Analysis of community connectivity in spatial transcriptomics data

Carter Allen[1,2,†], Kyeong Joo Jung[3,†], Yuzhou Chang[1,2], Qin Ma[1,2], and Dongjun Chung[1,2*]

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

*email: chung.911@osu.edu
†These authors contributed equally to this work.

Abstract

The advent of high throughput spatial transcriptomics (HST) has allowed for unprecedented characterization of spatially distinct cell communities within a tissue sample. While a wide range of computational tools exist for detecting cell communities in HST data, none allow for characterization of community connectivity, i.e., the relative similarity of cells within and between found communities – an analysis task that can elucidate cellular dynamics in important settings such as the tumor microenvironment. To address this gap, we introduce the concept of analysis of community connectivity (ACC), which entails not only labeling distinct cell communities within a tissue sample, but understanding the relative similarity of cells within and between communities. We develop a Bayesian multi-layer network model called BANYAN for integration of spatial and gene expression information to achieve ACC. We use BANYAN to implement ACC in invasive ductal carcinoma, and uncover distinct community structure relevant to the interaction of cell types within the tumor microenvironment. Next, we show how ACC can help clarify ambiguous annotations in a human white adipose tissue sample. Finally, we demonstrate BANYAN's ability to recover community connectivity structure via a simulation study based on real sagittal mouse brain HST data.

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

Key Words: Spatial transcriptomics; analysis of community connectivity; stochastic block model; Bayesian models; network analysis; data integration

Author Summary

The proliferation of spatial transcriptomics technologies have prompted the development of numerous statistical models for characterizing the makeup of a tissue sample in terms of distinct cell sub-populations. However, existing methods regard inferred sub-populations as static entities and do not offer any ability to discover the relative similarity of cells within and between communities, thereby obfuscating the true interactive nature of cells in a tissue sample. We develop BANYAN: a statistical model for implementing analysis of community connectivity (ACC), i.e., the process of inferring the similarity of cells within and between sub-populations. We demonstrate the utility of ACC through the analysis of a publicly available breast cancer data set, which revealed distinct community structure between tumor suppressive and invasive cancer sub-populations. We then showed how ACC may help elucidate ambiguous sub-population annotations in a publicly available human white adipose tissue data set. Finally, we implement a simulation study to validate BANYAN's ability to recover true community connectivity structure in HST data.

1 Introduction

The advent of spatial transcriptomics has allowed for the unprecedented characterization of tissue architecture in terms of spatially resolved transcript abundance [Asp et al., 2020]. In particular, *high throughput spatial transcriptomics* (HST) technologies such as the 10X Visium platform have become popular due to their transcriptome-wide sequencing depth. The proliferation of HST data has lead to the development of several computational tools for discerning cell sub-populations in HST data, while considering both gene expression and spatial information. The existing tools span a range of methodological categories, including neural networks [Chang et al., 2021, Hu et al., 2021, Canozo et al., 2022], graph clustering algorithms [Dries et al., 2019, Hao et al., 2020, Pham et al., 2020], and Bayesian statistical models [Zhao et al., 2021, Allen et al., 2021].

While each of these methodological categories presents unique advantages, they are fundamentally limited in that they do not explicitly model the interactive nature of cell sub-populations in a tissue sample [Barresi and Gilbert, 2019]. In other words, the sub-populations derived from existing methods are considered static, and no information is provided on how they relate to one another. Meanwhile, it is known that communication within and between groups of cells is a fundamental driver of healthy and diseased processes in a complex tissue [Armingol et al., 2021]. Moreover, Canozo et al. [2022] report substantial heterogeneity within traditional mouse olfactory bulb layer annotations, driven in part by spatial variation in intercellular communication patterns. However, detecting higher resolution cell sup-populations with existing tools is challenging as there is no principled methodology for determining which cell sup-populations may be members of a common broader phenotype (e.g., immune or cancer cell sub-types) based on similar yet distinct gene expression or spatial location patterns. As a consequence, current tools cannot be used to study the *community connectivity structure* of cell sub-populations, i.e., the relative similarity among cells within and between sub-populations.

By studying community connectivity structure in HST data, we may obtain valuable insights into the interactive dynamics of cell sup-populations in challenging settings such as the tumor microenvironment. For example, instead of simply labeling categories of immune cells and cancer cells in a tumor, we can describe how these important cell sub-populations relate to one another, and how tertiary intermediate sub-populations may be mediating important dynamics within the tumor microenvironment. Furthermore, characterizing community connectivity structure may help inform more biologically informative annotations of ambiguous sub-populations by relating them to more clearly defined sub-populations. Doing so may allow for a more biologically meaningful interpretation of all HST cell clusters in the common case when only a few cell clusters correspond clearly to a known cell type.

To address these gaps, we propose BANYAN (**B**ayesian **AN**alysis of communit**Y** connectivity in sp**A**tial single-cell **N**etworks): a statistical network model capable of discerning community connectivity structure in HST data. BANYAN draws inspiration from the vast field of biological network analysis [Guzzi and Roy, 2020], and is built on the supposition that HST data is most accurately represented as similarity networks that reflect similarity between cell spots in terms of spatial location and transcriptional profiles. BANYAN introduces the notion of analysis of community connectivity (ACC) to HST data analysis through implementation of a Bayesian multi-layered stochastic block model [Nowicki and Snijders, 2001, Valles-Catala et al., 2016] that infers sub-populations based on transcriptional and spatial similarity between cell spots. We offer convenient implementation and interactive visualization functionality via the `R` package `banyan` freely available at `https://github.com/carter-allen/banyan`.

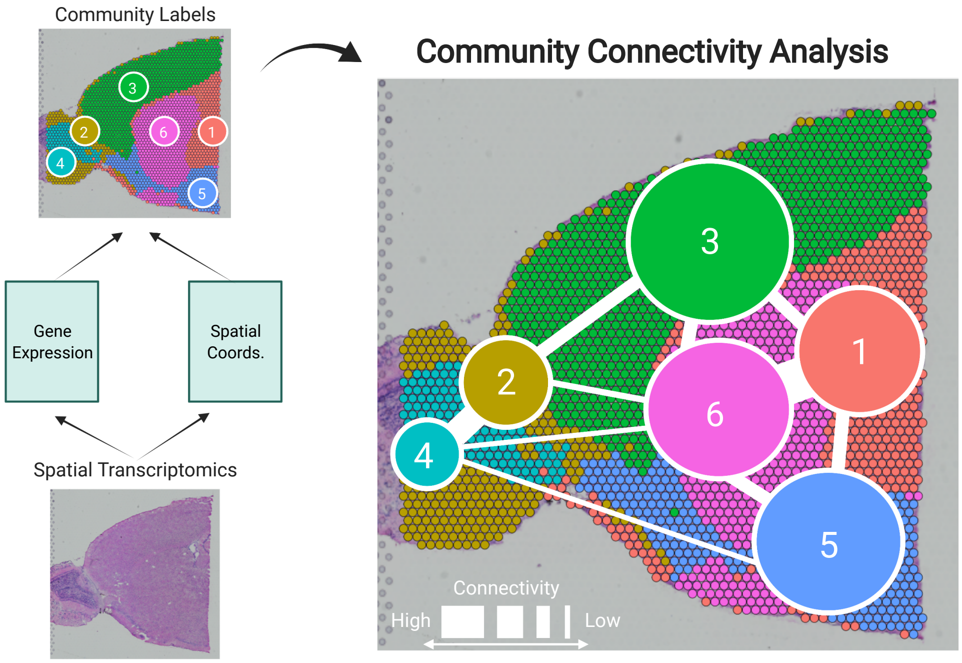


Figure 1: **Analysis of community connectivity**. Spatial transcriptomics platforms yield gene expression and spatial coordinate matrices, which may be used to derive community labels. analysis of community connectivity is achieved using BANYAN, which integrates gene expression profiles with spatial locations to infer the connectivity within and between communities.

2 Results

2.1 BANYAN allows for the analysis of community connectivity in HST data

BANYAN is the first HST computational tool to allow for *analysis of community connectivity* (ACC), i.e., the process of inferring the similarity of cells within and between sub-populations. A graphical representation is given in Figure 1, and the workflow to acheive ACC can be summarized as follows. First, given cell spot-level gene expression features and spatial coordinate data from HST platforms such as 10X Visium, we construct two spot-spot nearest neighbors networks. These networks are then integrated into a multi-layer graph data structure. Then, we fit the Bayesian multi-layer stochastic block model (MLSBM). The estimated parameters from this model allow us to (i) characterize community structure using cell spot sub-population labels, (ii) quantify uncertainty in predicted sub-population labels to identify ambiguous community structure regions, and (iii) infer the community structure of the tissue sample by quantifying the relative similarity between cell spots within and between sub-populations.

2.2 Discovering community structure in invasive ductal carcinoma

Accounting for roughly 25% of all non-dermal cancers in women, breast cancer ranks as the most common non-dermal female-specific cancer type, and narrowly the most common cancer type across both sexes [WCRF, 2020]. Of all sub-types, invasive ductal carcinoma (IDC) is the most common and most severe, accounting for roughly 80% of all breast cancers in women [Harris et al., 2012]. While previous authors have used spatial trancriptomics to study IDC samples relative to ductal carcinoma in situ (DCIS) samples [Yoosuf et al., 2020], IDC has yet to be studied through the lens of community structure due to the lack of computational tools available for performing ACC with HST data.

To illustrate ACC in the tumor microenvironment, we applied BANYAN to a publicly available IDC sample sequenced with the 10X Visium platform [10x Genomics, 2020]. We identified five spatially distinct cell spot sub-populations (Figure 2A), with associated uncertainty measures (Figure 2B). We then identified community structure by computing posterior estimates of within and between-community connectivity parameters, displayed in Figures 2C and 2D, respectively. Finally, to interpret each sub-population in terms of IDC biology, we computed the most differentially expressed genes between each sub-population and all others using the Wilcoxon Rank Sum test (Figure 2E).

Figure 2E displays a clear block structure in the expression of sub-population marker genes, indicative of strong community structure signal in the data. These marker genes can be used to obtain a number of interesting biological insights regarding the community structure of the IDC sample. For instance, the *S100A11* gene, a marker for sub-population 1, has been shown to be a diagnostic marker in breast cancers [Liu et al., 2010] and has been implicated in aggressive tumor progression [McKiernan et al., 2011]. Further, *KRT8* is used to differentiate aggressive grades of IDCs [Walker et al., 2007]. While outside of the context of IDCs, *DEGS2* has been shown to play a role in the invasion and metastasis of colorectal cancer [Guo et al., 2021]. Taken together, these marker genes suggest sub-population 1 contains a relatively high abundance of aggressive and invasive cancer cell types. On the other hand, sub-population 2 featured marker genes such as *MALAT1* that are associated with tumor suppressive behaviors in IDCs [Kim et al., 2018]. Another marker gene for sub-population 2, *CCDC80*, has been linked with tumor suppressive functions, albeit not in the context of IDCs [Ferraro et al., 2013].

Given these brief characterizations of sub-populations 1 and 2 available from the existing literature, we may hypothesize that these groups of cell spots are in some sense opposed in terms of their role within the tumor based on their transcriptional profiles. Indeed, these sub-populations also reside spatially at opposite ends of the tumor slice. We may investigate the similarity or dissimilarity of these sub-populations 1 and 2 using the between-community connectivity parameters presented in Figure 2D. We find that the estimate of this parameter is near zero (as evidenced by the black coloring of the entry (1,2) in Figure 2D), supporting our hypothesized dissimilarity between sub-populations 1 and 2.

In fact, sub-population 1 featured very low between-community connectivity with all other sub-populations besides sub-population 4, which occupies a heterogeneous “background” position in the spatial landscape of the tissue sample (Figure 2A) and therefore featured relatively high connectivity with all other communities. This spatial heterogeneity is accompanied by relatively low within-community connectivity (Figure 2C), which indicates that spot-spot similarities are less common between cell spots in sub-population 4 than in other sub-populations. In Figure 2E, it can be seen that many of the marker genes for sub-population 2 are shared by sub-population 4, including *MALAT1*, suggesting a similarity between these two sub-populations in terms of transcriptional profiles. In addition to the marker genes shared with sub-population 2, sub-population 4 features several of its own distinct marker genes, namely the immunoglobulin heavy chain-encoding RNAs *IGHG1* and *IGHG3*, which have themselves been shown to feature tumor suppressive tendencies via promotion of B cell specific immunoglobulin [Hsu et al., 2019], and have been associated with increased patient survival [Larsson et al., 2020]. This observation of functional similarity between sub-populations 2 and 4 is validated by Figure 2D, which clearly shows the highest estimated between-community connectivity in the data occurring between sub-populations 2 and 4. Taken together, these observations may lead us to reason that the sub-population 1 vs. 2 dynamic described previously is linked via the more heterogeneous yet still tumor suppressive-like sub-population 4.

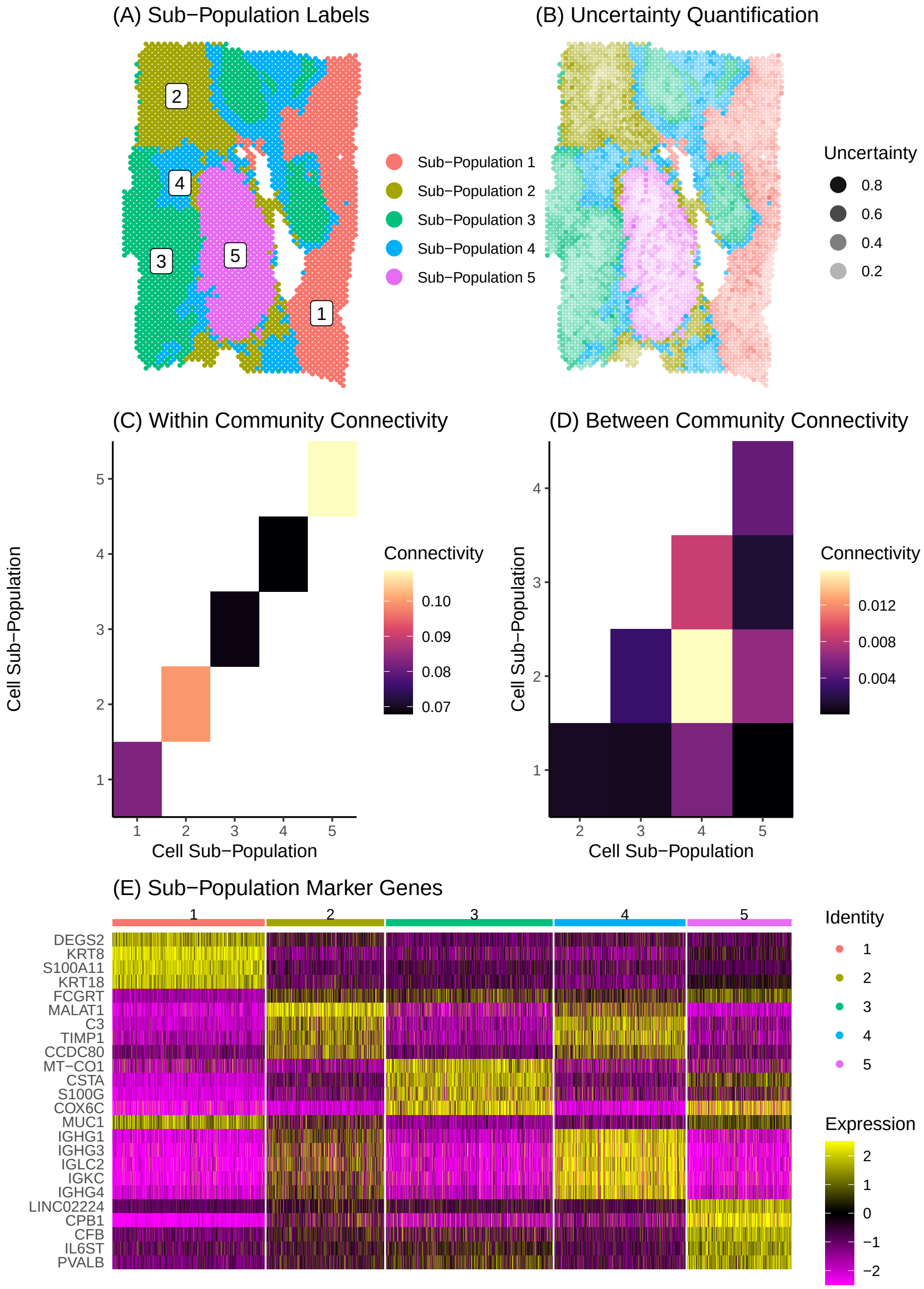


Figure 2: **Community structure in invasive ductal carcinoma.** (A) Inferred cell spot sub-population labels from BANYAN. (B) Relative uncertainty measures distinguish uncertain (dark) from certain (light) labels. (C) Within-community connectivity parameters reflect the homogeneity of sub-populations. Higher connectivity values reflect higher homogeneity within sub-populations. (D) Between-community connectivity parameters reflect the relative similarity of cell spots between sub-populations. Higher connectivity values reflect more similarity between sub-populations. (E) Sub-population markers genes.

While these observations would require further experimental validation to confirm, they showcase the unique ability of BANYAN to describe community structure in the data.

2.3 Characterizing ambiguous annotations in human white adipose tissue

Next, to demonstrate the application of community connectivity to inform ambiguous sub-populations, we applied BANYAN to the analysis of a human white adipose tissue (WAT) sample sequenced with the 10X Visium platform [Bäckdahl et al., 2021]. In contrast to the IDC data set considered in Section 2.2, WAT samples are characterized by weak histological organization, thus challenging the manual annotation of cell spots. In Figure 3A, we display the manual annotations from [Bäckdahl et al., 2021] for an individual WAT sample (ID: ADI24). Of the 2,747 cell spots in the original sample, 6 were annotated "unknown" and 1,520 were labeled "unspecific." Hence, over 50% of cell spots were unable to be clearly annotated due to ambiguous expression profiles or heterogeneous cell type mixtures within cell spots as a result of the resolution of the 10X Visium platform, a matter further complicated by the weak histological organization of WAT samples. Since existing approaches to labeling sub-populations in HST data fail to account for the community structure of tissue samples, disambiguating unknown cell spots *post hoc* remains challenging with existing tools.

In Figure 3B, we show the sub-population labels for the entire WAT sample ADI24 derived from BANYAN, where the unannotated cell spots (i.e., those classified as either "unknown" or "unspecific" by [Bäckdahl et al., 2021]) are highlighted in bold. We find that BANYAN identified residual heterogeneity within the 1,526 unannotated cell spots, with 513 unannotated cell spots labeled as sub-population 1, 178 unannotated cell spots labeled as sub-population 2, 64 unannotated cell spots labeled as sub-population 3, 99 unannotated cell spots labeled as sub-population 4, and 672 unannotated cell spots labeled as sub-population 5. We quantified the average uncertainty measures derived from BANYAN across the original annotation groups, and found the highest uncertainty occurring in the "unknown" or "unspecific" categories (Figure 3C), further validating the low signal contained in this subset of cell spots.

In Figure 3D, we present a heatmap depicting differentially expressed marker genes for each BANYAN sub-population using only the unannotated cell spots. The results from this analysis suggest that residual heterogeneity exists within the unannotated cell spot subset, with distinct marker genes present for sub-populations 1 through 4. Meanwhile, BANYAN sub-population 5 lacked clear markers, suggesting this sub-population could be reflective of a smaller ambiguous subset of cell spots within the original unannotated subset. In particular, as shown in the within-community structure displayed in Figure 3E, we find sub-population 3 to exhibit the highest within-community connectivity. This high within-community connectivity is supported by the contiguous spatial organization of sub-population 3 (Figure 3B) as well as high expression of distinct marker genes such as *VCAN* (Figure S1), which are suggestive of veriscan producing adipocytes associated with the development obesity-related inflammation of adipose tissue [Han et al., 2020].

When assessing the between-community structure of the WAT sample inferred by BANYAN as shown in Figure 3F, we find additional evidence for sub-population 5 being a heterogeneous unspecific sub-population (e.g., low within-community connectivity in Figure 3E and high between-community connectivity in Figure 3F). However, sub-population 5 did feature relatively high connectivity with sub-population 4, as evidenced by the bright coloring of entry (4,5) of Figure 3F. Sub-population 4 was marked by significant differential expression of adipose-resident immune cell related genes (Figure S1) such as *IGKC* [Vijay et al., 2020]. This suggests sub-population 5 may play an important role in mediating the function of immune cells within body fat compartments [Vijay et al., 2020]. While additional studies correlating these sub-populations with true single-cell data such as scRNA-seq would aid in further elucidation of these ambiguous sub-populations,

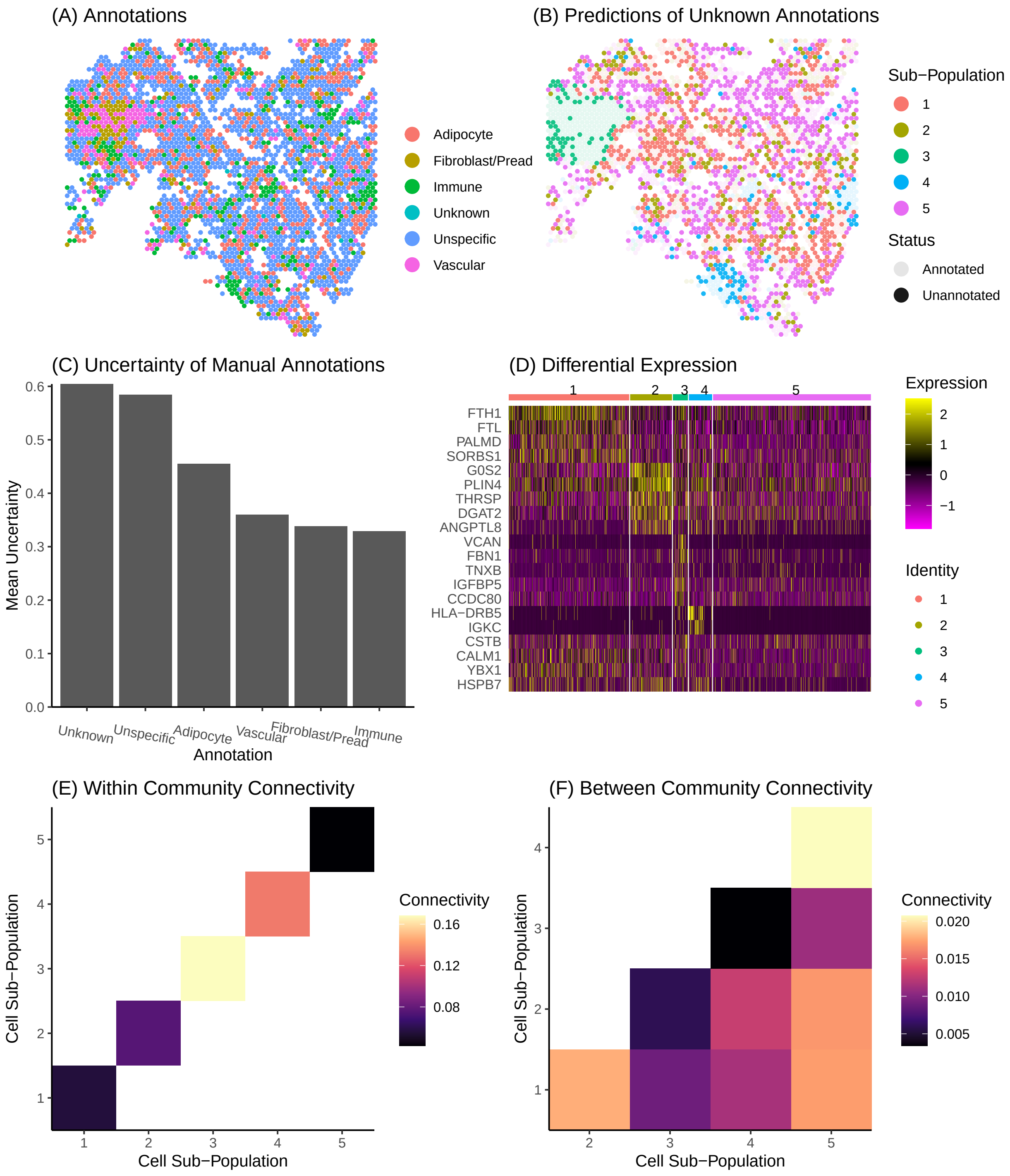


Figure 3: **Characterization of ambiguous cell spot annotations.** (A) Manual annotations. (B) BANYAN labels of unknown or unspecific (i.e., unannotated) cell spots from (A). (C) Mean BANYAN uncertainty scores by each annotation group. (D) Differential expression of BANYAN sub-populations using the unannotated subset. (E) Within-community connectivity between cell spots belonging to each sub-population. (F) Between-community connectivity between cell sub-populations.

leveraging community structure allows for more informative interpretation even in low-resolution HST data.

2.4 Simulation studies show BANYAN identifies community structure in low-signal settings

Finally, we designed a simulation study to validate the performance of the Bayesian multi-layer stochastic block model (SBM) employed by BANYAN, in the sense of identifying cell sub-populations and recovering community connectivity structure. We adopted a publicly available sagittal mouse brain data set [10x Genomics, 2019] sequenced with the 10X Visium platform. We manually allocated the $N = 2696$ total cell spots in the original sagital mouse brain data set into one of $K = 4$ simulated ground truth tissue segments, resulting in 4 spatially contiguous mouse brain layers (Figure 4A). The result of this was a ground-truth community structure that is reflective of sub-populations found in real HST data. We then formed the spot-spot spatial neighbors networks, using $R = 51$, the closest odd integer to $\sqrt{2696}$. We then simulated the spot-spot gene expression similarity network from a stochastic block model with community structure given by

$$\mathbf{\Theta} = \begin{bmatrix} \theta & 0.1 & 0.1 & 0.1 \\ 0.1 & \theta & 0.1 & 0.1 \\ 0.1 & 0.1 & \theta & 0.1 \\ 0.1 & 0.1 & 0.1 & \theta \end{bmatrix}, \tag{1}$$

where the *signal to noise ratio* (SNR) of the simulated gene expression network is given by SNR $= \theta/0.1$. SNR values much greater than 1 give rise to a strong community structure in the simulated data, while SNR values close to 1 result in a weaker community structure. We do not consider values of SNR below 1, as the resultant dissortative community structure is not reflective of cell type structure in HST data. We simulated gene expression networks for a range of SNR settings using $\theta = (0.105, 0.11, 0.13, 0.15, 0.17, 0.20, 0.22, 0.25)$ and fit two model variants: (i) a single-layer model considering gene expression information only, and (ii) the full model using both gene expression and spatial networks.

Figure 4B displays the average adjusted rand index (ARI) – a measure of accuracy in cell spot labels relative to the ground truth labels in Figure 4A, for the single-layer non-spatial approach and the multi-layer spatial approach. We find that at low SNR settings (e.g., below 1.5) the spatial model outperforms the non-spatial model in recovering ground truth cell spot labels. This is indicative of BANYAN's ability to leaverage spatial information to detect community structure in low-signal data. At higher SNR settings, the strong community structure signal contained in the simulated gene expression layer is sufficient for accurate community structure recovery, and the spatial information does not provide any further information.

We showcase the community structure results for one particular simulated data set at a setting of SNR $= 1.3$, reflective of a relatively low signal setting. Figure 4C displays the inferred community structure from the single-layer non-spatial model, while Figure 4D shows the same for the multi-layer spatial model. We find that at this moderately low signal setting, the non-spatial model is unable to accurately recover true cell spot labels, while the spatial model predicts cell spot labels almost perfectly. These results are indicative of the ability for spatial information to aid in disambiguating community structure in low-signal data settings.

While the simulated HST data generated from the community structure encoded in Equation (1) features a homogeneous community structure (i.e., uniform within and between-community connectivity parameters), BANYAN is capable of detecting more heterogeneous community structures.

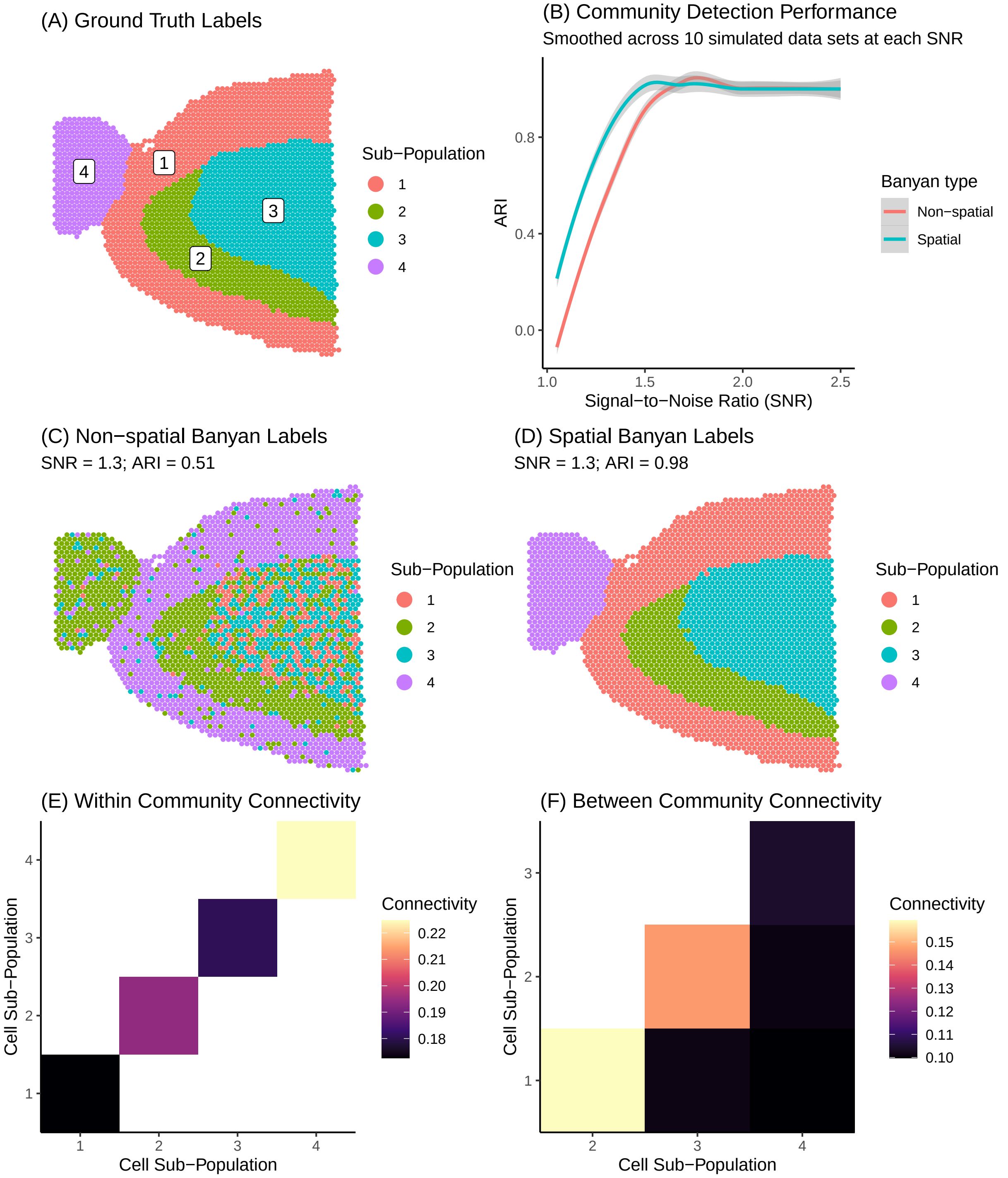


Figure 4: Segmenting sagittal mouse brain tissue sample to four different clusters. (A) Ground truth labels. (B) Average adjusted rand index (ARI) vs. signal-to-noise ratio (SNR). (C) Estimated cluster labels from a single layer (non-spatial). (D) Estimated cluster labels from corresponding multi layer (spatial). (E) Within-community connectivity estimates. (F) Between-community connectivity estimates.

To illustrate this, we generated a simulated mouse brain HST data set from

$$\mathbf{\Theta} = \begin{bmatrix} 0.30 & 0.15 & 0.10 & 0.10 \\ 0.15 & 0.30 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.30 & 0.15 \\ 0.10 & 0.10 & 0.15 & 0.30 \end{bmatrix}, \tag{2}$$

which features 50% stronger connectivity between sub-population pairs (1,2) and (3,4) than other between-community pairs. This setting reflects the common real data scenario wherein cliques of sub-populations form. In Figures 4E and 4F, we display the estimated within and between-community connectivity parameters, respectively. We find that the BANYAN model correctly identified the sub-population cliques (1,2) and (3,4) as sharing higher between-community connectivity than the remaining sub-population pairs, showcasing the model's ability to identify heterogeneous community structures. In Figure S2 of the Supplementary Materials, we demonstrate how the Bayesian information criterion (BIC) identifies the true number of communities in the simulated data, validating the use of statistical model fit criteria for choosing the number of sub-populations in the absence of prior knowledge.

3 Discussion

We have proposed BANYAN: a network-based statistical framework for analysis of community connectivity in HST data. We applied BANYAN to human breast cancer and white adipose tissue to illustrate its utility in applied settings. In the breast cancer case study, we found a strong community structure, with sub-populations marked by both invasive cancer and cancer suppressive marker genes. Using community structure parameters, we also identified an intermediate sub-population between these two. In the white adipose tissue case study, we demonstrated the use of BANYAN to disambiguate unknown or unspecific cell spot labels. In our simulation study, we validate BANYAN's ability to accurately identify the tissue architecture, especially in the low signal setting, and to recover the community connectivity structure. We provide the `R` package `banyan` for convenient implementation of the proposed workflow. The `banyan` package efficiently implements Bayesian estimation using custom Gibbs sampling algorithms implemented in `C++` using `Rcpp`. We have also developed interactive and static visualization functions for interrogation of BANYAN sub-population labels, uncertainty measures, and community structure. The `banyan` package interfaces seamlessly with standard `Seurat` workflows, and is freely available at `https://github.com/carter-allen/banyan`.

There a number of ways our work may be extended. First, often the SBM is refined to accommodate heterogeneous degree distributions among nodes, i.e., *degree correction* [Karrer and Newman, 2011]. By making this methodological extension to the multi-layer stochastic block model (MLSBM) at the core of Banyan, one could relax our assumption that each cell spot features the same number of neighbors and thereby allow for certain cells spots to feature more connections to the rest of the tissue than other cell spots, such as those on the periphery of the tissue sample. Learning the degree of each cell spot would then inform the detection of highly connected "hub" regions, or weakly connected "satellite" regions of a tissue sample. Second, the inherent complexity of network data structures leads to a heavy computational burden for large HST experiments. While we implement our proposed MCMC sampling algorithm using efficient `Rcpp` routines, BANYAN still requires significantly more computational time than non-network statistical methods Allen et al. [2021], Zhao et al. [2021]. Further optimization would help to reduce computational burden of community connectivity analysis. Finally, while BANYAN provides the first statistical framework for quantifying community connectivity structure in HST data, further extensions could be made to link BANYAN with methods for predicting cell-cell interactions using data such as ligand-receptor pair status of cells. By doing so, one could refine the general notion of cell spot connectivity to cell spot interaction, which is of major interest in HST data analysis.

4 Methods

4.1 Data pre-processing

To represent the interactive nature of cells and cell types, we adopt two cell-cell similarity networks as our primary data objects: one for gene expression and another for spatial location. To form the cell spot-cell spot gene expression similarity network, we first apply standard pre-processing steps including scaling, removal of technical artifacts, and identification of highly variable genes [Hao et al., 2020, seu, 2021a,b]. We then embed each of the N total cell spots in a lower-dimensional space using principal components analysis (PCA) applied to the top $2,000$ most variable genes. To form the cell spot-cell spot gene expression similarity matrix, we represent each cell as a node and connect each cell to its R closest neighboring cells in the gene expression principal component space using a binary edge. We utilize the same approach to construct the spatial cell spot-cell spot similarity network, where principal components are replaced with 2-dimensional spatial coordinates. The resultant data structure is two networks with N nodes, each of degree R. By default, we adopt the widely used heuristic of choosing R as the closet odd integer to $\sqrt{N}$ [Stork et al., 2001], which allows the number of neighboring spots to increase as the size of the tissue sample increases. With the typical HST experiment yielding a total number of cell spots between $2,000$ and $3,000$, this heuristic leads to consideration of between third and fourth order neighborhood structures (Fig 5). Overall, we view R as a tuning parameter that may be adjusted depending on the amount of information sharing desired across a tissue sample.

4.2 Model

We develop the core statistical model within BANYAN as an extension of the widely used stochastic block model (SBM) [Snijders and Nowicki, 1997], a flexible generative model for network data that allows for the assessment of community structure based on the frequency of binary edges among and between subsets nodes. We define $\mathbf{A}^1$ as the $N \times N$ binary adjacency matrix encoding the gene expression similarity network, and $\mathbf{A}^2$ as the binary adjacency matrix encoding the spatial similarity network. The matrix elements A^1_{ij} and A^2_{ij} indicate the presence or absence of a binary un-directed edge between nodes i and j for gene expression and spatial information, respectively. We define $\mathcal{A} = \{\mathbf{A}^1, \mathbf{A}^2\}$ as the multi-layer graph that encodes similarity between cell spots in terms of both gene expression and spatial information. While we focus on integration of spatial and gene expression information, our proposed framework may be extended to L layers to incorporate other sources of information from multiplexed experimental assays.

Given the multi-layer graph data $\mathcal{A}$, we assume that the absence or presence of edges in each layer between each pair of nodes i and j follows a Bernoulli distribution with probability of an edge θ_{z_i,z_j}, where $z_i \in \{1, ..., K\}$ denotes the latent cell sub-population assignment for cell i. We refer to such a model as a multi-layer stochastic block model (MLSBM). Formally, we assume for $l = 1, 2$,

$$A^l_{ij} | \mathbf{z}, \mathbf{\Theta} \overset{ind}{\sim} \text{Bernoulli}(\theta_{z_i,z_j}) \text{ for } i < j = 1, ..., N, \tag{3}$$

where $\mathbf{z} = (z_1, ..., z_N)$, and $\mathbf{\Theta}$ is a $K \times K$ *connectivity matrix* with diagonal elements θ_{rs} for $r = s = 1, ..., K$ controlling the probability of an edge occurring between two cells in the same sub-population, and off-diagonal elements θ_{rs} for $r < s = 1, ..., K$ controlling the probability of an edge occurring between two nodes in different cell sub-populations. Importantly, Model 3 implies that connections among cell spots in the gene expression and spatial layers are governed by a common set of community structure parameters $\mathbf{z}$ and $\mathbf{\Theta}$. Given Model 3 and data $\mathcal{A}$, our primary

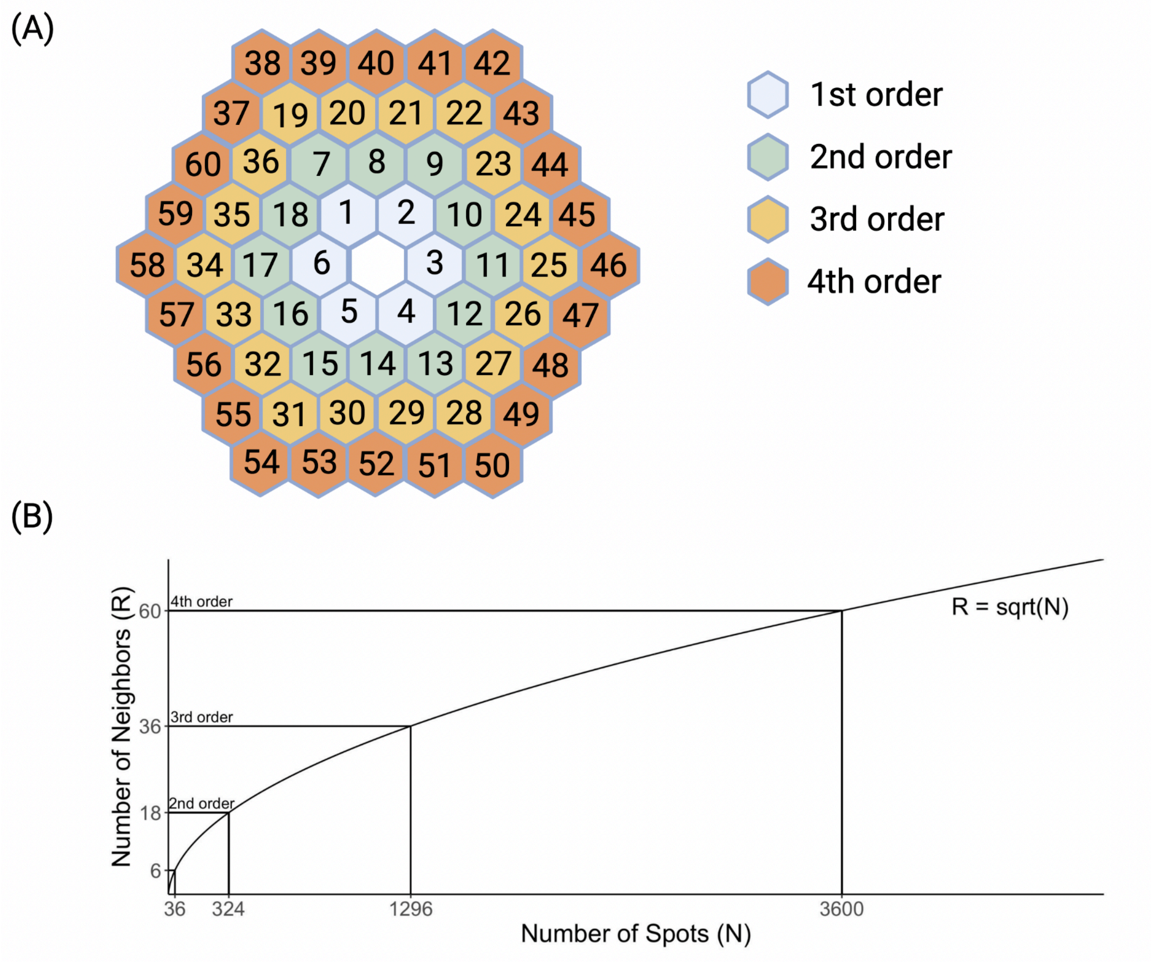


Figure 5: **Graphical depiction of relationship between number of neighbors and neighbor order.** (A) Hexagonal neighborhood structure for an interior cell spot shown with 1st through 4th order neighbors. (B) Suggested relationship between the number of cell spots (N) and the number of nearest neighbors (R).

inferential objective is to characterize cell sub-populations and the cell-cell interaction both *within* and *between* them by estimating the parameters $\mathbf{z}$ and $\mathbf{\Theta}$, which we accomplish using a Bayesian approach as described below.

4.3 Bayesian Inference

4.3.1 Priors

To achieve a fully Bayesian parameter estimation scheme, we assign prior distributions to all model parameters. We adopt available conjugate priors to obtain closed-form full conditional distributions of all model parameters, allowing for straightforward Gibbs sampling. For the latent cell sub-population indicators $z_1, ..., z_N$, we assume a conjugate multinomial-Dirichlet prior with $z_i \overset{iid}{\sim} \text{Categorical}(\boldsymbol{\pi})$ for $i = 1, ..., N$, and $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$, where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ controls the relative size of each cell sub-population to allow for a heterogeneous distribution of cell type abundances. We adopt a conjugate Beta-Bernoulli prior for $\mathbf{\Theta}$ by assuming $\theta_{rs} \overset{iid}{\sim} \text{Beta}(\beta_1, \beta_2)$ for $r < s = 1, ...K$. As a default, we opt for weakly informative priors by setting $\alpha_1 = \alpha_2 = ... = \alpha_K = 1$ and $\beta_1 = \beta_2 = 1$ [Gelman et al., 2013].

4.3.2 Markov chain Monte Carlo (MCMC) algorithm

The model proposed in Sections 4.2 and 4.3.1 allows for closed-form full conditional distributions of all model parameters. Thus, we adopt the following Gibbs sampling algorithm for parameter estimation. In practice, we recommend initializing the indicators $z_i, ..., z_N$ using a heuristic graph clustering method such as the Louvain algorithm [Blondel et al., 2008] applied to $\mathbf{A}^1$ to facilitate timely model convergence.

1. Update $\boldsymbol{\pi}$ from its full conditional $(\boldsymbol{\pi}|\mathbf{A}, \mathbf{z}, \mathbf{\Theta}) \sim \text{Dirichlet}(a_1, ..., a_N)$, where $a_k = \alpha_k + n_k$, and n_k is the number of nodes assigned to cell sub-population k at the current MCMC iteration, i.e., $n_k = \sum_{i=1}^N I_{z_i=k}$.

2. For $r \leq s = 1, ..., K$, update θ_{rs} from

$$(\theta_{rs}|\mathbf{A}, \mathbf{z}, \boldsymbol{\pi}) \sim \text{Beta}(\beta_1 + A[rs], \beta_2 + n_{rs} - A[rs]) \tag{4}$$

   where $A[rs]$ are the number of observed edges between communities r and s across both layers, and $n_{rs} = 2(n_r n_s - n_r I(r = s))$ are the number of possible edges between communities r and s, n_r is the number of nodes assigned to cell sup-population r, and $I(r = s)$ is the indicator function equal to 1 if $r = s$ and 0 otherwise.

3. For $i = 1, ..., N$, update z_i from $(z_i|z_{-i}, \mathbf{A}, \boldsymbol{\pi}, \mathbf{\Theta}) \sim \text{Categorical}(\boldsymbol{\rho_i})$, where $\boldsymbol{\rho_i} = (\rho_{i1}, ..., \rho_{iK})$ and

$$\rho_{ik} = \pi_k \left(\prod_{l=1}^{2} \prod_{j \neq i} \theta_{z_i, z_j}^{A_{ij}^l} (1 - \theta_{z_i, z_j})^{1 - A_{ij}^l} \right) \left(\prod_{l=1}^{2} \prod_{h \neq i} \theta_{z_i, z_h}^{A_{ih}^l} (1 - \theta_{z_i, z_h})^{1 - A_{ih}^l} \right). \tag{5}$$

4.3.3 Model selection

The choice of number of cell sub-populations K is a critical step in the analysis of HST data. In some cases, K may be chosen based on *a priori* biological knowledge of the cell types are expected to exist in a tissue sample, or the desire to investigate a known number of sub-populations within a more homogeneous tissue sample. In the absence of such prior information, K may be chosen using statistical model fit criteria, such as the Bayesian information criterion (BIC) [Schwarz et al., 1978].

4.3.4 Label switching

Label switching is a ubiquitous issue faced by models whose likelihood is invariant to permutations of a latent categorical variable such as $\mathbf{z}$. Consequently, stochastically equivalent permutations of $\mathbf{z}$ may occur over the course of MCMC sampling, causing the estimates of all community-specific parameters to be conflated, thereby jeopardizing the accuracy of model parameter estimates. Previous approaches for addressing label switching rely on re-shuffling posterior samples after completion of the MCMC algorithm [Papastamoulis, 2016]. However, such methods rely on prediction and thereby are subject to to prediction error. To protect against label switching within the MCMC sampler, we adopt the canonical projection of $\mathbf{z}$ proposed by Peng and Carvalho [2016], who restrict updates of $\mathbf{z}$ to the reduced sample space $\mathcal{Z} = \{\mathbf{z} : \text{ord}(\mathbf{z}) = (1, ..., K)\}$, wherein label switching is less likely due to the restricted sample space. In practice, we manually permute $\mathbf{z}$ at each MCMC iteration such that community 1 appears first in $\mathbf{z}$, community 2 appears second in $\mathbf{z}$, *et cetera*. Finally, we estimate $\mathbf{z}$ using the maximum *a posteriori* (MAP) estimate across all post-burn MCMC samples [Gelman et al., 2013].

4.4 Analysis of community connectivity

Estimation of the MLSBM model parameters $\mathbf{\Theta}$ and $\mathbf{z}$ with the corresponding maximum *a posteriori* estimates $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{z}}$ allow for inference of community connectivity structure in HST data. While the estimated community labeling vector $\hat{\mathbf{z}}$ is what we use to define communities, the elements of $\hat{\mathbf{\Theta}}$ describe how cell spots within and between communities relate to one another, thereby characterizing community connectivity. Specifically, elements $\hat{\theta}_{rs}$ reflect the estimated probability of a randomly chosen cell spot in community r sharing a nearest neighbors edge in $\mathcal{A}$ with a cell spot in community s. When $r = s$, $\hat{\theta}_{rs}$ reflects the average connectivity within a community, which may be used to assess the relative homogeneity of a community. Heterogeneous communities tend to have lower average within-community connectivity, while more homogeneous communities tend to have higher within-community connectivity. Likewise, when $r \neq s$, $\hat{\theta}_{rs}$ represents the probability of connection between cell spots in two distinct communities. This between-community connectivity measurement allows us to discern closely related communities that may contain similar cell types from more distinct communities. Taken together, these between and within-community connectivity parameters capacitate analysis of community connectivity.

4.5 Uncertainty quantification

Discrete clustering approaches for community structure identification inherently fail to account for heterogeneity within each spot cluster by assuming all cell spots within the same cluster are stochastically equivalent. To address these shortcomings of existing approaches, we utilize the inferential benefits of Bayesian modeling to derive two biologically relevant measures: (i) continuous phenotypes and (ii) uncertainty scores. With continuous phenotype measures, we may assess the propensity of a given cell spot for a cell type other than its most likely cell type, thus allowing for identification of possible intermediate cell states. Relatedly, uncertainty measures allow us to distinguish high confidence from low confidence cell type assignments.

Conceptually, we choose c_{ik}, the continuous phenotype for cell spot i towards cell sub-population k, to be proportional to the posterior probability $P(z_i = k | z_{-i}, \mathbf{A}, \boldsymbol{\pi}, \mathbf{\Theta})$. Considering terms only related to z_i, we define c_{ik} as

$$c_{ik} = \hat{\pi}_k \left(\prod_{l=1}^{2} \prod_{j \neq i} \hat{\theta}_{k,\hat{z}_j}^{A_{ij}^l} (1 - \hat{\theta}_{k,\hat{z}_j})^{1 - A_{ij}^l} \right) \left(\prod_{l=1}^{2} \prod_{h \neq i} \hat{\theta}_{k,\hat{z}_h}^{A_{ih}^l} (1 - \hat{\theta}_{k,\hat{z}_h})^{1 - A_{ih}^l} \right), \tag{6}$$

for $i = 1, ..., N$ and $k = 1, ..., K$, where $\hat{\pi}_k$, $\hat{\theta}_{z_i,z_j}$, and $\hat{z}_i$ are the posterior estimates of π_k, θ_{z_i,z_j}, and z_i, respectively. We define the uncertainty measure for cell spot i as $u_i = 1 - c_{i\hat{z}_i}$, i.e., the sum of cell spot i continuous phenotypes for all cell types besides $\hat{z}_i$.

Supplementary Information

Figure S1: Differentially expressed markers for BANYAN sub-populations across all cell spots in WAT sample. Differential expression p-values were computed using the Wilcoxon Rank-Sum test. Sub-populations 1 through 4 display clear marker genes while sub-population 5 remained more ambiguous.

Figure S2: Identification of true number of communities in simulated data using BIC. Higher values indicate better model fit.

Acknowledgements

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

References

Seurat - guided clustering tutorial. `https://satijalab.org/seurat/articles/pbmc3k_tutorial.html`, 2021a. Accessed: 2021-05-27.

Analysis, visualization, and integration of spatial datasets with seurat. `https://satijalab.org/seurat/articles/spatial_vignette.html#acknowledgments-1`, 2021b. Accessed: 2021-05-27.

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

10x Genomics. Human breast cancer (block a section 1); spatial gene expression dataset by space ranger 1.1.0. `https://support.10xgenomics.com/spatial-gene-expression/datasets/1.0.0/V1_Breast_Cancer_Block_A_Section_1`, 2020.

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

Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

Michaela Asp, Joseph Bergenstrahle, and Joakim Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays*, 42(10):1900221, 2020.

Jesper Bäckdahl, Lovisa Franzén, Lucas Massier, Qian Li, Jutta Jalkanen, Hui Gao, Alma Andersson, Nayanika Bhalla, Anders Thorell, Mikael Rydén, et al. Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell metabolism*, 33(9):1869–1882, 2021.

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

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008 (10):P10008, 2008.

Francisco Jose Grisanti Canozo, Zhen Zuo, James F Martin, and Md Abul Hassan Samee. Cell-type modeling in spatial transcriptomics data elucidates spatially variable colocalization and communication between cell-types in mouse brain. *Cell Systems*, 13(1):58–70, 2022.

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

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

Angelo Ferraro, Filippo Schepis, Vincenza Leone, Antonella Federico, Eleonora Borbone, Pierlorenzo Pallante, Maria Teresa Berlingieri, Gennaro Chiappetta, Mario Monaco, Dario Palmieri, et al. Tumor suppressor role of the cl2/dro1/ccdc80 gene in thyroid carcinogenesis. *The Journal of Clinical Endocrinology & Metabolism*, 98(7):2834–2843, 2013.

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

Wei Guo, Cuiyu Zhang, Panpan Feng, Mingying Li, Xia Wang, Yuan Xia, Dawei Chen, and Jingxin Li. M6a methylation of degs2, a key ceramide-synthesizing enzyme, is involved in colorectal cancer progression through ceramide synthesis. *Oncogene*, 40(40):5913–5924, 2021.

Pietro Hiram Guzzi and Swarup Roy. *Biological Network Analysis: Trends, Approaches, Graph Theory, and Algorithms*. Elsevier, 2020.

Chang Yeop Han, Inkyung Kang, Ingrid A Harten, John A Gebe, Christina K Chan, Mohamed Omer, Kimberly M Alonge, Laura J den Hartigh, Diego Gomes Kjerulf, Leela Goodspeed, et al. Adipocyte-derived versican and macrophage-derived biglycan control adipose tissue inflammation in obesity. *Cell reports*, 31(13):107818, 2020.

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

Jay R Harris, Marc E Lippman, C Kent Osborne, and Monica Morrow. *Diseases of the Breast*. Lippincott Williams & Wilkins, 2012.

Huan-Ming Hsu, Chi-Ming Chu, Yu-Jia Chang, Jyh-Cherng Yu, Chien-Ting Chen, Chen-En Jian, Chia-Yi Lee, Yueh-Tao Chiang, Chi-Wen Chang, and Yu-Tien Chang. Six novel immunoglobulin genes as biomarkers for better prognosis in triple-negative breast cancer by gene co-expression network analysis. *Scientific reports*, 9(1):1–12, 2019.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Jongchan Kim, Hai-Long Piao, Beom-Jun Kim, Fan Yao, Zhenbo Han, Yumeng Wang, Zhenna Xiao, Ashley N Siverly, Sarah E Lawhon, Baochau N Ton, et al. Long noncoding rna malat1 suppresses breast cancer metastasis. *Nature genetics*, 50(12):1705–1715, 2018.

Christer Larsson, Anna Ehinger, Sofia Winslow, Karin Leandersson, Marie Klintman, Ludvig Dahl, Johan Vallon-Christersson, Jari Häkkinen, Cecilia Hegardt, Jonas Manjer, et al. Prognostic implications of the expression levels of different immunoglobulin heavy chain-encoding rnas in early breast cancer. *NPJ breast cancer*, 6(1):1–13, 2020.

Xiang-Guo Liu, Xiao-Ping Wang, Wan-Feng Li, Shuo Yang, Xin Zhou, Si-Jie Li, Xiang-Jun Li, Dong-Yun Hao, and Zhi-Min Fan. Ca2+-binding protein s100a11: a novel diagnostic marker for breast carcinoma. *Oncology reports*, 23(5):1301–1308, 2010.

Eadaoin McKiernan, Enda W McDermott, Dennis Evoy, John Crown, and Michael J Duffy. The role of s100 genes in breast cancer progression. *Tumor Biology*, 32(3):441–450, 2011.

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

Panagiotis Papastamoulis. label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software*, 69(1):1–24, 2016.

Lijun Peng and Luis Carvalho. Bayesian degree-corrected stochastic blockmodels for community detection. *Electronic Journal of Statistics*, 10(2):2746–2779, 2016.

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

Gideon Schwarz et al. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464, 1978.

Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

David G Stork, Richard O Duda, Peter E Hart, and D Stork. Pattern classification. *A Wiley-Interscience Publication*, 2001.

Toni Valles-Catala, Francesco A Massucci, Roger Guimera, and Marta Sales-Pardo. Multilayer stochastic block models reveal the multilayer structure of complex networks. *Physical Review X*, 6(1):011036, 2016.

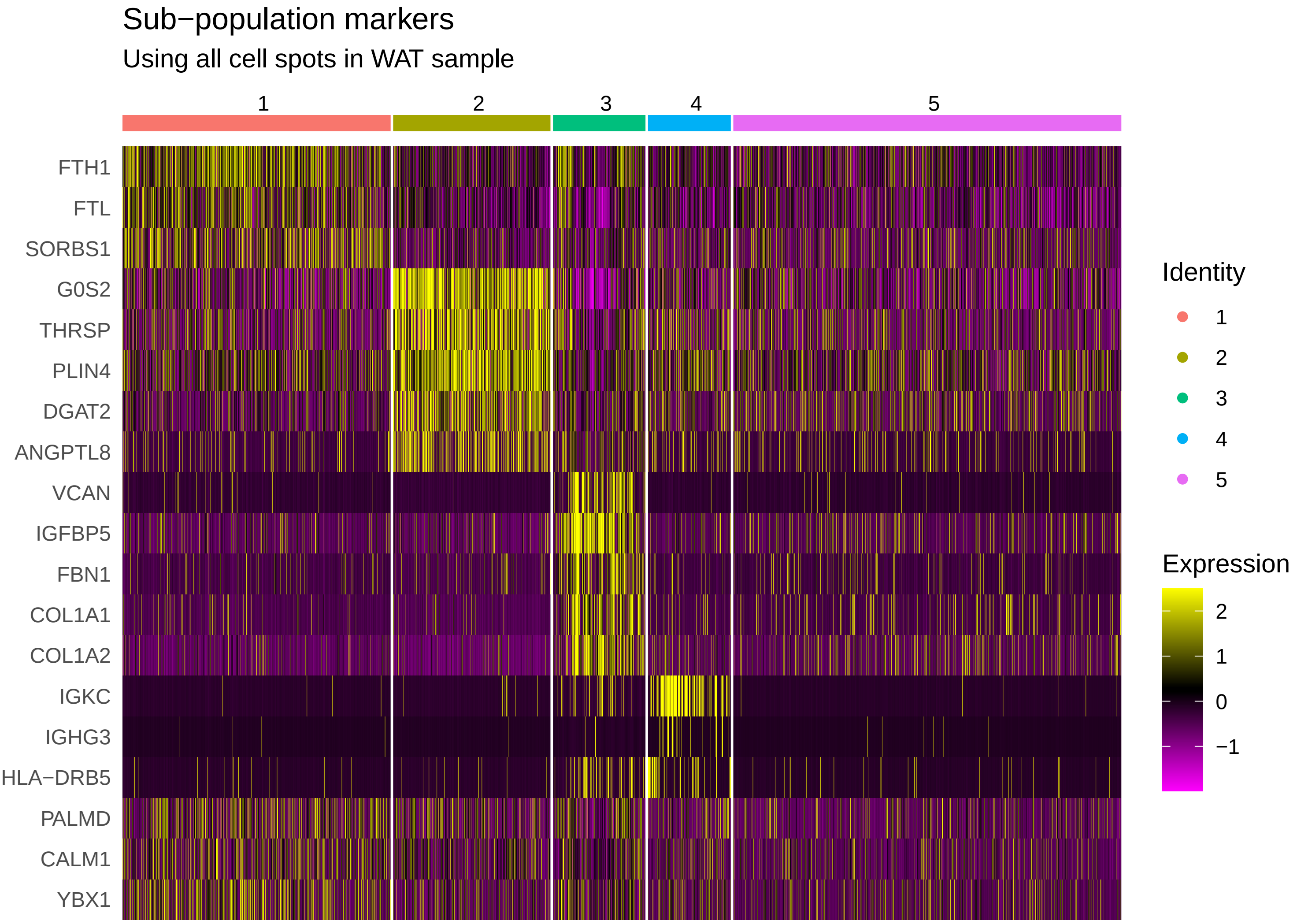
Jinchu Vijay, Marie-Frédérique Gauthier, Rebecca L Biswell, Daniel A Louiselle, Jeffrey J Johnston, Warren A Cheung, Bradley Belden, Albena Pramatarova, Laurent Biertho, Margaret Gibson, et al. Single-cell analysis of human adipose tissue identifies depot-and disease-specific cell types. *Nature metabolism*, 2(1):97–109, 2020.

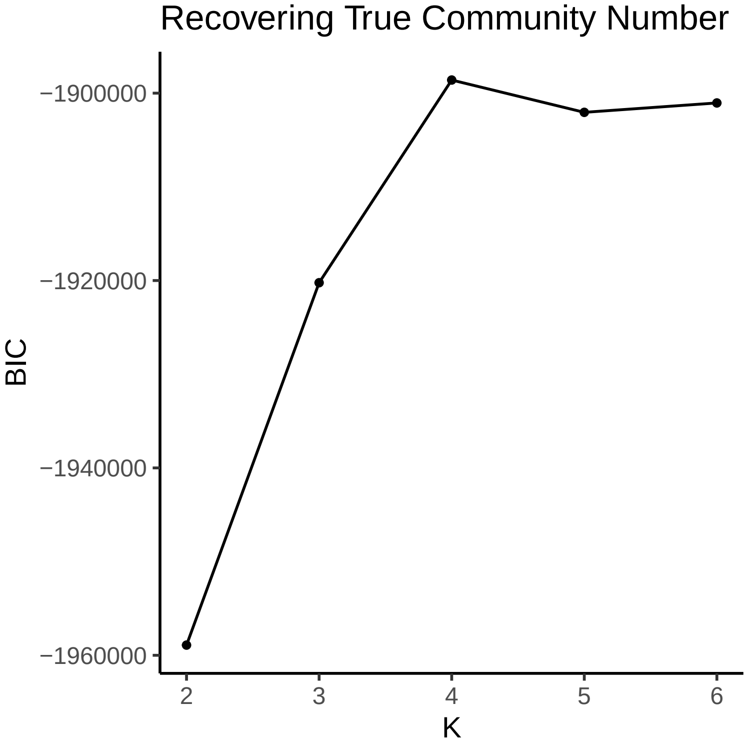
Logan C Walker, Gavin C Harris, Andrew J Holloway, Grant W McKenzie, J Elisabeth Wells, Bridget A Robinson, and Christine M Morris. Cytokeratin krt8/18 expression differentiates distinct subtypes of grade 3 invasive ductal carcinoma of the breast. *Cancer genetics and cytogenetics*, 178(2):94–103, 2007.

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

Niyaz Yoosuf, José Fernández Navarro, Fredrik Salmén, Patrik L Ståhl, and Carsten O Daub. Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Research*, 22(1):1–10, 2020.

Edward Zhao, Matthew R Stone, Xing Ren, Thomas Pulliam, Paul Nghiem, Jason H Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, 2021.


Sub−population markers
Using all cell spots in WAT sample
1
2
3
4
5
FTH1
FTL
SORBS1
G0S2
THRSP
PLIN4
DGAT2
ANGPTL8
VCAN
IGFBP5
FBN1
COL1A1
COL1A2
IGKC
IGHG3
HLA−DRB5
PALMD
CALM1
YBX1
Identity
1
2
3
4
5
Expression
2
1
0
−1


Recovering True Community Number
BIC
−1900000
−1920000
−1940000
−1960000
2
3
4
5
6
K