Accurate Imputation of Pathway-specific Gene Expression in Spatial Transcriptomics with PASTA

Ruoxing Li[1,2], Peng Yang[1,3], Mauro Di Pilato[4], Jianjun Zhang[5], Christopher R Flowers[6], Lulu Shang[1,*] and Ziyi Li[1,*]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
[2]Department of Biostatistics, The University of Texas Health Science Center, Houston, TX 78284, USA.
[3]Department of Statistics, Rice University, Houston, TX 77005, USA.
[4]Department of Immunology, The Unviersity of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
[5]Department of Thoracic/Head and Neck Medical Oncology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
[6]Department of Lymphoma - Myeloma, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
[*]Correspondence: Zli16@mdanderson.org, LShang@mdanderson.org

March 18, 2025

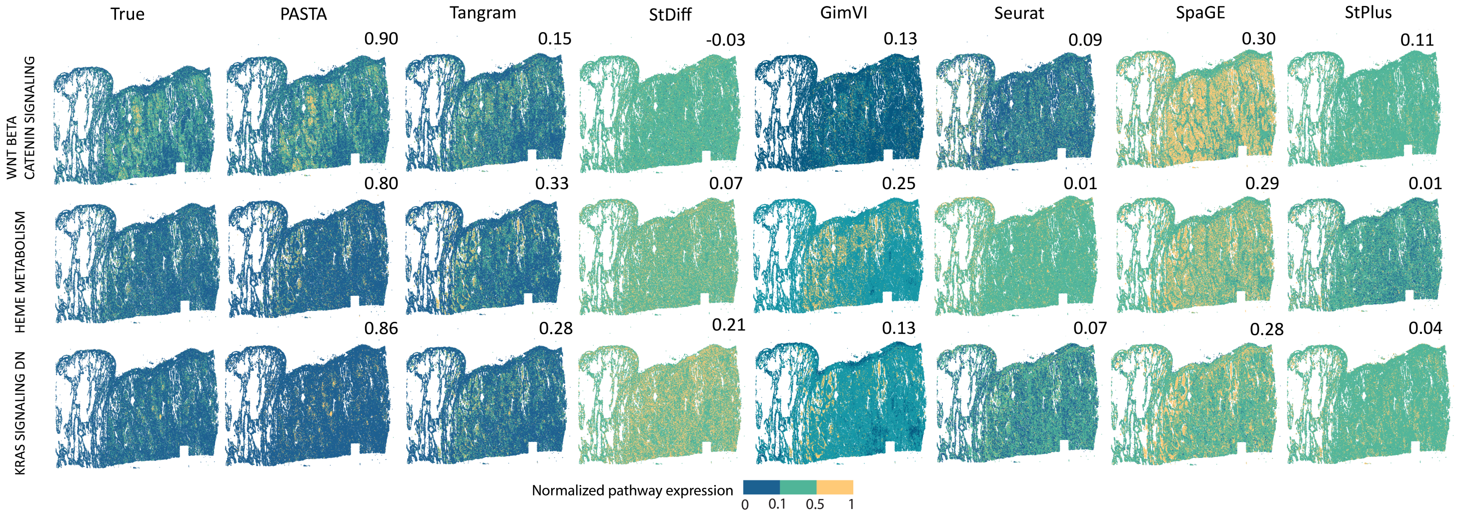

Figure S1: **Imputed pathway expression by PASTA and six other existing methods using a human lung cancer spatial transcriptomics (ST) dataset.** We collected a ST dataset from human lung cancer tissue with 480 genes processed by 10x Genomics Xenium [Janesick et al., 2023]. We annotated the dataset manually referring to the PanglaoDB [Franzén et al., 2019] for the seven clusters that Xenium provides. The corresponding scRNA-seq dataset was collected from the UCSC cell Browser website. Three pathways are presented: WNT BETA CATENIN SIGNALING, HEME METABOLISM, and KRAS SIGNALING DN (from top to bottom, respectively). Normalized pathway expression are visualized for the cells. PASTA-imputed pathway expression shows consistently higher correlations with true values, as indicated by the strong correlations between the predicted and actual pathway expression.

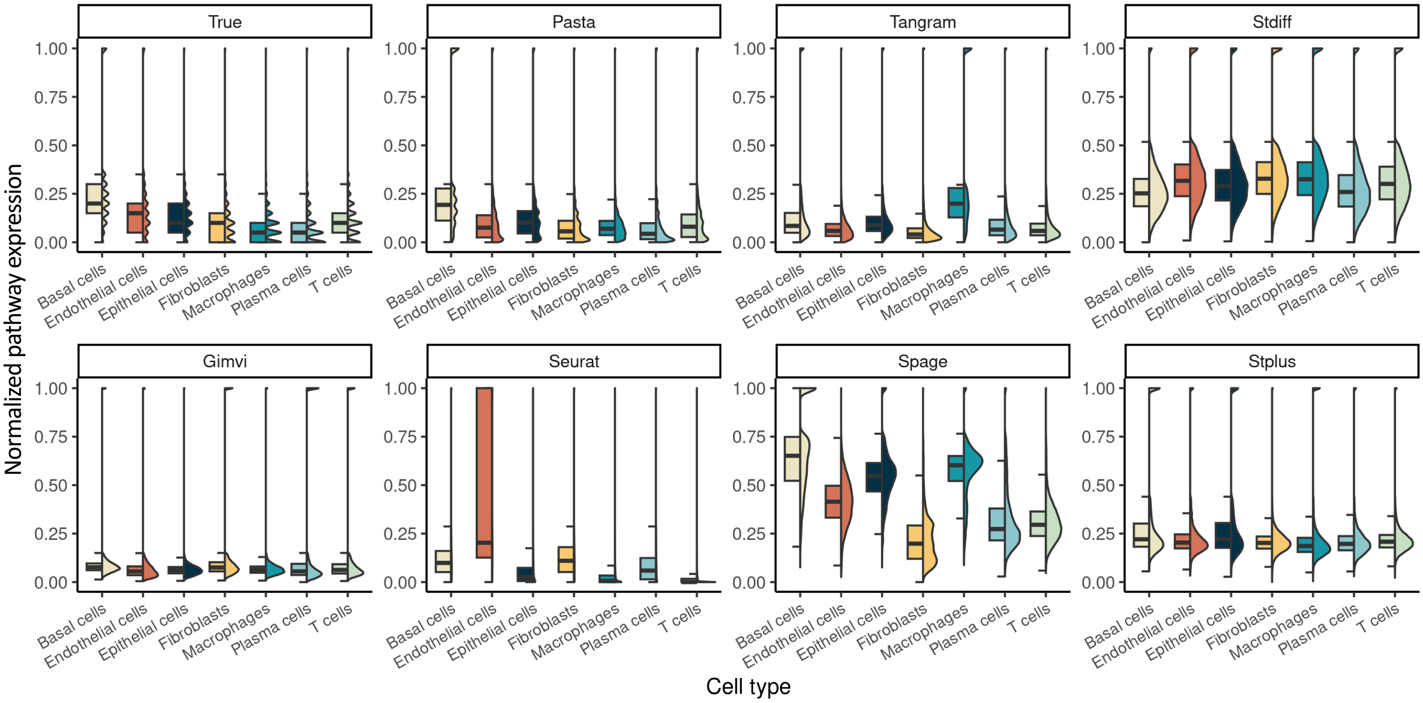

Figure S2: **Distributions of a hallmark pathway expression.** We collected a ST dataset from human lung cancer tissue with 480 genes, and applied the impuataion methods on the ST dataset. We compare the distribution of WNT BETA CATENIN SIGNALING pathway predictions among different cell types. PASTA predicts expression distributions among all the cell types similar to the ground truth. In contrast, the other methods fail in predicting one or more cell types.

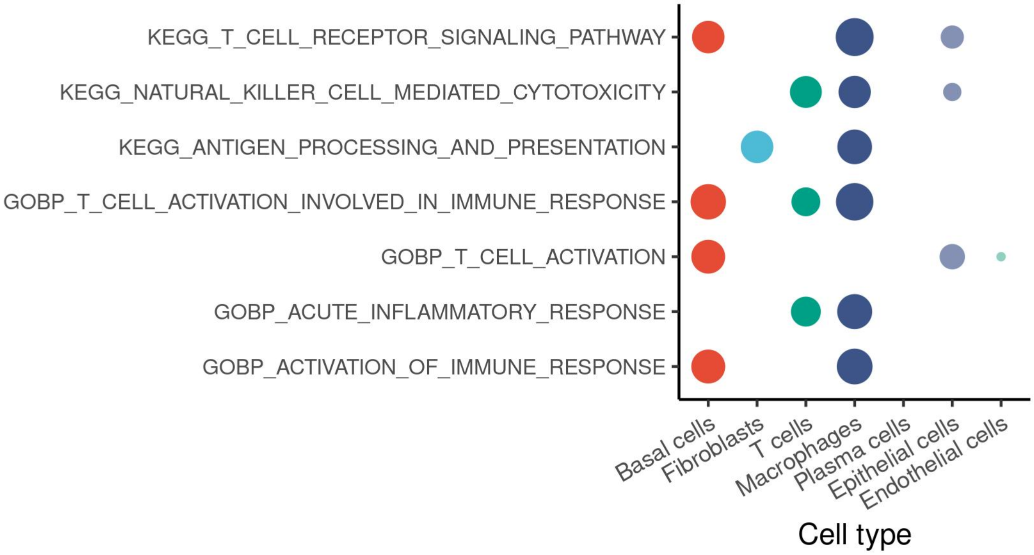

Figure S3: **Enrichment of immune pathways predicted by PASTA in different cell types.** We manually selected a group of immune-related pathways from GOBP (Gene Ontology Biological Process) and KEGG (Kyoto Encyclopedia of Genes and Genomes) databases. Each dot indicates a significant pathway enrichment in a particular cell type, with the size of the dot reflecting the significance level: the larger the dot, the more significant the enrichment. In general, PASTA successfully predicted that those pathways were enriched in immune cells.

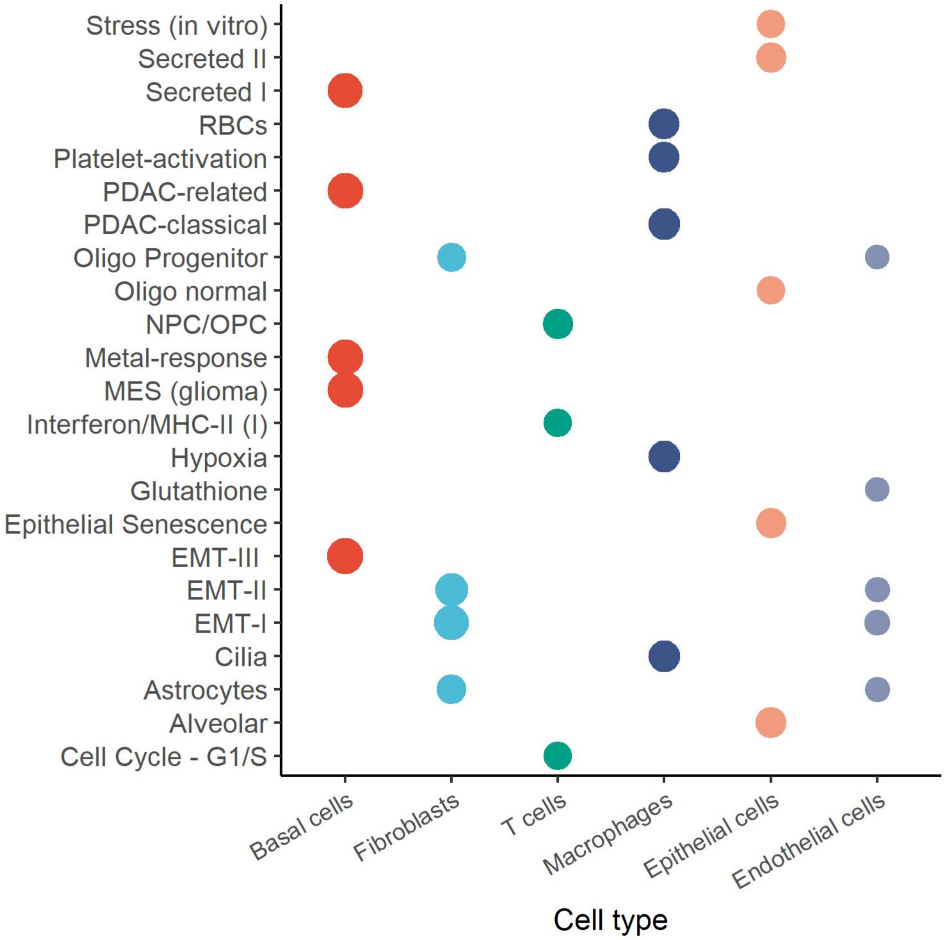

Figure S4: **Top 5 of enriched predicted meta-programs (MPs) in each cell type.** The MPs are developed by [Gavish et al., 2023], consisting of genes that are upregulated in subpopulations of cells within many tumors. MPs are presented along the Y axis, and the names are reflecting its functional annotations. RBCs: red blood cells; PDAC: pancreatic ductal adenocarcinoma; NPC/OPC: neuroal/oligodendrocyte (Oligo) progenitors cells; EMT: epithelial-mesenchymal transition; MES: mesenchymal. We applied PASTA on predicting the MPs. PASTA successfully connects almost all the basal cells to all the MPs. Here we presented the top 5 enriched MPs in each cell type. Each dot indicates a significant pathway enrichment in a particular cell type, with the size of the dot reflecting the significance level: the larger the dot, the more significant the enrichment.

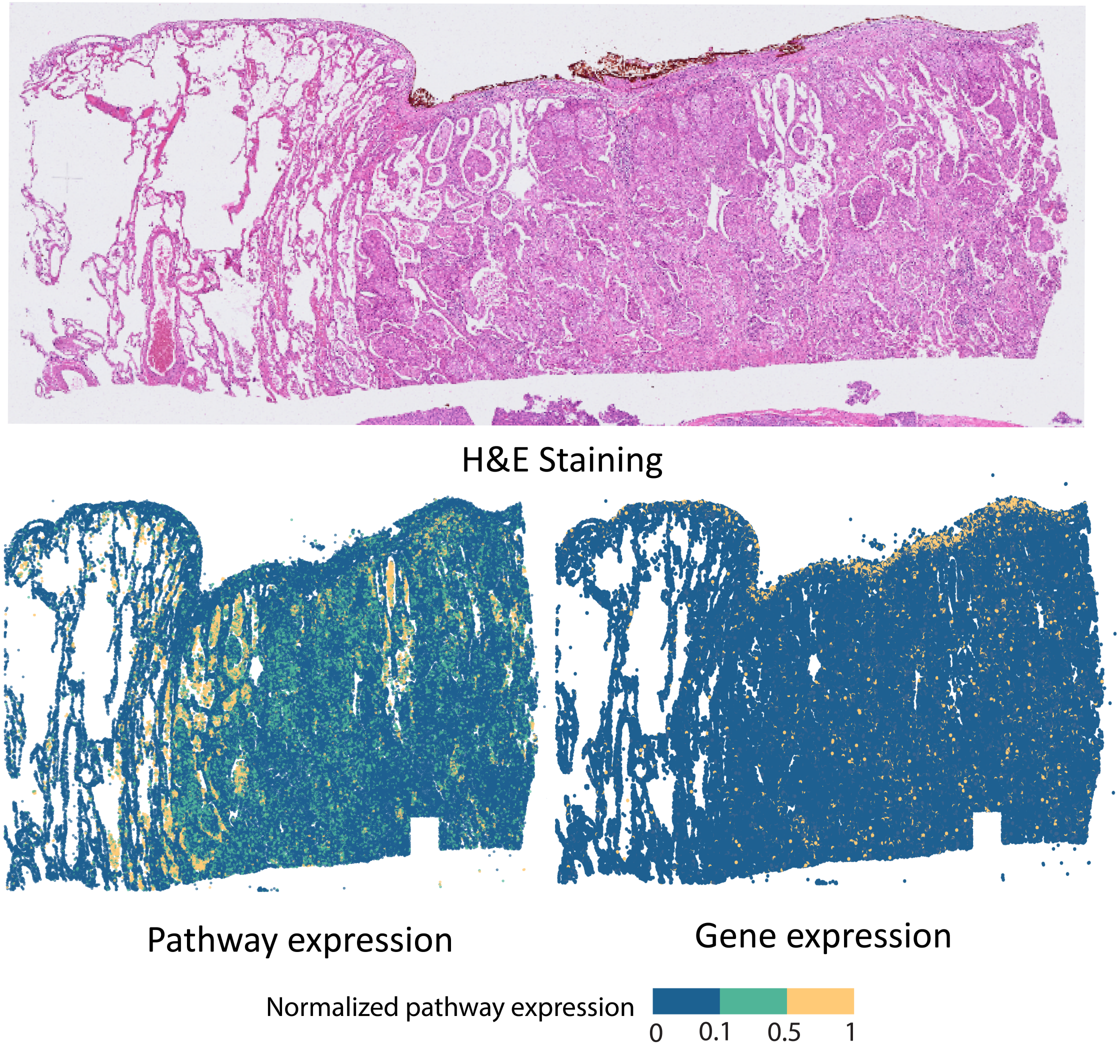

Figure S5: **HE staining of the human lung tissue, predicted pathway, and gene expression figures.** The pathway is collected from HALLMARK gene sets and the gene is RETN. The pathway and gene are both related to the tumor cell expression. Single gene expression can fail on identify the tissue structure, while pathway that aggregates genes with similar function can provide more information about the tissue structure.

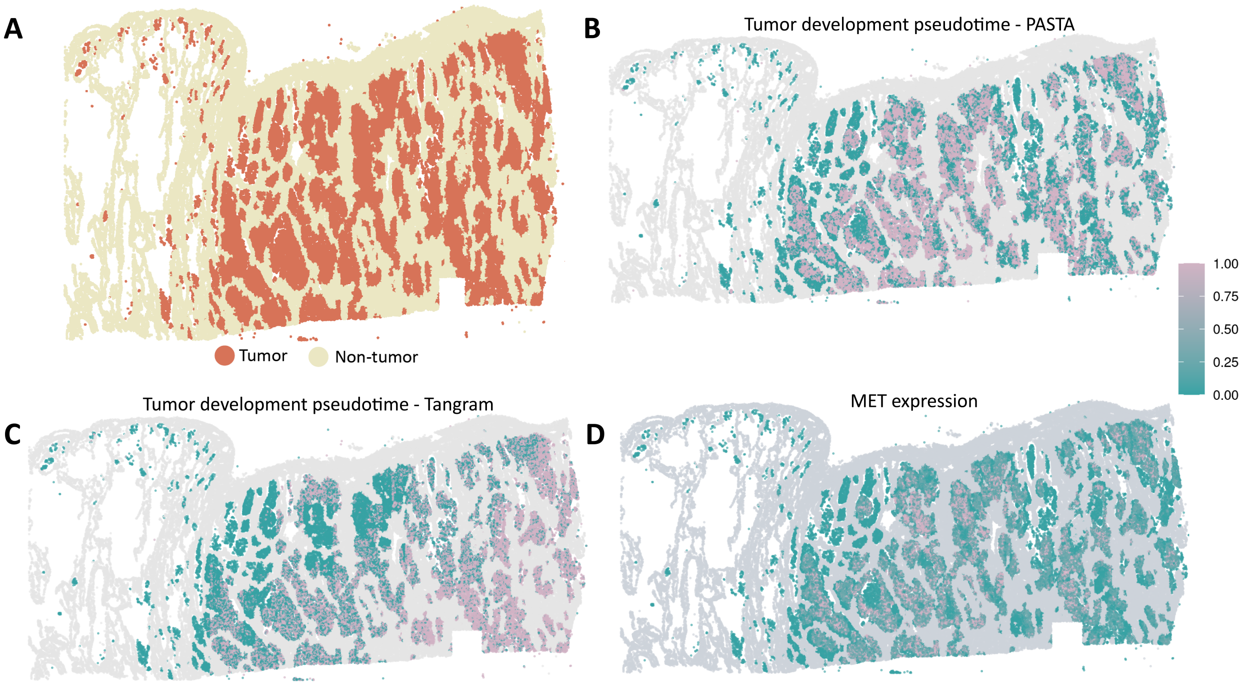

Figure S6: **Trajectory analysis of tumor development.** A. Domain detected by IRIS. One domain contains mainly basal cells and we identify it as tumor domain. B. Pseudotime of tumor development predicted using the pathway-by-cell predicted by PASTA. C. Pseudotime of tumor development predicted using the pathway-by-cell predicted by Tangram. D. Gene expression of MET. MET gene is related to tumor development. The pseudotime estimation from PASTA and Tangram showed different patterns: Our method revealed that more developed tumor cells were concentrated in the center, whereas Tangram's predictions placed them at the right edge. We examined the expression of the MET gene, which is associated with abnormal cell proliferation. The distribution of the gene expression aligned closely with our method's predictions, demonstrating that our approach effectively captures the underlying biological development among the tumor cells.

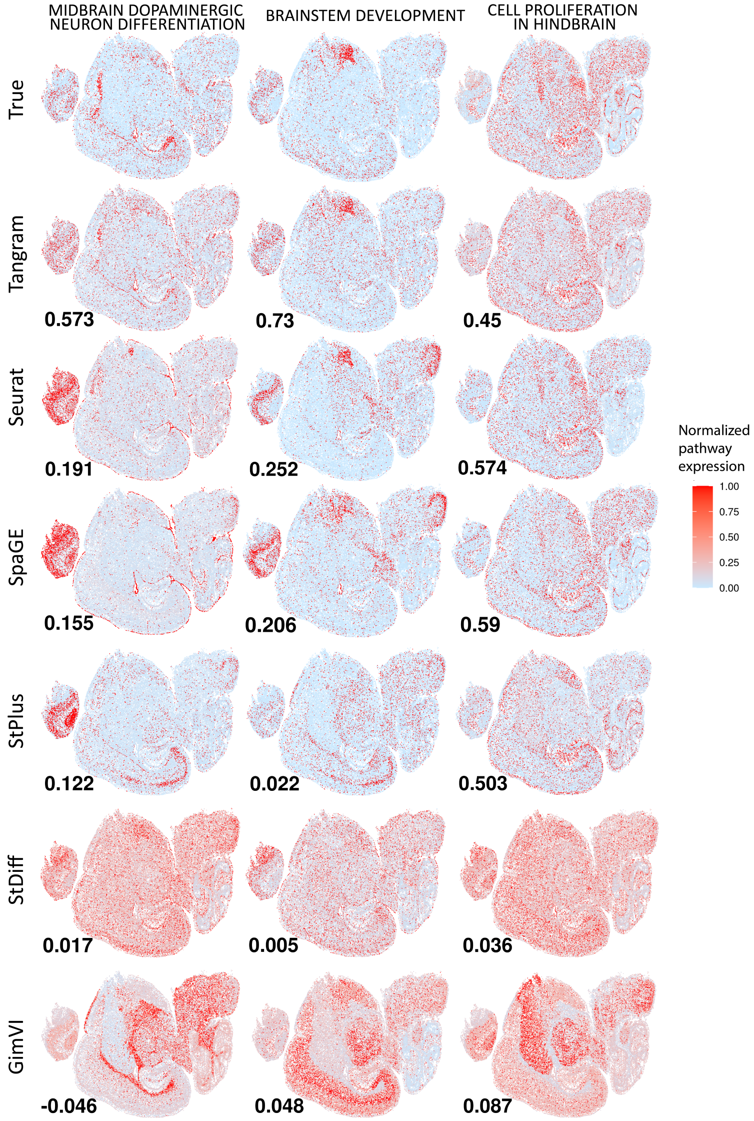

Figure S7: **Predicted pathway expressions of the Merfish mouse brain data.** The ST dataset contains 82075 cells with a panel of 1122 genes. The cell types were identified by the authors manually, including mostly neurons. The corresponding mouse primary visual cortex scRNA-seq data is collected from the Allen Institute. The three pathways are collected from GOBP database. Pearson correlations are shown in the figure between the prediction and the ground truth. The existing method did not show ideal performance on predicting the pathway expression, indicating by the low correlations.

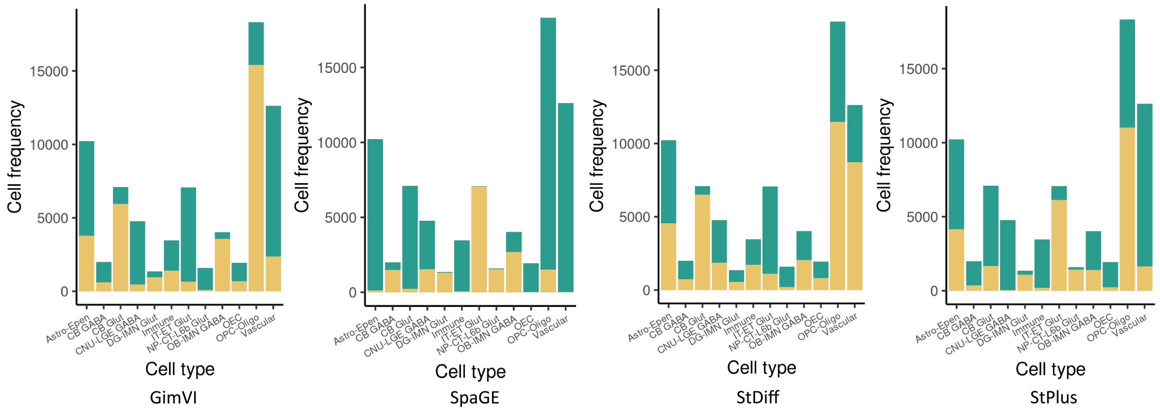

Figure S8: **Cell distributions in the two K-means clusters.** We performed the K-means clusters using the predicted pathway-by-cell matrix, setting the $K = 2$. The pathways were collected from GOBP database, which are related to the brain development. Distributions of major cells are presented. Except for PASTA, the imputation methods all generated cell mixtures in the two clusters. Astro-Epen, astrocyte and ependymal-like cells; GABA, Gamma-aminobutyric acid; Glut, glucose transporters; CB, cerebellum; IT-ET, inhibitory-excitatory; IMN, immature neurons; DG, dentate gyrus; OEC, olfactory ensheathing cell; OPC, oligodendrocyte progenitor cells.

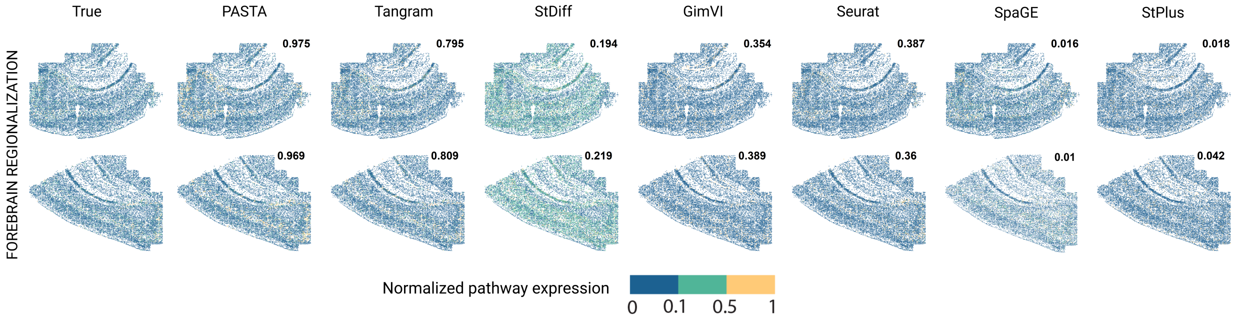

Figure S9: Predicted FOREBRAIN REGIONALIZATION pathway expression in the ISS mouse ST data. The pathway is from GOBP database. We applied the imputation methods on two different slices of the ISS mouse ST data. The two datasets contain 136 genes profiled from 14,066 and 15,823 cells, respectively. We clustered the dataset using Seurat and annotated each cluster manually referring to the PanglaoDB. During the training process, we did not incorporate cell type information for PASTA, as the majority of the cells were neurons that are distributed across the spatial domain. PASTA still generates a high correlation compared to the other methods.

References

O. Franzén, L.-M. Gan, and J. L. Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. Database, 2019:baz046, 2019.

A. Gavish, M. Tyler, A. C. Greenwald, R. Hoefflin, D. Simkin, R. Tschernichovsky, N. Galili Darnell, E. Somech, C. Barbolin, T. Antman, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. Nature, 618(7965):598–606, 2023.

A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, S. R. Williams, M. Rouault, G. Beliakoff, C. A. Morrison, M. F. Oliveira, J. T. Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nature communications, 14(1):8353, 2023.