Supplemental Material

A unified analysis of atlas single cell data

Hao Chen [1,2], Nam D. Nguyen [1], Matthew Ruffalo [1], and Ziv Bar-Joseph [*1,3]

[1]Ray and Stephanie Lane Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2]Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607, USA

[3]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[*]To whom correspondence should be addressed. Email: zivbj@cs.cmu.edu

Supplemental Notes

Supplemental Note 1: Silhouette coefficient for data modalities

We calculated the mean silhouette coefficient over all the nodes in the t-SNE space of each embedding by considering their source of data modalities. Specifically, the silhouette Coefficient was first calculated for each node:

$$\frac{b-a}{\max(a,b)},$$

where a is the mean intra-modality distance for the node and b is the mean nearest-modality distance for the node, defined as the distance between a node and the nearest modality that the node is not a part of. The mean Silhouette Coefficient over all nodes is then calculated, which ranges between -1 and 1. A silhouette coefficient near 0 indicates different data modalities are overlapping globally, while a silhouette coefficient near 1 indicates nodes are clustered based on their modalities. We consider Slide-seq expression and Slide-seq spatial as one modality to avoid considering the trivial case where two representations of the same data overlap. Thus there are three groups, scRNA-seq, scATAC-seq, and Slide-seq, in the calculation of silhouette coefficients.

Supplemental Note 2: Comparisons with cell-based embedding methods

To compare with GIANT, we performed cell embeddings on the HuBMAP dataset using six popular cell-based embedding methods, including Harmony (Korsunsky et al. 2019) (clustering based correction), LIGER (Liu et al. 2020) (non-negative matrix factorization based), Scanorama (Hie et al. 2019) (mutual nearest neighbor based), scVI (Lopez et al. 2018) (deep learning based), Seurat v4 (Y Hao et al. 2021) (mutual nearest neighbors based), and GLUE (ZJ Cao et al. 2022) (deep learning based). We note that there are many other methods developed for multi-modality single cell data integration, however, may not be suitable to be applied to the datasets we used. For example, both multiVI (**ashuach2021multivi**) and MOFA+ (Argelaguet et al. 2020) require input data where multi-omics measurements are derived from the same set of cells. Such data are not available in our dataset.

For Harmony, LIGER, Scanorama, scVI, and Seurat, the gene count matrix for scRNA-seq and gene activity matrix for scATAC-seq were used as input as suggested by Seurat and LIGER (Y Hao

et al. 2021; Liu et al. 2020). For GLUE, the gene count matrix for scRNA-seq and the cell-by-bin count matrix for scATAC-seq were used as input, and the default regulatory graph was constructed to connect ATAC bins to genes if they overlap in either the gene bodies or promoter regions. The reciprocal PCA method was used for Seurat to improve speed. The parameters of different methods were set as default. We computed the same silhouette coefficient for each of the cell-based methods.

To evaluate the ability of the cell-based method in discovering functional gene modules across datasets, for each cell-based method, we clustered cells into cell clusters based on its embeddings using the Leiden algorithm (Traag et al. 2019) with the resolution of 10. Then a co-expression network is built for each cell cluster using the same method as in GIANT. The WGCNA module detection method (Langfelder et al. 2008) is then applied to detect gene modules from each co-expression network by hierarchical clustering of genes and dynamic tree cut. GO enrichment was then performed in each gene module. We aggregated the enriched GO terms for each method.

The GLUE method also generates gene embeddings along with the cell embeddings, however, there is only one embedding for each gene in the whole dataset (vs. one embedding for each gene in each cell cluster for GIANT), which cannot reflect tissue or cell type-specific functions of the gene. We therefore only compared with the cell embeddings of GLUE.

Supplemental Note 3: Comparison with node2vec and Gene2vec gene embeddings

We applied two graph embedding methods, node2vec (Grover et al. 2016) and Gene2vec (Du et al. 2019), to generate embeddings for gene nodes using the same gene graphs provided as input to GIANT. Both methods use the Skip-gram model (Mikolov et al. 2013) for learning node embeddings, though they adopt different strategies. node2vec captures graph structure through biased random walks, while Gene2vec focuses on directly embedding nodes based on their connections in the graphs. For node2vec, we set the return parameter (p) to 1 and the in-out parameter (q) to 0.5 in random walks. We then applied Leiden clustering with a resolution parameter of 1 (same as we used for GIANT) on the resulting embeddings to identify embedding components.

Supplemental Note 4: Mean average precision

We computed the mean average precision (MAP) of cell types in different embedding spaces (ZJ Cao et al. 2022). For each gene (cell), MAP calculated the average precision (AP) within its K ordered nearest neighbors. Supposing that the gene (cell) i comes from cell type/tissue class $y^{(i)}$ and the cell types/tissue classes where its K ordered nearest neighbors come from are $y_1^{(i)}, y_2^{(i)}, ..., y_K^{(i)}$, then AP of gene (cell) i is defined as follows:

$$AP^{(i)} = \begin{cases} \frac{\sum_{k=1}^{K} 1_{y^{(i)}=y_k^{(i)}} \cdot \frac{\sum_{j=1}^{k} 1_{y^{(i)}=y_j^{(i)}}}{k}}{\sum_{k=1}^{K} 1_{y^{(i)}=y_k^{(i)}}}, & \text{if } \sum_{k=1}^{K} 1_{y^{(i)}=y_k^{(i)}} > 0 \\ 0, & \text{otherwise} \end{cases}.$$

where $1_{y^{(i)}=y_k^{(i)}}$ is an indicator function that equals 1 if $y^{(i)} = y_k^{(i)}$ and 0 otherwise. We set K to 0.1% of the total number of genes (cells) in the whole embedding space. The cell type/tissue average MAP score was calculated by averaging the AP of genes firstly within each cell type/tissue class and then across cell types/tissue classes.

Supplemental Note 5: Evaluation on the Allen Brain Cell Atlas

We collected single cell mouse brain data from the Allen Brain Cell Atlas (Yao et al. 2023), which contains 2,341,350 cells from scRNA-seq and 3,938,808 cells from MERFISH spatial transcriptomics. Both modalities were derived from entire adult mouse brains, encompassing the same tissue regions and cell populations. In each modality, cells were grouped into 34 clusters from the original publication, with each cell cluster annotated as a cell type. For this analysis, we focused on the intersecting set of 500 genes shared between the two modalities. We built a gene coexpression graph for each scRNA-seq cell cluster and MERFISH cell cluster using the approach described in Methods. Differentially expressed (DE) genes of cell clusters were computed as described in Methods, and the top 30 DE genes of each cell cluster were used to build a cell cluster dendrogram. We then run GIANT to embed genes from the gene graphs.

To compare GIANT with other methods, we again run the cell embedding methods (Harmony, LIGER, scVI, Scanorama, Seurat, and GLUE) on this dataset. Due to the large size of the dataset, we randomly sampled 5000 cells from each cell type in each modality. For batch-effect correction

methods, two modalities were treated as two batches.

Supplemental Note 6: Local Inverse Simpson's Index for modality mixing in neighborhoods

We assessed modality mixing in the embdding space using the Local Inverse Simpson's Index (LISI), which was introduced in (Korsunsky et al. 2019). LISI measures the effective number of modalities in a neighborhood. Neighborhoods represented by only a single modalities get an LISI of 1, while neighborhoods with an equal number of nodes from two modalities get an LISI of 2. LISI are computed from neighborhood lists per node from k-nearest neighbor (k-NN) graphs in the embedding space. In our experiments, we used the parameters of k=90 nearest neighbors at a fixed perplexity of 30 to compute the weighted k-NN graphs. A LISI score was computed for each gene or cell node and the average LISI score across nodes is reported for each method.

Supplemental Note 7: Enrichment of Reactome pathways in embedding components

Data of the associations of human genes to 1736 Reactome pathways (Gillespie et al. 2022) were collected from the GSEA database (Subramanian et al. 2005). We then performed the enrichment analysis of the 1736 pathways in the embedding components identified from the GIANT's embedding using the methods described in the section "Splitting embedding space into components and identifying enriched GOs and TF regulons in them". To summarize the enrichment analysis results, we associated the embedding components with tissues as described in the same section. Then for each tissue, we identified the top 10 most frequently enriched pathways in the embedding components associated with this tissue. The P-values of these pathways were reported in Figure S5.

Supplemental Note 8: Comparisons with gene clustering on the original data matrices

For each tissue, gene expression profiles of all the cells in the scRNA-seq data and Slide-seq data, as well as gene activity profiles of all the cells in the scATAC-seq data were firstly concatenated into a cell by gene matrix. Following the Scanpy pipeline, each cell in the data matrix was normalized by total counts over all genes, and each value in the data matrix was then logarithmized. To

cluster genes in the data matrix of each tissue, we applied hierarchical clustering on the data matrix and identified 150 gene clusters for each tissue, which is roughly the average number of gene modules associated with each tissue in our GIANT embedding. We finally performed GO enrichment analysis in each gene cluster. For comparative analysis, a heatmap was generated summarizing GO enrichment results for all tissues similar to that in Figure 3A (Fig. S6), following the same procedure except for adopting a P-value threshold of 0.05 to include as many enriched GO terms as possible for the hierarchical gene clustering method.

Supplemental Note 9: Comparison with Limma

We first concatenated the gene expression profiles of all cells from the scRNA-seq and Slide-seq data, along with the gene activity profiles from the scATAC-seq data, into a cell-by-gene matrix. Each cell in the matrix was then normalized by total counts across all genes, and the resulting values were logarithmized. Treating each cell as a sample, Limma (Ritchie et al. 2015) was applied to correct for batch effects in the data matrix. To compare with GIANT, we performed Leiden clustering on the cells and then used the same clustering methods to group genes into gene modules using the corrected data matrix, which resulted in 42 cell clusters and 889 gene modules in total. GO enrichment analysis was then conducted for each gene module, and the modules were associated with specific tissues using the same method described in the section "Genes closely embedded in the space show tissue-specific GO enrichment". A heatmap (Fig. S7) was generated summarizing GO enrichment results (GO terms with P-values < 0.05) for all tissues similar to that in Figure 3A.

Supplemental Note 10: Application on the human fetal atlases

We applied GIANT on another dataset of human fetal atlas that contains scRNA-seq data of 4,062,965 cells from 15 tissues (J Cao et al. 2020), including Adrenal, Cerebellum, Cerebrum, Eye, Heart, Intestine, Kidney, Liver, Lung, Muscle, Pancreas, Placenta, Spleen, Stomach, Thymus, and scATAC-seq data (Domcke et al. 2020) of 720,613 cells from the same 15 tissues.

We used the cell clusters provided by the atlas with cell type annotations. Specifically, cells in the scRNA-seq data were clustered using the Louvain clustering (Blondel et al. 2008) on the UMAP space per tissue. Clusters were annotated based on cell type-specific marker gene expression. Cells in the scATAC-seq data were annotated with cell types by leveraging the annotations on the

scRNA-seq data of the same tissues, on the basis of gene-level accessibility scores computed for the scATAC-seq data. We finally got 172 cell clusters of 15 tissues for the scRNA-seq data and 100 cell clusters of 15 tissues for the scATAC-seq data.

The same algorithms were used to construct a gene co-expression graph for each scRNA-seq cell cluster and a gene-TF hypergraph for each scATAC-seq cell cluster. To get the list of peaks for each scATAC-seq cell cluster, specificity scores for peak and cell type pairs were obtained from the original publication (Domcke et al. 2020). We consider the top 500,000 peaks for each cell type with the highest specificity scores.

To construct the dendrogram, the top 50 differentially expressed or active genes were computed from the gene expression matrix of each scRNA-seq cell cluster and the gene-level accessibility matrix of each scATAC-seq cell cluster, which was computed in the original publication (Domcke et al. 2020).

To make results comparable between the HuBMAP dataset and the human fetal dataset, we consider the same set of genes selected from the HuBMAP dataset for generating the gene embeddings of the human fetal dataset.

After the gene embeddings were obtained. Embedding components were again identified from this human fetal dataset. To assess the consistency between the embedding components obtained from the two datasets, we computed the enrichment of the 1,199 HuBMAP embedding components in the human fetal dataset. Specifically, for each HuBMAP embedding component, we treated the set of genes appearing in the embedding component as a gene module. The enrichment of the gene modules in the set of genes of each embedding component of the human fetal dataset was then computed. Specifically, for each HuBMAP embedding component, we considered the genes within that component as gene set A. For each embedding component (set B) in the human fetal dataset, we performed a hypergeometric test to calculate the overrepresentation of set A within set B. In this analysis, the universe consisted of the intersection of gene scopes between the HuBMAP dataset and the human fetal dataset. The P-value of the test is calculated as the probability of observing x or more common genes between set A and set B by chance:

$$P\text{-value} = \sum_{i=x}^{\min(k,n)} \frac{\binom{k}{i}\binom{M-k}{n-i}}{\binom{M}{n}},$$

where M is the total number of genes in the universe, n is the number of genes in set B, and k is the number of genes in set A.

Cell embeddings for the human fetal dataset were generated using the aforementioned cell based methods, including Harmony, LIGER, Scanorama, scVI, Seurat v4, and GLUE. The gene count matrix for scRNA-seq and gene activity matrix for scATAC-seq were used as input for all the methods as the cell-by-bin count matrices for scATAC-seq data are not available. Cells with at least 200 total gene counts were kept for the cell-based embedding.

We conducted additional analyses by performing different methods exclusively on the scRNA-seq data from the human fetal dataset, and then computed the cell type average MAP scores. To ensure balanced representation across cell types, we sampled 1000 cells per cell type for this analysis.

Supplemental Note 11: Identifying dataset specific gene modules

We treat the gene embedding components identified from each atlas as the gene modules. To measure the similarity among gene modules from different atlases, we calculated the Intersection over Union (IoU) of gene members for each gene module in a given tissue of one atlas with respect to each gene module in the same tissue of the other atlas. Only genes present in the intersection of the gene feature spaces of both atlases were considered. The top 10% of gene modules with the smallest IoU in each tissue for each atlas were selected as atlas-specific gene modules. This analysis identified atlas-specific gene modules for seven tissues shared between the two atlases.

Enrichment of cell differentiation, cell cycle, metabolic process, and biosynthetic process GO terms (hypergeometric test) were calculated for each gene module. The highest enrichment levels among all the gene modules associated with each tissue are presented in Figure S13.

Supplemental Figures

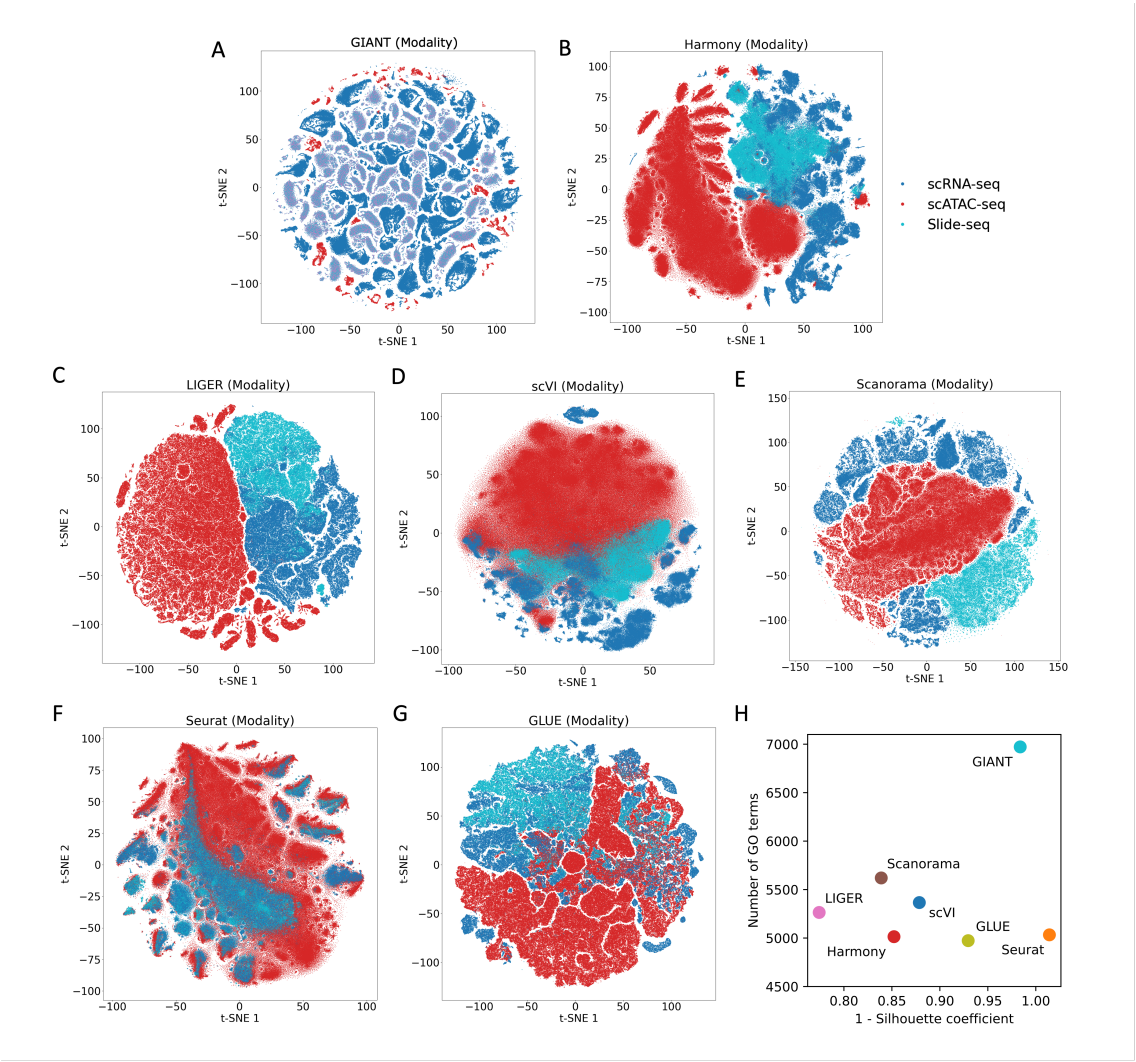


Figure S1: (A) Replication of Figure 2A for comparative purposes. (B-G) Visualization of cell embeddings on the HuBMAP dataset from six cell-based methods. Each point in the visualization represents a cell. Cells are colored by their data modalities. (H) The comparisons between the GIANT gene embeddings and the other cell-based embeddings. More unique GO terms are found enriched in the GIANT embedding components than in the gene modules identified from the competing cell-based embeddings. For GIANT, GO enrichment because of multiple copies of a single gene is ignored. On the other hand, GIANT shows a larger 1 - silhouette coefficient for data modalities than most of the cell-based embeddings, indicating more overlapping between different modalities globally.

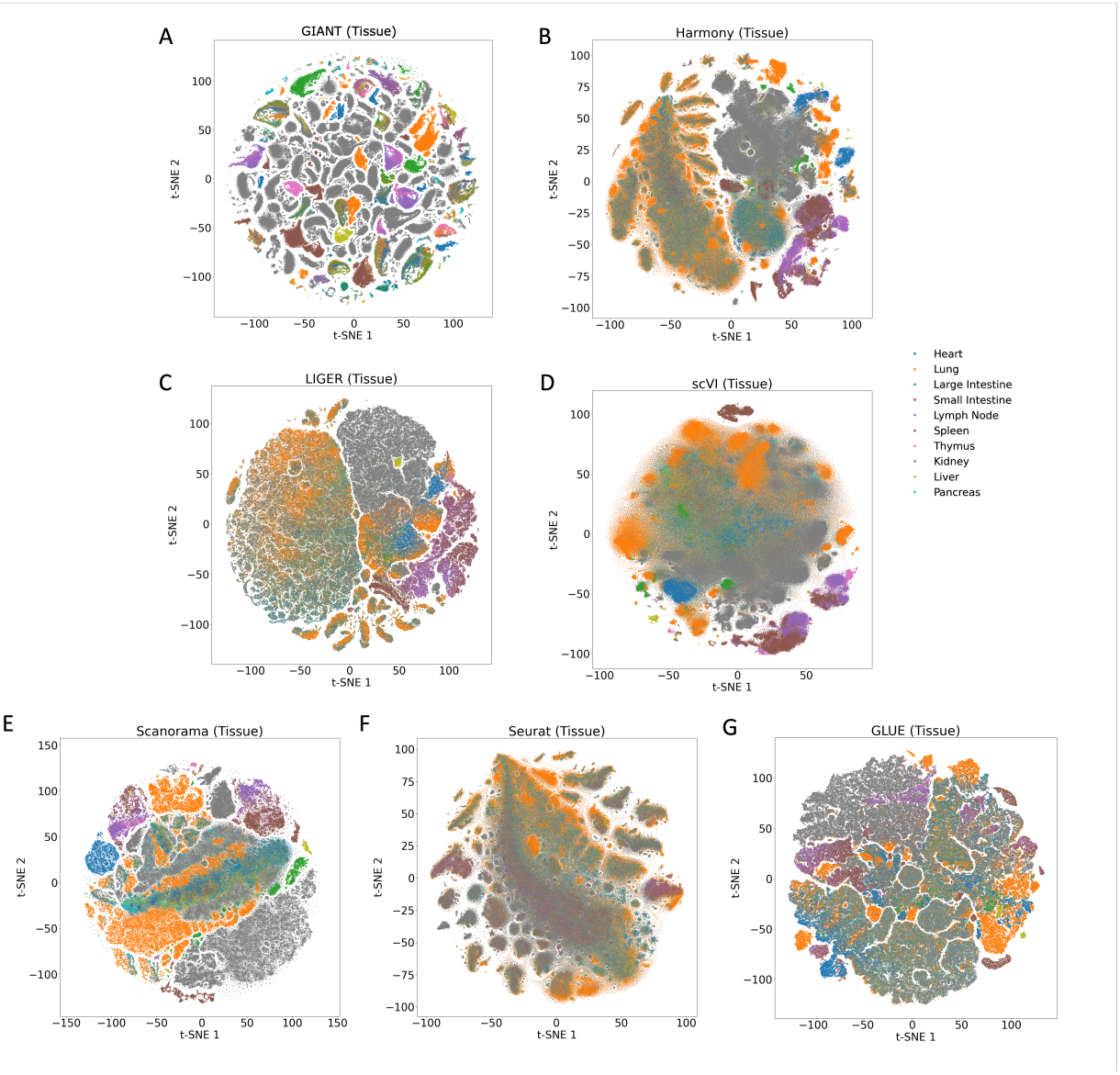


Figure S2: (A) Replication of Figure 2B for comparative purposes. (B-G) The visualization of cell embeddings on the HuBMAP dataset for six cell-based methods (indicated in the titles of subfigures). Each point in the visualization represents a cell. Cells are colored by their tissues.

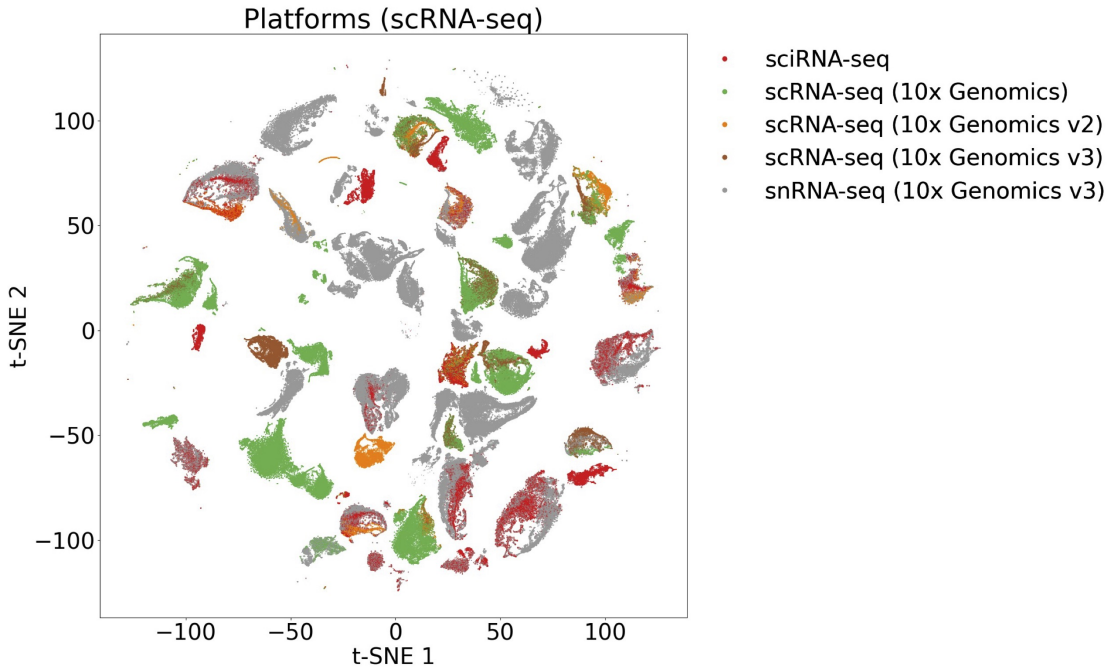


Figure S3: Visualization of GIANT gene embeddings on the HuBMAP dataset. Only the scRNA-seq modality is shown in this plot. Genes are colored by the platforms that the datasets were generated from.

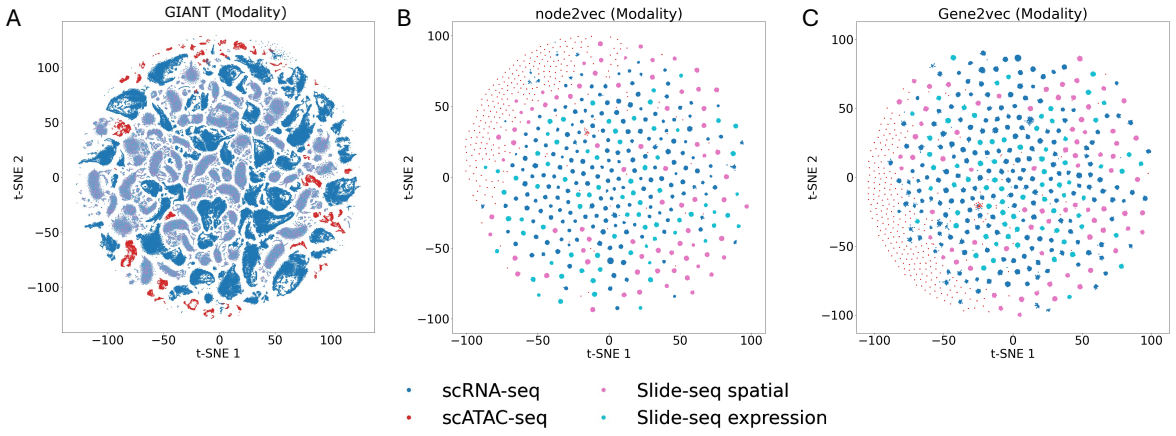


Figure S4: (A) Replication of Figure 2A for comparative purposes. (B) Visualization of node2vec gene embeddings generated from the same gene graphs used as input for GIANT in (A). Gene nodes from different graphs are grouped into distinct clusters. (C) Gene2vec gene embeddings on the same gene graphs, with a similar pattern of gene nodes clustering by graphs. Each dot represents a gene, colored by data modalities.

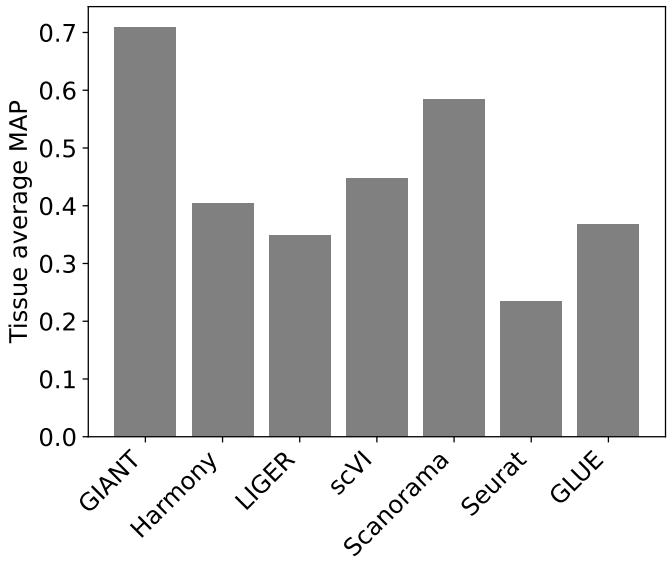


Figure S5: Tissue average MAP scores derived from embeddings of different methods on the HuBMAP dataset.

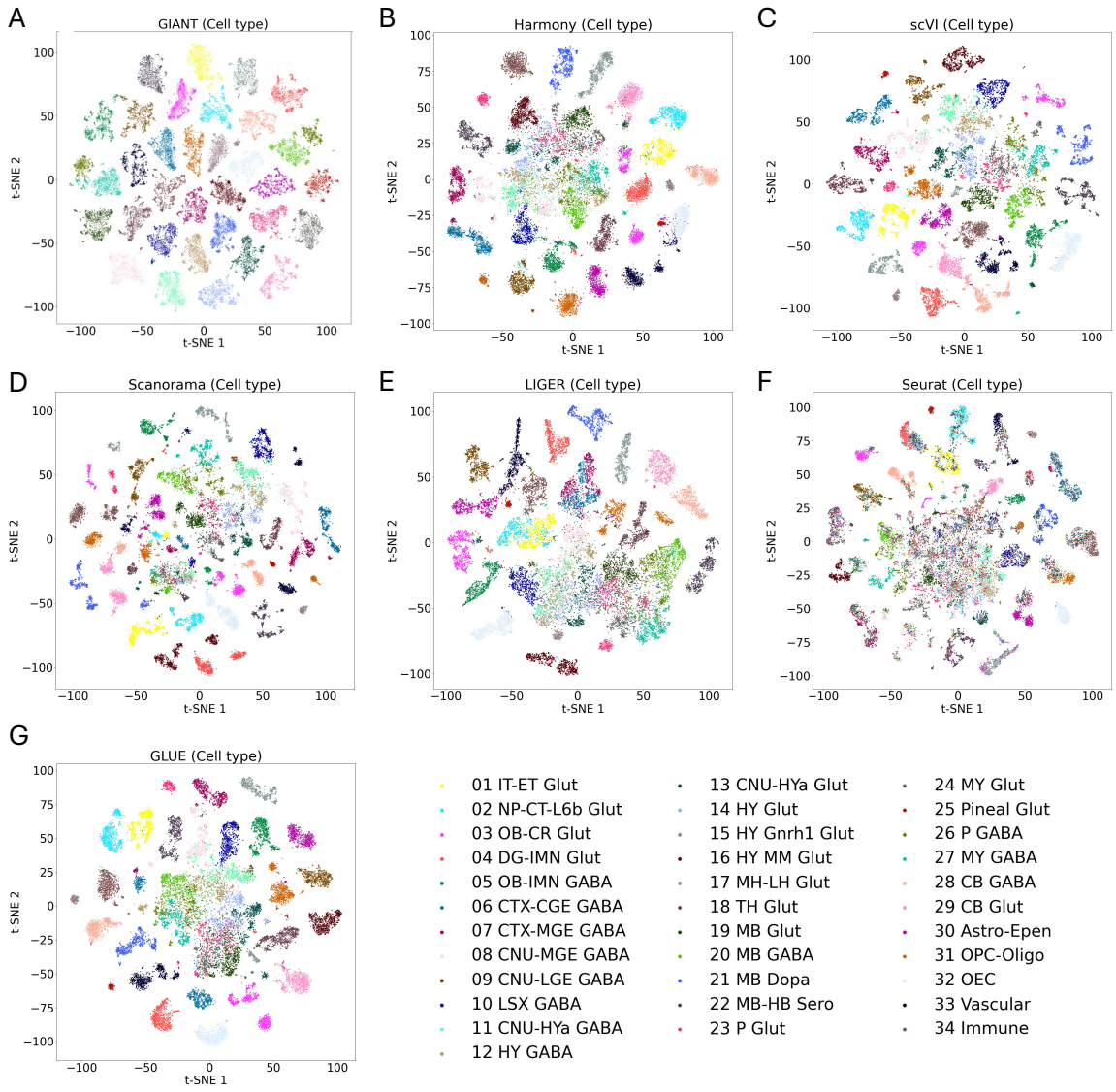


Figure S6: (A) Visualizations of GIANT gene embeddings on the Allen Brain Cell Atlas dataset. Each dot represents a gene, with colors indicating cell types the genes are from. (B-G) show visualizations of cell embeddings generated by six cell-based methods (as specified in the subfigure titles) on the same dataset. Each dot represents a cell, colored by its corresponding cell type. The figure legend provides the cell type taxonomy used in the original dataset publication (Yao et al. 2023).

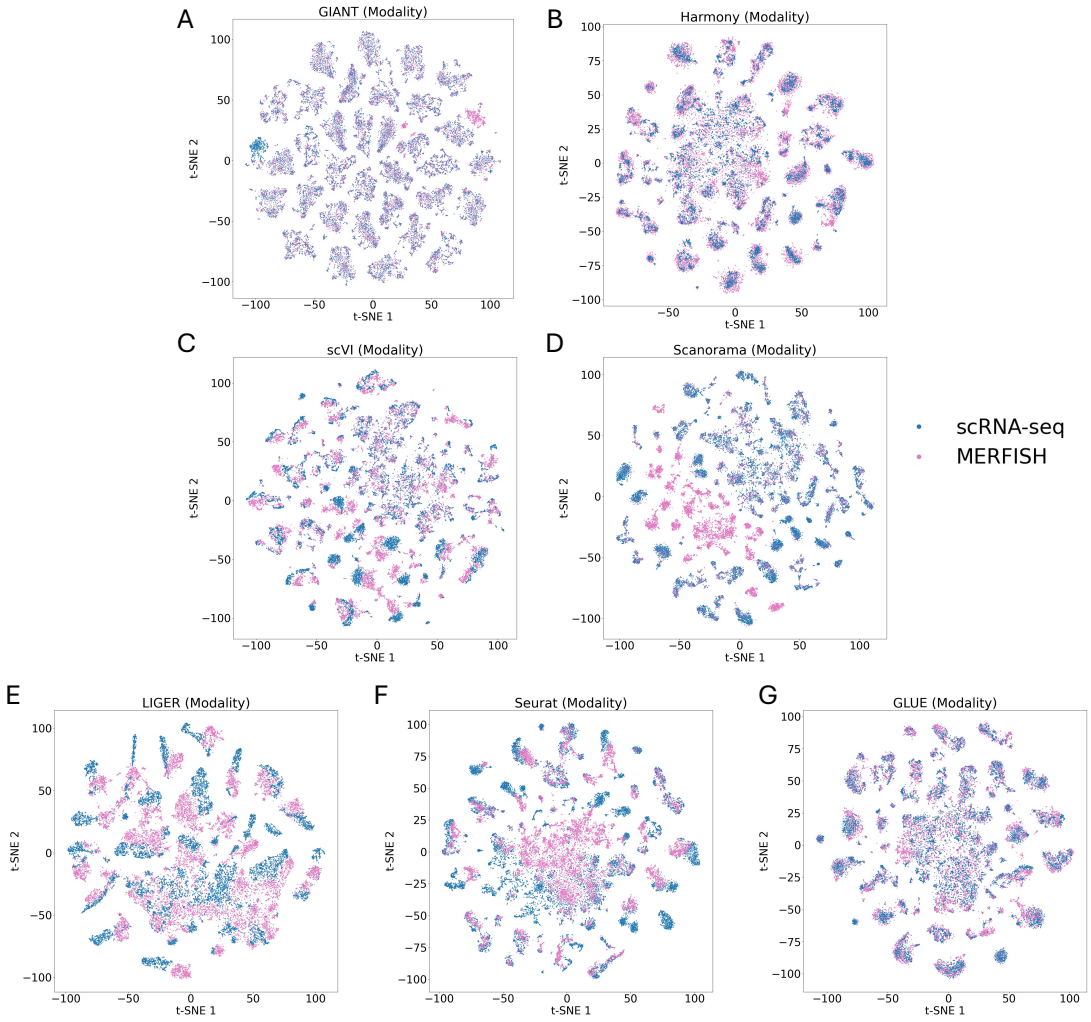


Figure S7: (A) Visualizations of GIANT gene embeddings on the Allen Brain Cell Atlas dataset. Each dot represents a gene, with colors indicating data modalities. (B-G) show visualizations of cell embeddings generated by six cell-based methods (as specified in the subfigure titles) on the same dataset. Each dot represents a cell, colored by data modalities.

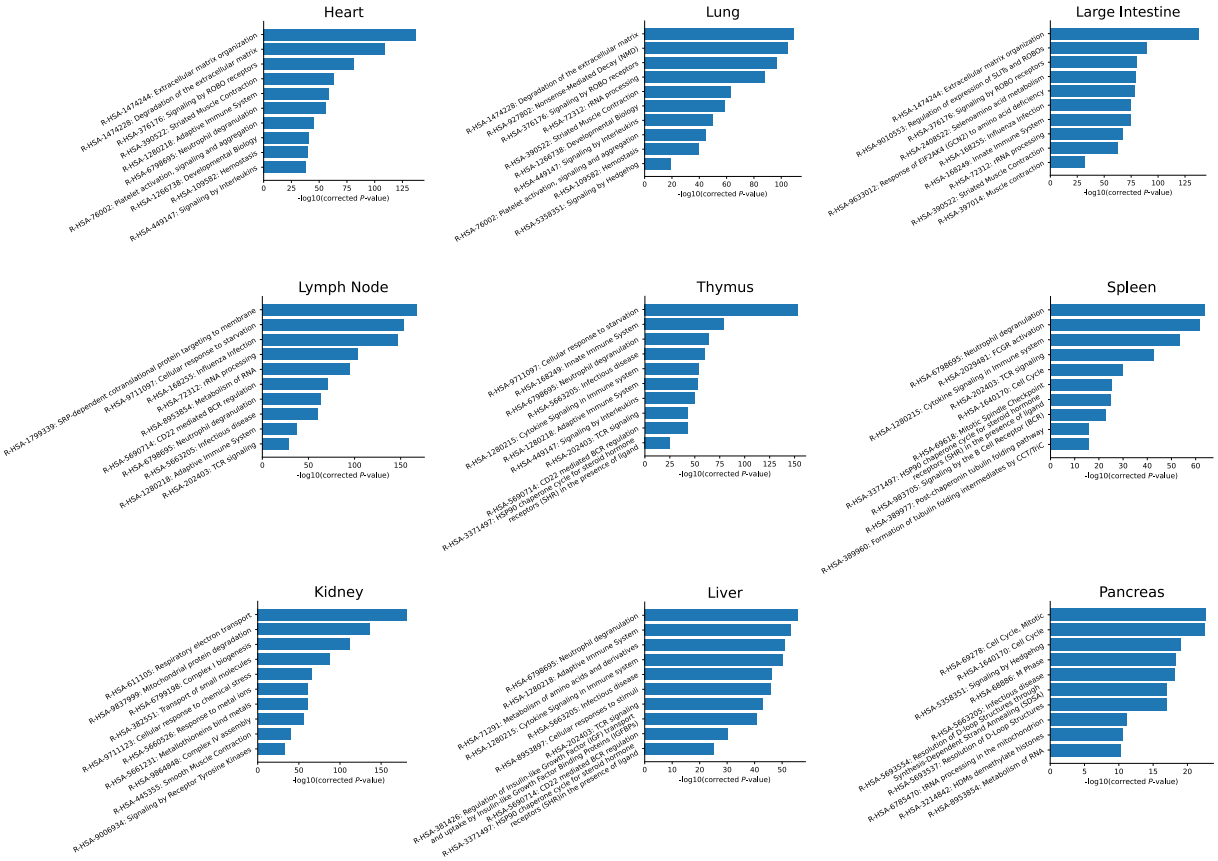


Figure S8: The ten most frequently enriched Reactome pathways in embedding components of each tissue. The bar plots display the lowest P-value for each pathway among the embedding components associated with each tissue.

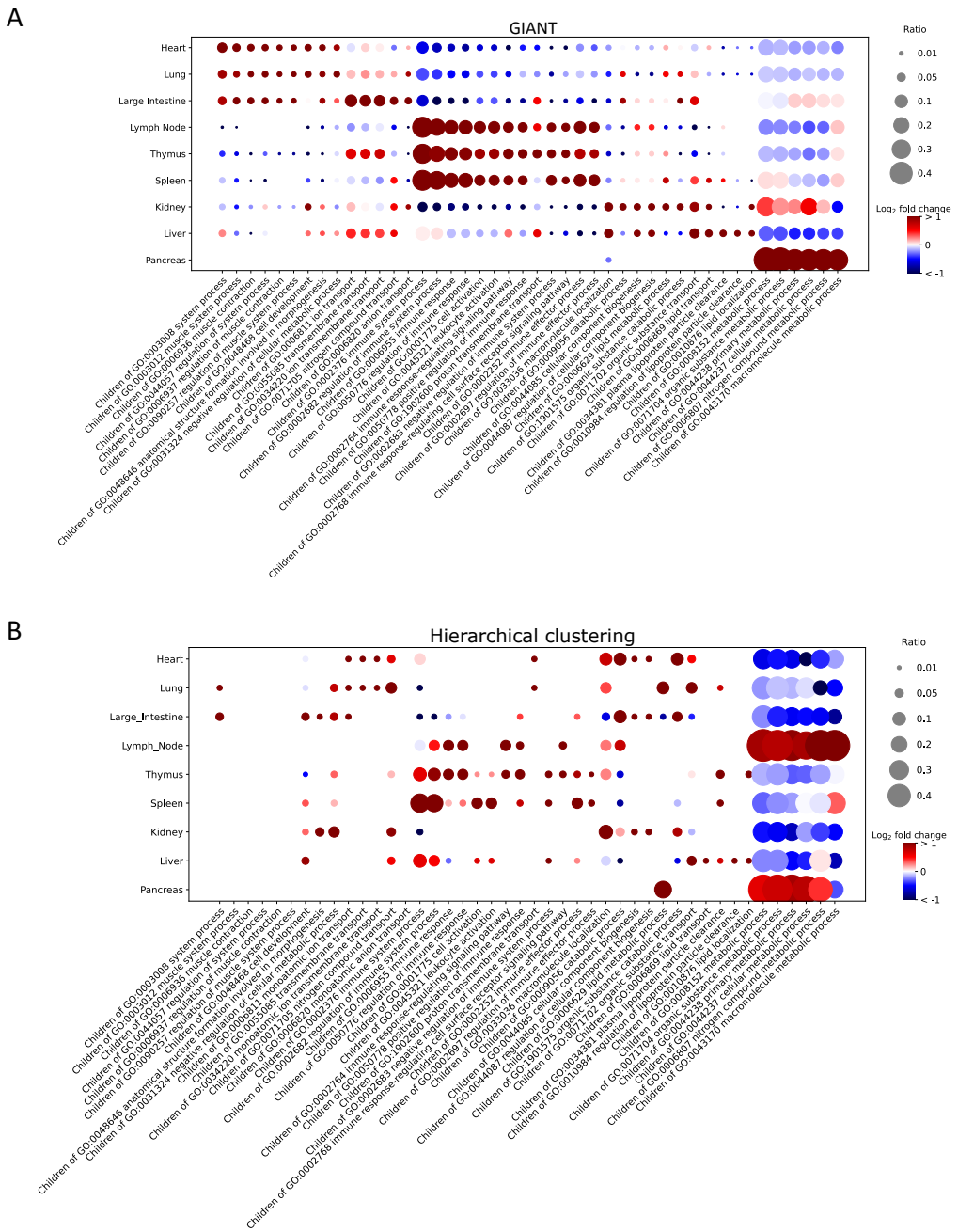


Figure S9: (A) Replication of Figure 3A for comparative purposes. (B) Categories of GO terms enriched in the gene clusters associated with different tissues, which are identified from hierarchical clustering on the original data matrices. Dot size reflects the ratio of child terms within the GO category among all enriched terms for the tissue. Dot color indicates the fold change of the ratio compared to the average value of ratios in the column. Fewer tissue-specific GO categories showed enrichment in (B). For example, GO terms "muscle system process" and "regulation of muscle contraction" as well as their children terms are absent for the gene clusters associated with Heart. GO term "leukocyte (white blood cell) activation" and its descendants are missing for gene clusters associated with Lymph Node.

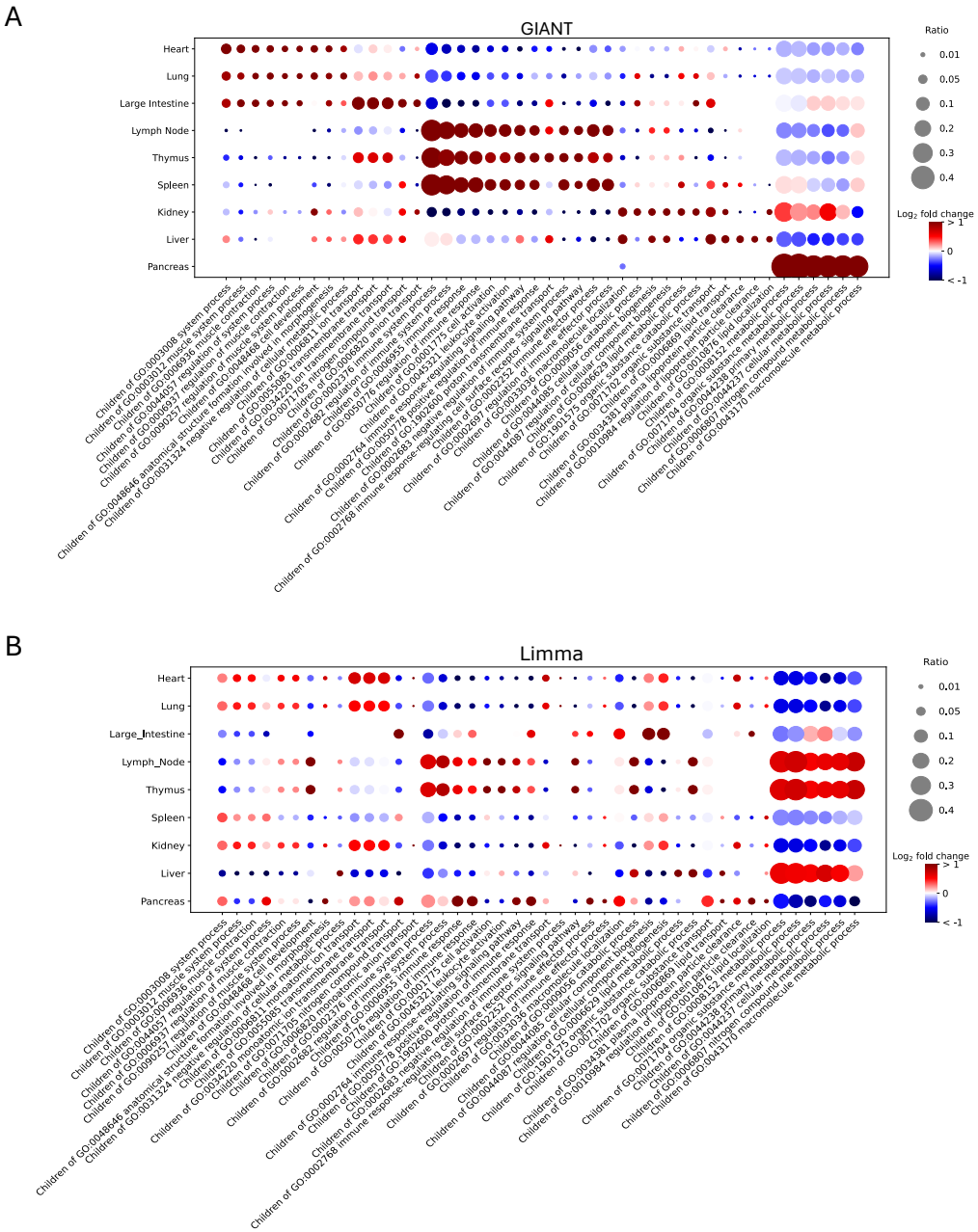


Figure S10: Comparison with Limma. (A) Replication of Figure 3A for comparison. (B) Categories of GO terms enriched in the gene modules associated with different tissues, which are identified from the Limma-corrected data matrix. Dot size reflects the ratio of child terms within the GO category among all enriched terms for the tissue. Dot color indicates the fold change of the ratio compared to the average value of ratios in the column. Some prominent GO categories in (A), reflecting known tissue functions, are less evident in (B). For example, "muscle system process" and "regulation of muscle contraction" are not prominent in Large Intestine gene modules, and the ratios for "immune system process" and "regulation of immune response" are lower in Spleen. Additionally, certain GO categories not typically associated with tissue-specific functions appear in the corresponding gene modules, such as "Immune response" in Pancreas, which is a non-immune tissue, and "muscle system process" in Kidney and Spleen, which have few muscle cells. These discrepancies may be due to inaccuracies in cell clustering from the Limma-corrected data matrix.

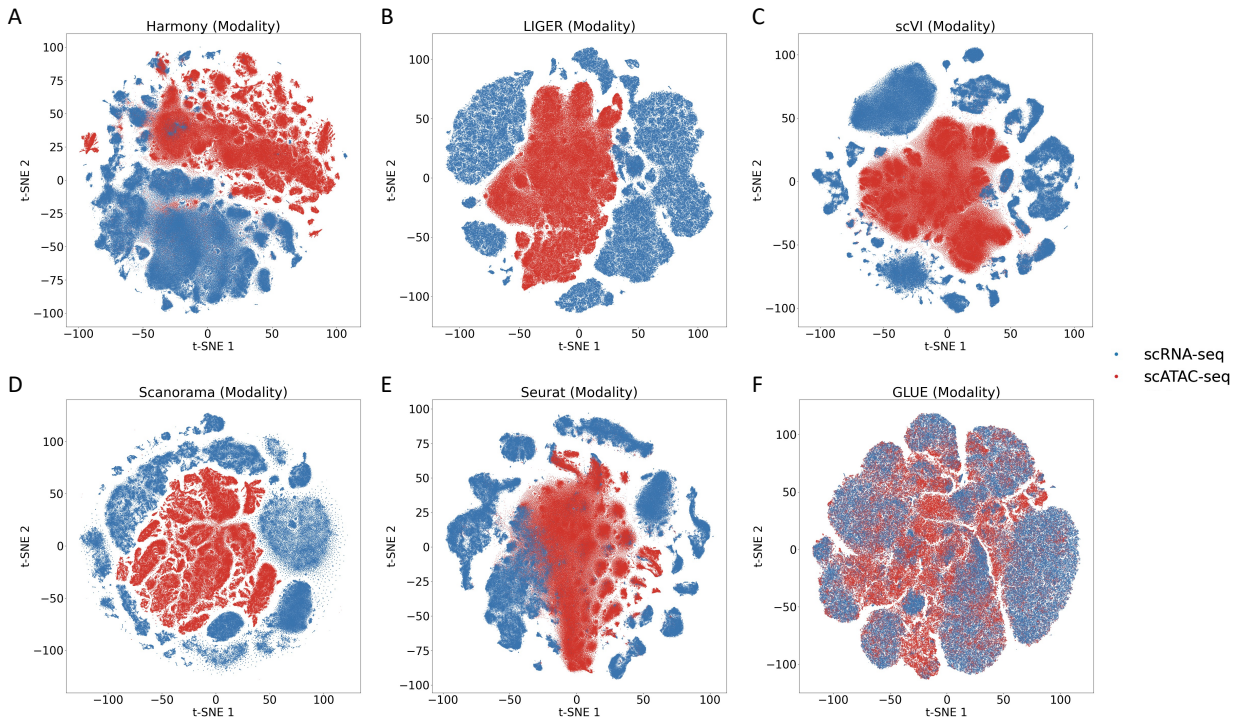


Figure S11: (A-F) Visualization of cell embeddings on the human fetal dataset from six cell-based methods (indicated in the titles of subfigures). Each point in the visualization represents a cell. Cells are colored by their data modalities.

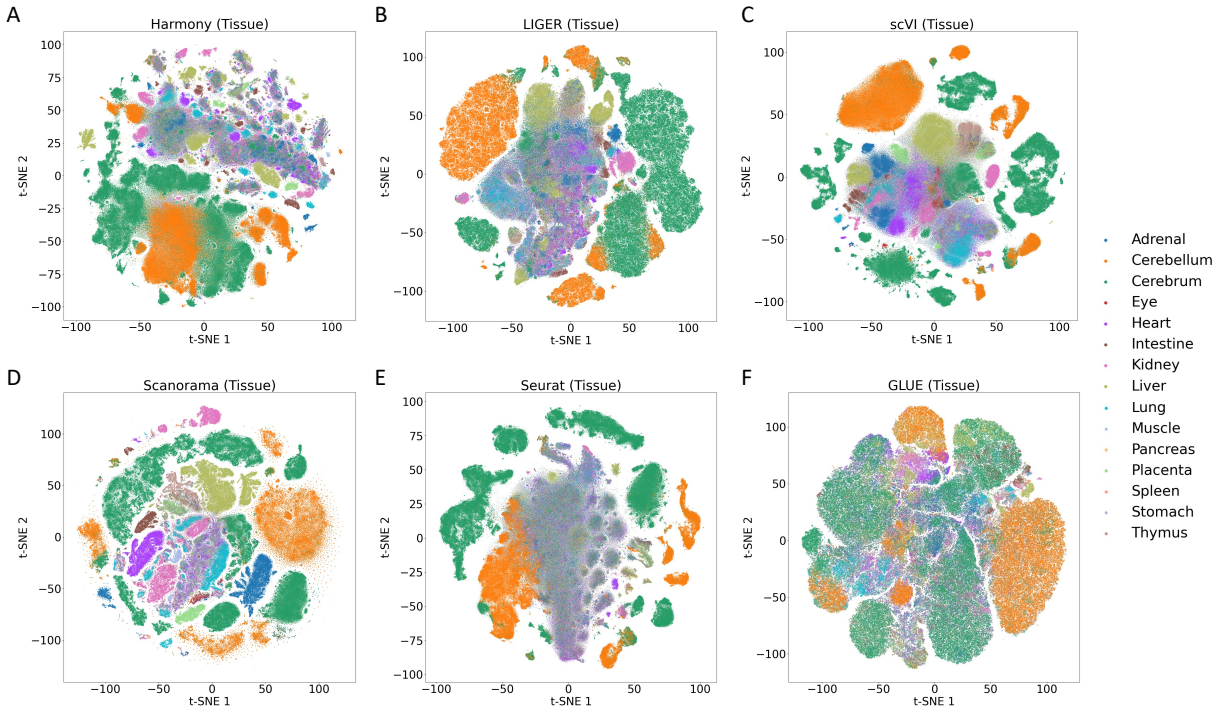


Figure S12: (A-F) Visualization of cell embeddings on the human fetal dataset from six cell-based methods (indicated in the titles of subfigures). Each point in the visualization represents a cell. Cells are colored by their tissues.

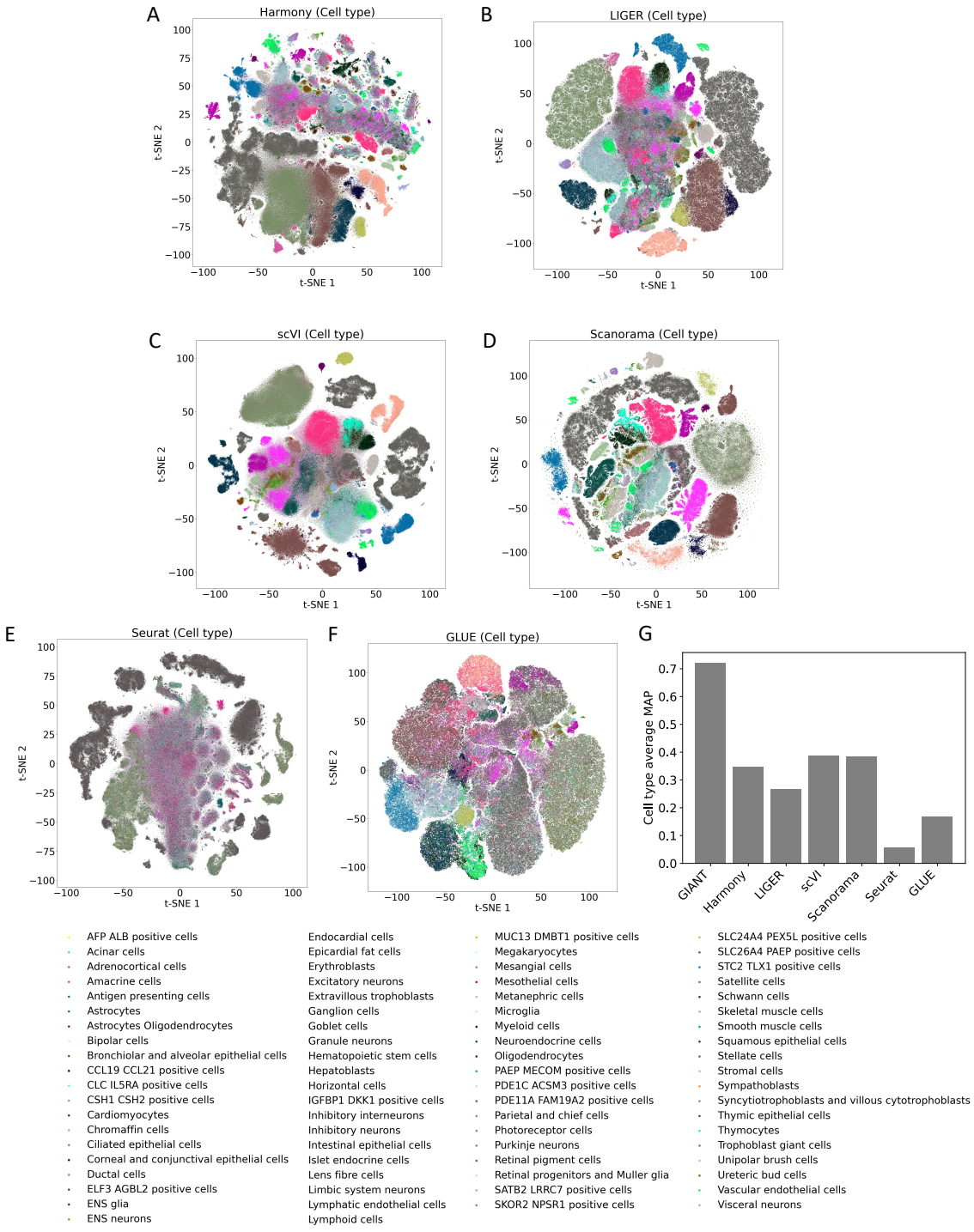


Figure S13: (A-F) Visualization of cell embeddings on the human fetal dataset from six cell-based methods (indicated in the titles of subfigures). Each point in the visualization represents a cell. Cells are colored by their tissues. (G) Cell type average MAP scores derived from embeddings of different methods.

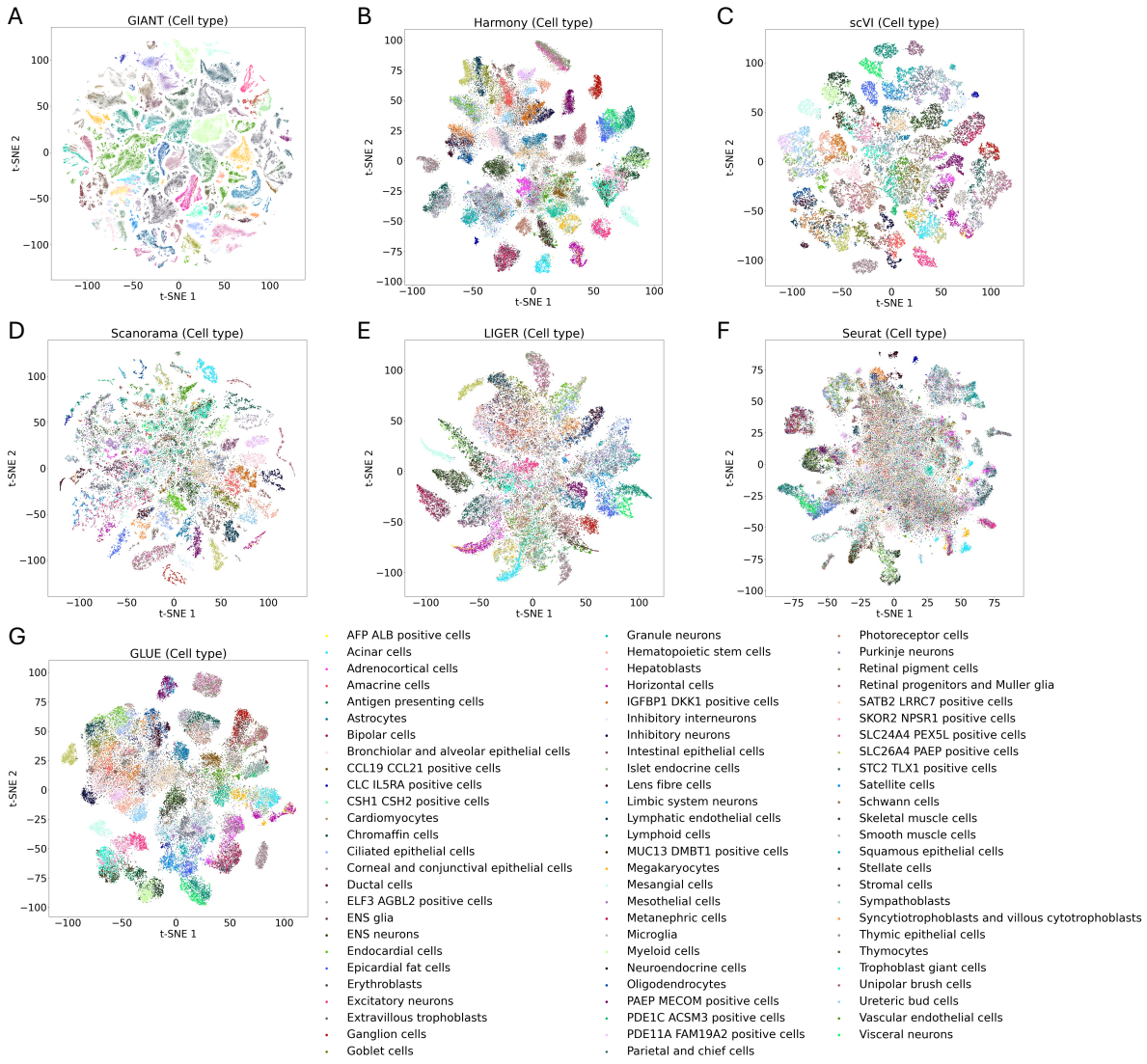


Figure S14: (A) Visualizations of GIANT gene embeddings on the human fetal dataset with only data from the scRNA-seq modality. Each dot represents a gene, with colors indicating cell types. (B-G) show visualizations of cell embeddings generated by six cell-based methods (as specified in the subfigure titles) on the same dataset. Each dot represents a cell, colored by data modalities.

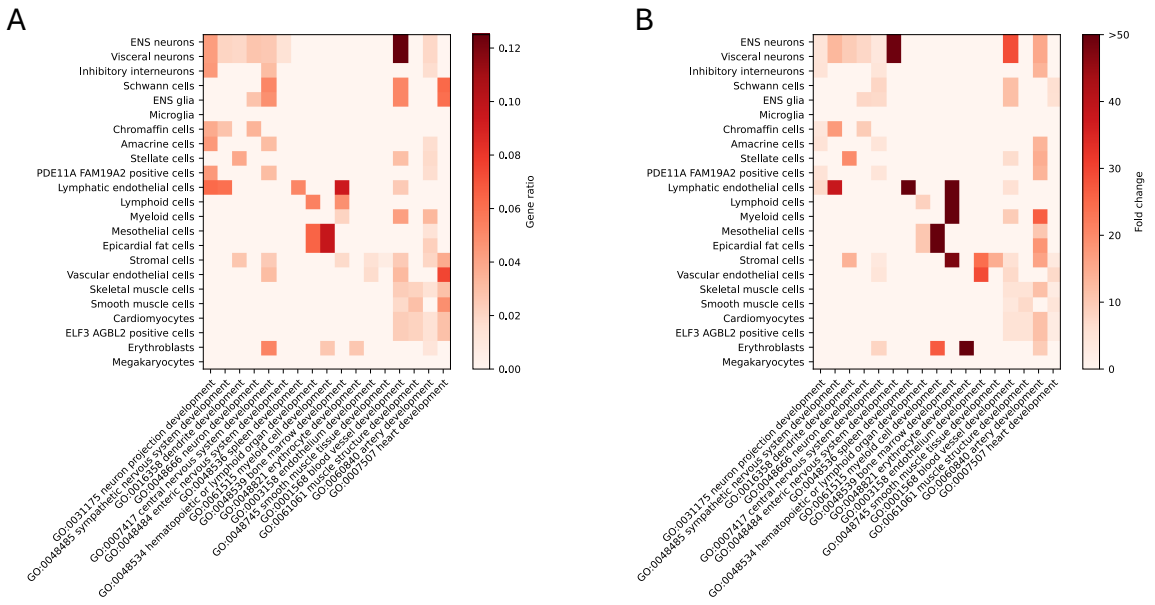


Figure S15: Supplementary data for Figure 4E. (A) Gene ratios for each significant GO term across cell types. The gene ratio represents the proportion of genes associated with a specific GO term relative to the total number of genes in the embedding component. (B) Fold change of each significant GO term across cell types, calculated as the ratio of the gene ratio to the background ratio (the proportion of genes associated with the GO term in the overall gene set).

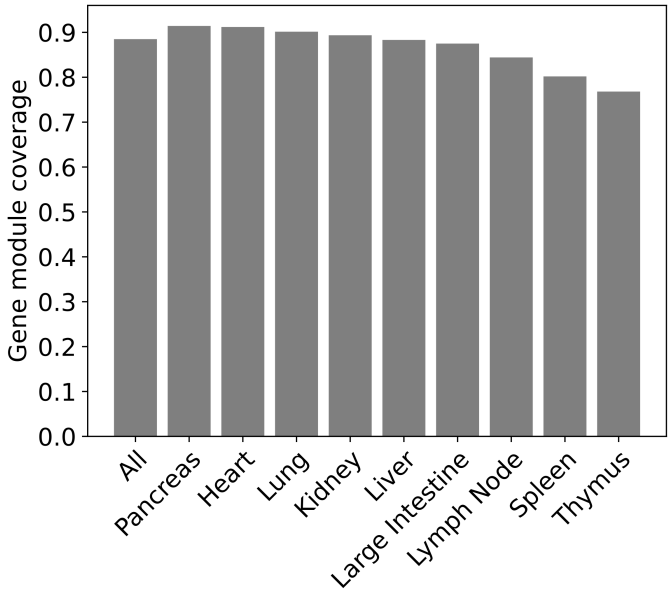


Figure S16: The percentage of gene modules associated with each HuBMAP human tissue that have enrichment in the embedding components of the human fetal atlas dataset.

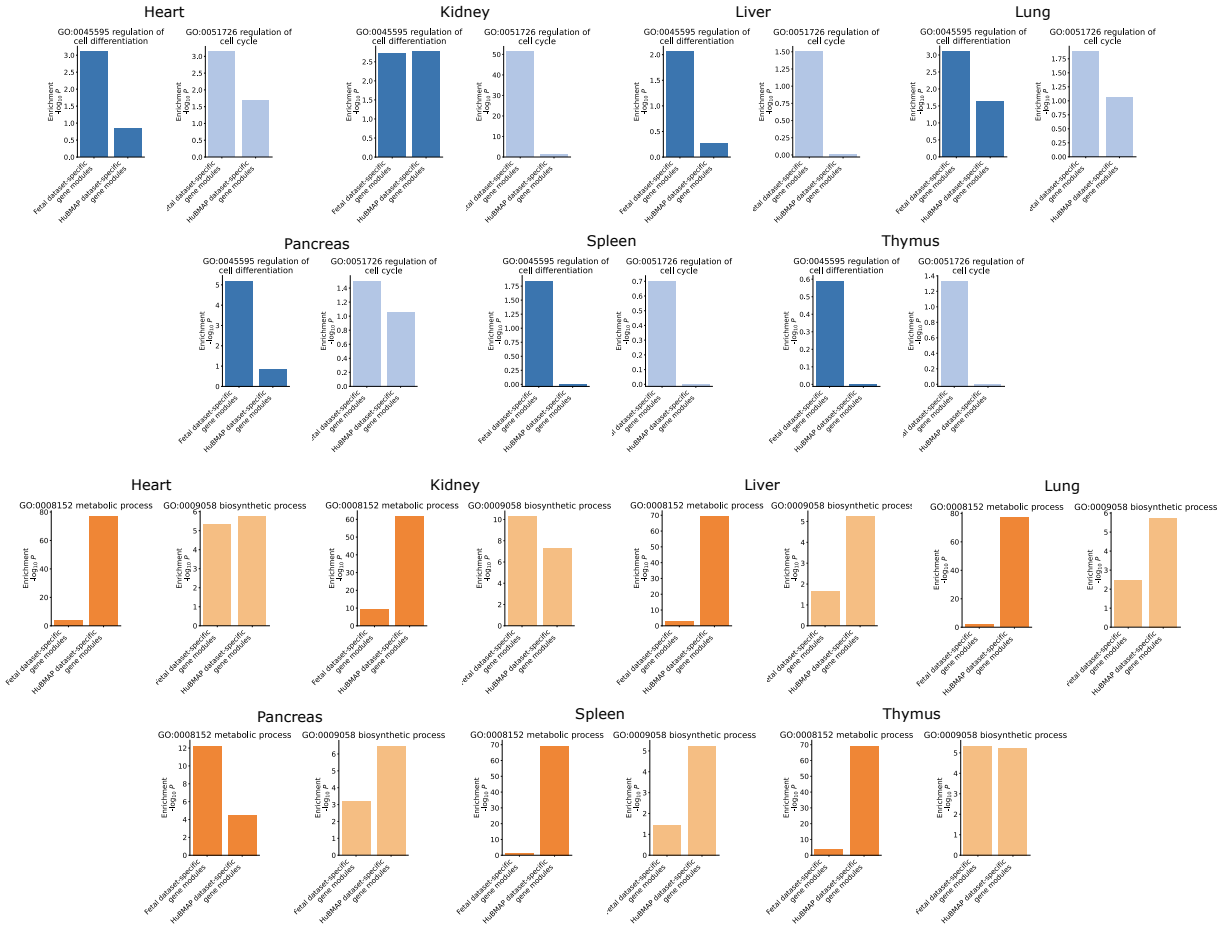


Figure S17: (A) Enrichment of cell differentiation and cell cycle GO terms for the top 10% unique gene modules of each tissue identified from each dataset. Differentiation and cell cycle GO terms are more often enriched in gene modules specific to the human fetal dataset across tissues. (B) Enrichment of metabolic and biosynthetic processes GO terms for the dataset-specific gene modules. Metabolic and biosynthetic processes GO terms often have higher enrichment levels in HuBMAP dataset-specific gene modules across tissues.

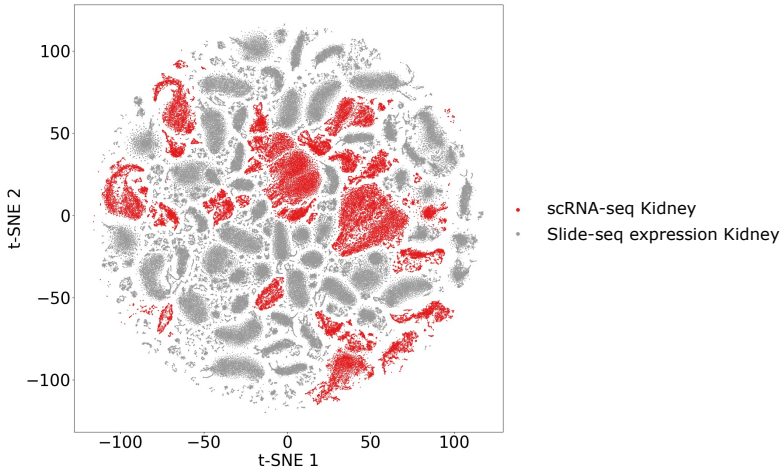
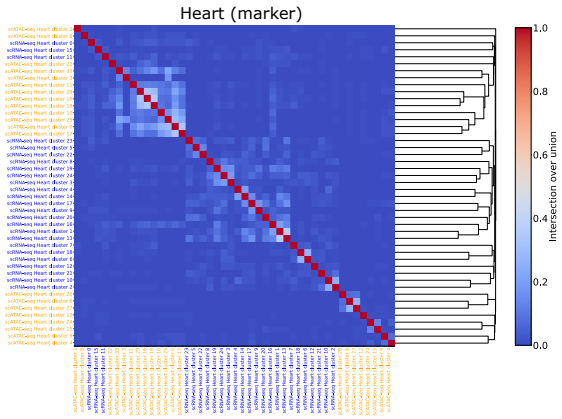
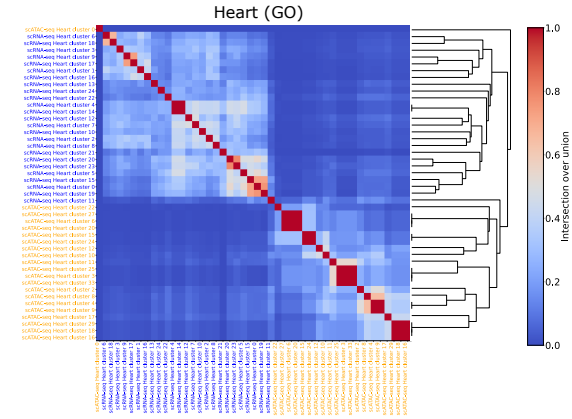
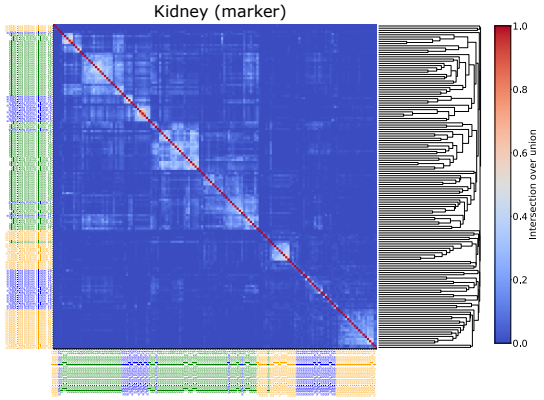
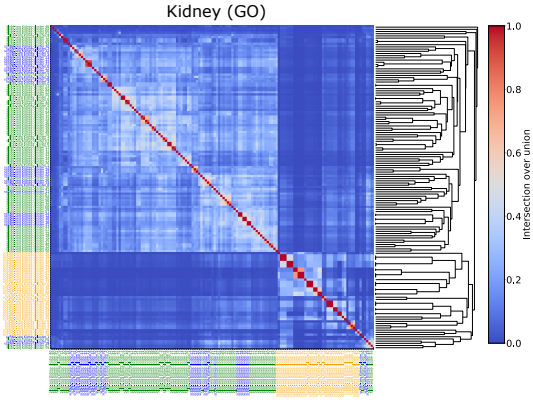


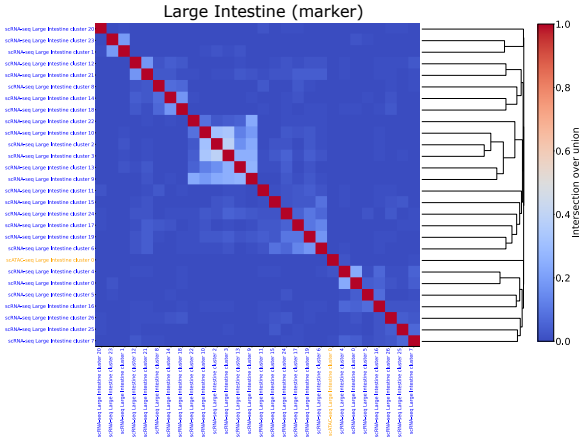
Figure S18: t-SNE plot for gene embeddings from Kidney tissue in scRNA-seq and Slide-seq expression data modalities.

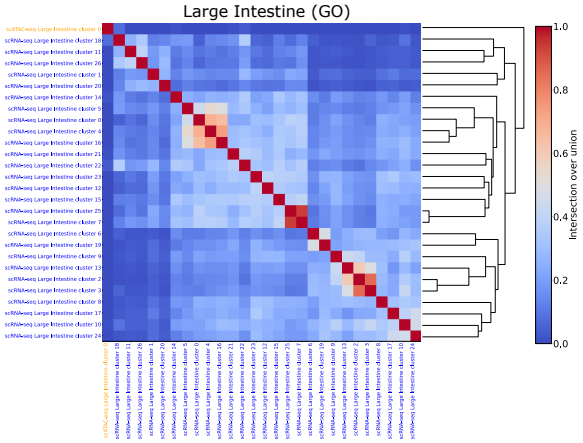

Heart (marker)
Intersection over union
1.0
0.8
0.6
0.4
0.2
0.0


Heart (GO)
Intersection over union
1.0
0.8
0.6
0.4
0.2
0.0


Kidney (marker)
Intersection over union
1.0
0.8
0.6
0.4
0.2
0.0


Kidney (GO)
Intersection over union
1.0
0.8
0.6
0.4
0.2
0.0


Large Intestine (marker)
Intersection over union
1.0
0.8
0.6
0.4
0.2
0.0


Large Intestine (GO)
Intersection over union
1.0
0.8
0.6
0.4
0.2
0.0

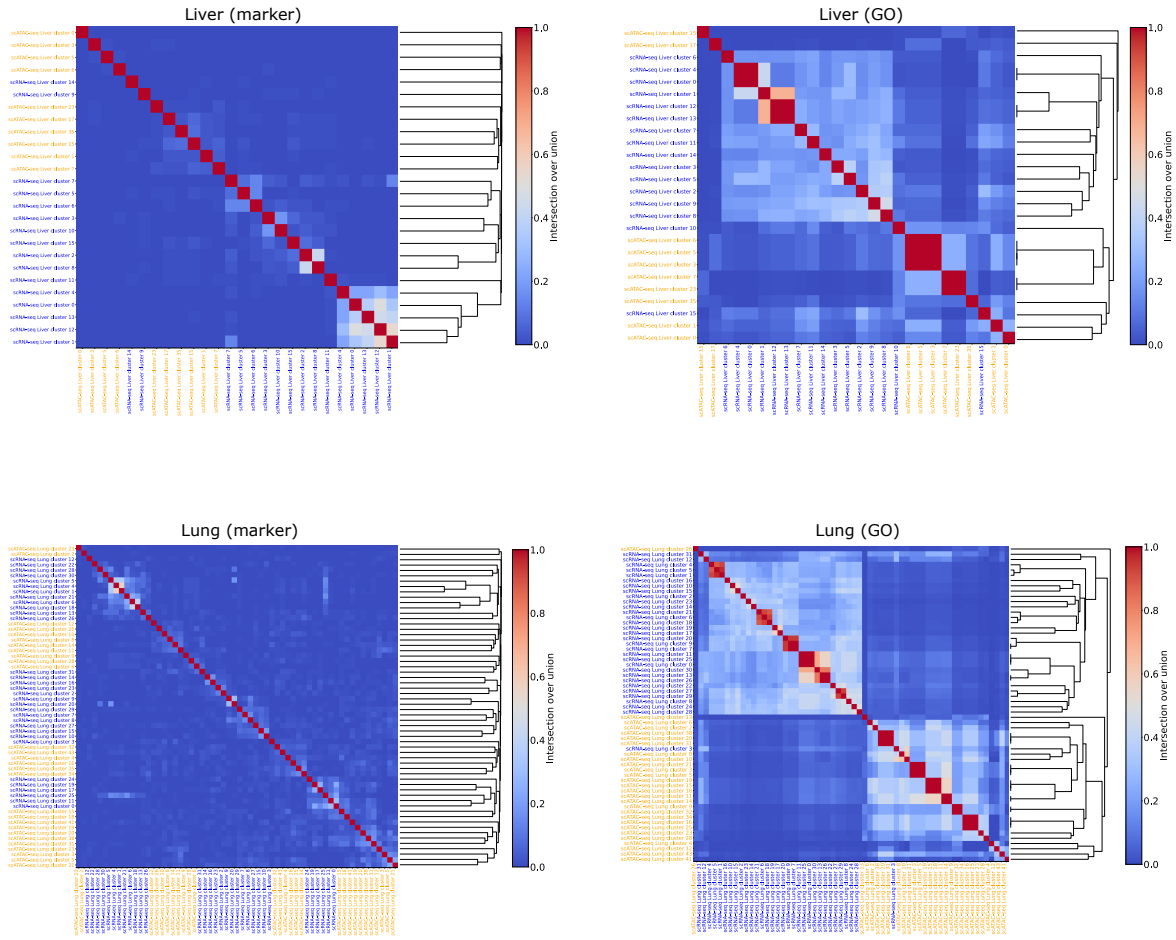


Figure S19: Marker gene overlap and enriched GO term overlap across cell clusters within each tissue. The heatmaps in the left column show the intersection over union (IoU) of marker genes (determined by differential expression as described in Methods) between cell clusters in each tissue, with cell clusters arranged by hierarchical clustering based on their IoU values. Cell clusters from scRNA-seq, scATAC-seq, and Slide-seq are labeled in blue, yellow, and green on the axes. The heatmaps in the right column show the IoU of enriched GO terms in the embedding components of different cell clusters (similar to Fig. 3A, embedding components were associated with the top three cell clusters for genes assigned to the respective components), with cell clusters arranged by hierarchical clustering based on IoU values. Cell clusters from scRNA-seq, scATAC-seq, and Slide-seq are labeled in blue, yellow, and green on the axes. Clustering patterns were found for cell clusters from the same modality. Heatmaps are shown for tissues where at least two modalities show GO enrichment in their embedding components.

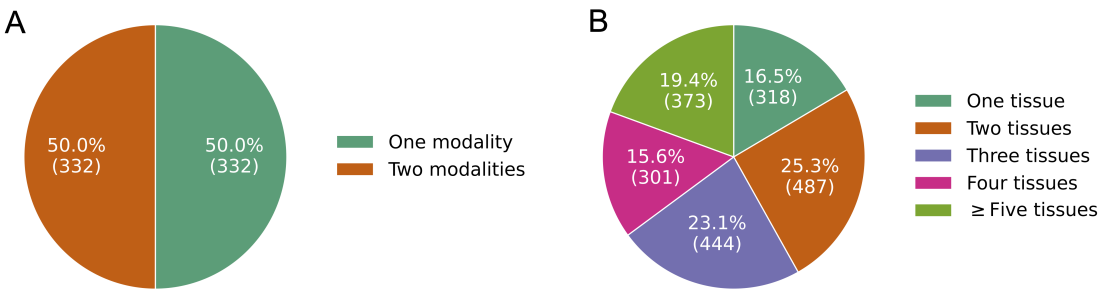


Figure S20: (A) The number of genes that have predicted functions arising out of neighbors from one modality or at least two modalities. 664 genes that appear in at least two of the three data modalities are considered. (B) The number of genes that have predicted functions arising out of neighbors from different numbers of tissues. 1,924 genes that appear in at least two tissues are considered.

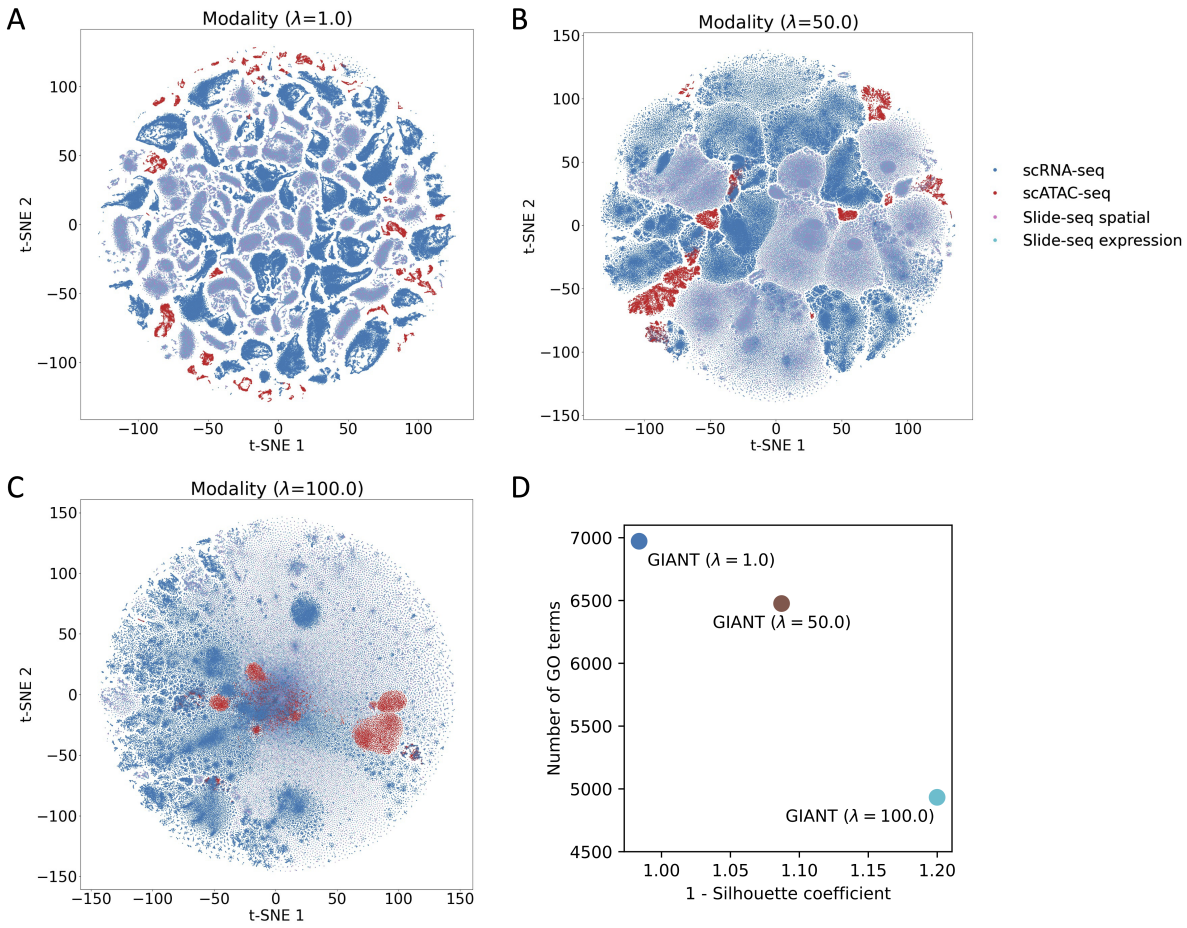


Figure S21: (A-C) Visualization of GIANT gene embeddings on the HuBMAP dataset for different regulation strength parameters (λ). Genes are colored by their data modalities. (D) Larger λ can increase the level of modality merging, indicated by the larger 1-Silhouette coefficients, however, the GO terms that are found enriched in the embedding components are decreased, which indicates lower qualities of these embedding components.

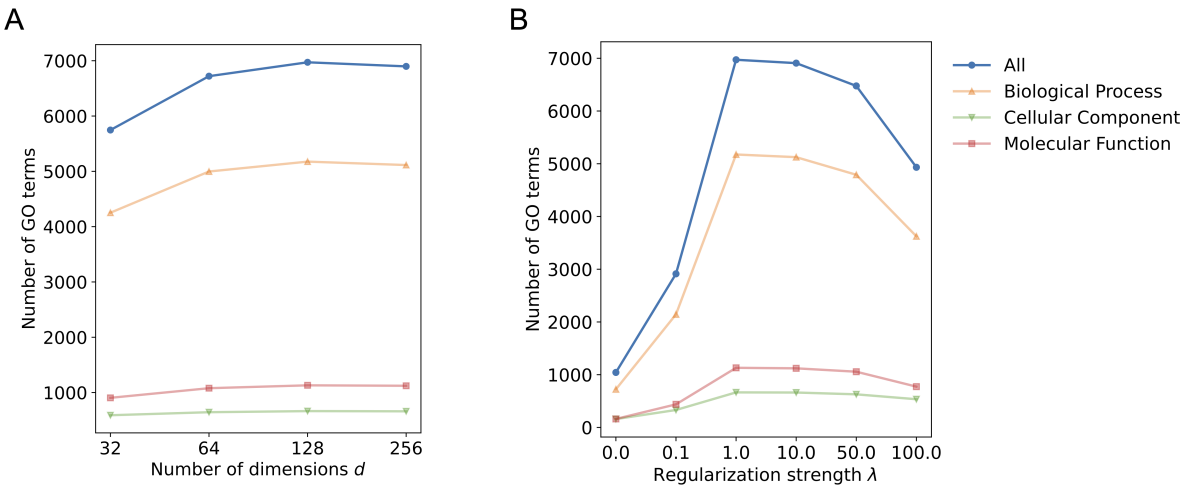


Figure S22: Effect on the performance of GIANT of the choices of hyperparameters for the embedding dimensions (d) and regulation strength for balancing two learning objectives (λ). We measure the performance using the number of different types of GO terms that can be discovered from different embedding components (Methods). (A) While the performance improves when the number of dimensions is up to 128, the performance is overall stable across a reasonable range of values. (B) The performance drops when the model focuses on only the single graph learning objective ($\lambda = 0.0$) or applies too much regularization, though when λ is around 1.0 to 10.0, the model achieves the best performance.

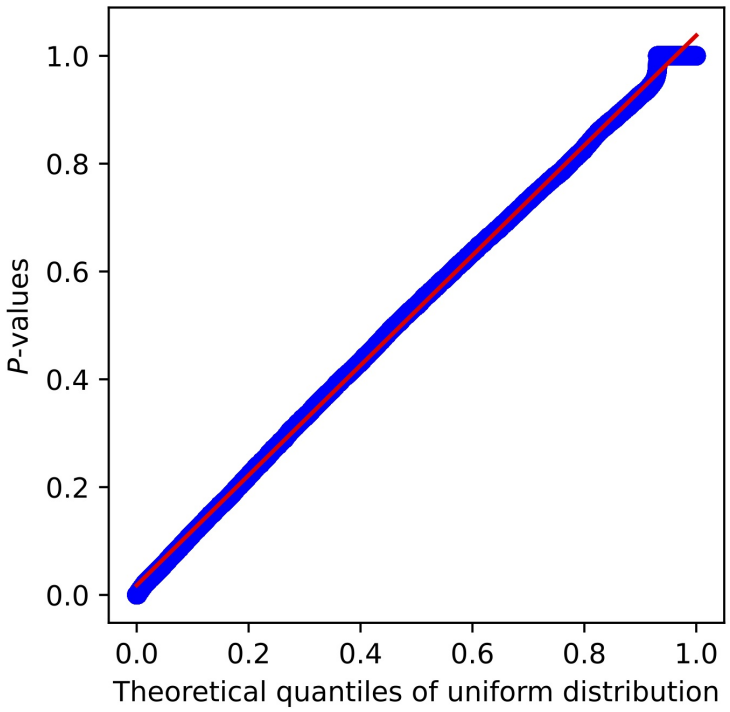


Figure S23: The quantiles of the distribution of P-values against those of uniform distributions. The plot indicates the distribution of P-values is uniform.

References

Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, and Stegle O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* 21: pp. 1–17.

Blondel VD, Guillaume JL, Lambiotte R, and Lefebvre E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical mMchanics: Theory and Experiment* 2008: P10008.

Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F, et al. 2020. A human cell atlas of fetal gene expression. *Science* 370: eaba7721.

Cao ZJ and Gao G. 2022. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology* 40: pp. 1458–1466.

Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, et al. 2020. A human cell atlas of fetal chromatin accessibility. *Science* 370: eaba7612.

Du J, Jia P, Dai Y, Tao C, Zhao Z, and Zhi D. 2019. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics* 20: pp. 7–15.

Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C, Matthews L, Gong C, et al. 2022. The reactome pathway knowledgebase 2022. *Nucleic Acids Research* 50: pp. D687–D692.

Grover A and Leskovec J (2016). "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* 184: pp. 3573–3587.

Hie B, Bryson B, and Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology* 37: pp. 685–691.

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh Pr, and Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* 16: pp. 1289–1296.

Langfelder P and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: pp. 1–13.

Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, and Welch JD. 2020. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature Protocols* 15: pp. 3632–3662.

Lopez R, Regier J, Cole MB, Jordan MI, and Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nature Methods* 15: pp. 1053–1058.

Mikolov T, Chen K, Corrado G, and Dean J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* 3781.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43: e47–e47.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102: pp. 15545–15550.

Traag VA, Waltman L, and Van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* 9: pp. 1–12.

Yao Z, Velthoven CT van, Kunst M, Zhang M, McMillen D, Lee C, Jung W, Goldy J, Abdelhak A, Aitken M, et al. 2023. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* 624: pp. 317–332.