Large-scale proteomic inference at single-cell resolution by scInfer

Tianyi Zhao[1, 2, #], Yuzhi Sun[3, #], Renjie Liu[2, 3, #], Liyuan Zhang[3], Chengcheng Zhang[3], Yuran Jia[3], Liang Cheng[4, *], Guohua Wang[3, *], Yadong Wang[1, *]

# These authors contributed equally: Tianyi Zhao, Yuzhi Sun, Renjie Liu.
* Corresponding authors: Liang Cheng (liangcheng@hrbmu.edu.cn), Guohua Wang (ghwang@hit.edu.cn), Yadong Wang (ydwang@hit.edu.cn).
1 School of Medicine and Health, Harbin Institute of Technology, Harbin, China.
2 Harbin Institute of Technology Zhengzhou Research Institute, Zhengzhou, China.
3 Faculty of Computing, Harbin Institute of Technology, Harbin, China.
4 Harbin Medical University, Harbin, China.

Supplementary Notes

Note1: Detailed analysis of beta cell subtype classification

We performed enrichment analysis for each P-GP and R-GP to determine the associated biological functions of the gene programs. Then visual the average expression of these gene programs as activity distributions (Supplementary Fig. 5). In the ProHi-Beta subpopulation, the high activity of P-GP13 represents enhanced ribosomal translation, activation of secretory-related structures, and increased mitochondrial energy metabolism. This subtype likely corresponds to a beta cell subpopulation specialized for insulin synthesis, hence we named Proinsulin-High Beta cells (ProHi-Beta). In the EpiRes-Beta subpopulation, the high activity of P-GP5 indicates a state of active chromatin remodelling, possibly maintaining beta cell differentiation or responding to metabolic and inflammatory stress. Therefore, we named Epigenetic-Responsive Beta cells (EpiRes-Beta). In the EMCI-Beta subpopulation, P-GP2, P-GP4, P-GP11, and P-GP12 exhibited significant activity changes. P-GP2, P-GP11, and P-GP12 showed high activity, representing mitochondrial oxidative phosphorylation and ATP production to support insulin secretion, as well as endoplasmic reticulum. P-GP4 displayed low activity, indicating suppressed translation and ribosomal pathways, suggesting reduced protein synthesis capacity with resources preferentially allocated to secretion. Thus, we named ER-Mitochondria Coupled Insulin Secretory Beta cells (EMCI-Beta). In the StrRes-Beta subpopulation, we defined it based on P-GP2, P-GP10, and R-GP3. The high activity of P-GP2 and R-GP3 reflects endoplasmic reticulum stress and protein quality control, heightened mitochondrial metabolism, and oxidative stress defence. The low activity of P-GP10 indicates compensatory suppression of secretory function. This subtype may represent a critical transitional state during the compensation-to-decompensation shift in prediabetic beta cells. Targeting its stress pathways could potentially delay beta cell failure, and its predominant presence in the aged beta group further supports this hypothesis. Thus we named Stress responsive cell subtypes (StrRes-Beta). In the MSHy-Beta subpopulation, we defined it based on P-GP14 and P-GP15. The high activity of these two gene programs reflects mitochondrial energy metabolism and neurosecretory-cytoskeletal coordination. Thus, we named it Metabolic-Secretory Hyperactive Beta cells (MSHy-Beta). In the SPAS-Beta subpopulation, we characterized it using P-GP5, P-GP7, P-GP10, P-GP13, and P-GP14. The highly active gene programs represent chromatin structure regulation, autoimmunity, and highly active secretory pathways, while the low-activity programs indicate suppressed ribosome function, biosynthesis, and insufficient mitochondrial energy supply. This suggests that the subtype may represent a transitional state in prediabetes or early disease progression. Therefore, we named Stress-Prone Autoimmune-Sensitive Beta cells (SPAS-Beta). For the MAIS-Beta subpopulation, this was the only group isolated from the aged cohort that closely resembled subpopulations in the adult group. We analysed it based on P-GP3, P-GP5, P-GP6, P-GP8, P-GP10, and P-GP12. The high activity of these gene programs reflects enhanced mitochondrial and energy metabolism, activation of the ER-Golgi secretory system, and active chromatin remodelling and transcriptional regulation. The low activity indicates reduced glycolysis and downregulated antioxidant defence systems. These functional features suggest that this subtype likely corresponds to a functionally mature insulin-secreting beta cell population, hence we named Metabolically Active Insulin-Secreting Beta cells (MAIS-Beta). Details on beta cell subtypes, gene programs, and selected enrichment analysis functions are provided in Supplementary Table 9.

Note2: scInfer reveals a more comprehensive change across different tumour states

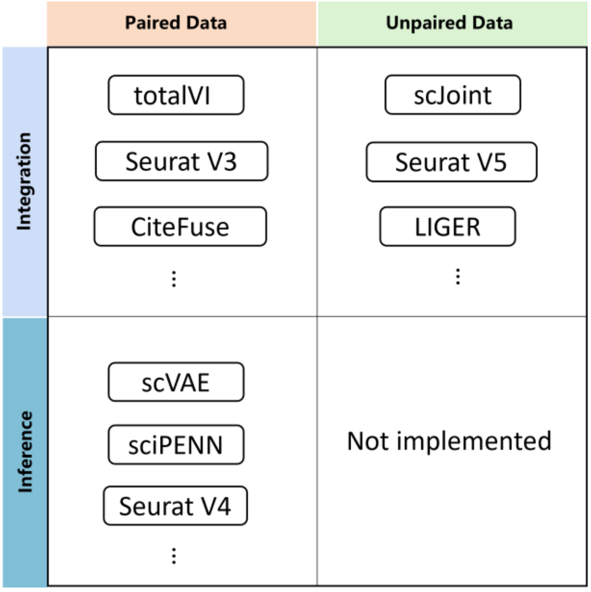
scInfer can be used to study the microenvironments associated with different stages of breast cancer. We used the single-cell RNA expression atlas measured by Bhupinder *et al.* [37]. which includes samples from human breast tissue in the normal, preneoplastic, and tumour states (a total of more than 300,000 cells from 8 normal samples, 4 BRCA1+ carrier tissue samples and 8 triple-negative breast cancer samples). We utilized scInfer, referencing Leduc 2022, Specht, and Khan which are from breast tissue to infer the corresponding proteomic data for the Bhupinder dataset. Using transcriptomic data combined with inferred proteomics data can better perform t-SNE dimensionality reduction and clustering of cells (Supplementary Fig. 6). The inferred proteomics data can be reached by Supplementary Table7-1 and Supplementary Table7-2. Subsequently, we conducted a joint analysis of transcriptomic and proteomic data across the different microenvironments associated with the normal, preneoplastic, and cancerous states. Here, we focused on mesenchymal and epithelial cells closely associated with the breast cancer microenvironment.

We conducted differential expression analysis (N-B1 and B1-TN) on the transcriptomic and proteomic data, respectively. We then summarized the gene sets on the basis of upregulation and downregulation as well as cell type, and created an UpSet plot. As shown in Supplementary Fig. 7a, in epithelial cells, the ATOX1, CSTB, and NUCKS1 genes were significantly upregulated at both the RNA and protein levels. Previous studies have shown that the ATOX1 gene can alter cell migration capabilities, which is closely related to the progression of breast cancer [38–40]. The CSTB gene has been identified as crucial for the proteolytic cascade associated with tumour progression, including tumour growth, invasion, and metastasis [41]. Additionally, the NUCKS1 gene has been confirmed to participate in tumor suppression [42,43]. In epithelial cells, 27 genes were significantly downregulated at both the RNA and protein levels, with functions related primarily to metabolic regulation, survival, and anti-apoptosis [44–48]. Similarly, as shown in Supplementary Fig. 7b, 25 genes in mesenchymal cells were downregulated at both stages, with functions also related to metabolic regulation and anti-apoptosis [49–53]. However, no genes exhibiting consistent upregulation were identified in mesenchymal cells; the specific gene sets are shown in Supplementary Table 3. Thus, scInfer efficiently analyses breast cancer progression at the cellular level from a proteomic perspective.

We then analysed the RNAs and proteins that exhibited significant expression changes at both the N-B1 and B1-TN stages within the same cell type. Specifically, we first selected the top 200 genes (100 upregulated and 100 downregulated, sorted by logarithmic fold change) that presented significant changes (p-adjusted less than 0.05) at each stage. Detailed data records can be found in Supplementary Table 3. We then took the overlap of the two gene sets and created a Sankey plot to illustrate the continuous changes in the expression of the RNAs and proteins at each stage. A substantial portion of RNAs in epithelial and mesenchymal cells exhibited a trend of downregulation followed by upregulation, while some showed upregulation followed by no change (Supplementary Fig. 7c). However, proteins predominantly displayed a continuous upregulation trend in epithelial cells, whereas a continuous downregulation trend was more evident in mesenchymal cells. As executors of cellular functions, proteins exhibit a more stable trend of change than do RNAs. Notably, during the two stages (N-B1 and B1-TN), we identified a gene, HMGA1, in epithelial cells that exhibited a persistent decline in RNA expression while protein abundance continuously increased. HMGA1 is closely related to the migration, invasion, and metastasis of TNBC cells [54], and its high protein expression can serve as a biomarker to predict metastasis incidence [55], histological grade, clinical stage [56], and survival time [57]. The protein abundance inferred by scInfer aligns with that reported in previous studies, indicating that high expression of the HMGA1 protein is highly correlated with breast cancer progression, while scRNA-seq showed contrasting results. This highlights the importance of scInfer in complementing multiomics information to comprehensively investigate the role of the tumour microenvironment.

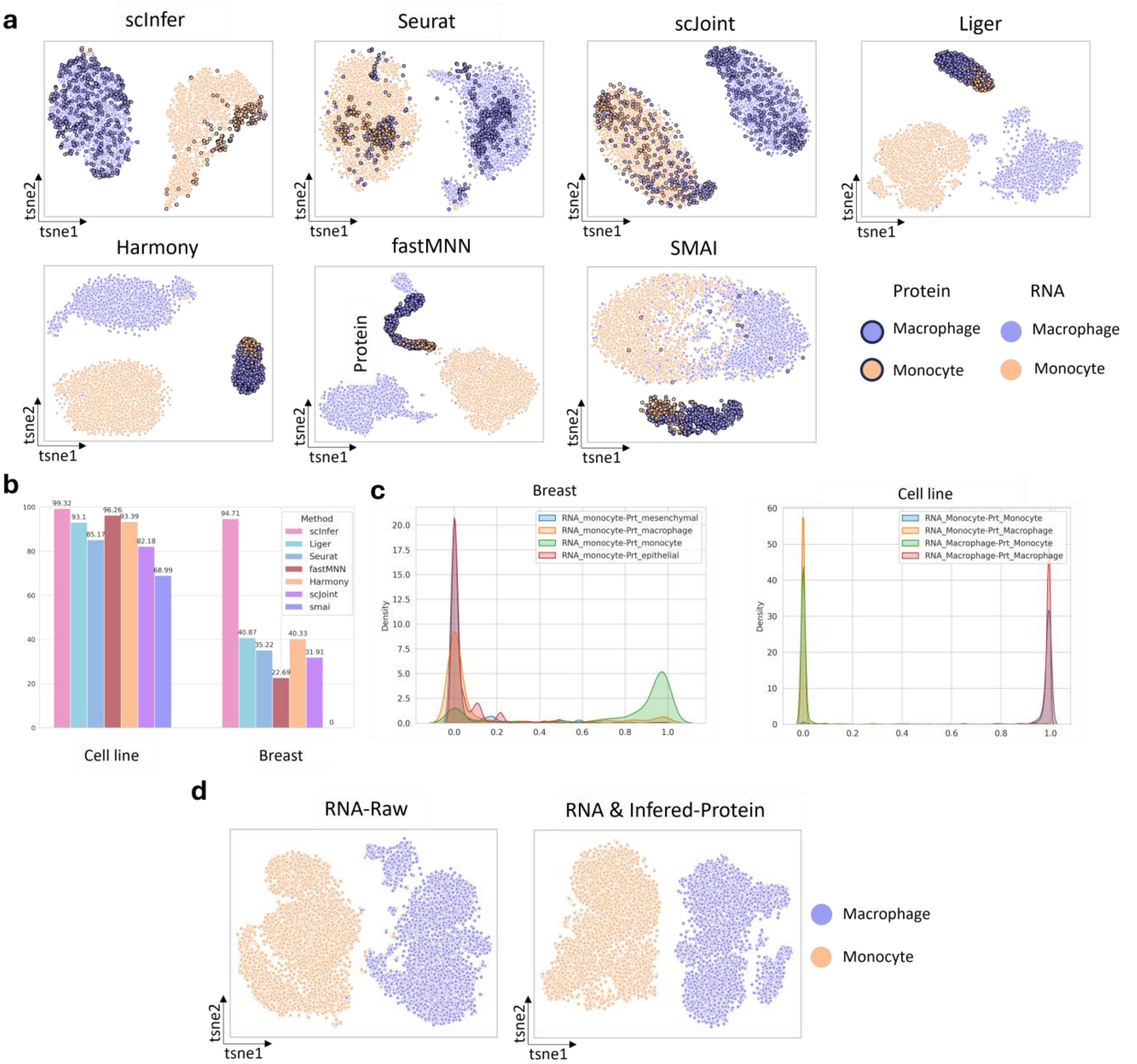
We summarized the genes whose expression was continuously upregulated or downregulated and conducted functional enrichment analysis. As shown in Supplementary Fig. 7d, we identified many pathways with similar functions in both cell types, such as extracellular exosomes, which play important roles in cancer development and metastasis. We also discovered functions unique to specific cell types; for example, apoptosis-related pathways were found in epithelial cells, whereas mitochondrial function-related pathways, such as ATP:ADP antiporter activity, were identified in mesenchymal cells. These pathways are closely related to the development and environment of cancer cells. Detailed data records can be found in Supplementary Table 3.

Supplementary Figures

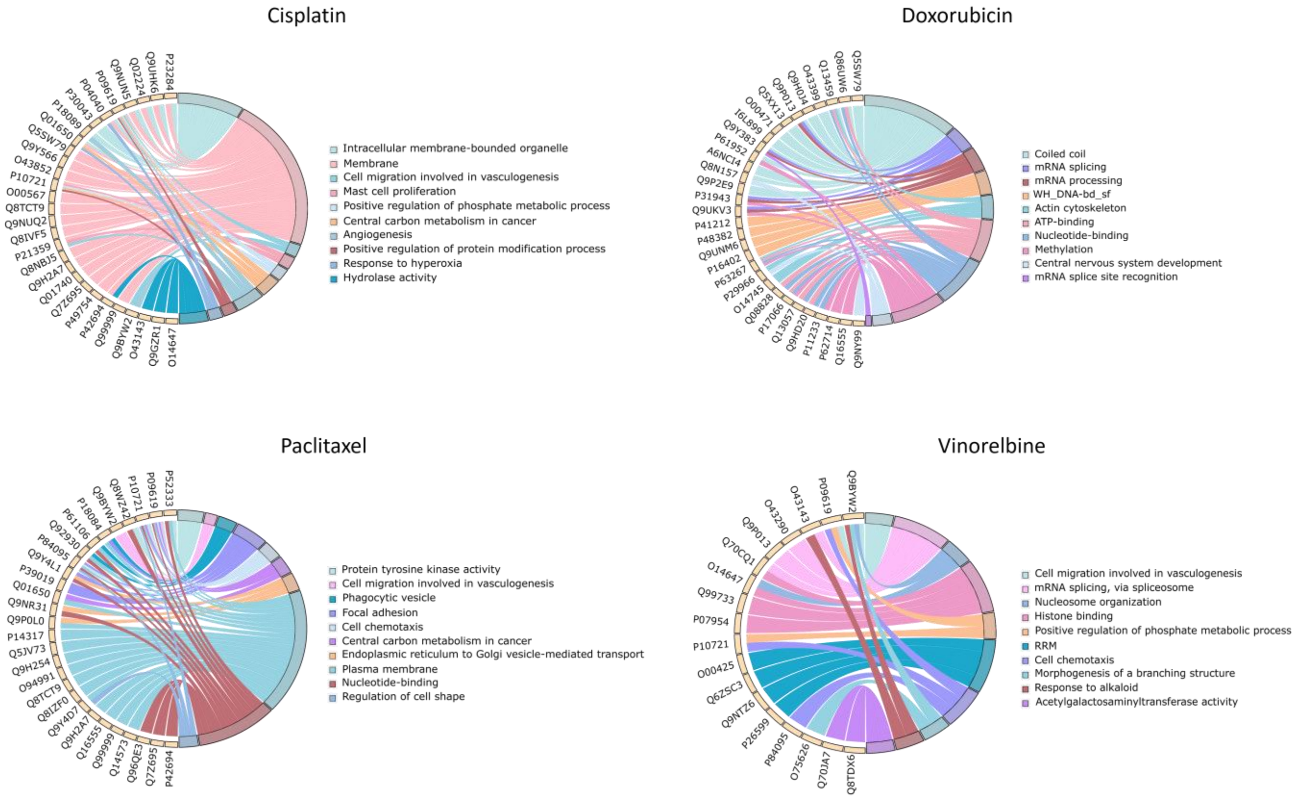


	Paired Data	Unpaired Data
Integration	totalVI Seurat V3 CiteFuse ⋮	scJoint Seurat V5 LIGER ⋮
Inference	scVAE sciPENN Seurat V4 ⋮	Not implemented

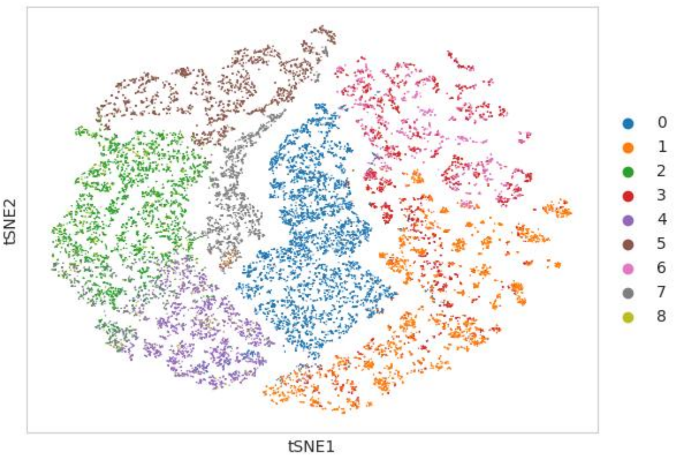
Supplementary Fig. 1 Existing methods. The two columns in the figure represent paired and unpaired data, whereas the two rows represent integration and inference methods, forming four quadrants. Within each quadrant, the existing methods are listed.



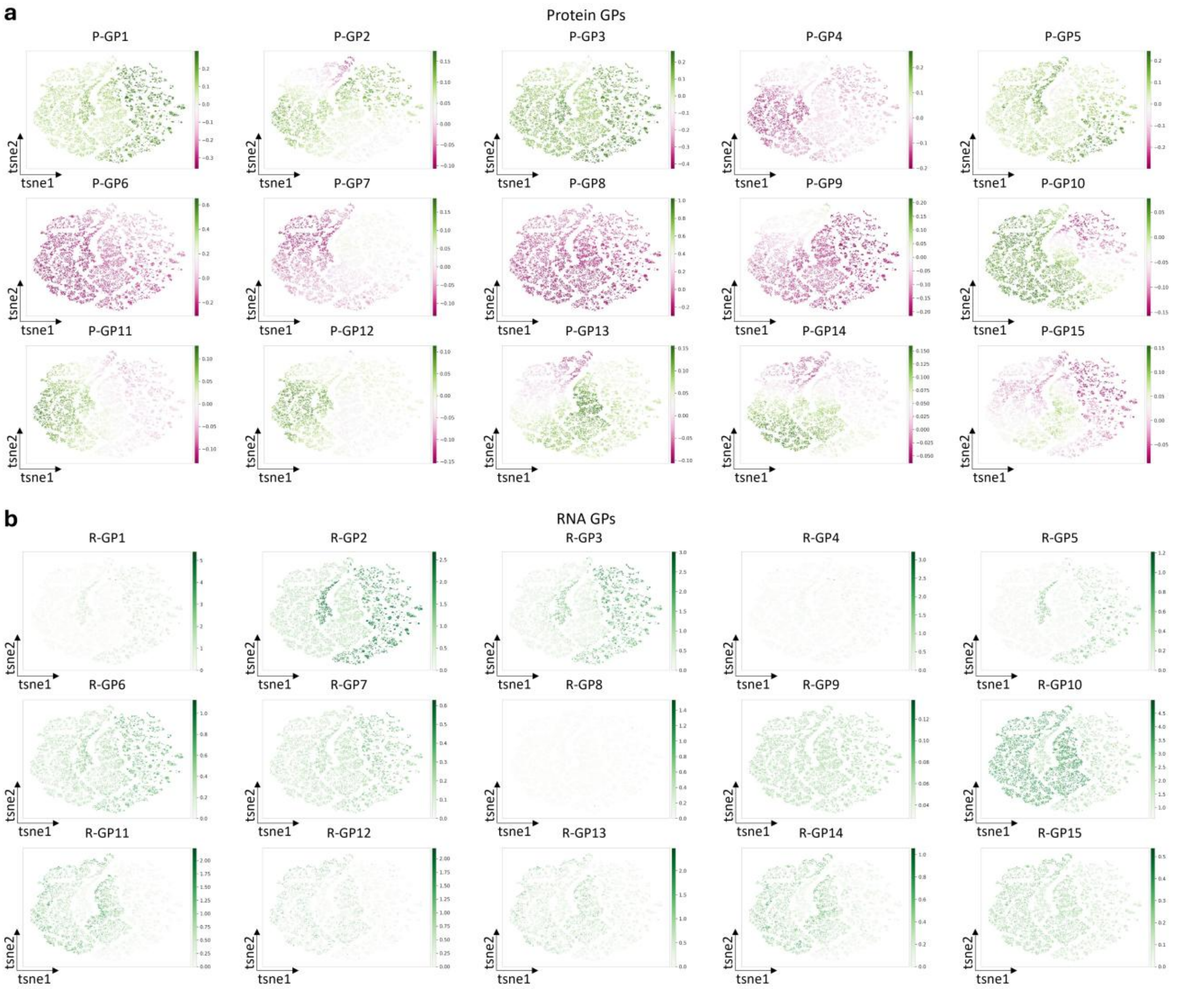
Supplementary Fig. 2 Evaluations on unpaired datasets. a. T-SNE scatter plot displays the result of integrating unpaired transcriptomics and proteomics data in cell line task. The scatter points without edges represent transcriptomic cells, whereas the scatter points with black edges represent proteomic cells. Different cell types are distinguished by fill colours. It is preferable for cells of the same type from two omics to cluster closely together, whereas different types should be more distant. **b.** Bar plot of the cell type matching accuracy. Cosine similarity is calculated on the basis of embedded features to find the most matching cells. The proportion of consistent matches among cell types is calculated as the accuracy. **c.** Distribution of cosine similarity for all cells. The Cartesian product of cell collections from two types of omics is performed to obtain cell pairs, followed by the calculation of the cosine similarity for each cell pair and kernel density visualization of the similarity across all cell pairs. **d.** T-SNE plot of using transcriptomics and combined with proteomics predicted by scInfer in cell line task.



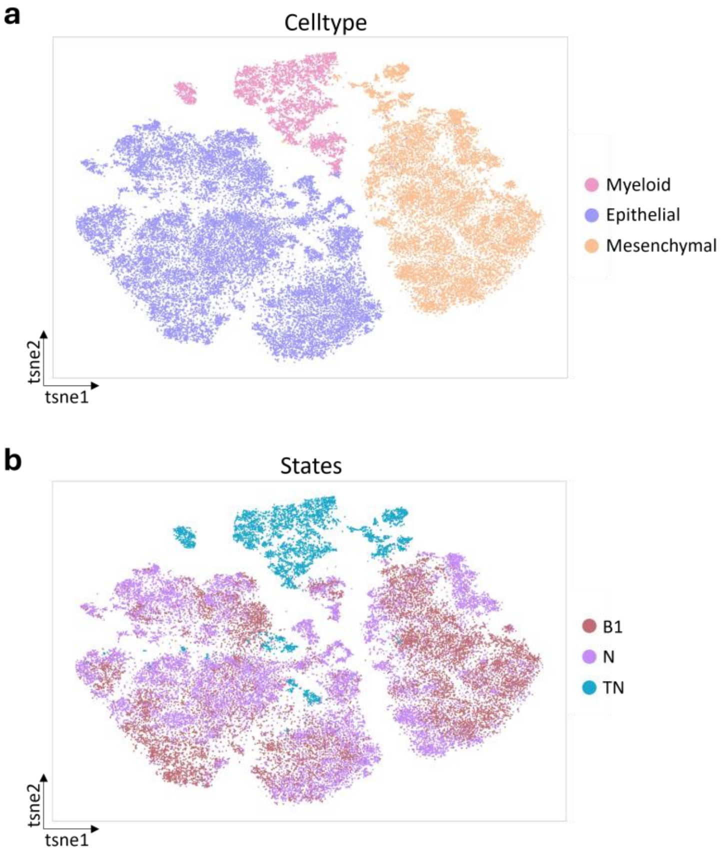
Supplementary Fig. 3 Differential Enrichment Analysis Plot. Under different medications, we first conduct enrichment analysis on the differentially expressed proteins and RNAs between the treatment group and the control group. Then, we extract the enrichment analysis results that appeared in the protein analysis but were not involved in the RNA analysis results.



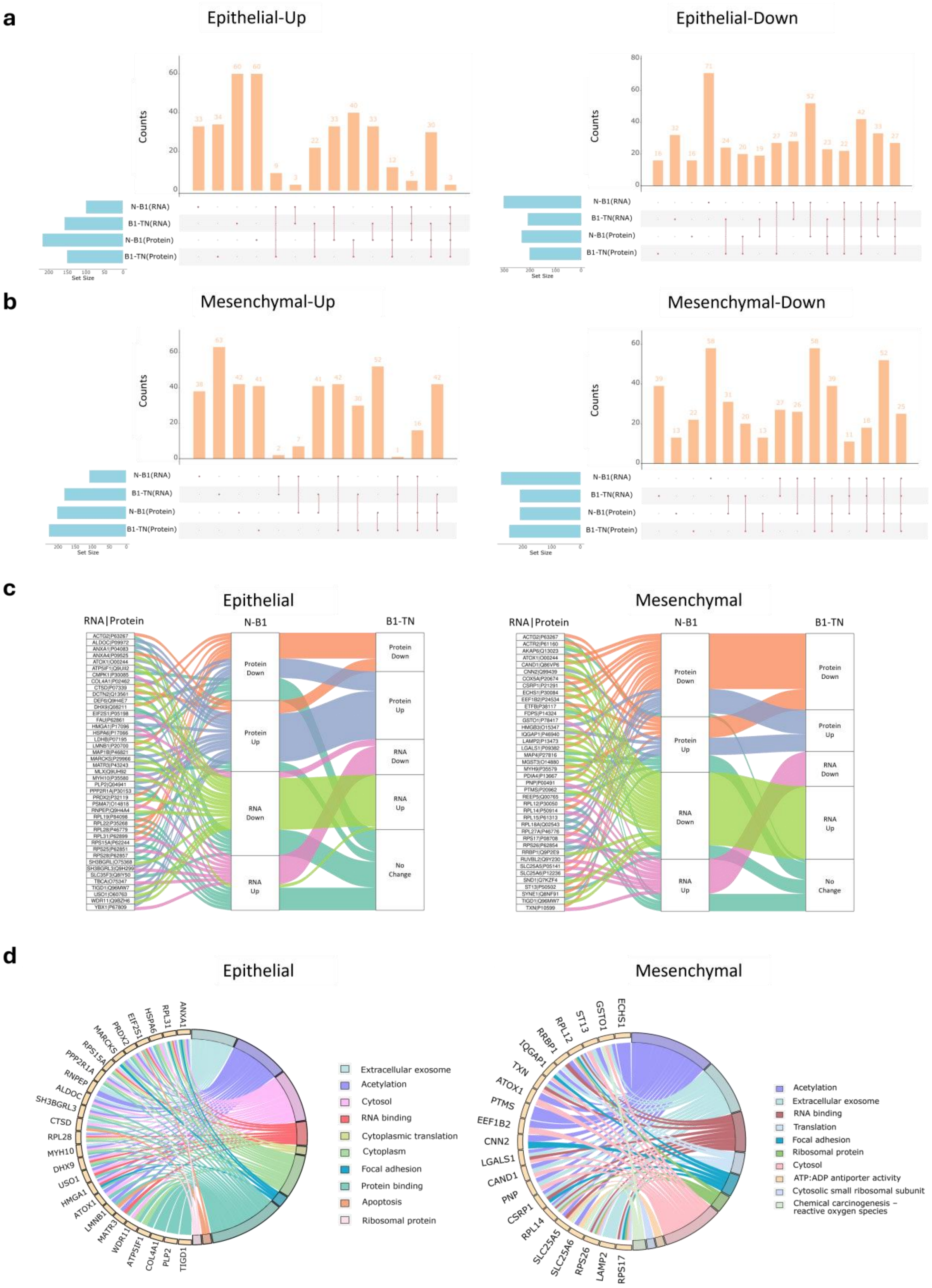
Supplementary Fig. 4 Cell clustering results using Leiden algorithm on gene program activity data.



Supplementary Fig. 5 Gene programs activity distribution on RNA and protein. a. Protein GPs. **b.** RNA GPs.



Supplementary Fig. 6 T-SNE plot of using transcriptomics and combined with proteomics predicted by scInfer in N, B1, and TN group cells.



Supplementary Fig. 7 Single-cell RNA and protein changes in different environments. a. UpSet plot showing the number of differentially expressed RNAs and proteins in epithelial cells at different cancer stages. N represents the normal (healthy) state, B1 indicates the preneoplastic state (BRCA1+/–), and TN denotes cancer patients. For example, N-B1(RNA) indicates the differentially expressed RNAs between the preneoplastic state and the normal state, with the upper bar plot showing the number of genes (33) and the line connecting the points indicating the overlap of different groups. **b.** UpSet plot showing the number of differentially expressed RNAs and proteins in mesenchymal cells at different cancer stages. **c.** Sankey plot illustrating the continuous changes in RNA and protein levels from the normal to preneoplastic to cancerous state. The top 200 differentially expressed RNAs and proteins shared between the N-B1 and B1-TN states were selected for plotting, with No change indicating no significant differences. **d.** RNAs and proteins whose expression continuously changed, as shown in Fig. 5c, were selected for functional enrichment analysis. For example, ALDOC was selected for enrichment analysis since its protein expression is upregulated in both N-B1 and B1-TN. The same pathways are represented in the same colour in the two enrichment analysis diagrams.

Supplementary Tables

The supplementary tables are stored on Zendo(https://doi.org/10.5281/zenodo.14986872)