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

Hyeongseon Jeon[1,2,5], Carter Allen[1,2,5], José Antonio Ovando-Ricárdez[3], Yuzhou Chang[1,2], Lorena Rosas[3], Natalia-Del Pilar Vanegas[3], Hao Cheng[1], Juan Xie[1,2,4], Cankun Wang[1], Ana L. Mora[3,6], Mauricio Rojas[3,6], Qin Ma[1,2,6], and Dongjun Chung[1,2,6]

[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, U.S.A.
[3]Dorothy M. Davis Heart and Lung Research Institute, Division of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, The Ohio State University, Columbus, Ohio, U.S.A.
[4]The Interdisciplinary PhD program in Biostatistics, The Ohio State University, Columbus, Ohio, U.S.A.
[5]Co-first authors
[6]Co-corresponding authors

High throughput spatial transcriptomics (HST) technologies provide unprecedented opportunity to identify spatially resolved cell sub-populations in tissue samples. However, existing methods preclude joint analysis of multiple HST samples, do not allow for differential abundance analysis (DAA), and ignore uncertainty quantification. To address this, we developed MAPLE: a hybrid deep learning and Bayesian modeling framework for joint detection of spatially informed sub-populations, DAA, and uncertainty quantification. We demonstrate the capability of MAPLE to achieve these multi-sample analyses through four case studies that span a variety of organs in both humans and animal models. An `R` package `maple` is available on GitHub at `https://github.com/carter-allen/maple`.

Correspondence: ***chung.911@osu.edu***

Introduction

Spatial transcriptomics was named the 2020 Nature Methods method of the year for its unprecedented ability to characterize transcriptomic data while retaining the positional context of cells in a tissue (1). 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, in addition to gene expression profiles (2). This critical need follows from the known importance of spatial proximity to cell fate (3). 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 abundance of cell sub-populations between different conditions (e.g., knock-out vs. wild type), groups (treatment responders vs. non-responders), or throughout a temporal study window in developmental biology. Also known as *differential abundance analysis* (DAA) (4, 5), these studies can inform important biological processes such as treatment response or disease progression. 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 sub-population identification in spatial transcriptomics data (6–10), there are no formal methods available for implementing DAA in multi-sample HST data.

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 prohibitively 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 (11). SpaGCN (12), scGNN (13), RESEPT (10), and STAGATE (14), among others, have provided deep learning-based approaches for deriving spatially informed dimension reductions of HST data. Each of these methods trains a spatially aware graph neural network to produce low-dimensional embeddings of cell spots, and have shown advantages of these spatially-aware features compared to standard non-spatial embeddings like principal components analysis (PCA) (10, 12, 13).

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 (15). Notably, BayesSpace (9) and SPRUCE (16) are Bayesian multivariate finite mixture models that distinguish cell spot sub-populations using mixture components. One unique advantage of statistical mixture model-based approaches for identifying cell spot sub-populations is that they provide a flexible framework for robust characterization of sub-population 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. Despite their potential, no extension has been made from these single-sample methods to the problem of joint analysis of multiple HST samples and implementation of DAA using sample-specific covariates such as disease, treatment, or sex. Critically, multi-sample analysis cannot be done simply by combining spots from multiple samples together. Instead, these statistical models need to carefully re-

DRAFT

flect the design structure of multiple samples in the sense of information sharing and spatial correlation modeling, among others. 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 integrated with the spatially aware deep learning features discussed above. Finally, while statistical models provide a formalized framework for derivation of model-based uncertainty measures for all parameters including sub-population labels – a critical task for characterizing biological nuance in tissue architecture, especially in the case of spot-level resolution platforms like 10X Visium, neither BayesSpace nor SPRUCE utilizes their model architecture to provide uncertainty measures.

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 joint sub-population identification and implementation of DAA using multi-sample HST. 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 implement DAA. Second, MAPLE is the first computational framework to leverage the benefits of both deep learning and statistical modeling in HST data analysis via a two-stage approach, wherein a graph neural network such as scGNN, STAGATE, or any embedding method of choice is first used to derive a low-dimensional set of spatially aware gene expression features, and then a Bayesian finite mixture model is fit to these features for robust and interpretable identification of cell spot sub-populations and implementation of DAA. Finally, MAPLE accompanies cell spot sub-population labels with uncertainty measures defined in terms of posterior probabilities from the core Bayesian finite mixture model, which can be used to characterize ambiguous cell spot labels that often occur on the boundaries between spatially contiguous sub-populations.

MATERIALS AND METHODS

Spatially aware feature mining using deep learning.

While MAPLE is compatible with any dimension reduction approach, we have considered the use of spatially aware deep learning methods such as scGNN and STAGATE. 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 (17). 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 (13) and the application of scGNN on spatial transcriptomics data refer to (10). For alternative methods such as STAGATE (14) or SpaGCN (18), we refer to the original publications for detailed methodology and list the essence of their methodology design as follows. STAGATE is a graph attention auto-encoder framework by learning the similarity of neighboring spots to produce embeddings and identify spatial domains. STAGATE first constructs an undirected shared neighbor network (SNN) based on spatial information. Subsequently, it prunes the constructed SNN graph using pre-clustering results of gene expression. Lastly, the pruned SNN graph will be sent to graph attention auto-encoder to minimize the reconstruction loss of normalized expressions and output embeddings. SpaGCN is a graph convolutional deep neural network that simultaneously considers spatial location and histology information to cluster a spot and identify spatial domains. It first constructs the weighted graph based on spatial information, RGB value from the corresponding histological image, and gene expression. Specifically, a node is a spot and an edge is defined by calculating the Euclidean distance between connected two spots based on spatial information and RBG values. Next, SpaGCN uses a graph convolutional neural network to obtain embeddings which are output to the Louvain clustering algorithm for iteratively determining the spot clusters.

Joint identification of cell sub-populations in multi--sample HST data.

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

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

where π_{ik} is a mixing weight that represents the probability of cell spot i belonging to mixture component k, which may itself be modeled as a function of external covariates to achieve DAA as described in the subsequent sections, 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

DRAFT

model $\boldsymbol{\eta}_{ik}$ as

$$\boldsymbol{\eta}_{ik} = \underbrace{\boldsymbol{\mu}_k}_{\text{Cell spot sub-population k effect}} + \underbrace{\boldsymbol{\phi}_i}_{\text{Spatial effect}}, \quad (2)$$

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 (19) for the spot-specific random effects $\boldsymbol{\phi}_i$. That is, we assume

$$\boldsymbol{\phi}_i | \boldsymbol{\phi}_{-i}, \boldsymbol{\Lambda} \sim \mathrm{N}_g \left(\frac{1}{m_i} \sum_{l \in \delta_i} \boldsymbol{\phi}_l, \frac{1}{m_i} \boldsymbol{\Lambda} \right), \quad (3)$$

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 (20), 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 $\boldsymbol{\Sigma}_k \sim \mathrm{IW}(\nu_{0k}, \mathbf{S}_{0k})$. Further, we specify $\boldsymbol{\Lambda} \sim \mathrm{IW}(\lambda_0, \mathbf{D}_0)$. By default, we specify weakly-informative priors (21) 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 $E(\boldsymbol{\Sigma}_k) = \mathbf{I}_{g \times g}$, where $\mathbf{I}_{g \times g}$ is the $g \times g$ identity matrix. Similarly, we set $\lambda_0 = \lambda_{0k} = g + 2$ and $\mathbf{D}_0 = \mathbf{D}_{0k} = \mathbf{I}_{g \times g}$.

Models (1) and (2) feature 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 (16) 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 these discrete parameter estimates 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.

Differential abundance analysis (DAA) regression model.

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

$$\pi_{ik} = \frac{\exp(b_{0k} + \mathbf{x}_i^T \boldsymbol{\beta}_k + \psi_{ik})}{\sum_{h=1}^{K} \exp(b_{0h} + \mathbf{x}_i^T \boldsymbol{\beta}_h + \psi_{ih})} \text{ for } k = 1, ..., K, \quad (4)$$

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. Statistical significance may be assessed in a Bayesian manner by investigation of the posterior distributions of $\boldsymbol{\beta}_k$. To introduce spatial association into the component membership model, we assume univariate intrinsic CAR priors for ψ_{ik}:

$$\psi_{ik} | \psi_{-ik}, \nu_k^2 \sim \mathrm{N} \left(\frac{1}{m_i} \sum_{l \in \delta_i} \psi_{lk}, \frac{\nu_k^2}{m_i} \right), \text{ for } k = 2, ..., K, \quad (5)$$

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). We assign a conjugate multivariate normal prior for the remaining regression coefficients as $(b_{0k}, \beta_{1k}, ..., \beta_{pk}) \sim \mathrm{N}_{p+1}(\mathbf{b}_0^*, \mathbf{B}_0^*)$, where $\mathbf{b}_0^* = \mathbf{0}_{p+1}$, and $\mathbf{B}_0^* = \mathbf{I}_{(p+1) \times (p+1)}$, for $k = 2, ..., K$.

The utility of model (4) 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.

$$\underbrace{\pi_{ik}}_{P(z_i = k)} \propto \underbrace{b_{0k}}_{\text{(i) Intercept}} + \underbrace{\mathbf{x}_i^T \boldsymbol{\beta}_k}_{\text{(ii) Covariate effects}} + \underbrace{\psi_{ik}}_{\text{(iii) Spatial effect}} \quad (6)$$

Intuitively, as shown above in equation (6), we construct model (4) to serve three important functions: (i) adjustment

DRAFT

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 (22).

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 (23). 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 (24), 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

$$u_i = 1 - f(\mathbf{y}_i | \hat{\boldsymbol{\eta}}_{i\hat{z}_i}, \hat{\boldsymbol{\Sigma}}_{\hat{z}_i}) \hat{\pi}_{i\hat{z}_i} / \sum_{h=1}^{K} f(\mathbf{y}_i | \hat{\boldsymbol{\eta}}_{ih}, \hat{\boldsymbol{\Sigma}}_h) \hat{\pi}_{ih}, \quad \textbf{(7)}$$

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 (2) and (1), 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 (4) 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* $\hat{z}_i$.

Software implementation.

To aid in usability, we implement the proposed MAPLE approach in an interactive and user-friendly `R` package called `maple`, which is freely available on GitHub at https://github.com/carter-allen/maple. The `maple` package seamlessly integrates with standard Seurat (7) 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 pre-processing steps include normalization and adjustment for heterogeneous sequencing depth using sctransform (25). For batch correction of multi-sample HST experiments, we utilize the previously-developed method Harmony (26). After this pre-processing, users may then use a spatially aware embedding method of choice such as scGNN or STAGATE to compute low-dimensional cell spot embeddings from raw gene expression data using a spatially aware graph neural network (Figure S1). While these recently developed deep learning embedding methods may offer state of the art performance, MAPLE is compatible with any embedding approach, including standard methods like PCA. Given the resultant low-dimensional cell spot embedding derived using existing methods, MAPLE's unique contribution is to implement a spatial Bayesian finite mixture model (27, 28) to jointly infer sub-populations while sharing information across samples, and implementing DAA using a flexible embedded multinomial regression model. The `maple` package estimates model parameters using efficient Gibbs sampling routines implemented in `C++` using `Rcpp` (29), and interactively visualizes the resultant tissue architecture using the Shiny framework (30). 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 uncertainty measures that allows assessment of confidence in cell spot labels, 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.

RESULTS

Joint multi-sample 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 (31). The experimental design consisted of paired anterior-posterior samples, resulting in two sagittal anterior and two sagittal posterior samples. We integrated the sam-

DRAFT

ples and normalized gene expression features using standard approaches (32, 33), 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 (34), a reproduction of which we provide in Figure S2. 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. Differential analysis results of each sub-population versus all others are provided in Figure S3.

These results highlight a key benefit of joint analysis of multiple HST samples, namely that information may be shared across samples to better estimate sub-population-specific parameters. As detailed in Materials and Methods, MAPLE assumes that cell spot sub-populations are governed by a common set of parameters. Instead of identifying sub-populations in each sample separately, then comparing the inferred tissue architecture, MAPLE groups cell spots across all samples based on similar gene expression and spatial locations within each sample. As a result, not only do we leverage more data to estimate each sub-population, but sub-populations are more likely to be comparable across samples, as seen in Figure 2. This qualitative assessment of the benefits of multi-sample vs. single-sample analysis is augmented with quantitative results shown in Section "Spatially aware feature engineering facilitates accurate tissue architecture detection", for which ground truth manual annotations are available.

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 (35), 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, (35) 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 3A). Using the proposed MAPLE framework, we integrated all 12 samples (7), performed batch correction (26), and identified 10 distinct cell spot sub-populations (Figure 3B) using the top 16 batch-corrected principal components as feature inputs, where the number of sub-populations was chosen according to annotations by (35). We visualized associated uncertainty measures derived from MAPLE in Figure 3C, 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 3D.

A number of observations may be gleaned from Figure 3. 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 (36–38). 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 3D. 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). Differential analysis results of each sub-population versus all others are provided in Figure S4.

DAA 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 (39). 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 (40). 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 (41). While HST technology provides the opportunity to con-

DRAFT

duct 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 DAA of ER+ vs. TNBC primary tumor samples sequenced with the 10X Visium (41). 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. Batch effects were also minimized in these data by sequencing each group of tumor samples on a single slide (41). Using annotations available from (41), we identified $K = 7$ distinct cell spot sub-populations in the integrated sample (Figure 4A). Associated measures of uncertainty are visualized for each cell spot sub-population label in Figure 4B. Using visualization functions included in `maple`, we illustrate ER+ vs. TNBC tumor differences via alluvial plots (Figure 4C), 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

$$\pi_{ik} \propto b_{0k} + x_i\beta_k + \psi_{ik}, \quad \textbf{(8)}$$

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. Setting sub-population $k = 1$ as the reference category, we visualize box plots of the empirical posterior distributions of coefficients $b_{02}, \cdots, b_{07}$ in Figure 4D (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. Using these posterior distributions, we may assess statistical significance in a more unified framework relative to *post hoc* hypothesis tests. When estimated posterior distributions of regression coefficients do not include the null value of 0, we deem them "significant" in our DAA model. 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 4D (ii). As shown in Figure 4D (ii), estimated posterior distributions of 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 added biological annotations by investigating marker genes of each subpopulation. Using the Seurat framework, we identified the top differentially expressed genes (DEGs) for each sub-population versus all others (Figure S5). As observations from Figure 4C and 4D, sub-populations 3, 4, 5, and 6 indicated significant changes (i.e., estimated posterior distributions not containing 0) in conserved tissue architectures shared by ER+ and TNBC. For example, sub-population 3 in ER+ breast cancer showed a higher number and proportion of spots compared with TBNC (Table 1), indicating that sub-population 3 was a representative region for this tumor type. According to the previous study (42), *IGKC* was a B cell marker that shows a strong expression in ER+ breast cancer and improves the survival rate for ER+ breast cancer patients. In addition, the pathological annotation of sub-population 3 indicated heterogeneous cellular compositions, including invasive cancer, lymphocytes, and health stroma. Notably, DEGs of sub-population 3 demonstrated that the stromal cell-dependent collagen (e.g., *COL1A1*) potentially contributes to breast cancer aggression (43). HLA-B was also identified as DEGs of sub-population 3, indicating the T cell activation event (44). Overall, the DEGs of sub-population 3 can support the main tissue composition and functions for ER+ breast cancer.

Unlike sub-population 3, sub-population 5 was enriched in TNBC compared to ER+ breast cancer (Table 1), as evidenced by the positive estimate of the β_5 coefficient shown in Figure 4D (ii). The corresponding pathological feature indicates that sub-population 5 in TNBC shows a homogeneous composition and mainly contains ductal carcinoma *in situ* (DCIS). For example, *SOX11* is a classic marker of TNBC and also supports the SOX11+ DCIS functions (e.g., tumor progression and metastasis) of sub-population 5 (45). Other tumor markers can be supported by *TMSB15A* (46), *TMSB10* (47), *FABP7* (48), and *STMN1* (49). A visualization spatial expression patterns for a selection of genes is provided in Figure S6, and the description of sub-populations 4 and 5 can be found in Supplementary Section 2.

Based on these observations from Figure 4, sub-populations 3, 4, 5, and 6 indicated significant (i.e., estimated posterior distributions not containing 0) 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 (41).

In the ER+ sample (Figure 5A), 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

DRAFT

tumor-infiltrating lymphocytes, indicating potential responsiveness to treatment and higher survival outcome (50, 51). On the other side, spots from sub-populations 3 and 4 showed a lower proportion in TNBC (Figure 5B), 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.

To further investigate uncertainty quantification in HST data, we analyzed cell level composition information via deconvolution analysis based on the robust cell type decomposition RCTD framework (52). First, sample-matched scRNA-seq data were collected from the same data source (CID4535 for ER+ and CID44971 for TNBC) (41). 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 5C) 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 4C 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 5D 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.

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 (10), consisting of 12 samples from (53) and 4 samples from (54). For each data set we computed multi-dimensional cell spot gene expression embeddings using PCA, scGNN, positional variational autoencoder only (PVAE), a negative control version of scGNN where coordinates are randomly shuffled to destroy spatial structure (Shuffled), and STAGATE. For each method, we varied the number of dimensions used for the embedding from 3 to 18. We quantified agreement between ground truth expert annotations and MAPLE cell spot labels using the adjusted Rand Index (ARI) (55). In Figure 6A, we visualize the number of dimensions needed to identify tissue architecture. We find that scGNN and STAGATE were able to achieve higher average performance using fewer dimensions – reflective of the ability of these spatially aware embedding methods to capture heterogeneity more efficiently than spatially ignorant methods. To this end, we found PCA and PVAE required more dimensions, yet still did not realize the performance of the spatially aware methods.

To augment these aggregate trends, we present results from (54) sample 2-5. Figure 6B shows manual annotations of sample 2-5 from (54). Figures 6C and 6D show MAPLE labels using scGNN and STAGATE as embedding methods, respectively, where 4 dimensions are chosen for each. We find that both embedding approaches when used with MAPLE offer strong performance relative to manual labels. While STAGATE embeddings led to a higher ARI, scGNN offered lower noise tissue architecture labels. For a more detailed catalogue of the results from the 12 data sets (53), we have included visualizations of each MAPLE's performance using each embedding strategy in Figures S7-S9 for selected dimensions, and Table S1 containing the ARI from each setting applied to each of the 12 brain slices. In addition, we have characterized the ability for spatially aware embedding methods to offer increased performance relative to PCA in Figure S10.

Next, we sought to quantify the advantage of integrative multi-sample analysis with MAPLE relative to individual sample-level tissue architecture identification. We considered samples 151673, 151674, 151675, 151676 from (53). We applied MAPLE using 4-dimensional STAGATE embeddings to each sample separately (Figure 6E) and to all samples simultaneously (Figure 6F). We find that the multi-sample approach offered increased performance relative to manual annotations, while facilitating more straightforward comparison of shared sub-populations between samples. In Figure S11, we show further benchmarking results for data set 151676, including BayesSpace and SpaGCN, which did not offer comparable performance.

DISCUSSION

We have demonstrated the advantage of joint 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. This result is attributable to MAPLE's information sharing accross samples to estimate sub-population-specific param-

DRAFT

eters. The multi-sample mouse brain analysis also allowed for exploration of quantification of uncertainty in cell spot labels, which often occurs at the border between neighboring brain layers due to the spot-level resolution of the 10X Visium platform. We then 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. Next, 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 DAA framework identified an enriched sub-population of cell spots marked by genes associated with aggressive cancer sub-types, such as *TMSB15A* and *FABP7*. Statistical significance was able to be assessed via Bayesian posterior distributions of DAA model coefficients. Finally, we showed how combining recently developed spatially aware 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. These results also suggested an information saturation limit present in HST data, wherein the gained in performance from using spatially aware features relative to PCA diminishes as the number of dimensions used increases. The choice of number of dimensions used for embeddings is often *ad hoc* (56), and presents an opportunity to the field for future investigation.

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. Third, the current MAPLE framework is a two-step approach, where spatially aware dimension reduction is first implemented, and then Bayesian modeling is implemented based on this embedding. We plan to investigate other types of pipeline construction (e.g., feedback loops) for this hybrid modeling in the future. Fourth, currently information sharing across multiple samples is mainly implemented in the Bayesian modeling step while spatially aware dimension reduction is implemented separately for each sample. We plan to investigate alternative ways for information sharing in the future, e.g., sharing parameters across multiple samples in the spatially aware dimension reduction step. Fifth, while our case studies showed the superior performance and the generalizability of MAPLE, it will be of great interest to implement more in-depth retrospective studies to fully understand and validate the findings generated with MAPLE. Finally, as with any multi-sample experiment, the validity and reproducibility of differential analyses across groups will depend on the the experimental design and 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. While MAPLE is applicable to heterogeneous and homogeneous collections of tissue samples, the interpretation of differential abundance must be made with consideration for the design of the grouping variables of an experiment. In short, MAPLE establishes a flexible HST analysis framework that will only improve in utility along with the ongoing maturation of HST sequencing technologies.

CONCLUSION

We have developed MAPLE: a hybrid deep learning and statistical modeling framework for joint identification of spatially resolved sub-populations and DAA in multi-sample HST experiments. MAPLE extends previous developments for single-sample HST analysis (13, 16) 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, thereby offering the first formal implementation of DAA in HST analysis methods. While MAPLE is completely compatible with standard dimension reduction techniques such as PCA, it allows for the option of using recently developed spatially aware cell spot embedding methods, which we found to provide higher quality embeddings in highly organized tissues like the human brain. Finally, MAPLE is the first method to formally leverage statistical modeling to derive continuous uncertainty measures to accompany discrete sub-population labels. We believe that MAPLE can be the one-of-a-kind framework capable of handling a wide variety of HST analyses.

DATA AVAILABILITY

All analyses were conducted using the `R` package `maple`, which is available on GitHub at https://github.com/carter-allen/maple and compatible with Mac, Windows, and Linux platforms (GPL-2/3 license). All data analyzed in the manuscript are taken from previously published publicly available sources, and may be accessed via the included references.

DRAFT

SUPPLEMENTARY DATA

Supplementary Data are available at Online. A detailed step-by-step implementation of the Gibbs sampling algorithm is available in Supplementary Section 1. The description of sub-populations 4 and 5 from the ER+ vs. TNBC case study can be found in Supplementary Section 2. Additional figures for the real data applications are provided in Supplementary Section 3. Additional tables for the real data applications are provided in Supplementary Section 4. The differentially expressed genes and gene set enrichment analysis results (Gene Ontology terms and KEGG pathways) for each sub-population identified from the ER+ vs. TNBC case study are provided in the Supplementary Data S1.

FUNDING

This work has been supported through grant support from the National Human Genome Research Institute (R21 HG012482), the National Institute on Aging (U54 AG075931), the National Institute of General Medical Sciences (R01 GM122078 and R01 GM131399), the National Institute on Drug Abuse (U01 DA045300), the National Science Foundation (NSF1945971), and the Pelotonia Institute for Immuno-Oncology (PIIO). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

REFERENCES

1. Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature Methods*, 18 (1):9–14, 2021.
2. Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.
3. M. J. F. Barresi and S. F. Gilbert. *Developmental Biology*. Oxford University Press, New York, 12 edition, 2019.
4. Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, 2022.
5. Helena L Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D Robinson. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1):1–12, 2020.
6. Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):1–31, 2021.
7. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.
8. 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. doi: 10.1101/2020.05.31.125658.
9. 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.
10. 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. *Computational and structural biotechnology journal*, 20:4600–4617, 2022.
11. 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.
12. 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.
13. 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.
14. Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13 (1):1–12, 2022.
15. Zexian Zeng, Yawei Li, Yiming Li, and Yuan Luo. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biology*, 23(1):1–23, 2022.
16. 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. doi: 10.1101/2021.06.23.449615.
17. 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.
18. Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *bioRxiv*, 2020.
19. 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.
20. Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC press, New York, 2014.
21. Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, New York, 3 edition, 2013.
22. 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.
23. 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 and Oncology*, 11(1):1–22, 2018.
24. Darren J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6): 317–317, 2019.
25. 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.
26. 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.
27. 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.
28. Brian Neelon, Alan E Gelfand, and Marie Lynn Miranda. A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 63(5):737, 2014.
29. Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.
30. 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*, 2021. R package version 1.7.1.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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. *Nature Communications*, 12(1):1–13, 2021.
36. 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.
37. 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.
38. 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.
39. WCRF. Worldwide cancer data. https://www.wcrf.org/dietandcancer/worldwide-cancer-data/, 2020.
40. M Elizabeth H Hammond. Hormone receptors in breast cancer: Clinical utility and guideline recommendations to improve test accuracy, 2014.
41. 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.
42. Yan Mao, Qing Qu, Xiaosong Chen, Ou Huang, Jiayi Wu, and Kunwei Shen. The prognostic value of tumor-infiltrating lymphocytes in breast cancer: a systematic review and meta-analysis. *PloS one*, 11(4):e0152500, 2016.
43. Abdel Jelil Njouendou, Arnol Auvaker Zebaze Tiofack, Rovaldo Nguims Kenfack, Sidonie Noa Ananga, Esther Hortense Murielle Dina Bell, Gustave Simo, Joerg D Hoheisel, Jens T Siveke, and Smiths S Lueong. Sox2 dosage sustains tumor-promoting inflammation

DRAFT

to drive disease aggressiveness by modulating the fosl2/il6 axis. *bioRxiv*, 2022.
44. María del Mar Noblejas-López, Cristina Nieto-Jiménez, Sara Morcillo Garcia, Javier Pérez-Peña, Miriam Nuncia-Cantarero, Fernando Andrés-Pretel, Eva M Galán-Moya, Eitan Amir, Atanasio Pandiella, Balázs Győrffy, et al. Expression of mhc class i, hla-a and hla-b identifies immune-activated breast tumors with favorable outcome. *Oncoimmunology*, 8(10): e1629780, 2019.
45. Erik Oliemuller, Richard Newman, Siu Man Tsang, Shane Foo, Gareth Muirhead, Farzana Noor, Syed Haider, Iskander Aurrekoetxea-Rodríguez, Maria dM Vivanco, and Beatrice A Howard. Sox11 promotes epithelial/mesenchymal hybrid state and alters tropism of invasive breast cancer cells. *Elife*, 9:e58374, 2020.
46. 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.
47. Xin Zhang, Dong Ren, Ling Guo, Lan Wang, Shu Wu, Chuyong Lin, Liping Ye, Jinrong Zhu, Jun Li, Libing Song, et al. Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer. *Breast Cancer Research*, 19(1): 1–15, 2017.
48. 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.
49. Sayaka Obayashi, Jun Horiguchi, Toru Higuchi, Ayaka Katayama, Tadashi Handa, Bolag Altan, Tuya Bai, Pinjie Bao, Halin Bao, Takehiko Yokobori, et al. Stathmin1 expression is associated with aggressive phenotypes and cancer stem cell marker expression in breast cancer patients. *International journal of oncology*, 51(3):781–790, 2017.
50. 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.
51. 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.
52. 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.
53. Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.
54. Mor Kenigsbuch, Pierre Bost, Shahar Halevi, Yuzhou Chang, Shuo Chen, Qin Ma, Renana Hajbi, Benno Schwikowski, Bernd Bodenmiller, Hongjun Fu, et al. A shared disease-associated oligodendrocyte signature among multiple cns pathologies. *Nature neuroscience*, 25(7):876–886, 2022.
55. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985.
56. Yu Lin, Yan Wang, Yanchun Liang, Yang Yu, Jingyi Li, Qin Ma, Fei He, and Dong Xu. Sampling and ranking spatial transcriptomics data embeddings to identify tissue architecture. *Frontiers in genetics*, 13, 2022.

Tables

Table 1. The sub-population composition in the ER+ and TNBC samples. The first column shows the MAPLE labels and indicates seven shared sub-populations. The second and fourth columns are the numbers of spots correspondings to sub-populations in the ER+ and TNBC samples, respectively. The third and fifth columns are the proportions of sub-populations in the ER+ and TNBC samples, respectively. The last column shows the fold change for the proportion in ER+ over the proportion in TNBC.

Subpopulation	$ER+$	P_{ER+}	$TNBC$	P_{TNBC}	P_{ER+} / P_{TNBC}
1	205	0.20	140	0.12	1.65
2	102	0.10	126	0.11	0.91
3	223	0.22	141	0.12	1.78
4	214	0.21	73	0.06	3.31
5	73	0.07	200	0.17	0.41
6	79	0.08	289	0.25	0.31
7	132	0.13	190	0.16	0.78

Figures

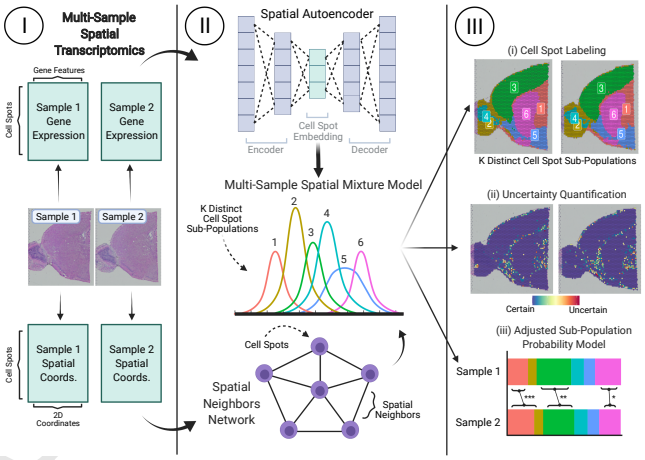

Fig. 1. MAPLE workflow. Panel I: multi-sample HST data yields sample-specific raw gene expression matrices and associated spatial coordinates. Panel II: gene expression data is fed to a spatially aware autoencoder graph neural network to produce a low-dimensional embedding of cell spots. Spatial coordinates 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, and implementing differential abundance analysis (DAA).

DR

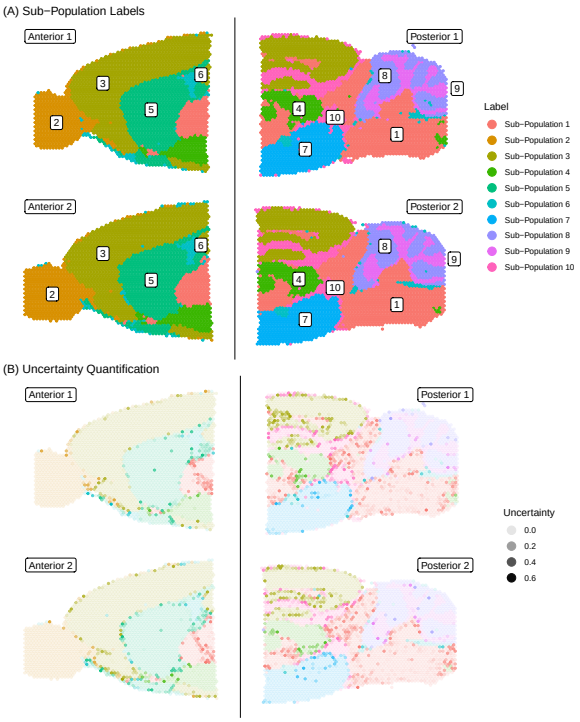

Fig. 2. MAPLE results from 4-sample sagittal mouse brain analysis. Experimental design consisted of 2 healthy anterior bras and 2 healthy posterior brain sections. Panel A: Cell spot labels obtained by MAPLE. Panel B: Uncertainty measures for cell spot labels.

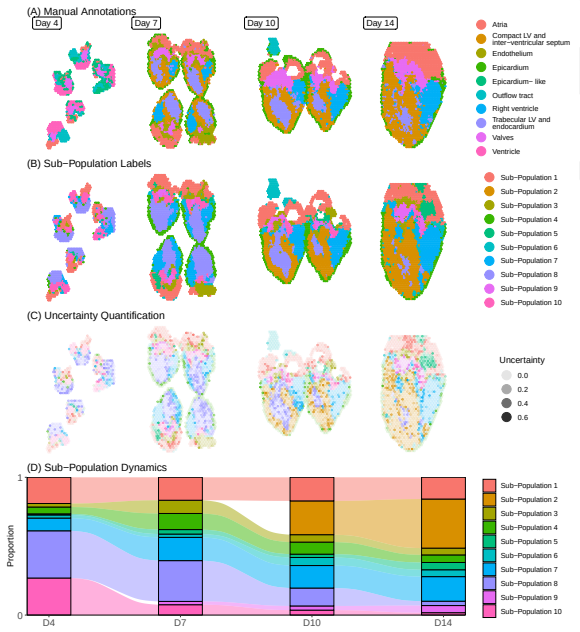

Fig. 3. Results from developmental chicken heart analysis. (A) Manual anatomical annotations. (B) Sub-population labels from MAPLE. (C) Uncertainty quantification. (D) Alluvial plot of sub-population dynamics of heart tissue over time.

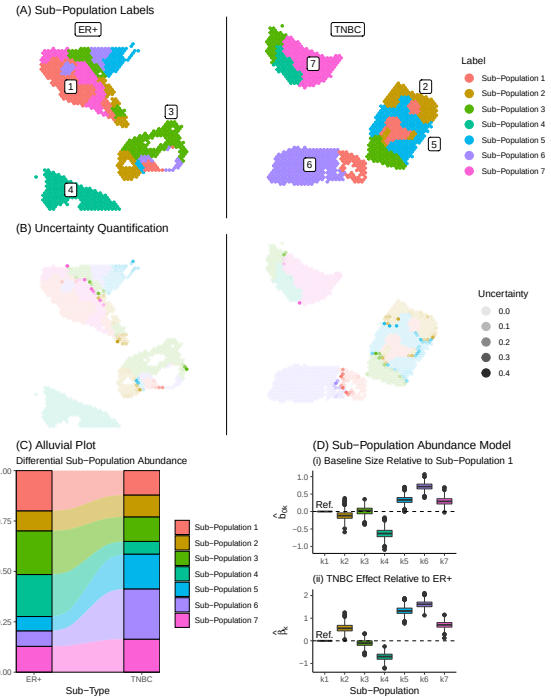

Fig. 4. Results from ER+ vs. TNBC multi-sample breast tumor analysis. (A) Cell spot labels from MAPLE. (B) Uncertainty quantification. (C) Alluvial plot of differential sub-population abundance between ER+ and TNBC sample. (D) Sub-Population abundance model assessing TNBC sample effect adjusting for baseline sub-population sizes.

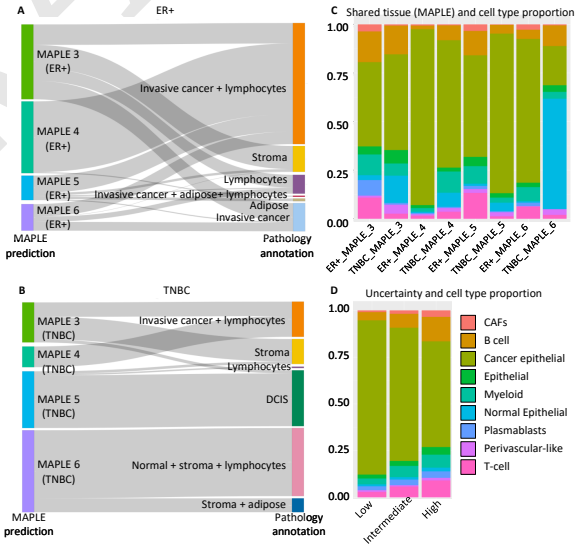

Fig. 5. Biological interpretation of significantly changed shared region between ER+ and TNBC samples. (A) Alluvial plot indicating the pathological composition (derived from H&E image observation) of the shared regions in ER+ cancer, including sub-populations 3, 4, 5, and 6. For example, MAPLE cluster 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.

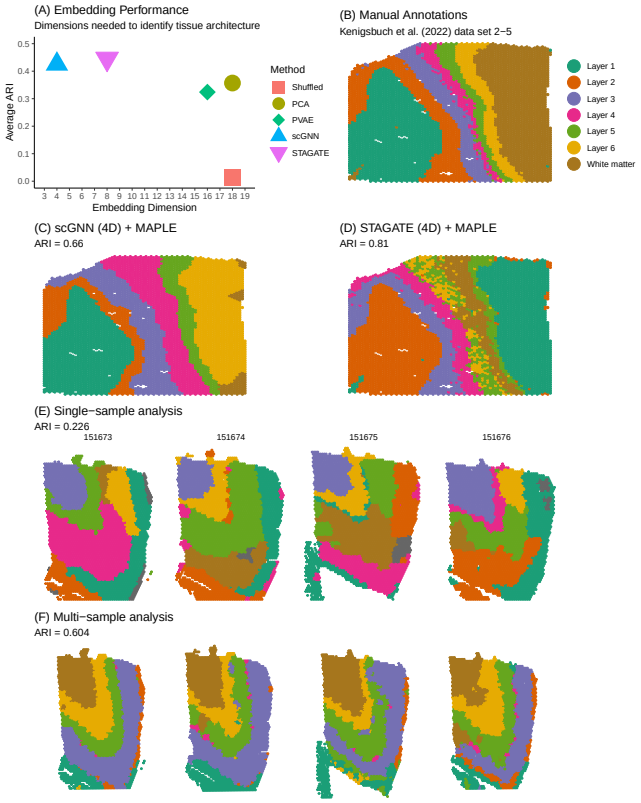

Fig. 6. Results from comparison of dimension reduction techniques on recovery of expert annotations of human brain layers using HST data from (53) and (54). (A) Comparison of the minimum number of dimensions needed to identify tissue architecture. scGNN and STAGATE use fewer dimensions and achieve higher performance using the adjusted Rand index (ARI) relative to manual annotations across 16 human brain data sets. (B) Manual annotations (54) data set 2-5. (C) MAPLE labels of (54) data set 2-5 using 4 scGNN embedding dimensions. (D) MAPLE labels of (54) data set 2-5 using 4 STAGATE embedding dimensions. (E) Single-sample analysis of 4 data sets from (53). (F) Multi-sample analysis of 4 data sets from (53) offers improved performance relative to single-sample analysis.

DRAFT