID-19 by applying chromium single-cell 3g reagent kits v.2 (10x Genomics). After preprocessing, the COVID-19 data sets include 63,246 cells and 33,539 genes covering six severe patients and three mild patients.

Software availability

The scHGR algorithm is implemented in Python and is available on GitHub (https://github.com/MengyuanZhaoo/scHGR) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (NSFC 62322215, 62172296), the Shenzhen Science and Technology Program (no. KQTD20200820113106007), and the Excellent Young Scientists Fund in Hunan Province (2022JJ20077). This study was supported in part by the high-performance computing center of Central South University. It was also supported in part by the high-performance computing clusters (PL-17161) of Shenzhen Institutes of Advanced Technology.

Author contributions: M.Z. conducted the experiments and wrote the manuscript. J.L. collected the data set used in this research. X.L. and K.M. contributed to the refinement of the manuscript. J.T. and F.G. provided guidance throughout the entire research and writing process.

References

Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A. 2019. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* **20:** 194. doi:10.1186/s13059-019-1795-z

Abdelazim MH, Abdelazim AH, Ismaiel WF, Alsobky ME, Younes A, Hadeya AM, Ramzy S, Shahin M. 2022. Effect of intra-nasal nitrilotriacetic acid trisodium salt in lowering elevated calcium cations and improving olfactory dysfunction in COVID-19 patients. *Eur Arch Otorhinolaryngol* **279:** 4623–4628. doi:10.1007/s00405-022-07424-5

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14:** 1083–1086. doi:10.1038/nmeth.4463

Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. 2021. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* **39:** 1202–1215. doi:10.1038/s41587-021-00895-7

Armingol E, Officer A, Harismendy O, Lewis NE. 2021. Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* **22:** 71–88. doi:10.1038/s41576-020-00292-x

Armingol E, Baghdassarian HM, Martino C, Perez-Lopez A, Aamodt C, Knight R, Lewis NE. 2022. Context-aware deconvolution of cell–cell

communication with tensor-cell2cell. *Nat Commun* **13:** 3665. doi:10.1038/s41467-022-31369-2

Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al. 2016. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intracell population structure. *Cell Syst* **3:** 346–360.e4. doi:10.1016/j.cels.2016.08.011

Cheng Y, Ma X. 2022. scGAC: a graph attentional architecture for clustering single-cell RNA-Seq data. *Bioinformatics* **38:** 2187–2193. doi:10.1093/bioinformatics/btac099

Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, Debnath O, Thürmann L, Kurth F, Völker MT, et al. 2020. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat Biotechnol* **38:** 970–979. doi:10.1038/s41587-020-0602-4

Dheir H, Yaylaci S, Sipahi S, Genc AC, Cekic D, Tuncer FB, Cokluk E, Kocayigit H, Genc AB, Salihi S, et al. 2021. Does macrophage migration inhibitory factor predict the prognosis of COVID-19 disease? *J Infect Dev Ctries* **15:** 398–403. doi:10.3855/jidc.14009

Dimosiari A, Patoulias D, Pantazopoulos I, Zakynthinos E, Makris D. 2023. Safety and efficacy of interleukin-1 antagonists in hospitalized patients with COVID-19. *Eur J Intern Med* **109:** 117–119. doi:10.1016/j.ejim.2022.11.014

Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT, et al. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* **38:** 737–746. doi:10.1038/s41587-020-0465-8

Elyanow R, Dumitrascu B, Engelhardt BE, Raphael BJ. 2020. netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res* **30:** 195–204. doi:10.1101/gr.251603.119

Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, et al. 2018. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* **46:** D380–D386. doi:10.1093/nar/gkx1013

Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, et al. 2020. Construction of a human cell landscape at single-cell level. *Nature* **581:** 303–309. doi:10.1038/s41586-020-2157-4

Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O, et al. 2019. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573:** 61–68. doi:10.1038/s41586-019-1506-7

Huang T, Huang X, Shi B, Yao M. 2019. GEREDB: gene expression regulation database curated by mining abstracts from literature. *J Bioinform Comput Biol* **17:** 1950024. doi:10.1142/S0219720019500240

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5:** e12776. doi:10.1371/journal.pone.0012776

Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, Myung P, Plikus MV, Nie Q. 2021. Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12:** 1088. doi:10.1038/s41467-021-21246-9

Kalucka J, de Rooij LPMH, Goveia J, Rohlenova K, Dumas SJ, Meta E, Conchinha NV, Taverna F, Teuwen L-A, Veys K, et al. 2020. Single-cell transcriptome atlas of murine endothelial cells. *Cell* **180:** 764–779.e20. doi:10.1016/j.cell.2020.01.015

Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. 2023. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614:** 742–751. doi:10.1038/s41586-022-05688-9

Karki R, Sharma BR, Tuladhar S, Williams EP, Zalduondo L, Samir P, Zheng M, Sundaram B, Banoth B, Malireddi RKS, et al. 2021. Synergism of TNF-α and IFN-γ triggers inflammatory cell death, tissue damage, and mortality in SARS-CoV-2 infection and cytokine shock syndromes. *Cell* **184:** 149–168.e17. doi:10.1016/j.cell.2020.11.025

La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, et al. 2021. Molecular architecture of the developing mouse brain. *Nature* **596:** 92–96. doi:10.1038/s41586-021-03775-x

Lee SH, Girard S, Macina D, Busa M, Zafer A, Belouchi A, Gros P, Vidal SM. 2001. Susceptibility to mouse cytomegalovirus is associated with deletion of an activating natural killer cell receptor of the C-type lectin superfamily. *Nat Genet* **28:** 42–45. doi:10.1038/ng0501-42

Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, et al. 2020. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* **26:** 842–844. doi:10.1038/s41591-020-0901-9

Lin Y, Cao Y, Kim H, Salim A, Speed T, Lin D, Yang P, Yang J. 2020. scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol Syst Biol* **16:** e9389. doi:10.15252/msb.20199389

Liu Z-P, Wu C, Miao H, Wu H. 2015. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database (Oxford)* **2015:** bav095. doi:10.1093/database/bav095

Liu X, Shen Q, Zhang S. 2023. Cross-species cell-type assignment from single-cell RNA-seq data by a heterogeneous graph neural network. *Genome Res* **33:** 96–111. doi:10.1101/gr.276868.122

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15:** 1053–1058. doi:10.1038/s41592-018-0229-2

Ma F, Pellegrini M. 2020. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* **36:** 533–538. doi:10.1093/bioinformatics/btz592

Mardi A, Meidaninikjeh S, Nikfarjam S, Majidi Zolbanin N, Jafari R. 2021. *Interleukin-1* in COVID-19 infection: immunopathogenesis and possible therapeutic perspective. *Viral Immunol* **34:** 679–688. doi:10.1089/vim.2021.0071

Michielsen L, Reinders MJT, Mahfouz A. 2021. Hierarchical progressive learning of cell identities in single-cell data. *Nat Commun* **12:** 2799. doi:10.1038/s41467-021-23196-8

Miller BF, Bambah-Mukku D, Dulac C, Zhuang X, Fan J. 2021. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Res* **31:** 1843–1855. doi:10.1101/gr.271288.120

Ngo HT, Dang VT, Nguyen NH-T, Bui AN-T, Pham PV. 2023. Comparison of cytotoxic potency between freshly cultured and freshly thawed cytokine-induced killer cells from human umbilical cord blood. *Cell Tissue Bank* **24:** 139–152. doi:10.1007/s10561-022-10022-8

Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, et al. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* **47:** D529–D541. doi:10.1093/nar/gky1079

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* **12:** 2825–2830.

Pliner HA, Shendure J, Trapnell C. 2019. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* **16:** 983–986. doi:10.1038/s41592-019-0535-3

Popko K, Malinowska I, Gorska E, Stelmaszczyk-Emmel A, Demkow U. 2013. Flow cytometry in detection of abnormalities of natural killer cell. *Adv Exp Med Biol* **756:** 303–311. doi:10.1007/978-94-007-4549-0_37

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. 2017. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171:** 1611–1624.e24. doi:10.1016/j.cell.2017.10.044

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Samprathi M, Jayashree M. 2021. Biomarkers in COVID-19: an up-to-date review. *Front Pediatr* **8:** 607647. doi:10.3389/fped.2020.607647

Semrau S, Goldmann JE, Soumillon M, Mikkelsen TS, Jaenisch R, van Oudenaarden A. 2017. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat Commun* **8:** 1096. doi:10.1038/s41467-017-01076-4

Shafer MER. 2019. Cross-species analysis of single-cell transcriptomic data. *Front Cell Dev Biol* **7:** 175. doi:10.3389/fcell.2019.00175

Shao X, Yang H, Zhuang X, Liao J, Yang P, Cheng J, Lu X, Chen H, Fan X. 2021. scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res* **49:** e122. doi:10.1093/nar/gkab775

So E, Hayat S, Nair SK, Wang B, Haibe-Kains B. 2023. GraphComm: a graph-based deep learning method to predict cell-cell communication in single-cell RNAseq data. bioRxiv doi:10.1101/2023.04.26.538432

Song Q, Su J, Zhang W. 2021. scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat Commun* **12:** 3826. doi:10.1038/s41467-021-24172-y

Suoangbaji T, Zhang VX, Ng IO-L, Ho DW-H. 2023. Single-cell analysis of primary liver cancer in mouse models. *Cells* **12:** 477. doi:10.3390/cells12030477

Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, Goldy J, Garren E, Economo MN, Viswanathan S, et al. 2018. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563:** 72–78. doi:10.1038/s41586-018-0654-5

Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21:** 12. doi:10.1186/s13059-019-1850-9

Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, Chang S, Conley SD, Mori Y, Seita J, et al. 2020. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587:** 619–625. doi:10.1038/s41586-020-2922-4

van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J Mach Learn Res* **9:** 2579–2605.

Wahiduzzaman M, Liu Y, Huang T, Wei W, Li Y. 2022. Cell-cell communication analysis for single-cell RNA sequencing and its applications in carcinogenesis and COVID-19. *Biosafety and Health* **4:** 220–227. doi:10.1016/j.bsheal.2022.03.001

Wang D, Cui P, Zhu W. 2016. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, pp. 1225–1234. Association for Computing Machinery.

Wang J, Sun H, Jiang M, Li J, Zhang P, Chen H, Mei Y, Fei L, Lai S, Han X, et al. 2021. Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Rep* **34:** 108803. doi:10.1016/j.celrep.2021.108803

Wang XY, Almet AA, Nie Q. 2022. Analyzing network diversity of cell–cell interactions in COVID-19 using single-cell transcriptomics. *Front Genet* **13:** 948508. doi:10.3389/fgene.2022.948508

Wang S, Sun S-T, Zhang X-Y, Ding H-R, Yuan Y, He J-J, Wang M-S, Yang B, Li Y-B. 2023. The evolution of single-cell RNA sequencing technology and application: progress and perspectives. *Int J Mol Sci* **24:** 2943. doi:10.3390/ijms24032943

Yu L, Cao Y, Yang JYH, Yang P. 2022. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol* **23:** 49. doi:10.1186/s13059-022-02622-0

Zeng Y, Wei Z, Pan Z, Lu Y, Yang Y. 2022. A robust and scalable graph neural network for accurate single-cell classification. *Brief Bioinformatics* **23:** bbab570. doi:10.1093/bib/bbab570

Zhai Y, Chen L, Deng M. 2023. scGAD: a new task and end-to-end framework for generalized cell type annotation and discovery. *Brief Bioinformatics* **24:** bbad045. doi:10.1093/bib/bbad045

Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, Wiens M, Walters P, Chan T, Hewitson B, et al. 2019. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* **16:** 1007–1015. doi:10.1038/s41592-019-0529-1

Zhang Y, Zhang F, Wang Z, Wu S, Tian W. 2022. scMAGIC: accurately annotating single cells using two rounds of reference-based classification. *Nucleic Acids Res* **50:** e43. doi:10.1093/nar/gkab1275

Zhao M, He W, Tang J, Zou Q, Guo F. 2021. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Brief Bioinformatics* **22:** bbab009. doi:10.1093/bib/bbab009

Zhao M, He W, Tang J, Zou Q, Guo F. 2022. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. *Brief Bioinformatics* **23:** bbab568. doi:10.1093/bib/bbab568

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8:** 14049. doi:10.1038/ncomms14049

Received August 28, 2023; accepted in revised form July 23, 2024.