Large-scale proteome inference from unpaired single-cell transcriptomic and proteomic data by msInfer

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: Yuzhi Sun (yuzhi@stu.hit.edu.cn), 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: msInfer reveals a more comprehensive change across different tumour states

msInfer 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.* [1]. 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 msInfer, 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. S5). 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.

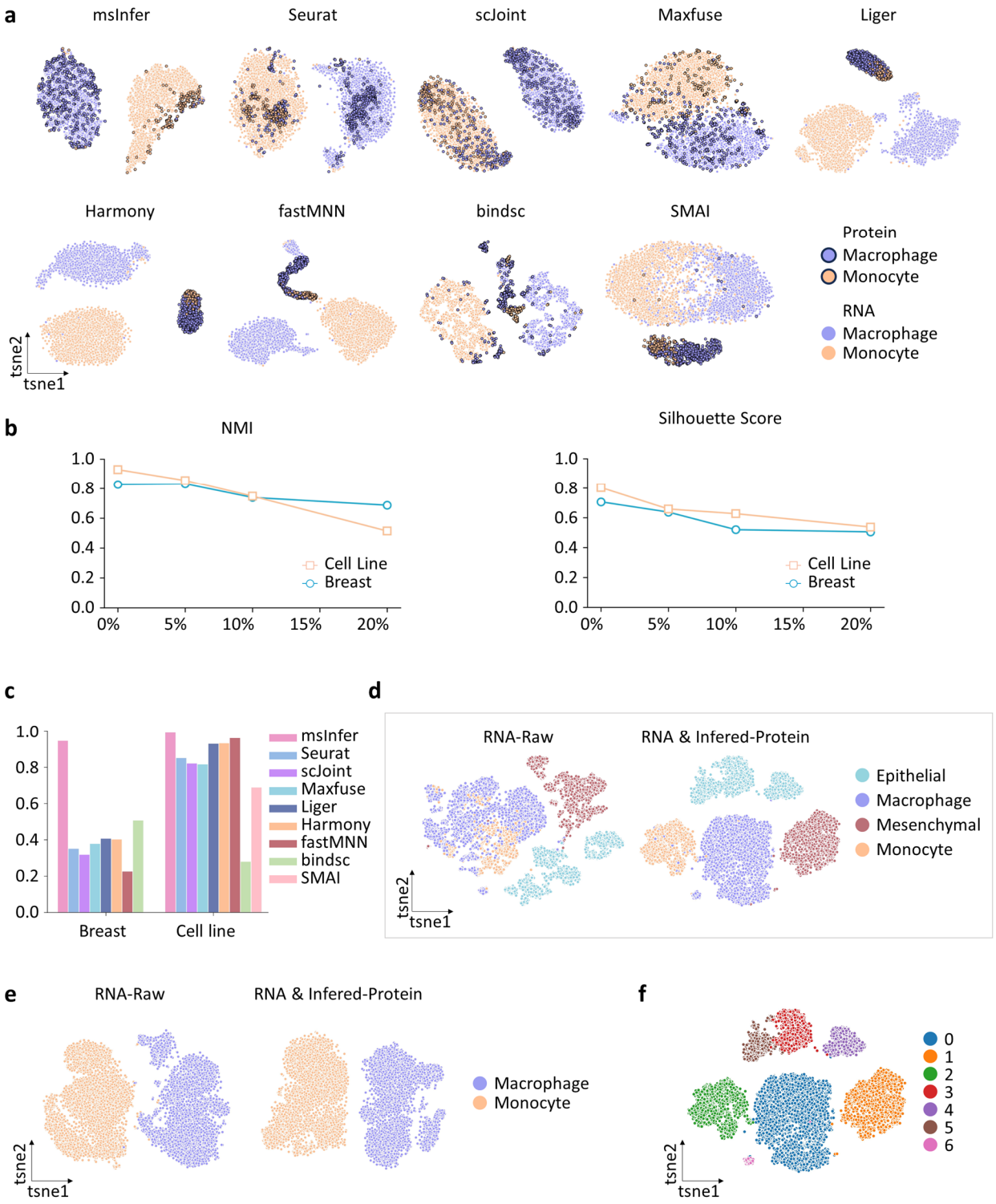
We conducted differential expression analysis (N-B1 and B1-TN) on the transcriptomic and proteomic data, respectively. We then summarized the gene sets on the basis of upregulation and downregulation as well as cell type, and created an UpSet plot. As shown in Supplementary Fig. S6a, in epithelial cells, 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 [2–4]. The CSTB gene has been identified as crucial for the proteolytic cascade associated with tumour progression, including tumour growth, invasion, and metastasis [5]. Additionally, the NUCKS1 gene has been confirmed to participate in tumour suppression [6,7]. 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 [8–12]. Similarly, as shown in Supplementary Fig. S6b, 25 genes in mesenchymal cells were downregulated at both stages, with functions also related to metabolic regulation and anti-apoptosis [13–17]. However, no genes exhibiting consistent upregulation were identified in mesenchymal cells. Thus, msInfer 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. 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. S6c). 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 [18], and its high protein expression can serve as a biomarker to predict metastasis incidence [19], histological grade, clinical stage [20], and survival time [21]. The protein abundance inferred by msInfer 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 msInfer in complementing multiomics information to comprehensively investigate the role of the tumour microenvironment.

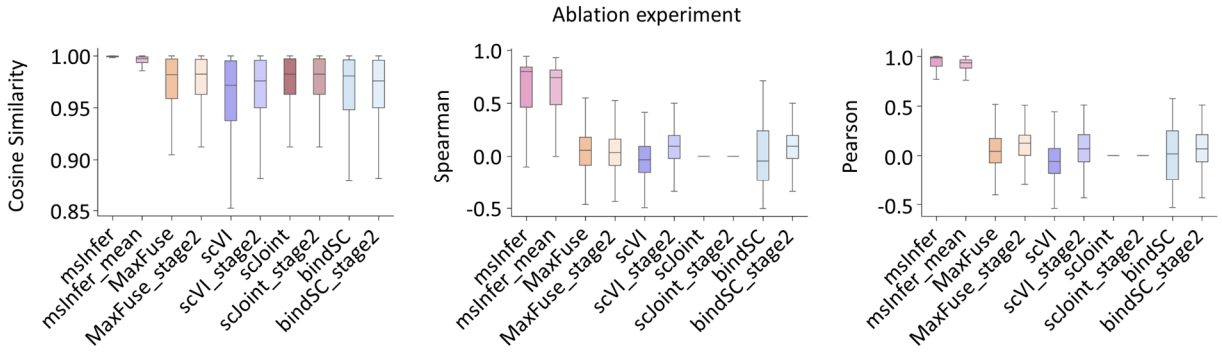
We summarized the genes whose expression was continuously upregulated or downregulated and conducted functional enrichment analysis. As shown in Supplementary Fig. S6d, we identified many pathways with similar functions in both cell types, such as extracellular exosomes, which play important roles in 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.

Supplementary Figures

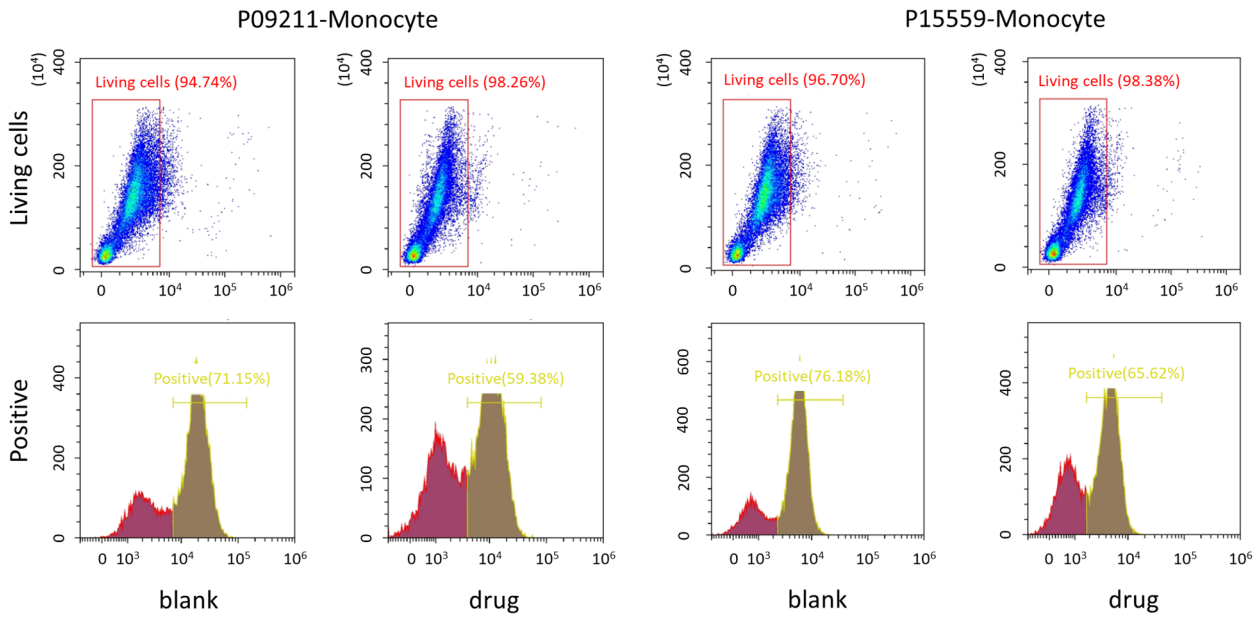


Supplementary Fig. S1 Validation on scRNA-seq and Mass Spectrometry-based Single-cell Proteomics Dataset. a. t-SNE visualization of transcriptomic and proteomic cells distributions with integrated embedding features on Cell Line task. **b.** Clustering performance (NMI and Sihouette Score) under disturbance. **c.** Accuracy of cell type matching. **d.** t-SNE visualization of cells with

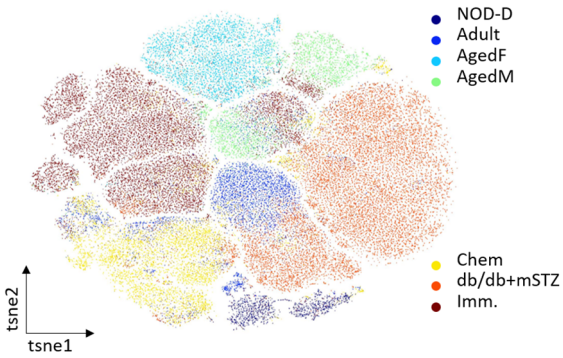
RNA data and RNA & inferred-protein data. **e.** t-SNE visualization of cells with RNA data and RNA & inferred-protein data on Cell Line task. **f.** Cell distribution map after Leiden clustering of transcriptome and proteome data from the breast dataset. derived from Wu et al. [50] .



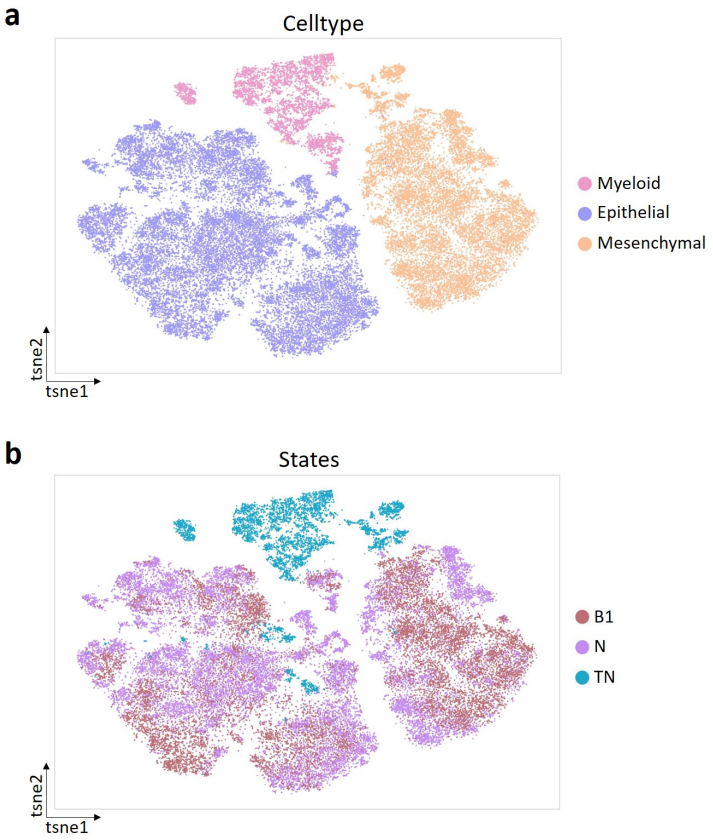
Supplementary Fig. S2 Ablation experiment on benchmark. Use mean calculation to replace stage 2 of msInfer for prediction, and use integrated methods combined with stage 2 for prediction.



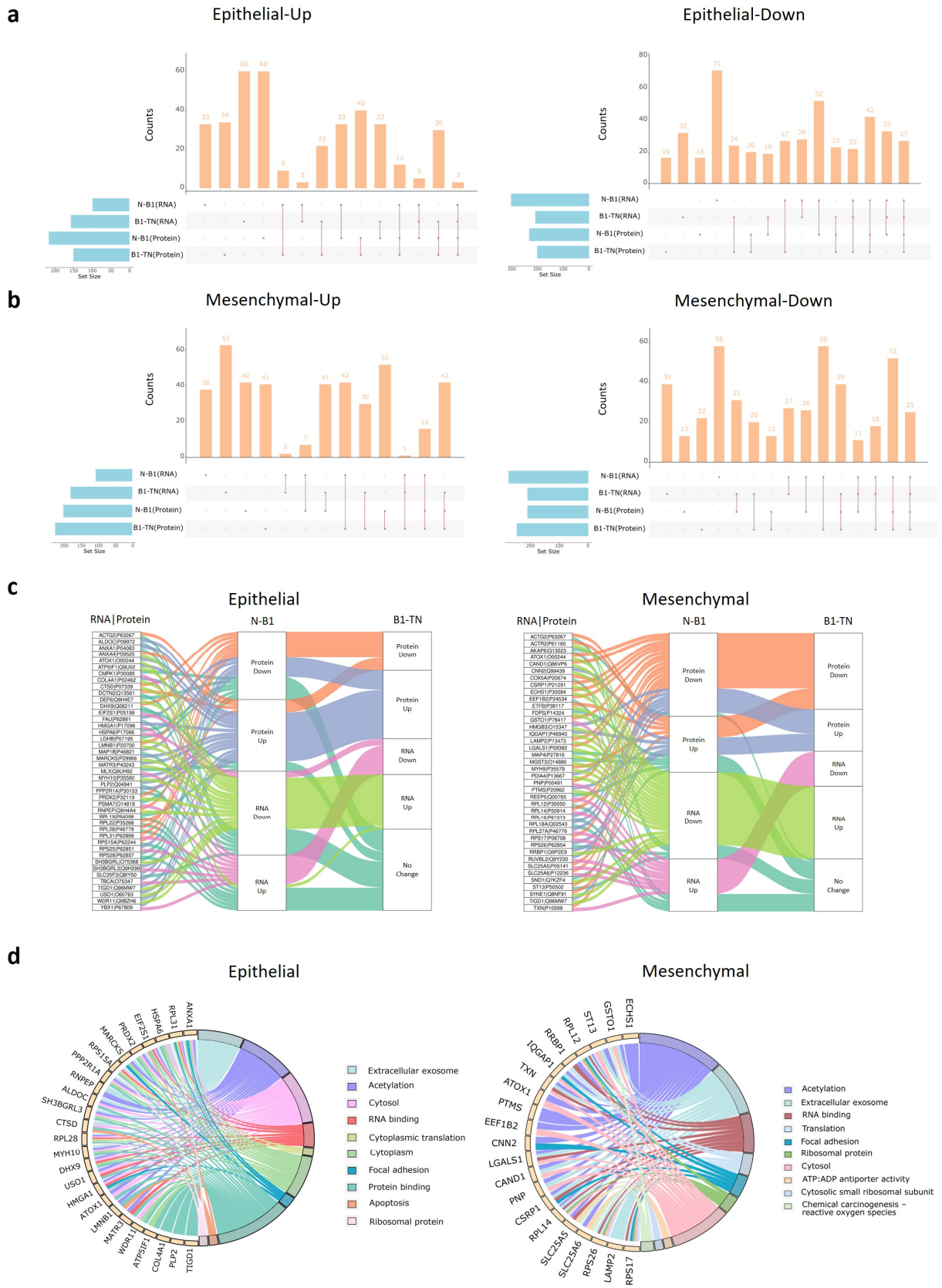
Supplementary Fig. S3 Changes in the levels of proteins P09211 and P15559 in monocytes before and after treatment with cisplatin measured by flow cytometry.



Supplementary Fig. S4 Scatter plot of beta atlas cells distribution only using transcriptome data



Supplementary Fig. S5 T-SNE plot of using transcriptomics and combined with proteomics predicted by msInfer in N, B1, and TN group cells.



Supplementary Fig. S6 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.