Supplementary Figures

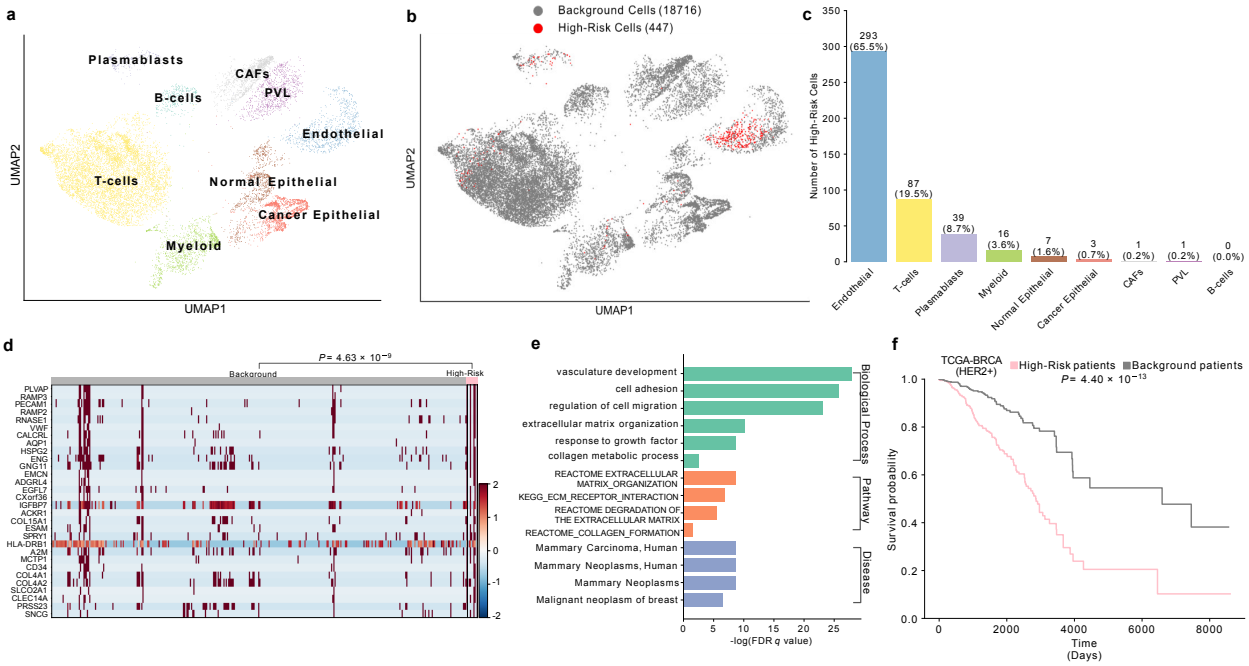

Fig. S1: **Full Analysis of HER2+ Breast Cancer Subtype. a,** UMAP plot showing the clustering of identified cell types, including cancer epithelial cells, cancer-associated fibroblasts (CAFs), perivascular-like (PVL) cells, normal epithelial cells, myeloid cells, T-cells, B-cells, plasmablasts, and endothelial cells. **b,** UMAP visualization of the HER2+ dataset, containing 19,163 cells. High-Risk cells (447) identified by SIDISH are shown in red, while Background cells (18,716) are depicted in gray. **c,** Bar plot showing the distribution of High-Risk cells across various cell types. **d,** Heatmap of differential gene expression analysis between High-Risk and Background cells. High-Risk cells exhibit significantly higher expression of marker genes ($P = 4.63 \times 10^{-9}$). P values were calculated using a one-sided Mann-Whitney U test to highlight differences in expression levels between High-Risk and Background cells. **e,** Functional enrichment analysis, including GO terms, pathways, and disease terms, highlights upregulated genes associated with tumor aggressiveness and invasive behavior. **f,** Kaplan-Meier survival analysis on the TCGA-BRCA bulk data demonstrates significantly poorer survival outcomes for patients with higher expression levels of marker genes identified by SIDISH ($P = 4.40 \times 10^{-13}$). High-Risk patients (pink) show a distinct separation from Background patients (gray), validating SIDISH's predictive power. P values were calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups across all cohorts.

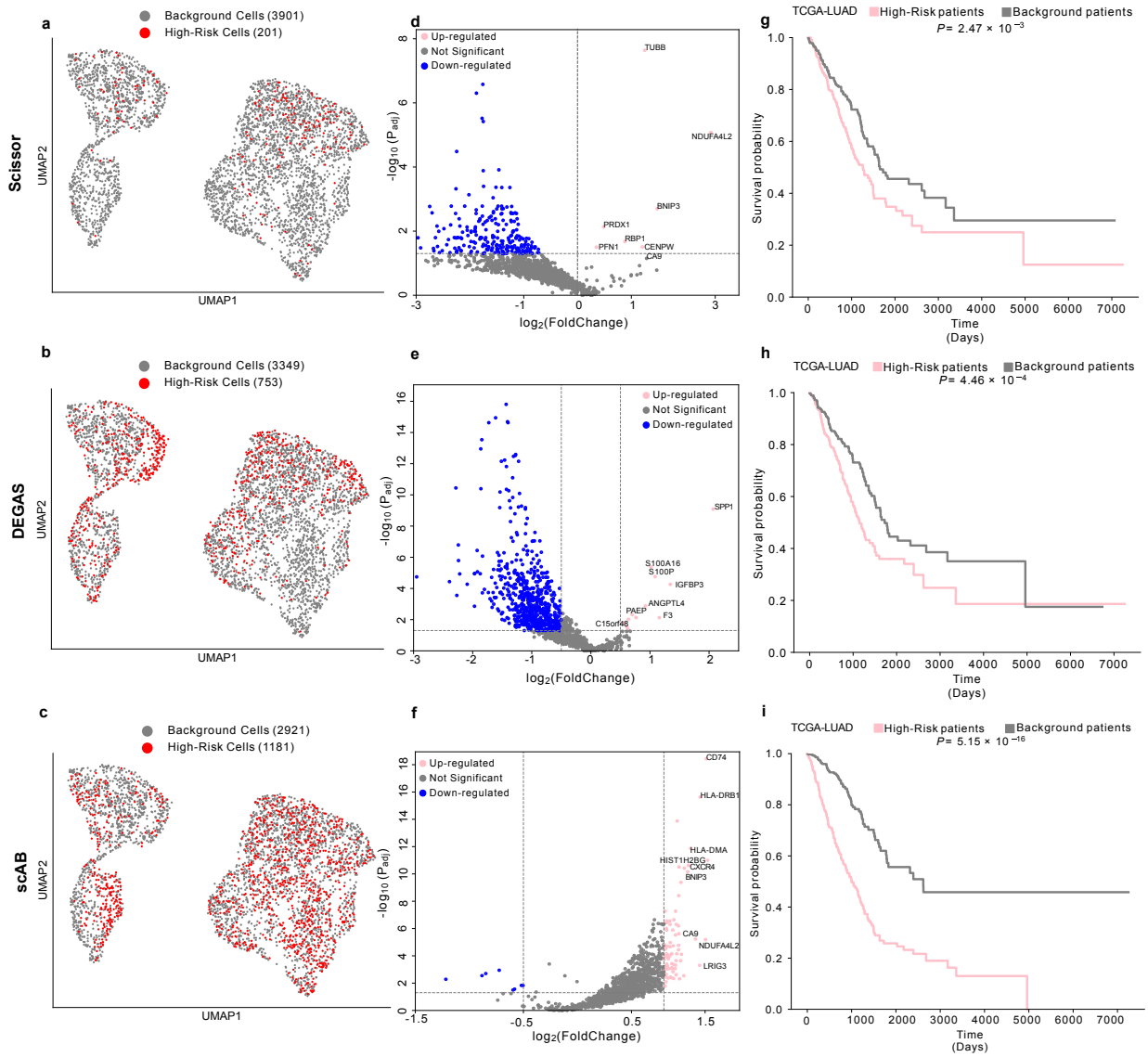

Fig. S2: **Overview of Existing Methods Results in the LUAD Dataset. a-c,** UMAP visualizations of High-Risk cells identified by Scissor (**a**), DEGAS (**b**), and scAB (**c**). For the Scissor method, cells classified as Scissor+ were considered High-Risk. For DEGAS, cells with a score above 0.5 were classified as High-Risk. For scAB, any cells not labeled as "Other" were classified as High-Risk. **d-f,** Volcano plots for Scissor (**d**), DEGAS (**e**), and scAB (**f**), highlighting upregulated and downregulated genes. **g-i,** Kaplan-Meier survival plots based on TCGA-LUAD patient data for Scissor (**g**), DEGAS (**h**), and scAB (**i**). Upregulated genes were used to stratify patients based on their expression levels. A broader separation between survival curves indicates better stratification between patients at higher risk of death and those at lower risk. *P*values were calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups.

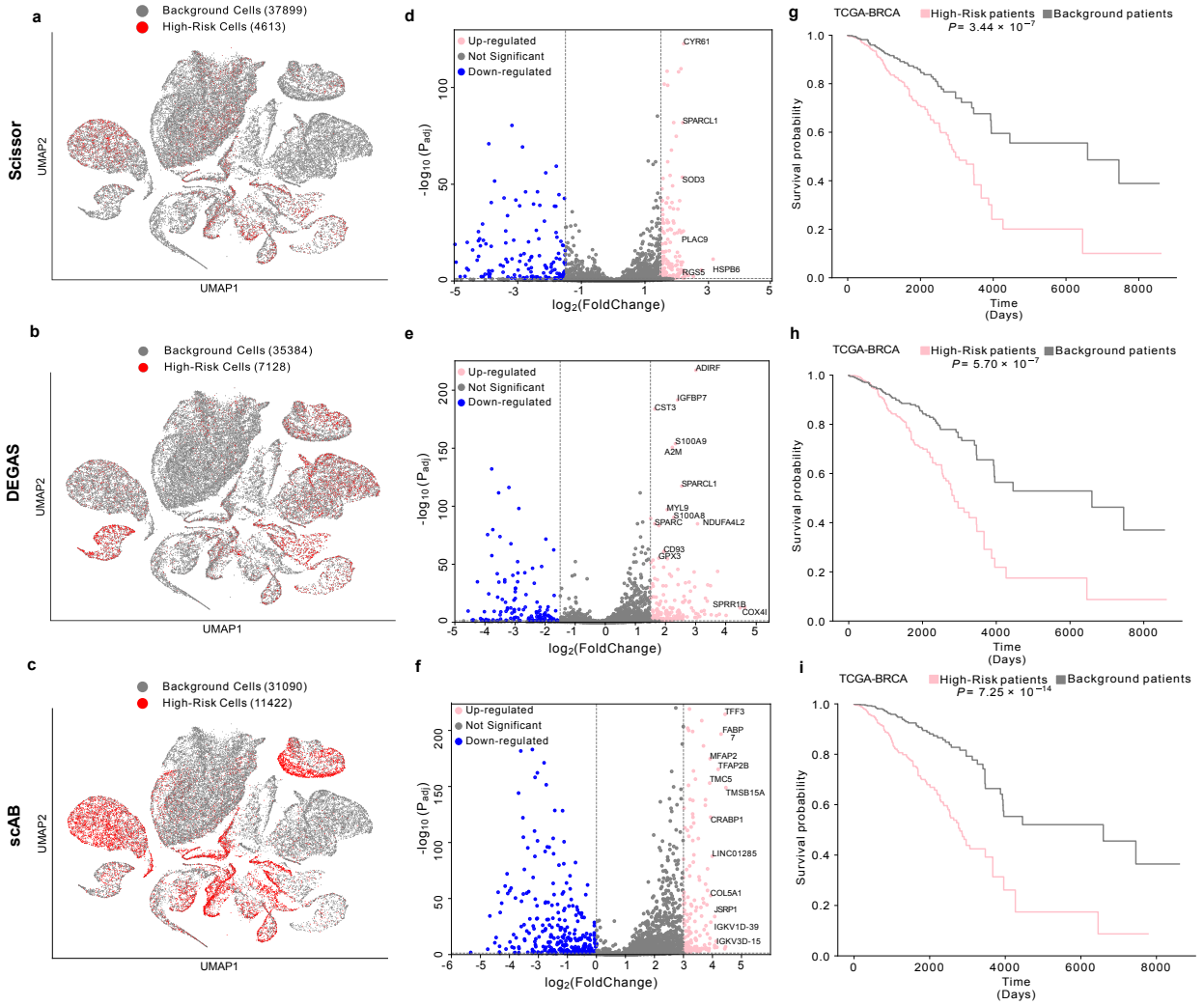

Fig. S3: **Overview of Existing Methods Results in the BRCA Dataset. a-c,** UMAP visualizations of High-Risk cells identified by Scissor (**a**), DEGAS (**b**), and scAB (**c**). For the Scissor method, cells classified as Scissor+ were considered High-Risk. For DEGAS, cells with a score above 0.5 were classified as High-Risk. For scAB, any cells not labeled as "Other" were classified as High-Risk. **d-f,** Volcano plots for Scissor (**d**), DEGAS (**e**), and scAB (**f**), highlighting upregulated and downregulated genes. **g-i,** Kaplan-Meier survival plots based on TCGA-BRCA patient data for Scissor (**g**), DEGAS (**h**), and scAB (**i**). Upregulated genes were used to stratify patients based on their expression levels. A broader separation between survival curves indicates better stratification between patients at higher risk of death and those at lower risk. *P* values were calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups.

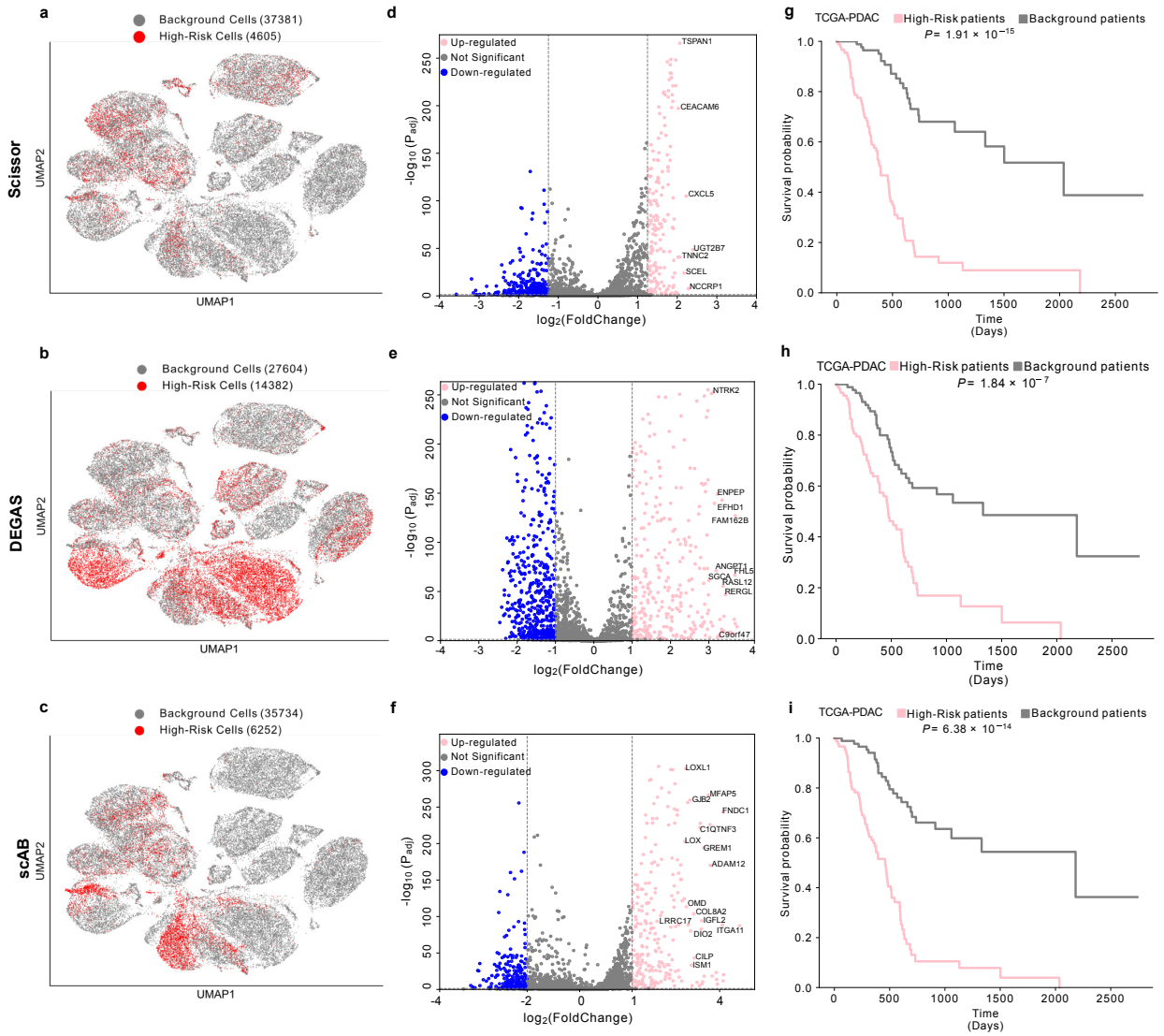

Fig. S4: **Overview of Existing Methods Results in the PDAC Dataset. a-c,** UMAP visualizations of High-Risk cells identified by Scissor (**a**), DEGAS (**b**), and scAB (**c**). For the Scissor method, cells classified as Scissor+ were considered High-Risk. For DEGAS, cells with a score above 0.5 were classified as High-Risk. For scAB, any cells not labeled as "Other" were classified as High-Risk. **d-f,** Volcano plots for Scissor (**d**), DEGAS (**e**), and scAB (**f**), highlighting upregulated and downregulated genes. **g-i,** Kaplan-Meier survival plots based on TCGA-PDAC patient data for Scissor (**g**), DEGAS (**h**), and scAB (**i**). Upregulated genes were used to stratify patients based on their expression levels. A broader separation between survival curves indicates better stratification between patients at a higher risk of death and those at a lower risk. *P* values were calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups.

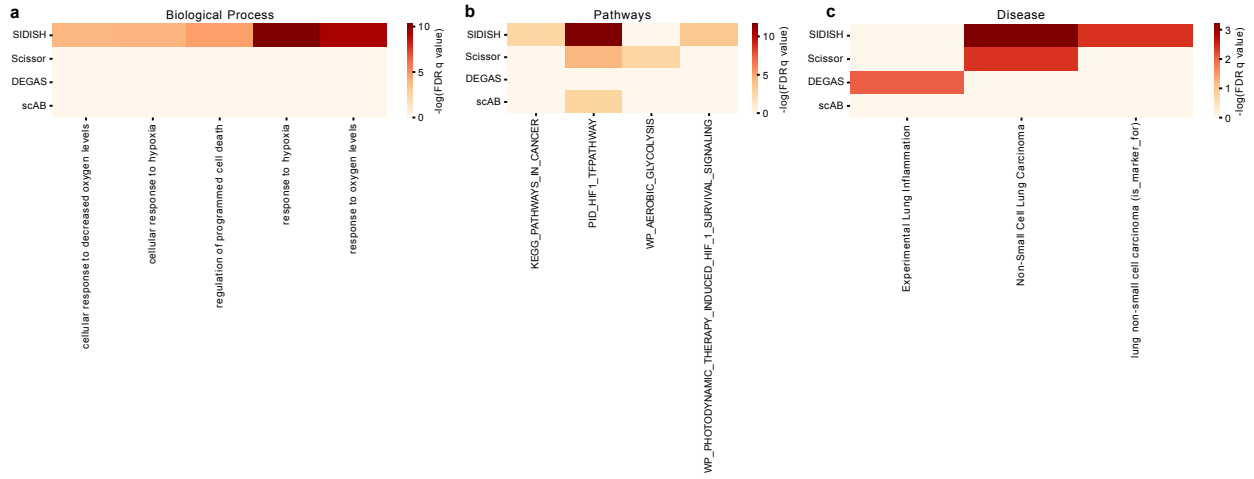

Fig. S5: **Enrichement Analysis benchmark between SIDISH, Scissor, DEGAS, and scAB in the LUAD scRNA-seq dataset**. **a,** GO enrichment analysis between all 4 methods, shows SIDISH marker genes being significantly more enriched than Scissor, DEGAS, and scAB. **b,** Pathway enrichment analysis between all 4 methods. **c,** Disease enrichment between all 4 methods.

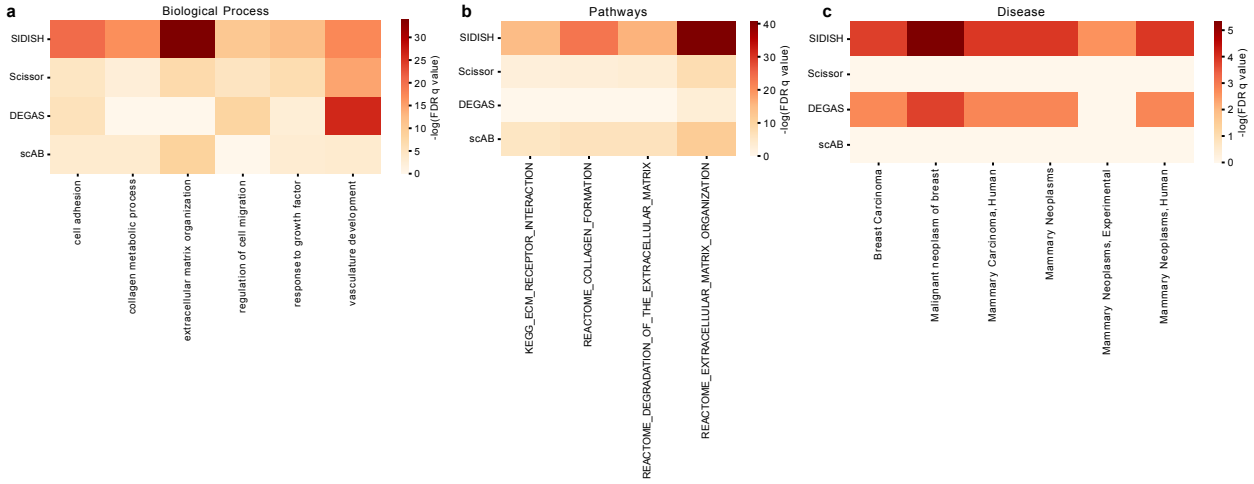

Fig. S6: **Enrichment Analysis Benchmark Between SIDISH, Scissor, DEGAS, and scAB in the BRCA scRNA-seq Dataset. a,** GO enrichment analysis comparing all four methods shows that SIDISH marker genes are significantly more enriched than those from Scissor, DEGAS, and scAB. **b,** Pathway enrichment analysis comparing all four methods. **c,** Disease enrichment analysis comparing all four methods.

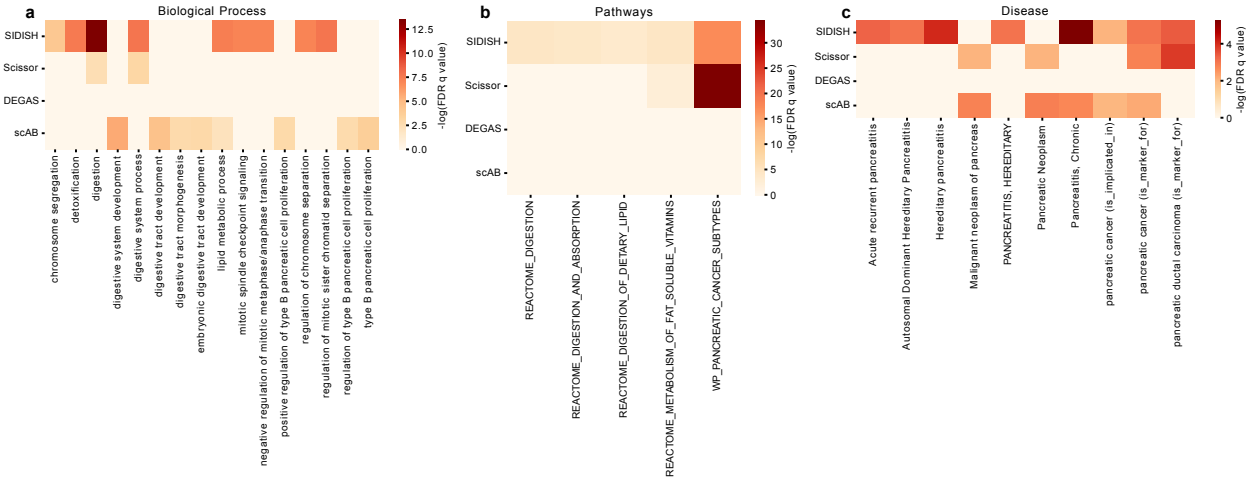

Fig. S7: **Enrichment Analysis Benchmark Between SIDISH, Scissor, DEGAS, and scAB in the PDAC scRNA-seq Dataset. a,** GO enrichment analysis comparing all four methods shows that SIDISH marker genes are significantly more enriched than those from Scissor, DEGAS, and scAB. **b,** Pathway enrichment analysis comparing all four methods. **c,** Disease enrichment analysis comparing all four methods.

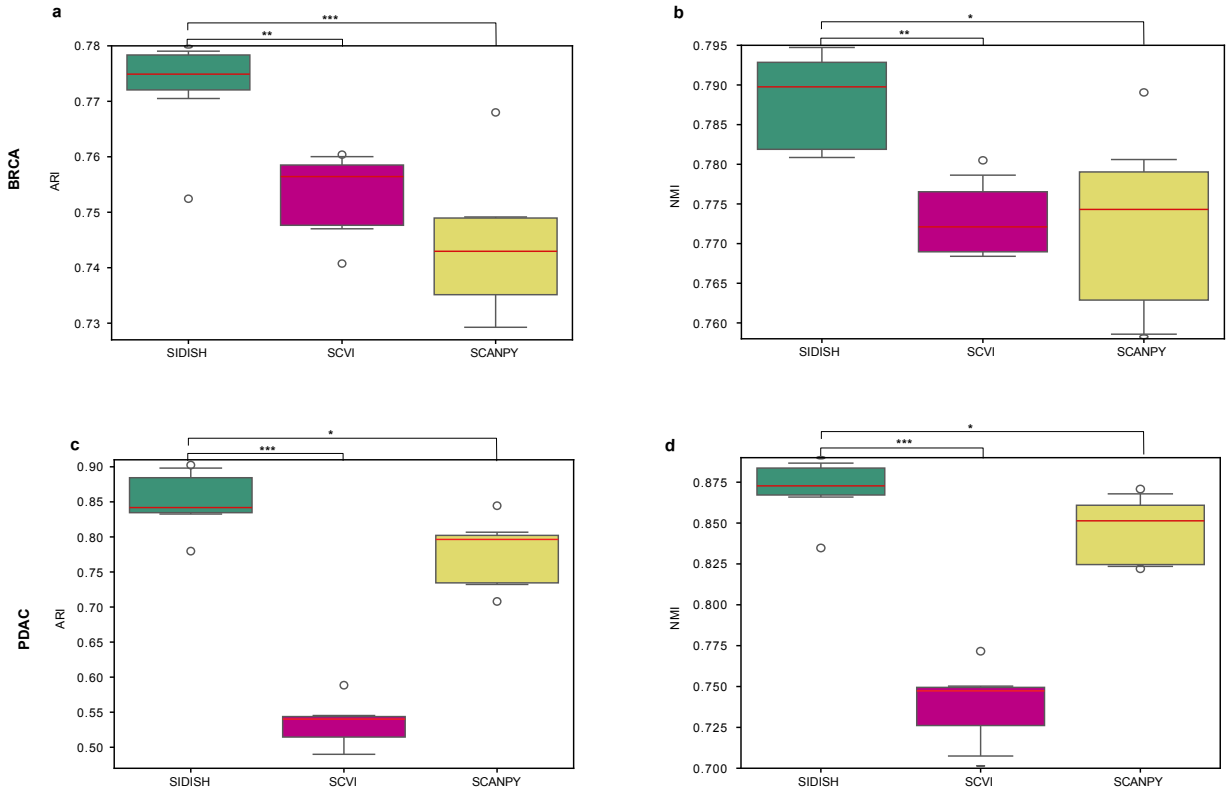

Fig. S8: **SIDISH Embeddings Achieve Superior Clustering Performance Compared to State-of-the-Art Tools. a-b,** Box plots demonstrate that SIDISH's embeddings achieve significantly better clustering performance than state-of-the-art tools such as Scanpy and SCVI in the BRCA dataset. SIDISH outperforms these tools in Adjusted Rand Index (ARI, **a**) and Normalized Mutual Information (NMI, **b**) metrics. **c-d,** In the PDAC dataset, SIDISH-derived embeddings also yield superior clustering performance compared to Scanpy and SCVI. Statistical significance was assessed using Mann-Whitney U tests: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$; n.s., not significant.

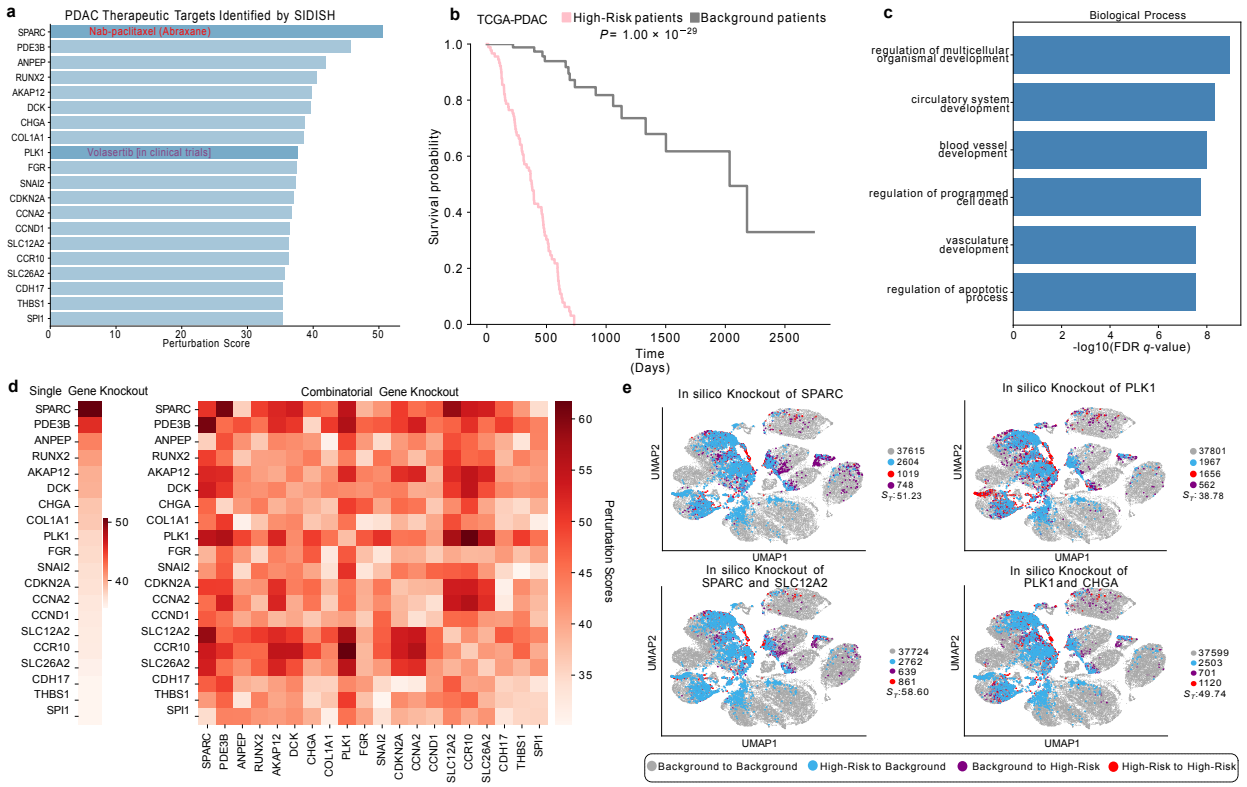

Fig. S9: **In silico Gene Perturbation with SIDISH Identifies Key Therapeutic Targets and Reduces High-Risk Cells in PDAC scRNA-seq Dataset. a,** Ranked bar plot of therapeutic targets identified by SIDISH in pancreatic ductal adenocarcinoma (PDAC) based on perturbation scores. Top-ranking genes, including *SPARC* and *PLK1*, exhibit the highest scores, highlighting their potential as therapeutic targets. An FDA-approved drug such as Nab-paclitaxel (red) is bound by *SPARC* facilitating drug delivery within the tumor, and a clinically relevant drug such as Volasertib (purple), targeting *PLK1* (currently in clinical trials for PDAC), is prominently featured, underscoring the therapeutic potential of SIDISHs in silico perturbation feature. **b,** Kaplan-Meier survival analysis of TCGA-PDAC patients demonstrates significantly poorer survival outcomes for High-Risk patients, characterized by higher expression of the top perturbed genes (pink), compared to Background patients with lower expression levels (gray) ($P = 1 \times 10^{-29}$). The P value was calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups. **c,** Gene Ontology (GO) enrichment analysis of the potential therapeutic candidate genes identified by SIDISH reveals significant enrichment in regulated processes associated with poor survival. **d,** Heatmaps of single-gene and combinatorial gene knockouts, ranked by perturbation scores. Single-gene knockouts, such as *SPARC* and *PLK1*, show substantial reductions in High-Risk cells, while combinatorial knockouts, including *SPARC and SLC12A2* and *PLK1 and CHGA*, result in greater reductions in High-Risk cell subpopulations. **e,** UMAP visualizations illustrating the spatial effects of in-silico knockouts on High-Risk and Background cell subpopulations. Knockout of *SPARC* achieved perturbation scores (S_T) of 51.23 and 38.78. Combinatorial knockouts, such as *SPARC and SLC12A2* (58.60) and *PLK1 and CHGA* (49.74), achieve even greater scores. The UMAP color legend categorizes transitions: ray for unchanged Background cells (Background to Background), blue for High-Risk cells converting to Background (High-Risk to Background), purple for Background cells converting to High-Risk (Background to High-Risk), and red for persistent High-Risk cells (High-Risk to High-Risk). S_T represents the perturbation score.

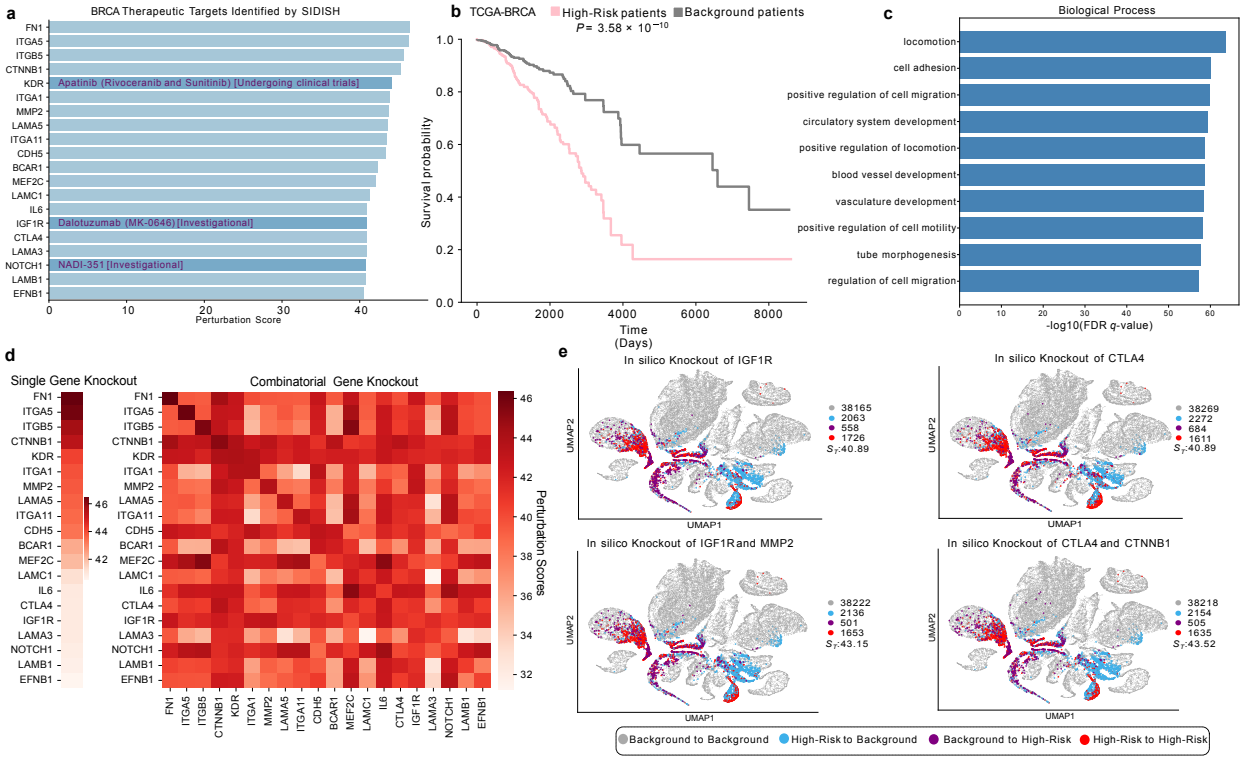

Fig. S10: **In silico Gene Perturbation with SIDISH Identifies Key Therapeutic Targets to Reduce High-Risk Cells in the BRCA scRNA-seq Dataset. a, a,** Ranked bar plot of therapeutic targets identified by SIDISH in BRCA based on perturbation scores. High-ranking genes, including *FN1*, *ITGA5*, and *ITGB5*, demonstrate strong potential as therapeutic targets. Clinically relevant drugs linked to SIDISH-identified targets include Apatinib (targeting *KDR*), Rivoceranib, and Sunitinib (targeting *KDR*), and investigational agents such as Atezolizumab (MK-0646, targeting *IGF1R*), and NADI-351 (targeting *NOTCH1*). These drugs are all being investigated for use in BRCA. **b,** Kaplan-Meier survival analysis of TCGA-BRCA patients demonstrates significantly poorer survival outcomes for High-Risk patients, characterized by higher expression of the top perturbed genes (pink), compared to Background patients with lower expression levels (gray) ($P = 3.58 \times 10^{-10}$). The P value was calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups. **c,** Gene Ontology (GO) analysis reveals biological processes that are critical to hindering cancer progression and therapeutic resistance. **d,** Heatmaps showing perturbation scores for single-gene and combinatorial gene knockouts. Single-gene knockouts, such as *IGF1R* and *CTLA4*, result in substantial reductions in High-Risk cells. Combinatorial knockouts, such as *IGF1R and MMP2* and *CTLA4 and CTNNB1*, obtain an even greater reduction in High-Risk cell subpopulations compared to single-gene perturbations. **e,** UMAP visualizations illustrating the spatial effects of in-silico gene perturbations on High-Risk and Background cells. Knockouts of *IGF1R* and *CTLA4* achieve equal perturbation scores (S_T) of 40.89%, respectively. Combinatorial knockouts, such as *IGF1R and MMP2* (43.15) and *CTLA4 and CTNNB1* (43.52), achieve higher perturbation scores. The UMAP color legend categorizes transitions: ray for unchanged Background cells (Background to Background), blue for High-Risk cells converting to Background (High-Risk to Background), purple for Background cells converting to High-Risk (Background to High-Risk), and red for persistent High-Risk cells (High-Risk to High-Risk). S_T represents the perturbation score.

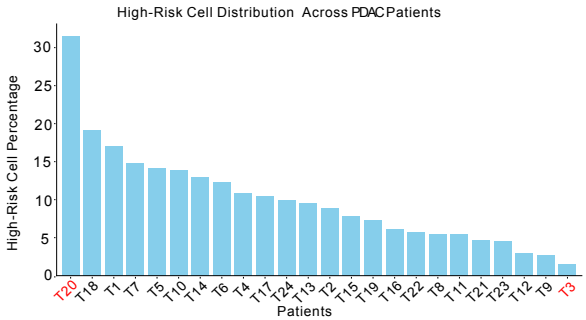

Fig. S11: **Comparison of proportions of High-Risk cells between patients in the PDAC scRNA-seq dataset.** Distribution of the proportion of High-Risk cells across patients ranked in decreasing order. Patient T20 has the highest amount of High-Risk cells among all PDAC patients. Patient T3 has the lowest.

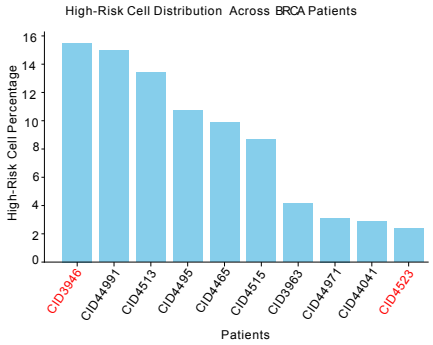

Fig. S12: **Comparison of proportions of High-Risk cells between patients in the BRCA (TNBC subset) scRNA-seq dataset. a,** Distribution of the proportion of High-Risk cells across patients, ranked in decreasing order. Patient CID3946 has the highest amount of High-Risk cells among all BRCA patients, while patient CID4523 has the lowest.

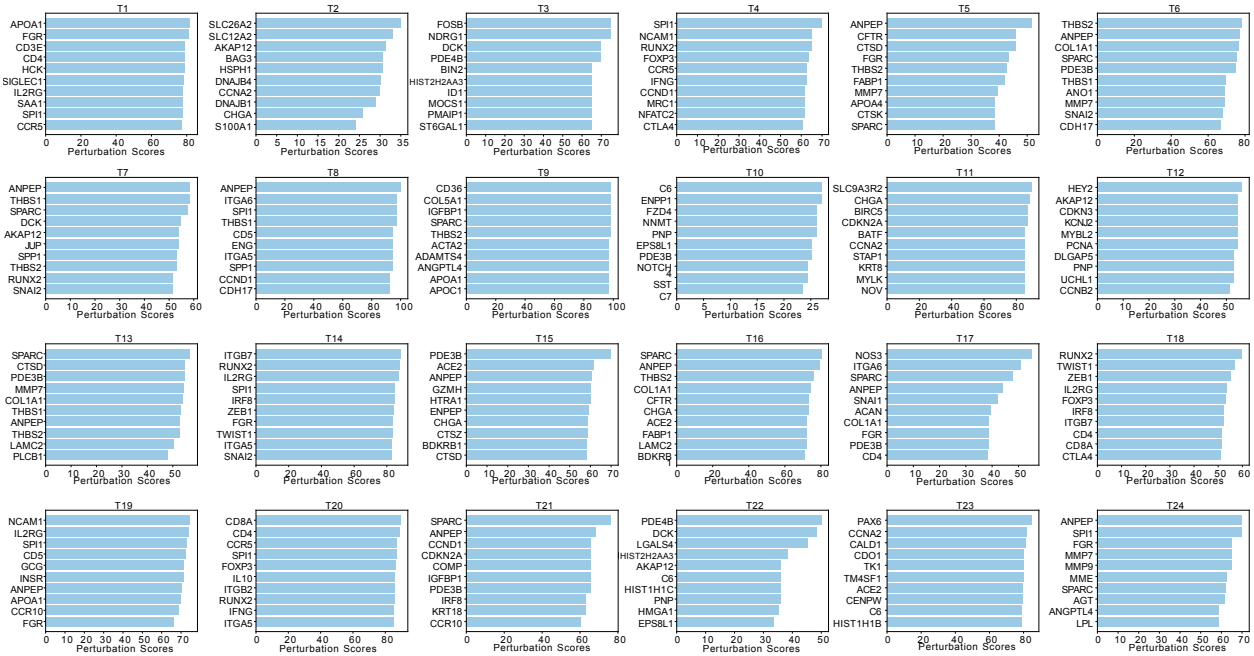

Fig. S13: **Top in silico perturbation targets identified by SIDISH for each patient in the PDAC scRNA-seq dataset.** Bar plots display the top 10 genes with the highest perturbation scores for each patient, highlighting key potential therapeutic candidates.

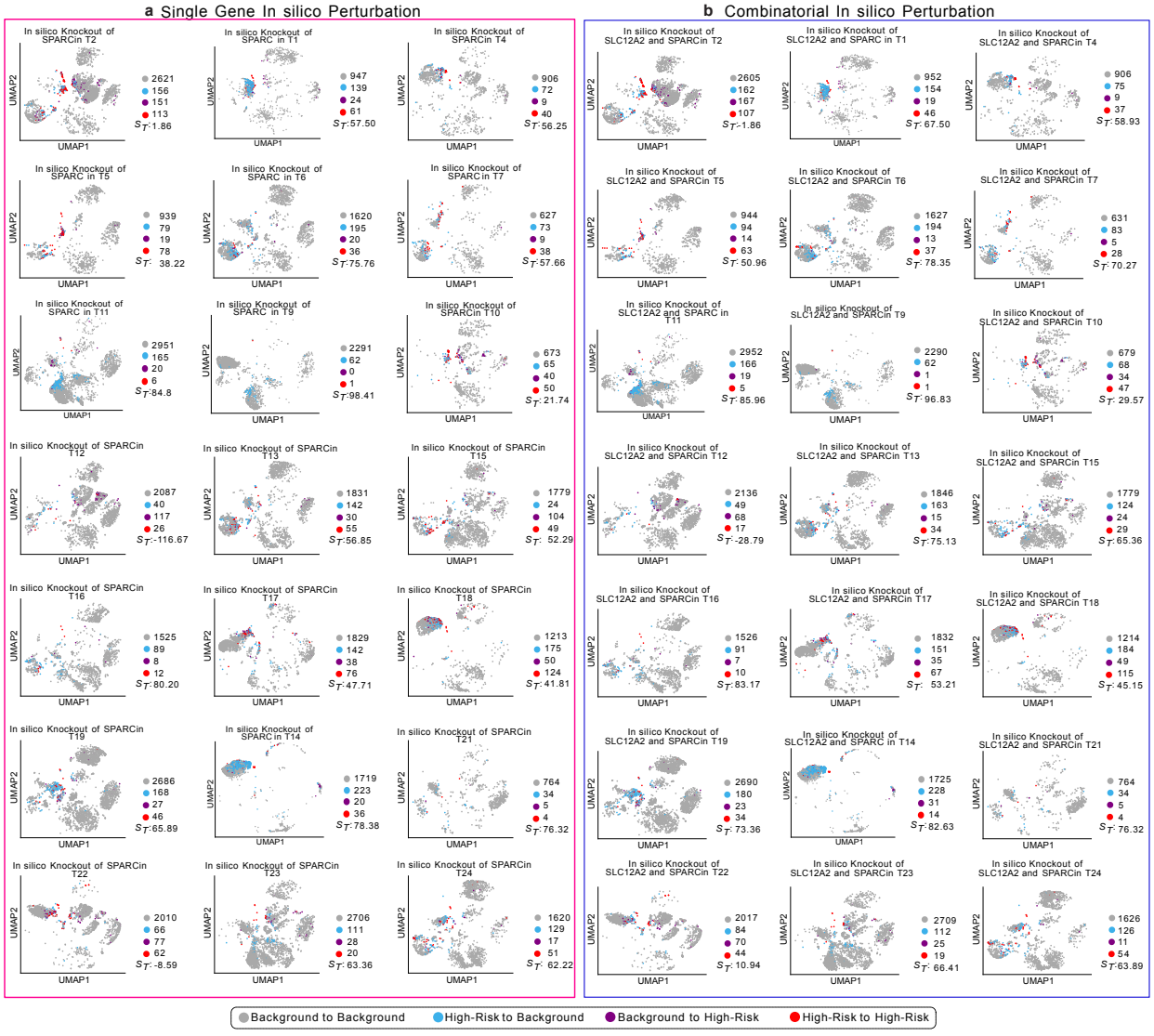

Fig. S14: **UMAP visualization of in silico single-gene knockout of SPARC and combinatorial gene knockout of SPARC and SLC12A2 in PDAC scRNA-seq patients.** The color legend in the UMAPs indicates: blue for High-Risk cells transitioned to Background cells (High-Risk to Background), red for persistent High-Risk cells (High-Risk to High-Risk), purple for Background cells transitioned to High-Risk cells (Background to High-Risk), and gray for unchanged Background cells (Background to Background). S_T represents the perturbation score.

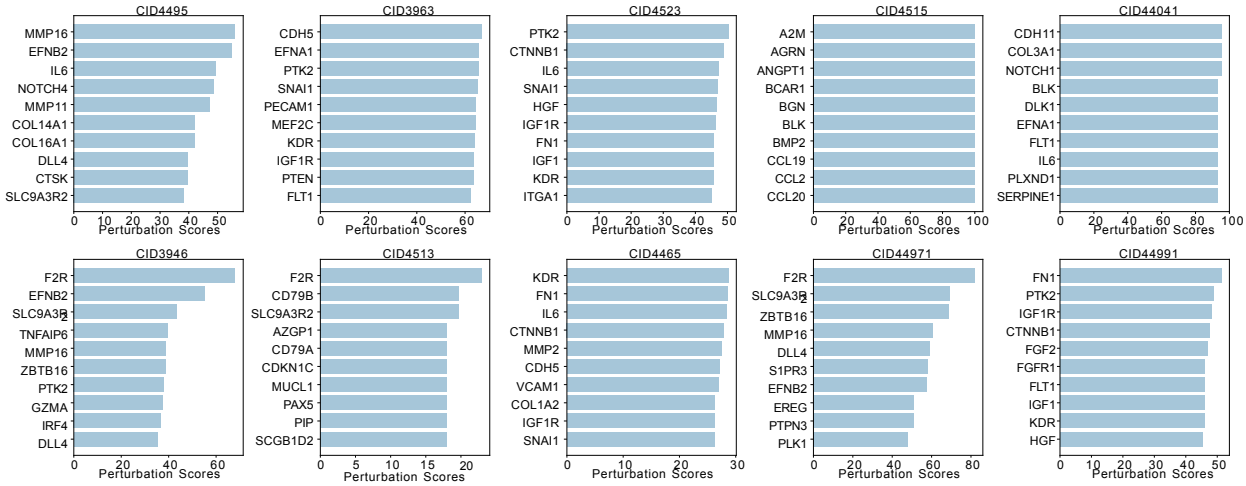

Fig. S15: **Top in silico perturbation targets identified by SIDISH for each patient in the BRCA scRNA-seq dataset.** Bar plots display the top 10 genes with the highest perturbation scores for each patient, highlighting key potential therapeutic candidates.

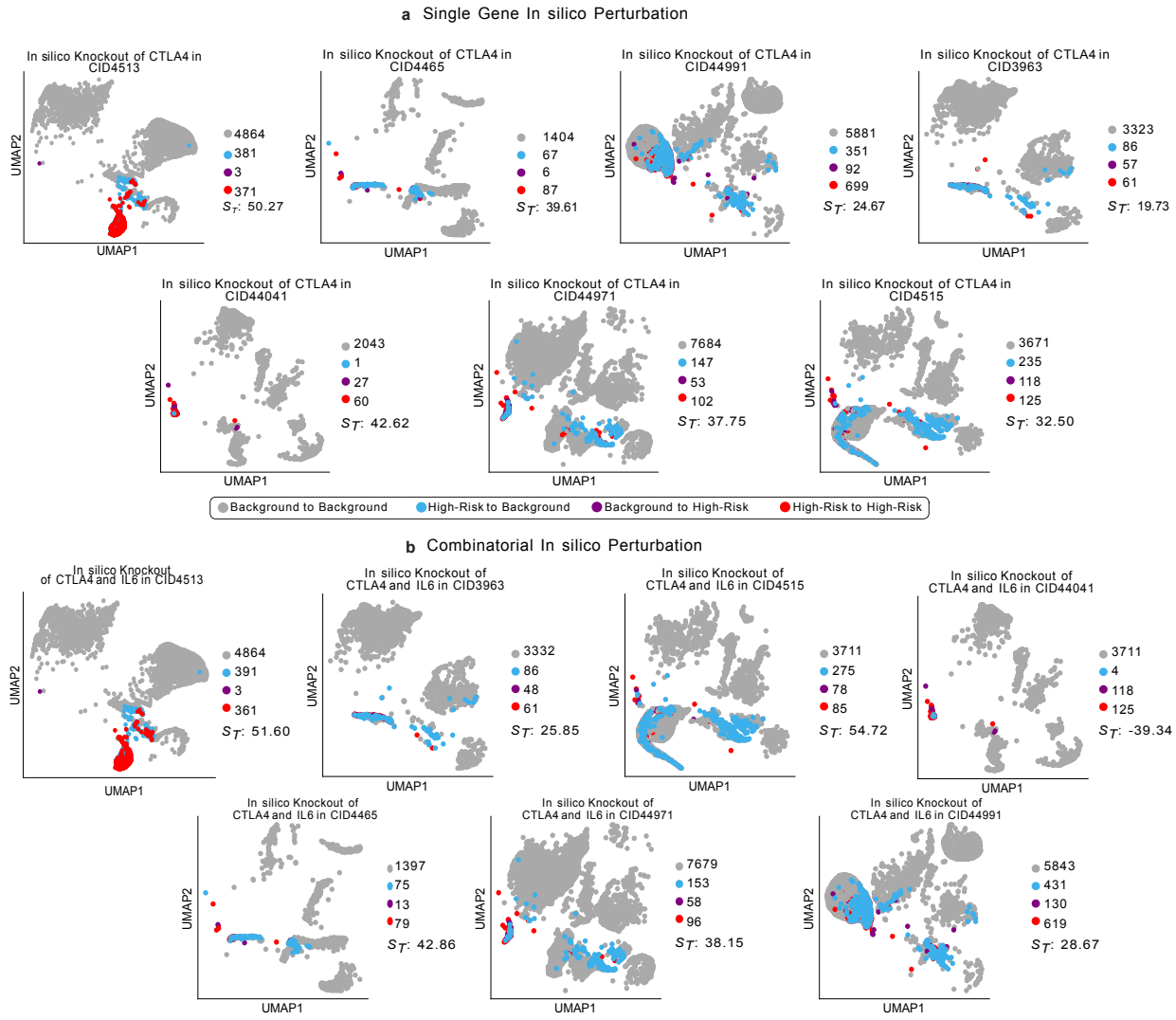

Fig. S16: **UMAP visualization of in silico single-gene knockout of CTLA4 and combinatorial gene knockout of CTLA4 and IL6 in BRCA scRNA-seq patients.** The color legend in the UMAPs indicates: blue for High-Risk cells transitioned to Background cells (High-Risk to Background), red for persistent High-Risk cells (High-Risk to High-Risk), purple for Background cells transitioned to High-Risk cells (Background to High-Risk), and gray for unchanged Background cells (Background to Background). S_T represents the perturbation score.

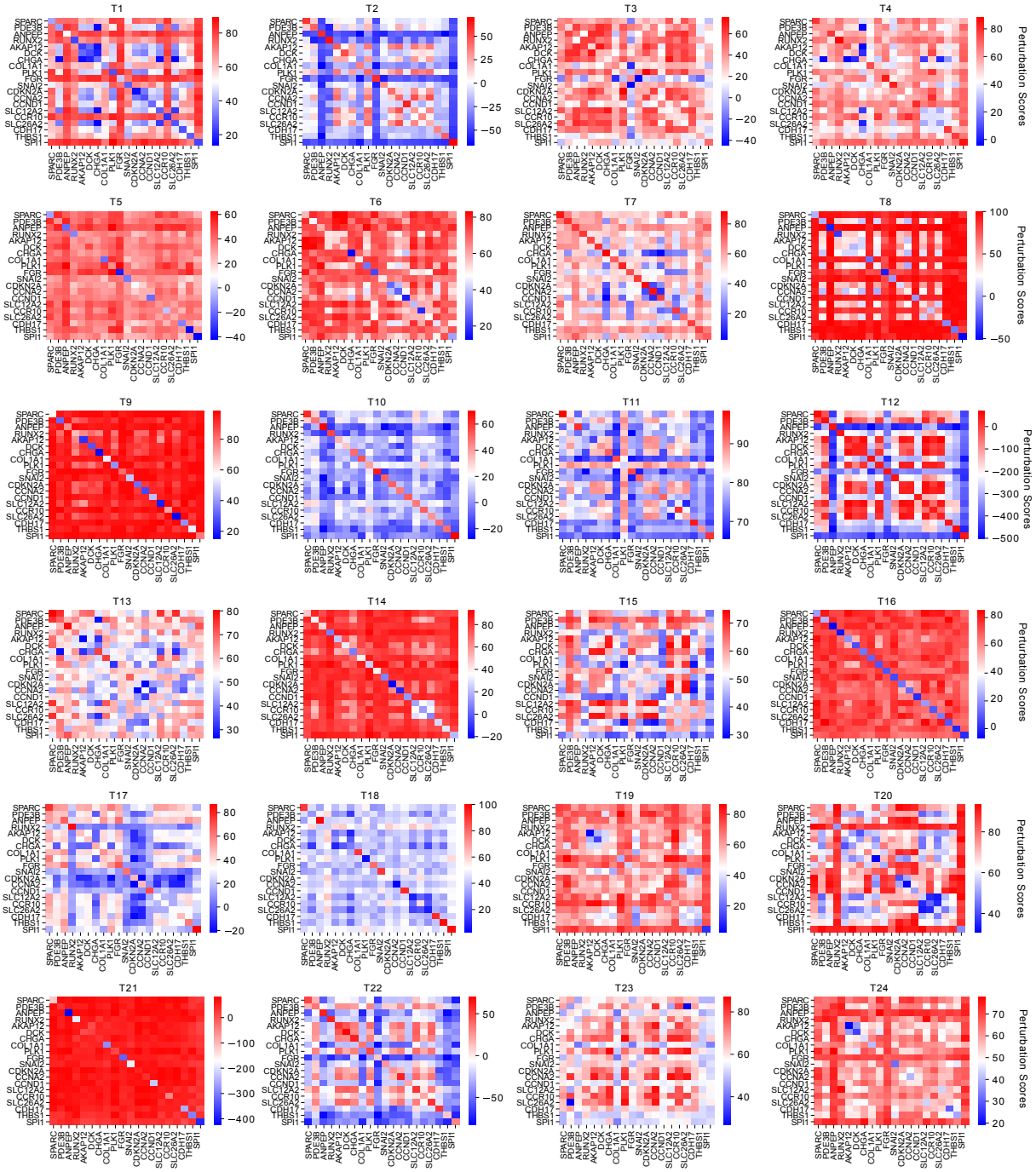

Fig. S17: **Heatmap of in silico combinatorial gene perturbation scores in PDAC scRNA-seq patients.** The heatmap displays perturbation scores for various knockout gene pairs. The same marker genes identified in the perturbation analysis of the entire PDAC scRNA-seq dataset were used for comparison purposes.

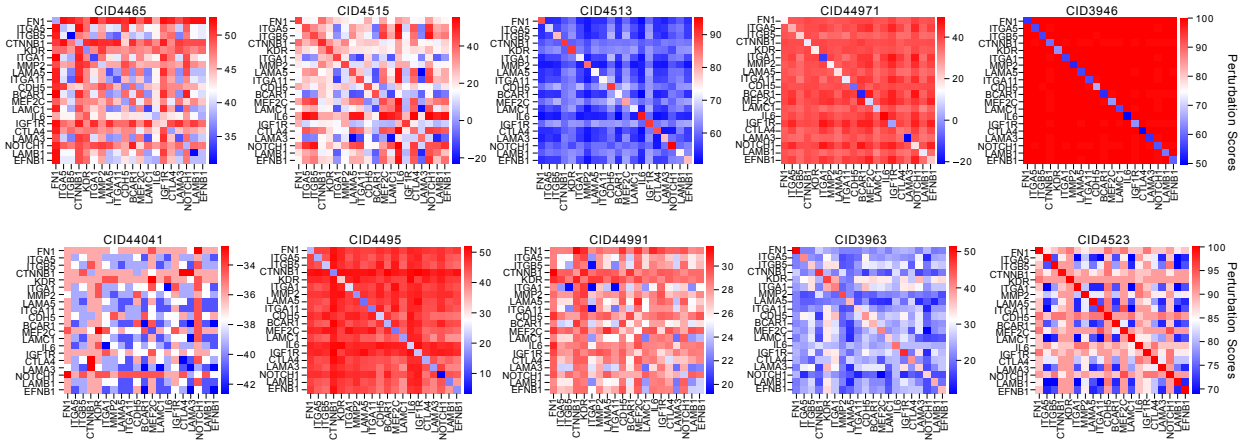

Fig. S18: **Heatmap of in silico combinatorial gene perturbation scores in BRCA scRNA-seq patients.** The heatmap displays perturbation scores for various knockout gene pairs. The same marker genes identified in the perturbation analysis of the entire BRCA scRNA-seq dataset were used for comparison purposes.

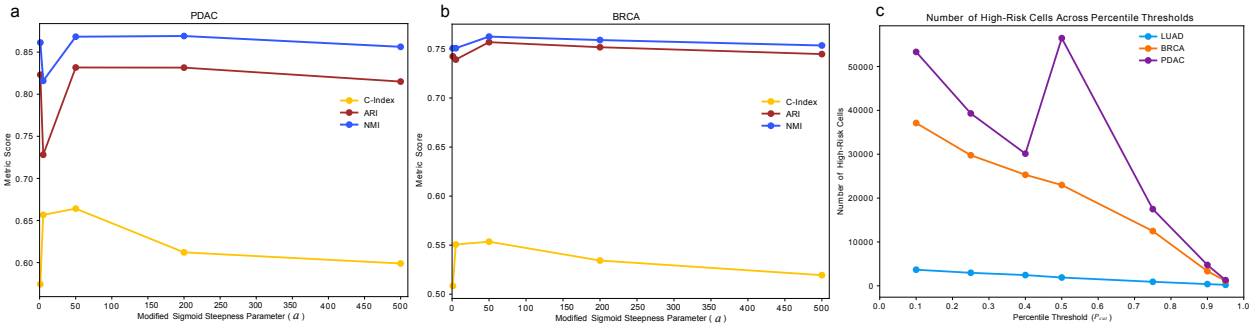

Fig. S19: **Effect of Hyperparameters Sigmoid Steepness (a) and Weibull Distribution Percentile ($P_{\mathbf{cut}}$) on Model Performance and High-Risk Cell Identification. a-b,** Plots showing the relationship between the sigmoid steepness hyperparameter (a) and performance metrics—C-Index (yellow), Adjusted Rand Index (ARI, red), and Normalized Mutual Information (NMI, blue)—highlighting its impact on SIDISH when trained on the PDAC dataset (**a**) and the BRCA dataset (**b**). Variations in a demonstrate its influence on model performance across datasets. **c,** Plot illustrating the effect of the Weibull distribution percentile hyperparameter (P_{cut}) on the number of High-Risk cells identified by SIDISH. As P_{cut} increases, the number of cells classified as High-Risk decreases, indicating a thresholding effect on cell selection.

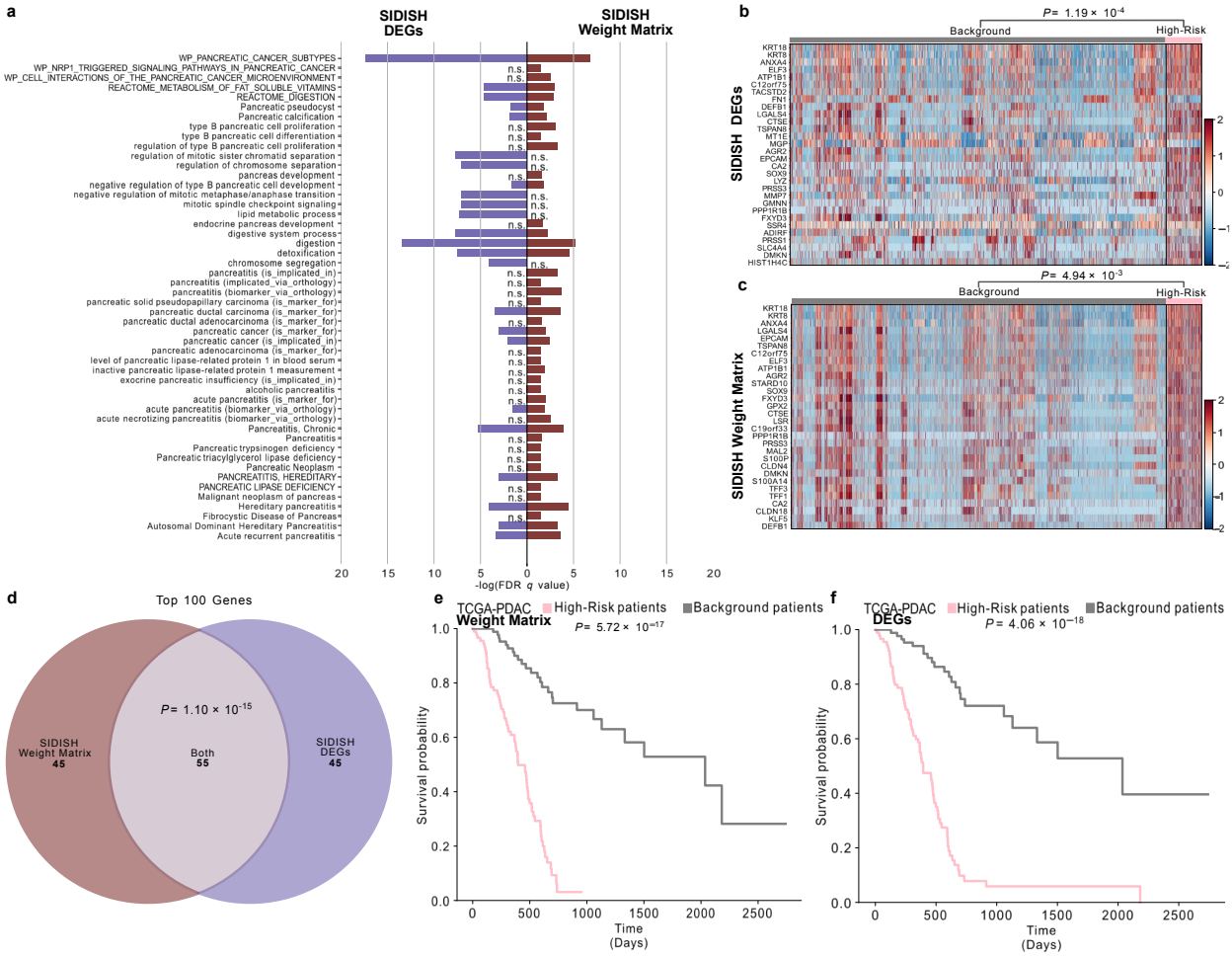

Fig. S20: **Comparison between marker genes derived from SIDISH's weight matrix and differential gene expression (DEG) analysis in the PDAC scRNA-seq dataset. a,** Diverging bar plot comparing the enrichment analysis of marker genes derived from DEG analysis (purple) and SIDISH's weight matrix (burgundy). The scale on the x-axis is the -log(FDR q-value), highlighting the biological relevance of both methods. **b,** Heatmap showing the expression of marker genes derived from DEG analysis in Background and High-Risk cells. A higher concentration of red in one group indicates greater gene expression, while blue indicates lower expression. **c,** Heatmap showing the expression of marker genes derived from SIDISH's weight matrix in Background and High-Risk cells. Color intensity reflects expression levels as in panel **b**. **d,** Venn diagram illustrating the overlap between marker genes identified through DEG analysis in purple and SIDISH's weight matrix in burgundy ($P = 1.10 \times 10^{-15}$). The P value assessing the statistical significance of this overlap was calculated using a hypergeometric test. **e-f,** Kaplan-Meier survival plots based on TCGA-PDAC patient data, stratified by expression levels of marker genes from DEG analysis (**e**) and SIDISH's weight matrix (**f**). A broader separation between survival curves indicates better stratification between patients at a higher risk of death and those at a lower risk. P values were calculated using the log-rank test to compare survival curves between High-Risk and Background patient groups.

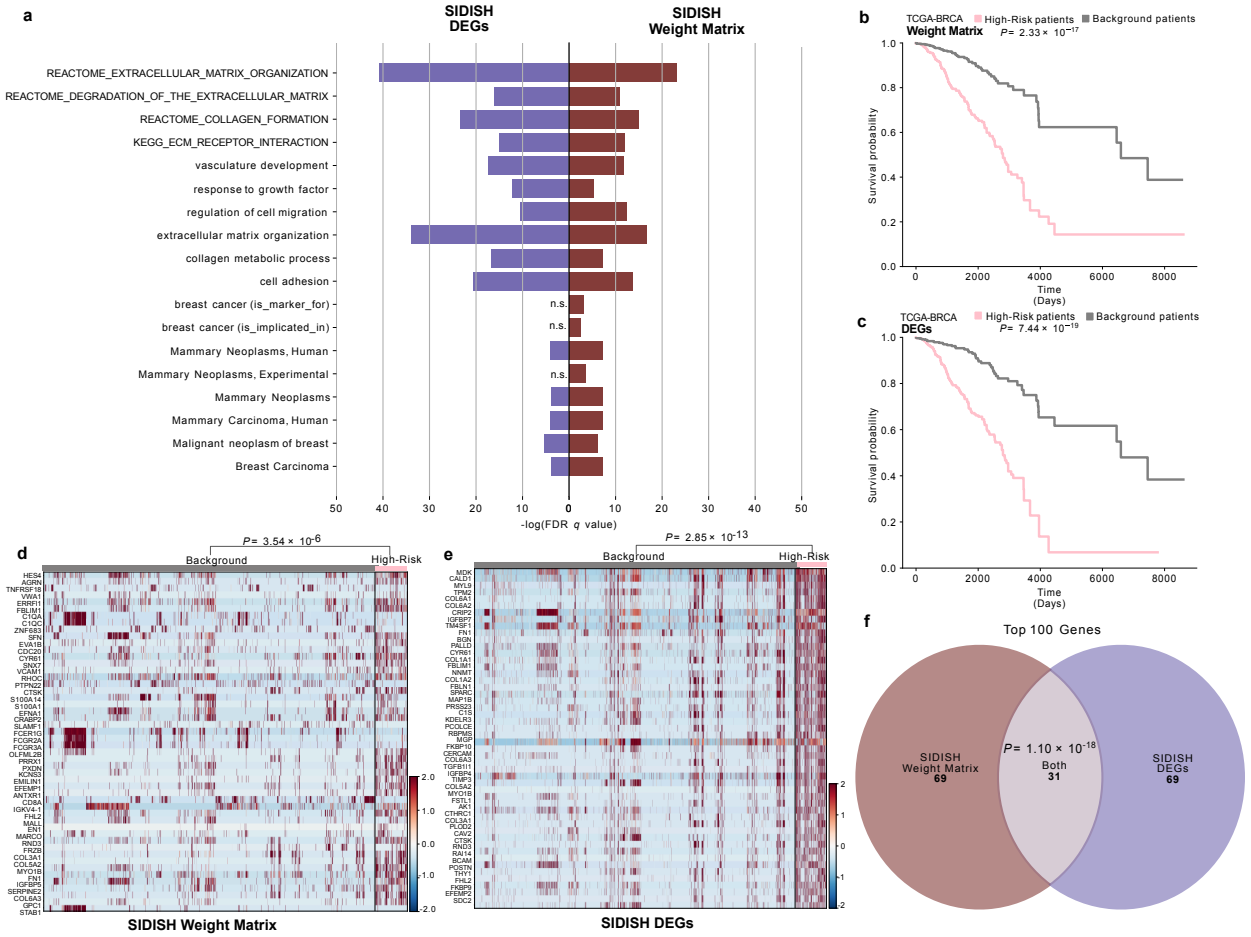

Fig. S21: **Comparison between marker genes derived from SIDISH's weight matrix and differential gene expression (DEG) analysis in the BRCA scRNA-seq dataset. a,** Diverging bar plot comparing the enrichment analysis of marker genes derived from DEG analysis (purple) and those obtained from SIDISH's weight matrix (burgundy). The scale on the x-axis is the -log(FDR q-value), highlighting the biological relevance of both methods. **b-c,** Kaplan-Meier survival plots based on TCGA-BRCA patient data, stratified by expression levels of marker genes from SIDISH's weight matrix (**b**) and DEG analysis (**c**). Broader separation between survival curves indicates better stratification of patients into High-Risk and Background groups. P values were calculated using the log-rank test to assess the significance of survival differences. **d,** Heatmap showing the expression of marker genes derived from SIDISH's weight matrix in Background and High-Risk cells. A higher concentration of red in one group indicates greater gene expression, while blue indicates lower expression. **e,** Heatmap showing the expression of marker genes derived from DEG analysis in Background and High-Risk cells, with color intensity reflecting expression levels as in panel **d**. **f,** Venn diagram illustrating the overlap between marker genes identified through DEG analysis in purple and SIDISH's weight matrix in burgundy ($P = 1.10 \times 10^{-18}$). The P value assessing the statistical significance of this overlap was calculated using a hypergeometric test.

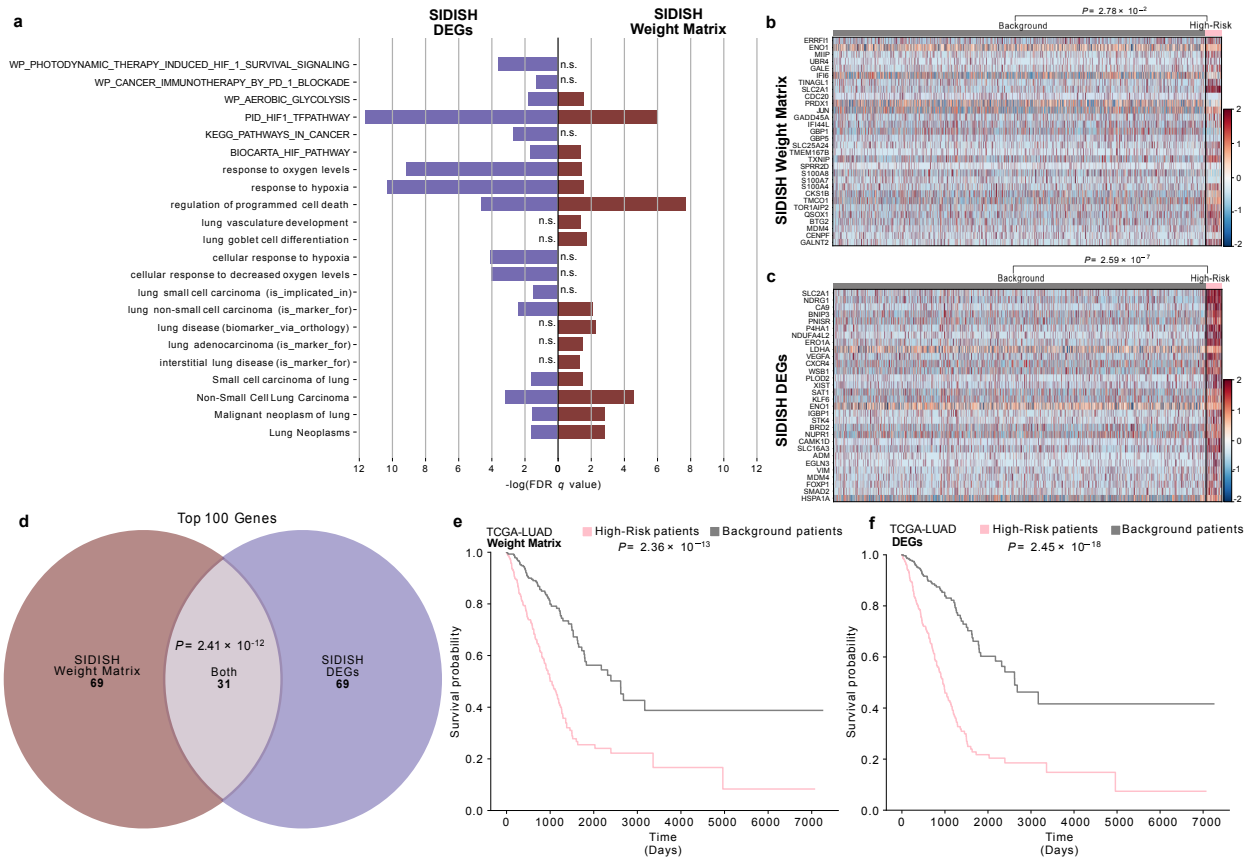

Fig. S22: **Comparison between marker genes derived from SIDISH's weight matrix and differential gene expression (DEG) analysis in the LUAD scRNA-seq dataset. a,** Diverging bar plot comparing the enrichment analysis of marker genes derived from DEG analysis (purple) and those obtained from SIDISH's weight matrix (burgundy). The scale on the x-axis is the -log(FDR q-value), highlighting the biological relevance of both methods. **b-c,** Kaplan-Meier survival plots based on TCGA-LUAD patient data, stratified by expression levels of marker genes from SIDISH's weight matrix (**b**) and DEG analysis (**c**). Broader separation between survival curves indicates better stratification of patients into High-Risk and Background groups. P values were calculated using the log-rank test to assess the significance of survival differences. **d,** Heatmap showing the expression of marker genes derived from SIDISH's weight matrix in Background and High-Risk cells. A higher concentration of red in one group indicates greater gene expression, while blue indicates lower expression. **e,** Heatmap showing the expression of marker genes derived from DEG analysis in Background and High-Risk cells, with color intensity reflecting expression levels as in panel **d**. **f,** Venn diagram illustrating the overlap between marker genes identified through DEG analysis in purple and SIDISH's weight matrix in burgundy ($P = 2.41 \times 10^{-12}$). The P value assessing the statistical significance of this overlap was calculated using a hypergeometric tests.