Supplementary information

HarveST: Heterogeneous Graph Learning Framework for Revealing Spatial Transcriptomics Patterns

Junning Feng[1], Tianwei Yu[2*], and Yanlin Zhang[1*]

[1]*Data Science and Analytics Thrust,Hong Kong University of Science and Technology (Guangzhou)*

[2]*School of Data Science, Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, Guangdong 518172, China*

[*]*Correspondence: yutianwei@cuhk.edu.cn, yanlinzhang@hkust-gz.edu.cn.*

1 Identification and Evaluation of Region-Specific Marker Genes

Accurate identification of region-specific marker genes is essential for elucidating the spatial functional heterogeneity of tissues in spatial transcriptomics studies. To achieve this, we applied the HarveST method alongside a baseline approach, ranking genes according to their specificity scores to systematically assess the performance and reliability of different methods in detecting spatially distinctive genes. A comparative analysis was conducted from two perspectives to provide a comprehensive evaluation.

First, the top three ranked genes were considered the most region-specific markers, as they exhibited the highest degree of spatial specificity within a given region. Their spatial distribution was visualized to assess the ability of each method to capture genes with strong regional specificity. Genes exhibiting highly localized expression, with minimal or no expression in other regions, are expected to be the most biologically relevant markers, and their identification serves as an indicator of the method's effectiveness in detecting high-confidence region-specific genes.

To further evaluate the stability and specificity of our approach, we extended the analysis to the top 20 ranked genes in each region. The degree of overlap between the genes identified by the HarveST method and those detected by the baseline approach was assessed, where the shared markers represented genes consistently identified by both methods, reinforcing the reliability of our approach in detecting well-established region-specific genes. Additionally, we examined the unique markers, referring to genes ranked among the top 20 but exclusively identified by our method.

To validate the biological significance of these uniquely identified genes, we conducted a literature-based evaluation and found that a substantial proportion had been previously reported as experimentally validated region-specific markers, despite being undetected by the baseline approach. This finding highlights the robustness and sensitivity of our method in capturing biologically meaningful spatial markers that may be overlooked by conventional approaches. Furthermore, the remaining uniquely identified genes, which have not yet been documented in existing literature, represent potential novel region-specific markers. These candidates warrant further experimental validation to confirm their spatial specificity and biological relevance, offering valuable insights for biomarker discovery, disease mechanism studies, and tissue function characterization.

Overall, this study not only corroborates previously identified region-specific genes but also expands the repertoire of potential spatial markers, demonstrating the superior detection capability, stability, and biological credibility of the HarveST method in spatial transcriptomics analysis.

2 Processing and Interpretation of Enrichment Analysis Results

This study employed the HarveST method to systematically analyze the enrichment characteristics of genes across distinct spatial regions, aiming to identify region-specific genes and elucidate their biological significance.

To quantify the degree of spatial specificity, a specificity score (score) was assigned to each gene within a given region. This score reflects the relative importance of a gene in that particular region, with higher values indicating stronger biological relevance. In addition to the specificity score, two distinct types of fold change (FC) metrics were computed to further

characterize the spatial variation of gene activity. The network fold change measures the relative enrichment of a gene's specificity score in a given region compared to all other regions, thereby capturing the extent to which a gene is regionally specialized. Meanwhile, the expression fold change quantifies the relative expression level of a gene in one region compared to its expression in other regions, providing insights into spatially regulated transcriptional differences.

To investigate the functional implications of region-specific genes, genes were ranked based on their specificity scores, and the top 100 genes from each region were selected for enrichment analysis. The Metascape platform was employed to perform functional annotation, leveraging multiple databases—including Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome—to systematically identify enriched biological processes and pathways associated with the selected genes.

Pathway enrichment was further examined through KEGG analysis, which enabled the visualization of overrepresented metabolic and signaling pathways. To elucidate the relationship between enriched pathways and the genes contributing to their enrichment, a Sankey diagram was generated. This was based on the "QC_GO_AllLists" dataset from the Metascape results, where pathways were ranked by logP value in ascending order, ensuring that the most statistically significant pathways were prioritized. The top 30 pathways, along with their corresponding enriched genes, were selected for visualization. This representation facilitates an intuitive understanding of gene-pathway associations and highlights potential functional interactions among region-specific genes.

By implementing this analytical framework, the study systematically identifies spatially enriched genes and delineates their involvement in key biological processes, signaling cascades, and disease-associated pathways. These findings contribute to a deeper understanding of regional transcriptional dynamics and provide a foundation for future investigations into regulatory mechanisms and biomarker discovery. Further integration with multi-omics data, such as transcriptomics, proteomics, and metabolomics, as well as network-based approaches, could refine these insights and uncover additional layers of biological complexity.

Table 1: Top 20 domain-specific marker genes of Pancreatic region identified by HarveST

Rank	Gene	Score	Scaled Score	p-value	Network FC	Expression FC
1	AC009078.2	0.00106	1.000	0.002	2.316	0.780
2	CTRC	0.00093	0.871	0.002	1.537	0.000
3	PNLIPRP2	0.00089	0.831	0.002	1.581	0.000
4	PNLIPRP1	0.00088	0.818	0.002	1.775	8107.872
5	PRSS3	0.00088	0.812	0.002	1.452	0.000
6	REG3G	0.00087	0.808	0.002	1.458	0.000
7	GP2	0.00087	0.805	0.002	1.285	0.000
8	PRSS3P1	0.00086	0.799	0.002	1.400	0.000
9	CPA2	0.00086	0.799	0.002	1.394	0.000
10	CEL	0.00086	0.795	0.002	1.260	4383.778
11	PRSS1	0.00084	0.781	0.002	1.281	1.023
12	CELA3B	0.00084	0.779	0.002	1.266	2.805
13	MUC6	0.00084	0.776	0.002	1.579	0.000
14	PNLIP	0.00083	0.763	0.002	1.605	0.000
15	CTRB1	0.00083	0.762	0.002	1.154	35601.240
16	AL034369.1	0.00082	0.758	0.002	2.257	0.000
17	CPB1	0.00082	0.755	0.002	1.189	0.000
18	CELP	0.00080	0.741	0.002	1.866	0.000
19	CELA3A	0.00080	0.737	0.002	1.181	1.068
20	CPA1	0.00079	0.731	0.002	1.223	1.833

Note:FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 2: Top 20 domain-specific marker genes of Pancreatic region identified by DESpace

Rank	Gene	LR	logCPM	FDR	Pancreatic FDR
1	KRT17	557.281	14.156	1.34E-118	2.39E-37
2	S100A6	479.973	15.916	3.81E-102	4.18E-41
3	KRT19	464.301	15.505	6.32E-99	1.20E-36
4	CTRB1	322.256	15.002	1.84E-68	6.11E-51
5	CTRB2	320.719	15.565	3.40E-68	1.55E-50
6	CEL	305.169	14.357	6.91E-65	3.09E-44
7	PRSS2	280.826	15.127	1.14E-59	8.72E-36
8	PRSS1	257.547	14.034	1.11E-54	2.96E-39
9	GP2	247.421	13.975	1.44E-52	5.06E-43
10	CPB1	240.443	14.265	4.29E-51	3.97E-46
11	CLPS	223.825	14.318	1.56E-47	1.67E-37
12	AC009078.2	219.799	12.854	1.08E-46	2.65E-45
13	CELA3A	218.793	13.986	1.67E-46	1.50E-37
14	REG3A	216.357	15.693	5.29E-46	3.96E-42
15	CPA2	213.002	13.549	2.65E-45	2.82E-33
16	CTRC	198.868	13.498	2.58E-42	7.15E-35
17	PNLIPRP1	189.853	12.962	1.92E-40	6.02E-35
18	PNLIP	188.256	12.992	4.08E-40	1.43E-36
19	REG3G	179.804	13.243	2.45E-38	1.94E-38
20	PNLIPRP2	177.484	13.184	7.50E-38	1.45E-36

Note: LR = Likelihood Ratio; logCPM = log Counts Per Million; FDR = False Discovery Rate. Genes are sorted by LR statistic in descending order. The Pancreatic FDR column shows the region-specific significance values.

Table 3: Top 20 domain-specific marker genes of Pancreatic region identified by Scanpy

Rank	**Gene**	**logFC**	**Adj. P-value**	**Score**
1	CTRB1	2.51	1.29×10^{-16}	9.41
2	CTRB2	1.72	4.19×10^{-14}	8.70
3	CEL	2.64	3.67×10^{-13}	8.37
4	REG3A	1.50	3.67×10^{-13}	8.39
5	PRSS2	2.01	3.23×10^{-12}	8.08
6	AC009078.2	3.64	4.91×10^{-12}	8.01
7	GP2	2.23	6.79×10^{-12}	7.95
8	CPB1	2.15	1.49×10^{-11}	7.84
9	PRSS1	2.34	3.23×10^{-10}	7.41
10	CTRC	2.60	2.31×10^{-9}	7.13
11	CELA3B	2.12	1.03×10^{-8}	6.91
12	CELA3A	1.95	1.76×10^{-8}	6.82
13	CLPS	1.79	2.82×10^{-8}	6.73
14	REG3G	2.18	6.03×10^{-8}	6.61
15	PNLIPRP2	2.49	9.42×10^{-8}	6.54
16	PRSS3P1	2.42	1.47×10^{-7}	6.46
17	PRSS3	2.34	3.98×10^{-7}	6.29
18	CPA2	2.31	3.98×10^{-7}	6.29
19	CUZD1	3.35	6.49×10^{-7}	6.21
20	PNLIPRP1	2.65	1.02×10^{-6}	6.13

Note: logFC = log fold change; Adj. P-value = Benjamini-Hochberg adjusted P-value; Score = Differential expression score. All genes were specifically associated with the Pancreatic region (Ground Truth = Pancreatic).

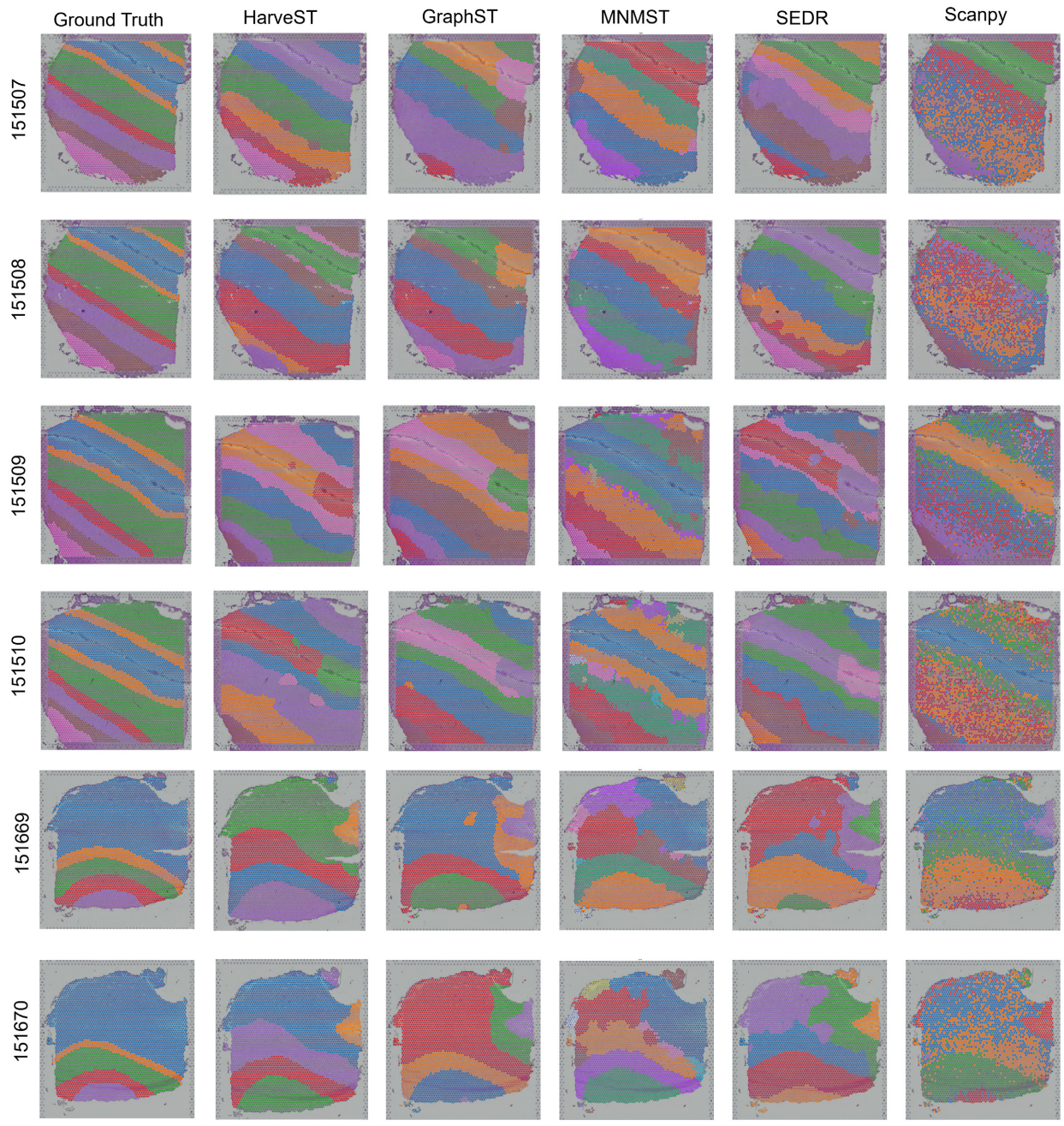

Figure S1: Comparative results of domain identification in dorsolateral prefrontal cortex (DLPFC) tissue sections (Visualization 1).

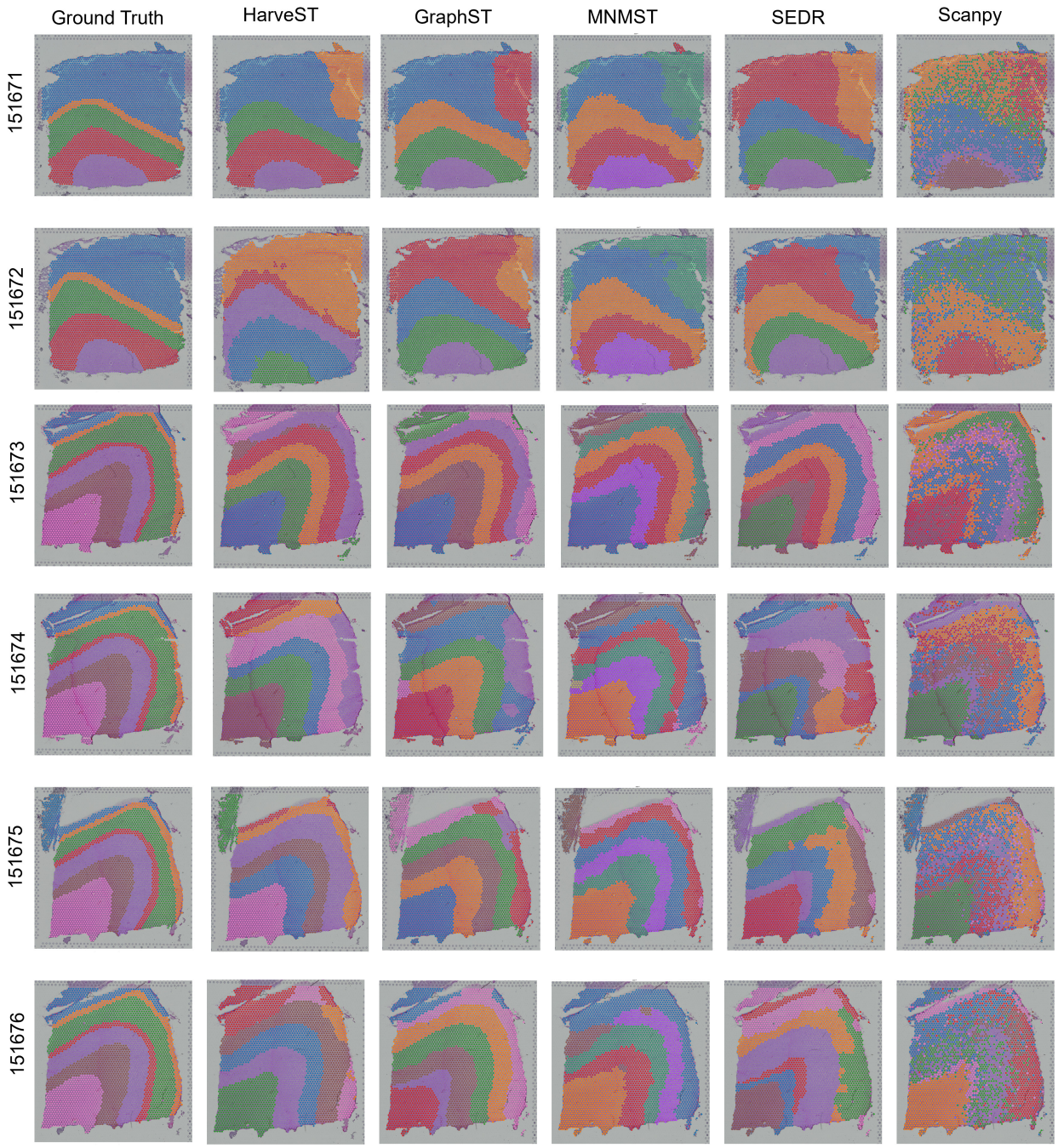

Figure S2: Comparative results of domain identification in dorsolateral prefrontal cortex (DLPFC) tissue sections (Visualization 2).

Figure S3: Comparison of tissue region identification performance among various spatial transcriptomics analysis methods on mouse olfactory bulb (MOB) Stereo-seq dataset.

Table 4: Top 15 domain-specific marker genes of Pancreatic region identified by SpaGCN

Rank	**Gene**	**In%**	**Out%**	**I/O ratio**	**In exp**	**FC**	**adj.p**
1	KRT19	0.928	0.884	1.05	3.633	2.95	0.014
2	CTRB1	0.985	0.891	1.106	4.39	3.147	0.027

Note: In% = in-group fraction; Out% = out-group fraction; I/O ratio = in/out group ratio; In exp = in-group mean expression; FC = fold change; adj.p = adjusted p-value. Genes are sorted by adjusted p-value in ascending order.

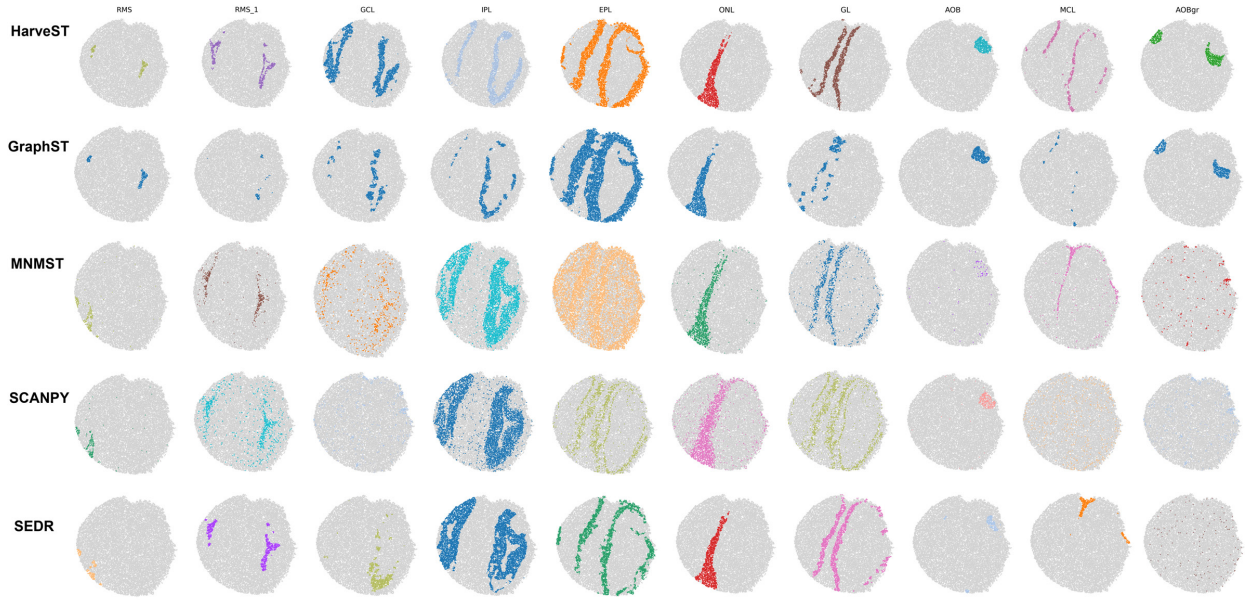

Figure S4: Comparison of tissue region identification performance among various spatial transcriptomics analysis methods on mouse olfactory bulb (MOB) Slide-seq dataset. The figure illustrates the identification results of ten major anatomical regions in mouse olfactory bulb using five different analysis methods: HarveST, GraphST, MNMST, SCANPY, and SEDR. The identified regions include two rostral migratory stream areas (RMS and RMS_1), granule cell layer (GCL), internal plexiform layer (IPL), external plexiform layer (EPL), olfactory nerve layer (ONL), glomerular layer (GL), accessory olfactory bulb (AOB), mitral cell layer (MCL), and accessory olfactory bulb granular layer (AOBgr). Each method's identification results are displayed in different colors on corresponding tissue sections. Our HarveST method (first row) successfully identifies all ten anatomical regions, each marked with a specific color, demonstrating excellent identification specificity and completeness. In contrast, other methods show significant limitations in region identification: GraphST primarily uses blue tones to identify partial regions, making distinction between different structures difficult; MNMST identifies some regions but with evident confusion and incomplete recognition; SCANPY recognizes only some major regions with complete absence in others; SEDR performs well in identifying certain regions (such as IPL, EPL, ONL) but shows incomplete identification in others. This comparison clearly demonstrates the superiority of the HarveST method on Slide-seq data, which presents higher technical challenges among spatial transcriptomics technologies. HarveST comprehensively and accurately delineates the complex anatomical structure of the mouse olfactory bulb, providing a powerful tool for precise analysis of high-resolution spatial transcriptomics data and highlighting our method's applicability and robustness across different types of spatial transcriptomics technologies.

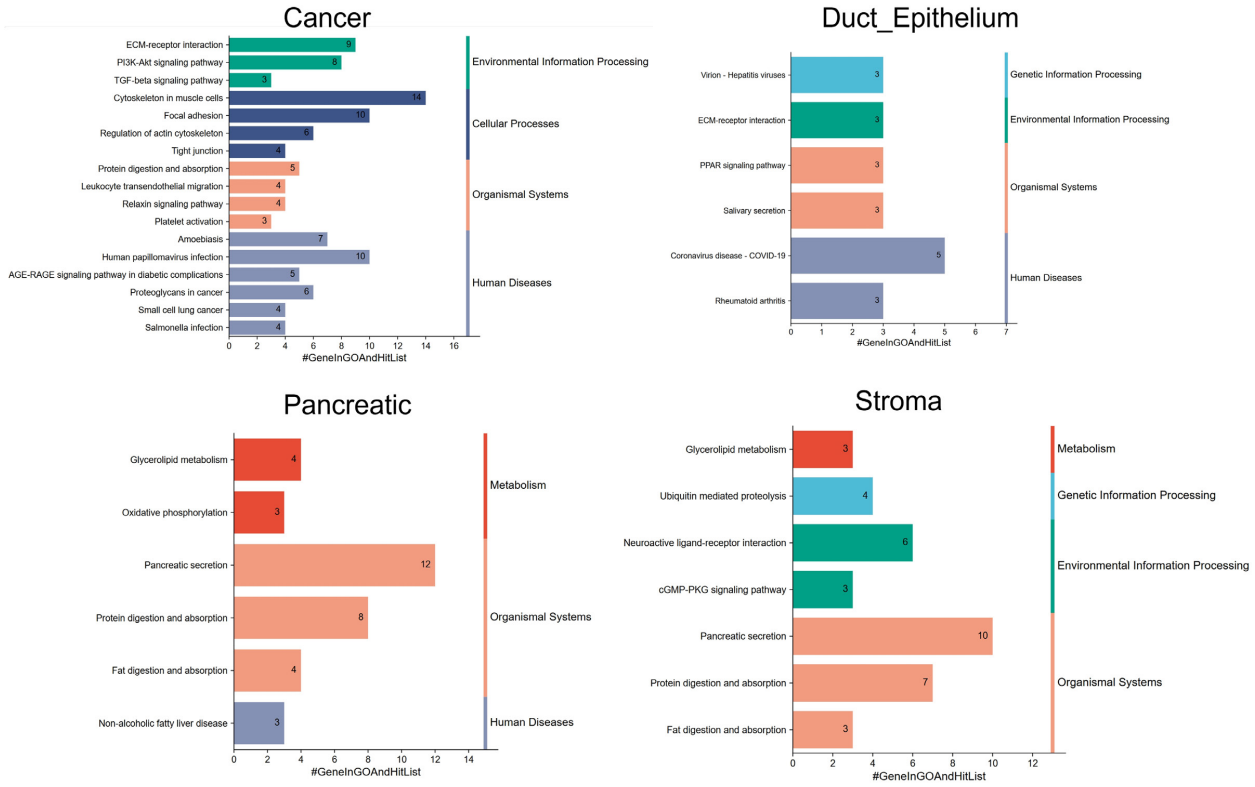

Figure S5: KEGG pathway enrichment analysis of region-specific genes identified by HarveST across different tissue compartments in pancreatic ductal adenocarcinoma (PDAC) spatial transcriptomics dataset. The figure illustrates KEGG pathway enrichment analysis of region-specific genes identified by our HarveST method in four major compartments of PDAC: Cancer, Ductal Epithelium, Pancreatic tissue, and Stroma. Genes specific to the Cancer region (top left) were significantly enriched in pathways related to ECM-receptor interaction, PI3K-Akt signaling pathway, and cytoskeleton-related processes, as well as human papillomavirus infection. The Ductal Epithelium region (top right) showed enrichment in virion-hepatitis viruses, ECM-receptor interaction, and PPAR signaling pathways. Pancreatic tissue-specific genes (bottom left) predominantly enriched in pancreatic secretion, protein digestion and absorption, and glycerolipid metabolism pathways. The Stroma region (bottom right) exhibited enrichment in pancreatic secretion, neuroactive ligand-receptor interaction, and ubiquitin-mediated proteolysis pathways.

Table 5: Top 20 domain-specific marker genes of Cancer region identified by HarveST

Rank	**Gene**	**Score**	**Scaled Score**	**p-value**	**Network FC**	**Expression FC**
1	LAMC2	0.00103	1.000	0.002	1.782	0.255
2	SFN	0.00099	0.951	0.002	1.707	0.000
3	KRT17	0.00098	0.944	0.002	1.502	12742.929
4	APOL1	0.00096	0.929	0.002	1.621	0.000
5	PLEC	0.00096	0.919	0.002	1.487	0.000
6	TRIM29	0.00094	0.903	0.002	1.885	0.614
7	TM4SF1	0.00093	0.898	0.002	1.432	0.000
8	HMGA1	0.00092	0.879	0.002	1.551	0.000
9	LAMB3	0.00091	0.878	0.002	1.786	0.000
10	CHPF	0.00089	0.850	0.002	1.554	0.000
11	S100A14	0.00088	0.847	0.002	1.328	0.000
12	ITGB4	0.00088	0.842	0.002	1.545	0.000
13	C19orf33	0.00087	0.834	0.002	1.488	0.388
14	FXYD3	0.00087	0.833	0.002	1.554	0.719
15	IL1RN	0.00086	0.824	0.002	1.779	0.000
16	AEBP1	0.00085	0.812	0.002	1.304	0.000
17	IFI27	0.00085	0.807	0.002	1.492	0.000
18	TAGLN	0.00085	0.807	0.002	1.319	0.000
19	PKM	0.00085	0.807	0.002	1.324	1.105
20	SLC16A3	0.00084	0.800	0.002	1.452	0.000

Note: FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 6: Top 20 domain-specific marker genes of Cancer region identified by DESpace

Rank	Gene	LR	logCPM	FDR	Cancer FDR
1	KRT17	557.281	14.156	1.34E-118	1.63E-116
2	S100A6	479.973	15.916	3.81E-102	7.95E-89
3	KRT19	464.301	15.505	6.32E-99	7.95E-89
4	MUC5B	430.868	14.785	8.31E-92	2.53E-47
5	GSTP1	328.460	14.149	1.00E-69	1.07E-66
6	CTRB1	322.256	15.002	1.84E-68	2.54E-23
7	CEL	305.169	14.357	6.91E-65	2.28E-21
8	PRSS1	257.547	14.034	1.11E-54	1.03E-21
9	GAPDH	251.421	14.581	2.14E-53	1.20E-44
10	REG3A	216.357	15.693	5.29E-46	1.36E-21
11	SPP1	200.969	13.478	9.51E-43	2.16E-26
12	REG1A	161.322	14.442	2.10E-34	3.29E-22
13	IFI6	163.266	14.188	8.25E-35	9.80E-23
14	ACTB	161.190	15.028	2.18E-34	4.10E-33
15	EEF2	132.816	14.064	2.79E-28	1.14E-29
16	PFN1	127.231	14.415	4.22E-27	3.95E-24
17	COL1A1	117.747	13.979	4.53E-25	2.57E-21
18	FTH1	104.047	14.338	3.92E-22	9.38E-23
19	TMSB10	103.931	14.071	4.05E-22	8.36E-23
20	C3	97.158	13.859	1.08E-20	2.18E-21

Note: LR = Likelihood Ratio; logCPM = log Counts Per Million; FDR = False Discovery Rate. Genes are sorted by LR statistic in descending order. The Cancer FDR column shows the region-specific significance values.

Table 7: Top 20 domain-specific marker genes of Cancer region identified by Scanpy

Rank	**Gene**	**$\log_2$FC**	**adj. p-value**	**Score**
1	KRT19	1.972	1.57E-35	13.227
2	S100A6	1.571	5.23E-30	12.178
3	KRT17	3.233	1.64E-25	11.267
4	TM4SF1	2.581	2.40E-22	10.579
5	PLEC	2.488	6.32E-22	10.467
6	LAMC2	3.198	4.55E-20	10.033
7	GSTP1	1.802	9.04E-20	9.939
8	SFN	2.956	1.16E-17	9.431
9	S100A14	2.098	5.01E-17	9.265
10	ACTB	1.106	1.30E-16	9.147
11	S100A11	1.873	1.30E-16	9.143
12	APOL1	2.667	3.19E-16	9.037
13	SYNGR2	1.799	1.86E-15	8.826
14	KRT18	1.558	4.68E-14	8.450
15	TRIM29	3.107	8.97E-14	8.367
16	HMGA1	2.360	1.45E-13	8.303
17	RPL28	0.837	1.35E-12	8.027
18	LAMB3	2.872	5.08E-12	7.857
19	ITGB4	2.324	4.74E-11	7.559
20	GAPDH	0.984	6.71E-11	7.508

Note: $\log_2$FC = $\log_2$ fold change; adj. p-value = adjusted p-value. Genes are sorted by Score in descending order. All genes are confirmed markers of Cancer region according to ground truth annotation.

Table 8: Top 15 domain-specific marker genes of Cancer region identified by SpaGCN

Rank	**Gene**	**In%**	**Out%**	**I/O ratio**	**In exp**	**FC**	**adj.p**
1	KRT19	1.000	0.918	1.089	4.490	3.685	6.59E-14
2	S100A6	0.985	0.966	1.020	4.624	2.706	4.30E-09
3	PLEC	0.876	0.463	1.894	2.549	4.314	6.85E-08
4	KRT17	0.861	0.517	1.666	3.149	6.177	6.85E-08
5	TM4SF1	0.861	0.517	1.666	2.780	4.563	3.74E-07
6	S100A11	0.891	0.558	1.596	2.739	3.644	4.19E-07
7	GSTP1	0.898	0.680	1.320	3.023	3.190	6.93E-06
8	S100A14	0.876	0.605	1.447	2.752	3.391	6.02E-05
9	ACTB	0.956	0.864	1.107	3.707	2.249	1.08E-04
10	SYNGR2	0.847	0.605	1.399	2.640	3.017	1.61E-04
11	KRT18	0.869	0.633	1.373	2.686	2.869	1.67E-04
12	IFI6	0.891	0.619	1.439	2.880	3.163	3.12E-04
13	GAPDH	0.920	0.823	1.117	3.288	2.089	6.44E-04
14	RPL28	0.920	0.891	1.032	3.385	1.770	9.22E-04
15	PFN1	0.920	0.748	1.229	3.085	2.171	1.32E-02

Note: In% = in-group fraction; Out% = out-group fraction; I/O ratio = in/out group ratio; In exp = in-group mean expression; FC = fold change; adj.p = adjusted p-value. Genes are sorted by adjusted p-value in ascending order.

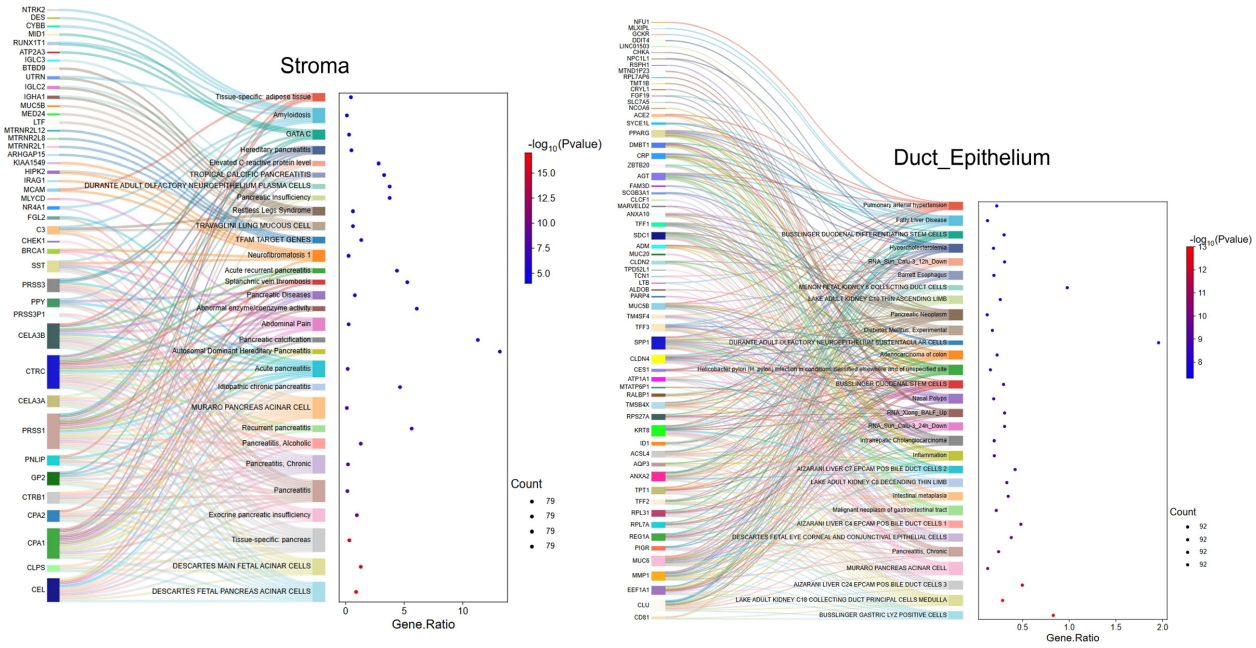

Figure S6: Sankey diagram analysis of disease/cell type associations of region-specific genes identified by HarveST in stroma and ductal epithelium regions of pancreatic ductal adenocarcinoma (PDAC) spatial transcriptomics dataset. The figure presents Sankey diagrams illustrating the associations between region-specific genes identified by our HarveST method in the stroma region (left) and ductal epithelium region (right) of PDAC and various diseases/cell types. The left side of each diagram displays region-specific genes, while the right side shows enriched diseases or cell types, with connection width indicating association strength and dot color intensity representing enrichment significance (-log10(P-value)). Genes from the stroma region predominantly enriched in tissue-specific adipose tissue, amyloidosis, hereditary pancreatitis, and pancreatic acinar cells, which are associated with pancreatic inflammation and fibrosis. In contrast, ductal epithelium-specific genes significantly enriched in pulmonary arterial hypertension, fatty liver disease, pancreatic neoplasm, and various epithelial cell types including duodenal differentiating stem cells, breast ductal stem cells, and liver bile duct cells.

Table 9: Top 20 domain-specific marker genes of Stroma region identified by HarveST

Rank	Gene	Score	Scaled Score	p-value	Network FC	Expression FC
1	MUC5B	0.00081	1.000	0.002	1.158	0.000
2	HMGB3P17	0.00078	0.971	0.002	1.283	0.000
3	CLPS	0.00078	0.969	0.002	1.122	1.401
4	PRSS1	0.00078	0.963	0.002	1.180	0.910
5	CPA1	0.00078	0.962	0.002	1.199	1.126
6	CTRB1	0.00077	0.953	0.002	1.076	0.000
7	GP2	0.00076	0.944	0.002	1.130	2017.943
8	MTRNR2L12	0.00076	0.941	0.002	1.159	1.313
9	C3	0.00076	0.939	0.002	1.133	6.972
10	PPY	0.00076	0.937	0.002	1.126	1.750
11	CEL	0.00076	0.934	0.002	1.106	0.000
12	CELA3A	0.00076	0.933	0.002	1.110	0.827
13	MTRNR2L1	0.00075	0.930	0.002	1.141	0.573
14	AD000090.1	0.00075	0.922	0.002	1.118	0.000
15	MTRNR2L8	0.00074	0.913	0.002	1.202	1.155
16	PRSS3	0.00074	0.908	0.002	1.217	0.000
17	PRSS3P1	0.00073	0.904	0.002	1.183	0.000
18	CPA2	0.00073	0.897	0.002	1.179	0.000
19	CELA3B	0.00073	0.896	0.002	1.094	0.900
20	CTRC	0.00070	0.858	0.004	1.146	0.000

Note: FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 10: Top 20 domain-specific marker genes of Stroma region identified by DESpace

Rank	Gene	LR	logCPM	FDR	Stroma FDR
1	KRT17	557.281	14.156	1.34E-118	6.38E-32
2	S100A6	479.973	15.916	3.81E-102	8.51E-31
3	KRT19	464.301	15.505	6.32E-99	3.92E-31
4	GSTP1	328.460	14.149	1.00E-69	3.95E-20
5	GAPDH	251.421	14.581	2.14E-53	1.63E-18
6	AC009078.2	219.799	12.854	1.08E-46	4.71E-07
7	SPP1	200.969	13.478	9.51E-43	3.21E-08
8	MUC1	192.884	14.393	4.61E-41	2.60E-14
9	REG3G	179.804	13.243	2.45E-38	1.11E-05
10	IFI6	163.266	14.188	8.25E-35	4.20E-17
11	ACTB	161.190	15.028	2.18E-34	5.53E-11
12	EEF2	132.816	14.064	2.79E-28	8.69E-07
13	PFN1	127.231	14.415	4.22E-27	1.76E-11
14	FTH1	104.047	14.338	3.92E-22	2.41E-10
15	TMSB10	103.931	14.071	4.05E-22	2.30E-06
16	SPINK1	102.852	14.896	6.74E-22	1.17E-20
17	ATP1A1	64.627	13.527	9.57E-14	5.41E-06
18	RPLP2	60.268	14.440	8.00E-13	8.02E-05
19	EEF1A1	26.317	14.017	9.19E-06	4.43E-05
20	RPL11	19.809	14.383	2.00E-04	1.06E-04

Note: LR = Likelihood Ratio; logCPM = log Counts Per Million; FDR = False Discovery Rate. Genes are sorted by LR statistic in descending order. The Stroma FDR column shows the region-specific significance values.

Table 11: Top 20 domain-specific marker genes of Stroma region identified by Scanpy

Rank	Gene	$\log_2$FC	adj. p-value	Score
1	MT-RNR2	0.854	3.48E-07	6.763
2	MT-ND4	0.616	1.55E-06	6.438
3	MT-CO3	0.679	5.76E-04	5.068
4	CTRB2	0.756	1.17E-03	4.846
5	PRSS2	0.901	1.19E-03	4.828
6	HMGB3P17	1.437	2.25E-03	4.660
7	AD000090.1	0.809	1.76E-02	4.068
8	MUC5B	0.951	1.91E-02	4.038
9	MTRNR2L1	0.736	2.27E-02	3.981
10	MTRNR2L12	0.828	5.28E-02	3.698
11	CPA1	0.973	5.88E-02	3.658
12	MTCO1P40	0.174	7.68E-02	3.568
13	MTRNR2L8	0.939	8.75E-02	3.508
14	C3	0.679	1.05E-01	3.434
15	CLPS	0.676	1.18E-01	3.383
16	MT-CO1	0.203	1.18E-01	3.383
17	PRSS1	0.842	1.63E-01	3.235
18	CTRB1	0.597	1.91E-01	3.173
19	MTND4P12	0.477	1.99E-01	3.154
20	MT-CYB	0.390	2.11E-01	3.129

Note: $\log_2$FC = $\log_2$ fold change; adj. p-value = adjusted p-value. Genes are sorted by Score in descending order. All genes are confirmed markers of Stroma region according to ground truth annotation. The mitochondrial genes (MT- prefix) predominance suggests higher metabolic activity in stromal cells.

Table 12: Top 20 domain-specific marker genes of Duct Epithelium region identified by HarveST

Rank	Gene	Score	Scaled Score	p-value	Network FC	Expression FC
1	CES1	0.00106	1.000	0.002	1.821	0.000
2	PIGR	0.00102	0.965	0.002	2.049	15995.830
3	TFF3	0.00098	0.924	0.002	1.698	0.500
4	ID1	0.00095	0.894	0.002	1.706	0.000
5	DMBT1	0.00093	0.874	0.002	1.445	0.000
6	TCN1	0.00093	0.872	0.002	1.541	14781.010
7	TFF1	0.00092	0.860	0.002	1.792	19333.630
8	CRP	0.00091	0.856	0.002	1.506	0.000
9	AGT	0.00091	0.850	0.002	2.389	0.000
10	MMP1	0.00090	0.845	0.002	1.441	1.182
11	CLU	0.00089	0.835	0.002	1.429	0.000
12	SPP1	0.00089	0.830	0.002	1.468	40664.360
13	AQP3	0.00086	0.806	0.002	1.671	1.745
14	MUC5B	0.00085	0.790	0.002	1.227	0.000
15	DDIT4	0.00085	0.788	0.002	1.500	0.000
16	TPT1	0.00083	0.772	0.002	1.343	0.897
17	TMSB4X	0.00083	0.771	0.002	1.317	0.000
18	ATP1A1	0.00082	0.758	0.002	1.270	18206.130
19	RPS27A	0.00081	0.755	0.002	1.254	0.000
20	CLDN2	0.00080	0.746	0.002	1.358	2.005

Note: FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 13: Top 20 domain-specific marker genes of Duct Epithelium region identified by DESpace

Rank	Gene	LR	logCPM	FDR	Duct Epithelium FDR
1	KRT17	557.281	14.156	1.34E-118	3.14E-20
2	MUC5B	430.868	14.785	8.31E-92	3.34E-69
3	CTRB1	322.256	15.002	1.84E-68	2.19E-22
4	CTRB2	320.719	15.565	3.40E-68	1.63E-24
5	CEL	305.169	14.357	6.91E-65	1.52E-25
6	PRSS2	280.826	15.127	1.14E-59	3.29E-26
7	PRSS1	257.547	14.034	1.11E-54	9.51E-18
8	GP2	247.421	13.975	1.44E-52	1.25E-15
9	CPB1	240.443	14.265	4.29E-51	2.31E-14
10	CLPS	223.825	14.318	1.56E-47	4.33E-14
11	AC009078.2	219.799	12.854	1.08E-46	4.33E-14
12	CELA3A	218.793	13.986	1.67E-46	4.33E-14
13	CPA2	213.002	13.549	2.65E-45	7.66E-16
14	CTRC	198.868	13.498	2.58E-42	1.43E-14
15	PRSS3P1	197.858	13.737	4.06E-42	4.27E-17
16	PNLIPRP1	189.853	12.962	1.92E-40	4.23E-14
17	CELA3B	184.579	13.847	2.45E-39	3.38E-14
18	SERPINA1	180.500	15.184	1.79E-38	1.27E-32
19	DMBT1	167.679	13.804	9.50E-36	5.77E-35
20	LCN2	102.616	13.901	7.39E-22	2.60E-21

Note: LR = Likelihood Ratio; logCPM = log Counts Per Million; FDR = False Discovery Rate. Genes are sorted by LR statistic in descending order. The Duct Epithelium FDR column shows the region-specific significance values.

Table 14: Top 20 domain-specific marker genes of Duct Epithelium region identified by Scanpy

Rank	**Gene**	**$\log_2$FC**	**adj. p-value**	**Score**
1	MUC5B	2.572	5.98E-18	9.727
2	CES1	2.754	4.67E-12	8.150
3	DMBT1	2.372	9.36E-10	7.391
4	PIGR	2.879	3.65E-09	7.178
5	TGM2	2.409	1.69E-08	6.918
6	CRISP3	3.867	4.10E-08	6.772
7	TCN1	2.097	4.44E-08	6.743
8	TFF3	2.455	8.96E-08	6.625
9	ID1	2.298	1.77E-07	6.509
10	SERPINA1	1.094	9.75E-07	6.234
11	CRP	2.077	9.75E-07	6.222
12	ATP1B1	2.317	1.86E-06	6.091
13	TFF1	2.492	1.86E-06	6.089
14	MMP1	1.977	1.86E-06	6.086
15	APCS	2.168	1.88E-06	6.075
16	AGT	2.953	3.05E-06	5.987
17	HLA-B	1.697	1.81E-05	5.664
18	MUC3A	2.340	4.50E-05	5.498
19	FOS	2.897	5.29E-05	5.461
20	LCN2	1.399	6.89E-05	5.407

Note: $\log_2$FC = $\log_2$ fold change; adj. p-value = adjusted p-value. Genes are sorted by Score in descending order. All genes are confirmed markers of Duct Epithelium region according to ground truth annotation.

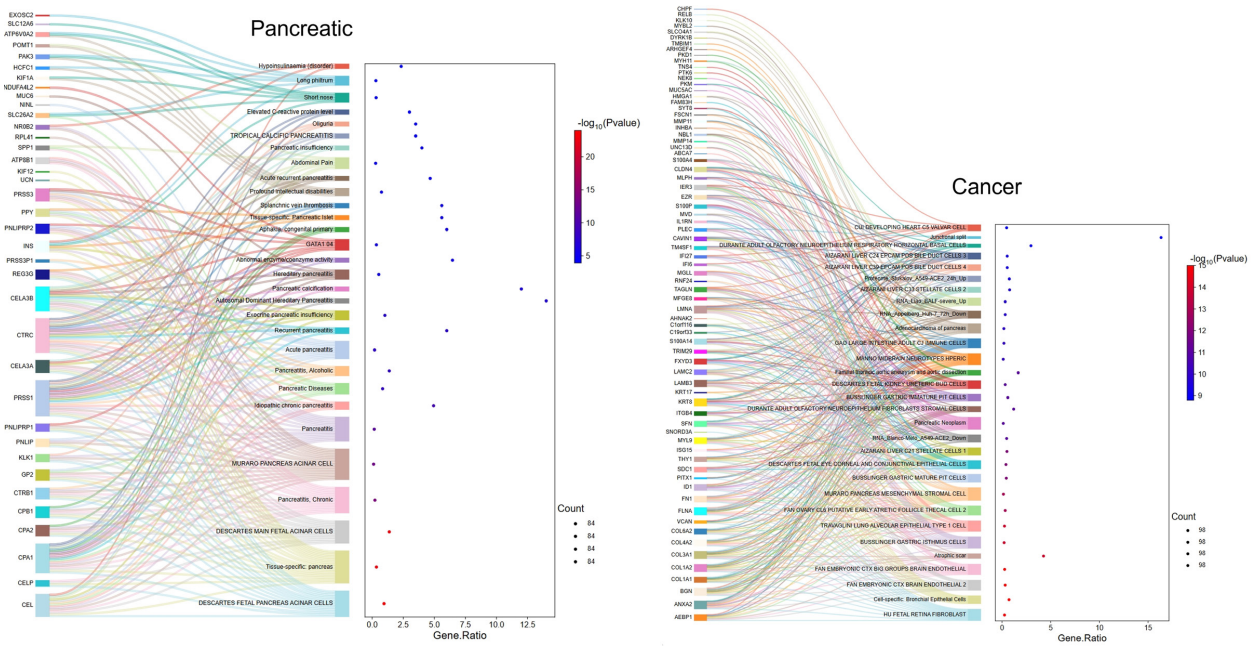

Figure S7: Sankey diagram analysis of disease/cell type associations of region-specific genes identified by HarveST in pancreatic tissue and cancer regions of pancreatic ductal adenocarcinoma (PDAC) spatial transcriptomics dataset. The figure presents Sankey diagrams illustrating the associations between region-specific genes identified by our HarveST method in the pancreatic tissue region (left) and cancer region (right) of PDAC and various diseases/cell types. The left side of each diagram displays region-specific genes, while the right side shows enriched diseases or cell types, with connection width indicating association strength and dot color intensity representing enrichment significance (-log10(P-value)). Genes from the pancreatic tissue region predominantly enriched in hypoglycemia disorder, elevated C-reactive protein level, pancreatic insufficiency, acute/chronic pancreatitis, and pancreatic acinar cells, all associated with pancreatic endocrine and exocrine functions, showing extremely high statistical significance (-log10(P-value)>20). In contrast, cancer region-specific genes significantly enriched in various developmental and differentiated cell types, including developing heart cells, olfactory neuroepithelial cells, liver bile duct cells, and gastrointestinal epithelial cells, suggesting reactivation of multiple developmental programs in cancer cells.

Table 15: Top 20 domain-specific marker genes of DCIS/LCIS1 region identified by HarveST

Rank	**Gene**	**Score**	**Scaled Score**	**p-value**	**Network FC**	**Expression FC**
1	ZNF350-AS1	0.00017	1.000	0.002	2.341	0.000
2	ERP27	0.00015	0.880	0.002	2.138	1.230
3	KRT37	0.00015	0.823	0.002	1.799	0.191
4	BNIP3	0.00014	0.813	0.002	1.743	2.181
5	AC087379.2	0.00014	0.794	0.002	1.892	1.703
6	DNAH5	0.00014	0.780	0.002	2.028	0.000
7	HMGB3	0.00014	0.759	0.002	1.602	1.222
8	CITED2	0.00014	0.757	0.002	1.656	0.422
9	S100A8	0.00013	0.725	0.002	1.988	0.808
10	ST8SIA6-AS1	0.00013	0.723	0.002	1.627	0.000
11	S100G	0.00013	0.716	0.002	1.401	1.125
12	RERG	0.00013	0.713	0.002	1.467	1.113
13	SPTSSB	0.00013	0.713	0.002	1.802	0.987
14	DRAIC	0.00013	0.693	0.002	1.587	0.000
15	CD163L1	0.00013	0.692	0.002	1.701	0.751
16	CKB	0.00013	0.685	0.002	1.630	1.109
17	KRT17	0.00013	0.685	0.002	2.019	1.067
18	SERPINA5	0.00013	0.685	0.002	1.583	0.996
19	CNTD1	0.00013	0.683	0.002	1.774	1.162
20	BACE2	0.00013	0.681	0.002	1.479	0.879

Table 16: Top 20 domain-specific marker genes of DCIS/LCIS1 region identified by DESpace

Rank	**Gene**	**LR Statistic**	**logCPM**	**P-value**	**FDR**
1	MGP	10643.848	14.121	0	0
2	IGFBP5	6513.865	13.792	0	3.79×10^{-167}
3	RPL17	4268.698	12.665	0	1.43×10^{-118}
4	TFF3	4566.735	12.353	0	5.10×10^{-290}
5	S100G	5436.800	10.632	0	1.87×10^{-235}
6	DSP	3250.829	10.771	0	2.36×10^{-290}
7	TFF1	3429.611	11.035	0	3.50×10^{-282}
8	SERPINA3	3601.616	11.109	0	5.07×10^{-205}
9	S100A6	3051.007	12.194	0	0
10	RPL38	2823.015	12.284	0	1.46×10^{-151}
11	RPL36A	2749.367	11.981	0	1.19×10^{-130}
12	RPS23	2469.447	12.181	0	2.82×10^{-130}
13	RPL26	2811.564	11.717	0	7.41×10^{-270}
14	LDHA	2780.101	10.782	0	3.73×10^{-264}
15	UQCRB	2301.430	11.158	0	3.11×10^{-121}
16	IER3	2660.607	10.838	0	2.19×10^{-182}
17	ACADSB	2648.621	10.352	0	4.49×10^{-168}
18	RERG	2094.600	9.617	0	3.98×10^{-148}
19	HMGB3	1538.812	9.394	2.05×10^{-315}	4.62×10^{-147}
20	BNIP3	1201.077	9.185	5.45×10^{-243}	1.71×10^{-151}

Note: LR = Likelihood Ratio statistic; logCPM = log counts per million; FDR = False Discovery Rate. Genes are ranked by LR statistic. Most genes showed statistically significant differences (P-value = 0) except where noted.

Table 17: Top 20 domain-specific marker genes of DCIS/LCIS_1 region identified by Scanpy

Rank	**Gene**	**Log Fold Change;**	**djusted P-value**	**Score**
1	MGP	3.101	8.00E-77	19.108
2	GAPDH	0.943	4.76E-67	17.833
3	RPL26	1.145	3.71E-63	17.285
4	RPS18	0.667	2.57E-61	17.030
5	DSP	1.371	1.20E-60	16.925
6	RPL19	0.586	2.06E-58	16.607
7	LDHA	1.328	8.57E-57	16.377
8	TFF1	1.939	5.02E-55	16.118
9	SERPINA3	1.598	8.12E-53	15.796
10	RPL23A	0.679	4.75E-52	15.680
11	ACADSB	1.474	1.13E-49	15.317
12	RPL30	0.593	4.30E-48	15.075
13	RPL38	0.628	1.12E-47	15.000
14	TFF3	1.344	1.28E-47	14.988
15	RPL37A	0.540	4.54E-47	14.899
16	IGFBP5	0.819	3.37E-45	14.604
17	RPS23	0.651	5.97E-44	14.402
18	RPL36A	0.715	1.43E-43	14.337
19	PGK1	1.055	7.88E-43	14.216
20	S100A6	1.147	5.08E-42	14.081

Table 18: Top 20 domain-specific marker genes of DCIS/LCIS1 region identified by SpaGCN

Rank	Gene	In%	Out%	I/O ratio	In exp	FC	adj.p
1	IGFBP5	1.000	0.994	1.006	4.204	2.131	7.07E-41
2	DSP	1.000	0.906	1.104	2.274	2.497	7.07E-41
3	RPL26	1.000	0.969	1.032	2.905	2.107	2.66E-40
4	HSP90AB1	1.000	0.994	1.006	3.146	1.796	9.83E-40
5	XBP1	1.000	0.994	1.006	2.968	1.859	5.63E-39
6	RPL17	1.000	0.994	1.006	3.259	1.712	7.83E-39
7	RPL23A	1.000	0.994	1.006	3.113	1.770	6.29E-38
8	RPL38	1.000	0.987	1.013	3.040	1.752	1.88E-37
9	RPL36A	1.000	0.987	1.013	2.842	1.821	1.41E-36
10	EFNA1	1.000	0.937	1.067	2.065	2.097	2.73E-36
11	ACADSB	1.000	0.893	1.120	1.933	2.324	7.01E-36
12	RPL14	1.000	0.981	1.019	2.641	1.678	7.01E-36
13	UQCRB	1.000	0.981	1.019	2.282	1.962	1.34E-35
14	TFF1	1.000	0.987	1.013	2.693	2.954	6.70E-35
15	YWHAE	1.000	0.925	1.082	1.859	1.891	1.60E-34
16	LDHA	1.000	0.956	1.046	2.265	2.191	2.23E-34
17	NME2	1.000	0.987	1.013	2.575	1.751	2.53E-34
18	COX7C	1.000	0.975	1.026	2.423	1.847	5.19E-34
19	TCEAL4	1.000	0.987	1.013	2.546	1.952	1.00E-33
20	HSPA1A	1.000	0.962	1.039	2.318	2.055	2.95E-33

Note: In% = in-group fraction; Out% = out-group fraction; I/O ratio = in/out group ratio; In exp = in-group mean expression; FC = fold change; adj.p = adjusted p-value. Genes are sorted by adjusted p-value in ascending order. DCIS = Ductal Carcinoma In Situ; LCIS = Lobular Carcinoma In Situ. Notably, all marker genes are expressed in 100% of cells within the DCIS/LCIS1 region.

Table 19: Top 20 domain-specific marker genes of DCIS/LCIS2 region identified by HarveST

Rank	**Gene**	**Score**	**Scaled Score**	**p-value**	**Network FC**	**Expression FC**
1	MMP7	0.00023	1.000	0.002	3.145	0.914
2	KRT5	0.00023	0.977	0.002	4.058	0.765
3	S100A2	0.00022	0.936	0.002	3.803	0.000
4	KRT14	0.00022	0.916	0.002	3.253	1.164
5	KRT17	0.00021	0.894	0.002	3.398	1.087
6	SAA1	0.00021	0.882	0.002	2.993	0.977
7	KRT6B	0.00020	0.853	0.002	3.935	2.628
8	MIA	0.00018	0.732	0.002	4.097	0.000
9	FOSB	0.00017	0.704	0.002	2.706	0.000
10	GABRP	0.00017	0.671	0.002	3.786	1.072
11	SHROOM1	0.00016	0.631	0.002	2.163	1.075
12	IRX1	0.00015	0.577	0.002	3.806	0.729
13	KRT23	0.00014	0.565	0.002	2.645	0.000
14	WFDC2	0.00014	0.563	0.002	1.670	0.901
15	FOS	0.00014	0.557	0.002	1.602	0.467
16	LAMB3	0.00014	0.557	0.002	2.867	0.000
17	KRT81	0.00014	0.540	0.002	2.773	0.950
18	SPTSSB	0.00014	0.533	0.002	1.915	1.263
19	SLPI	0.00014	0.531	0.002	2.786	1.026
20	LAMC2	0.00014	0.528	0.002	3.278	1.035

Note: FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 20: Top 20 domain-specific marker genes of DCIS/LCIS2 region identified by DESpace

Rank	**Gene**	**LR Statistic**	**logCPM**	**P-value**	**DCIS/LCIS2 FDR**
1	MGP	10643.848	14.121	0	3.29×10^{-55}
2	IGFBP5	6513.865	13.792	0	1.49×10^{-15}
3	S100G	5436.800	10.632	0	2.68×10^{-42}
4	TFF3	4566.735	12.353	0	1.71×10^{-13}
5	RPL17	4268.698	12.665	0	4.71×10^{-4}
6	SERPINA3	3601.616	11.109	0	7.73×10^{-203}
7	TFF1	3429.611	11.035	0	5.83×10^{-25}
8	DSP	3250.829	10.771	0	6.10×10^{-25}
9	S100A6	3051.007	12.194	0	1.35×10^{-45}
10	RPL38	2823.015	12.284	0	1.56×10^{-12}
11	RPL26	2811.564	11.717	0	1.36×10^{-43}
12	RPL36A	2749.367	11.981	0	3.19×10^{-15}
13	LDHA	2780.101	10.782	0	1.15×10^{-3}
14	IER3	2660.607	10.838	0	0.896
15	ACADSB	2648.621	10.352	0	1.30×10^{-54}
16	RPS23	2469.447	12.181	0	5.39×10^{-19}
17	UQCRB	2301.430	11.158	0	1.05×10^{-12}
18	RERG	2094.600	9.617	0	4.86×10^{-8}
19	HMGB3	1538.812	9.394	2.05×10^{-315}	1.01×10^{-2}
20	BNIP3	1201.077	9.185	5.45×10^{-243}	0.626

Note: LR = Likelihood Ratio statistic; logCPM = log counts per million; FDR = False Discovery Rate. Genes are ranked by LR statistic. Most genes showed statistically significant differences (P-value = 0) except where noted. Non-significant FDR values (FDR > 0.05) are shown in plain decimal format.

Table 21: Top 20 domain-specific marker genes of DCIS/LCIS2 region identified by Scanpy

Rank	**Gene**	**logfoldchanges**	**pvals_adj**	**scores**
1	SERPINA3	2.8202045	2.95E-14	8.859036
2	MMP7	3.5550125	7.25E-12	8.139769
3	CPB1	3.0980315	1.08E-11	8.041341
4	KRT5	3.7562945	2.58E-11	7.8987155
5	SAA1	3.3312678	2.72E-11	7.864205
6	S100G	2.3547359	1.22E-10	7.651002
7	TNFSF10	1.6649883	1.33E-10	7.619865
8	MGP	2.075939	4.07E-10	7.4564815
9	KRT14	3.5200076	5.12E-10	7.4102945
10	RPL26	1.0226438	5.48E-10	7.3872876
11	FOS	2.1784728	9.21E-10	7.305293
12	KRT17	3.4189346	3.73E-09	7.081365
13	KRT6B	4.223693	8.23E-09	6.961228
14	S100A2	3.8135648	9.66E-09	6.9295716
15	ACADSB	1.651405	1.10E-08	6.9021535
16	S100A6	1.0071373	1.24E-08	6.874044
17	WFDC2	2.0481818	1.24E-08	6.869114
18	SHROOM1	2.6617312	2.08E-08	6.7880707
19	CFB	1.2746075	3.50E-08	6.7055573
20	BTG2	1.534204	1.59E-07	6.4607844

Table 22: Top 20 domain-specific marker genes of DCIS/LCIS2 region identified by SpaGCN

Rank	Gene	In%	Out%	I/O ratio	In exp	FC	adj.p
1	S100G	1.000	0.940	1.064	2.505	3.704	2.35E-07
2	ACADSB	1.000	0.960	1.042	2.067	2.817	9.77E-07
3	TNFSF10	1.000	0.980	1.020	2.443	2.524	1.72E-06
4	POLR2K	1.000	0.960	1.042	1.569	1.792	4.91E-06
5	CSTA	1.000	0.960	1.042	1.996	2.155	4.91E-06
6	CTPS2	1.000	0.780	1.282	1.259	1.915	5.02E-06
7	RPL22L1	1.000	0.940	1.064	1.527	1.975	5.60E-06
8	ABHD2	1.000	0.920	1.087	1.571	1.917	8.54E-06
9	TOMM6	0.964	0.920	1.048	1.407	1.579	8.84E-06
10	IFIT1	1.000	0.900	1.111	1.368	1.717	1.10E-05
11	CSNK1A1	1.000	0.980	1.020	1.674	1.568	1.66E-05
12	FOS	1.000	0.900	1.111	1.674	2.091	1.84E-05
13	TCIM	0.929	0.780	1.190	0.942	1.639	1.99E-05
14	C1QBP	1.000	0.680	1.471	1.050	1.761	3.05E-05
15	ERLIN2	0.964	0.880	1.096	1.437	1.890	6.26E-05
16	RBIS	1.000	0.980	1.020	1.468	1.531	1.17E-04
17	SHROOM1	0.929	0.440	2.110	0.869	1.884	1.45E-04
18	SNCG	1.000	0.800	1.250	1.044	1.623	1.82E-04
19	TXNDC17	1.000	0.920	1.087	1.328	1.613	1.96E-04
20	SEC61G	1.000	0.940	1.064	1.415	1.562	3.52E-04

Note: In% = in-group fraction; Out% = out-group fraction; I/O ratio = in/out group ratio; In exp = in-group mean expression; FC = fold change; adj.p = adjusted p-value. Genes are sorted by adjusted p-value in ascending order. DCIS = Ductal Carcinoma In Situ; LCIS = Lobular Carcinoma In Situ. SHROOM1 shows the highest in/out ratio (2.110), indicating strong region specificity.

Table 23: Top 20 domain-specific marker genes of IDC2 region identified by HarveST

Rank	Gene	Score	Scaled Score	p-value	Network FC	Expression FC
1	SLITRK6	0.00021	1.000	0.002	2.684	0.229
2	C6orf141	0.00018	0.825	0.002	2.316	1.174
3	CRISP3	0.00018	0.820	0.002	1.968	1.146
4	VTCN1	0.00016	0.703	0.002	1.883	1.107
5	ABCC11	0.00014	0.628	0.002	2.040	2.739
6	CEACAM6	0.00014	0.627	0.002	1.878	0.000
7	MAB21L4	0.00014	0.612	0.002	1.957	1.231
8	PIP	0.00014	0.602	0.002	1.998	0.000
9	CRISP2	0.00014	0.598	0.002	2.486	10.355
10	SHISA2	0.00014	0.588	0.002	1.649	0.000
11	CYP4B1	0.00013	0.580	0.002	1.735	1.004
12	PTGES	0.00013	0.556	0.002	1.586	2.304
13	ABCC12	0.00013	0.553	0.002	2.178	3.981
14	NAT1	0.00013	0.550	0.002	1.596	0.000
15	KIAA2012-AS1	0.00013	0.550	0.002	2.047	1.784
16	AC093001.1	0.00013	0.546	0.002	1.607	1.088
17	CBWD5	0.00012	0.529	0.002	1.535	0.579
18	PDLIM1	0.00012	0.526	0.002	1.426	2.260
19	SERHL2	0.00012	0.522	0.002	1.336	0.679
20	ANKRD30B	0.00012	0.518	0.002	1.551	0.000

Note: FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 24: Top 20 domain-specific marker genes of IDC2 region identified by scanpy

Rank	**Gene**	**Log Fold Change**	**Score**
1	CRISP3	4.119	31.715
2	SLITRK6	3.872	30.980
3	H3F3A	1.021	29.291
4	IGFBP5	1.099	29.034
5	C6orf141	3.170	28.526
6	S100A13	1.131	28.044
7	PSMA6	1.331	27.472
8	VTCN1	2.480	27.221
9	SERHL2	1.794	26.951
10	S100A16	1.045	26.701
11	PPIA	0.720	25.671
12	S100A11	0.744	24.308
13	NUPR1	0.904	24.222
14	KRT8	0.789	24.020
15	EFHD1	1.400	23.543
16	CTTN	0.950	23.523
17	DBI	1.186	23.499
18	CEACAM6	2.226	23.312
19	PTMA	0.563	23.261
20	AZGP1	1.104	23.233

Note: All genes are from the IDC2 region (Ground Truth: IDC_2). Scores represent the statistical significance of each gene as a marker for the region.

Table 25: Top 20 domain-specific marker genes of IDC2 region identified by DESpace

Rank	**Gene**	**LR Statistic**	**logCPM**	**P-value**	**FDR**
1	KRT18	7174.337	13.343	0	0
2	IGHG4	2549.845	13.350	0	0
3	IGHM	2011.497	10.613	0	0
4	CCL19	1404.839	9.012	1.17×10^{-286}	7.67×10^{-286}
5	TRBC2	1267.538	9.217	3.19×10^{-257}	1.77×10^{-256}
6	PTGDS	873.612	8.950	4.70×10^{-173}	1.64×10^{-172}
7	LTB	898.241	8.869	2.67×10^{-178}	9.61×10^{-178}
8	TRAC	929.729	9.012	5.19×10^{-185}	1.93×10^{-184}
9	CD52	956.363	9.082	1.09×10^{-190}	4.18×10^{-190}
10	CD79A	851.450	8.566	2.45×10^{-168}	8.34×10^{-168}
11	CORO1A	775.567	9.225	3.34×10^{-152}	1.05×10^{-151}
12	JAK3	578.208	8.663	1.99×10^{-110}	5.14×10^{-110}
13	TBC1D10C	569.547	8.618	1.33×10^{-108}	3.40×10^{-108}
14	CD37	614.329	8.759	4.77×10^{-118}	1.28×10^{-117}
15	RAC2	482.446	8.668	2.68×10^{-90}	6.38×10^{-90}
16	TMC8	446.190	8.574	1.03×10^{-82}	2.35×10^{-82}
17	ARHGAP45	462.619	8.843	3.79×10^{-86}	8.86×10^{-86}
18	CXCR4	396.087	8.871	2.86×10^{-72}	6.25×10^{-72}
19	RASGRP2	402.696	8.455	1.21×10^{-73}	2.66×10^{-73}
20	MS4A1	525.443	8.422	2.54×10^{-99}	6.21×10^{-99}

Note: LR = Likelihood Ratio statistic; logCPM = log counts per million; FDR = False Discovery Rate. All genes showed statistically significant differences (P-value = 0) except where noted.

Table 26: Top 20 domain-specific marker genes of IDC3 region identified by HarveST

Rank	Gene	Score	Scaled Score	p-value	Network FC	Expression FC
1	CD79A	0.00026	1.000	0.002	3.925	0.000
2	MS4A1	0.00024	0.923	0.002	4.318	0.000
3	TBC1D10C	0.00021	0.770	0.002	2.859	1.021
4	RASGRP2	0.00021	0.763	0.002	3.375	1.218
5	CD52	0.00020	0.755	0.002	2.453	0.992
6	JCHAIN	0.00020	0.750	0.002	2.346	1.569
7	CCL19	0.00020	0.749	0.002	2.437	0.000
8	PTGDS	0.00020	0.748	0.002	2.495	0.000
9	TRBC2	0.00020	0.740	0.002	2.347	0.000
10	RAC2	0.00020	0.722	0.002	2.689	0.000
11	JAK3	0.00019	0.695	0.002	2.543	1.293
12	IL7R	0.00019	0.681	0.002	2.755	0.942
13	CXCR4	0.00019	0.679	0.002	2.356	0.000
14	LTB	0.00019	0.676	0.002	2.338	0.000
15	CD3E	0.00018	0.650	0.002	2.506	38.002
16	ACAP1	0.00018	0.650	0.002	2.497	0.000
17	TRAC	0.00018	0.646	0.002	2.175	0.793
18	TRBC1	0.00018	0.645	0.002	2.282	0.000
19	SELL	0.00018	0.644	0.002	2.810	0.982
20	BIRC3	0.00018	0.631	0.002	2.655	0.000

Note: FC = Fold Change; Network FC refers to the fold change of Random Walk with Restart (RWR) scores in heterogeneous graph by Harvest, while Expression FC indicates the fold change in gene expression levels. All genes showed statistically significant differences ($p < 0.002$).

Table 27: Top 20 domain-specific marker genes of IDC3 region identified by DESpace

Rank	**Gene**	**LR Statistic**	**logCPM**	**P-value**	**FDR**
1	IGHG4	2549.845	13.350	0	0
2	KRT18	7174.337	13.343	0	0
3	IGHM	2011.497	10.613	0	0
4	CCL19	1404.839	9.012	1.17×10^{-286}	7.67×10^{-286}
5	TRBC2	1267.538	9.217	3.19×10^{-257}	1.77×10^{-256}
6	PTGDS	873.612	8.950	4.70×10^{-173}	1.64×10^{-172}
7	CD79A	851.450	8.566	2.45×10^{-168}	8.34×10^{-168}
8	LTB	898.241	8.869	2.67×10^{-178}	9.61×10^{-178}
9	TRAC	929.729	9.012	5.19×10^{-185}	1.93×10^{-184}
10	CD52	956.363	9.082	1.09×10^{-190}	4.18×10^{-190}
11	CORO1A	775.567	9.225	3.34×10^{-152}	1.05×10^{-151}
12	JAK3	578.208	8.663	1.99×10^{-110}	5.14×10^{-110}
13	TBC1D10C	569.547	8.618	1.33×10^{-108}	3.40×10^{-108}
14	CD37	614.329	8.759	4.77×10^{-118}	1.28×10^{-117}
15	RAC2	482.446	8.668	2.68×10^{-90}	6.38×10^{-90}
16	TMC8	446.190	8.574	1.03×10^{-82}	2.35×10^{-82}
17	ARHGAP45	462.619	8.843	3.79×10^{-86}	8.86×10^{-86}
18	RASGRP2	402.696	8.455	1.21×10^{-73}	2.66×10^{-73}
19	CXCR4	396.087	8.871	2.86×10^{-72}	6.25×10^{-72}
20	MS4A1	525.443	8.422	2.54×10^{-99}	6.21×10^{-99}

Note: LR = Likelihood Ratio statistic; logCPM = log counts per million; FDR = False Discovery Rate. All genes showed statistically significant differences (P-value = 0) except where noted.

Table 28: Top 20 domain-specific marker genes of IDC3 region identified by Scanpy

Rank	**Gene**	**logFC**	**Adj. P-value**	**Score**
1	IGHA1	2.47	2.51×10^{-20}	10.30
2	TRBC2	2.79	2.12×10^{-19}	10.03
3	JCHAIN	2.80	1.06×10^{-17}	9.59
4	IGHM	2.76	1.06×10^{-17}	9.56
5	IGKC	2.20	1.30×10^{-17}	9.52
6	CD79A	3.63	2.71×10^{-16}	9.16
7	IGHG4	1.78	2.71×10^{-16}	9.17
8	IGHG3	2.20	1.42×10^{-15}	8.96
9	RPL11	0.80	1.73×10^{-14}	8.67
10	RPL21	0.81	1.73×10^{-14}	8.66
11	CD52	2.71	6.03×10^{-14}	8.50
12	PTGDS	2.64	9.36×10^{-14}	8.43
13	PTPRC	2.14	1.56×10^{-13}	8.36
14	IGLC2	2.10	1.56×10^{-13}	8.35
15	CCL19	2.78	1.56×10^{-13}	8.35
16	TRAC	2.28	7.41×10^{-13}	8.14
17	IGHG1	1.58	3.27×10^{-12}	7.96
18	TBC1D10C	2.85	5.72×10^{-12}	7.88
19	TXNIP	1.34	9.81×10^{-12}	7.81
20	JAK3	2.54	1.04×10^{-11}	7.79

Note: logFC = log fold change; Adj. P-value = Adjusted P-value; Score = Differential expression score. All genes were identified as significantly associated with IDC3 region (Ground Truth = IDC_3).

Table 29: Top 20 domain-specific marker genes of IDC_3 region identified by DESpace

Rank	Gene	LR	logCPM	FDR	IDC_3 FDR
1	KRT18	7174.337	13.343	0.000	6.43E-47
2	IGHG4	2549.845	13.350	0.000	2.75E-47
3	IGHM	2011.497	10.613	0.000	7.71E-66
4	CCL19	1404.839	9.012	7.67E-286	5.59E-63
5	TRBC2	1267.538	9.217	1.77E-256	8.61E-99
6	CD52	956.363	9.082	4.18E-190	1.58E-108
7	TRAC	929.729	9.012	1.93E-184	6.27E-67
8	LTB	898.241	8.869	9.61E-178	3.74E-77
9	PTGDS	873.612	8.950	1.64E-172	8.14E-47
10	CD79A	851.450	8.566	8.34E-168	1.70E-57
11	CORO1A	775.567	9.225	1.05E-151	9.85E-56
12	CD37	614.329	8.759	1.28E-117	9.80E-64
13	JAK3	578.208	8.663	5.14E-110	3.51E-52
14	TBC1D10C	569.547	8.618	3.40E-108	2.66E-54
15	MS4A1	525.443	8.422	6.21E-99	2.50E-79
16	RAC2	482.446	8.668	6.38E-90	2.44E-56
17	ARHGAP45	462.619	8.843	8.86E-86	2.70E-60
18	TMC8	446.190	8.574	2.35E-82	3.39E-52
19	RASGRP2	402.696	8.455	2.66E-73	1.92E-55
20	CXCR4	396.087	8.871	6.25E-72	2.05E-52

Note: LR = Likelihood Ratio; logCPM = log Counts Per Million; FDR = False Discovery Rate. Genes are sorted by LR statistic. For FDR values of 0.000, the actual values are below computational precision. The IDC_3 FDR column shows the region-specific significance values.

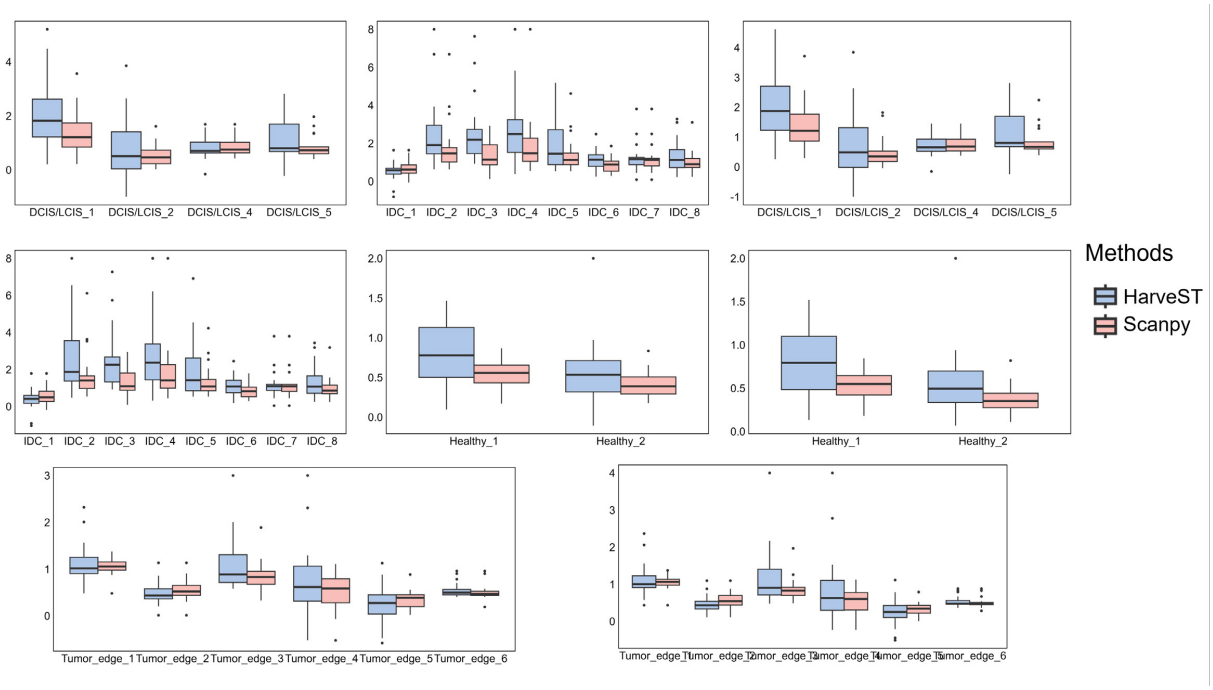

Figure S8: Boxplots depicting the relative spatial specificity of the top 20 region-specific genes identified in each region of the HBRC dataset. The y-axis represents the ratio of Moran's I within the designated region to its Moran's I across all other regions, with higher values indicating stronger regional specificity. The results obtained using the HarveST method are shown in blue, while those from the baseline Scanpy approach are presented in pink. This comparison highlights differences in the sensitivity and effectiveness of the two methods in capturing spatially distinctive gene expression patterns across diverse tissue microenvironments.

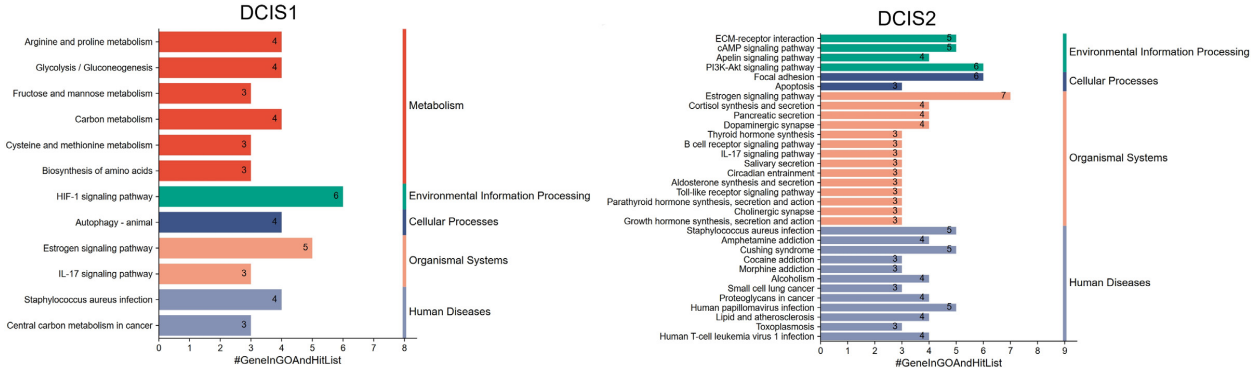

Figure S9: KEGG pathway enrichment analysis of the top 100 region-specific genes identified by HarveST in the DCIS/LCIS regions of the HBRC dataset. The results demonstrate a strong enrichment of metabolic pathways, including glycolysis, carbon metabolism, and amino acid biosynthesis, as well as signaling pathways such as HIF-1, estrogen, and IL-17, which are known to be associated with tumor microenvironment dynamics. Additionally, pathways related to human diseases, including cancer-associated signaling cascades and infection-related pathways, are significantly enriched. These findings highlight the capability of HarveST to accurately capture biologically relevant and spatially distinctive gene expression patterns, providing valuable insights into the molecular landscape of DCIS/LCIS.

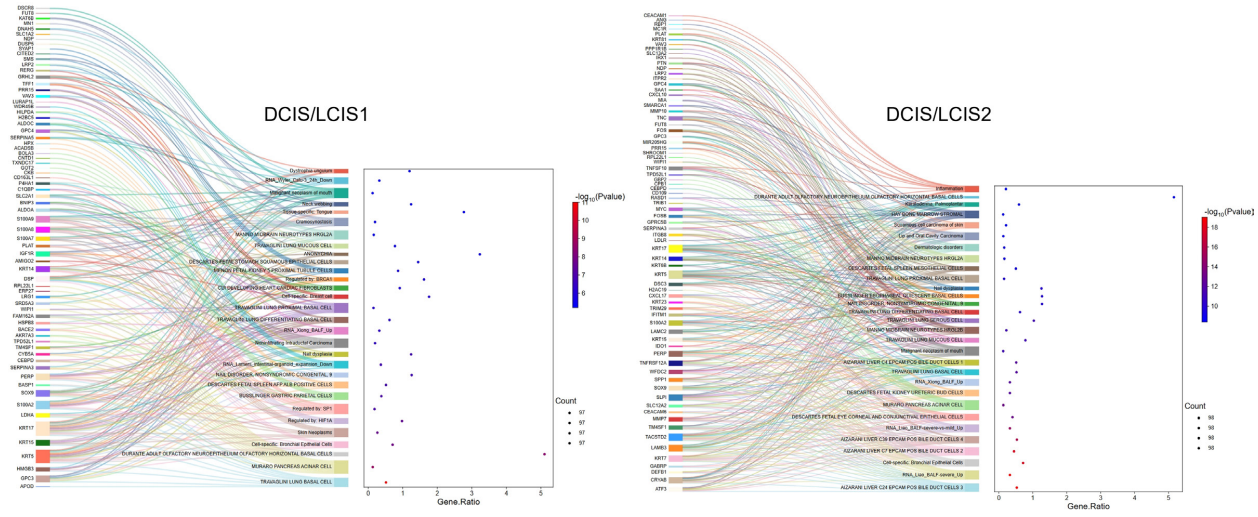

Figure S10: Sankey diagrams illustrating the functional enrichment of the top-ranked region-specific genes identified in the DCIS/LCIS1 and DCIS/LCIS2 regions of the HBRC dataset. The left panel represents the results for DCIS/LCIS1, while the right panel corresponds to DCIS/LCIS2. Each diagram visualizes the associations between region-specific genes and their enriched biological processes or cellular components, with pathway significance represented by the color gradient corresponding to $-(\text{P-value})$. The enrichment results reveal a strong involvement of these genes in tumor-associated processes, including epithelial differentiation, inflammation, and neoplastic transformation, further underscoring the biological relevance of the spatially resolved gene signatures identified by HarveST.

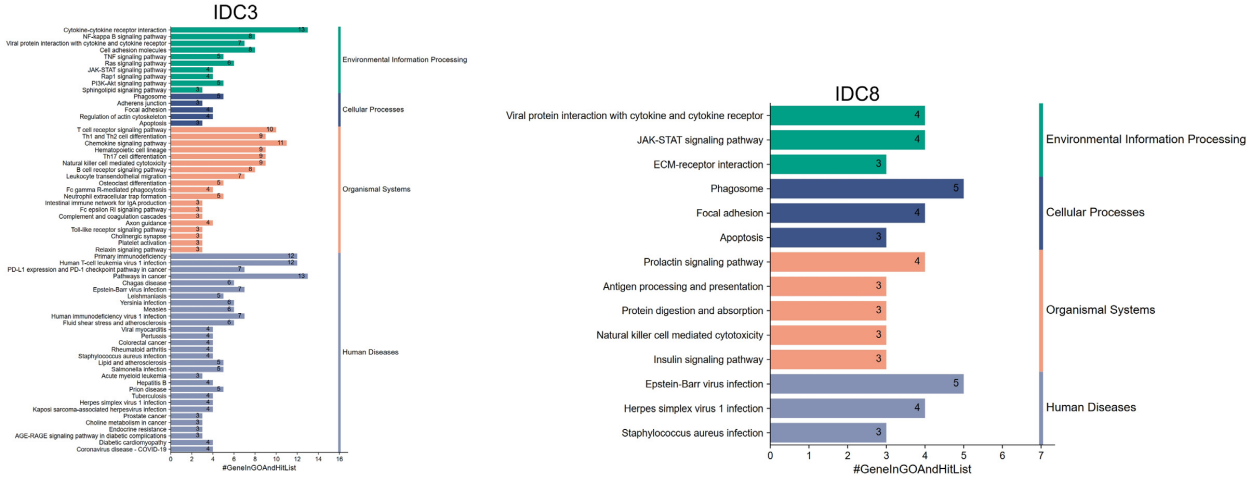

Figure S11: KEGG pathway enrichment analysis of IDC region-specific genes identified by HarveST in the breast cancer spatial transcriptomics dataset. The figure illustrates KEGG pathway enrichment analysis of the top 100 region-specific genes identified by our HarveST method in invasive ductal carcinoma (IDC3 and IDC8) regions. In the IDC3 region (left panel), genes were significantly enriched in pathways related to environmental information processing and immune function, including cytokine-cytokine receptor interaction, NF-kappaB signaling, and viral protein interaction, as well as multiple human disease-associated pathways. The IDC8 region (right panel) exhibited similar enrichment patterns, with notable overrepresentation in viral protein interaction with cytokine and cytokine receptor, JAK-STAT signaling pathway, and Epstein-Barr virus infection. These findings reveal distinct molecular signatures of different IDC regions in immune modulation and environmental information processing, demonstrating the effectiveness of HarveST in identifying region-specific molecular mechanisms within the breast cancer microenvironment. The results provide valuable molecular insights into breast cancer heterogeneity and potential therapeutic targets for precision medicine approaches.

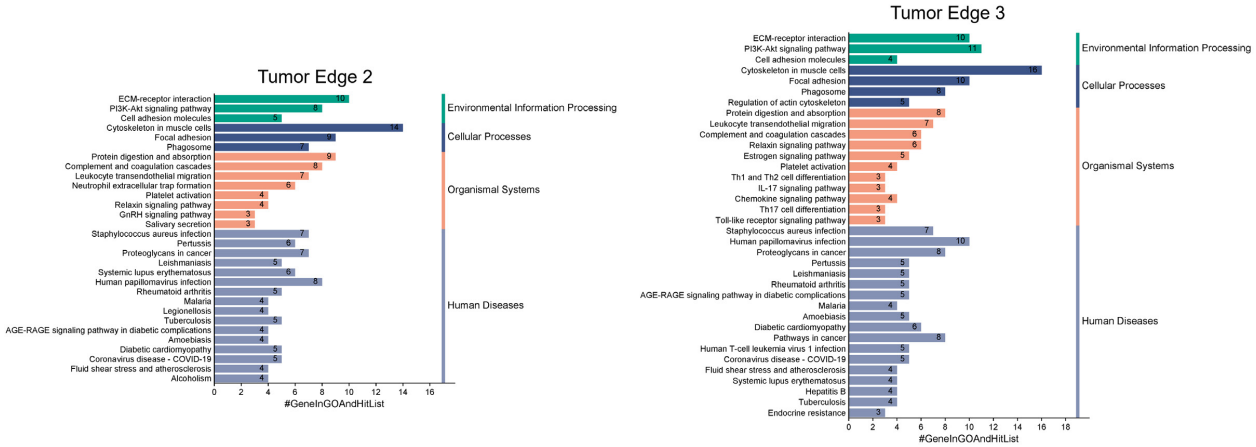

Figure S12: KEGG pathway enrichment analysis of tumor edge region-specific genes identified by HarveST in breast cancer spatial transcriptomics dataset. The figure illustrates KEGG pathway enrichment analysis of region-specific genes identified by our HarveST method in two tumor edge regions (Tumor Edge 2 and Tumor Edge 3) of the breast cancer dataset. Both regions demonstrate significant enrichment in pathways related to environmental information processing and cellular processes, particularly ECM-receptor interaction, PI3K-Akt signaling pathway, and cytoskeleton-related pathways. In terms of organismal systems, both regions show enrichment in pathways associated with immune and inflammatory responses, including leukocyte transendothelial migration, platelet activation, and complement and coagulation cascades. Regarding human diseases, pathways related to Staphylococcus aureus infection, human papillomavirus infection, rheumatoid arthritis, and proteoglycans in cancer are significantly enriched in both regions. Notably, Tumor Edge 3 exhibits additional enrichment in immune-related pathways such as Th1 and Th2 cell differentiation, IL-17 signaling pathway, and chemokine signaling pathway, suggesting stronger immunomodulatory characteristics in this region. These findings reveal critical molecular features of tumor edge regions in the breast cancer microenvironment, particularly their important roles in extracellular matrix remodeling, cytoskeletal dynamics, and immune regulation.

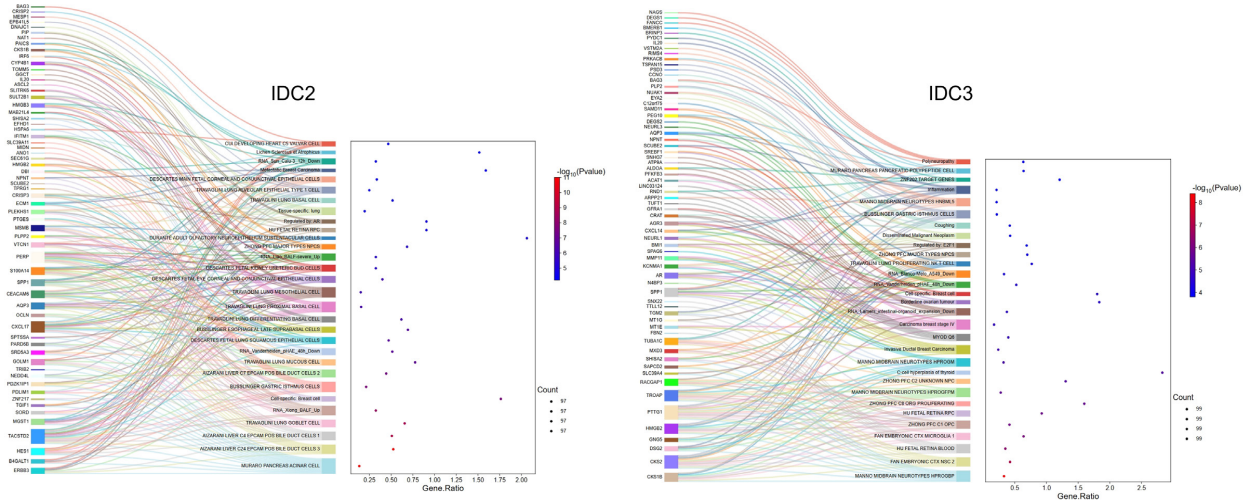

Figure S13: Sankey diagram analysis of cell type associations of region-specific genes identified by HarveST in IDC regions of breast cancer spatial transcriptomics dataset. The figure presents Sankey diagrams illustrating the associations between region-specific genes identified by our HarveST method and various cell types in invasive ductal carcinoma regions IDC2 (left) and IDC3 (right). The left side of each diagram displays differentially expressed genes, while the right side shows enriched cell types, with connection width indicating association strength. Genes from the IDC2 region predominantly enriched in various epithelial cell types including alveolar epithelial cells, tracheal epithelial cells, and airway epithelial cells, whereas IDC3 region genes significantly enriched in pancreatic acinar cells, kidney proximal tubule cells, and myeloid cells. Colored dots adjacent to each cell type represent enrichment significance (-log10(P-value)).