MAPLE: A Hybrid Framework for Multi-Sample Spatial Transcriptomics Data

Carter Allen[1,2], Yuzhou Chang[1,2], Qin Ma[1,2], and Dongjun Chung[1,2*]

[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH, U.S.A.
[2] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA.

*email: chung.911@osu.edu

Abstract

High throughput spatial transcriptomics (HST) technologies have allowed for identification of distinct cell sub-populations in tissue samples, i.e., tissue architecture identification. However, existing methods do not allow for simultaneous analysis of multiple HST samples. Moreover, standard tissue architecture identification approaches do not provide uncertainty measures. Finally, no existing frameworks have integrated deep learning with Bayesian statistical models for HST data analyses. To address these gaps, we developed MAPLE: a hybrid deep learning and Bayesian modeling framework for detection of spatially informed cell spot sub-populations, uncertainty quantification, and inference of group effects in multi-sample HST experiments. MAPLE includes an embedded regression model to explain cell sub-population abundance in terms of available covariates such as treatment group, disease status, or tissue region. We demonstrate the capability of MAPLE to achieve accurate, comprehensive, and interpretable tissue architecture inference through four case studies that spanned a variety of organs in both humans and animal models.

Availability: An `R` package `maple` is available at `https://github.com/carter-allen/maple`.
Contact: chung.911@osu.edu
Supplementary information: Supplementary data are available online.

Key Words: Spatial transcriptomics; Multi-sample; Tissue architecture identification; Uncertainty quantification; Conditionally autoregressive models; Mixture models; Graph neural networks; Bayesian models

1 Introduction

Spatial transcriptomics was named the 2020 Nature Methods method of the year for its unprecedented ability to characterize transcriptomic data and infer the positional context of cells in a tissue [Marx, 2021]. A recent review has pointed to an urgent need for the improvement of tissue architecture identification algorithms through the use of the spatial location of cells within a tissue sample, in addition to gene expression profiles [Rao et al., 2021]. This critical need follows from the known importance of spatial proximity to cell fate [Barresi and Gilbert, 2019]. Of the available spatial transcriptomics platforms, *high throughput spatial transcriptomics* (HST) technologies, e.g., the 10X Visium platform, showcased their ability to offer transcriptome-wide sequencing with widespread commercial availability.

Often, it is of great biological interest to compare the relative gene expression abundance of cell spot sub-populations between different conditions (e.g., knock-out vs. wild type) or groups (treatment responders vs. non-responders). However, in the context of HST, these approaches are non-trivial due to the irreconcilable differences in spatial architecture across samples. Further, while a variety of methods have been proposed for cell spot sub-population identification in spatial transcriptomics data [Dries et al., 2019, Hao et al., 2020, Pham et al., 2020, Zhao et al., 2021, Chang et al., 2021b] there are no available methods for multi-sample analysis of HST data. Specifically, no existing methods simultaneously infer spatially-informed cell spot sub-populations in each sample while sharing information across samples.

Recently, important advancements have been made in computational approaches for two critical phases of HST data analysis, namely feature engineering and cell spot sub-population identification. With regard to feature engineering, gene expression matrices generated by HST platforms are extremely high dimensional, with roughly 30,000 unique genes being measured at several thousand cell spots in a tissue sample. This has led to the need for computational methods that derive a parsimonious set of high-information features for use in downstream analyses [Erfanian et al., 2021]. SpaGCN [Hu et al., 2021] and scGNN [Wang et al., 2021], among others, have provided deep learning-based approaches for deriving spatially informed dimension reductions of HST data. These methods each trains a spatially aware graph neural network to produce low dimension embeddings of cell spots, and have shown advantages of these spatially-aware features compared to standard non-spatial embeddings like principal components analysis (PCA) [Chang et al., 2021b, Hu et al., 2021, Wang et al., 2021].

Following a parallel development, there has been notable sophistication of computational approaches for discerning cell spot sub-populations in HST (i.e., tissue architecture identification) while considering both gene expression profiles and spatial locations of cell spots. Most notably, BayesSpace [Zhao et al., 2021] and SPRUCE [Allen et al., 2021] are Bayesian multivariate finite mixture models that distinguish cell spot sub-populations using mixture components. One critical advantage of statistical mixture model-based approaches for identifying cell spot sub-populations is that they provide a flexible framework for robust characterization of mixture component membership in terms of available metadata such as spatial information. In the case of BayesSpace, spatial information is encoded into the prior distribution for mixture component label parameters, while SPRUCE adopts a spatially correlated random effects approach to induce spatially informed mixture component assignments. However, no extension has been made from these single-sample methods to the problem of joint analysis of multiple HST samples, a setting in which it is extremely natural to use sample-specific covariates such as disease, treatment, or sex to explain cell spot sub-population abundance. Furthermore, these existing statistical models are limited in that they model either principal component reductions or normalized gene expression features, and have yet to be applied to the spatially aware deep learning features discussed previously.

To address these gaps while leveraging recent advances in HST data analysis methodology, we developed MAPLE (**M**ulti-s**A**mple s**P**atia**L** transcriptomics mod**E**l): a hybrid machine learning and Bayesian statistical modeling framework for multi-sample spatial transcriptomics data. MAPLE represents a number of marked advantages over existing computational methods for HST data analysis. First, and most importantly, MAPLE is the first framework developed explicitly for the simultaneous analysis of multiple HST samples. It includes critical multi-sample design considerations such as information sharing *across* samples to aid in parameter estimation, accommodation of spatial correlation in gene expression patterns *within* samples, and an integrated robust multinomial regression model to explain differences between samples in cell spot sub-population composition using available covariates. Second, MAPLE is the first computational framework to leverage the benefits of both deep learning and statistical modeling in HST data analysis, wherein a graph neural network is used to derive a low-dimension set of spatially aware gene expression features, and a Bayesian finite mixture model is fit to these features for robust and interpretable identification of cell spot sub-populations. Finally, MAPLE accompanies cell spot sub-population labels with uncertainty measures defined in terms of posterior probabilities from the Bayesian finite mixture model, which can be used to characterize ambiguous cell spot sub-population boundaries and discern between high and low confidence assignments.

2 Results

2.1 MAPLE offers novel methodology and interactive software for multi-sample HST analysis

We developed MAPLE as a hybrid framework that includes (i) a graph neural network for extracting informative features from multi-sample spatial transcriptomics data, and (ii) a Bayesian finite mixture model for detection of spatially informed cell spot sub-populations and comparison of cell spot sub-population composition between samples while adjusting for possible confounding factors in multi-sample experimental designs. A user-friendly `R` package `maple` for identification of cell spot sub-populations in HST data is freely available through GitHub (https://github.com/carter-allen/maple). The `maple` package seamlessly integrates with standard Seurat [Hao et al., 2020] workflows through a unified modeling interface and interactive visualization functions.

As shown in Figure 1, MAPLE accepts multi-sample HST data input in the form of an integrated `Seurat` data object, where batch correction and adjustment for technical artifacts such as sequencing depth can be accomplished using standard approaches [Hao et al., 2020, Hafemeister and Satija, 2019, Korsunsky et al., 2019]. Users may then pass data to scGNN [Chang et al., 2021b, Wang et al., 2021] to compute low-dimensional cell spot embeddings from raw gene expression data using a spatially aware graph neural network, or use other standard dimension reduction methods such as PCA. Given the resultant low-dimension cell spot embedding, MAPLE then implements a spatial Bayesian finite mixture model [Frühwirth-Schnatter and Pyne, 2010] as detailed in Materials and Methods. Briefly, MAPLE assumes the presence of K cell spot sub-populations (i.e., model mixture components) across the integrated sample. To assign cell spots to sub-populations, MAPLE assumes that the low-dimensional gene expression embeddings for each cell spot follow a multivariate normal distribution with sub-population-specific parameters and spatial correlation in gene expression profiles between spots accounted for through the use of spatially-correlated random effects [Besag, 1974]. MAPLE then iteratively estimates the parameters of each sub-population-specific multivariate normal distribution and assigns cell spots to their most probable sub-population given these estimated parameters. In addition to discrete sub-population labels, MAPLE provides continuous uncertainty measures defined using Bayesian posterior probabilities that reflect the model's confidence in the inferred sub-population for each cell spot. Similarly,

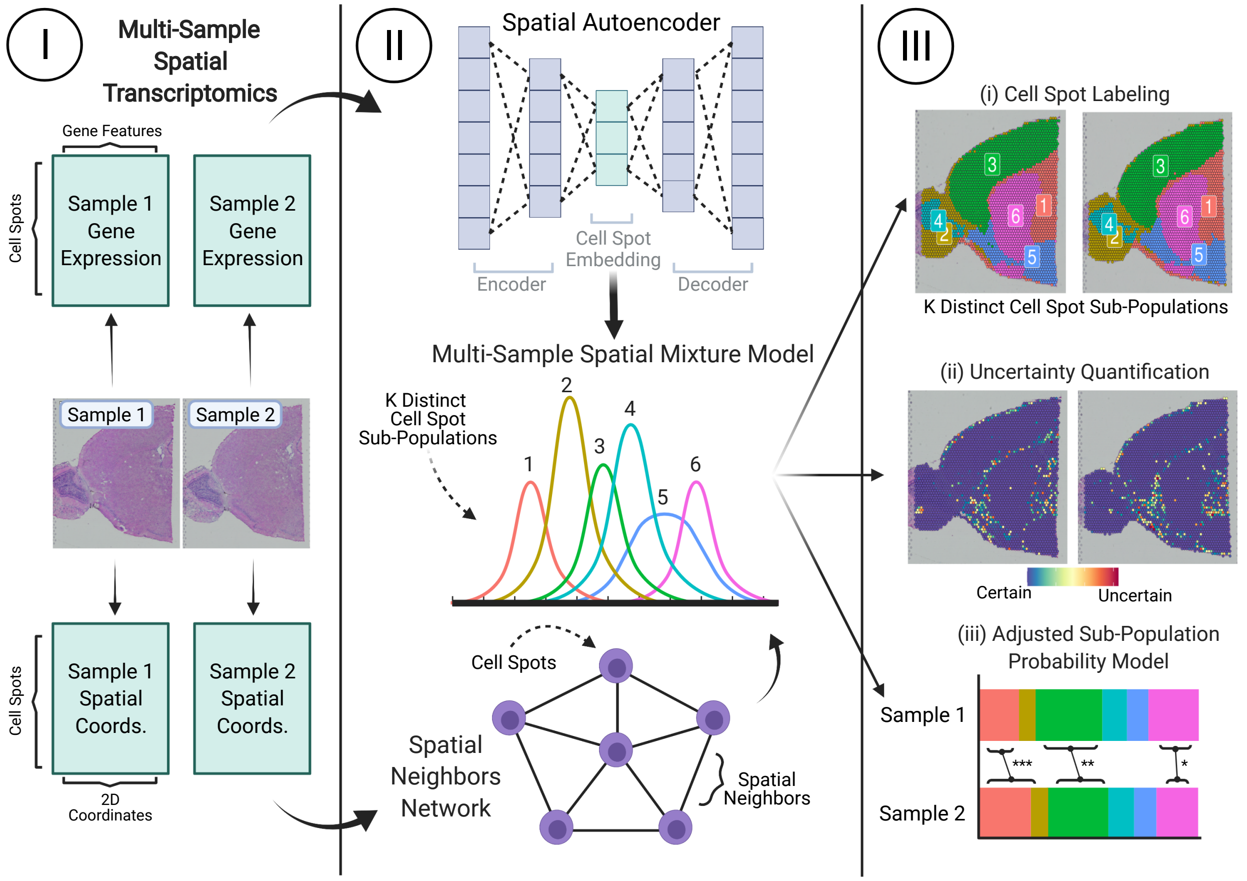

Figure 1: MAPLE workflow. Panel I: multi-sample HST data yields sample-specific raw gene expression matrices and associated spatial coordinates. Panel II: gene expression data is fed to a spatially aware autoencoder graph neural network to produce a low-dimensional embedding of cell spots. Spatial coordinates are used to construct a neighbors-network between cell spots within each tissue sample. Data is then passed to a Bayesian finite mixture model that allows for information sharing *between* samples while only considering spatial correlation *within* samples. Panel III: parameter estimates from the Bayesian finite mixture model are used for annotating cell spots with sub-population labels, quantifying associated uncertainty of inferred labels, and assessing significant differences between samples or groups of samples in cell spot sub-population abundances.

MAPLE provides continuous phenotypes, i.e., continuous measures of propensity for each cell spot towards each sub-population. Taken together, these uncertainty measures and continuous phenotypes augment traditional discrete cell spot labels and more closely reflect the continuous nature of cell type differentiation.

Embedded in the MAPLE framework is an explanatory regression model to assess the effect of covariates of interest such as treatment group, disease status, or sex on sample-specific cell spot sub-population membership probabilities. In addition to accounting for spatial correlation in sub-population labels, this model allows for comparison of sub-population abundance between samples while controlling for possible confounding factors such as the size of a given sub-population. By estimating model parameters in a Bayesian framework, we may rigorously assess possible differences between samples in terms of posterior probabilities.

To aid in usability, we implement the Bayesian model in an interactive `R` package called `maple`. The `maple` package estimates model parameters using efficient Gibbs sampling routines implemented in `C++` using `Rcpp` [Eddelbuettel and François, 2011], and interactively visualizes the resultant tissue architecture using the Shiny framework [Chang et al., 2021a]. Run-time for a typical HST data analysis with roughly 5,000 cell spots requires approximately 1 minute per 1,000 iterations on an M1 Apple iMac desktop with 8GB RAM, making it feasible to analyze HST data on a personal computer. After parameter estimation, users may (i) interactively visualize the inferred cell spot labels and annotate cell spot sub-populations, (ii) compute and interactively visualize novel uncertainty scores that allows assessment of confidence in cell spot label, and (iii) visualize the relative changes between samples or groups of samples in cell spot sub-population abundances, while accounting for any specified covariates of interest. Taken together, `maple` offers an unprecedented combination of rigorous statistical modeling and interactive visualization capability for the field of HST data analysis.

2.2 Integrative analysis allows for improved tissue architecture detection

To demonstrate the integrative analysis available with MAPLE, we considered four sagittal mouse brain samples sequenced and made publicly available through 10X Genomics [10x Genomics, 2019]. The experimental design consisted of paired anterior-posterior samples, resulting in two sagittal anterior and two sagittal posterior samples. We integrated the samples and normalized gene expression features using standard approaches [Butler et al., 2018, Stuart et al., 2019] and embedded cell spots in a low-dimensional space using principal components analysis. We then applied MAPLE to infer tissue architecture while sharing information between samples.

In Figure 2A, we present the inferred cell spot labels obtained by MAPLE. These results illustrate one important advantage of MAPLE, namely the ability to identify cell spot sub-populations that are shared across samples. In particular, MAPLE identifies cell spot sub-populations 1 and 3 as being sub-populations that are bisected by the anterior-posterior divide of the experimental design. This provides a distinct advantage over non-integrative methods, which fail to implement information sharing across samples. In Figure 2B we present associated uncertainty measures, which quantify our confidence in the cell spot labels presented in Figure 2A. We notice that higher uncertainty often occurs (i) between bordering cell spot sub-populations, such as the border between sub-populations 3 and 6 in the anterior region, or (ii) where a "satellite" group of cell spots is located far from the majority of cell spots of the same label, such as the group of sub-population 1 cell spots located in the top half of the posterior samples and contained within a larger surrounding region of sub-population 3.

While manual ground truth annotations are not available for this data set, we compared the sub-populations identified by MAPLE with known anatomy of the mouse brain made available by the Allen Brain Atlas [Daigle et al., 2018], a reproduction of which we provide in Figure S2.

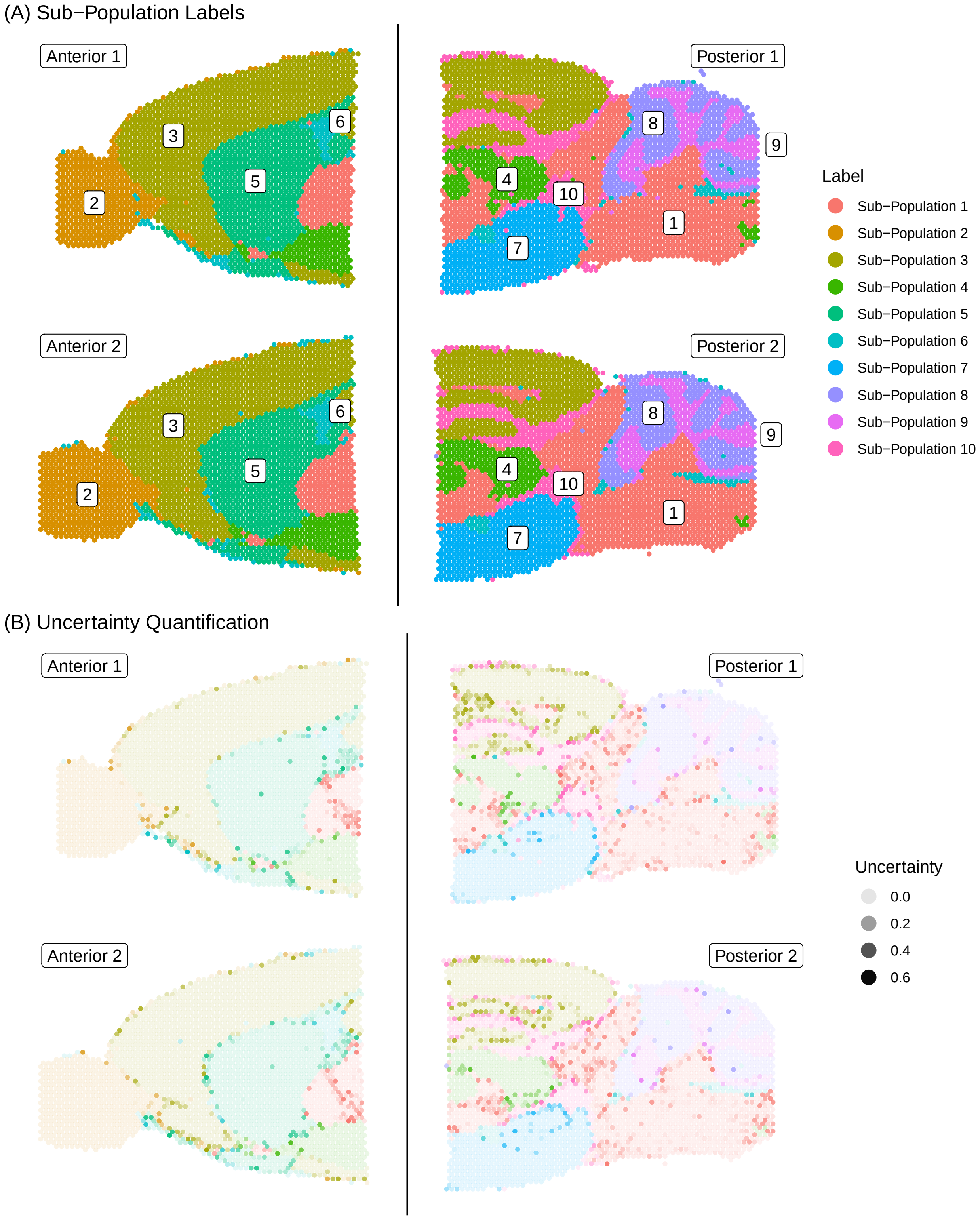

Figure 2: MAPLE results from 4-sample sagittal mouse brain analysis. Experimental design consisted of 2 healthy anterior brain sections and 2 healthy posterior brain sections. Panel A: Cell spot labels obtained by MAPLE. Panel B: Uncertainty measures for cell spot labels.

After consulting Figure S2, we may label each cell spot sub-population with relevant anatomical regions. For instance, in the anterior section, sub-population 2 found in anterior 1 and 2 samples corresponds clearly to the main olfactory bulb of the mouse brain. Likewise, in the posterior section, sub-populations 8 and 9 correspond to a region of cerebellum and grey matter. Meanwhile, more heterogeneous sub-populations were found, such as sub-population 1, which encapsulates the medulla, pons, and midbrain.

2.3 Differential analysis identifies distinct tissue architecture in ER+ vs. triple negative breast tumors

Accounting for roughly 25% of all non-dermal cancers in women, breast cancer ranks as by far the most common non-dermal female-specific cancer type, and narrowly the most common cancer type across both sexes [WCRF, 2020]. Breast cancers are commonly classified according to their estrogen and progesterone receptor status, in addition to their production of the HER2 protein signaled by the ERBB2 gene [Hammond, 2014]. While it is known that estrogen receptor positive (ER+) or progesterone receptor positive (PR+) tumors generally feature more favorable patient outcomes than triple negative tumors (TNBC) (ER-/PR-/HER2-) due to their responsiveness to hormone therapies, little is known about differences between these cancer sub-types in terms of tissue architecture [Wu et al., 2021]. While HST technology provides the opportunity to conduct spatially resolved transcriptome-wide sequencing of tumor samples, comparative analyses between sub-types has been limited by the lack of multi-sample HST analysis methods.

To address this gap, we applied MAPLE to the analysis of ER+ vs. TNBC primary tumor samples sequenced with the 10X Visium by Wu et al. [2021]. To achieve a balanced design, we considered 3 ER+ tumor sections and 3 TNBC tumor sections, totaling $N = 2187$ cell spots across all sections. We pre-processed raw cell spot RNA read-counts through normalization, embedding with scGNN, and batch correction with Harmony [Korsunsky et al., 2019]. Using annotations available from Wu et al. [2021], we identified $K = 7$ distinct cell spot sub-populations in the integrated sample (Figure 3A). Associated measures of uncertainty are visualized for each cell spot sub-population label in Figure 3B. Using visualization functions included in `maple`, we illustrate ER+ vs. TNBC tumor differences via alluvial plots (Figure 3C), allowing for comparison of the relative sub-population compositions of each cancer sub-type.

We further quantified ER+ vs. TNBC tissue architecture differences using MAPLE's embedded multinomial regression model. Briefly, to explain the propensity of cell spots towards sub-populations as a function of cancer sub-type, we specified

$$\pi_{ik} \propto b_{0k} + x_i\beta_k + \psi_{ik}, \tag{1}$$

where π_{ik} is the probability of cell spot i belonging to sub-population k, for $i = 1, ..., 2187$ and $k = 1, ..., 7$, b_{0k} is an intercept to account for heterogeneous global sub-population sizes, x_i is TNBC indicator equal to 1 if cell spot i belongs to a TNBC tumor section and 0 otherwise, β_k is a coefficient measuring the effect of TNBC status on propensity towards sub-population k relative to ER+, and ψ_{ik} is a spatially correlated random effect detailed in Section 4. Setting sub-population $k = 1$ as the reference category, we visualize box plots of the empirical posterior distributions of coefficients $b_{02}, ..., b_{07}$ in Figure 3D (i). These parameters measure the global (i.e., across all ER+ and TNBC tumor slices) differences in cell spot sub-population sizes relative to sub-population 1, which was found to encompass $n_1 = 345$ cell spots. We see that sub-populations 2 and 3 are not significantly smaller or larger than sub-population 1, while sub-population 4 was significantly smaller than sub-population 1, and sub-populations 5, 6, and 7 are significantly larger than sub-population 1. These global differences in sub-population sizes establish a baseline for comparison of sub-population abundances between ER+ and TNBC tumor slices.

To assess such differences between ER+ and TNBC tumor slices while accounting for global sub-population sizes, we display the empirical posterior distributions of coefficients $\beta_2, ..., \beta_7$ in Figure 3D (ii). These coefficients imply significantly higher abundances of sub-population 2, 5, 6, and 7 in TNBC tissue samples relative to ER+ tissue samples, while sub-population 4 was found to be significantly less represented in TNBC tumor slices relative to ER+, adjusting for baseline sub-population sizes. While the sub-populations derived from MAPLE are at first abstract entities, we may add biological annotations by investigating marker genes of each sub-population. Using the Wilcoxon Rank Sum test, we found the top 5 marker genes for each sub-population. Of particular interest were sub-populations 4 and 5. Sub-population 4 featured marker genes that have been found to be associated with tumor suppressive tendencies, such as KRT19 [Saha et al., 2018], as well as marker genes that have been associated with significantly longer patient survival such as TPT1 [Uhlén et al., 2015b] and NPY1R [Uhlén et al., 2015a]. Sub-population 4 was also significantly enriched in ER+ relative to TNBC, as seen negative estimate of the β_4 coefficient in Figure 3D (ii). Meanwhile, sub-population 5 was marked by genes that are associated with more aggressive and TNBC tumors, such as TMSB15A [Darb-Esfahani et al., 2012] and FABP7 [Liu et al., 2012]. Further, sub-population 5 was enriched in TNBC compared to ER+, as evidenced by the positive estimate of the β_5 coefficient shown in Figure 3D (ii). Overall, these results are indicative of MAPLE capturing previously documented differences between ER+ and TNBC, but through the novel lens of tissue architecture derived from HST data. Detailed results from each differential expression analysis are provided in Table S2 and Figure S4.

Based on these observations from Figure 3, sub-populations 3, 4, 5, and 6 indicated significant changes of conserved tissue architectures shared by ER+ and TNBC. Specifically, MAPLE sub-populations 3 and 4 had a higher proportion in the ER+ sample, while MAPLE sub-populations 5 and 6 occupied a higher proportion in TNBC. Therefore, we further investigated the pathological differences based on the H&E image observations from the previous study [Wu et al., 2021].

In the ER+ sample (Figure 4A), we observed spots from sub-populations 3 and 4 had a higher proportion compared to sub-population 5 and 6, and showed a complex tissue composition, including tumor, stroma, pure lymphocytes site, and tumor-infiltrating lymphocytes, indicating potential responsiveness to treatment and higher survival outcome [Wang et al., 2016, Karn et al., 2020]. On the other side, spots from sub-populations 3 and 4 showed a lower proportion in TNBC (Figure 4B), and their tissue composition lacked tumor-infiltrating lymphocytes, indicating a potential non-responsiveness to treatment and poor survival rate. Regarding sub-populations 5 and 6, the results indicated the non-invasive tumor and normal tissue were enriched in the TNBC sample, respectively. By the H&E pathological features and sub-populations, we concluded that significantly differential proportion in shared sub-populations from two cancer subsets could reflect a different tissue composition, which may contribute to capturing the region with unique pathological features of two samples.

In addition to pathological level analysis and due to low resolution of Visium spatial data, we next investigated the cell level composition information via deconvolution analysis based on the RCTD framework [Cable et al., 2021]. First, sample-matched scRNA-seq data were collected from the same data source (CID4535 for ER+ and CID44971 for TNBC) [Wu et al., 2021]. Then, we used the RCTD framework to deconvolute cell spots and obtain the cellular composition (nine major cell types in the tumor microenvironment). The results (Figure 4C) indicated that sub-population 3 in the ER+ sample had a higher T cell proportion than the TNBC sample while the T cell proportion was decreased in the same TNBC sub-population, supporting the pathological evidence. Interestingly, the sub-population 6 in ER+ tumor showed cancer epithelial dominant proportion, but this sub-population was enriched with normal epithelial cells in TNBC. The significant difference in cellular compositions from a shared sub-population further demonstrated differential proportion in

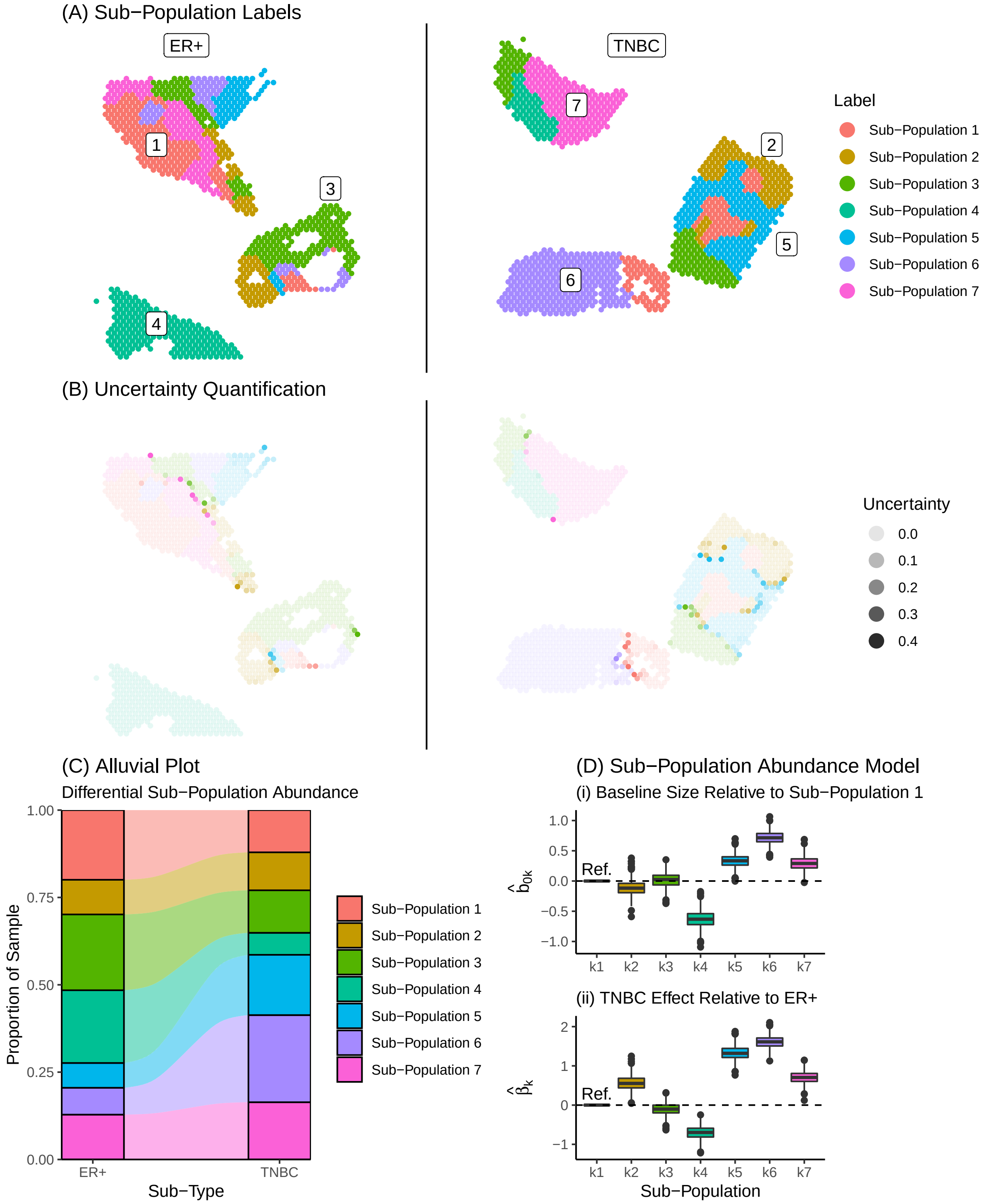

Figure 3: Results from ER+ vs. TNBC multi-sample breast tumor analysis. (A) Cell spot labels from MAPLE. (B) Uncertainty quantification. (C) Alluvial plot of differential sub-population abundance between ER+ and TNBC sample. (D) Sub-Population abundance model assessing TNBC sample effect adjusting for baseline sub-population sizes.

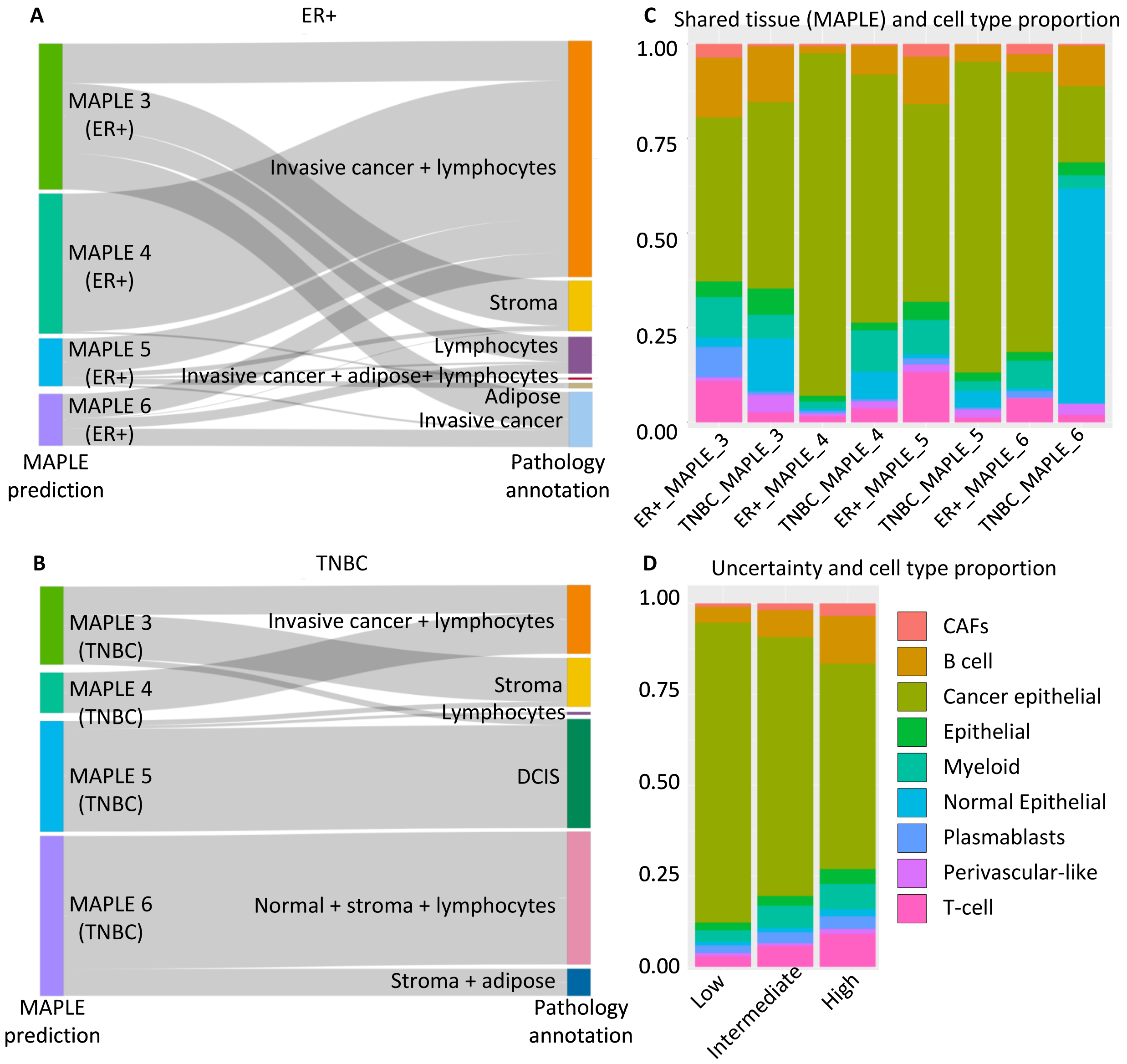

Figure 4: Biological interpretation of significantly changed shared region between ER+ and TNBC samples. (A) Alluvial plot indicating the pathological composition (derived from H&E image observation) of the shared regions in ER+ cancer, including sub-populations 3, 4, 5, and 6. For example, MAPLE cluster 3 in the ER+ sample contains a more complex component, including partial invasive cancer with infiltrating lymphocytes, pure lymphocytes, pure stroma, and adipose tissue. Node length (bars on the two sides) indicates the number of spots, meaning a longer bar representing more spots. (B) Figure showing similar information with Panel A, but the sample is TNBC. (C) Bar plot showing the cell type composition (derived from the RTCD deconvolution method) in terms of sub-populations 3, 4, 5, and 6 for ER+ and TNBC samples. (D) Bar plot indicates the relationship between uncertain measurement and cell-type compositions.

Figure 3C might reflect various cell components in the shared sub-populations of the two samples.

Lastly, we explained the biological insights of uncertainty measurement. As a result, Figure 4D indicates a spot uncertainty value responded to tissue composition diversity regarding ER+ sample. For instance, more than 80% of cell-type proportions in low uncertainty value were dominated by cancer epithelial cells. However, spots identified with a higher uncertainty value potentially indicated a more diverse cell composition, and, in this case, T cells and B cells were enriched in higher uncertainty spots compared to low and intermediate uncertainty spots. We conclude that uncertainty measurement can reflect cellular diversity. Overall, MAPLE could integrate multiple spatial transcriptomics samples from cancer, and differential proportion analysis indicated diverse regions with multiple cell types, contributing to identifying unique pathological features.

2.4 Spatiotemporal analysis reveals anatomical development trends of chicken hearts

To demonstrate the capability of MAPLE to accommodate spatially and temporally resolved HST experiments, we considered data from Mantri et al. [2020], who sequenced developing chicken hearts at four time points using the 10X Visium platform. A total of 12 heart tissue samples were sequenced, with 5 hearts sequenced on day 4, 4 hearts sequenced on day 7, 2 hearts sequenced on day 10, and 1 heart sequenced on day 14. At each time point, Mantri et al. [2020] annotated anatomical regions of the heart with a total of 10 distinct cell spot sub-populations, including prominent regions like the atria, valves, and left and right ventricles (Figure 5A). Using the proposed MAPLE framework, we integrated all 12 samples [Hao et al., 2020], performed batch correction [Korsunsky et al., 2019], and identified 10 distinct cell spot sub-populations (Figure 5B) using the top 16 batch-corrected principal components as feature inputs, where the number of sub-populations was chosen according to annotations by Mantri et al. [2020]. We visualized associated uncertainty measures derived from MAPLE in Figure 5C, which distinguished between areas of high and low confidence in the identified tissue architecture. We then tracked changes in sub-population abundance throughout the developmental window using the alluvial plot in Figure 5D.

A number of observations may be gleaned from Figure 5. Generally, MAPLE identified both horizontally and vertically distinct regions of the heart, consistent with the well known anatomical structure of the organ in both chicks and humans [Wittig and Münsterberg, 2016, 2020, Martinsen, 2005]. Notably, using only spatially-resolved transcriptomic data, MAPLE was able to accurately recover manually-labeled anatomical regions over the course of the developmental period (ARI = 0.42). Prominent regions such as the atria were identified clearly at each time point by MAPLE (sub-population 1), and were found to have consistent representation in cell spot abundance throughout the developmental period, as evidenced by the stable dynamics of sub-population 1 in Figure 5D. MAPLE also identified irregular sub-population patterns such as the epicardium cell spots present on the boundaries of each tissue sample after day 4 (sub-population 4).

2.5 Spatially aware feature engineering facilitates accurate tissue architecture detection

We sought to compare the effect of using spatially aware gene expression features generated from scGNN for tissue architecture identification versus standard spatially unaware features such as principal components (PCs). We considered the set of 16 manually annotated human brain samples analyzed in Chang et al. [2021b], and for each data set we computed multi-dimensional cell spot gene expression embeddings using PCA, SpaGCN, and scGNN, while varying the number of dimensions from 3 to 18. We quantified agreement between ground truth expert annotations and MAPLE cell spot labels using the adjusted Rand Index (ARI) [Hubert and Arabie, 1985]. In Figure 6A, we present the smoothed trends of tissue architecture identification as measured by ARI vs.

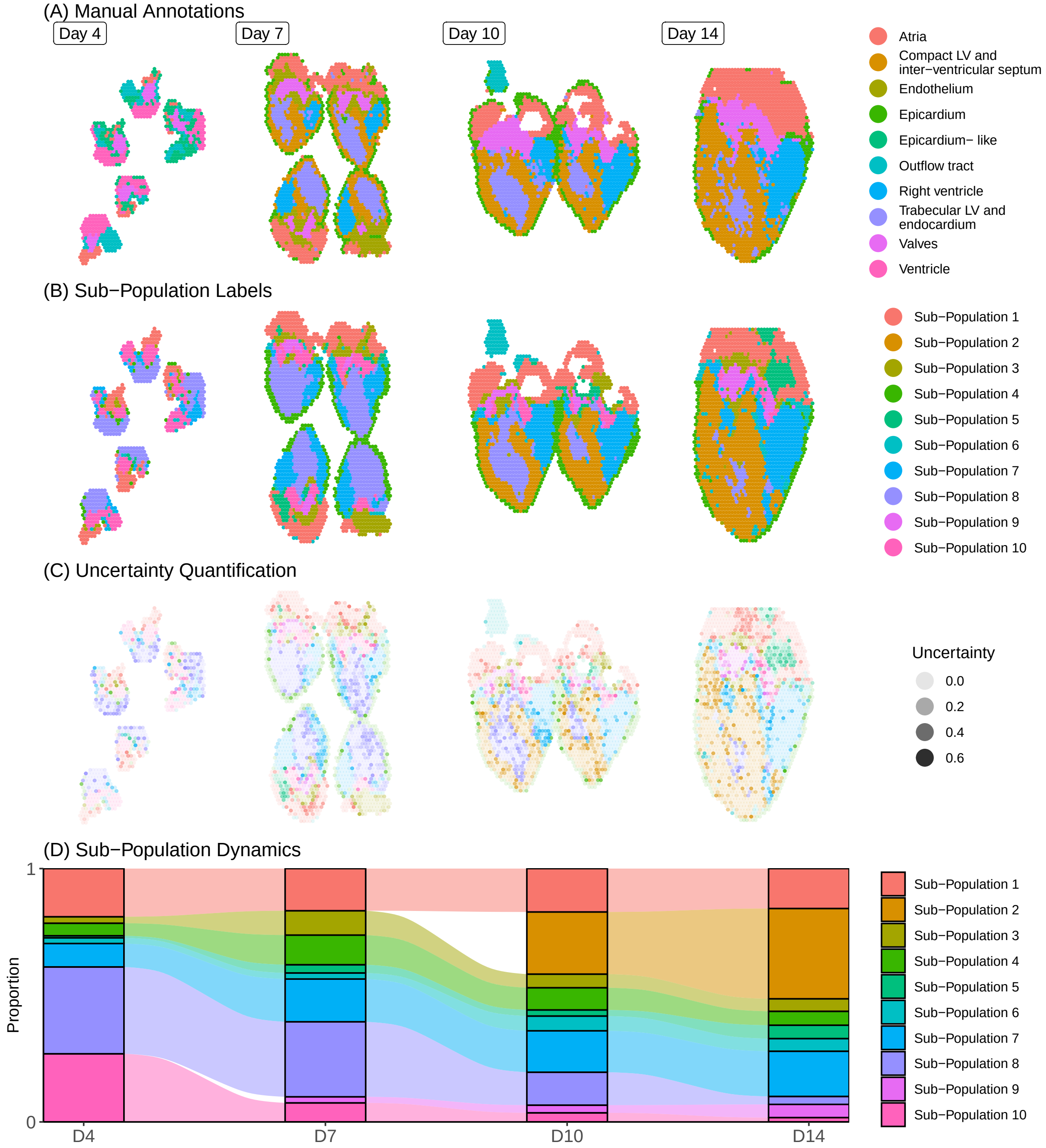

Figure 5: Results from developmental chicken heart analysis. (A) Manual anatomical annotations. (B) Sub-population labels from MAPLE. (C) Uncertainty quantification. (D) Alluvial plot of sub-population dynamics of heart tissue over time.

dimensionality of cell spot embeddings for each of the three dimension reduction methods. When smoothed over the 16 human brain data sets, we found that use of scGNN features for the MAPLE model led to the highest agreement between expert annotations of cell spot labels and cell spot labels predicted by MAPLE. This trend persisted as the dimensionality of cell spot embeddings increased from 3 to 18, although at higher dimensions (i.e., above 14), the results of each dimension reduction method converged to within the margin of error. For low-dimensional embeddings (i.e., less than 8), PC features performed significantly worse than the other two spatially aware dimension reduction methods. This result implies the ability of spatially aware feature engineering methods such as scGNN and SpaGCN to more accurately capture tissue architecture heterogeneity in low-dimensional settings compared to spatially unaware approaches such as PCA. In addition to these quantitative results, the tissue architecture identifications obtained from application of MAPLE to SpaGCN and scGNN features demonstrated a marked improvement in spatial smoothing relative to PC features: a characteristic of the ground truth expert annotations – an observation supported by the average trends of Moran's I statistic shown across all 16 data sets in Figure 6B. This trend further supports the need for spatially aware feature engineering methods for downstream HST data analysis.

To further characterize the effect of each feature engineering method on tissue architecture predictions obtained by MAPLE, we studied one particular human brain sample (sample 18-64) in detail. In Figure 6C, we plot manually annotated cell spot layers for this human brain sample. In Figures 6D-6F, we show the predicted cell spot labels from MAPLE using 3-dimensional cell spot embeddings derived from each feature engineering approach. We find that in this very low-dimensional setting, MAPLE is unable to identify any tissue architecture using PC features (ARI = 0.01). However, 3-dimensional cell spot embeddings derived from SpaGCN and scGNN lead to moderate detection of the tissue architecture patterns identified by manual annotations (ARI = 0.39, and ARI 0.45, respectively). This pattern persisted for 8-dimensional embeddings (Figures 6G-6I) and 18-dimensional embeddings (Figures 6J-6L).

3 Discussion

We have developed MAPLE: a hybrid deep learning and statistical modeling framework for identification of spatially resolved sub-populations in multi-sample HST experiments. MAPLE extends previous developments for single-sample HST analysis [Allen et al., 2021] to the multi-sample case, allowing for robust and interpretable characterization of cell spot sub-populations across multiple tissue samples. MAPLE includes a flexible embedded multinomial regression model that allows for assessment of the effect of experimental factors such as disease status or treatment effect on the relative abundance of sub-populations of interest in HST samples. While MAPLE is completely compatible with standard dimension reduction techniques such as principal components analysis, it allows for the option of using a recently developed spatially aware cell spot embedding method, scGNN, which we found to provide higher quality embeddings in highly organized tissues like the human brain. As a result, MAPLE is a one of a kind framework capable of handling a wide variety of HST analyses.

In Section 2.2, we demonstrated the advantage of multi-sample analysis with MAPLE on four sagittal mouse brain tissue samples. We found that MAPLE was not only able to recover well known mouse brain structure with 10 distinct cell spot sub-populations, but it was also able to reconstruct shared sub-populations between anterior and posterior brain tissues. Next, in Section 2.3, we applied MAPLE to the analysis of six breast tumor samples to elucidate differences between ER+ and TNBC tumors using spatially resolved transcriptomics. We identified 7 distinct cell spot sub-populations shared between three ER+ and three TNBC tumor slices. MAPLE's differential analysis framework identified an enriched sub-population of cell spots marked by genes associated

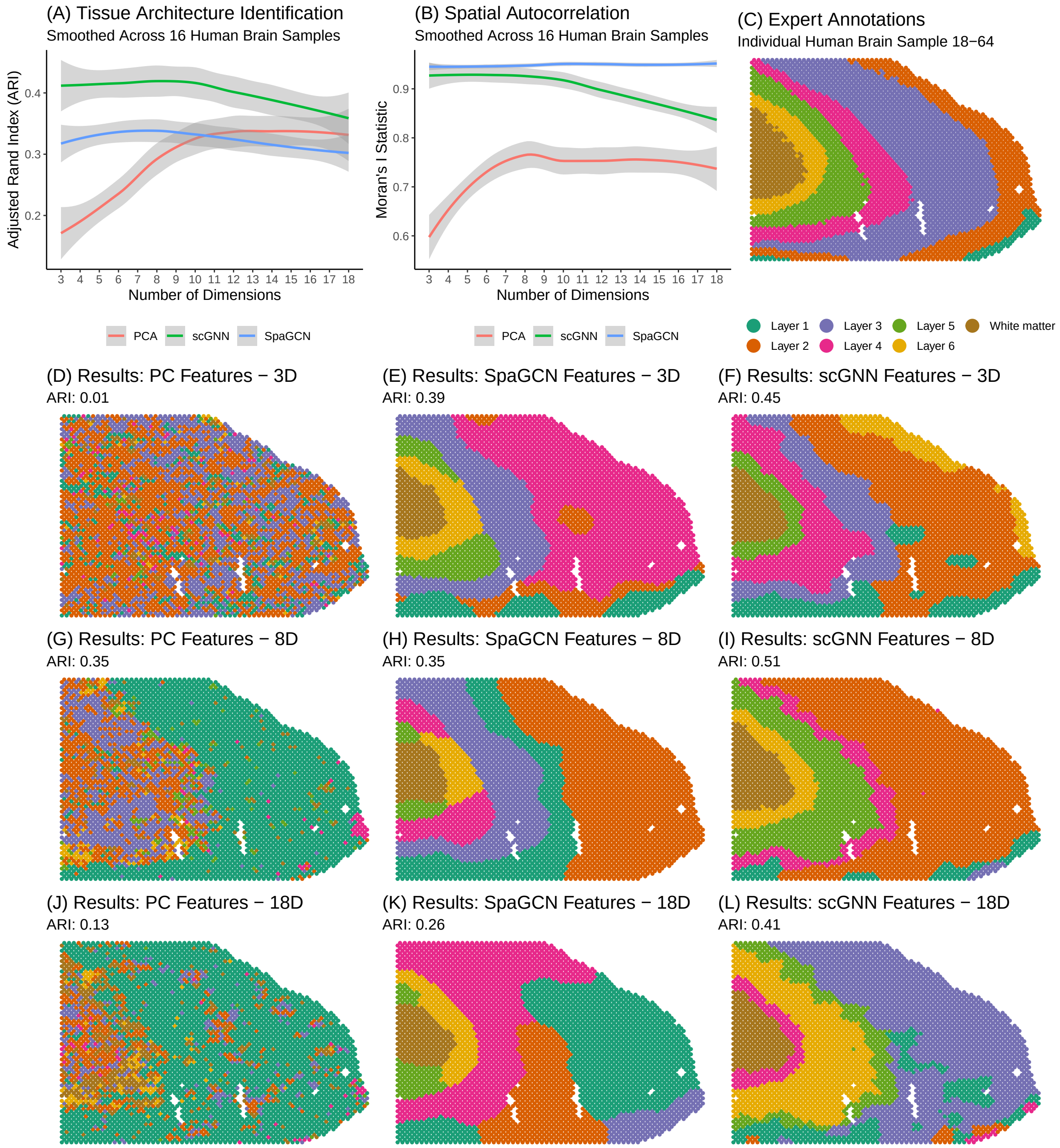

Figure 6: Results from comparison of dimension reduction techniques on recovery of expert annotations of human brain layers. Panel A: manual annotation recovery accuracy from MAPLE using each dimension reduction method applied to 16 human brain data sets for dimensionality from 3 to 18. Panel B: spatial autocorrelation in cell spot sub-population labels. Panel C: ground truth expert annotations for individual sample 18-64. Panels D-F: 3D predicted cell spot labels. Panels G-I: 8D predicted cell spot labels. Panels J-L: 18D predicted cell spot labels.

with aggressive cancer sub-types, such as TMSB15A and FABP7. In Section 2.4, we illustrated how the multi-sample analysis framework introduced by MAPLE allows for accommodation of longitudinal experimental designs, which may be especially useful to areas such as developmental biology. Finally, in Section 2.5, we showed how combining recently developed deep learning methods for cell spot embedding with the MAPLE's Bayesian finite mixture model for tissue architecture identification leads to improved performance in recovery of expert annotations across 16 human brain tissue samples.

Despite the many advantages to MAPLE outlined in this work, there are still certain drawbacks to our approach. First, MAPLE is limited by the current resolution of commercially available HST platforms such as Visium, which was the platform used for all case studies in this paper. While the notion of cell spot sub-populations are extremely useful in a variety of settings, direct inference of cell types will require true cell-level resolution HST platforms. Second, while the novel inferential capabilities such as uncertainty quantification and sub-population abundance modeling introduced by MAPLE are critical for robust HST analysis, they come at a computation price. As with any model-based method, inference with MAPLE is more computationally time consuming than common heuristic approaches such as k-means, hierarchical clustering, or graph clustering. However, model-based analysis, especially in a Bayesian framework, is often more robust and transparent than heuristic methods. For these reasons, we argue the benefits of MAPLE relative to these methods are well worth the added computational complexity. Finally, as with any multi-sample experiment, the validity and reproducibility of differential analyses across groups will depend on the number of samples in each group. As HST sequencing platforms advance to accommodate larger sequencing slides, multi-sample experimental designs should begin to include more tissue samples, increasing the validity and reproducibility of inferences obtained with MAPLE. In short, MAPLE establishes a definitive HST analysis framework that will only improve in utility along with the ongoing maturation of HST sequencing technologies.

4 Materials and Methods

Spatially Aware Feature Mining

We extract spatially aware gene expression features using scGNN: a novel graph neural network framework for single-cell RNA-Seq analyses proposed by Wang et al. [2021] for use in the context of HST data. The cell spot embedding component of scGNN consists of two phases, as depicted in Figure S1. First, cell spot coordinates and the top spatially varying gene expression features are reconciled into a single spot-spot adjacency network with homogeneous node degree of six via a positional variational autoencoder [Zhong et al., 2021]. Then, this network object is passed to multi-layered graph convolutional networks (GCNs) that are structured to create a graph autoencoder, where the focus is on reconstruction of a cell spot-cell spot similarity nearest neighbors graph using a G-dimensional learned latent embedding, with penalty functions chosen to incorporate both gene expression profiles and spatial coordinates of cell spots. For more details on scGNN refer to Wang et al. [2021].

Cell Spot Sub-Population Identification

We detect spatially informed cell sub-populations in each tissue sample using a Bayesian multivariate finite mixture model with prior distributions specified to induce correlation in mixture component assignments between neighboring cell spots within each sample. First, we let $l = 1, ..., L$ index the individual tissue samples in a given spatial transcriptomics data set, where the total number of cell spots sequenced in each sample is denoted n_l. We index each cell spot in the multi-sample data set as $i = 1, ..., N$, where the total number of cell spots present in the experiment is given by

$N = \sum_{l=1}^{L} n_l$. For each cell spot, we denote gene expression features with the length-g vector $\mathbf{y}_i$. We assume $\mathbf{y}_i$ arises from a K component finite mixture model given by

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_{ik} f(\mathbf{y}_i | \boldsymbol{\eta}_{ik}, \boldsymbol{\Sigma}_k), \tag{2}$$

where π_{ik} is a mixing weight that represents the probability of cell spot i belonging to mixture component k, and $f(\mathbf{y}_i | \boldsymbol{\eta}_{ik}, \boldsymbol{\Sigma}_k)$ denotes a g-dimensional multivariate normal density with length-g location vector $\boldsymbol{\eta}_{ik}$ and $g \times g$ variance-covariance matrix $\boldsymbol{\Sigma}_k$. To allow for spatial heterogeneity in average gene expression profiles within sub-populations, we model $\boldsymbol{\eta}_{ik}$ as

$$\boldsymbol{\eta}_{ik} = \underbrace{\boldsymbol{\mu}_k}_{\text{Cell spot sub-population k effect}} + \underbrace{\boldsymbol{\phi}_i}_{\text{Spatial effect}}, \tag{3}$$

where $\boldsymbol{\mu}_k$ is a length-g mean gene expression profile for mixture component k, and spatial autocorrelation in features among neighboring cell spots is achieved through assuming multivariate conditionally autoregressive (MCAR) priors [Besag, 1974] for the spot-specific random effects $\boldsymbol{\phi}_i$. That is, we assume

$$\boldsymbol{\phi}_i | \boldsymbol{\phi}_{-i}, \boldsymbol{\Lambda} \sim \mathrm{N}_g \left(\frac{1}{m_i} \sum_{l \in \delta_i} \boldsymbol{\phi}_l, \frac{1}{m_i} \boldsymbol{\Lambda} \right), \tag{4}$$

where $\boldsymbol{\phi}_{-i}$ denotes the spatial random effects for all spots except spot i, $\boldsymbol{\Lambda}$ is a $g \times g$ variance-covariance matrix for the elements of $\boldsymbol{\phi}_i$, m_i is the number of neighbors of spot i, and δ_i is the set of all neighboring spots to cell spot i. We enforce $\mathrm{Cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) = 0$ when spots i and j are from different tissue samples. As described in Banerjee et al. [2014], we ensure a proper posterior distribution for each $\boldsymbol{\phi}_i$ by enforcing a sum-to-zero constraint on the elements of each $\boldsymbol{\phi}_i$ for $i = 1, ..., n$. We complete a fully Bayesian model specification by assuming conjugate priors for the mixture component-specific multivariate normal model as $\boldsymbol{\mu}_k \sim \mathrm{N}_g(\boldsymbol{\mu}_{0k}, \mathbf{V}_{0k})$ and $\boldsymbol{\Sigma}_k \sim \mathrm{IW}(\nu_{0k}, \mathbf{S}_{0k})$. By default, we specify weakly-informative priors [Gelman et al., 2013] by setting $\boldsymbol{\mu}_{0k} = \mathbf{0}_{g\times 1}$, $\mathbf{V}_{0k} = \mathbf{S}_{0k} = \mathbf{I}_{g\times g}$, and $\nu_{0k} = g + 2$, which gives $E(\boldsymbol{\Sigma}_k) = \mathbf{I}_{g\times g}$.

Models (2) and (3) features a number of desirable properties in the context of multi-sample HST data analysis. First, information sharing between samples is achieved by common cell spot sub-population parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, thus supporting the utility of integrated multi-sample analysis relative to sample-specific analyses. Relatedly, in addition to inferring cell spot sub-population labels, we may characterize sub-populations using posterior estimates of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, which represent the average gene expression profile and gene-gene correlation of each sub-population, respectively. Additionally, the contribution of each observation $\mathbf{y}_i$ to each cell spot sub-population is governed by the continuous probabilities $\pi_{i1}, ..., \pi_{iK}$. These parameters allows for (i) uncertainty quantification of inferred cell spot sub-population labels and (ii) explanation of cell spot sub-population membership in terms of available covariates. Finally, we note that while we focus this work on modeling continuous features derived from feature mining approaches, our proposed framework may be extended to accommodation of normalized gene expression values directly by allowing for multivariate skew-normal mixture component densities as detailed in [Allen et al., 2021] and implemented in the R package `maple`.

To facilitate posterior inference, we introduce the latent cell spot sub-population indicators $z_1, ...z_n$, where $z_i \in \{1, ..., K\}$ denotes to which cell spot sub-population cell spot i belongs. We assume $z_i \sim \mathrm{Categorical}(\pi_{i1}, ..., \pi_{iK})$. We assign cell spots to discrete cell spot sub-populations using the maximum *a posteriori* estimates $\hat{z}_1, ..., \hat{z}_n$. However, unlike existing methods, we accompany the discrete estimates $\hat{z}_1, ..., \hat{z}_n$ with continuous uncertainty measures to account for (i) the semi-continuous nature of cell type differentiation and (ii) the statistical uncertainty inherent to cell spot sub-population identification.

Cell Spot Sub-Population Membership Models

To quantify the effect of sample-level covariates on tissue architecture, we use an embedded multinomial logit regression model for the spot-level mixture component probabilities π_{ik}:

$$\pi_{ik} = \frac{\exp(b_{0k} + \mathbf{x}_i^T \boldsymbol{\beta}_k + \psi_{ik})}{\sum_{h=1}^{K} \exp(b_{0h} + \mathbf{x}_i^T \boldsymbol{\beta}_h + \psi_{ih})} \text{ for } k = 1, ..., K, \tag{5}$$

where b_{0k} is an intercept to adjust for varying cell spot sub-population sizes, $\mathbf{x_i}$ is a p-length vector of covariates specific to spot i, $\boldsymbol{\beta}_k$ an associated p-length vector of regression coefficients for mixture component k, and ψ_{ik} is a spatial random effect allowing spatially-correlated variation with respect to $\mathbf{w}_i^T \boldsymbol{\beta}_k$. Since specification of a reference category is necessary to ensure an identifiable model formulation in terms of $K-1$ non-redundant categories, we specify mixture component 1 as the reference category and fix $b_{01} = 0$, $\boldsymbol{\beta}_1 = \mathbf{0}_{p\times 1}$, and $\psi_{i1} = 0$ for all $i = 1, ..., N$ accordingly. To introduce spatial association into the component membership model, we assume univariate intrinsic CAR priors for ψ_{ik}:

$$\psi_{ik} | \psi_{-ik}, \nu_k^2 \sim \mathrm{N}\left(\frac{1}{m_i} \sum_{l \in \delta_i} \psi_{lk}, \frac{\nu_k^2}{m_i} \right), \text{ for } k = 2, ..., K, \tag{6}$$

where ν_k^2 is a mixture component-specific variance for ψ_{ik}, and $\mathrm{Cov}(\psi_{ik}, \psi_{jk}) = 0$ for all $k = 2, ..., K$ when cell spots i and j are from different samples. This ensures that spatial correlation in mixture component assignments is not introduced between distinct spatial entities (i.e., tissue slices).

The utility of model (5) lies largely in its generalizability to arbitrary spatial transcriptomics experimental designs, where spot-level covariates $\mathbf{x}_i$ may be specified based on available metadata of a given experiment. By adopting a regression approach, we may assess the effect of experimental covariates such as treatment group or disease status, while adjusting for possible sample-specific confounders like sex, age, or batch identifiers.

$$\underbrace{\pi_{ik}}_{P(z_i = k)} \propto \underbrace{b_{0k}}_{\text{(i) Intercept}} + \underbrace{\mathbf{x}_i^T \boldsymbol{\beta}_k}_{\text{(ii) Covariate effects}} + \underbrace{\psi_{ik}}_{\text{(iii) Spatial effect}} \tag{7}$$

Intuitively, as shown above in equation (7), we construct model (5) to serve three important functions: (i) adjustment for varying cell spot sub-population sizes through the intercept b_{0k} to account for the fact that we do not necessarily wish for the probability of a randomly selected cell spot belonging to a certain cell spot sub-population to be proportional to the size of that sub-population; (ii) assessment of covariate effects or adjustment for any other confounders of cell spot sub-population membership probability through $\mathbf{x}_i^T \boldsymbol{\beta}_k$; and (iii) the introduction of spatial correlation among neighboring cell spots within the same tissue sample via ψ_{ik}. Critically, by adopting a multinomial regression approach, we avoid the pitfalls of univariate comparison of cell spot sub-population proportions across samples such as unaccounted for negative bias in cell spot sub-population proportions within samples [Buettner et al., 2020].

Uncertainty Quantification

Existing approaches for assigning cell spots to sub-populations in HST data fail to account for the inherent uncertainty in cell spot sub-population identification that may be introduced by a variety of sources, including biological, technical, or statistical factors. Biologically speaking, while the notion of discrete cell spot sub-populations is useful for describing complex tissue samples, it is known that cells move between states in a more continuous fashion than is implied by discrete clustering algorithms [Fang et al., 2018]. Technically, HST sequencing platforms are known to suffer from a number of technical limitations, including resolution and sequencing depth. The issue of resolution leads to each sequencing unit (i.e., cell spot) containing possibly more than one cell, while the sequencing depth issue refers to the non-uniform distribution of the number of genes sequenced across all sequencing units of a tissue sample. Both of these factors introduce technical noise into the gene expression profiles derived from each cell spot. Finally, statistical uncertainty results from the fact that despite their utility, all models are approximations of reality.

While addressing these sources of uncertainty is an problem for the field of spatial transcriptomics [Burgess, 2019], we argue that computational methods for cell spot sub-population identification should at least attempt to reflect these sources of uncertainty in reported cell spot sub-population assignments. To this end, we propose an *uncertainty score* defined in terms of Bayesian posterior probabilities. For each cell spot $i = 1, ..., N$, let $\hat{z}_i$ be the maximum *a posteriori* (MAP) estimate of z_i. We define u_i, the associated uncertainty score of such assignment as

$$u_i = 1 - f(\mathbf{y}_i|\hat{\boldsymbol{\eta}}_{i\hat{z}_i}, \hat{\boldsymbol{\Sigma}}_{\hat{z}_i})\hat{\pi}_{i\hat{z}_i} / \sum_{h=1}^{K} f(\mathbf{y}_i|\hat{\boldsymbol{\eta}}_{ih}, \hat{\boldsymbol{\Sigma}}_h)\hat{\pi}_{ih}, \tag{8}$$

where $f(\mathbf{y}_i|\hat{\boldsymbol{\eta}}_{i\hat{z}_i}, \hat{\boldsymbol{\Sigma}}_{\hat{z}_i})$ denotes a g-dimension multivariate normal density with mean vector $\hat{\boldsymbol{\eta}}_{i\hat{z}_i}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\hat{z}_i}$ evaluated at $\mathbf{y}_i$, and $\hat{\boldsymbol{\eta}}_{i\hat{z}_i}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{z}_i}$ are the MAP estimates of $\boldsymbol{\eta}_{ik}$ and $\boldsymbol{\Sigma}_k$ as defined in equations (3) and (2), respectively. Likewise, $\hat{\pi}_{i\hat{z}_i}$ represents the estimated propensity of cell spot i towards sub-population $\hat{z}_i$ according to the cell spot sub-population membership model defined in equation (5) evaluated using $\hat{\boldsymbol{\beta}}_{\hat{z}_i}$, the MAP estimate of $\boldsymbol{\beta}_{z_i}$. Intuitively, u_i represents the residual affinity of cell spot i towards all other cell spot sub-populations *besides* $\hat{z}_i$.

Acknowledgements

This work has been supported through grant support from the National Institute of General Medical Sciences (R01 GM122078 and R01 GM131399), National Institute on Drug Abuse (U01 DA045300), and National Institute on Aging (U54 AG075931). Additional support has been provided by awards NSF1945971 from the National Science Foundation.

Supplementary Materials

A detailed step-by-step implementation of the Gibbs sampler proposed in Section 4 is available in Supplementary Section 1. Additional figures for the case studies presented in Section 2 are provided in Supplementary Section 2. Supplementary web tables are linked from the publisher's website.

References

10x Genomics. Mouse brain serial section 1 (sagittal-anterior); spatial gene expression dataset by space ranger 1.0.0. `https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Mouse_Brain_Sagittal_Anterior`, 2019.

Carter Allen, Yuzhou Chang, Brian Neelon, Won Chang, Hang J Kim, Zihai Li, Qin Ma, and Dongjun Chung. A bayesian multivariate mixture model for spatial transcriptomics data. *bioRxiv*, 2021.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

M. J. F. Barresi and S. F. Gilbert. *Developmental Biology*, volume 12. Sinauer Associates, 2019.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Maren Buettner, Johannes Ostner, Christian L Mueller, Fabian J Theis, and Benjamin Schubert. sccoda: A bayesian model for compositional single-cell data analysis. *bioRxiv*, 2020.

Darren J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6):317–317, 2019.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, pages 1–10, 2021.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2021a. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.7.1.

Yuzhou Chang, Fei He, Juexin Wang, Shuo Chen, Jingyi Li, Jixin Liu, Yang Yu, Li Su, Anjun Ma, Carter Allen, et al. Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning. *bioRxiv*, 2021b.

Tanya L Daigle, Linda Madisen, Travis A Hage, Matthew T Valley, Ulf Knoblich, Rylan S Larsen, Marc M Takeno, Lawrence Huang, Hong Gu, Rachael Larsen, et al. A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. *Cell*, 174(2):465–480, 2018.

S Darb-Esfahani, R Kronenwett, G Von Minckwitz, C Denkert, M Gehrmann, A Rody, J Budczies, JC Brase, MK Mehta, H Bojar, et al. Thymosin beta 15a (tmsb15a) is a predictor of chemotherapy response in triple-negative breast cancer. *British journal of cancer*, 107(11):1892–1900, 2012.

Ruben Dries, Qian Zhu, Chee-Huat Linus Eng, Arpan Sarkar, Feng Bao, Rani E George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *BioRxiv*, page 701680, 2019.

Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08. URL `https://www.jstatsoft.org/v40/i08/`.

Nafiseh Erfanian, A. Ali Heydari, Pablo Ianez, Afshin Derakhshani, Mohammad Ghasemigol, Mohsen Farahpour, Saeed Nasseri, Hossein Safarpour, and Amirhossein Sahebkar. Deep learning applications in single-cell omics data analysis. *bioRxiv*, 2021. doi: 10.1101/2021.11.26.470166. URL `https://www.biorxiv.org/content/early/2021/11/27/2021.11.26.470166`.

Pu Fang, Xinyuan Li, Jin Dai, Lauren Cole, Javier Andres Camacho, Yuling Zhang, Yong Ji, Jingfeng Wang, Xiao-Feng Yang, and Hong Wang. Immune cell subset differentiation and tissue inflammation. *Journal of hematology & oncology*, 11(1):1–22, 2018.

Sylvia Frühwirth-Schnatter and Saumyadipta Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336, 2010.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.

M Elizabeth H Hammond. Hormone receptors in breast cancer: Clinical utility and guideline recommendations to improve test accuracy, 2014.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie Jane Lee, Aaron J Wilk, Charlotte Darby, Michael Zagar, et al. Integrated analysis of multimodal single-cell data. *bioRxiv*, 2020.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, pages 1–10, 2021.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Thomas Karn, Carslen Denkert, Karsten Ernst Weber, Uwe Holtrich, Claus Hanusch, BV Sinn, Brandon W Higgs, Paul Jank, Hans-Peter Sinn, Jens Huober, et al. Tumor mutational burden and immune infiltration as independent predictors of response to neoadjuvant immune checkpoint inhibition in early tnbc in geparnuevo. *Annals of Oncology*, 31(9):1216–1222, 2020.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12): 1289–1296, 2019.

Rong-Zong Liu, Kathryn Graham, Darryl D Glubrecht, Raymond Lai, John R Mackey, and Roseline Godbout. A fatty acid-binding protein 7/rxrβ pathway enhances survival and proliferation in triple-negative breast cancer. *The Journal of pathology*, 228(3): 310–321, 2012.

Madhav Mantri, Gaetano J Scuderi, Roozbeh Abedini Nassab, Michael FZ Wang, David McKellar, Jonathan T Butcher, and Iwijn De Vlaminck. Spatiotemporal single-cell rna sequencing of developing hearts reveals interplay between cellular differentiation and morphogenesis. *bioRxiv*, 2020.

Brad J Martinsen. Reference guide to the stages of chick heart embryology. *Developmental dynamics: an official publication of the American Association of Anatomists*, 233(4):1217–1237, 2005.

Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, 2021.

Duy Truong Pham, Xiao Tan, Jun Xu, Laura F Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J Ruitenberg, and Quan Hoang Nguyen. stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*, 2020.

Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596 (7871):211–220, 2021.

Subbroto Kumar Saha, Kyeongseok Kim, Gwang-Mo Yang, Hye Yeon Choi, and Ssang-Goo Cho. Cytokeratin 19 (krt19) has a role in the reprogramming of cancer stem cell-like cells to less aggressive and more drug-sensitive cells. *International journal of molecular sciences*, 19(5):1423, 2018.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220), 2015a.

Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220), 2015b.

Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1–11, 2021.

Ke Wang, Jianjun Xu, Tao Zhang, and Dan Xue. Tumor-infiltrating lymphocytes in breast cancer predict the response to chemotherapy and survival outcome: a meta-analysis. *Oncotarget*, 7(28):44288, 2016.

WCRF. Worldwide cancer data. `https://www.wcrf.org/dietandcancer/worldwide-cancer-data/`, 2020.

Johannes G Wittig and Andrea Münsterberg. The early stages of heart development: insights from chicken embryos. *Journal of cardiovascular development and disease*, 3(2):12, 2016.

Johannes G Wittig and Andrea Münsterberg. The chicken as a model organism to study heart development. *Cold Spring Harbor Perspectives in Biology*, 12(8):a037218, 2020.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021.

Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, pages 1–10, 2021.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, 18(2):176–185, 2021.