Statistical methods for the analysis of spatial gene expression data

Edward Zhao

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Raphael Gottardo, Chair

Timothy Thornton

Jonathan Wakefield

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Statistical methods for the analysis of spatial gene expression data

Edward Zhao

Chair of the Supervisory Committee:

Raphael Gottardo

Statistics

In recent years, there has been rapid development of spatial gene expression and spatial transcriptomics technologies, though corresponding advances in computational and statistical tools for the analysis of the data generated from these technologies have lagged. Initial approaches often neglected to consider differences between spatial transcriptomics and its predecessors, thus leading to analyses that may not fully realize the potential of spatial transcriptomics to generate biological insights. The overall aim of my dissertation research is to develop statistical methods for the analysis of spatial gene expression data. Specifically, I present new approaches for spatial clustering and resolution enhancement of spatial transcriptomics data as well as a joint model for spatial transcriptomics and single-cell RNA sequencing data.

Table of Contents

Chapter 1: Background

The human genetic code is stored in DNA while human phenotypes can often be described at the molecular level through the expression levels of proteins. The major pathway for information to flow from DNA to protein involves another molecule, RNA. While proteins provide the most direct view into understanding biological mechanisms, RNA has proven to be much easier to profile, first with low-throughput approaches based on Sanger sequencing in the 20th century and later with hybridization-based microarrays. Though microarrays enable researchers to profile expression at the whole-transcriptome level, they suffer from challenges in sensitivity, specificity, and reproducibility[1]. Since then, bulk RNA sequencing (RNA-seq) and later single-cell RNA sequencing (scRNAseq) technologies were developed, overcoming the limitations of microarrays to a large extent and becoming vital components of the biologist's tookit for understanding the genetic mechanisms underlying disease[2].

scRNAseq achieves high-throughput and high-resolution profiling of gene expression, but because tissue is dissociated for sample preparation, spatial information is not retained. Knowledge of the spatial location of transcript expression can provide vital insights into biological function and pathology. Recent methods for high-throughput profiling of gene expression while retaining spatial information allow analyses to be made within the context of the biological tissue[3]. Studies performed with the Spatial Transcriptomics (ST) platform and the improved Visium platform have already generated insights into diverse areas such as tumor heterogeneity[4,5], brain function[6], and the pathophysiology of sepsis[7].

Since the microarray era, statistical methods have played an important role in the analysis of transcriptomics data, from the initial quantification to the final generation of inferences[2,8]. As the field continues to develop with new technologies, datasets, and biological questions, the

development of statistical tools to meet these challenges remains an important area of research as ever. In the following chapters, I describe a series of challenges to data analysis for spatial transcriptomics data and my proposed tools for addressing them. Specifically, we aim to develop methods for clustering analysis, resolution enhancement, and integration with complementary data. More in-depth background of each challenge is presented in chapter-specific introductions.

Chapter 2: Cluster gene expression spots into biologically relevant regions

Introduction

There is a need for new statistical methods for the analysis of spatial gene expression data that efficiently use the available spatial information. Clustering is an important step in the analysis of such data that allows downstream analyses such as cell type or tissue annotation and differential expression to provide unbiased biological insights. Existing analyses of spatial gene expression data often rely on clustering methods for non-spatial scRNA-seq data[4,6]. Popular clustering methods used in scRNAseq analysis include k-means[9], Louvain[10], mclust[11], and SC3[12]. Louvain clustering involves partitioning a nearest-neighbor distance graph while optimizing modularity. K-means and mclust involve fitting a centroid for each cluster, where k-means optimizes for Euclidean distance while mclust uses the multivariate normal likelihood. The SC3 method is based on k-means but builds upon it by finding consensus between k-means partitions run on various distance metrics, ranges of dimensionality, and transformations applied to the data.

The additional spatial information available from ST and Visium is not utilized by the aforementioned clustering methods but can help address the analytical challenges of sparsity and noise by smoothing over adjacent spots, which are more likely to have similar transcriptomic profiles. Zhu et al. (2018) proposed a hidden Markov random field model (HMRF) for clustering of low-resolution *in situ* hybridization data into distinct spatial domains by jointly modeling gene expression and the spatial neighborhood structure[13]. This approach was later adapted for use with high-throughput spatial transcriptomics data through the selection of spatially differentially expressed genes prior to clustering[14]. Another recently developed spatial clustering algorithm is

stLearn, which uses deep learning features extracted from the histopathological images as well as the expression of neighboring spots to spatially smooth the data[15].

Here, we introduce BayesSpace, which enables spatial clustering by modeling a low-dimensional representation of the gene expression matrix and encouraging neighboring spots to belong to the same cluster via a spatial prior (Fig. 1A). Our method draws from previously developed spatial statistics methods for image analysis and microarray data[16,17]. Compared with previous approaches, BayesSpace allows for a more flexible specification of the clustering structure and error term than alternative approaches. From a user perspective, BayesSpace is accessible in that it takes the widely used Bioconductor SingleCellExperiment object as input[18], does not require the additional task of preselection of marker genes, and involves minimal parameter tuning.

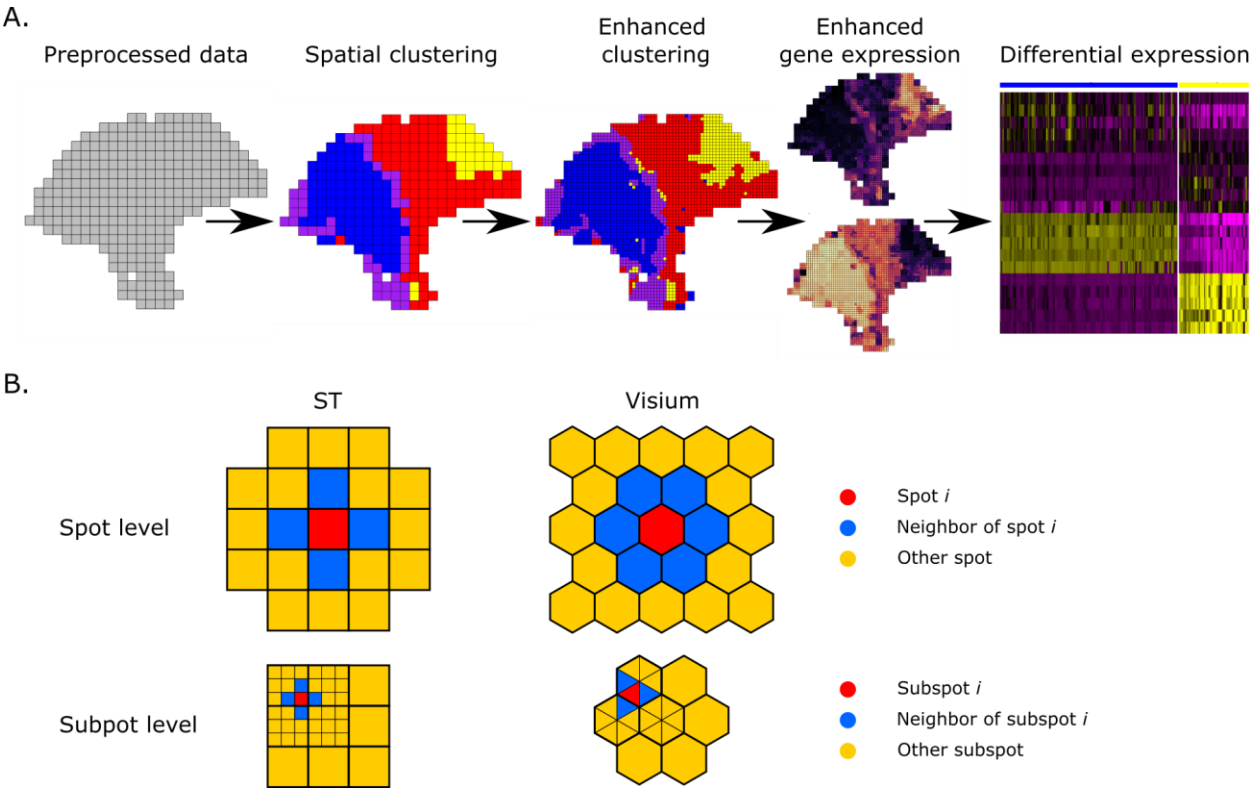


Figure 1. The BayesSpace workflow. (A) The BayesSpace workflow begins with preprocessed ST or Visium data. The data are spatially clustered to infer regions with similar expression profiles. These clusters can be refined via enhanced clustering to provide a higher resolution spatial map. Enhanced clustering also provides the basis for predicting gene expression at the higher resolution, which can be used in further differential expression analyses. (B) From geometric representations of the spatial distribution of spots in the ST and Visium technologies, neighbors can be identified for each spot based on shared edges (top). Each spot can be subdivided into subspots, which again have natural edge-based neighbors (bottom).

In both real and simulated data, we show that BayesSpace improves the identification of spatially distributed tissue domains through spatial clustering.

Methods

SPATIAL CLUSTERING MODEL

BayesSpace implements a fully Bayesian model with a Markov random field (MRF) prior to encourage spots of the same cluster to be close to one another. Such models have been previously used in image analysis[16,17]. The origins of this approach can be traced to the Ising model, which was originally developed to model atomic spins[19]. The spins can be in either the positive or negative state, and under the ferromagnetic setting, the system prefers spins (z) that are aligned between particles that are adjacent on the lattice:

$$P(z_i) \propto \exp\left(\frac{\gamma}{|\langle i\ j\rangle|} \times 2 \sum_{\langle ij\rangle} z_i z_j\right),$$

Where $z_i \in \{+1, -1\}$, $\langle i\ j\rangle$ denotes all neighbors j of particle i, and γ is a term representing the inverse temperature of the system. Applying this to image analysis, pixels can be classified as

one of two colors and adjacent pixels are encouraged to belong to the same color. The Potts model generalizes the two-color Ising model to q colors[20]:

$$P(z_i) \propto \exp\left(\frac{\gamma}{|\langle i\ j\rangle|} \times 2 \sum_{\langle i\ j\rangle} I(z_i = z_j)\right), \quad (1)$$

Where $z_i \in \{1, \ldots, q\}$. The Ising and Potts models can be further generalized to allow for an external field at each pixel i.

We adapt the q-color Potts model to solve the clustering analysis problem in spatial transcriptomics, where each node is a spot. ST and Visium spots are arranged on square and hexagonal lattices, which provide a natural way to define a neighborhood structure (Fig. 1B). For each spot i, a low d-dimensional representation $\boldsymbol{y}_i$ (e.g. PCs) of the gene expression vector can be obtained. This information forms the external field component of the Potts model. We model the expression data as follows:

$$(\boldsymbol{y}_i | z_i = k, w_i) \sim N(y_i; \mu_k, w_i^{-1}\boldsymbol{\Lambda}^{-1}).$$

Here, $z_i \in \{1, \ldots, q\}$ denotes the latent cluster that i belongs to, $\boldsymbol{\mu}_k$ the mean vector for cluster k, $\boldsymbol{\Lambda}$ precision matrix, and w_i an unknown (observation-specific) scaling factor. The number of clusters q is determined by prior biological knowledge when available or otherwise by the BIC (Supplementary Figures 9 and 10)[21].

We assume a common (fixed) precision matrix across clusters since the number of unknown parameters in the precision matrix quickly rises with higher number of clusters and number of PCs modeled. In practice, we found that the variable precision model often required strong priors for parameter estimation. We also assume that the common precision matrix is unconstrained since there is correlation between PCs after conditioning on cluster, even though

PCs are marginally uncorrelated (Supplementary Figure 20). On real data, variable and independent precision models both performed poorly relative to the unconstrained, fixed precision model.

We place the following priors on $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}$, and w_i:

$$\boldsymbol{\mu}_k \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}),$$

$$\boldsymbol{\Lambda} \overset{i.i.d.}{\sim} \text{Wishart}_d(\alpha, \text{diag}\,(\beta)_d^{-1}),$$

$$w_i \overset{i.i.d.}{\sim} \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right),$$

where $\boldsymbol{\mu}_0$, $\boldsymbol{\Lambda}_0$, α and β are fixed hyperparameters. By default, we set $\boldsymbol{\mu}_0$ to be the empirical mean vector of the data, which is the generally the zero vector for PCA input. $\boldsymbol{\Lambda}_0$ is set to 0.01 times the identity matrix to provide a weak prior that will be dominated by the data when there are spots assigned to the cluster. Similarly, we set $\alpha = 1$ and $\beta = 0.01$ to provide a weak prior for the precision matrix. We also assume $\boldsymbol{y_i}$ and w_i are independent. As such when marginalizing over w_i, our normal likelihood becomes a multivariable *t* distribution with 0 mean and covariance matrix $\frac{\nu}{\nu-2}\boldsymbol{\Lambda}^{-1}$:

$$(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, z_i = k) \sim T_\nu(\boldsymbol{\mu}_k, w_i^{-1}\boldsymbol{\Lambda}^{-1}),$$

This formulation allows us to use a simple Gibbs sampling from full conditionals for updating most of the parameters, since the observations are normally distributed when conditioning on w_i. ν denotes a fixed degrees of freedom parameter to control the heaviness of tails and is set to $\nu = 4$, which has been previously been shown to overcome the influence of outlier spots during clustering[17].

The w_i's can also be interpreted as weights; the model will estimate a small weight value for any potential outlying data value. This provides robustness against outliers that can be commonly

encountered in these types of data (Supplementary Figure 2). In the BayesSpace software package, we additionally provide an option for a simplification of the model where all weights are fixed to be $w_i = 1$, resulting in Gaussian marginal errors.

Estimation of the parameters is done using a Markov chain Monte Carlo (MCMC) method. We initialize $\boldsymbol{z}$ using a non-spatial clustering method such as mclust by default[11]. Alternative initializations can also be supplied as a label vector. Then, iteratively and sequentially, each $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}$, and w_i is updated via Gibbs sampling and each z_i is updated via Metropolis-Hastings (MH).

We update $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k$, and w_i via Gibbs sampling from the full conditionals:

$$\boldsymbol{\mu}_k \sim N\left(\left(\boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}\sum_{i:\{z_i=k\}} w_i\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \sum_{i:\{z_i=k\}} w_i\boldsymbol{y}_i\right), \left(\boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}\sum_{i:\{z_i=k\}} w_i\right)^{-1}\right)$$

$$\boldsymbol{\Lambda} \sim \text{Wishart}_d\left(n + \alpha, \left(\text{diag}(\beta)_d + \sum_i w_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right)^{-1}\right)$$

$$w_i \sim \text{Gamma}\left(\frac{d+\nu}{2}, \frac{\nu + (\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Lambda}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)}{2}\right)$$

As with standard MH, each z_i is updated by taking into account both the likelihood and spatial prior information.

Given n spots and a set cluster number q, the cluster label vector $\boldsymbol{z}$ can take values in $\{1, \ldots, q\}^n$. Using the result of the Hammersley-Clifford theorem[22], we can factorize the high-dimensional joint distribution of $\boldsymbol{z}$ into the product of conditionals $p\left(z_i | z_{\langle i\ j\rangle}\right)$, since the conditional distribution of each spot depends only on its neighbors by the Markov property of the Potts model MRF. To explore this parameter space, we use the Metropolis-Hastings algorithm. For each spot i, a new cluster label z_i' is proposed from $\{1, \ldots, q\} \backslash z_i$ and accepted or rejected based

on the ratio of posterior distributions $\frac{p(z_i'|y_i)}{p(z_i|y_i)} = \frac{p(y_i|z_i')\pi(z_i')}{p(y_i|z_i)\pi(z_i)}$, where the likelihood $p(y_i|z_i)$, representing the external field, is given by $(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}, z_i = k, w_i) \sim N(\boldsymbol{\mu}_k, w_i^{-1}\boldsymbol{\Lambda}^{-1})$ and the spatial smoothing prior is given by the Potts model as specified in (1). In this way, neighboring spots are encouraged to belong to the same cluster. Given that ST and Visium spots are arranged on a regular lattice, there is a natural way to define spatial neighbors (Fig. 1B). ST spots can have up to 4 neighbors while Visium spots can have up to 6. By default, we use $\gamma = 2$ for ST and $\gamma = 3$ for Visium, accounting for the higher number of neighbors. Model fitting diagnostics are provided in Supplementary Figures 2 and 20.

After iterating for a fixed number of iterations, the mode of the chain (after discarding a specified number of burn-in iterations) for each z_i is assigned as the cluster label for the corresponding spot i.

Updating each spot one at a time, as described above, may suffer from poor mixing due to difficulty moving between different modes. An alternative sampling scheme involves multiple spots at once using the Swendsen-Wang algorithm[23]. The algorithm introduces an additional random variable for each pair of neighboring spots, the bond variable ξ_{ij} for neighboring spots i and j:

$$\xi_{ij} = \begin{cases} 1 & \text{with probability } p_{ij} \text{ if } z_i = z_j \\ 0 & \text{otherwise} \end{cases},$$

Where $p_{ij} = 1 - \exp\left(-\frac{\gamma}{|\langle i\ j\rangle|} \times 2I(z_i = z_j)\right)$. This allows the formation of patches, where each patch is defined by a set of spots that are connected by open bonds ($\xi_{ij} = 1$). All spots in a patch P, which belong to the same cluster z_P, are then switched by Metropolis-Hastings to a new cluster z_P' with acceptance probability:

$$\frac{\exp\{\sum_{i\in P}\log p(y_i|z'_P)\}}{\exp\{\sum_{i\in P}\log p(y_i|z_P)\}}.$$

In practice, the generation of patches requires substantial computational work and the Swendsen-Wang algorithm can perform poorly in the presence of a highly informative external field, as is generally the case for the gene expression clustering problem[24].

DATA DESCRIPTION AND PREPROCESSING

We applied BayesSpace to a real biological dataset that included twelve human DLPFC samples from three individuals run on the Visium platform[6]. Briefly, each sample contained approximately 4,000 spots that were manually annotated to belong to one of the six DLPFC layers or the white matter. A SingleCellExperiment object containing the counts and spatial coordinates for all samples was downloaded using the `spatialLIBD::fetch_data()` method and then subset by sample.

Raw gene expression counts were log-transformed and normalized using library size as is commonly done for expression data[25,26]. Principal component analysis (PCA) was then performed on the top 2,000 most highly variable genes. 2,000 HVGs provided the best clustering performance in our benchmarks (Supplementary Figure 21). In downstream analyses, we modeled the top 15 principal components (PCs) from the Visium libraries, and we modeled the top 7 PCs from the sample prepared on the ST platform (melanoma). The choice to model PCs rather than the full gene expression profile allows for a more tractable probabilistic model, avoiding the need for cumbersome multivariate discrete distributions. PCs are commonly used in clustering analysis of gene expression data. Here, we recommend modeling the top 15 PCs to capture as much of the variability in the data as possible while limiting the rapid increase in space that occurs with higher dimensions, though users may choose to model a different number of PCs or HVGs using the BayesSpace R package. Modeling more than 15 PCs did not provide

substantial improvements in clustering performance but increased runtime and memory usage in our benchmarks (Supplementary Figure 21). In the melanoma sample, many of the higher PCs exhibited higher numbers of extreme outliers (Supplementary Figure 22) and significantly less variance, suggesting that they most likely represent technical variability. Since the older ST technology has lower coverage, sequencing depth and throughput, fewer PCs are necessary for modeling.

COMPARISON OF OTHER CLUSTERING METHODS

We applied the other clustering methods as follows:

- *k*-means: The base R `stats::kmeans()` function was used with default parameters.
- mclust: The "EEE" model was used to match the model used by BayesSpace. All other parameters were kept at the defaults.
- Louvain: Louvain clustering was performed with Seurat, following their suggested workflow. The shared nearest neighbors graph was constructed with `seurat::FindNeighbors()`, and clustering was done via `seurat::FindClusters()`. We tuned the resolution parameter manually to obtain the specified number of clusters for each sample.
- Giotto/HMRF: We adapted the Giotto workflow described in their online tutorial (http://spatialgiotto.rc.fas.harvard.edu/giotto.visium.brain.html). Specifically, we filtered genes expressed in fewer than 10 spots after encountering numerical issues without this filter. We used Giotto's internal functions to normalize expression, identify HVGs and perform PCA, and create a Delaunay network. The Delaunay network was created with maximum distance set to 'auto', which reproduced our defined neighborhood structure. We note that this network detection method does not always identify the correct spatial

network for Visium samples. Finally we used `Giotto::doHMRF()` to obtain spatial domains (cluster assignments) from spatially expressed genes (using k-means binarization; `Giotto::binspect()`) in the DLPFC samples and the top 15 PCs in the remaining samples. The parameter beta, which controls the strength of interaction between spots, was set to 2 for the melanoma sample and simulations, 3 for the OC sample and simulations, and 18 for the DLPFC samples. The DLPFC sample beta parameter was higher due to the higher dimensionality of the input (genes rather than PCs). We chose to use the genes rather than PCs for this dataset due to the poor performance of Giotto on PCs here. The results using PCs are also run using beta = 3 and shown in Supplementary Figure 27.

- stLearn: stLearn was applied to the DLPFC samples as described in their online tutorial (https://stlearn.readthedocs.io/en/latest/stSME_clustering.html#Human-Brain-dorsolateral-prefrontal-cortex-(DLPFC)). To control for the choice of input genes, we ran stLearn's SME normalization directly on the top 15 PCs we computed from the top 2,000 HVGs . However, in order to compare with the author's recommended practices, we additionally applied the method to the raw counts of all genes, as shown in the online tutorial.

SIMULATION

We additionally evaluate the performance of BayesSpace in simulation, comparing BayesSpace spot-level clustering to other non-spatial and spatial clustering methods: k-means, Louvain, mclust, SC3, and Giotto (Simulation 1). We could not evaluate stLearn in simulation due to the need for an image as input. The simulated data are based on the melanoma and OC samples introduced in the earlier results. Eight replicates of simulated melanoma and OC PCs are

generated from *t*-distributions with means, precision, and spot labels determined by the spot-level clustering results of the real melanoma and OC samples respectively (Fig. 3B, Supplementary Fig. 23). Other clustering methods are implemented as described in the supplement with the true cluster number provided as input. BayesSpace is also implemented with the true cluster number provided as input. Performance is assessed using the adjusted Rand index (ARI) between the ground truth spot labels and the clustering results. The ARI equals 1 when there is perfect correspondence and equals 0 when the correspondence is the similarity that is expected under randomness.

EXTENSION TO MULTIPLE SAMPLES

A study may involve multiple biological replicates that are profiled using spatial transcriptomics. In such cases, it may be advantageous to jointly cluster the replicates, rather than clustering individually and then attempting to reconcile the cluster labels through permutation. Here, we propose a BayesSpace extension to enable joint clustering of multiple samples. To account for technical differences that may exist between replicates, we use the Harmony algorithm which projects data from multiple replicates into a common embedding by iterating between a clustering step and a cluster-specific linear batch correction step. Afterwards, the harmonized PCs can be used in place of the original PCs in the likelihood component of the BayesSpace model. For the spatial neighborhood component, neighbors are defined only within each replicate.

Results

Recently, Maynard et al. (2020) presented Visium spatial expression profiles of twelve dorsolateral prefrontal cortex (DLPFC) samples, as well as manual annotations of the six cortical layers and white matter for each sample as part of the spatialLIBD package[6] (Fig. 2A). Maynard

et al. annotated the DLPFC layers by considering cytoarchitecture and selected gene markers. Here, we evaluate BayesSpace's ability to identify distinct layer-specific expression profiles and compare its performance to other spatial and non-spatial clustering methods. Specifically, we compare the performance of four non-spatial algorithms commonly applied to scRNA-seq data – *k*-means, mclust[11], Louvain[10], and SC3[12] ; two recently published spatial clustering algorithms – HMRF (as implemented in the Giotto package)[14] and stLearn[15]; and the clustering partitions originally reported by Maynard et al. in the spatialLIBD package, which involve Walktrap clustering of spatial coordinates and PCs calculated from highly variable genes (HVGs) or known layer-specific marker genes. Following the methodology of Maynard et al., we use the ARI to quantify the similarity between cluster labels and the manual annotations, which are considered the ground truth.

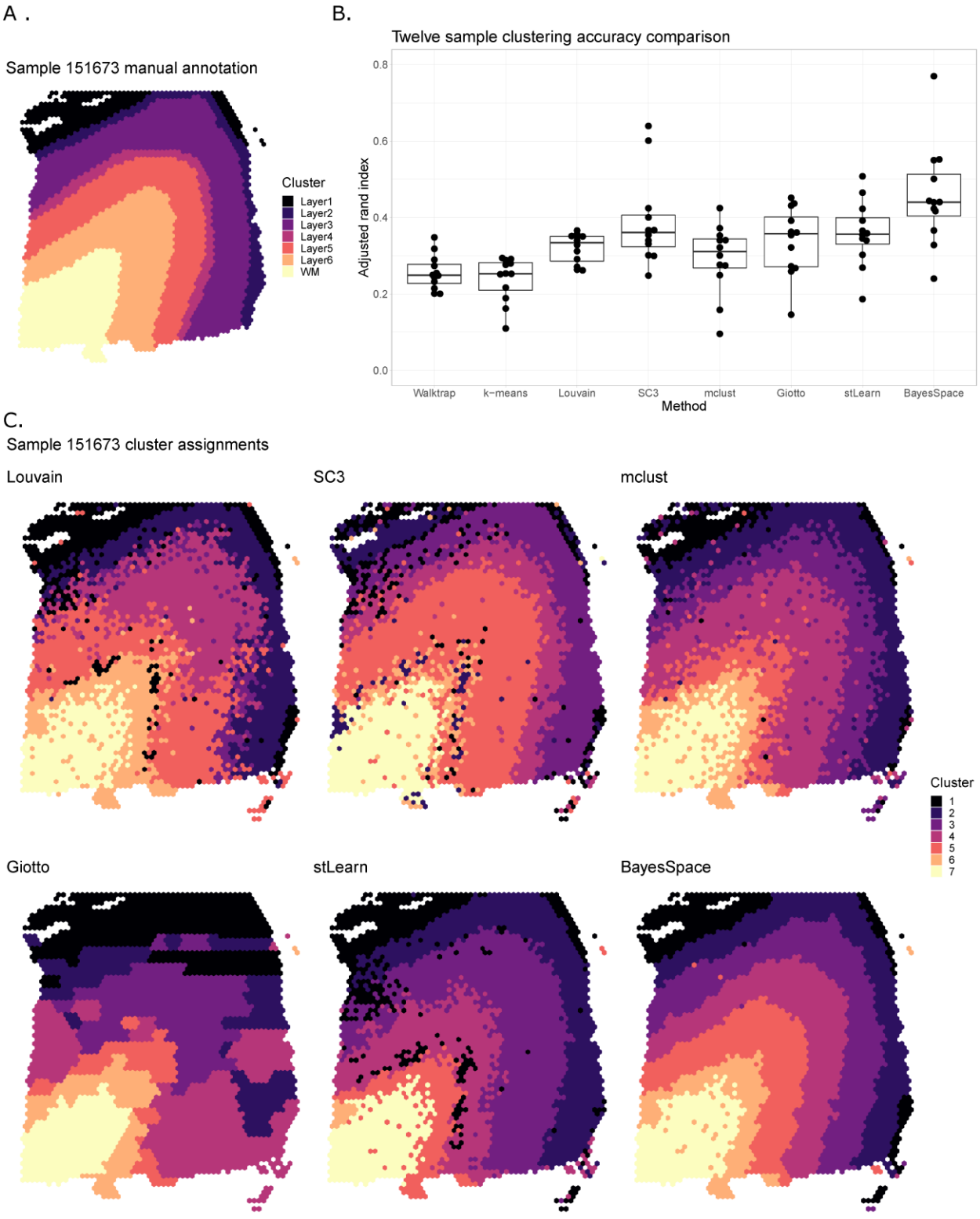

A .
Sample 151673 manual annotation
Cluster
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
B.
Twelve sample clustering accuracy comparison
Adjusted rand index
0.0
0.2
0.4
0.6
0.8
Walktrap
k−means
Louvain
SC3
mclust
Giotto
stLearn
BayesSpace
Method
C.
Sample 151673 cluster assignments
Louvain
SC3
mclust
Giotto
stLearn
BayesSpace
Cluster
1
2
3
4
5
6
7

Figure 2. BayesSpace improves computational resolution of layers in the dorsolateral prefrontal cortex. (A) Ground truth. We highlight the manually annotated six DLPFC layers and white matter in sample 151673 from the spatialLIBD dataset. Annotated layers for the remaining samples can be found in the original publication[6]. (B) Summary of clustering accuracy in all twelve samples. ARI is used to compare the similarity between cluster labels from each method against the manually annotated layers for all twelve samples. In the boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively. (C) Cluster assignments generated by non-spatial (top) and spatial (bottom) methods for sample 151673.

BayesSpace substantially outperforms the original spatialLIBD clustering partitions, as well as all non-spatial clustering algorithms and spatial clustering methods developed for spatial transcriptomics data (Fig. 2B). BayesSpace and the non-spatial methods were applied on 15 PCs calculated from the top 2000 HVGs. The spatial clustering methods, Giotto and stLearn, were implemented based on the original authors' recommended parameters (Supplementary Notes). We also show Giotto and stLearn results using precomputed PCs from BayesSpace to provide a more controlled comparison, though we find this does not improve either method's performance (Supplementary Figure 1). As an example, in sample 151673, we find that only SC3 (ARI = 0.42), mclust (ARI = 0.42), stLearn (ARI = 0.37) and BayesSpace (ARI = 0.55) generate clusters that qualitatively follow the expected layer pattern (Fig. 2C). Most clustering partitions besides BayesSpace exhibit substantial noise and lack of clear spatial separation between clusters. In contrast, BayesSpace leverages spatial information to smooth the data and provides distinct layers of clusters. The *t*-distributed error model that BayesSpace uses is particularly robust

against outliers in the clusters, which may be driven by technical artifacts caused during sample preparation or downstream analyses (Supplementary Figure 2). Additionally, BayesSpace's runtime and memory footprints are comparable to other spatial clustering methods, requiring 27 minutes of wall time and 9.6 GB of memory in this sample (Supplementary Figure 3).

In Simulation 1, where we simulated data modeled on two of our experimental datasets (See Methods for details), the results show that BayesSpace spot-level clustering consistently outperforms all other methods in both the simulated melanoma and ovarian datasets (Fig. 6A). Giotto, another spatial clustering method, also outperforms all non-spatial methods but provides slightly worse performance relative to BayesSpace. Among the non-spatial methods, mclust and Louvain clustering had fair performance.

Finally, we examine the performance of BayesSpace on jointly clustering multiple samples using samples 151673, 151674, 151675, and 151676 from the DLPFC dataset. These samples are serial section replicates and show consistent layer structure across samples (Fig. 7A). However, the low-dimensional UMAP representation of the data shows noticeable batch effects at the sample level, which may arise from technical issues in the data collection (Fig. 7B). Harmony performs well in this dataset to generate batch corrected PCs which are then spatially clustered using BayesSpace (Fig. 7C-D). The results of the joint clustering show consistent structure across samples. Comparing to the ground truth, we obtain ARIs of 0.56, 0.59, 0.6, and 0.57 for the four samples. This represents a substantial improvement over the single sample clustering analyses, where we obtained values of 0.55, 0.44, 0.55, and 0.37 respectively for the four samples.

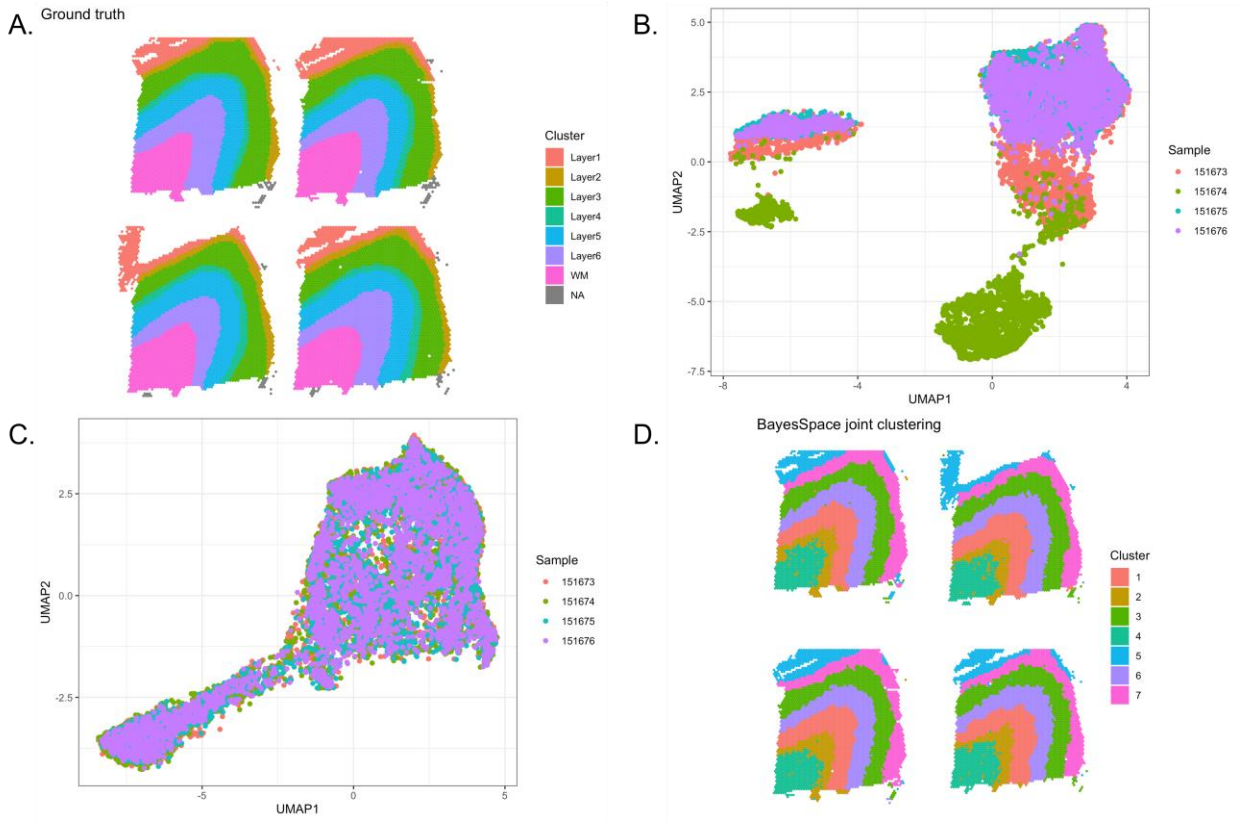


Figure 7. Joint clustering of multiple samples improves clustering accuracy. (A) Similar layer patterns can be seen in the four replicates, but (B) there are noticeable differences in the expression profiles of the replicates. (C) The harmonization procedure corrects for these batch effects. (D) BayesSpace joint clustering results are plotted spatially.

Discussion

We have demonstrated the utility of BayesSpace in identifying spatial clusters with similar expression profiles by efficiently using spatial information to inform the clustering of expression data. BayesSpace outperforms existing non-spatial and spatial clustering methods on both real and simulated data.

While there are similarities in the specification of the spatial prior between BayesSpace and Giotto/HMRF, we highlight several differences between the methods. To our knowledge, BayesSpace is the first spatial transcriptomics model-based clustering method that uses a *t*-distributed error model to identify spatial clusters that are more robust to the presence of outliers caused by technical noise. BayesSpace also uses Markov-chain-Monte-Carlo (MCMC) to estimate model parameters, while HMRF uses expectation-maximization (EM) which might not explore the space as efficiently[27]. BayesSpace also differs from Giotto/HMRF in using a fixed rather than a variable precision matrix across clusters, which we found to improve the stability of estimates without compromising clustering performance (Supplementary Figure 20), and in using a more reliable method for detecting the spatial neighborhood network.

In our work, the smoothing parameter γ is not estimated statistically, but instead guided by expert knowledge of tissue biology, as was previously recommended in other work[28]. Setting $\gamma = 3$ appeared to provide a level of smoothing that matched well with biologically relevant tissue structures. Statistical estimation of γ poses a challenge since the intractable normalization constant depends on γ and would therefore not cancel out when calculating the acceptance ratio for Metropolis-Hastings proposals of γ (as it did for estimation of $\boldsymbol{z}$). One approach to avoid this issue is by using the pseudo-likelihood to estimate γ. In the limiting case when $\gamma \to 0$, the pseudo-likelihood is equivalent to the likelihood, but when γ is far from 0, use of the pseudo-

likelihood to estimate γ tends to perform poorly[29,30]. Another approach to estimating γ is the double Metropolis-Hasting algorithm, which is an approximate method[30]. As the name suggests, Metropolis-Hastings is used twice (or more): once to sample a new γ' and $m \geq 1$ times to sample a new auxiliary variable $\boldsymbol{z}'$ from the current iteration's value for $\boldsymbol{z}$ that shares the state space as $\boldsymbol{z}$. Then, $\gamma'$is accepted with probability $\frac{p(\boldsymbol{z}'|\gamma)p(\boldsymbol{z}|\gamma')}{p(\boldsymbol{z}|\gamma')p(\boldsymbol{z}'|\gamma)}$, allowing the intractable normalization constant to be cancelled out. In practice, simulating $\boldsymbol{z}'$ involves $m \times n$ updates from $\boldsymbol{z}$ since the updates are done for each component of $\boldsymbol{z}$. This makes the procedure computationally expensive when the data contains a large number of spots. Estimating instead of fixing γ in the simple Potts model does not seem to be favorable given the computational cost and the desire for results to be guided by biological phenomenon. However, it can be argued that the simple Potts model is an overly simplified model of biology that does not take into account variation in propensities of different spatial domains (i.e. clusters) to be neighbors. In a generalized Potts model where the interaction strength is variable, estimation of the parameters via double Metropolis-Hastings may be advantageous. Such models will be most useful for modeling single-cell resolution spatial data, where clusters can credibly be identified as cell types, which may have different propensities for interaction, rather than spatially domains.

Our work focused on the ST and Visium platforms from 10x Genomics. However, BayesSpace should be applicable to other platforms where spots are arranged on a lattice. Beyond the considerations with γ, slight modifications may also be needed so that our spatial model can be used with a different neighborhood structure. Since BayesSpace models a lower dimensional representation of the data (i.e. PCA), it should also be applicable to other dimensional reduction techniques such as UMAP and possibly applied to other data types such as protein markers and multiomics.

Finally, BayesSpace's ability to jointly cluster multiple samples highlights the potential benefits to clustering accuracy when multiple replicates are collected in a study. It may be possible to further improve upon this by constructing neighborhoods in three dimensions. This would require spatially aligning serial section replicate samples by allowing for distortions and other transformations, which may be possible through techniques such as large deformation diffeomorphic metric mapping[31]. Further investigation is needed to investigate the relative importance of neighbors between and within samples in order to appropriately weight their contributions to the Potts model.

A challenge that remains is addressing the limited resolution of current spatial transcriptomics technology, which we tackle in the following chapter.

Chapter 3: Computationally enhance the resolution of spatial transcriptomics

Introduction

The primary technological limitation of current spatial gene expression platforms such as ST and Visium is resolution, with the unit of observation being spots that are 100μm in diameter on the ST platform and 55μm in diameter on the Visium platform. As such, the number of cells within a spot may range from 1 to 30 on the Visium platform, and up to 200 on the older ST platform, depending on the biological tissue[32]. Alternative approaches include fluorescence *in situ* hybridization technologies, such as seqFISH, MERFISH, and CosMx, and other recently developed spatial sequencing methods, such as Slide-seq and ZipSeq[33–37]. While these methods provide increased resolution, most are lower throughput, less sensitive, rely on custom protocols, or are not widely available.

In this chapter, we extend BayesSpace to use the neighborhood structure in spatial transcriptomics data to increase the resolution to the sub-spot level (Fig. 1A). In contrast to existing deconvolution methods using scRNA-seq data[38–40], the enhanced-resolution modeling of BayesSpace – which approaches single-cell resolution with the Visium platform – does not require independent single-cell data and allows us to infer the spatial arrangement of the subspots. While integration with scRNA-seq is appealing, it may be costly if using matched samples or introduce bias if using publicly available references. Furthermore, the deconvolved mixtures are still only spatially resolved at the original scale of the ST or Visium technology and the neighborhood structure of the cell types can’t be recovered.

We use immunohistochemistry as a ground truth in two cancer samples to validate that our enhanced-resolution clustering identifies a tissue structure consistent with cell surface markers, and we report examples of transcriptional heterogeneity in the tumor microenvironment

not achievable by immunohistochemical analyses alone. Furthermore, using *in silico* spatial transcriptomics datasets generated from aggregating single-cell RNA-seq, we show that BayesSpace can recover the true spatial structure at near single-cell resolution.
Using immunohistochemistry and an *in silico* dataset constructed from scRNA-seq data, we show that BayesSpace resolves tissue structure that is not detectable at the original resolution and identifies transcriptional heterogeneity inaccessible to histological analysis.

Our method draws from the existing literature for using Bayesian statistics to achieve super-resolution images[28,41–43]. In general, these approaches consider the setting where image data is only observed at the pixel level, but the goal is to restore an image at subpixel resolution by leverging spatial information through a MRF prior. For example, Ripley proposed a subpixel restoration method using a Ising model prior where a fixed number of subpixels, each with its own cluster label, constitute each pixel[28]. The observed data of each large pixel is then modeled as a normal distribution with mean given by the weighted mean of the cluster labels for the constituent subpixels. In another Bayesian approach, Gavin and Jennison do not explicitly model subpixels but instead allow the edges of clusters to cross within pixels[41]. The approach uses a prior that encourages the total edge length of clusters to be small, which is intuitively similar to the Potts model approach of encouraging neighbors to belong to the same cluster.

Methods

SPATIAL CLUSTERING MODEL AT ENHANCED RESOLUTION

In order to enhance the resolution of the clustering map, we segment each spot into subspots and leverage spatial information using the Potts model spatial prior. Specifically, we segment each ST spot into 9 subspots and each Visium spot into 6 subspots (Fig. 1B). For ST, we used 9 subspots to help increase the resolution of the data from the lower resolution

technology, since ST spots are 100 μm in diameter while Visium spots are 55 μm in diameter. This translates into a more than 3-fold difference in area. The choices of 6 and 9 subspots also exploit the triangular and square lattice geometries of the Visium and ST platforms respectively. In the IDC and OC samples, Visium spots are estimated to contain a median of around 20 cells, so subspots will generally represent the expression of a few cells, rather than potentially dozens of cells at the spot level (Supplementary Figures 7 and 8).

Relative to the spot-level clustering method, the model specification and parameter estimation is largely similar for the enhanced resolution clustering, though the unit of analysis is now the subspot rather than the spot. Since gene expression is not observed at the subspot level, it is modeled as another latent variable that is also estimated through MCMC. The latent expression of each subspot j that is part of spot i is denoted $\boldsymbol{y}_{ij}^{*}$, subject to the constraint that $\sum_j \boldsymbol{y}_{ij}^{*} = \boldsymbol{y}_i$. This constrant ensures that the sum of the subspot expression values is equal to the observed expression value of that spot. Now, we wish to sample from the following target distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{y}^{*} | \boldsymbol{y}),$$

Which can be done using MCMC by iteratively sampling from the joint distribution of each variable given the data $\boldsymbol{y}$. Here we simply describe how we sample from $(\boldsymbol{y}^{*} | \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{y})$ since all other steps remain the same as before, replacing $\boldsymbol{y}$ by $\boldsymbol{y}^{*}$. Each $\boldsymbol{y}_{ij}^{*}$ is initialized to equal $\boldsymbol{y}_i$ and then updated via random walk Metropolis-Hastings. In each iteration and for each spot, the new proposal is given by $\boldsymbol{y}_{ij}^{*\prime} = \boldsymbol{y}_{ij}^{*} + \boldsymbol{\varepsilon}_{ij}$ for each subspot such that $\boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \sigma^2 I_d)$ where σ^2 is a fixed parameter and $\sum_j \boldsymbol{\varepsilon}_{ij} = \mathbf{0}$. In effect, this jitters the latent expression value of each subspot within a spot while keeping the total expression of the spot fixed. The proposal is accepted or rejected using the standard MH step. Note that the latent expression of all subspots in a spot is

updated jointly in a single move. We set σ^2 so that the acceptance rate ranges from 25% to 40% of iterations on average to maximize the efficiency of the Metropolis-Hastings algorithm[44]. Aside from replacing $\boldsymbol{y}_i$ with $\boldsymbol{y}_{ij}^*$, all other steps of the MCMC algorithm remain the same as in the spot-level clustering method. Model fitting diagnostics are provided in Supplementary Figure 20. Intuitively, the enhancement procedure reassigns the total expression within a spot to its constituent subspots by leveraging spatial information, ultimately generating a higher resolution spatial clustering map.

Mapping high-resolution PCs to high-resolution gene expression space.

While BayesSpace can provide higher-resolution maps of spatial transcriptomic patterns, the modeling is done on the PC space and an additional step is necessary to map the principal component values back to the original log-normalized gene expression space. BayesSpace implements two options for predicting high-resolution gene expression: linear regression and non-linear regression using XGBoost (default) [45,46]. In either case, a model is trained for each gene where the outcome is the measured gene expression at the spot level, and the predictors are the PCs generated from the original data. The fitted model can then be used to predict gene expression from the high-resolution PCs estimated using enhanced resolution clustering. The enhanced gene expression values can be visualized spatially and analyzed via differential expression methods (Fig. 1A).

Data description

We apply BayesSpace to enhance several datasets. The first dataset involved melanoma samples run on the ST platform[4]. From this dataset, we analyzed the second replicate from biopsy 1 since it contained regions annotated as lymphoid tissue and was also described extensively in the original paper. Biopsy 1 contains 293 spots covered by tissue. Counts matrices

were downloaded from the Spatial Research website (https://www.spatialresearch.org/wp-content/uploads/2019/03/ST-Melanoma-Datasets_1.zip). Array row/column coordinates were obtained from the column labels of these matrices. An aligned image and pixel coordinates were not available for this sample.

The second dataset is publicly available from the 10x Genomics website and includes matching Visium spatial gene expression (3,493 spots) and IF staining of an endometrial adenocarcinoma of the ovary. The sample was stained with anti-Cytokeratin antibody, anti-human CD45 antibody, and DAPI. The counts matrix and spatial data generated by the SpaceRanger pipeline, along with the immunofluorescence stain image, were downloaded from 10x Genomics' website. Data and details of the sample preparation and staining are available at https://support.10xgenomics.com/spatial-gene-expression/datasets/1.2.0/Parent_Visium_Human_OvarianCancer.

The third dataset is an invasive ductal carcinoma prepared on the Visium platform (4,727 spots) and stained with anti-human CD3 antibody and DAPI. Tissue was collected from a middle-aged woman of Asian descent with a grade III invasive ductal carcinoma (IDC) (AJCC/UICC stage T2N0M0; the tumor was ER^+, PR^-, and $Her2^+$). 10x Genomics procured this sample from BioIVT:ASTERAND (Westbury, NY) and prepared it for sequencing as described in the Visium Spatial Protocols. Specifically, the tissue section was embedded and cryosectioned into 10 µm sections, then fixed and stained with DAPI and antibodies for CD3. A Visium Gene Expression library was prepared following the Visium Spatial Reagent Kits User Guide and sequenced on a NovaSeq 6000 to a mean depth of 40,795 reads per spot. The gene expression library was aligned and quantified with the Space Ranger pipeline (v2020.1023.1). Regions of invasive carcinoma, carcinoma *in situ*, and benign hyperplasia were grossly delineated by tissue

and nuclear morphology using the DAPI stain in consultation with two breast cancer pathologists. Visium spots were assigned to each annotated region using Adobe Photoshop and EBImage, by assigning each spot to an annotated region based on the color value at its pixel coordinate.

For the second and third datasets, images were analyzed using Indica Labs HALO (High-plex FL module v3.2.1). The HALO algorithm segmented the nuclei based on user-defined DAPI pixel intensities, then added a 0.5 pixel virtual cell boundary. After segmentation, we assigned cells to Visium spots where possible. If the annotated center of a cell fell within the boundary of a Visium spot (defined as a 55 μm radius around the spot center), we assigned the cell to that spot. In both datasets, we identified regions of poor focus or overexposure where cell segmentation was not possible. We masked these regions from our validation analyses of resolution enhancement due to correspondingly low quality immunofluorescent stain intensity. We used EBImage[47] to measure the intensity of the CD3 (IDC) and CD45 (OC) stains from their respective channels in the full resolution immunofluorescent image, averaging the intensity over the pixels encompassed by each spot (defined as the 55μm circle centered at the image pixel coordinates reported by Space Ranger) and subspot (defined as one-sixth of each spot). BayesSpace clusters (spot-level) were then binned into CD3 or CD45 “high” and “low” clusters based on the median average intensity of the respective stain in the spots (Supplementary Figures 9 and 10).

The final dataset included ten human skin SCCs profiled on either the ST or the Visium platform[48]. Among the two samples run on the Visium platform, we chose to analyze patient 4 (P4) since the data quality was higher as shown in the original paper. Sample P4 contains 722

spots covered by tissue. A counts matrix containing expression profiles from all samples and tissue positions were downloaded from GEO (accession GSE144239).

Differential expression analysis

Differential expression analysis was performed with the FindMarkers() and FindAllMarkers() functions provided with the Seurat R package,[49,50] using MAST[51] to test for significance. In the IDC sample, we used FindAllMarkers() to find genes differentially expressed in each cluster (Supplementary Figure 13). We also tested for differential expression among the non-tumor clusters (1, 7, 10; Supplementary Figure 14) and among the invasive tumor clusters (3, 4, 5, 6, 9; Supplementary Figure 17). In our analyses, we use the two-sided Wilcoxon rank-sum test as implemented in the R Seurat package to identify the top differentially expressed genes and also use Seurat for heatmap visualizations of the centered and scaled gene expression values[52].

Simulations

We perform two simulations to evaluate the performance of BayesSpace resolution enhancement. In the second simulation, we simulate 20 replicates from *t*-distributions with means, precision, and labels based on the real melanoma and OC samples, but unlike the spot-level simulation in the previous chapter, we generate subspots using the enhanced clustering results as the ground truth (Fig. 3C, Fig. 4D). The simulated subspot-level PCs are averaged to provide spot-level PCs that are given as input to BayesSpace. We can use the modal ground truth label of the subspots within each spot to generate an optimal spot-level clustering for each dataset (Supplementary Figure 23). The ARI between this optimal spot-level clustering and the subspot-level ground truth represents the highest ARI that can be achieved when all subspots within a spot must belong to the same cluster, as is the case with spot-level clustering.

In the second simulation, we sample data from real single cells rather than simulating PCs. Here, we sample single cells from scRNA-seq profiling of high-grade serous ovarian cancer (HGSOC) patients[53].[53]. The single cells can be sampled into subspots on the OC Visium sample, providing another way to evaluate the performance of BayesSpace clustering and enhancement relative to other methods without relying on model-based data generation. Given the limited number of single cells, we use only the positions from a portion of the OC Visium sample. The ground truth cluster labels are derived from expert single-cell level annotation of tumor and stroma compartments within the IF stain image associated with the OC sample. In each subspot, the ground truth is assigned using the modal annotation of the single cells located within the subspot. Consequently, the ground truth assignment takes into account the gaps between spots in spatial transcriptomics technologies, and the clusters represent realistic biological spatial domains.

To add complexity to the simulation, we separate the tumor compartment into two ground truth clusters and introduce two additional intra-tumoral clusters that represent heterogeneity within tumors. Thus, the simulation includes a total of five spatial ground truth clusters, including the stroma compartment cluster. The single-cell sampling strategy is shown in Supplementary Table 1, with single cells randomly drawn from single cell clusters into corresponding spatial clusters in each of the eight simulation replicates. Since raw counts were not available in the HGSOC dataset, pseudo-counts are obtained by back transforming the log-normalized counts, and the simulated data are generated by aggregating across all subspots within a spot. The data are then processed to generate PCs as described for real data in the Methods. Since the HGSOC single-cell clusters are very well separated, we also add random noise to each simulated PC equal to 25% of its variance, thus adding additional complexity to

our simulation. This process also makes our simulated more realistic when comparing the generated PCs to PCs derived from experimental data (Supplementary Figure 22).

EXTENSION TO INTEGRATE WITH IMAGES

In addition to spatial expression profiling, the Visium technology also captures an image of the tissue. This can be either an hemotoxylin and eosin stain (H&E) image or an immunofluorescence image. In either case, these image data are at much higher resolution than the spatial profiling, though with limited throughput. For example, H&E images are stored in RGB format meaning the data has dimension $\delta = 3$ while IF images as done in Visium also typically have two to three color channels. For each subspot, we can extract features from the imaging data to augment each d-dimensional $\boldsymbol{Y}_{ij}^{*}$ vector by δ, where the additional δ features are observed and known, rather than needing to be estimated. All other parameters such as the cluster means and labels are modeled and estimated as done in the original model described above, though the data are now $d + \delta$ in dimension. Intuitively, the image features provide more information to guide the estimation of cluster labels and latent subspot-level expression in addition to the spatial neighborhood information encoded in the Potts model.

We explore two approaches to feature extraction from images. The first involves taking a simple average in each color channel over all pixels corresponding to each subspot. This approach is used for the OC dataset's immunofluorescence image. The second involves using an autoencoder to learn a nonlinear low-dimensional coding of the image data. The autoencoder is an artificial neural network that includes an encoder and a decoder[54]. The encoder takes the image data as input and attempts to identify an efficient representation. Then the decoder, which mirrors the encoder, uses the efficient representation to reconstruct the original image as output, and the neural network is trained to minimize the mean squared error between the input and

output images. In this analysis, we chose a single layer for both the encoder and decoder, and we chose a dimension of 64 for the learned representation. For the SCC dataset, the input are three color channel image tiles corresponding to the subspots that each measure 161 by 161 pixels.

After feature extraction using either the mean or the autoencoder, PCA is applied so that the image data are on a comparable scale to the expression data and then the top three PCs are modeled by BayesSpace.

Results

INCREASED RESOLUTION CLUSTERING LEADS TO THE IDENTIFICATION OF KNOWN TISSUE STRUCTURES MISSED BY OTHER METHODS.

We used BayesSpace to analyze a melanoma ST sample first annotated and described by Thrane et al[4]. Since the manual annotation identified regions of melanoma, stroma, and lymphoid tissue and leaves an additional area unannotated (Fig. 3A), we ran spatial clustering with k=4 clusters (Fig. 3B). The resulting clusters correspond well with the manually annotated tissue types. Furthermore, the melanoma tissue is split into the central region of the tumor and an outer ring of mixed tumor and lymphoid tissue. BayesSpace enhanced spatial clustering provides a higher resolution map of the tissue types (Fig. 3C). Notably, the enhancement identifies lymphoid regions along the tumor border and possible immune infiltration into the tumor that cannot be discerned at the original resolution. These regions are also largely not identified by other clustering methods (Supplementary Fig. 4). While most clustering methods identified heterogeneity between the periphery and the center of the tumor, only SC3, Giotto, and subspot-level BayesSpace identify lymphoid regions proximal to the tumor, with BayesSpace providing higher resolution and more robust signal (Supplementary Fig. 4). Finally, we also ran

BayesSpace at the spot level using 5 and 6 clusters, identifying potential heterogeneity within the stroma region (Supplementary Fig. 4).

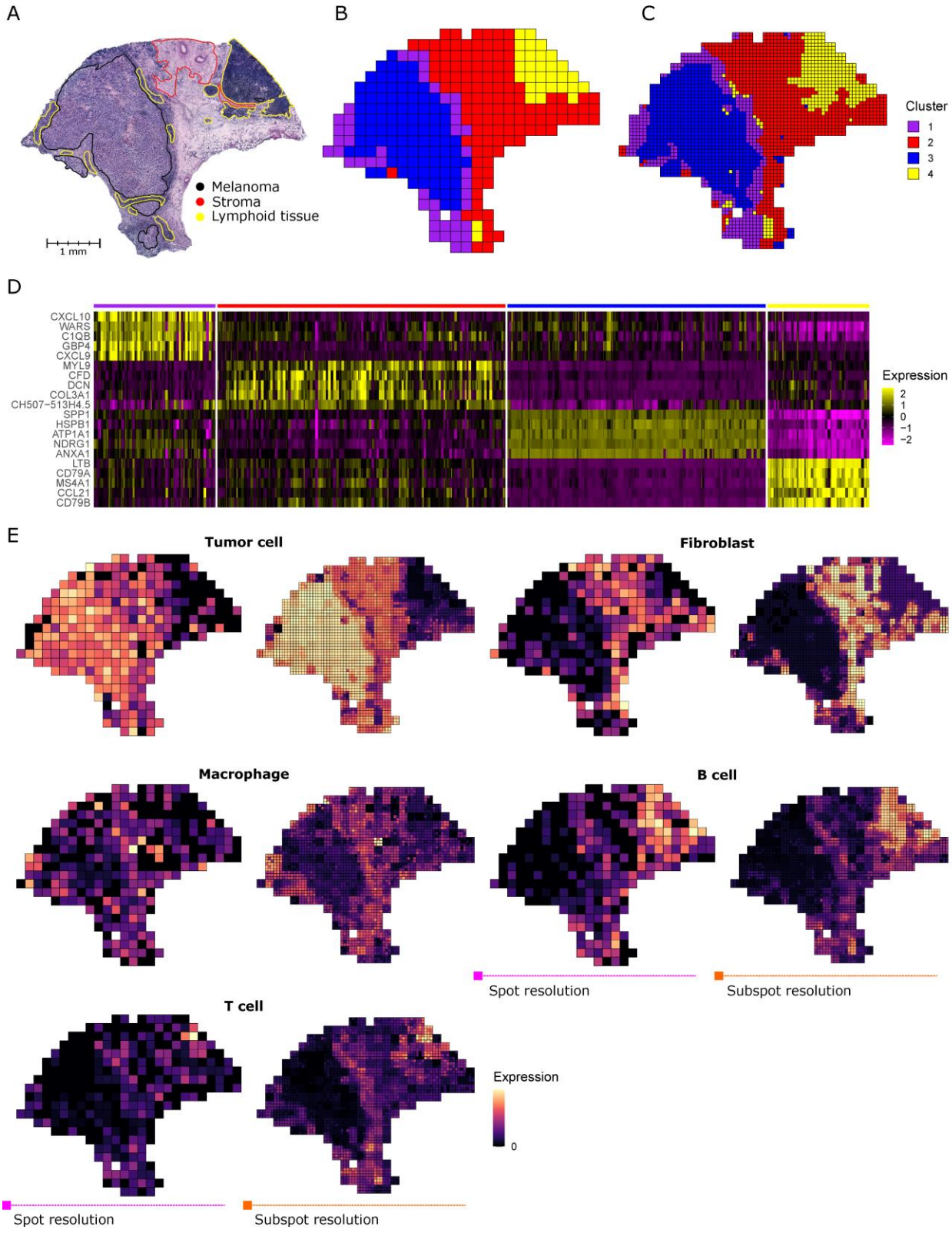

A
B
C
Melanoma
Stroma
Lymphoid tissue
1 mm
Cluster
1
2
3
4
D
CXCL10
WARS
C1QB
GBP4
CXCL9
MYL9
CFD
DCN
COL3A1
CH507−513H4.5
SPP1
HSPB1
ATP1A1
NDRG1
ANXA1
LTB
CD79A
MS4A1
CCL21
CD79B
Expression
2
1
0
−1
−2
E
Tumor cell
Fibroblast
Macrophage
B cell
Spot resolution
Subspot resolution
T cell
Expression
0
Spot resolution
Subspot resolution

Figure 3. Enhanced resolution clustering identifies tumor-proximal lymphoid tissue in a melanoma sample. (A) The original histopathological annotations of the H&E stained tissue (N = 1 tissue section, n = 293 spots) find a section of melanoma (black) adjacent to tumor-proximal lymphoid tissue (yellow) and a region of stroma (red) separating these from a larger section of tumor-distal lymphoid tissue (yellow)[4]. Adapted from *Cancer Research*, 2018, **78**(20), 5970-5979, Thrane, K. *et al*., Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma, with permission from AACR. Spatial clustering (B) and enhancement (C) generate biologically meaningful spatial domains corresponding to the original annotations. The enhanced resolution clustering identifies tumor-proximal lymphoid tissue (cluster 4; yellow) which is not resolved at spot-level clustering. (D) Differential expression analysis between the four clusters highlights the spatial differences in the expression of immune genes, cancer markers, and genes encoding extracellular matrix proteins. (E) For each of the five major cell types, the observed total spot-level expression (as measured by the summed log-normalized counts) of the defined marker genes (left) is shown alongside the corresponding enhanced resolution expression (right). We show the spatial expression plots for tumor cells (*PMEL*), fibroblasts (*COL1A1*), macrophages (*CD14*, *FCGR1A*, *FCGR1B*), B cells (*CD19*, *MS4A1*), and T cells (*CD2*, *CD3D*, *CD3E*, *CD3G*, *CD7*).

Using the enhanced PCs, we can generate high resolution maps of individual genes or expression profiles for major cell types as described in the Methods. Differential expression analysis performed on enhanced resolution gene expression indicates that the lymphoid regions have a distinct expression profile. We see elevated expression of lymphocyte markers such as *CD52* and *MS4A1* and lower expression of melanoma markers such as *MCAM* and *SPP1* relative

to the surrounding tumor border (Supplementary Figure 4). Enhanced resolution differential expression analysis between the four clusters highlights additional spatial variation in gene expression (Figure 3D). In the stroma (cluster 2), expression is higher for extracellular matrix proteins such as *DCN* and *COL3A1*. Furthermore, we reveal intra-tumor heterogeneity between the border and center of the tumor (clusters 1 and 3 respectively), with higher chemokine (*CXCL9*, *CXCL10*) activity at the border and elevated expression of genes related to cell proliferation (*HSPB1*) and metastasis (*ATP1A1*) at the center[55,56].

We define tumor cell (*PMEL*), fibroblast (*COL1A1*), B cell (*CD19*, *MS4A1*), T cell (*CD2*, *CD3D*, *CD3E*, *CD3G*, *CD7*), and macrophage (*CD14*, *FCGR1A*, *FCGR1B*) expression profiles based on one or more marker genes from existing literature[57]. The enhanced expression profiles provide noticeably higher spatial resolution (Figure 3E). In particular, we can more clearly see immune expression on the periphery of the tumor. The contrast between *PMEL* expression in the tumor, stroma, and lymphoid tissue is also more apparent with enhanced resolution.

IMMUNOHISTOCHEMISTRY VALIDATES ENHANCED-RESOLUTION CLUSTERS

In order to validate our enhanced-resolution clustering and gene expression, we analyzed an unreported breast cancer sample: an estrogen receptor-positive (ER+), progesterone receptor-negative (PR-), HER2-amplified (HER+) invasive ductal carcinoma (IDC) prepared on the Visium platform with immunofluorescence staining for DAPI (staining nuclei) and CD3 (staining T cells) (Supplementary Figure 5). We additionally analyzed a dataset published by 10x Genomics: an endometrial adenocarcinoma of the ovary (ovarian cancer; OC) sequenced on the Visium platform and stained with immunofluorescence for DAPI, pan-cytokeratin (panCK; staining epithelial tissue), and CD45 (staining leukocytes) (Supplementary Figure 6). After pathologist examination, out-of-focus and overexposed regions were excluded from the analysis

(Methods; Supplementary Figures 7, 8). Cell segmentation of the in-focus areas (IDC n=2,929/4,727 spots; OC n=2,041/3,493 spots) identified a median of 21 cells per spot in the IDC tissue and 19 cells per spot in the OC tissue, along with a median of 3 cells per subspot in both tissues (Supplementary Figures 7, 8).

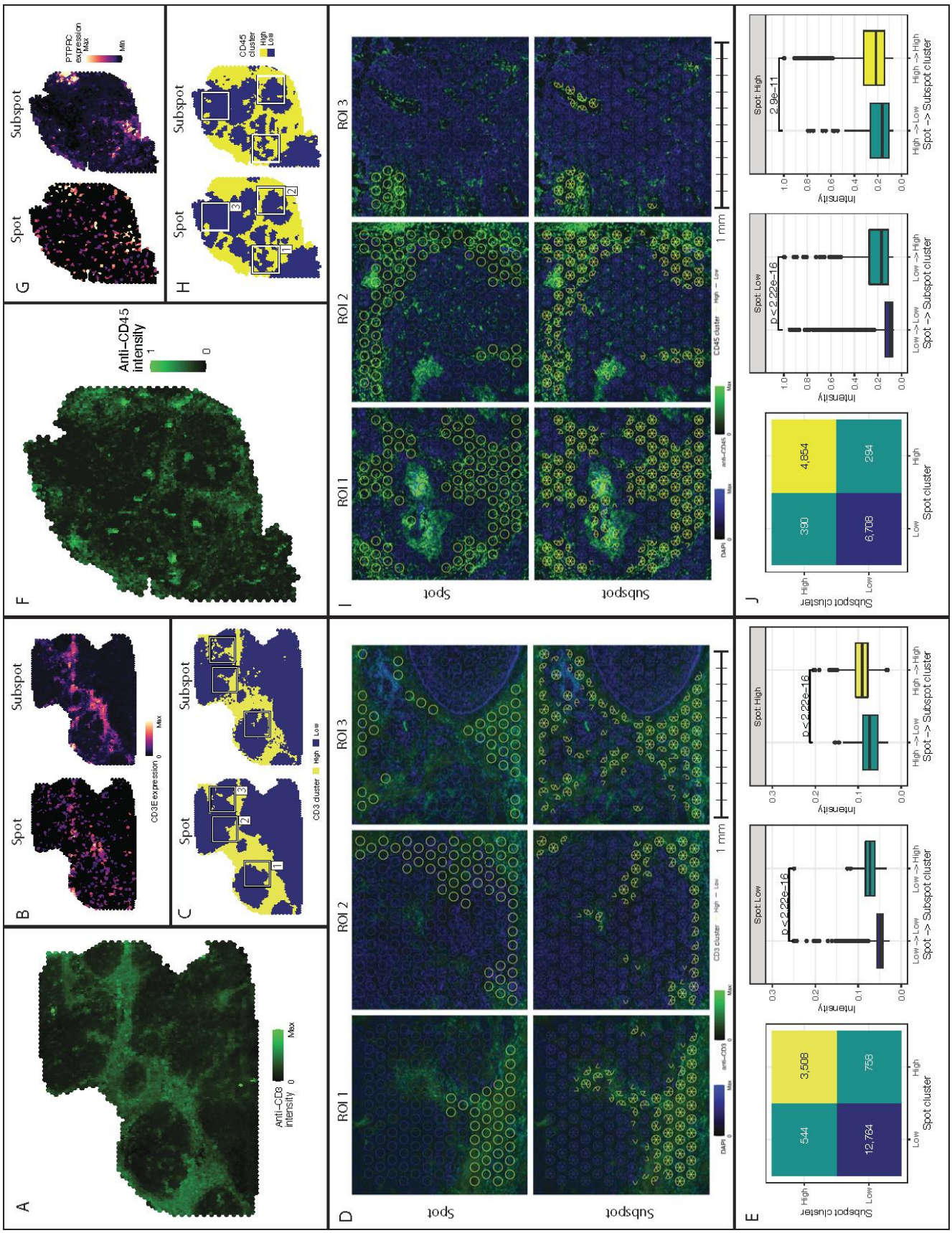

A
Anti-CD8 intensity
Max
0
B
Spot
Subspot
CD3E expression
Max
0
C
Spot
Subspot
1
2
3
CD3 cluster
High
Low
D
ROI 1
ROI 2
ROI 3
Spot
Subspot
DAPI
Max
0
CD3 cluster
High
Low
1 mm
E
Subspot cluster
High
Low
544
3,508
12,764
758
Low
High
Spot cluster
Intensity
0.3
0.2
0.1
0.0
Spot: Low
p < 2.22e−16
Low −> Low
Low −> High
Spot −> Subspot cluster
Spot: High
p < 2.22e−16
High −> Low
High −> High
Spot −> Subspot cluster
F
Anti−CD45 intensity
1
0
G
Spot
Subspot
PTPRC expression
Max
Min
H
Spot
Subspot
1
2
3
CD45 cluster
High
Low
I
ROI 1
ROI 2
ROI 3
Spot
Subspot
DAPI
Max
0
CD45 cluster
High
Low
anti−CD45
Max
0
1 mm
J
Subspot cluster
High
Low
380
6,708
294
4,854
Low
High
Spot cluster
Intensity
1.0
0.8
0.6
0.4
0.2
0.0
Spot: Low
p < 2.22e−16
Low −> Low
Low −> High
Spot −> Subspot cluster
Spot: High
2.9e−11
High −> Low
High −> High
Spot −> Subspot cluster

Figure 4. Immunohistochemistry validates BayesSpace enhancement in an invasive ductal carcinoma (IDC) sample and an ovarian cancer (OC) sample. A) Average intensity of the anti-CD3 immunofluorescent stain in the IDC. The intensity was scaled to the range [0, 1] for visualization. B) Log-normalized gene expression of *CD3E* measured on the Visium platform (left; "spot") and enhanced with BayesSpace (right; "subspot"). C) Dichotomized clustering of the Visium gene expression. After clustering the tissue section into ten clusters, the clusters were binned by their median anti-CD3 stain intensity into CD3 “high” and CD3 “low” clusters, shown here. The red squares outline three regions of interest (ROIs) where the enhanced clustering found areas of increased heterogeneity. D) Zoomed-in views of the n = 3 ROIs. Each panel shows a $1mm^2$ area of the IF image. DAPI intensity is shown in blue and anti-CD3 intensity in green. Overlaid on each panel in the top row is the spot-level clustering. Each circle corresponds to the position and size (55 μm diameter) of a spot on the Visium array and is colored based on whether it belongs to a CD3 “high” (yellow) or CD3 “low” (blue) cluster. The bottom row contains a similar overlay of the enhanced-resolution subspot clustering, where the circles are now sub-divided into six wedges corresponding to the positions of the subspots in the BayesSpace model. As in the spot overlay, the subspots are colored based on their cluster membership. E) Summary of subspot reassignment after enhancement. On the left we show a contingency table describing the number of subspots (n = 17,574) that belong to a CD3 “high” or “low” cluster at the spot level and at the subspot level. Using two-sided Wilcoxon rank-sum tests, we also show that the anti-CD3 intensity in subspots that are reassigned to a “high” cluster is significantly higher ($p < 2.22e-16$) than in those that remain in a “low” cluster (center), and that subspots which are reassigned to a “low” cluster” have a significantly lower ($p < 2.22e-16$) anti-CD3 intensity than those that remain in a “high” cluster (right). F-J) The panels for the OC

mirror those for the IDC, with anti-CD45 intensity replacing anti-CD3 and *PTPRC* (CD45) gene expression replacing *CD3E*. In panel E, we show n = 12,246 subspots. In panel I, we show n = 3 ROIs. In panel J, using two-sided Wilcoxon rank-sum tests, we show that the anti-CD45 intensity in subspots that are reassigned to a "high" cluster is significantly higher ($p < 2.22e-16$) than in those that remain in a "low" cluster (center), and that subspots which are reassigned to a "low" cluster" have a significantly lower ($p = 2.9e-11$) anti-CD45 intensity than those that remain in a "high" cluster (right). All reported p-values are unadjusted values.

We applied BayesSpace to cluster the IDC sample into 10 clusters and the OC sample into 8 clusters at spot and subspot resolution, selecting the number of clusters based on the negative log-likelihood curve (Supplementary Figures 9, 10). We analyzed the anti-CD3 and anti-CD45 intensity in the in-focus area of each tissue section (Fig. 4A and 4F, respectively), finding that the IF signal correlated well with the corresponding enhanced gene expression (Pearson's r=0.53 in IDC; Figure 4B and 4G). In both samples, we identified clusters enriched for the respective immune IF signal and dichotomized the clusters into CD3/CD45-rich and CD3/CD45-poor areas (Fig. 4C and 4H, Supplementary Figure 9, 10). From this, we identified regions of interest (ROIs) between the spot-level and enhanced clustering – areas where the enhancement increased the observed heterogeneity and many subspots flipped from immune-rich to immune-poor, or vice versa. We highlight six of these ROIs in Figure 4D and 4I to demonstrate that the enhanced clustering qualitatively improves the concordance of the clustering with the underlying immnunohistochemical stain. Specifically, we present regions where, compared to the coarser spot-level clustering, the enhanced-resolution clustering detects subspots with high underlying IF stain intensity and refines the boundary between immune-rich and immune-poor areas.

To quantify the improvement at enhanced resolution, we compared the distribution of IF intensity between subspots that changed classification after enhancement (e.g. immune-rich at the spot level, and immune-poor after enhancement) and subspots that maintained their classification (e.g. immune-rich at both the spot and subspot level). We found a significant difference in the intensity of subspots that changed classification compared to those that maintained their spot-level status (Fig. 4E and 4J), indicating that BayesSpace's resolution enhancement improves the accuracy of expression-based clustering with respect to an orthogonal immunohistochemistry signal.

BayesSpace distinguishes intratumoral heterogeneity in invasive ductal carcinoma

We further analyzed the IDC tissue section to identify clusters of biological relevance. Pathologist annotation identified regions of predominantly invasive carcinoma (IC), carcinoma *in situ* (CIS), and benign hyperplasia, from which we derived ground truth labels for each spot (Fig. 5A, Supplementary Figure 11). The clusters were largely consistent with the histopathological annotations (cluster purity=0.839; Fig. 5B, Supplementary Figures 9 and 11), and we identified five clusters that corresponded to annotated regions of predominantly IC (3, 4, 5, 6, 9), one cluster that encompassed all annotated regions of CIS (8), one cluster that coincided with the annotated benign hyperplasia and an invasive-appearing area (2), and three clusters corresponding to predominantly non-tumor areas (1, 7, and 10; Supplementary Figure 11). We note that without H&E stains or an immunofluorescent stain for a tumor marker, the tumor-stroma interface could not be fully delineated histologically, and BayesSpace's enhanced clustering identifies heterogeneity within the tissue that is not reflected in the annotated boundaries but clearly supported by key tumor marker genes (Fig. 5C-E). This further supports our previous validation with immunofluorescence (Fig. 4).

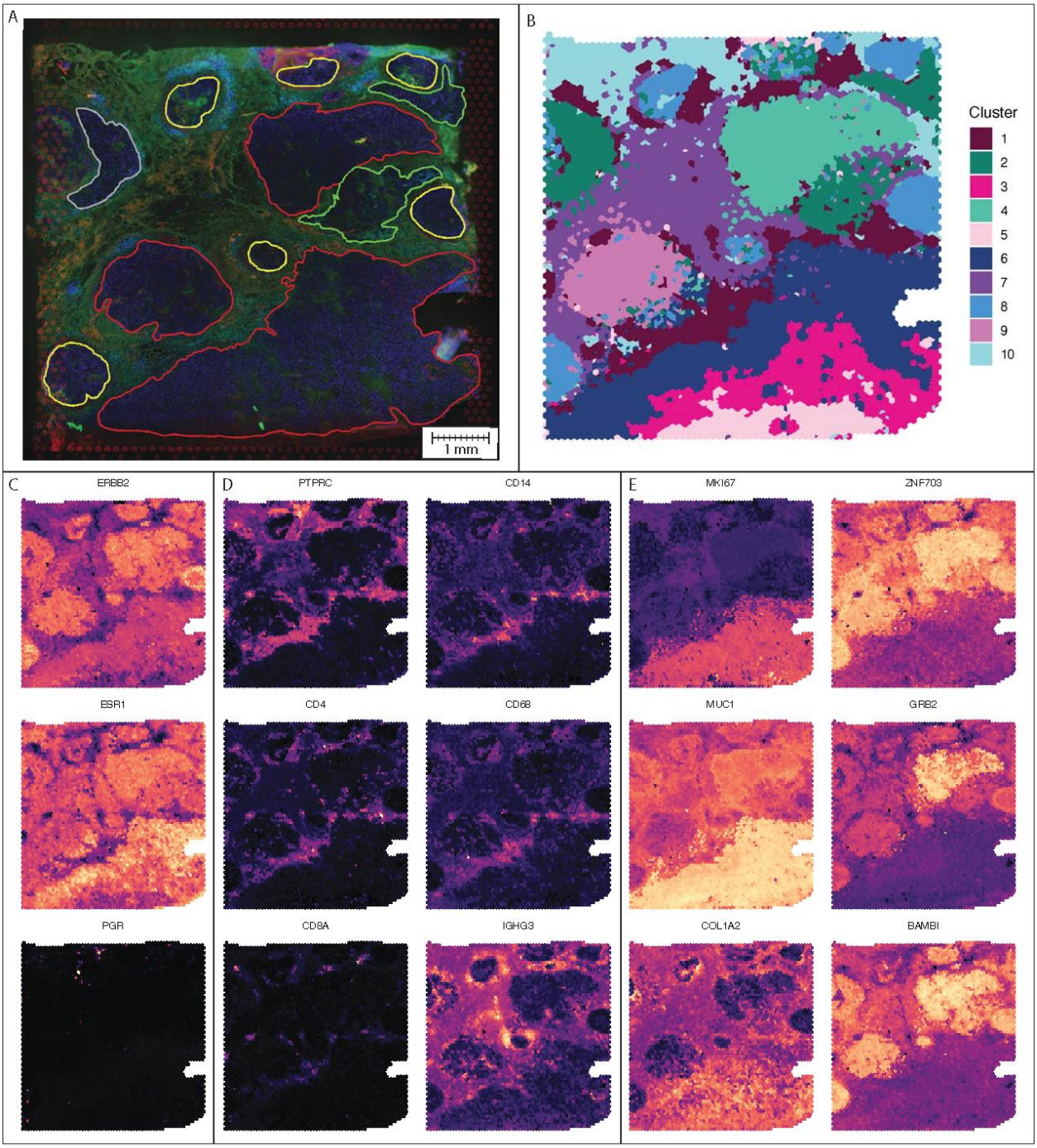

A
1 mm
B
Cluster
1
2
3
4
5
6
7
8
9
10
C
ERBB2
ESR1
PGR
D
PTPRC
CD14
CD4
CD68
CD8A
IGHG3
E
MKI67
ZNF703
MUC1
GRB2
COL1A2
BAMBI

Figure 5. BayesSpace identifies transcriptional heterogeneity within an invasive ductal carcinoma. A) Immunofluorescent imaging of the tissue section (N = 1 tissue section, n = 4,727 spots) and histopathological annotations. DAPI intensity is shown in blue, anti-CD3 intensity in green, and the Visium fiducial frame in red. Annotated regions of invasive carcinoma are outlined in red, carcinoma *in situ* in yellow, benign hyperplasia in green, and unclassified tumor in grey. B) Enhanced BayesSpace clustering. C) Spatial expression of the genes coding for HER2 (*ERBB2*), and estrogen receptor (*ESR1*) and progesterone receptor (*PGR*). D) Spatial expression of immune genes CD45 (*PTPRC*), *CD4*, *CD8A*, *CD14*, *CD68,* and *IGHG3*. E) Spatial expression of proliferation marker Ki-67 (*MKI67*), markers of tumor progression *MUC1* and *COL1A2*, oncogene *ZNF703*, growth factor receptor protein *GRB2*, and TGFβ pseudoreceptor *BAMBI*.

Spatial expression patterns of known marker genes and differential expression analysis among these clusters were largely in accord with the clinical and histopathological annotations. Consistent with the clinical report of ER+/PR-/HER2+ IDC, we observed high expression of genes coding for HER2 (*ERBB2*) and ER (*ESR1*) throughout the tumor clusters and minimal expression of the gene coding for PR (*PGR*) in the sample (Fig. 5C, Supplementary Figure 12). The non-tumor clusters 1, 7, and 10 were characterized by the expression of immune genes, with the leukocyte common antigen CD45 (*PTPRC*) highly expressed in these clusters. We found that these clusters corresponded to distinct spatial transcriptional patterns. Cluster 1 was enriched for signatures of cell-mediated immunity, including marker genes expressed by T-cells (*CD4, CD8A, CD8B*) and macrophages (*CD14, CD68*), while clusters 7 and 10 were enriched for genes involved in humoral immunity, particularly immunoglobulin chains (e.g. *IGHG3*; Fig. 5D,

Supplementary Figures 12-15). Compared to the other non-tumor clusters, Cluster 7 was also enriched for expression of HER2 and tumor-associated genes, such as *ZNF703*, suggesting this cluster represents a mixture of tumor and immune cells. Analysis of non-tumor subspots (cluster 1, 7 and 10) with CIBERSORT was consistent with the differential expression results, predicting subspots in cluster 1 to have a greater abundance of T cells while clusters 7 and 10 had higher proportions of B and plasma cells (Supplementary Figure 16).

We found similar heterogeneity within the invasive tumor clusters. Clusters 3, 5, and 6 displayed elevated expression of known markers of cell proliferation, including Ki-67 (*MKI67*) and cyclins, as well as genes associated with tumor progression, invasion, and proliferation, including *COL1A2*[58–60], *MUC1*[61–64], and *MMP11*[59,60,65] (Fig. 5E, Supplementary Figures 13, 15, 17). Clusters 4 and 9 showed increased expression of *ZNF703*, an oncogene in the more aggressive, ER+ luminal B breast cancer subtype[66,67] as well as *GRB2*, a gene implicated in breast cancer tumorigenesis[68,69], and *BAMBI*, a pseudoreceptor for *TGFβ*[70], whose signaling pathway is implicated in progression to invasion[61] (Fig. 5E). These spatial expression patterns suggest a transcriptional heterogeneity among compartments of invasive tumor inaccessible to histopathological analysis, demonstrating the superiority of spatial transcriptomic data over immunofluorescence alone.

Enhanced resolution clustering resolves keratinocyte structure in squamous cell carcinoma

Finally, we also used BayesSpace to analyze a squamous cell carcinoma (SCC) Visium sample first described by Ji et al.[48].Pathologist annotated H&E stained tissue identified the tumor borders and other major tissue structures (Supplementary Figure 18). We defined expression profiles for the major cell types present in the sample based on known marker genes from

literature: keratinocytes (*KRT1*, *KRT5*, *KRT10*, *KRT14*), melanocytes (*MLANA*, *DCT*, *PMEL*), myeloid cells (*LYZ*), and T cells (*CD2*, *CD3D*, *CD3E*, *CD3G*, *CD7*)[48,57]. The keratinocytes were further separated into basal keratinocytes (*KRT5*, *KRT14*) and suprabasal keratinocytes (*KRT1*, *KRT10*) since *KRT5* and *KRT14* form heterodimers that localize to the basal layer of the epidermis while *KRT1* and *KRT10* form heterodimers that localize to the suprabasal layer[71]. We show that our enhanced spatial gene expression plots delineate the border between the basal and suprabasal layers more precisely than the spot-level plots (Supplementary Figures 18 and 19), and similarly find that the enhanced expression of marker genes for melanocytes, myeloid cells, and T cells better match the expected patterns based on the annotated tissue structures (Supplementary Figure 18).

The high-resolution gene expressions are inferred from the enhanced clustering output of BayesSpace with twelve clusters (Supplementary Figure 18). This number is chosen based on the elbow of the pseudo-log-likelihood by cluster number plot as a way to measure how well the clustering partition fits the data (Supplementary Figure 19). We see high immune cell expression in clusters 4 and 8. Comparing these two clusters, differential expression analysis on the enhanced expression reveals that immunoglobulin genes are upregulated in cluster 8 (Supplementary Figure 18). The marker genes for cluster 4, such as *CXCL10*, *CXCL11*, and *ISG15*, are induced by interferon-γ. This suggests that plasma cells are enriched in cluster 8 while inflamed epithelial cells are present in cluster 4, which is a region in close proximity to T cell markers that are likely driving this inflammation. We also compare clusters 2, 3, and 6, in which the tumor is located (Supplementary Figure 19). In cluster 2, we see upregulation of *IGFBP2*, *IGFBP3*, and *IGFBP6*. These genes are members of the insulin-like growth factor-binding protein (IGFBP) family, which bind and stabilize *IGF1*, an important growth factor that

can promote tumor growth[72]. *IGFBP3* is also a known marker of apoptosis[73]. Similarly to cluster 4, in cluster 3 the chemokine *CXCL10* is upregulated, a chemokine that has been shown to be associated with response to radiotherapy and survival in SCC patients[74]. Clusters 3 and 6 also show spatial expression of *LCE3D*, *SPRR2A*, *SPRR2D*, and *SPRR3*. These genes are members of the epidermal differential complex, which contains genes responsible for keratinocyte development and are upregulated in cutaneous SCC[75].

BayesSpace enhances gene expression patterns to near single-cell resolution on in silico spatial data

We conducted two simulations to demonstrate that BayesSpace resolution enhancement outperforms existing methods.

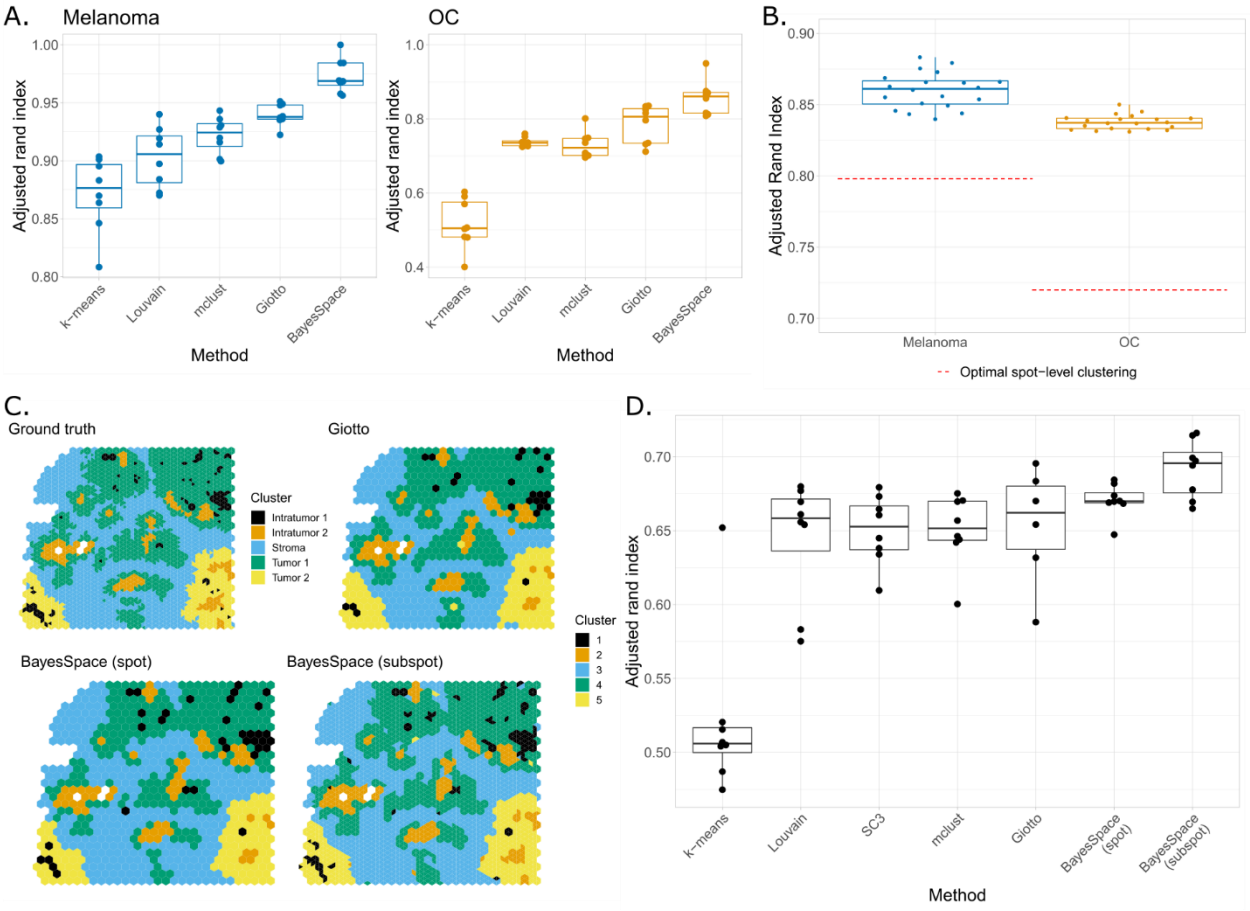


Figure 6. BayesSpace outperforms spatial and non-spatial clustering methods in simulated data. (A) In N = 8 replicates simulated from the melanoma sample and N = 8 replicates simulated from the OC sample, BayesSpace spot-level clustering outperforms other clustering methods. (B) In N = 20 replicates for the simulation done at the sub-spot level, BayesSpace enhanced clustering outperforms the optimal spot-level clustering (red dotted line). (C) In the third simulation using single cell data, the ground truth is derived from expert annotation of an IF staining image corresponding to the OC sample (top left). Examples of the clustering partitions generated by BayesSpace at the spot and subspot levels as well as by the next best method (Giotto) are also shown. (D) BayesSpace clustering at the spot level slightly outperforms competing methods while BayesSpace enhancement to the subspot level generally provides substantially higher performance than other methods in recapturing the ground truth clusters among the N = 8 simulation replicates. In all boxplots, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively.

In the first simulation, we show that BayesSpace enhanced resolution clustering outperforms the optimal clustering that can be achieved at the spot-level in melanoma and ovarian samples that are simulated at the subspot level (Fig. 6B). In each dataset, the enhanced clustering ARI exceeds the optimal spot-level clustering in all 20 simulated replicates. This indicates that BayesSpace is able to increase the resolution of the data to better recapture finer details of the ground truth.

In the second simulation, we demonstrate that BayesSpace enhanced resolution clustering can increase the resolution of data that are simulated from real, aggregated single cells (See Methods for details). BayesSpace better captures the spatial distribution of clusters than the

optimal spot-level clustering, as illustrated in the spatial representation of the enhanced clustering results from one replicate (Fig. 6C). In regions with high mixing of cell types, there is little to no information available to resolve the cluster labels at the subspot level, but BayesSpace is still able to closely approximate the overall tissue structure at the spot level. In these cases, though it is easy to miss the isolated cells due to the signal being diluted out from the aggregation of multiple cells at the spot level, we find that BayesSpace is still able to recover some of these populations. The simulation results further support our melanoma analyses where our enhanced analysis recovers lymphoid structure near the tumor that was not apparent at the spot level. In all, BayesSpace enhanced clusters better recapture the ground truth than all other methods, again highlighting the superior performance of our method (Fig. 6D) and showing that BayesSpace is able to successfully enhance the resolution of spot level data.

INTEGRATION WITH IMAGE DATA IMPROVES RESOLUTION ENHANCEMENT

We apply the image-assisted resolution enhancement to the OC and SCC datasets. For the OC dataset, using 8 clusters, we perform resolution enhancement with only expression data and resolution enhancement with both expression and imaging data (Fig. 8A). The image-assisted version contains noticeably more mixing between clusters. As before, we dichotomize clusters to be either high or low for CD45 and then compare with the anti-CD45 stain to validate the results. In all three ROIs examined in Figure 4I, we observe a high level of correspondence between the anti-CD45 stain intensity and subspots being classified in CD45 high clusters (Fig. 8B). These results represent further improvement over both the spot-level clustering and non-image enhancement. After mapping the subspot-resolution PCs to genes, we note that the predicted *PTPRC* expression has higher correlation to the anti-CD45 stain (Fig. 8C) when the subspot

model includes image data (R = 0.468) compared to the observed spot-level data (R = 0.151) and the original subspot-level model (R = 0.400).

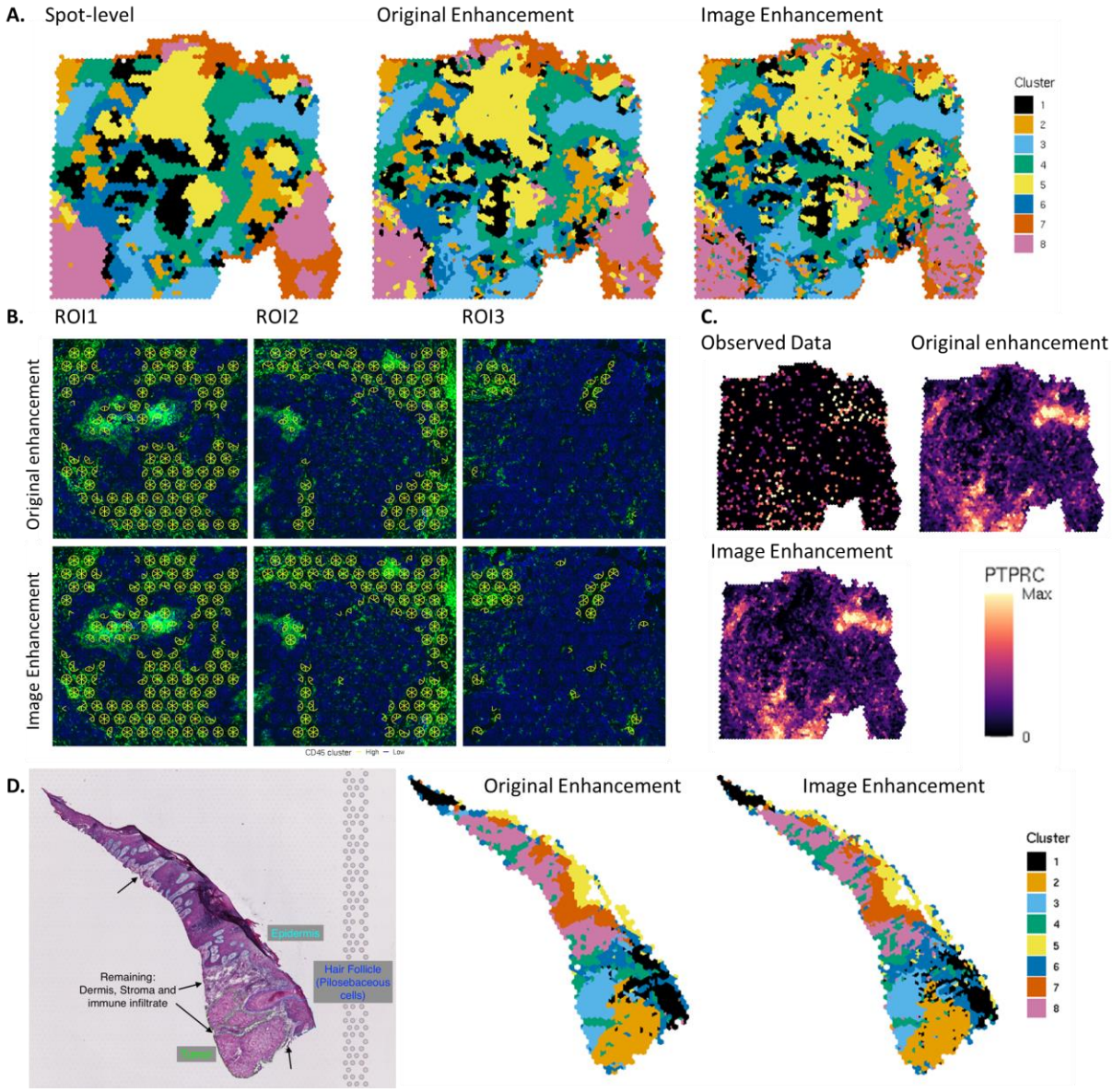


Figure 8. Subspot-level enhancement improves when incorporating image features. (A) Additional heterogeneity in the clusters is present from the spot-level clustering to the original enhancement to the image enhancement. (B) In all three ROIs, image enhancement clusters better correspond to anti-CD45 stain intensity. (C) Both subspot-level enhancement procedures show higher correlation between *PTPRC* and anti-CD45 stain intensity compared to the observed spot-level data. (D) Histopathology features shown in the H&E can be compared with enhanced resolution clusters.

For the SCC dataset, the autoencoder was used to extract an efficient low-dimensional representation of the image data. Comparing the clusters enhanced using only spatial gene expression with the clusters enhanced using image data as well, the image model clusters appear to identify more biologically relevant features (Fig. 8D). Notably, the finger-like stromal structures that are prominent in the top half of the sample are better captured by the image model clusters.

Discussion

To the best of our knowledge no previous study has achieved sub-spot resolution of spatial transcriptomics data without requiring the use of additional information besides spatial coordinates. The immunohistochemical analyses in the IDC and OC tissue sections provide validation that our subspot model accurately refines and reflects the spatial structure of the underlying tissue. Enhancement of gene expression analysis at subspot resolution allows downstream differential expression analyses to compare finer and more biologically meaningful clusters. Our analyses of differential expression in the IDC tissue section identify transcriptional heterogeneity within regions of invasive tumor that appear histologically indistinct. While the

histological analysis of this tissue was limited by the available immunofluorescent stains, notably lacking a tumor marker or H&E stains, our results suggest the potential for spatial transcriptomics and BayesSpace to capture previously uncharacterized spatial patterns of gene expression.

Furthermore, we demonstrate a way to incorporate information from images into a model for spatial gene expression data. Whereas the gene expression data is low-resolution and high-throughput, the images are high-resolution and low-throughput, so it is advantageous to model both sets of information together. While we explored two approaches for image feature extraction in this work, more sophisticated image processing and computer vision approaches should be pursued.

BayesSpace seamlessly integrates into the spatial transcriptomics analysis workflow by taking as input preprocessed data via the widely used Bioconductor SingleCellExperiment data structure. The output is likewise stored in a SingleCellExperiment object that can be used for downstream analyses. The methods are all implemented as an R package that is openly accessible on Bioconductor.

The resolution enhancement approaches single cell resolution, with approximately 3 cells per subspot for data acquired with the Visium platform, without the need for external single cell data. However, there is potential for the enhanced data to be integrated with external single cell data through deconvolution or label transfer methods. For example, it may be possible to enhance the resolution of spot-level cell type proportion estimates by using a Dirichlet regression model with enhanced PCs as predictors. Integration with single cell data has the potential to improve our ability to resolve cell types in dense and complex tissues. This approach is explored in the next chapter.

Chapter 4: Jointly model spatial transcriptomics with non-spatial single cell RNA-sequencing

Introduction

Spatial transcriptomics and scRNAseq can be complementary technologies. While the former can spatially resolve gene expression, it only measures the pooled expression of cells within a spot, which may number in the dozens[32]. This makes it challenging to identify patterns of expression within cell types of interest. In contrast, the scRNAseq profiles the gene expression of single cells and enables identification of fine cellular subpopulations, but the preparation process destroys all spatial localization information, which is necessary for understanding biological phenomenon within the context of the tissue[76]. A tool that jointly models paired spatial transcriptomics and scRNAseq data may be able to overcome the weaknesses of the individual technologies to ultimately deliver new biological insights.

Existing methods for integrating scRNAseq and spatial transcriptomics do not fully take advantage of the spatial nature of spatial transcriptomics data. Deconvolution methods for spatial transcriptomics evolved from (non-spatial) bulk RNA sequencing (RNA-seq) methods. These methods use cell-type labelled scRNAseq data to derive reference cell types that can be used to deconvolve bulk RNA-seq samples, or in this case spots. Such methods developed for spatial transcriptomics include SPOTlight and stereoscope, which use nonegative matrix factorization and negative binomial probabilistic modeling respectively to estimate cell type proportions within a spot, while largely ignoring the spatially correlated nature of the data and only resolving cell type proportions at the spot level[38,77]. Deconvolution methods also do not aim to map single cells onto spatial locations. Another method, Tangram, uses deep learning to align scRNAseq cells to locations on a spatial transcriptomics reference in a way that maximizes the correlation

between each gene in the two datasets, though resolution is again limited to the spot level[78]. While the method can be thought of as maximizing spatial correlation, it does not actually consider the spatial coordinates of the spots when performing the algorithm. Furthermore, the mapping is done independently from other analysis steps, even though one modality could help inform the other to cluster and annotate cells and spots, for example. More importantly, while Tangram claims to be able to map single-cells in the spatial space, it only uses spot-level data, and as such spatial information content is limited.

Methods

MODEL DESCRIPTION

We introduce an extension to BayesSpace that allows for joint modeling of scRNAseq and spatial transcriptomics data for high resolution clustering, deconvolution, and label transfer (Fig. 9)[79].

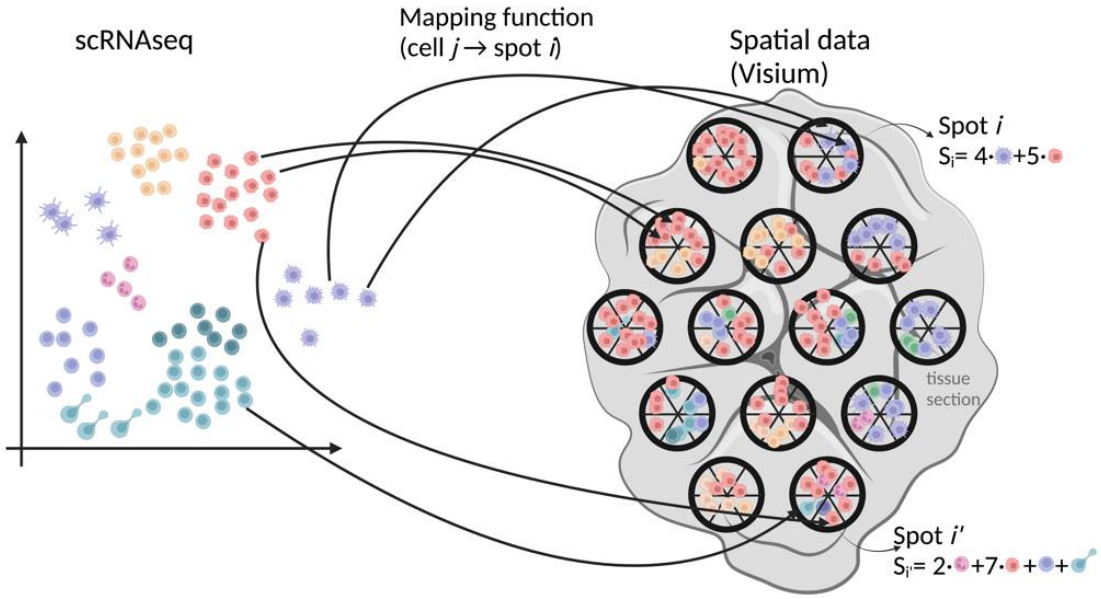


Figure 9. Integration of single-cell (scRNAseq) and spatial (Visium) transcriptomics. The goal is to estimate the function that maps single-cell data onto the SRT data. After mapping, each cell has an estimated x-y coordinate in the 2D spatial space. A spot can be written as the sum of the single-cell measurements within the spot.

We propose to jointly model low-dimensional representations of $\boldsymbol{Y}$, the spatial gene expression and $\boldsymbol{X}$, the single cell gene expression. Here, we use the first 15 PCs as in the original BayesSpace model, but other dimension reduction techniques could also be used. We assume that the data have been appropriately normalized so that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are on a comparable scale. As in BayesSpace, spots can be grouped into q clusters, and conditional on the cluster label, the observed data are approximately multivariate normal, i.e. Y_i is multivariate normal with mean $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\Lambda}$ given that $z_i = k$. In BayesSpace, we use a t-distribution instead of a normal distribution to account for outliers. For simplicity, we describe the model with a normal distribution, but the extension to the t-distribution would be straightforward, following the approach described in the original BayesSpace methods. As in the original BayesSpace approach, we use a Potts prior on the vector $\boldsymbol{z}$ to encourage neighboring spots to be of the same type, and parameters can be estimated via MCMC to make inferences about spot types and associated model features. However, by jointly modeling single cells with spatial transcriptomics, we aim to 1) learn the mapping from single cells to the spatial data and 2) jointly estimate spot types at the same time.

Conditional on the assumption that we know which cells map to which spots, let us define $\boldsymbol{S}_i = \sum_j I(l_j = i)\boldsymbol{X}_j$ the sum of the expression of all single cells j with locations l_j mapping to a spot i. Here $\boldsymbol{l} = [l_1 \quad ... \quad l_m]$ is a vector with length m equal to the number of

single cells, and each l_j can take values in $\{1,2,\dots,n\}$, where n is the total number of spots. Assuming that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been appropriately calibrated, we expect $\boldsymbol{Y}$ and $\boldsymbol{S}$ to share the same mean expression values, since spot expression values represent the aggregate of all cells within the spot in spatial transcriptomics data. As such, we define the following model:

$$\begin{pmatrix}\boldsymbol{Y}_i\\ \boldsymbol{S}_i\end{pmatrix}\Big| z_i = k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}, \boldsymbol{l}\Big) \sim N\left(\begin{bmatrix}\boldsymbol{\mu}_k\\ \boldsymbol{\mu}_k\end{bmatrix}, \boldsymbol{\Lambda}\right), \tag{2}$$

for each spot i belonging to cluster k with shared mean expression level μ_k and common precision $\boldsymbol{\Lambda}$ that captures correlation within and across data modalities. The model can allow a flexible and variable number of cells per spot. The number of cells per spot can be defined uniformly, based on the library size of the spot, or based on cell segmentation using a paired image. Prior specification and parameter estimation for $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{z}$ would be the same as in BayesSpace, the main difference being that both scRNAseq and spatial data would be used, leading to a more robust estimation of unknown parameters thanks to the sharing. Specifically, the full conditional distribution of μ is given by:

$$\boldsymbol{\mu}_k|\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\psi} \sim N\Bigg(\left(\boldsymbol{\Lambda}_0 + n_k(\boldsymbol{\phi} + \boldsymbol{\rho} + \boldsymbol{\rho}^T + \boldsymbol{\psi})\right)^{-1}\Bigg(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \boldsymbol{\phi}\sum_i \boldsymbol{Y}_i + \boldsymbol{\rho}^T\sum_i \boldsymbol{Y}_i + \boldsymbol{\rho}\sum_i \boldsymbol{S}_i + \boldsymbol{\psi}\sum_i \boldsymbol{S}_i\Bigg), \boldsymbol{\Lambda}_0 + n_k(\boldsymbol{\phi} + \boldsymbol{\rho} + \boldsymbol{\rho}^T + \boldsymbol{\psi})\Bigg),$$

Where we define $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, and $\boldsymbol{\rho}$ as components of the precision matrix $\boldsymbol{\Lambda} = \begin{bmatrix}\boldsymbol{\phi} & \boldsymbol{\rho}\\ \boldsymbol{\rho}^T & \boldsymbol{\psi}\end{bmatrix}$, and $\boldsymbol{\Lambda}_0$ and μ_0 are hyperparameters defined as in the original BayesSpace model. The Metropolis-Hastings algorithm can be used to estimate the spatial locations for single cells, exploring the parameter space by proposing moves for the cells across the spatial landscape of spots. We assume an

uninformative prior for the spatial distribution l_j of each cell j, with discrete uniform probability over all spots. Each proposed move will deterministically result in a proposed update for S, which is accepted or rejected based on the distribution specified in (2). We can update single cell locations with a series of three proposals generated by a location switch with a cell in 1) a neighboring spot, 2) a spot within the same cluster, and 3) just a random spot in the sample. Examples of these moves are shown in Supplementary Figure 24. The remaining parameters $\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{z}$ are estimated as in the original BayesSpace algorithm via Gibbs sampling and the Metropolis-Hastings algorithm[79]. As in the original BayesSpace model, the estimation can also be extended to subspot resolution by additionally estimating latent subspot level expression. The proposed model has several advantages. From a biological point of view, it would allow us to map single cells onto spatial maps at high resolution. Our model directly accounts for the fact that spot data are the sum of individual cell measurements. From a computational point of view, the super-resolution modeling would make that more tractable since we would only need to sum over a couple of cells rather than potentially more than 10, which would lead to difficult optimization problems when estimating the mapping function of cells to spatial data.

Given that there are differences between scRNAseq and Visium technology, there may be differences between gene expression measurements from the two platforms arising from technical factors. We propose two methods to account for these platform effects. First, existing batch effect correction tools such as Harmony can be used to project the two types of data into a shared space[80]. A second approach involves estimating the transformation as part of the joint model. Rather than modeling $\boldsymbol{Y}_i$ and $\boldsymbol{S}_i$ to have a common mean, we can instead apply a linear transformation $a_d + b \times \boldsymbol{S}_i$ to each dimension d, with a dimension-specific intercept $a_d \in \mathbb{R}$ and a common slope $b \in \mathbb{R}^+$. Note that the slope transformation here introduces a Jacobian term in

the log-likelihood: $+n \times d \times \log b$. The intercepts and slope can also be estimated using random walk Metropolis-Hastings. Thus, we model Y_i and the linearly transformed S_i and to have a common mean parameter.

The BayesSpace joint model of scRNAseq and spatial transcriptomics will generate several useful outputs. First, cell type annotations of the scRNAseq data can be transferred to the spatial transcriptomics data, allowing the user to localize cell subpopulations of interest within the spatial context of biological tissue. This can be done both discretely (by assigning the most likely cell type label to a spot or subspot) or continuously (by mapping the posterior probability of a cell type in a location). The newly aligned scRNAseq data also represents a new spatial dataset, which may be useful in cases when genes of interest are not adequately captured in the spatial data but are in the scRNAseq data. Finally, the additional information provided by scRNAseq data may improve clustering results, which were the primary output of the original BayesSpace method.

SIMULATION AND DATA ANALYSIS

We design two simulations to evaluate the performance of the joint model in mapping single cells to space. In Simulation 1, from a normal distribution, we simulate 100 spots arranged on a square lattice that belong to three clusters (Fig. 10A). Each element of each 15-dimensional true $\boldsymbol{\mu}_k$ is generated from a uniform distribution with support on $[-10,10]$. The simulated $\boldsymbol{Y}_i$ with $z_i = k$ are normally distributed with mean $\boldsymbol{\mu}_k$and an independent covariance matrix with value 10 for all diagonal elements. Assuming 6 cells per spot, the single cell data $\boldsymbol{X}_i$ are then simulated by adding Gaussian noise with variance 1 to $\boldsymbol{Y}_i/6$ (Fig. 10B). The BayesSpace integration approach is run for 10,000 iterations using three clusters and default hyperparameters.

We evaluate results by considering the accuracy of mapping cells back to the correct location and the correct cluster.

In the second simulation, we have a more realistic setting of data from two serial sections (Fig. 11A). We take samples 151673 and 151674 from the DLPFC Visium dataset and treat 151674 as the single-cell dataset by removing the spatial location information (i.e. dissociating the sample *in silico*). Then, the BayesSpace integration approach is used to attempt to recover the lost spatial location information from sample 151674 by mapping the sample 151674 spots to sample 151673 locations. 10,000 iterations are run using the true cluster number of 7. Given that the spatial locations of the two samples are not exactly the same, we cannot directly evaluate mapping accuracy to the correct location. Instead, we propose two metrics to assess performance. The first involves transferring the true layer labels from 151674 to 151673 based on the location mapping. Then the ARI between the transferred labels and ground truth layers of 151673 is calculated, where higher ARI indicates better performance. The second metric involves manually aligning the two samples (Fig. 11B) and then calculating the mean over spots of the Euclidean distance between the mapped location on sample 151673 and the true location on 151674. Here, a smaller Euclidean distance indicates better performance.

Finally, we utilize the BayesSpace joint model to map single cells to spatial locations in real biological data of the human tonsil. In this analysis, we jointly model single cell and Visium data derived from a single donor (BCLL-10-T) first described in by Massoni-Badosa et al (2022) as part of the Human Cell Atlas[81]. The single cell data from donor BCLL-10-T include 32,076 cells which have been annotated with 9 cell type labels by the original authors (Fig. 12A). Since epithelial cells were underrepresented in this dataset, we also include 354 epithelial cells from other donors. The spatial data include a Visium sample with a total of 3,079 spots along

with annotations of major biological structures within the tissue (Fig. 12B). As a quality control step, we remove spots with under 2,000 total counts from the analysis as these spots represent low quality data generally arising from tears and folds in the tissue or other technical artifacts such as an edge effect. To reduce computational complexity, we first assume a simple model with one cell per spot and downsample the single cell dataset. Normalization and gene selection are done for the single cell and spatial data following the standard procedure of scaling by total counts and selecting highly variable genes. PCA is performed on the combined dataset and the top 15 PCs are used as input to the BayesSpace joint model with 7 clusters and run for 10,000 iterations. Finally, we present preliminary results running the joint model at the subspot resolution with 1 cell per subspot.

Results

PERFORMANCE OF JOINT MODEL IN SIMULATION

In simulation 1, single cells were simulated based on the spatial expression PCs at the true location of each single cell, though noticeable differences are present between $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ (Fig. 10B). Figure 9C shows that the cluster labels $\boldsymbol{z}$ are correctly estimated. Furthermore, we also display the mapped locations of 100 single cells, one generated from each spatial location. While not all cells are correctly mapped to the location from which they were generated, we note that there is substantial improvement in location accuracy from the initialization (0%) to the locations in the final iteration (63.67%) when calculating the accuracy over all 600 single cells (Fig. 10D). This value improves substantially when we take the mode over the chain as the point estimate of location, with an accuracy of 91.7%. Furthermore, we obtain very high accuracy (99.0%) when only considering if a cell is mapped to the correct cluster (Fig. 10E).

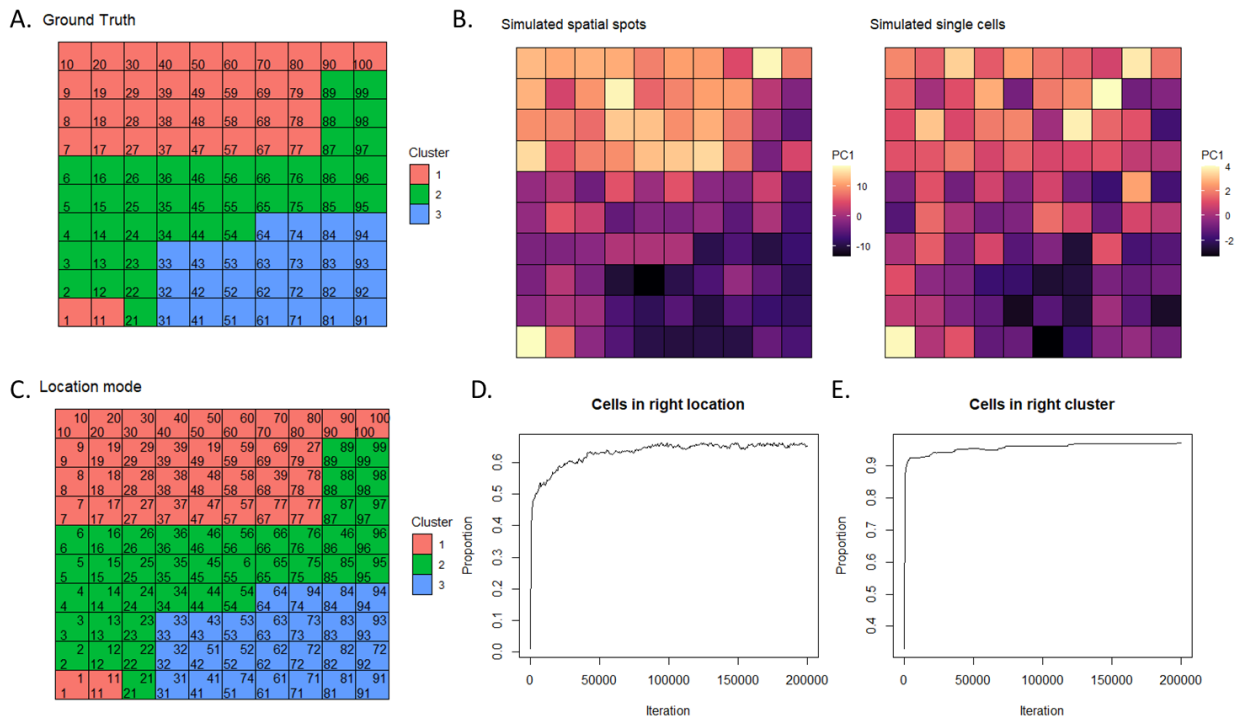


Figure 10. BayesSpace joint model performance in simulation 1. (A) The ground truth clusters are plotted spatially and each spot is numbered in the bottom left corner. (B) The first principal component of the simulated spot and single cell data are shown. (C) The results from the joint model are shown, with spots colored by the estimated cluster label. The top right corner annotation indicates the mapped location for a cell that was generated from the spot. Thus, cells mapped to the correct location have matching numbers between the top right and bottom left. For visual clarity, only cells 1-100 are shown out of a total of 600. (D) The cell location accuracy in each iteration is shown. (E) In each iteration, the proportion of cells assigned to a spot that is the correct cluster is shown.

In practice, with present technology, single cell and spatial gene expression profiling observations will not come from the same cells or even the same locations as the tissue section cannot be reused for more than one experiment. Instead, in the best-case scenario of paired data,

the single cell and spatial profiling will be done on adjacent sections of the tissue as in samples 151673 and 151674 of the DLPFC dataset. In simulation 2, to mimic the problem of mapping scRNAseq to spatial data, we hide the spatial locations of sample 151674, which serves as the simulated "single cell" dataset. As noted earlier in Chapter 2, batch effects exist between the two samples (Fig. 11C), which we can also account for using the linear transformation between the spatial and "single cell" PCs. Notably, PC2 of 151674 is shifted substantially downward relative to 151673, while the scale of the PCs appears to be comparable between the samples for both PC1 and PC2.

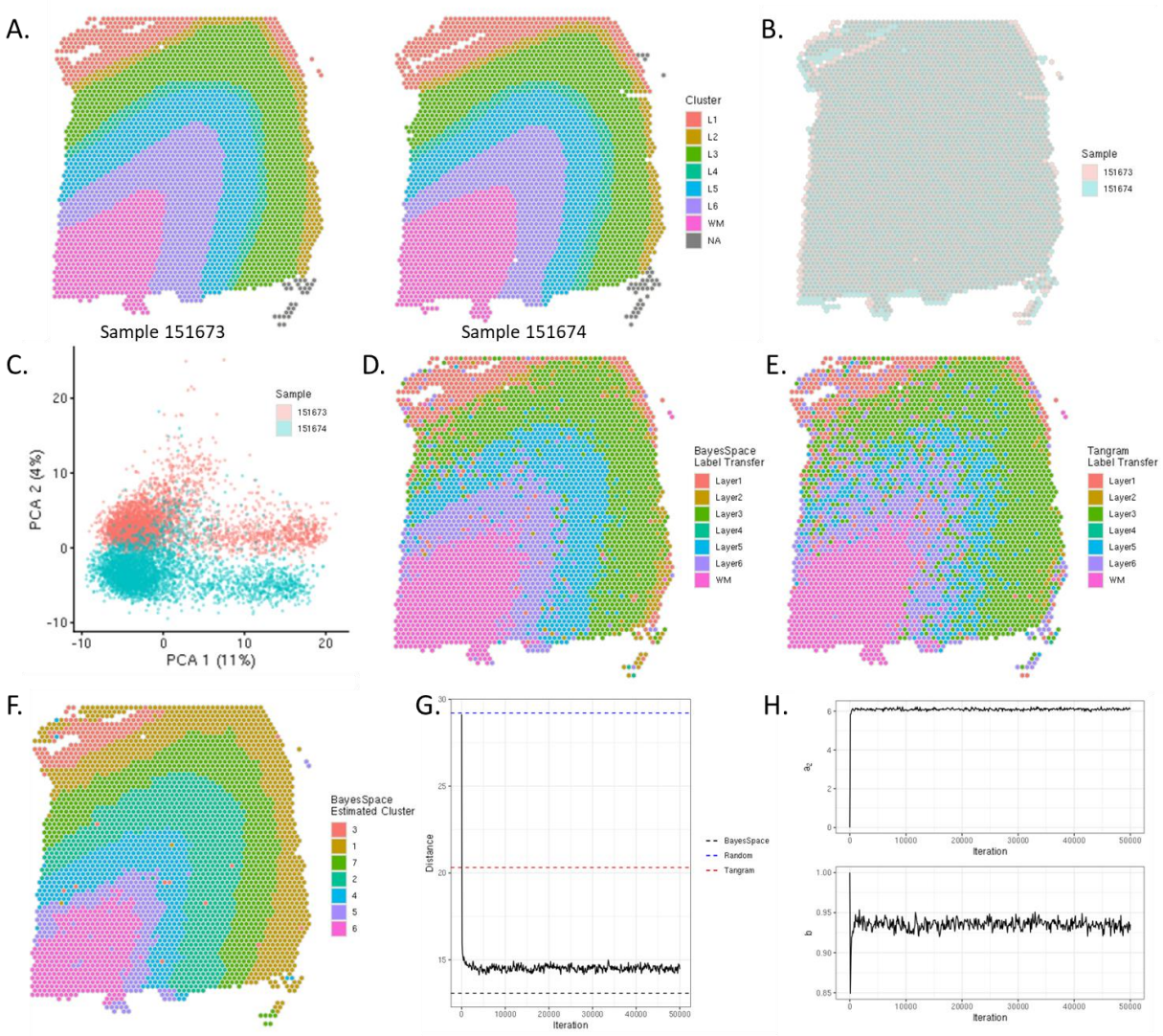


Figure 11. BayesSpace joint model performance in simulation 2. (A) Ground truth layers for samples 151673 and 151674 show striking similarity. (B) Spatial alignment between samples is good, though not perfect. (C) A sample specific batch-effect is evident from the PCA plot, particularly for PC2. Label transfer results from BayesSpace (D) show clearer layer structure than those from Tangram (E). (F) Jointly estimated clusters from BayesSpace. (G) The trace of the distance metric for BayesSpace is shown, along with lines for the final estimates of BayesSpace, Tangram, and a random mapping. (H) Trace plots for the intercept for the second PC (top) and the slope (bottom).

We apply the BayesSpace joint model as well as Tangram to map data from sample 151674 to spatial locations on sample 151673. Comparing ground truth layer labels transferred from sample 151674 to ground truth layers of 151673, BayesSpace finds good correspondence between the two partitions with ARI = 0.545 (Fig. 11D). In contrast, label transfer using Tangram resulted in noisier clusters with less apparent layer structure, resulting in ARI = 0.383 (Fig. 11E). These label transfer results can be compared to the jointly estimated clusters, which BayesSpace also provides (Fig. 11F). While the jointly estimated clusters have clear layer structure, the correspondence with the ground truth is slightly lower (ARI = 0.489) than the label transfer approach. The jointly estimated clusters may be identifying other biological phenomenon not captured within the layer information. Also note that the label transfer approach benefits from having accurate labels pre-defined in sample 151674. When evaluating mapping accuracy using the distance metric, BayesSpace again outperforms Tangram (Fig. 11G), with a mean distance of 13.65 between true and mapped locations for BayesSpace and 20.32 for

Tangram. For comparison, a random mapping would have a mean distance of approximately 29. BayesSpace quickly improves from this random starting point, and averaging over the MCMC chain provides slight further improvement compared to the result in individual iterations. Finally, we evaluate the estimation of the transformation with trace plots for the intercept of PC2 (a_2) and the slope (b). The intercept quickly reaches its stationary distribution near 6, which corresponds well to the vertical shift observed in the PCA plot in Figure 11C. The slope is estimated to be close to 1, again reflecting the lack of noticeable scale differences between PCs from the two samples (Fig. 11H). Plotting the transformed $\boldsymbol{S}$ values against $\boldsymbol{Y}$, we see that intercepts are estimated well, and the common slope assumption is reasonable. (Supplementary Figure 25).

BayesSpace maps single cells from tonsil to spatial transcriptomics

We use the BayesSpace integration model to map tonsil scRNAseq data to paired Visium data. Compared to the original authors' annotations (Fig. 12B), clusters estimated jointly by the single cell and spatial data reveal additional heterogeneity within the sample while still capturing the major annotated spatial domains (Fig. 12C). For example, clusters 4 and 5 correspond well to follicles while clusters 1 and 7 correspond well to the epithelial region. Clusters 4 and 5 mapping to follicles may correspond to the light and dark zones of germinal centers respectively. Germinal center B cells (GCBCs) in the light zone are positive for CD83 while those in the dark zone are known to have high CXCR4 expression, matching the expression patterns observed in the spatial data (Fig. 12E)[82]. The inter-follicular zone also contains additional heterogeneity, with cluster 3 surrounding follicles, cluster 6 lying near the epithelial layer, and cluster 2 in regions with vascular structures.

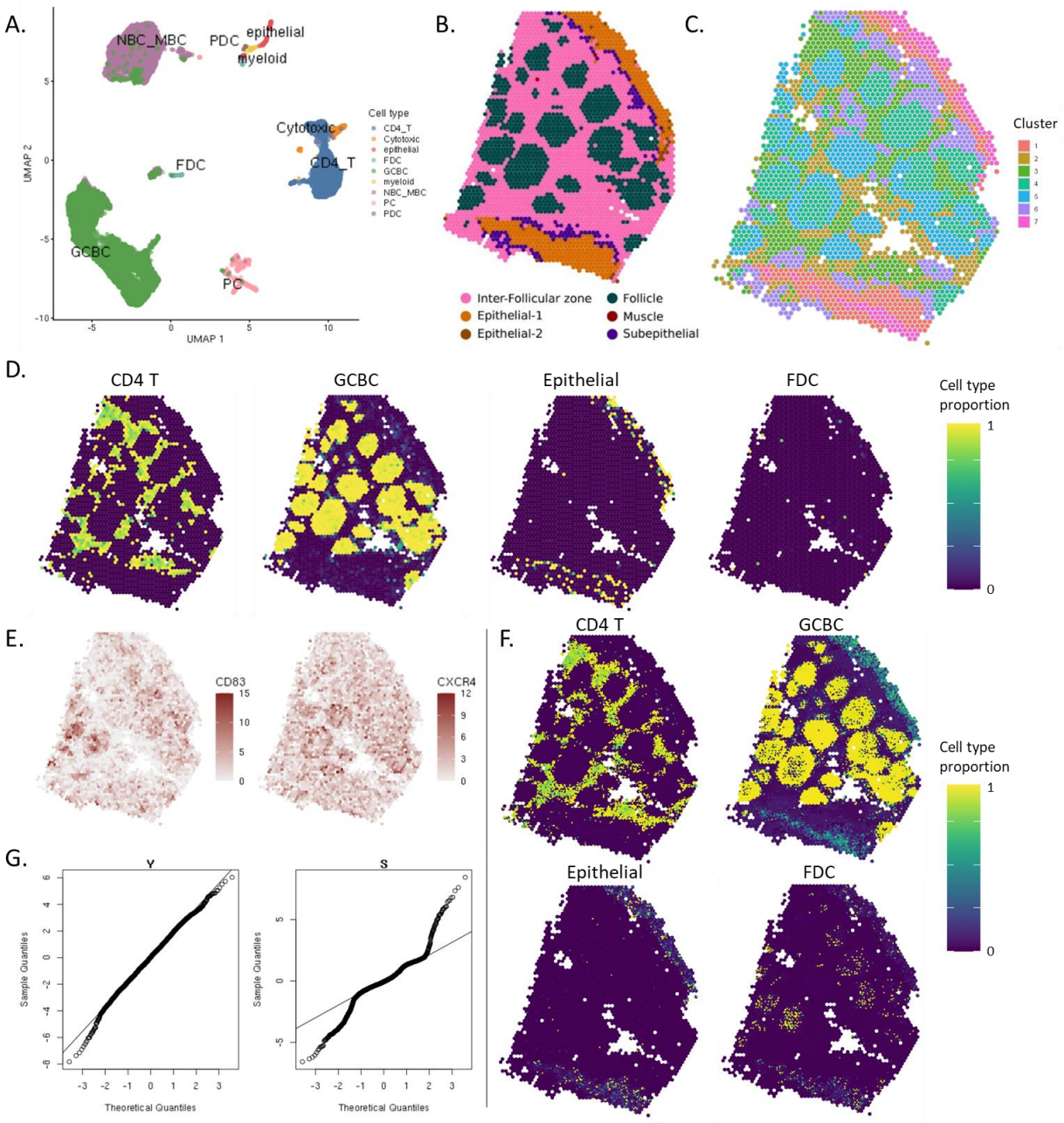


Figure 12. BayesSpace joint model to map tonsil single cell data onto space. (A) The UMAP representation of the scRNAseq data colored by cell type as annotated by the original authors[81]. We focus on the germinal center B cells (GCBCs), follicular dendritic cells (FDCs), epithelial cells, and CD4 T cells in this analysis. (B) Reproduction of figure from Massoni et al. (2022) showing annotated tissue structures[81]. (C) Joint estimated clusters. (D) Cell type proportions as

estimated through label transfer using the joint model. (E) Spatial plots of the expression for *CD83*, a marker for light zone B cells, and *CXCR4*, a marker for dark zone B cells. (F) Cell type proportions using the subspot resolution joint model. (G) Normal quantile-quantile plots for the spatial (left) and transformed single cell (right) data.

We transfer the labels from the single cell dataset to space, visualizing the probability of each cell type for all spots (Fig. 12D). For CD4 T cells, GCBCs and epithelial cells the cell type mapping matches biological expectation. Specifically, T cells are concentrated in T cell zones which surround follicles, GCBCs lie within follicles, and epithelial cells are localized to the epithelial layer. However, follicular dendritic cells (FDCs) do not have a clear spatial mapping in our analysis, though they are expected to localize to follicles. This may be due to FDCs being a small population within the scRNAseq data or because they typically only contribute a small fraction to the total expression within regions to which they should be localized. We attempt to address this by running the mapping at subspot resolution. While the localization of CD4 T cells, GCBCs, and epithelial cells remain largely the same as before, we note that FDCs now correctly localize to regions annotated as follicles (Fig. 12F). GCBCs previously dominated the estimated proportion in those areas.

Finally, we assess the normality of the residuals. The spatial data ($\boldsymbol{Y}$) are well approximated by the normal distribution, though there are substantial deviations from normality for the linearly-transformed single cell data ($\boldsymbol{S}$), with indication of heavier tails. Explanations for this deviation as well as potential remedies are explored in the Discussion.

Discussion

In this chapter, we present a novel approach for integrating scRNAseq and spatial transcriptomics data by mapping single cells to spatial locations through a joint model of the two data modalities. In both the DLPFC and the tonsil analyses, we demonstrate the ability of the model to jointly estimate cluster labels that match well with biological structures as well as transfer pre-existing scRNAseq annotations onto space. This latter feature can be particularly valuable when there are multiple levels or resolutions of annotations that can be transferred. For example, in the tonsil analysis, we found heterogeneity in the follicles to which GCBCs were mapped, indicating the existence of GCBC subpopulations that could be labelled more finely. Furthermore, differential expression can be performed for a cell type that is mapped to different spatial clusters to elucidate how cell type-specific gene expression varies over space.

While there are clear strengths to this integration approach, we also note a few weaknesses. Given that the single cell and spatial data are from two distinct platforms, it is challenging to jointly model the two data types in the setup of a clustering problem, as we see from the poor fit of the single cell residuals to the normal model. Using a t-distribution instead of the Gaussian may be more appropriate for fitting the heavy tailed residuals, though it may be challenging to find a degree of freedom that fits data from both modalities well. Instead, there may be some easier solutions. The first approaches the solution from the experimental design perspective and involves generating well-matched single-cell and spatial data. 10x Genomics recently released a new kit to generate SRT and single-cell data from formalin-fixed, paraffin-embedded (FFPE) tissue with the same chemistry, providing greater concordance between the two modalities. This will significantly accelerate the generation of paired spatial and single-cell data from adjacent tissue slices. We expect that the BayesSpace joint model will be a useful tool in the analysis of such data.

While it is advantageous to have well-matched single-cell and spatial data to minimize differences between the two, this is not always possible. Even when paired data exist, there can be significant technological variability. As part of the joint model, we proposed to learn transformations to make single-cell data more comparable to spatial transcriptomics. Here, we described a simple linear transformation, but non-linear transformations could also be used. Such transformations may be important for unmatched data that is more challenging to align using a linear transformation. By changing the underlying distribution of the PCs, a better-fitting joint model for single cells and Visium data may be possible.

Chapter 5: Future directions

In this dissertation, I developed a series of Bayesian statistical methods to address current challenges in the analysis of spatial transcriptomics data.

As the field of spatial transcriptomics continues to grow, more complex datasets are being generated[81] and new spatial expression profiling technologies are being developed[35]. Increasingly, studies involve multiple samples and serial section replicates, where samples are profiled using spatial transcriptomics alone or as part of a larger multi-omics effort. To enable analyses of these data, we have developed methods than allow joint clustering of multiple samples as well as joint modeling of data from different modalities. Further methods development will be necessary to jointly cluster serial section replicates in three dimensions, potentially by using a 3-D Potts model. As the size of these datasets continue to grow, there will be an increasing need for computationally efficient analysis tools. In the context of MCMC, computational time may be improved by parallelization and jointly updating multiple parameters when possible. In addition, rather than sweeping over all spots systematically in each iteration, we can instead only consider a spot-specific parameter for an update with a geometrically distributed probability unless the parameters of its neighbors change, as done in the clock method first described by Ripley and Kirkland (1990)[28].

In Visium data, the H&E images have much higher resolution data than the accompanying spatial expression data. Most studies only use these images for validation of results derived from analysis of the expression data[4,81,83]. As we showed in Chapter 2, including high-resolution image data as part of the analysis can be used to improve the resolution of the expression data as well. While we used only rudimentary approaches for image processing, more sophisticated methods for feature extraction from images may yield further insights. For

example, cell segmentation algorithms detect cell boundaries and would enable characterization of cells beyond a simple mean intensity of each color channel[84]. Furthermore, convolutional neural networks can potentially be trained to use images of cells to predict their phenotypes[85]. We provide a framework to easily include these additional image features as part of the resolution enhancement model.

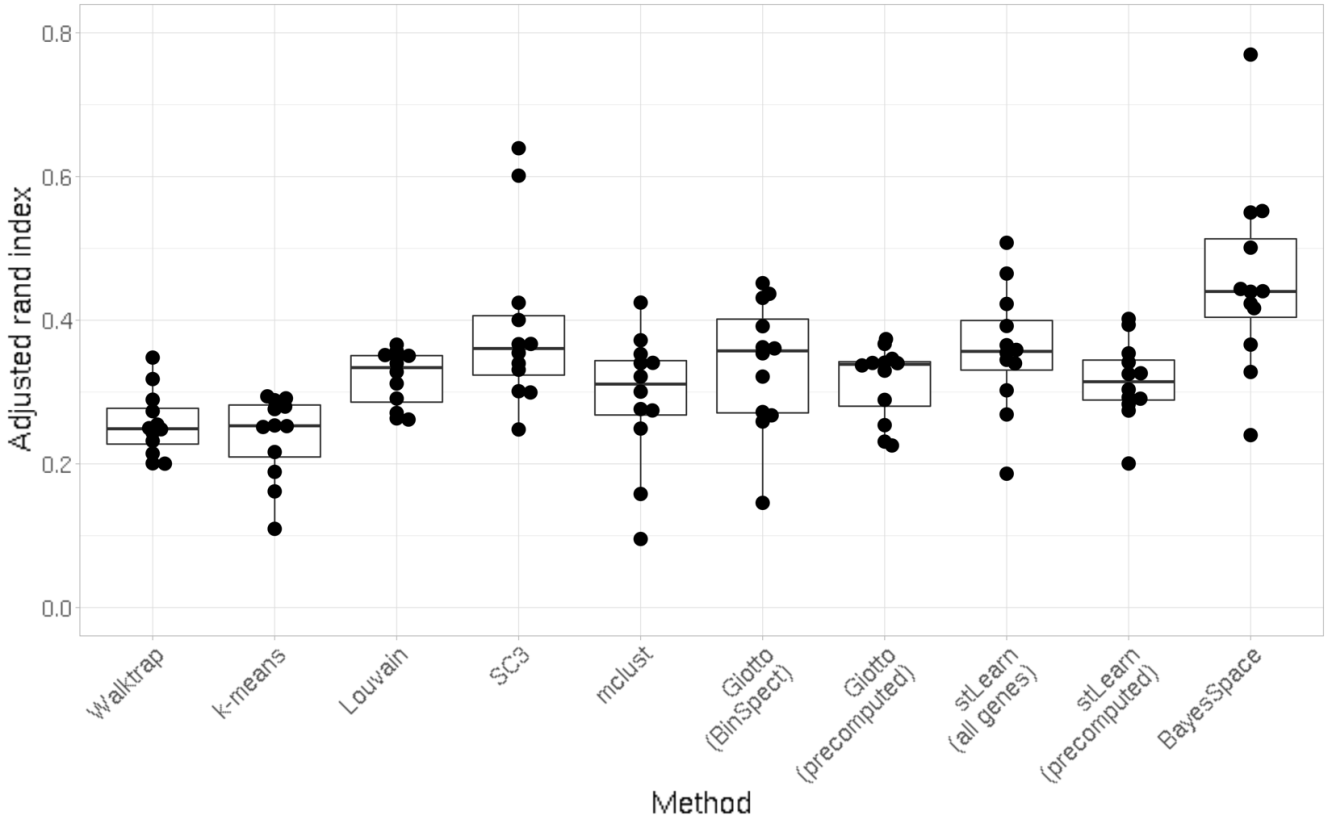
Recently, commercial image-based spatial expression profiling technologies have been developed, though it is not yet clear if these emerging technologies will surpass Visium, the current leading commercial spatial transcriptomics product[35,86]. While image-based spatial transcriptomics will bring forth a new set of analytical challenges, the BayesSpace framework will still be applicable for spatial clustering analysis of data from these new technologies though updates to the specification of the Potts model may be necessary. Notably, image-based technologies will profile expression at single-cell resolution, thus obviating the need for resolution enhancement. However, the mapping of scRNAseq data to space will still be important, as emerging image-based technologies only profile around 100 to 1000 genes, rather than the whole transcriptome. Mapping approaches may enable the imputation of the whole transcriptome from the much lower plex of the spatial expression data.

Finally, spatially resolved transcriptomics data provide a natural way to study how cell proximity impacts gene expression. Such information can allow researchers to, for example, better understand interactions between immune and cancer cells within the tumor microenvironment, which is eventually important for predicting clinically relevant outcomes such as disease progression[87]. Research on methods for cell-cell communication inference in the context of spatial transcriptomics has been limited and the lack of single-cell resolution in Visium data has made this work challenging[88,89]. However, the resolution enhancement that

BayesSpace provides may enable methods developed for single cell resolution spatial technologies to also be applicable to Visium. Ultimately, the development of statistical models for cell-cell interaction will be important in facilitating biological interpretation and accurate statistical inference.

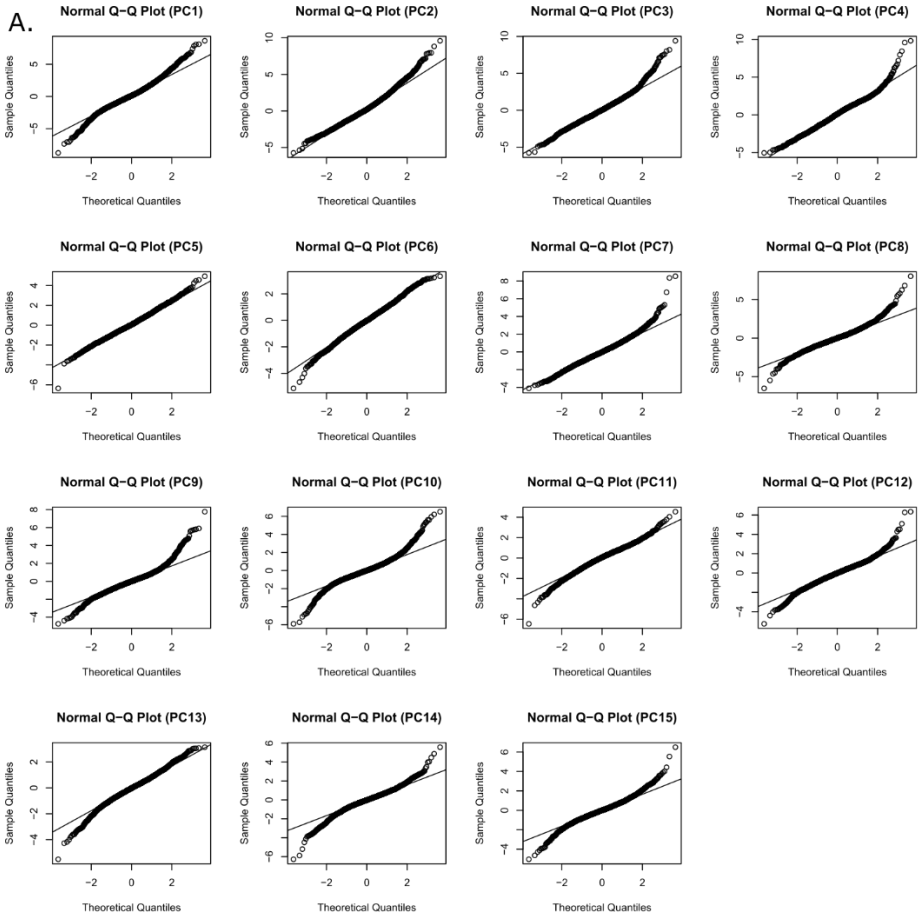
Supplementary Figures

Supplementary Figure 1. For a fairer comparison with BayesSpace, Giotto and stLearn are run on the precomputed PCs from BayesSpace. This approach does not improve the performance of either method. The comparison is run on the N = 12 tissue sections from the DLPFC dataset. In the boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively.

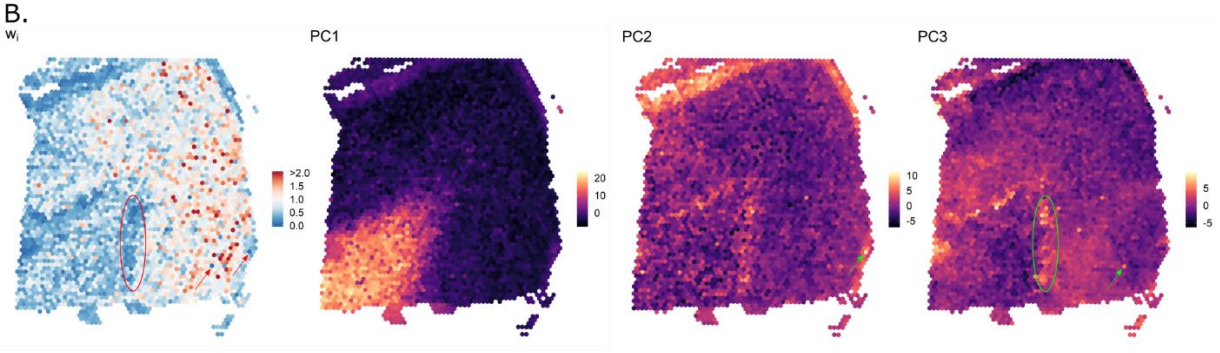


Supplementary Figure 2. Assessing BayesSpace modeling assumptions. (A) Q-Q plots of model residuals for sample 151673 demonstrate that most of the principal components are heavy-tailed, indicating the need for the model to assume *t*-distributed errors. (B) The spatial distribution of w_i's is shown for sample 151673. Notice that many of the visible outlier spots seen in PC2 and PC3 (some of which are annotated by the green oval and green arrows) are downweighted by BayesSpace as indicated by the w_i value. The corresponding locations on the w_i plot are annotated in red.

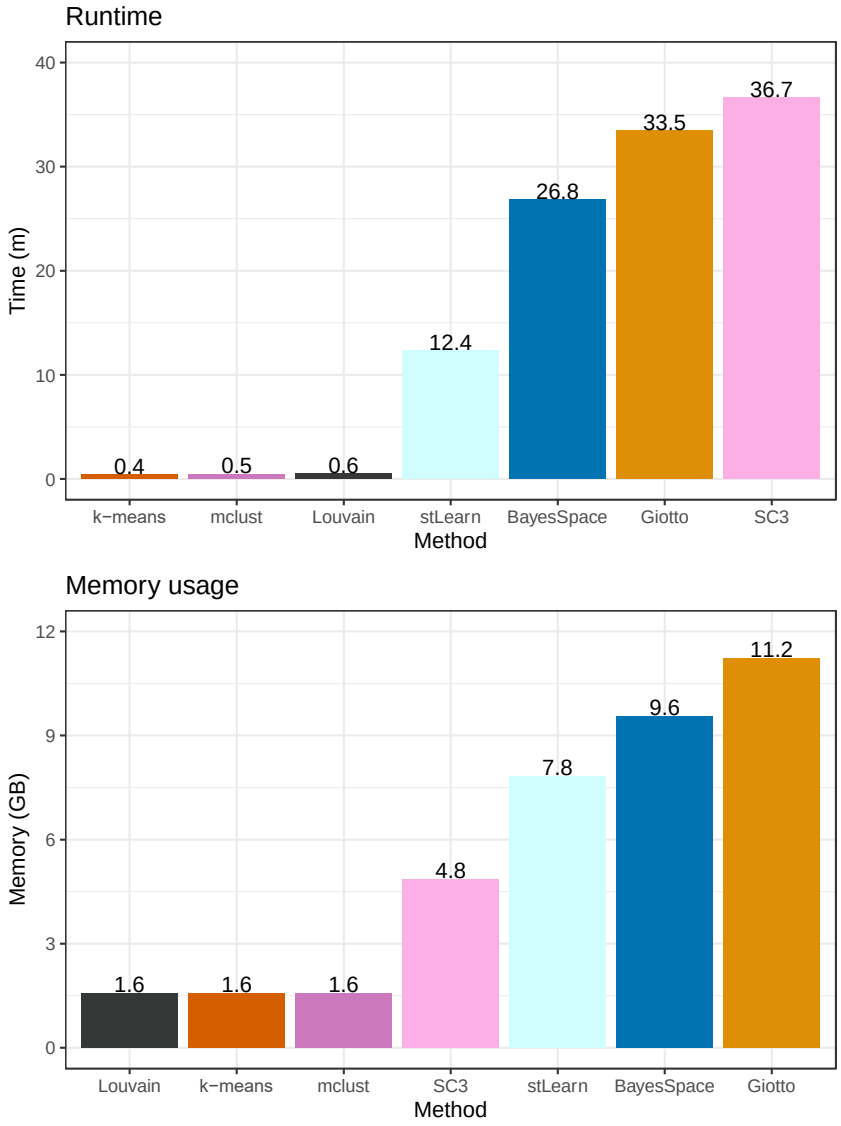
A.


Normal Q−Q Plot (PC1)
Normal Q−Q Plot (PC2)
Normal Q−Q Plot (PC3)
Normal Q−Q Plot (PC4)
Normal Q−Q Plot (PC5)
Normal Q−Q Plot (PC6)
Normal Q−Q Plot (PC7)
Normal Q−Q Plot (PC8)
Normal Q−Q Plot (PC9)
Normal Q−Q Plot (PC10)
Normal Q−Q Plot (PC11)
Normal Q−Q Plot (PC12)
Normal Q−Q Plot (PC13)
Normal Q−Q Plot (PC14)
Normal Q−Q Plot (PC15)
Sample Quantiles
Theoretical Quantiles

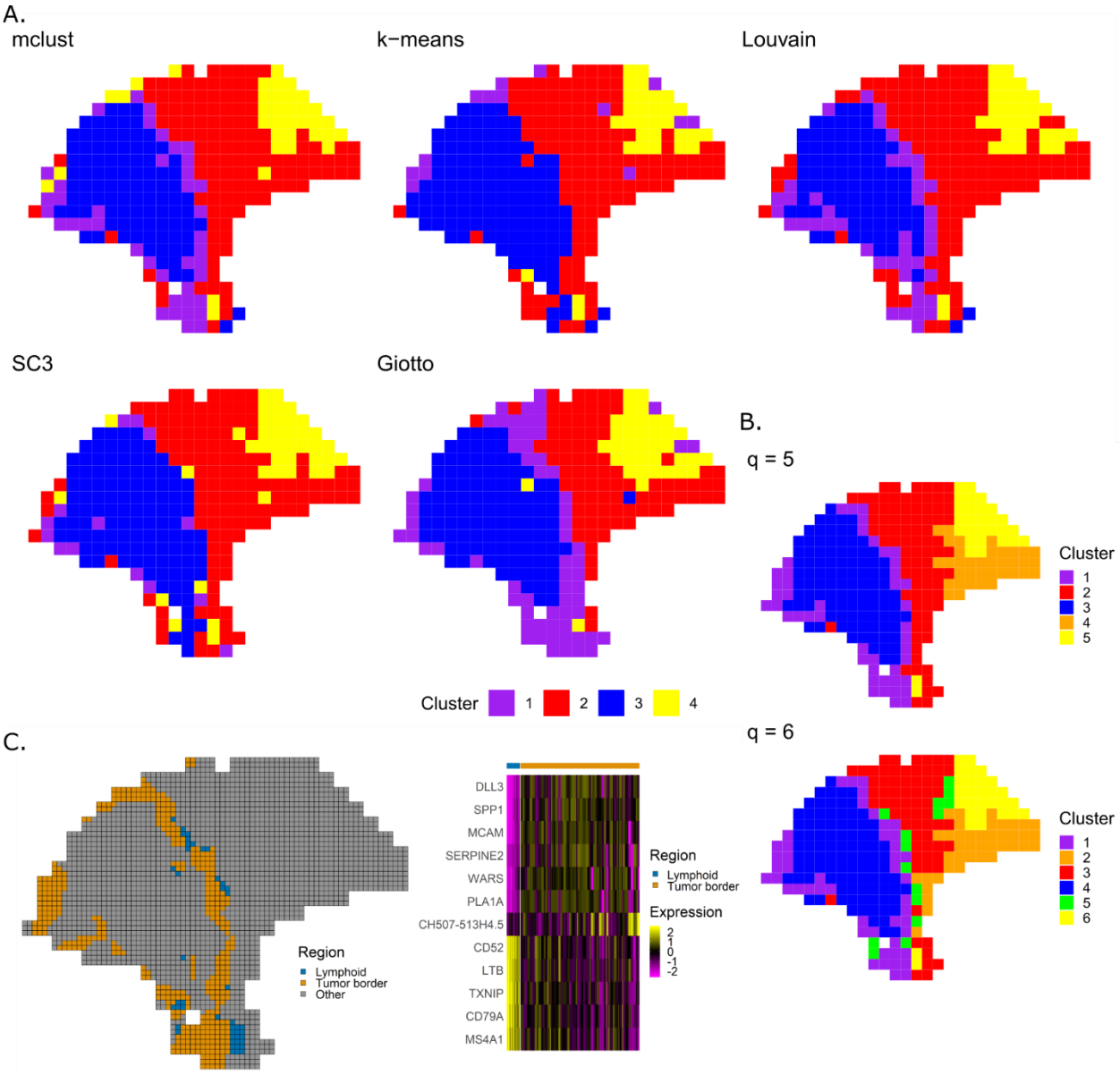
B.


w_i
PC1
PC2
PC3
>2.0
1.5
1.0
0.5
0.0
20
10
0
10
5
0
-5
5
0
-5

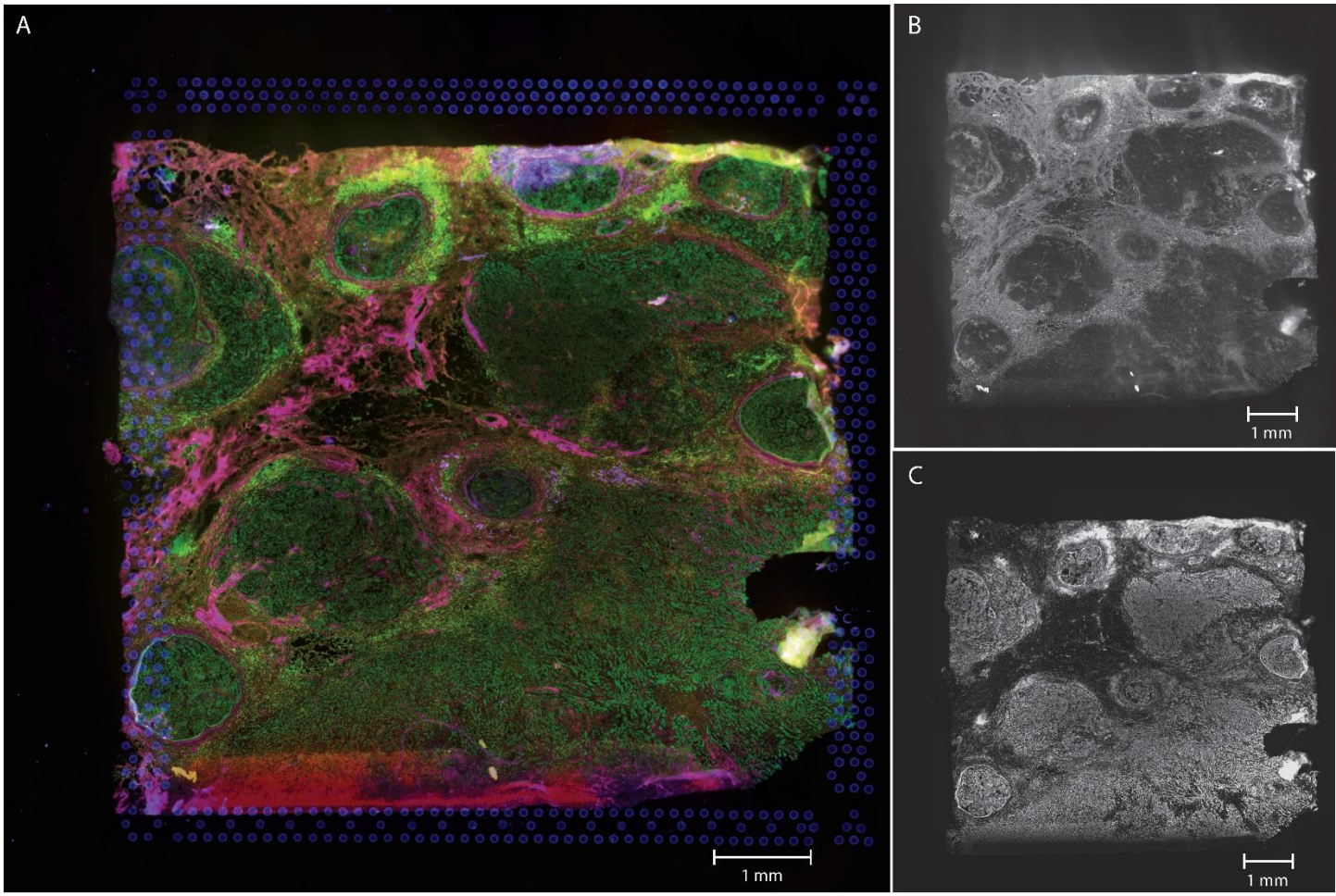
Supplementary Figure 3. Runtime and memory benchmarking of the evaluated clustering algorithms. All algorithms completed in under 40 minutes and required less than 12 GB of memory. A) Runtime of each algorithm (in minutes) on sample 151673 from the Maynard et al. DLPFC dataset. B) Memory consumption of each algorithm on the same sample.



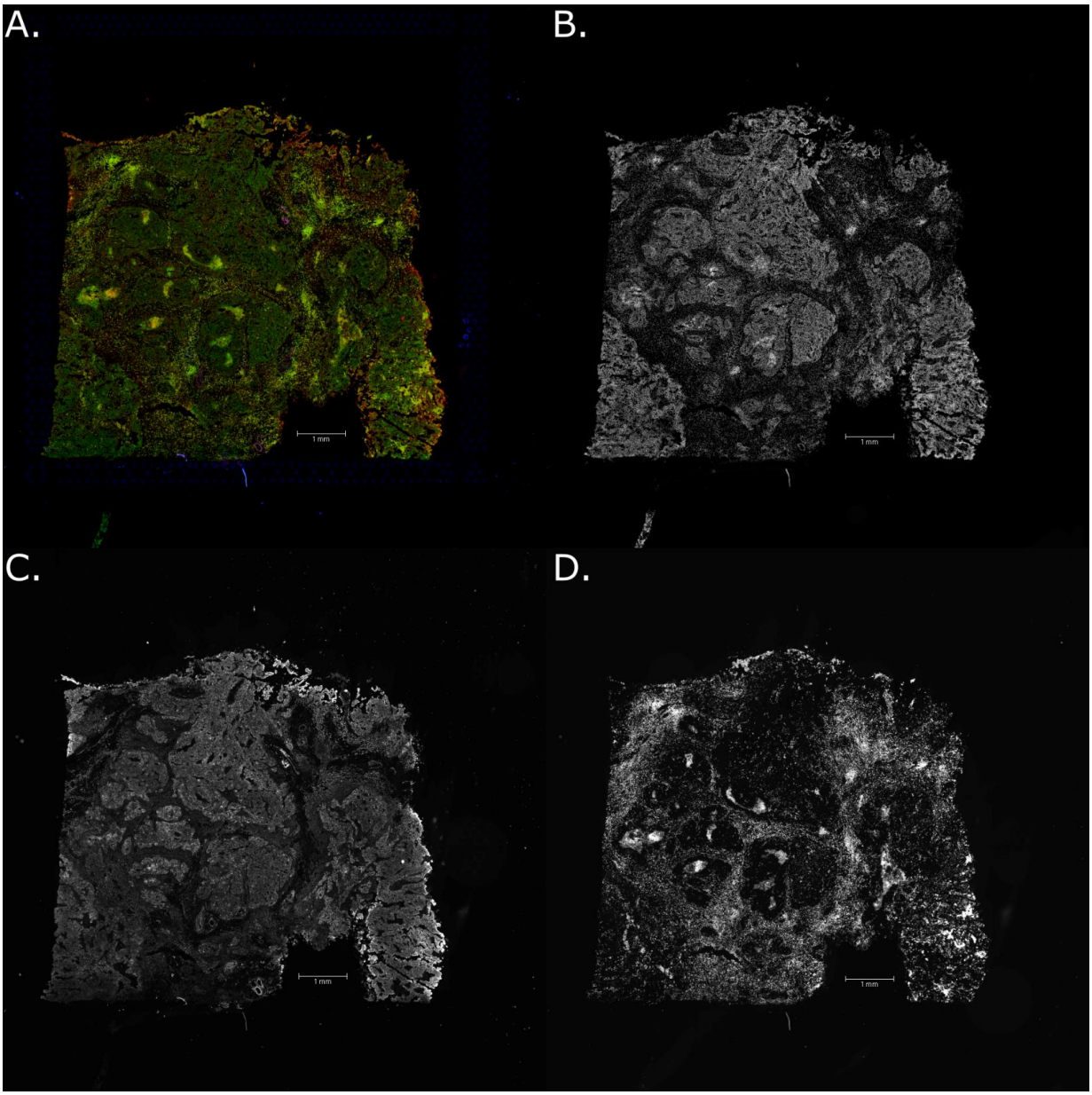
Supplementary Figure 4. (A) The results of other clustering methods are shown. For all methods, the spots are partitioned into four clusters to match the BayesSpace results in Figure 3. (B) BayesSpace results with 5 and 6 clusters are shown. (C) The expression of lymphoid regions identified near the tumor are compared to the expression of the remaining tumor border (left). The analysis reveals that lymphocyte markers are expressed more in the areas identified as lymphoid tissue (right).



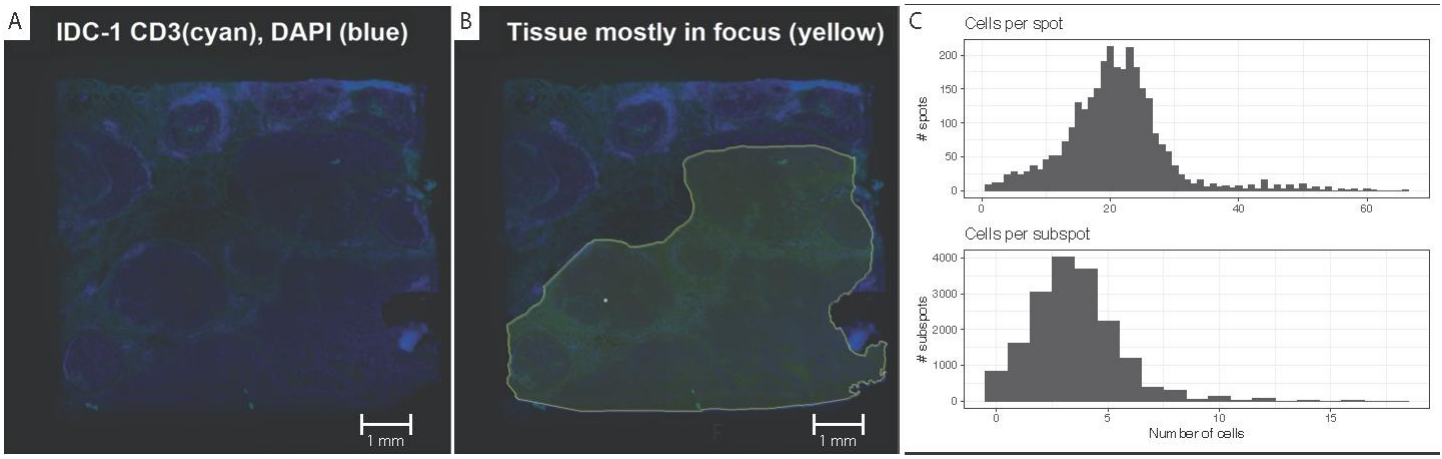
Supplementary Figure 5. Immunofluorescence staining of the invasive ductal carcinoma (IDC) sample (N = 1 tissue section, n = 4,727 spots). A) Four-channel image, depicting intensity of DAPI (green), the fiducial frame (blue), and CD3 (yellow). The FITC filter (magenta) does not correspond to an antibody stain. B-C) Single-channel images of anti-CD3 (B) and DAPI (C) stains. Intensity was scaled to [0, 1].



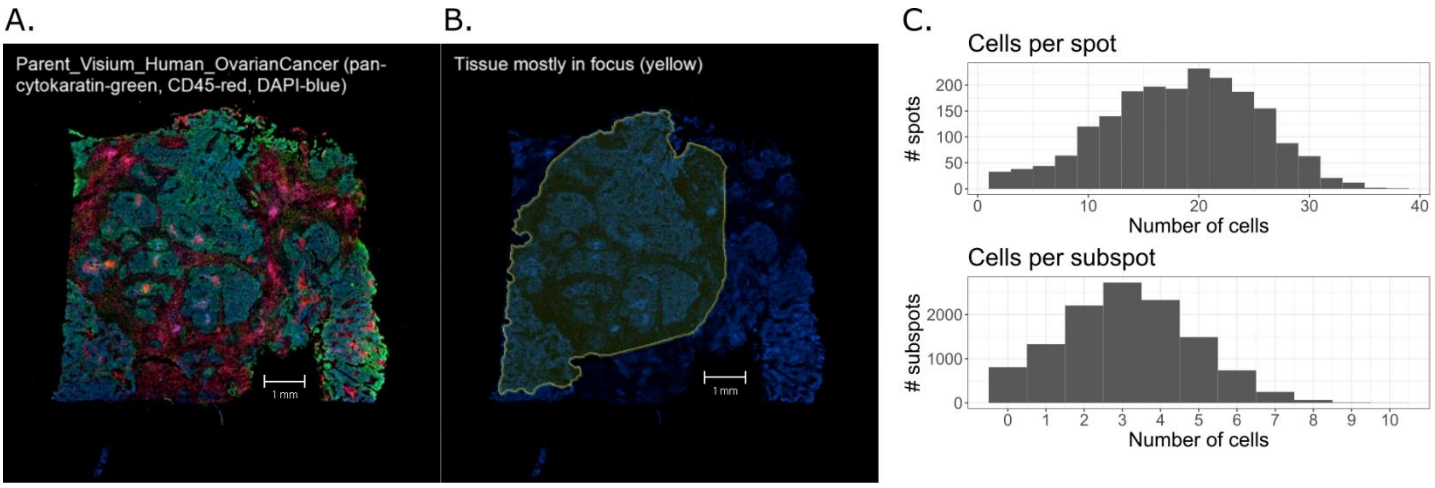
Supplementary Figure 6. Immunofluorescence staining of the ovarian cancer (OC) sample (N = 1 tissue section, n = 3,493 spots). A) Four-channel image, depicting intensity of DAPI (blue), cytokeratin (FITC/green), the fiducial frame, and CD45 (Cy5/red). B-D) Single-channel images of DAPI (B), anti-Cytokeratin (C), and anti-CD45 (D) stains.



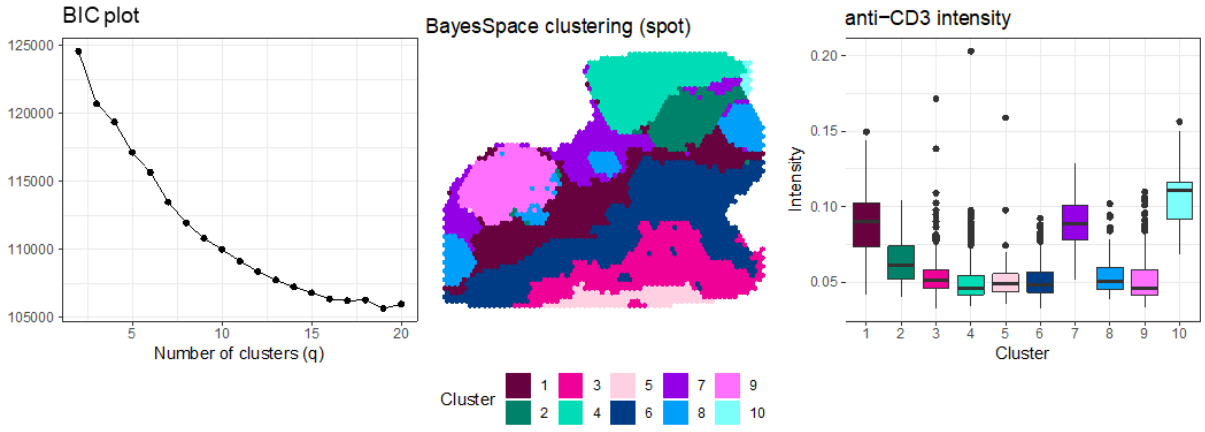
Supplementary Figure 7. Cell segmentation of invasive ductal carcinoma (IDC) based on immunofluorescence (IF) staining. A) Two stains, CD3 (cyan) and DAPI (blue) were used to segment cells. B) Cell segmentation was performed on the outlined tissue region that was in focus (in yellow); subsequent analyses of CD3 intensity in Figure 5 were restricted to this region (n = 2,929 in focus spots of 4,727 total). C) The distribution of cells per spot (top) and subspot (bottom). We found a median of 21 cells per spot (MAD=4) and 3 cells per subspot (MAD=1).



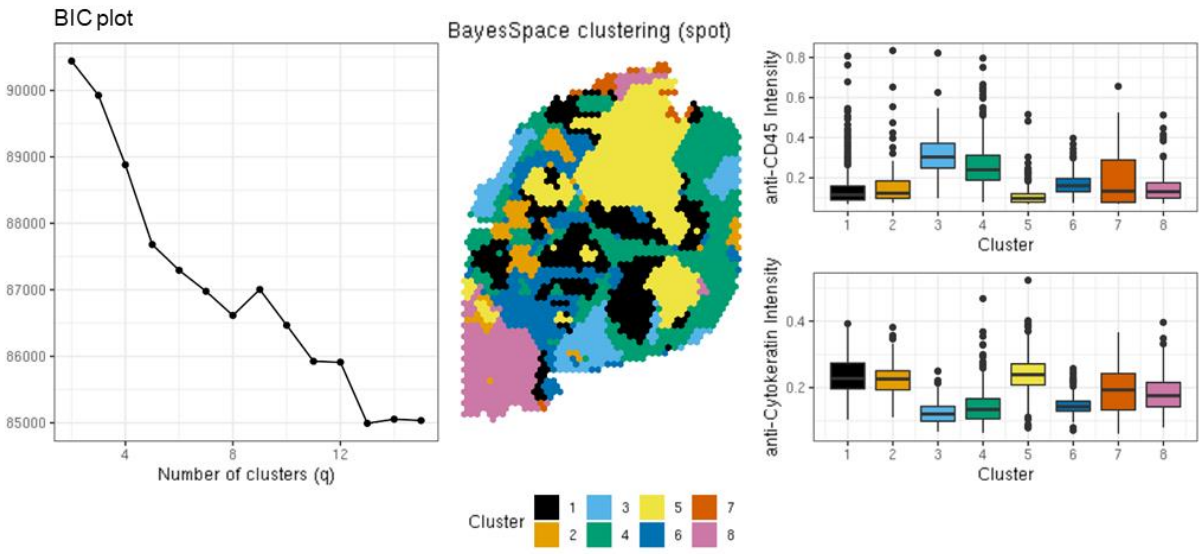
Supplementary Figure 8. Cell segmentation of the OC sample based on immunofluorescence (IF) staining. A) Three stains, Cytokeratin (green), CD45 (red), and DAPI (blue) were used to segment cells. B) Only a subset of the tissue was sufficiently in focus for cell segmentation (outlined in yellow); subsequent analyses of CD45 intensity in Figure 5 were restricted to this region (n = 2,041 in focus spots of 3,493 total). C) The distribution of cells per spot (top) and subspot (bottom). We found a median of 19 cells per spot (MAD=5) and 3 cells per subspot (MAD=1).



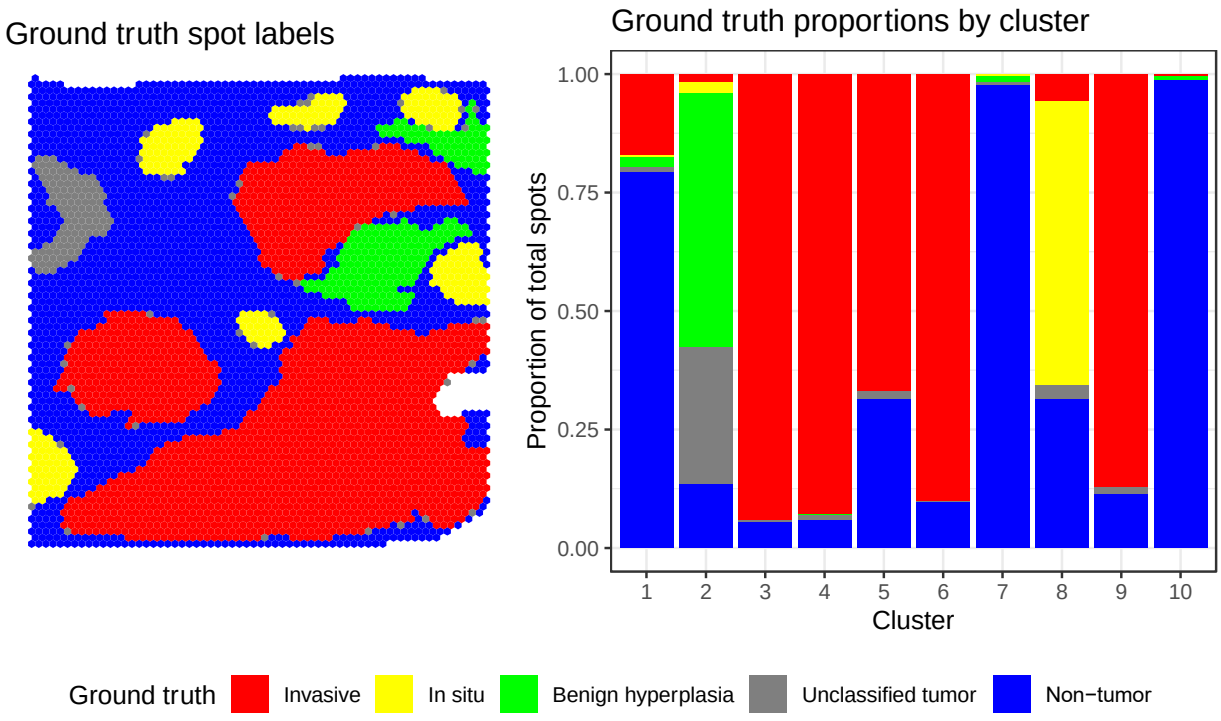
Supplementary Figure 9. Spot-level clustering of invasive ductal carcinoma (IDC) and anti-CD3 stain intensity. A) BIC plot. The elbow at q=10 was selected as the number of clusters to analyze. B) Spot-level BayesSpace clustering of IDC, masked to exclude out-of-focus image regions. C) Anti-CD3 stain intensity within each cluster (n = 2,929 spots). Clusters 1, 7, and 10 were deemed CD3 "high" clusters when comparing the clustering to underlying stain signal. In the boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively. Data beyond the end of the whiskers represent outlying points.



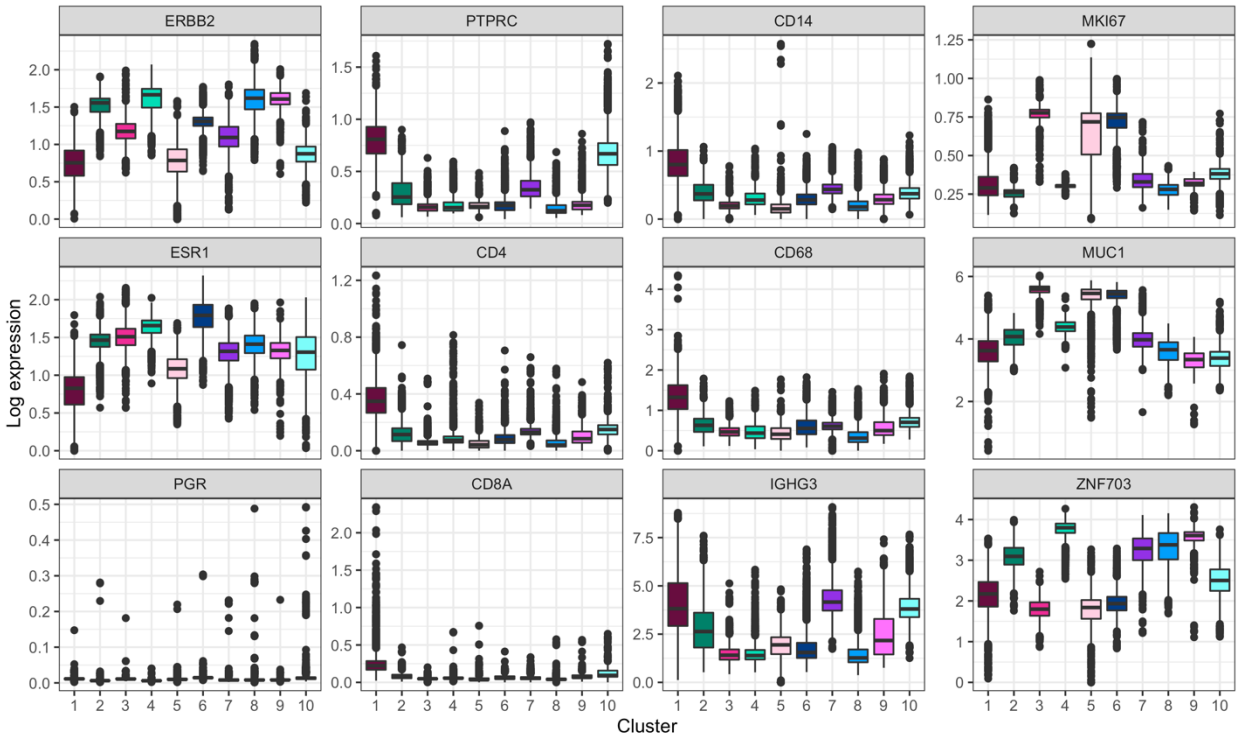
Supplementary Figure 10. Spot-level clustering of the OC sample and anti-CD45 stain intensity. A) BIC plot. The elbow at q=8 was selected as the number of clusters to analyze. B) Spot-level BayesSpace clustering of the OC sample, masked to exclude out-of-focus image regions. C) Anti-CD45 stain intensity within each cluster (n = 2,041 spots). Clusters 3, 4, and 6 were deemed CD45 "high" clusters when comparing the clustering to underlying stain signal. While cluster 6 has only slightly higher anti-CD45 intensity than clusters 7 and 8, its anti-Cytokeratin intensity was substantially lower, justifying its inclusion as a CD45 "high" cluster. In each boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively. Data beyond the end of the whiskers represent outlying points.



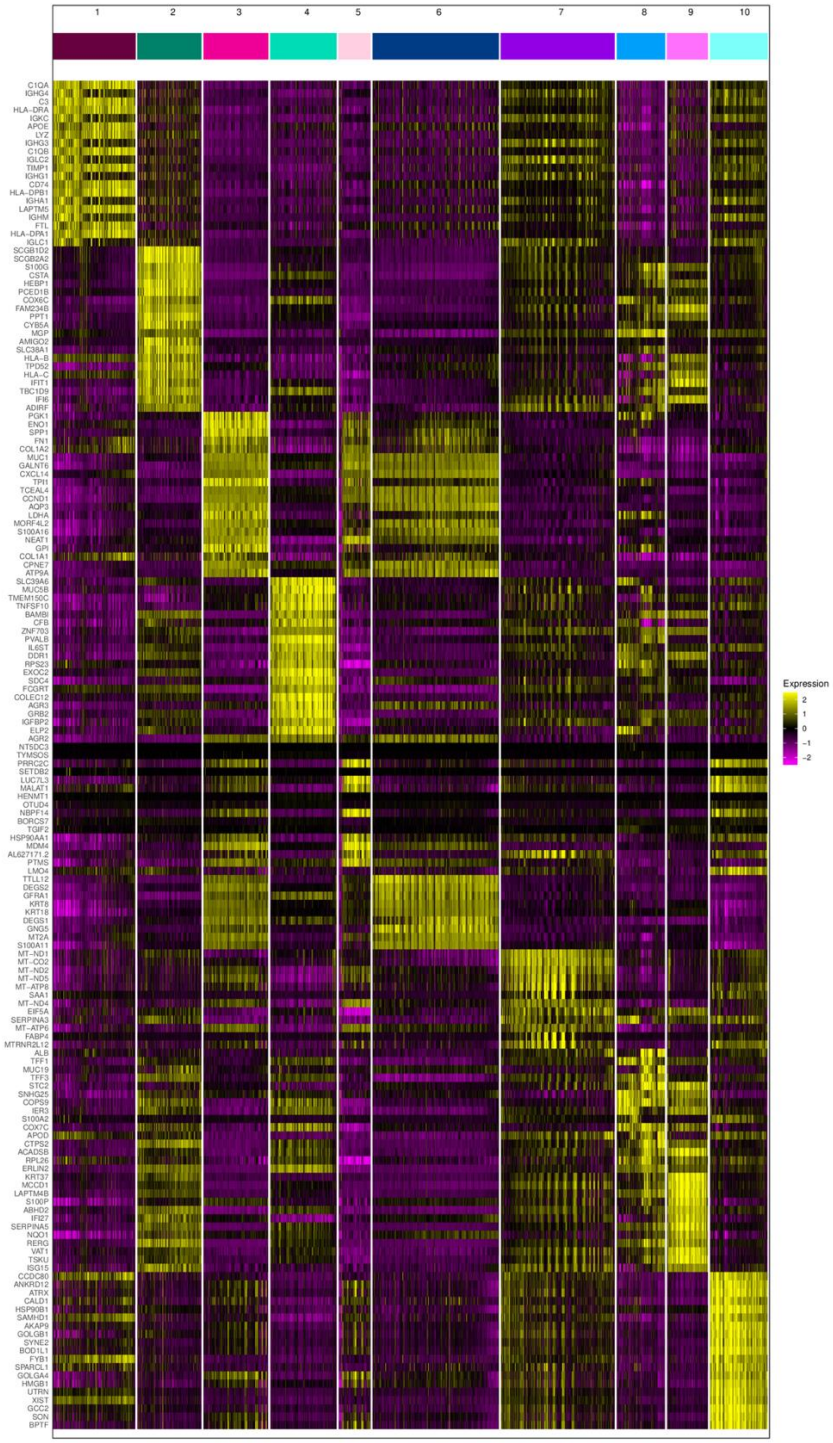
Supplementary Figure 11. Ground truth for IDC1 spots based on histological annotations. A) Assigned ground truth label for each spot, determined by whether they fell within a corresponding region on the annotated image. B) Proportion of ground truth labels within each cluster. Clusters generally corresponded to a single dominant ground truth label.



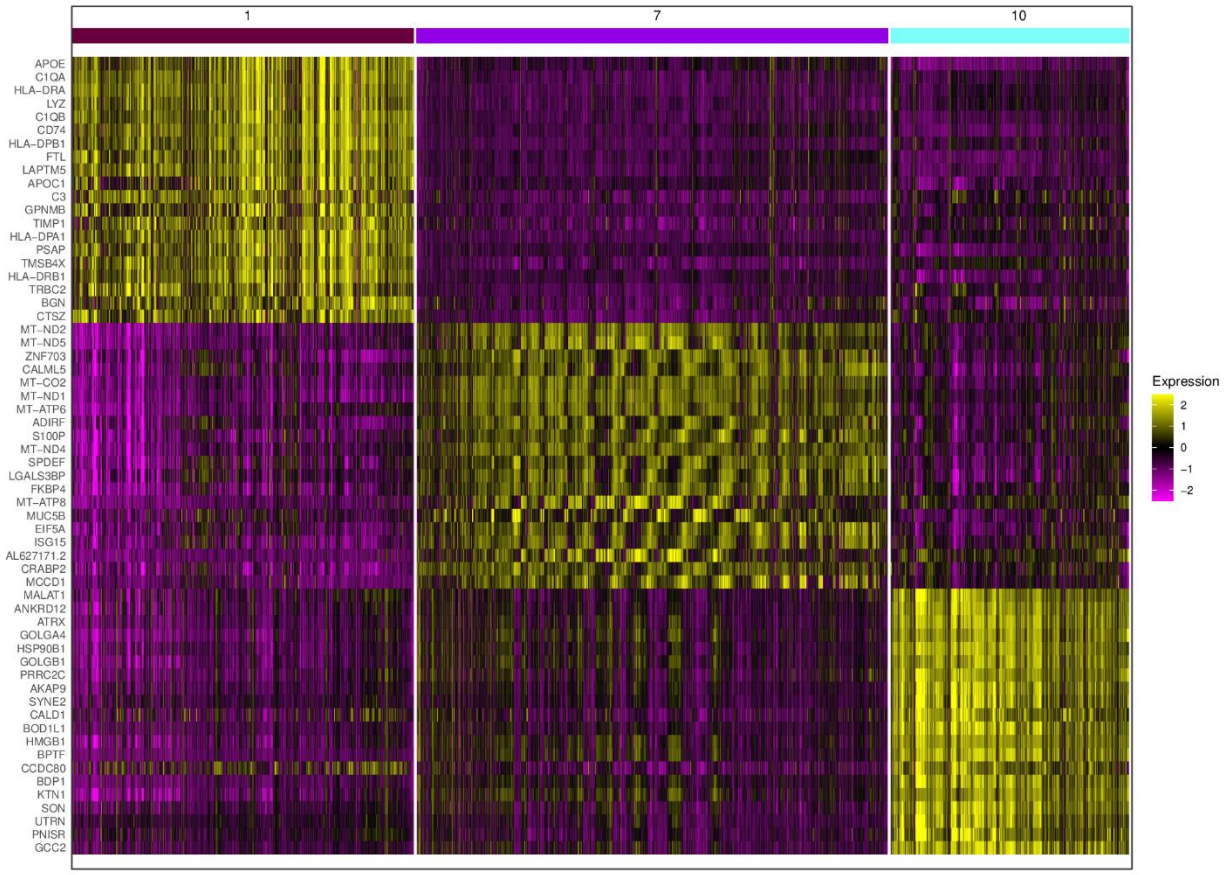
Supplementary Figure 12. Expression of highlighted marker genes in IDC clusters. Each panel corresponds to a spatial expression plot in Figure 6 and shows a boxplot of the respective marker gene's enhanced expression across the ten BayesSpace clusters. In each boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively. Data beyond the end of the whiskers represent outlying points.



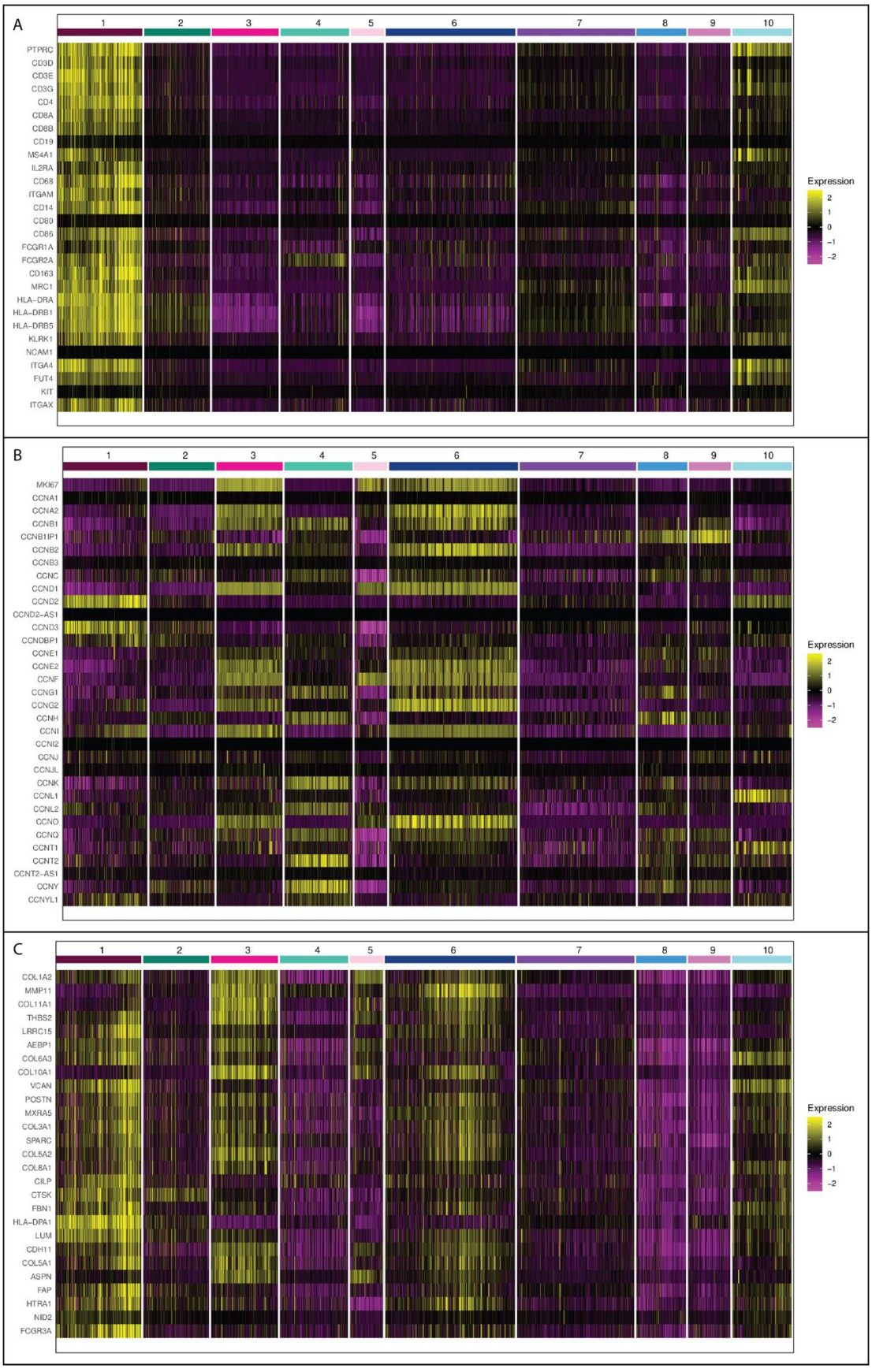
Supplementary Figure 13. Differentially expressed genes in ten BayesSpace clusters in ductal carcinoma. We show the top 20 differentially expressed genes (with respect to average log-fold-change) in each BayesSpace cluster.



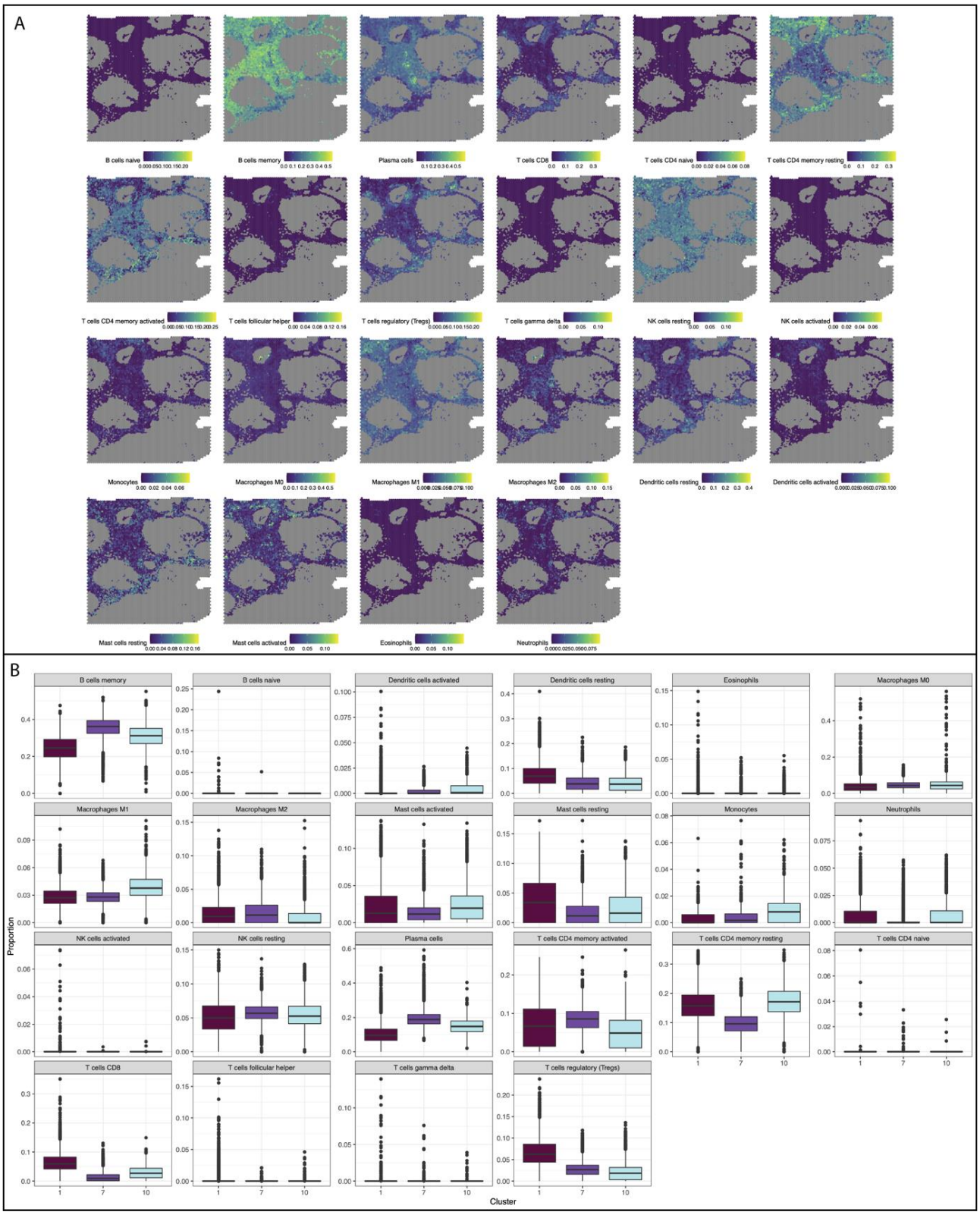
Supplementary Figure 14. Differentially expressed genes and heterogeneity among non-tumor clusters (IDC). In a comparison among the three non-tumor clusters, we highlight the top 20 differentially expressed genes in each cluster.



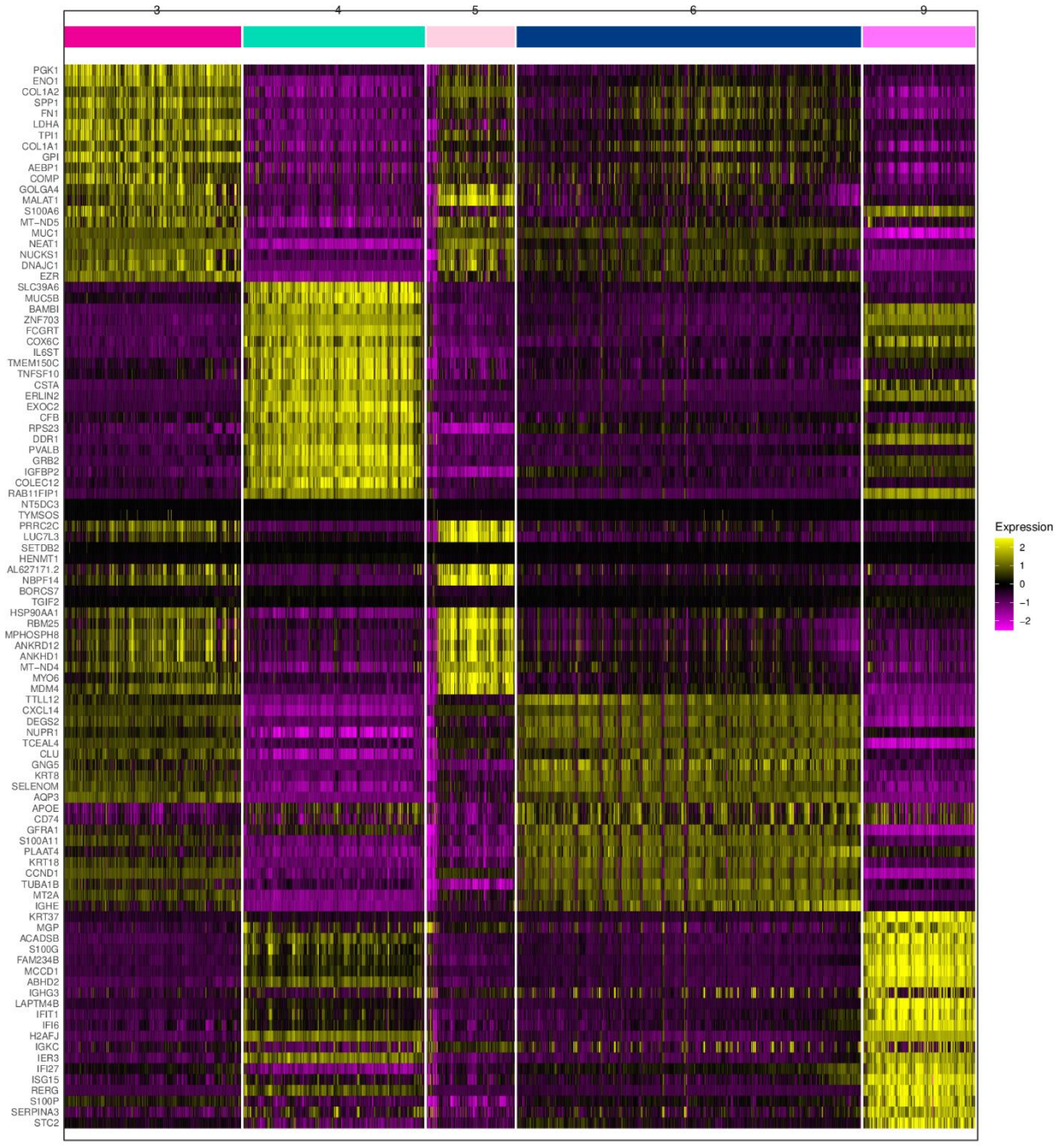
Supplementary Figure 15. Expression of relevant marker genes in ductal carcinoma clusters. A) Marker genes for common immune cell types. B) Markers of cell proliferation - Ki-67 and cyclins. C) Markers of ductal carcinoma invasion. Genes were included if they were reported to be associated with the progression to invasion by both Hu *et al.* (2008). and Knudsen *et al.* (2012).


A
1
2
3
4
5
6
7
8
9
10
PTPRC
CD3D
CD3E
CD3G
CD4
CD8A
CD8B
CD19
MS4A1
IL2RA
CD68
ITGAM
CD14
CD80
CD86
FCGR1A
FCGR2A
CD163
MRC1
HLA-DRA
HLA-DRB1
HLA-DRB5
KLRK1
NCAM1
ITGA4
FUT4
KIT
ITGAX
Expression
2
1
0
−1
−2
B
1
2
3
4
5
6
7
8
9
10
MKI67
CCNA1
CCNA2
CCNB1
CCNB1IP1
CCNB2
CCNB3
CCNC
CCND1
CCND2
CCND2-AS1
CCND3
CCNDBP1
CCNE1
CCNE2
CCNF
CCNG1
CCNG2
CCNH
CCNI
CCNI2
CCNJ
CCNJL
CCNK
CCNL1
CCNL2
CCNO
CCNQ
CCNT1
CCNT2
CCNT2-AS1
CCNY
CCNYL1
Expression
2
1
0
−1
−2
C
1
2
3
4
5
6
7
8
9
10
COL1A2
MMP11
COL11A1
THBS2
LRRC15
AEBP1
COL6A3
COL10A1
VCAN
POSTN
MXRA5
COL3A1
SPARC
COL5A2
COL8A1
CILP
CTSK
FBN1
HLA-DPA1
LUM
CDH11
COL5A1
ASPN
FAP
HTRA1
NID2
FCGR3A
Expression
2
1
0
−1
−2

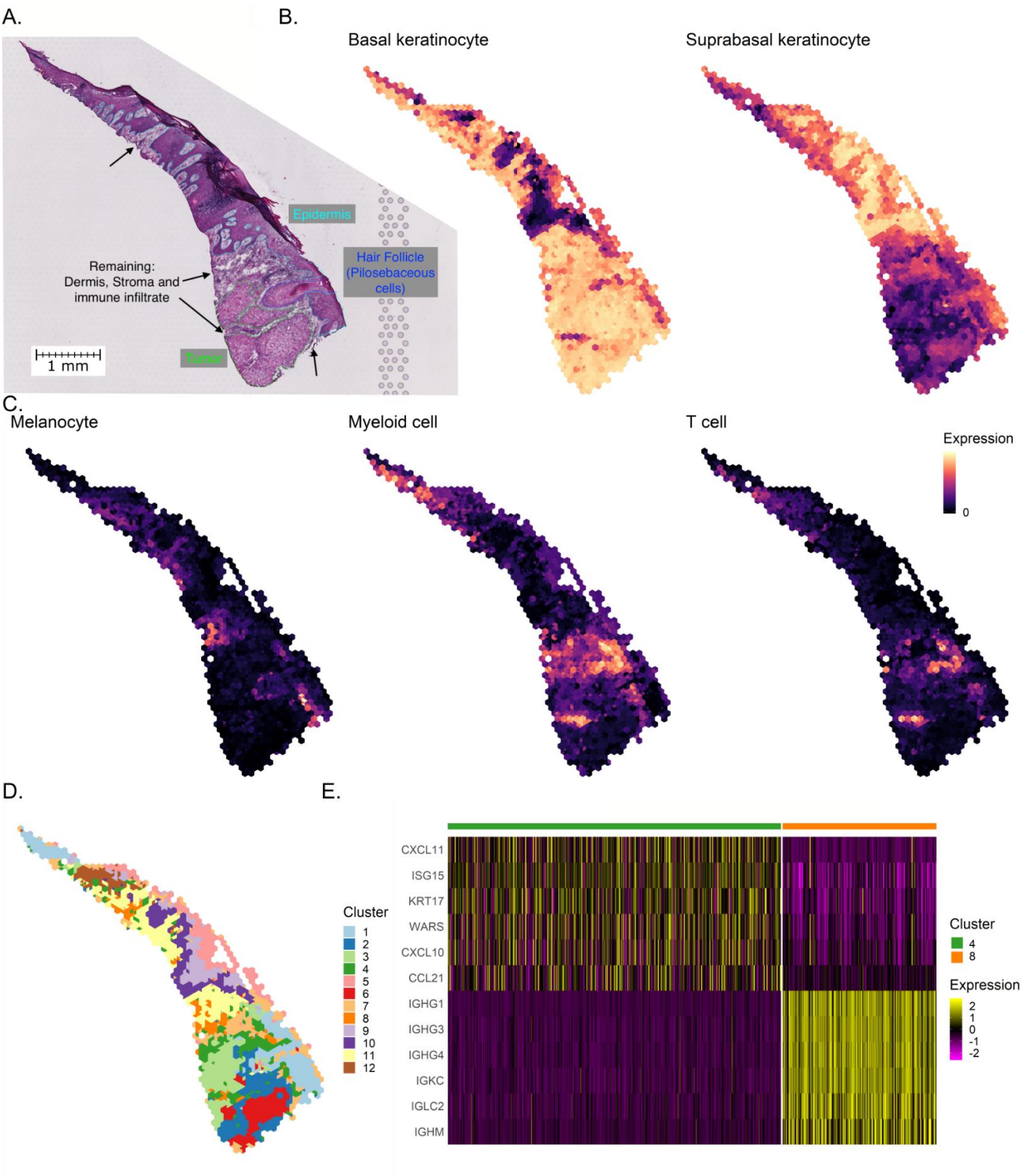
Supplementary Figure 16. CIBERSORT cell type proportions in invasive ductal carcinoma. A) Proportions of each CIBERSORT cell type in each subspot plotted spatially. B) Boxplots comparing the distribution of proportions of each cell type across subspots in each cluster (n = 10,335 subspots). In each boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively. Data beyond the end of the whiskers represent outlying points.


A
B cells naive
B cells memory
Plasma cells
T cells CD8
T cells CD4 naive
T cells CD4 memory resting
T cells CD4 memory activated
T cells follicular helper
T cells regulatory (Tregs)
T cells gamma delta
NK cells resting
NK cells activated
Monocytes
Macrophages M0
Macrophages M1
Macrophages M2
Dendritic cells resting
Dendritic cells activated
Mast cells resting
Mast cells activated
Eosinophils
Neutrophils
B
B cells memory
B cells naive
Dendritic cells activated
Dendritic cells resting
Eosinophils
Macrophages M0
Macrophages M1
Macrophages M2
Mast cells activated
Mast cells resting
Monocytes
Neutrophils
NK cells activated
NK cells resting
Plasma cells
T cells CD4 memory activated
T cells CD4 memory resting
T cells CD4 naive
T cells CD8
T cells follicular helper
T cells gamma delta
T cells regulatory (Tregs)
Proportion
Cluster

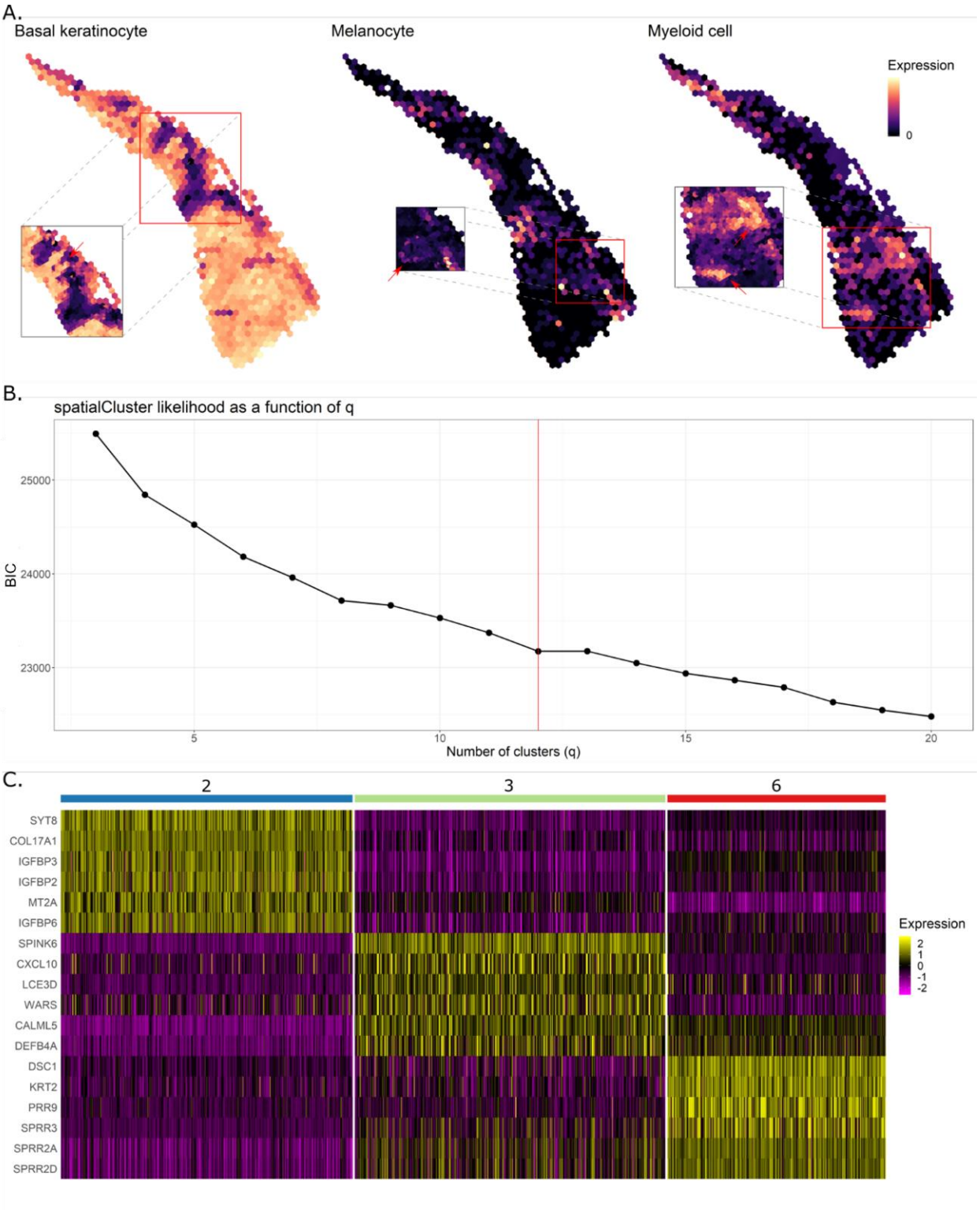
Supplementary Figure 17. Differentially expressed genes and heterogeneity among invasive tumor clusters (IDC). In a comparison among the five invasive tumor clusters, we highlight the top 20 differentially expressed genes in each cluster.



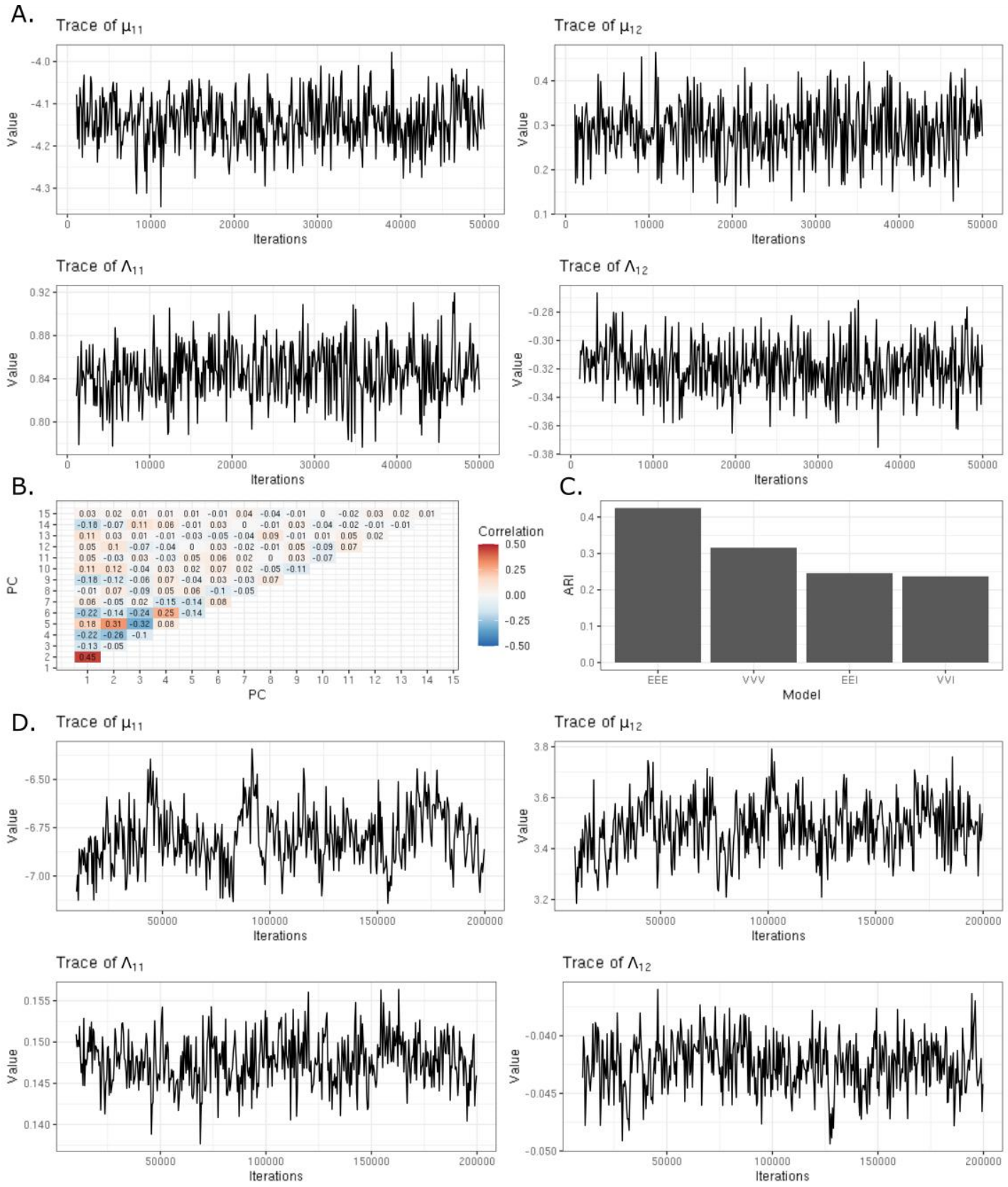
Supplementary Figure 18. BayesSpace identifies transcriptionally distinct immune clusters. (A) Manual histopathological annotations of the stained H&E tissue (N = 1 tissue section, n = 722 spots) differentiate the tumor (green), the epidermis (aqua), and a hair follicle (blue) from the remaining regions, which contain the dermis, stroma, and immune infiltrate. (B) Keratinocyte expression is high throughout the slide, but we see spatially distinct expression of basal (*KRT5*, *KRT14*) and suprabasal (*KRT1*, *KRT10*) keratinocytes at enhanced resolution. (C) Enhanced resolution maps of melanocyte (*MLANA*, *DCT*, *PMEL*), myeloid cell (*LYZ*), and T cell (*CD2*, *CD3D*, *CD3E*, *CD3G*, *CD7*) marker expression are biologically supported by the histopathological annotations. (D) The spots can be partitioned into twelve clusters, most of which display clear spatial patterns. (E) Clusters 4 and 8 have enriched immune cell expression but display substantial differences in expression of immunoglobulin genes and genes regulated by interferons.


A.
B.
Basal keratinocyte
Suprabasal keratinocyte
Epidermis
Hair Follicle (Pilosebaceous cells)
Remaining: Dermis, Stroma and immune infiltrate
Tumor
1 mm
C.
Melanocyte
Myeloid cell
T cell
Expression
0
D.
Cluster
1
2
3
4
5
6
7
8
9
10
11
12
E.
CXCL11
ISG15
KRT17
WARS
CXCL10
CCL21
IGHG1
IGHG3
IGHG4
IGKC
IGLC2
IGHM
Cluster
4
8
Expression
2
1
0
-1
-2

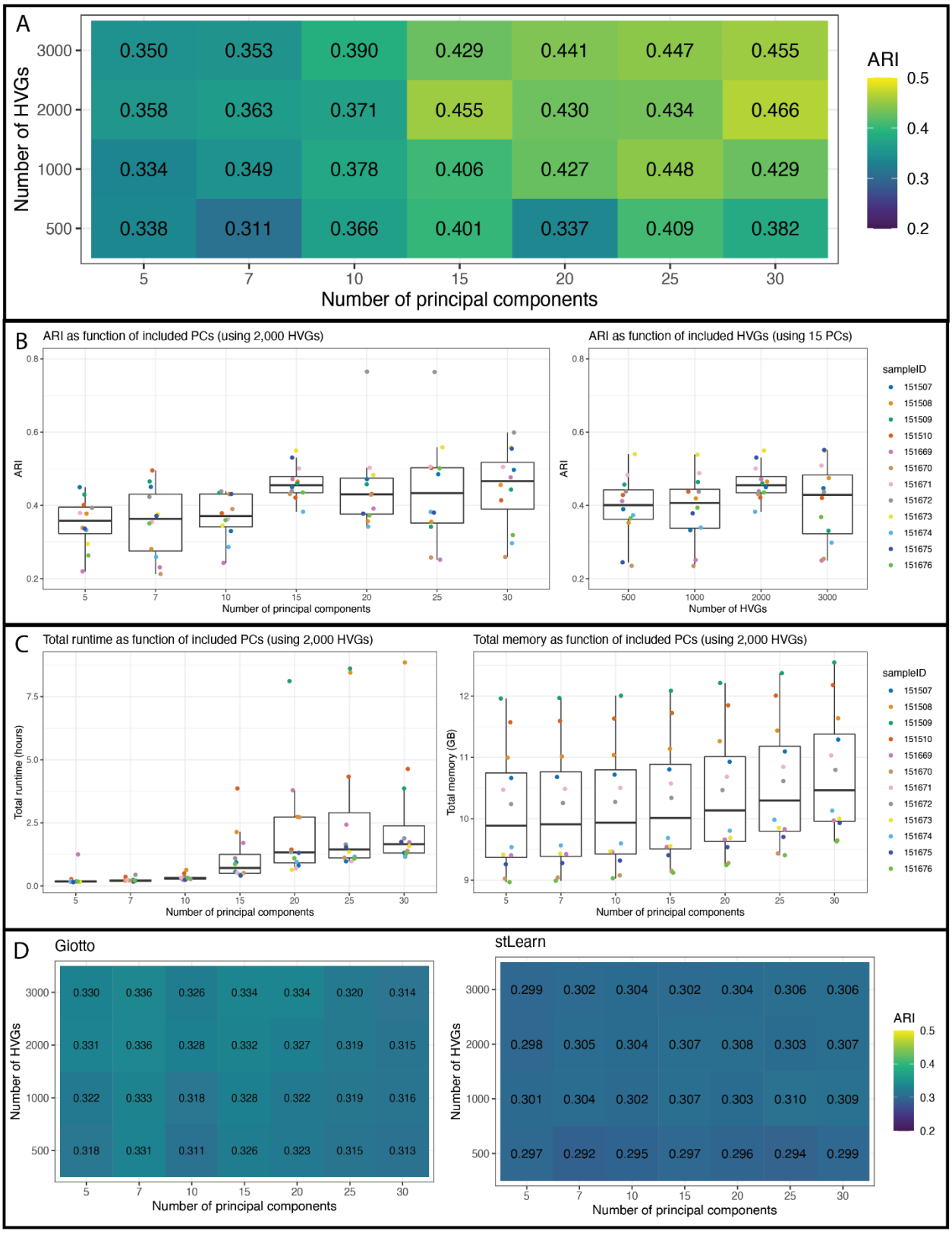
Supplementary Figure 19. (A) The spot-level marker expression of basal keratinocytes, melanocytes, and myeloid cells are shown. The areas in red boxes are also shown at enhanced resolution in the inset within the black box. In each case, enhanced resolution spatially refines the marker expression. (B) `spatialCluster()` is run for 1,000 iterations with number of clusters set between 3 and 20. The BIC over iterations is plotted for each cluster number. $q = 12$, denoted by the vertical red line, is a reasonable choice for the elbow. (C) Differential expression analysis between the three tumor clusters highlights heterogeneity within the tumor.


A.
Basal keratinocyte
Melanocyte
Myeloid cell
Expression
0
B.
spatialCluster likelihood as a function of q
BIC
25000
24000
23000
5
10
15
20
Number of clusters (q)
C.
2
3
6
SYT8
COL17A1
IGFBP3
IGFBP2
MT2A
IGFBP6
SPINK6
CXCL10
LCE3D
WARS
CALML5
DEFB4A
DSC1
KRT2
PRR9
SPRR3
SPRR2A
SPRR2D
Expression
2
1
0
-1
-2

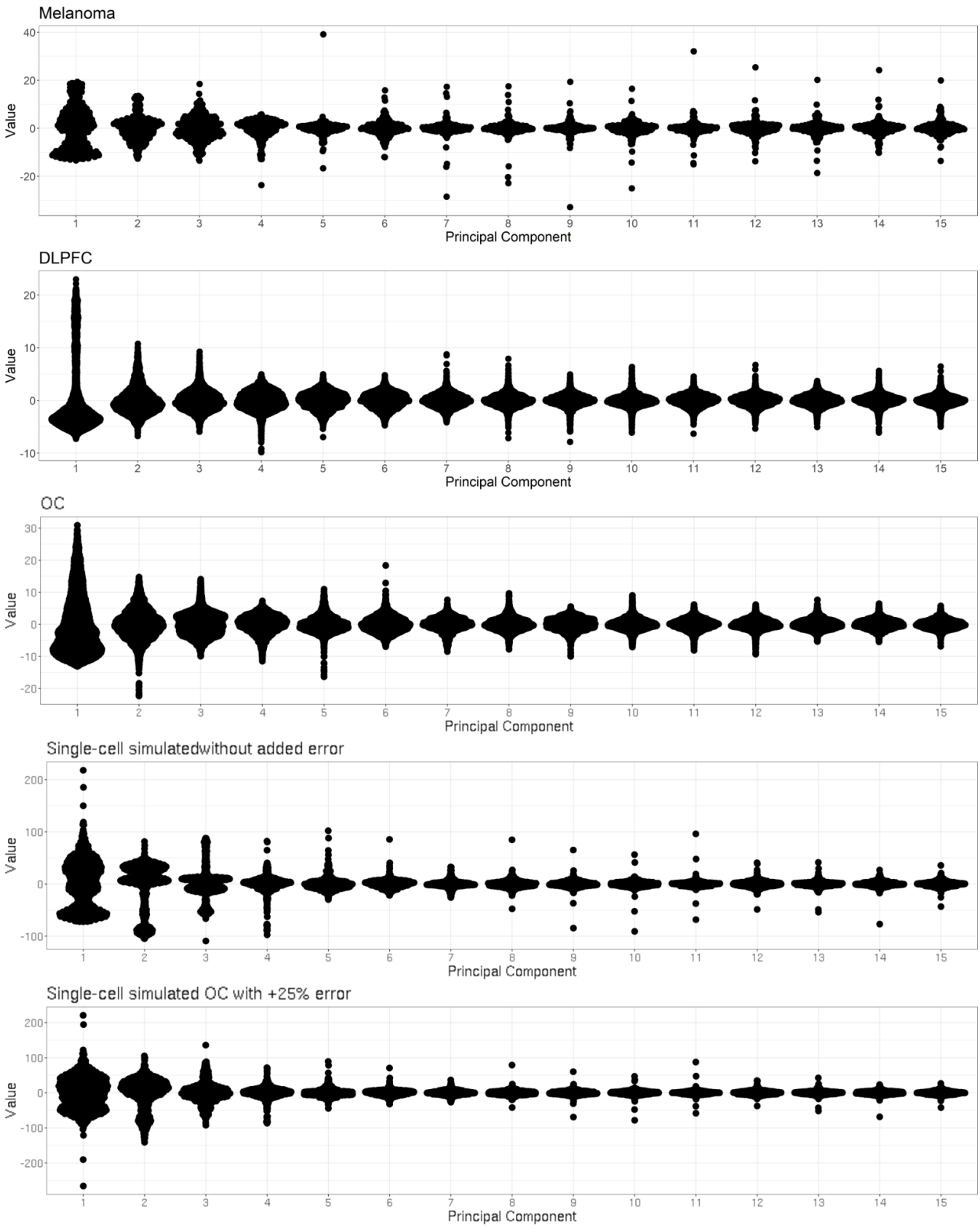
Supplementary Figure 20. Further assessment of BayesSpace modeling assumptions. A) Examples trace plots of mean (top) and precision (bottom) parameters for DLPFC sample 151673 are shown. Overall, the plots show that the estimates are stable and have good mixing. The MCMC chains converge as estimated by the Gelman-Rubin convergence diagnostic with $\hat{R} < 1.1$. B) The estimated correlation matrix (which is derived from the precision matrix) of the sample 151673 shows that there is correlation between some PCs after conditioning on cluster (top). C) The EEI and VVI mclust models, which assume an independent correlation structure, also perform worse in sample 151673 than the EEE and VVV mclust models, which respectively allow for fully parameterized fixed and variable covariances across clusters (bottom). EEE is the best performing model and the default setting for BayesSpace. D) Example trace plots of mean (top) and precision (bottom) parameters for enhancement of the OC sample to subspot level are shown. The trace for μ_{11} shows patterns of some dependence between iterations though this is not a problem given sufficient iterations. Here, the MCMC chains converge as estimated by the Gelman-Rubin convergence diagnostic with $\hat{R} < 1.1$.


A.
Trace of μ_{11}
Trace of μ_{12}
Trace of Λ_{11}
Trace of Λ_{12}
Value
Iterations
B.
PC
Correlation
C.
ARI
Model
EEE
VVV
EEI
VVI
D.
Trace of μ_{11}
Trace of μ_{12}
Trace of Λ_{11}
Trace of Λ_{12}
Value
Iterations

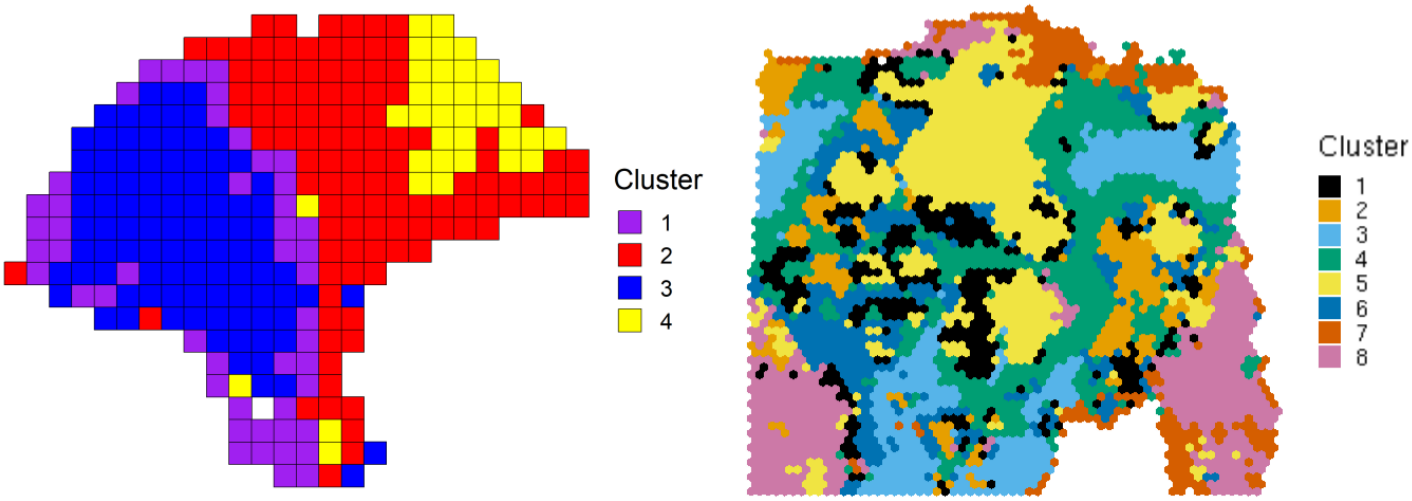
Supplementary Figure 21. Justifying the choice of 15 principal components (PCs) and 2,000 highly variable genes (HVGs). BayesSpace is robust to the number of included HVGs and demonstrates little improvement in clustering accuracy beyond 15 PCs (A, B). Additionally, runtime increases with the number of PCs (C) and other spatial clustering algorithms appear insensitive to these parameters (D). A) The heatmap shows the ARI between the BayesSpace clustering and the ground truth in the 12 dorsolateral prefrontal cortex (DLPFC) samples with different numbers of modeled PCs and included HVGs. Values are the median ARI over all 12 samples. B) The boxplots show the ARI between the BayesSpace clustering and the ground truth in the 12 DLPFC samples as a function of either the number of included PCs or HVGs when holding the other fixed at our chosen parameter (2,000 HVGs and 15 PCs, respectively). C) The boxplots show the total runtime (left) and memory consumption (right) of BayesSpace as the number of included PCs increases. For panels B and C, in each boxplot, the center line, box limits, and whiskers denote the median, upper/lower quartiles, and 1.5x interquartile range respectively. D) We compare the sensitivity of the two other spatial clustering algorithms, Giotto (left) and stLearn (right), to the number of modeled PCs and included HVGs. As in panel (A), each value is the median ARI between the respective clustering and the ground truth in the 12 DLPFC samples.


A
Number of HVGs
3000
2000
1000
500
0.350 0.353 0.390 0.429 0.441 0.447 0.455
0.358 0.363 0.371 0.455 0.430 0.434 0.466
0.334 0.349 0.378 0.406 0.427 0.448 0.429
0.338 0.311 0.366 0.401 0.337 0.409 0.382
5 7 10 15 20 25 30
Number of principal components
ARI
0.5
0.4
0.3
0.2
B
ARI as function of included PCs (using 2,000 HVGs)
ARI
Number of principal components
ARI as function of included HVGs (using 15 PCs)
Number of HVGs
sampleID
151507
151508
151509
151510
151669
151670
151671
151672
151673
151674
151675
151676
C
Total runtime as function of included PCs (using 2,000 HVGs)
Total runtime (hours)
Number of principal components
Total memory as function of included PCs (using 2,000 HVGs)
Total memory (GB)
Number of principal components
D
Giotto
Number of HVGs
3000 0.330 0.336 0.326 0.334 0.334 0.320 0.314
2000 0.331 0.336 0.328 0.332 0.327 0.319 0.315
1000 0.322 0.333 0.318 0.328 0.322 0.319 0.316
500 0.318 0.331 0.311 0.326 0.323 0.315 0.313
Number of principal components
stLearn
Number of HVGs
3000 0.299 0.302 0.304 0.302 0.304 0.306 0.306
2000 0.298 0.305 0.304 0.307 0.308 0.303 0.307
1000 0.301 0.304 0.302 0.307 0.303 0.310 0.309
500 0.297 0.292 0.295 0.297 0.296 0.294 0.299
Number of principal components
ARI

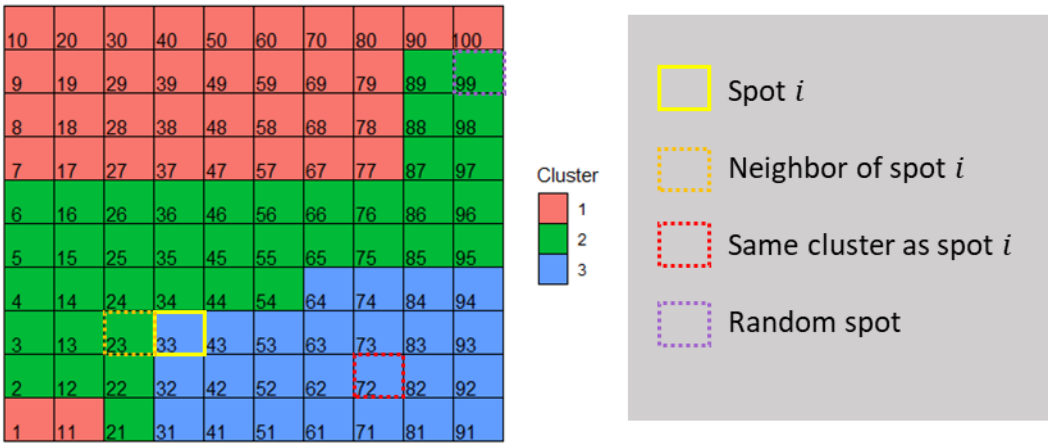
Supplementary Figure 22. The higher PCs of the melanoma sample have large outliers. Only PCs 1 through 7 are used in modeling the melanoma sample to capture as much biological variation as possible while limiting the impact of outliers. For reference, the distribution of PCs from DLPFC sample 151673 and the OC sample are shown below. These Visium samples do not have large outliers in the higher PCs. The PCs of the single-cell simulation based on the OC sample show distinct modes in their distributions that are not commonly seen in real data. Addition of random noise equal to 25% of the variance of each PC provides a more realistic PC distribution as well as a more challenging clustering problem.


Melanoma
Value
Principal Component
DLPFC
Value
Principal Component
OC
Value
Principal Component
Single-cell simulatedwithout added error
Value
Principal Component
Single-cell simulated OC with +25% error
Value
Principal Component

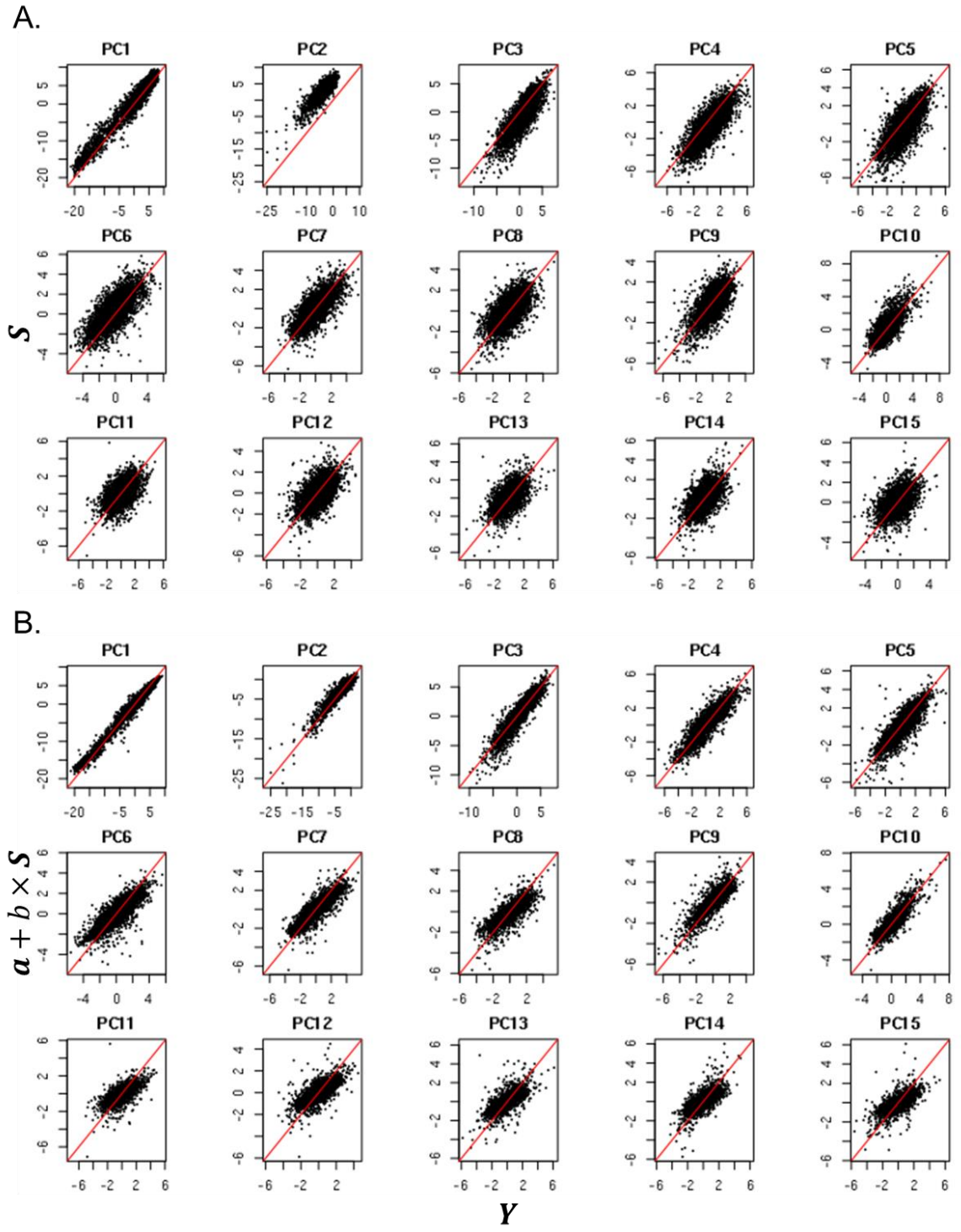
Supplementary Figure 23. The optimal spot-level clustering partitions for the melanoma (left) and OC (right) enhanced clustering simulations (Simulation 2) are shown.



Supplementary Figure 24. Moves considered for updating locations of single cells. For example, a cell mapped to spot i (spot 33) in the current iteration is proposed to swap with a cell from a neighboring spot (e.g. spot 23), spot with the same cluster label (e.g. spot 72), or a random spot (e.g. spot 99), in a series of three proposed moves for each cell.



Supplementary Figure 25. Untransformed (A) vs. linearly transformed (B) ***S*** vs. ***Y***. The transformation consists of a dimension-specific intercept and a common slope across all dimensions. The red line represents the identity line.



Supplementary Table

Supplementary Table 1. Single cell (SC) sampling scheme for spatial ground truth clusters. SC clusters were generated and annotated for cell type by Izar, et. al (2020). Single cells are sampled into corresponding spatial clusters without replacement.

Spatial Cluster	Spatial Cluster Name	Subspots in Spatial Cluster	Corresponding SC cluster	SC cell type	Cells in SC cluster
1	Intratumor 1	153	15	Dendritic cell	169
2	Intratumor 2	322	8	Fibroblast	360
3	Stroma	1893	11	Macrophage	2535
4	Tumor 1	1876	1	Malignant	3191
5	Tumor 2	532	2	Malignant	1613

References

1. Shendure, J. The beginning of the end for microarrays? *Nat. Methods 2008 57* **5,** 585–587 (2008).
2. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015,** pdb.top084970 (2015).
3. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).
4. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* **78,** 5970–5979 (2018).
5. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9,** 1–13 (2018).
6. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* 1–12 (2021). doi:10.1038/s41593-020-00787-0
7. Janosevic, D. *et al.* The orchestrated cellular and molecular responses of the kidney to endotoxin define a precise sepsis timeline. *Elife* **10,** (2021).
8. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods 2021 187* **18,** 723–732 (2021).
9. Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **28,** 129–137 (1982).
10. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008,** P10008 (2008).

11. Fraley, C., Raftery, A. E. & Murphy, T. B. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. (2012).

12. Kiselev, V. Y. *et al.* SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14,** 483–486 (2017).

13. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G. C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **36,** 1183–1190 (2018).

14. Dries, R. *et al.* Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *bioRxiv* 701680 (2019). doi:10.1101/701680

15. Pham, D. T. *et al.* stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv* 2020.05.31.125658 (2020). doi:10.1101/2020.05.31.125658

16. Besag, J. On the Statistical Analysis of Dirty Pictures. *J. R. Stat. Soc. Ser. B* **48,** 259–279 (1986).

17. Gottardo, R., Besag, J., Stephens, M. & Murua, A. Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics* **7,** 85–99 (2006).

18. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17,** 137–145 (2020).

19. Ising, E. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Phys. 1925 311* **31,** 253–258 (1925).

20. Potts, R. B. Some generalized order-disorder transformations. *Math. Proc. Cambridge Philos. Soc.* **48,** 106–109 (1952).

21. Schwarz, G. Estimating the Dimension of a Model.

https://doi.org/10.1214/aos/1176344136 **6,** 461–464 (1978).

22. Besag, J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. R. Stat. Soc. Ser. B* **36,** 192–225 (1974).

23. Swendsen, R. H. & Wang, J. S. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58,** 86 (1987).

24. Smith, D. & Smith, M. Estimation of Binary Markov Random Fields Using Markov chain Monte Carlo. *http://dx.doi.org/10.1198/106186006X97817* **15,** 207–227 (2012).

25. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17,** 75 (2016).

26. Mccarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. doi:10.1093/bioinformatics/btw777

27. Liu, W. Unsupervised learning approaches for the finite mixture models: EM versus MCMC. in *Proceedings - 2010 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2010* 498–501 (2010). doi:10.1109/CyberC.2010.96

28. Ripley, B. D. The use of spatial models as image priors. in 309–340 (Institute of Mathematical Statistics, 1991). doi:10.1214/lnms/1215460510

29. Geyer, C. J. & Thompson, E. A. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *J. R. Stat. Soc. Ser. B* **54,** 657–683 (1992).

30. Liang, F. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *http://dx.doi.org/10.1080/00949650902882162* **80,** 1007–1022 (2009).

31. Miller, M. I., Fan, J. & Tward, D. J. Multi scale diffeomorphic metric mapping of spatial transcriptomics datasets. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.* 4467–4475 (2021). doi:10.1109/CVPRW53098.2021.00504

32. Saiselet, M. *et al.* Transcriptional output, cell types densities and normalization in spatial transcriptomics. *J. Mol. Cell Biol.* (2020). doi:10.1093/jmcb/mjaa028

33. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* **11,** 360–361 (2014).

34. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348,** aaa6090-aaa6090 (2015).

35. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat. Biotechnol. 2022* 1–13 (2022). doi:10.1038/s41587-022-01483-z

36. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (80-.).* **363,** 1463–1467 (2019).

37. Hu, K. H. *et al.* ZipSeq: barcoding for real-time mapping of single cell transcriptomes. *Nat. Methods* **17,** 833–843 (2020).

38. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3,** 1–8 (2020).

39. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* 1–10 (2021). doi:10.1038/s41587-021-00830-w

40. M, E.-B., P, N., E, M., I, G. & H, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* (2021).

doi:10.1093/nar/gkab043

41. Gavin, J. & Jennison, C. A subpixel image restoration algorithm. *J. Comput. Graph. Stat.* **6,** 182–201 (1997).

42. Tipping, M. E. & Bishop, C. M. Bayesian image super-resolution. in *Proceedings of the 15th International Conference on Neural Information Processing Systems* (eds. Becker, S., Thrun, S. & Obermayer, K.) 1303–1310 (2002).

43. Aykroyd, R. G. & Green, P. J. Global and local priors, and the location of lesions using gamma-camera imagery. *Philos. Trans. R. Soc. London. Ser. A Phys. Eng. Sci.* **337,** 323–342 (1991).

44. Gelman, A., Roberts, G. O. & Gilks, W. R. Efficient Metropolis jumping rules. *Bayesian Stat.* **5,** 599–607 (1996).

45. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13–17–Augu,** 785–794 (Association for Computing Machinery, 2016).

46. Chen, T. *et al.* xgboost: Extreme Gradient Boosting. (2020).

47. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage--an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26,** 979–981 (2010).

48. Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *J. Clean. Prod.* **182,** 497–514.e22 (2020).

49. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).

50. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data Resource Comprehensive Integration of Single-Cell Data. *Cell* **177,** (2019).

51. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16,** 278 (2015).

52. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).

53. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* **26,** 1271–1279 (2020).

54. Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37,** 233–243 (1991).

55. Wang, H. X. *et al.* HSPB1 deficiency sensitizes melanoma cells to hyperthermia induced cell death. *Oncotarget* **7,** 67449–67462 (2016).

56. Mathieu, V. *et al.* The sodium pump α1 sub-unit: A disease progression-related target for metastatic melanoma treatment. *J. Cell. Mol. Med.* **13,** 3960–3972 (2009).

57. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352,** 189–196 (2016).

58. Mori, K. *et al.* CpG hypermethylation of collagen type I α 2 contributes to proliferation and migration activity of human bladder cancer. *Int. J. Oncol.* **34,** 1593–1602 (2009).

59. Knudsen, E. S. *et al.* Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia. *Breast Cancer Res. Treat.* **133,** 1009–1024 (2012).

60. Lee, S. *et al.* Differentially expressed genes regulating the progression of ductal

carcinoma in situ to invasive breast cancer. *Cancer Res.* **72,** 4574–4586 (2012).

61. Hu, M. *et al.* Regulation of In Situ to Invasive Breast Carcinoma Transition. *Cancer Cell* **13,** 394–406 (2008).

62. Hattrup, C. L. & Gendler, S. J. MUC1 alters oncogenic events and transcription in human breast cancer cells. *Breast Cancer Res.* **8,** (2006).

63. Besmer, D. M. *et al.* Pancreatic ductal adenocarcinoma mice lacking mucin 1 have a profound defect in tumor growth and metastasis. *Cancer Res.* **71,** 4432–4442 (2011).

64. Behrens, M. E. *et al.* The reactive tumor microenvironment: MUC1 signaling directly reprograms transcription of CTGF. *Oncogene* **29,** 5667–5677 (2010).

65. Zhang, X. *et al.* Insights into the distinct roles of MMP-11 in tumor biology and future therapeutics (Review). *Int. J. Oncol.* **48,** 1783–1793 (2016).

66. Holland, D. G. *et al.* ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. *EMBO Mol. Med.* **3,** 167–180 (2011).

67. Sircoulomb, F. *et al.* ZNF703 gene amplification at 8p12 specifies luminal B breast cancer. *EMBO Mol. Med.* **3,** 153–166 (2011).

68. Daly, R., Binder, M. & Sutherland, R. Overexpression of the Grb2 gene in human breast cancer cell lines. *Oncogene* **9,** 2723–2727 (1994).

69. Tari, A. M., Hung, M. C., Li, K. & Lopez-Berestein, G. Growth inhibition of breast cancer cells by Grb2 downregulation is correlated with inactivation of mitogen-activated protein kinase in EGFR, but not in ErbB2, cells. *Oncogene* **18,** 1325–1332 (1999).

70. Onichtchouk, D. *et al.* Silencing of TGF-β signalling by the pseudoreceptor BAMBI. *Nature* **401,** 480–485 (1999).

71. Sümer, C., Boz Er, A. B. & Dinçer, T. Keratin 14 is a novel interaction partner of keratinocyte differentiation regulatoreceptor-interacting protein kinase 4. *Turkish J. Biol.* **43,** 225–234 (2019).

72. Weroha, S. J. & Haluska, P. The Insulin-Like Growth Factor System in Cancer. *Endocrinology and Metabolism Clinics of North America* **41,** 335–350 (2012).

73. Grimberg, A. p53 and IGFBP-3: Apoptosis and cancer protection. *Mol. Genet. Metab.* **70,** 85–98 (2000).

74. Rentoft, M. *et al.* Expression of CXCL10 is associated with response to radiotherapy and overall survival in squamous cell carcinoma of the tongue. *Tumor Biol.* **35,** 4191–4198 (2014).

75. Hudson, L. G. *et al.* Microarray analysis of cutaneous squamous cell carcinomas reveals enhanced expression of epidermal differentiation complex genes. *Mol. Carcinog.* **49,** 619–629 (2010).

76. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22,** 627–644 (2021).

77. Elosua, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes. *bioRxiv* 2020.06.03.131334 (2020). doi:10.1101/2020.06.03.131334

78. Biancalani, T. *et al.* Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram. *bioRxiv* 2020.08.29.272831 (2020). doi:10.1101/2020.08.29.272831

79. Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat.*

Biotechnol. 1–10 (2021).

80. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods 2019 1612* **16,** 1289–1296 (2019).

81. Massoni-Badosa, R. *et al.* An Atlas of Cells in the Human Tonsil. *bioRxiv* **13,** 2022.06.24.497299 (2022).

82. Attaf, N., Baaklini, S., Binet, L. & Milpied, P. Heterogeneity of germinal center B cells: New insights from single-cell studies. *Eur. J. Immunol.* **51,** 2555–2567 (2021).

83. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *bioRxiv* 2020.02.28.969931 (2020). doi:10.1101/2020.02.28.969931

84. Petukhov, V. *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nat. Biotechnol. 2021 403* **40,** 345–354 (2021).

85. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016–December,** 770–778 (2015).

86. Janesick, A. *et al.* High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. *bioRxiv* 2022.10.06.510405 (2022). doi:10.1101/2022.10.06.510405

87. Smalley, I. & Smalley, K. S. M. Space Is the Place: Mapping the Cell–Cell Interactions That Predict Immunotherapy Responses in Melanoma. *Cancer Res.* **82,** 3198–3200 (2022).

88. Fischer, D. S., Schaar, A. C. & Theis, F. J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat. Biotechnol. 2022* 1–5 (2022).

doi:10.1038/s41587-022-01467-z

89. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis. *Cell Rep.* **29,** 202–211.e6 (2019).