SpatialSort: Characterizing Cellular Heterogeneity in the Tumour Microenvironment with Spatially Aware Clustering

by

Eric Lee

B.Sc., The University of British Columbia, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

December 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

SpatialSort: Characterizing Cellular Heterogeneity in the Tumour Microenvironment with Spatially Aware Clustering

submitted by **Eric Lee** in partial fulfillment of the requirements for the degree of **Master of Science** in **Bioinformatics**.

Examining Committee:

Dr. Andrew Roth, Assistant Professor, Departments of Computer Science and Pathology and Laboratory Medicine, UBC
Supervisor

Dr. Andrew Weng, Professor, Department of Pathology and Laboratory Medicine, UBC
Supervisory Committee Member

Dr. Gabriela V. Cohen Freue, Associate Professor, Department of Statistics, UBC
Supervisory Committee Member

Abstract

In the course of tumour progression, normal and malignant cells of various kinds of cell types engage in complex patterns of cell-cell interactions creating the tumor microenvironment. The dynamics in the tumour microenvironment are tumour-driven and fosters cellular heterogeneity to modulate cancer behavior. With an accurate classification of the cell type composition and a deep investigation of the cell-cell interactions, we can characterize the heterogeneity in the tumor microenvironment and potentially elucidate mechanisms of immune invasion, tumour growth, and metastases.

Analyses of single cells in suspension using mass cytometry is a current approach used to characterize previously unknown phenotypes, yet the data generated by this approach are disaggregated and do not retain the spatial structure of tumours. Emerging high-throughput spatial expression profiling technologies, such as imaging mass cytometry allow for spatially aware profiling of single cell expression in high-parameter space. Various cell type classification methods have been proposed for disaggregated data, however there is a need for spatially aware clustering methods.

We present SpatialSort, a scalable joint approach for spatially aware clustering of cell types and estimation of cell-cell interactions in the tumour microenvironment. This computational approach leverages a Markov random field model to allow spatially proximal cells linked in a neighbour graph to influence the cluster assignment of their neighbouring cells. Markov chain Monte Carlo sampling is employed to approximate the posterior distribution and perform probabilistic cell type identification. Cell to cell interactions will be encoded as interpretable model parameters representing the affinity between different cell types.

Through spatially aware clustering, we hope to characterize patient-specific phenotypic heterogeneity better than our current methods. As heterogeneity promotes therapeutic resistance, an improved understanding of cellular composition and cell-cell interaction profiles can potentially provide better prognosis for cancer patients.

Lay Summary

In the course of tumour progression, normal and malignant cells cross-interact within an evolving entity termed as the tumour microenvironment. The dynamics of these interactions fosters heterogeneity which affects tumour behavior and complicates treatment. With an accurate classification of the cell type composition and an investigation of the cell-cell interactions, it is possible to improve our characterization of the tumour microenvironment.

Various methods have been developed to cluster cell types using data from current technologies despite the shortcoming of not being able to retain spatial information. New technologies have been able to capture spatial structure however robust spatially aware clustering methods are yet to be developed.

We present SpatialSort, a scalable joint approach for spatially aware clustering of cell types and estimation of cell-cell interactions. Through the addition of spatial information, we hope to characterize tumour microenvironment heterogeneity better than current methods and potentially provide better prognosis for cancer patients.

Preface

I have completed the work detailed in this thesis under the supervision of Dr. Andrew Roth at the BC Cancer Research Centre. This project was proposed and initially designed by Dr. Andrew Roth. I jointly contributed to the modification of the model design with a student in statistics, Kevin Chern. I was the major contributor to the source code of the project with support from Kevin on mathematical derivations and the implementation of the double Metropolis Hastings algorithm. I implemented the experiments and performed benchmarking.

Statistical experiments on the synthetic data was done jointly with Kevin, while the statistical experiments and biological analysis on semi-real and real-world data was done by myself. The data used for analysis is publicly available and is acknowledged in this thesis. All public software packages are also acknowledged in this thesis.

Table of Contents

List of Tables

List of Figures

Glossary

CYTOF Cytometry by Time-of-Flight

DMH Double Metropolis Hastings

HMM Hidden Markov Model

HMRF Hidden Markov Random Field

IMC Imaging Mass Cytometry

MCMC Markov Chain Monte Carlo

METABRIC Molecular Taxonomy of Breast Cancer International Consortium

MH Metropolis Hastings

MPEAR Maximization of Posterior Expected Adjusted Rand

MRF Markov Random Field

SW Swendsen-Wang

TME Tumour Microenvironment

Acknowledgments

I would like to thank my supervisor, Dr. Andrew Roth, for his support in the many phases of this project. Your guidance has allowed me to explore creative solutions to tackle this challenging project and advance my understanding about building statistical software. I have been able to challenge myself to climb the steep learning curve virtually and independently.

I would also like to extend great appreciation to the committee members who have provided different valuable viewpoints and suggestions to assist in the implementation of this project. The committee meetings have been critical and constructive to allow for greater achievements.

A big thanks are to all the members in the Roth Lab who have brought interesting discussions from both the biological and computational literature to our weekly lab meetings and reading groups. The resilience from all the members continuing research during a pandemic at the very start of our first year together as a lab has motivated me to advance on.

It is also extended with great appreciation to the Canadian Institutes of Health Research to support my research through the Frederick Banting and Charles Best Canada Graduate Scholarships Masters award granted in April 2021.

Last but not least, I would like to give thanks to my family and friends who have accompanied me through my ups and downs.

The most special and biggest thanks are to my parents, Nina (Hsiu Mei) Wu and Shing Lee - who have always been my best supporters - to continuously support my every decision throughout my life. No words can describe how fortunate I am to be born into this loving family.

Chapter 1

Introduction

A human body consists of trillions of cells of hundreds of cell types that cycle through phases of growth, division, and death. When problems occur in the process, uncontrollable cell growth can occur that form tumours which can transform to a potentially malignant disease, which we refer to as cancer. The complexity arises because of cancer being a genetically and phenotypically diverse disease. A major area of research is the tumour microenvironment - an environment of heterogeneous cellular composition and activity.

In this chapter, we aim to introduce the concept of cellular heterogeneity in the tumor microenvironment by addressing cellular composition and cell-cell interactions as two main determinants of the tumour microenvironment. We will also introduce the recent technologies and methods, such as mass cytometry and imaging mass cytometry, in which both have been employed to conduct research on the analysis of the tumour microenvironment. Lastly, we will describe the research question, the hypothesis, and the main contributions on the topic of inferring the cellular composition and the cell-cell interactions which are made in this thesis.

1.1 Cellular Heterogeneity in the Tumour Microenvironment

1.1.1 Composition of the Tumour Microenvironment

Cancerous tumours are complex organs composed of cells of different types along with secreted proteins, blood vessels, and an extracellular matrix [21]. In the tumour ecosystem, it is not solely a blob of malignant cells but a heterogeneous mixture of cells such as immune cells, stromal cells and various other cell types that jointly contribute to the cellular composition [32]. The composition of cells can be viewed as a snapshot representation of the state of a tumour. Exploring the composition can allow us to understand the behavior of the tumour as well as make predictions on the outcomes of cancer [51].

The prognosis of cancer is however usually complicated by the genetic and phenotypic cellular diversity during tumour progression [3, 43]. Cellular diversity is observed in an complex evolving entity referred to as the Tumour Microenvironment (TME) [41]. The TME is an ecosystem that hosts a collection of resident host cells and infiltrating cells that are recruited to site [25]. The exact composition varies depending on the tumour type.

The TME is often viewed as an active promoter of tumour progression. As the tumour progresses, tumour cells often initiate changes to the cellular and physical properties of the microenvironment as well as the surrounding host tissues, thus enhancing cellular and phenotypic diversity [38].

The clinical behavior of some cancers, such as follicular lymphoma, has been shown to be more greatly impacted by the TME than the inherent properties of the malignant subclones [19]. The presence of specific cell types in the TME has been employed as a prognostic parameter to assess progression and survival.

The development of the tumour is also accompanied with a growth in the heterogeneity of phenotypes in cell types which are difficult to classify [5]. The heterogeneity of phenotypes can be succinctly described as a condition of various cells of the same type having different marker signatures. This usually leads to poor prognosis and clinical implications that are of high complexity [6]. Investigating the cellular composition of a tumour biopsy then becomes essential for providing a snapshot of the cell types comprising the tumour, yet the analysis is incomplete without describing the activity of the cell types [2].

1.1.2 Cell-Cell Interactions

In the course of tumour progression, normal cells and malignant (transformed) cells engage in dynamic patterns of cell-cell interactions in the tumour. The complex communication processes between different cell types together creates the Tumour Microenvironment (TME) [50]. In the TME, interactions occur predominantly when cells are spatially proximal to each other, with exceptions of secreted stimulus from distant sites [49]. Close distances between cells allows for the ligand-receptor binding and cell surface contact for cellular activity and signal transduction.

Various factors including cell type, cell count, spatial location, protein signals, and other factors determine type of interactions that cells partake in and thus the role of the cells [53]. For instance, the role of immune cells in the TME is known to be able to either suppress or promote tumour formation [52]. An common example of such would be a cytotoxic T cells invading malignant cells, which is characterized as an anti-tumour cell-cell interaction [36]. Due to the highly variable cellular composition and interactions, the TME is often characterized based on the collective behavior of the cells in the environment, such as being an anti-tumour immune microenvironment or an immune suppressive microenvironment.

The tumour type is a major factor dictating the roles of the cell types in the TME. An example of such is the subtypes in B cell lymphoma. B cell lymphomas are classified into subtypes based on range of malignant cell content, in which there is a spectrum from Hodgkin lymphoma being 1 percent malignant to Burkitt lymphoma being more than 90 percent [42]. In the case of Burkitt lymphoma, the TME is sparse with only a small percentage of macrophages with the rest being malignant cells. On the contrary, Hodgkin has a dense TME with little malignant cells present [15]. It can be observed that different levels of interactions are present in each of these subtypes, and a measure can be created to characterize the affinity of the interactions.

It is important to understand that patterns of interactions within the TME are known to be mainly tumour-driven and fosters cellular heterogeneity to modulate cancer behavior [33]. The consequence of the malignant phenotypes is that it can lead to patient specific TME compositions, serving as an effective barrier to treatment [17]. Taking a clinical perspective, investigating the cell-cell interactions that characterizes heterogeneity in the TME can potentially elucidate mechanisms of tumour cell invasion, tumour growth, and metastases.

1.2 Current Methods

1.2.1 Disaggregate Single Cell-omics

The exploration of the TME has been an ongoing area of research in which various methods have been developed. In the field of proteomics, the analysis of single cells in suspension using Cytometry by Time-of-Flight (CYTOF), also referred to as single-cell mass cytometry, is a current approach used to characterize previously unknown phenotypes and advance the understanding of tumour heterogeneity.

Single-cell mass cytometry is an instrument based on mass spectrometry that is designed for real time detection and quantification of markers on single cells [7]. The detection depends on a panel of metal isotope conjugated antibodies that stain single cells to identify target protein markers. Using this approach, we can obtain single-cell proteomic profiles from biological samples that are heterogeneous in nature, such as cancer biopsies [27].

Given the expression data from CYTOF, clustering is often a necessary task to identify cell types. An example of a current popular clustering method is Phenograph [28]. This method performs partitioning of high-dimensional single-cell expression data into subpopulations by employing the Louvian community detection method. It has been successfully applied to many experiments that require the clustering of heterogeneous cell populations.

However, a limitation to CYTOF is that tumour biopsies are dissociated into single cell suspensions rendering the generated data being disaggregated. The major drawback is then the loss of information about the original spatial structure of tumours which is a direct measurement of which cells are interacting with which. It should be noted that CYTOF is a destructive process where the original nuclei locations cannot be recovered after expression data are collected.

1.2.2 Spatially Resolved Single Cell-omics

Single-cell mass cytometry has allowed us to answer the questions of composition but leaves open the question of how cells are organized in the tissue [44]. In order to understand the functional behavior of the heterogeneous mixture of cells as a whole, there is a need to know which types of cells are likely to be near and interacting with each other in the TME [12].

New high-throughput spatial expression profiling technologies are allowing us to do spatial proteomics [31]. An example of such is Imaging Mass Cytometry (IMC) [23], which is an expansion of mass cytometry to incorporate image acquisition of the spatial structure. In order to collect the spatial context, IMC introduced a novel laser ablation device on the mass cytometer to allow for direct measurements of metal isotopes at each spot location on the tissue sections. This enables imaging for about 40 markers and retains spatial structure allowing for the analysis of cellular heterogeneity in a spatially aware sense.

Alternative imaging modalities that allow for spatial localization include multiplexed ion beam imaging (MIBI) [4] and co-detection by indexing (CODEX) [24]. Different methods are categorized by their antibody conjugation techniques [46]. MIBI uses metal conjugated antibodies which are ionized by high-energy beams to generate secondary ions detected by an imaging mass spectrometer over a five-log dynamic range. CODEX uses DNA barcodes, fluorescent dNTP analogs, and an in situ polymerization-based indexing procedure.

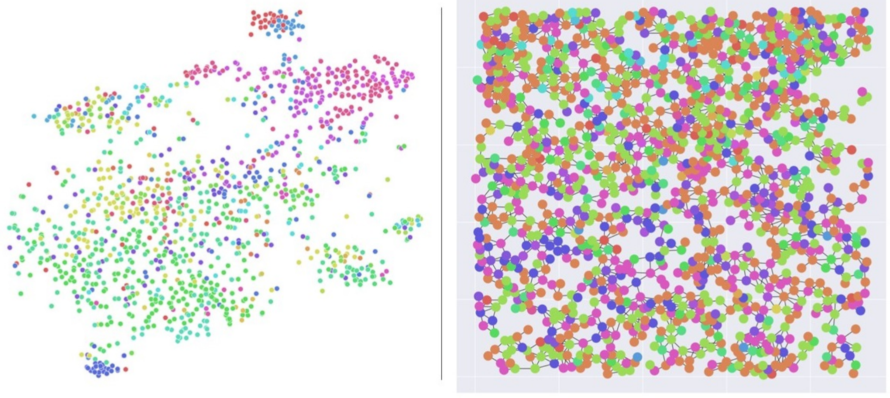


Figure 1.1: Disaggregate data visualized with a t-SNE graph (left) ; Spatial data visualized with a spatial neighbour graph (right). The graph on the left is a visualization of dimensionally reduced expression data color labeled by cell types. Cell type labelling is done by Phenograph. t-SNE axes are abstract coordinates that do not carry meaning. The graph on the right is a spatial omics approach where cells are plotted by spatial location. Non-random patterns of cells can be observed. Cells of the same types are clustered together to different extents depending on cell types.

Despite the effective capturing of spatial information, the analysis of spatially resolved data are still using tools designed for high parameter mass cytometry data which ignores the spatial information [14]. The cell clusters from IMC are still determined based solely on disaggregate data and do not incorporate spatial context. For instance, current analyses using IMC have been projecting clusters back to the IMC image to identify anatomical relationships [26]. The incorporation of spatial information as a parameter in clustering could potentially improve the accuracy in the assignment of cell type labels in clustering algorithms [9].

1.3 Objectives

1.3.1 Research Question

We formulate the research question as: Given the emerging high throughput spatial technologies, would spatially aware clustering reveal a more accurate assignment of cell type labels and effectively capture cell-cell interactions? In addition, which spatial models are the most appropriate for cell type clustering?

1.3.2 Hypothesis

We hypothesize that methods which cluster cells based on disaggregate single cell expression data can potentially cluster erroneously by not accounting for information on spatial structure. By addressing spatial information as well as the uncertainty of the cell type assignments in a new clustering method, we hypothesize there can be a better assignment of cell types and can uncover the interaction effects.

1.3.3 Thesis Contribution

In this thesis, we address our research question by developing a statistical framework, SpatialSort, which is a scalable joint approach for spatially aware clustering of cell types and estimation of cell-cell interactions. We fit our model to simulated data, semi-real data, and real-world data to demonstrate its utility in resolving cellular heterogeneity by providing cell type cluster assignments, cell-cell interaction matrices, and neighbour graph visualizations for patients of different subtypes of cancer.

Chapter 2

Spatially Aware Clustering

This chapter will be describing our novel approach to address the research question in Chapter 1. We will provide an overview of SpatialSort and address the methods in great detail. We will introduce the hidden Markov random field model that represents the underlying spatial structure of the cellular composition of a patient. We will also describe the inference procedure using algorithms from the Markov Chain Monte Carlo (MCMC) family, including Gibbs Sampling and Double Metropolis Hastings (DMH), as well as the Swendsen-Wang algorithm that allow for the probabilistic assignment of cell types and the estimation of cell-cell interactions for each patient subtype. Lastly, we will describe our method to make point estimates of cluster assignments.

2.1 Method Overview

Given the high-throughput spatial profiling of single cell expression profiles, we present a statistical method, SpatialSort, that probabilistically clusters cells to cell type clusters and estimates the cell-cell interactions between the cell type clusters.

In Figure 2.1, we show a visual representation of the overview of SpatialSort. This method requires five user inputs:

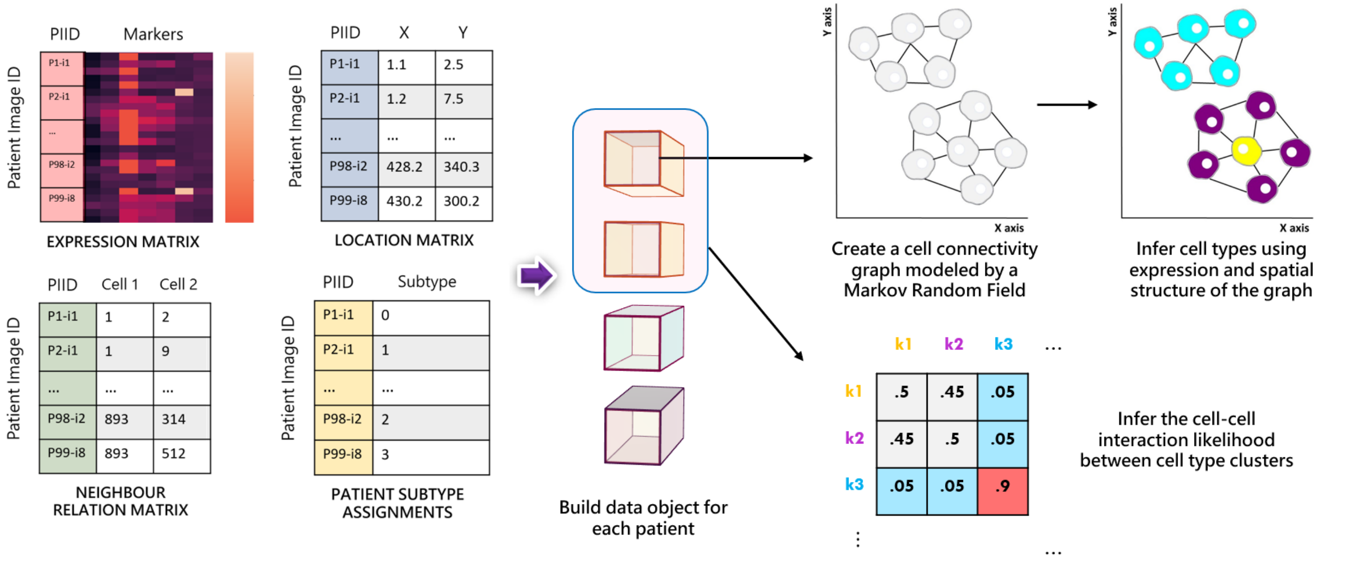
- A cell by marker expression matrix for the cells of all patients labeled by patient ID.
- A cell coordinate location matrix with the spatial coordinates of the nucleus or center of the imaged cell membrane for each cell in the same row order as the expression matrix.
- A cell neighbour relation matrix representing pairs of cells that are considered interacting in the spatial structure of a patient's cellular composition.
- A list of patient subtype assignments for each patient.
- A prior matrix quaternary coded to express prior belief of the marker expression levels for user expected cell types. This is an optional, yet important input for high performance.

Using the inputted information, SpatialSort employs a Bayesian approach to perform joint inference of cell type labels and cell-cell interaction for each patient subtype. This requires both the expression data and a neighbour graph representing cell connectivity which is built for each patient using the location and neighbour relation matrices provided.

The functions of SpatialSort are able to provide four major user outputs:

- A list of cell type cluster assignments for the cells of all patients.
- A cell-cell interaction matrix for each patient subtype describing the affinity of a cell type cluster to another cell type cluster, along with a list of interaction term point estimate values.

- A cell connectivity graph for each patient coloured by the cell type clusters along with its expression heat map.
- An expression heat map for each cluster.


Patient Image ID
PIID
Markers
P1-i1
P2-i1
...
P98-i2
P99-i8
EXPRESSION MATRIX
LOCATION MATRIX
NEIGHBOUR RELATION MATRIX
PATIENT SUBTYPE ASSIGNMENTS
Build data object for each patient
Y axis
X axis
Create a cell connectivity graph modeled by a Markov Random Field
Infer cell types using expression and spatial structure of the graph
Infer the cell-cell interaction likelihood between cell type clusters

PIID	X	Y
P1-i1	1.1	2.5
P2-i1	1.2	7.5
...	...	...
P98-i2	428.2	340.3
P99-i8	430.2	300.2

PIID	Cell 1	Cell 2
P1-i1	1	2
P2-i1	1	9
...	...	...
P98-i2	893	314
P99-i8	893	512

PIID	Subtype
P1-i1	0
P2-i1	1
...	...
P98-i2	2
P99-i8	3

	k1	k2	k3	...
k1	.5	.45	.05	
k2	.45	.5	.05	
k3	.05	.05	.9	
⋮				...

Figure 2.1: Overview of Spatially Aware Clustering. Conceptually, the four input matrices: expression, location, relation and patient subtype assignments are pre-processed to build patient-specific data objects that are shown in boxes colored by patient subtype. For each patient, a neighbour graph modeled by a Markov Random Field (MRF) is built to jointly infer cell types and cell-cell interaction likelihoods between cell type clusters which are interpretable parameters of the MRF.

2.2 Model Description

We propose a hierarchical Bayesian model to model expression data with spatial information, for instance IMC data.

Figure 2.2 shows the probabilistic graphical model, which is a graphical representation of the conditional dependencies present in this probabilistic model. Let Y denote a cell by marker expression matrix for N cells and M markers of P patients. Also, let C represent the known patient subtypes of P patients. Suppose θ to be the combination of mean and precision parameters of M marker expressions for K clusters. The latent cell types for N cells of P patients, X, is modelled with a Markov Random Field (MRF). Lastly, β represents the interaction term where it is supposedly unique for C patient subtypes.

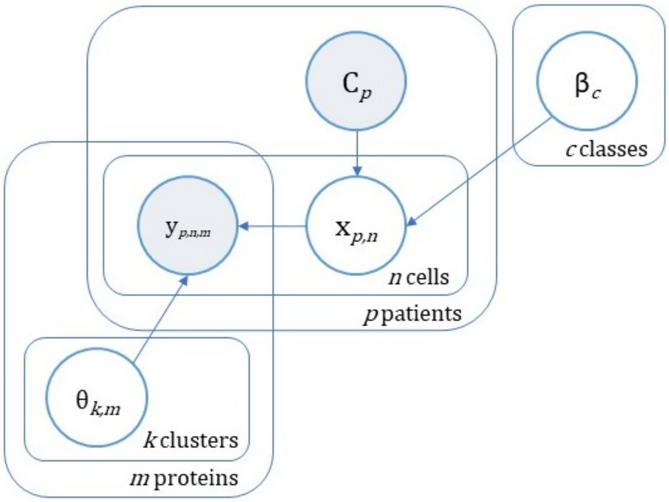


Figure 2.2: Probabilistic Graphical Model of SpatialSort. Shaded circles are observed variables, while non-shaded circles are latent variables to infer. Plates represent the repetition of variables with respect to the notation at the corner. Prior distributions and variable descriptions are in Table 2.1.

2.3 Prior Probability Distributions

In a Bayesian model, prior probability distributions, colloquially noted as priors, are carefully chosen to express some prior belief towards the uncertain values of variables before observing the data. We describe the priors for the variables in the following:

$$\begin{aligned} \theta_{k,m} &\sim NormalGamma(\mu_0, \lambda_0, \alpha_0, \beta_0) \\ x_{p,n}|\beta_{C_p=c} &\sim HotPotts(\beta_{C_p=c}) \\ \beta_c &\sim Beta(\alpha_0, \beta_0) \\ C_p &\sim Categorical(\pi_p) \\ y_{p,n,m}|x_{p,n}, \theta_{k,m} &\sim Gaussian(\mu_{x,m}, \sigma^2_{x,m}) \end{aligned}$$

In the expression data, each of the patients p has a marker expression vector of marker panel length m for each cell n, which we denote as $y_{p,n,m}$. It is conditioned on two parameters, which are the latent mean and precision parameter $\theta_{k,m}$ for each marker m in cluster k and the latent cell type cluster label $x_{p,n}$ for each cell n of patient p, and is distributed according to a Gaussian$(\mu_{x,m}, \sigma^2_{x,m})$ distribution where x is a realization of a cell type cluster label.

Table 2.1: Description of Variables and Prior Distributions for SpatialSort. A graphical representation is shown in Figure 2.2

Variable	Represents	State
$y_{p,n,m}$	Marker m expression of cell n of patient p	Observed
$x_{p,n}$	Cell type label of cell n of patient p	Latent
$\theta_{k,m}$	Combined mean and variance of marker m expression of cluster k	Latent
β_c	Cell-cell interaction parameter of class c	Latent
C_p	Patient subtype of patient p	Observed

2.3.1 Mean and Precision: Normal-Gamma Distribution

$\theta_{k,m}$ can be written as $(\mu_{k,m}, \tau_{k,m})$ in which it is distributed according to a Normal-Gamma distribution. This distribution is conjugate to the Gaussian distribution with an unknown mean μ and variance σ^2. It is also the conjugate prior for the Gaussian distribution that eases the computation in the posterior inference.

The distribution is defined as the follows:

$$\begin{aligned}\tau &\sim Gamma(\alpha, \beta)\\ Y|\tau &\sim Gaussian(\mu, \sqrt{1/(\lambda \cdot \tau)})\end{aligned}$$

in which α is the shape parameter and β is the rate parameter for the Gamma distribution, and the precision parameter τ has a Gamma distribution. The parameters for the Gamma distribution are selected using the variance of the expression data and parameter searching. Details are defined below.

$$\begin{aligned}&\lambda_0 = 0.1\\ &\mu_\tau = 1/(\lambda_0 \cdot s_Y^2)\\ &\sigma_\tau^2 = \begin{cases} 1 & \text{if uncertainty is low}\\ 100 & \text{if uncertainty is high}\end{cases}\\ &\alpha_0 = \mu_\tau \cdot \beta_0\\ &\beta_0 = \mu_\tau / \sigma_\tau^2\end{aligned}$$

where s_Y^2 is the sample variance of Y.

The mean parameter μ is determined by the prior matrix given as a user input. We also assume that the expression data is properly normalized in the preprocessing of the data. The input prior matrix has dimensions K clusters by M marker and discrete with quaternary codes. $\mu_{k,m}$ is translated to the 25th, 50th, and 75th percentiles for each marker expression if the prior matrix has codes 0, 1, 2 respectively. -1 is a special case which means a value of 0. Having a code of 2 for a particular marker of a cluster indicates prior knowledge of a high expression value. 1 and 0 follows the same idea and represents a user's belief of middle and low to no expression respectively. The code -1 is only used when coupled with a extremely uncertain precision which occurs in the case when we do not have prior knowledge on the expression of markers.

2.3.2 Cell Type: Hidden Markov Random Field Model

In order to classify the cell type of each cell, we designed a Hidden Markov Random Field (HMRF) where spatially proximal cells of a patient are linked in an undirected graph. A HMRF is a generalization of the Hidden Markov Model (HMM), defined as having a Markov Random Field (MRF) as the underlying stochastic process. MRFs are known as undirected graphical models G which have V nodes representing variables and E edges that connect pairs of nodes [10]. The variable of interest in inference is the hidden state of the nodes.

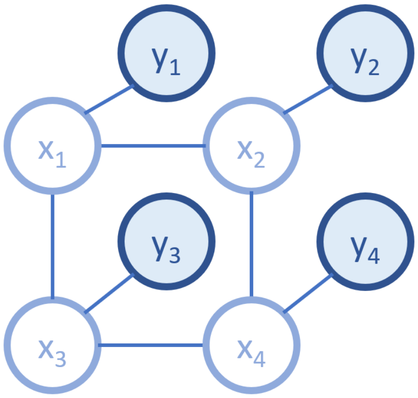


Figure 2.3: Hidden Markov Random Field. The variables y_i are observed variables, while the variables x_i are latent variables to infer.

HMRFs are commonly used in performing Bayesian image processing and analysis where it serves as a smoothing prior for object segmentation. Recent work in the spatial transcriptomic field have used the Potts model to identify transcriptional heterogeneity [54].

In the discrete MRF models, we chose to modify the Potts model to be the prior of X, which we coin the term ”HotPotts”. The Potts model was designed to explore the interactions between different internal elements of a system. The Potts model is a lattice graph in which each node is assigned a spin in a number of finite labels [48]. The combinations of spins and edges decide the interactions of nodes, which represents different states. The function of the energy, also known as the Hamiltonian, is defined as the follows:

$$h(w) = -J \cdot \sum_{i,j \in E(G)} \delta_{\sigma_i, \sigma_j}$$

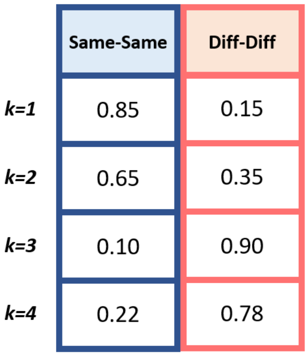
where J is the interaction strength and δ is equal to one when $\sigma_i = \sigma_j$ and zero otherwise.

This is closely related to our model in which each patient has a spatial structure with cells are represented by nodes and are linked by edges represent cells that are interacting. Yet, the definition of interacting cells here is dependent on the user's interpretation, which in this framework is user inputted. A method to draw edges between nodes can be to consider two cells as interacting if the cell area of both cells is overlapping in the processing of image segmentation.

Through performing inference on the MRF, the goal will be to uncover the latent cell types assignments of a single patient p where $X_p = (X_{p,n=1}, X_{1,2}, \ldots, X_{1,N_p})$. The MRF is parameterized by a latent interaction term β_{C_p} where $C_p = c$ that is a known patient subtype parameter. The β parameter is chosen to have a Beta distribution. When the shape parameters α, β are both equal to 1, it is identical to the Uniform distribution as an uninformed prior due to uncertainty. It can also be set to 0.1, to have distinct values of 0 or 1 to indicate how likely do clusters like to neighbour each other.

During initial exploratory analysis, three variants of MRF parameterization were considered: a one-parameter model where a β_c is scalar (referred to as the 1p model), a linear model where the dimension of a β_c is K (referred to as the 2k model), and a quadratic model where the dimension of β_c is $\binom{K}{2} + K - 1$. The quadratic model suffers from an exponential run time in the inference and was not considered afterwards.

We chose to experiment using the 1p and 2k models that we hypothesize to model cell-cell interactions effectively. Let us define the 2k model in which β is a matrix with K rows indicating the number of cell type clusters, and 2 columns. Let the first column to contain the affinity values where there is an interaction between the cells labeled k (referred to as same-same interactions). Also let the second column to contain the affinity values where there is an interaction of cell labeled k with other cell types $-k$ (referred to as diff-diff interactions). The affinities of label k have a distribution of Beta(1,1) and the property of the two types to sum up to one.



	Same-Same	Diff-Diff
k=1	0.85	0.15
k=2	0.65	0.35
k=3	0.10	0.90
k=4	0.22	0.78

Figure 2.4: The 2k β model as an interaction parameter in the MRF.

Using the 2k model, we can write the sufficient statistic for patient p as T_p and define it as:

$$T^p = (T^p_{k_1,k_2})_{\{k_1,k_2\}\in[2K]}$$

where

$$T^p_{\{k_1,k_2\}}(X_p) = \sum_{(u,v)\in E(G_p)} \mathbb{I}(\{X_{p,u}, X_{p,v}\} = \{k_1,k_2\})$$

in which the indicator function $\mathbb{I}$ evaluates to 1 when the statement is true, and 0 otherwise. Then, the MRF model for patient p has a density as:

$$\mathbb{P}(X_p \mid \beta_{C_p=c} = b) \propto \exp\left(\sum_{i,j} [b \circ T^p(X_p)]_{i,j}\right)$$

where the $\circ$ the Hadamard product between b and the sufficient statistic.

Effectively, a HMRF allows for the cluster assignment of each cell to be influenced by their neighboring cells, which is a natural way to model cellular organization in the tumour microenvironment.

2.4 Bayesian Inference

In this statistical framework, Bayesian inference was adopted as the inference method. Posterior inference was performed using Markov Chain Monte Carlo (MCMC) techniques.

For the updates on X, collapsed Gibbs sampling was used to integrate out the mean and precision parameters of the expression and the Swendsen-Wang (SW) algorithm was also employed to sample from the HotPotts distribution. The Double Metropolis Hastings (DMH) algorithm was used to update betas where sampling beta faces the problem of doubly intractable normalizing constants.

In this section, I will be addressing the proposed joint approach for spatially aware clustering to infer cell types assignments and estimate cell-cell interactions.

2.4.1 Updating X: MRF Inference

Estimating the parameters of a Gaussian distribution by the use of conjugate priors is a convenient and common approach in Bayesian inference. A prior is defined as a conjugate prior if the prior and the posterior are in the same probability distribution family. This property can be exploited to allow results to be derived in closed form and removes the need to deal with computationally intractable multi-dimensional integrals.

In the updates of the labels X of cells, we exploit the properties of conjugacy. Recall that the observed expression data Y is distributed according to a Gaussian distribution with parameters $(\mu_{x,m}, \sigma^2_{x,m})$. We model the parameters to have a Normal-Gamma distribution which is the conjugate prior of the Gaussian distribution [35]. Let us formulate the computation of the posterior as,

$$\begin{aligned} P(x_{p,i}, \theta | y_p, x_{p,-i}, \beta, C_p) &\propto P(y_p|\theta, x_p) P(x_p|\beta) P(\theta) P(\beta) \\ &= \Pi_{k\in K}\Big(\Pi_{j\in I_{p,k}} P(y_{p,j}|\theta_k, x_{p,j})\Big) P(\theta_k) P(x_p|\beta) P(\beta) \end{aligned}$$

where $I_{p,k}$ refers the indices of the cells of patient p where the cell label is of cluster k.

We calculate likelihood as a product of probabilities per cluster k. We can further expand this equation to incorporate conjugacy to form a collapsed likelihood.

The derivation is described as the follows:

$$\begin{aligned}
P(x_{p,i}|y_p, x_{p,-i}, \beta, C_p) &\propto \int (\Pi_{k\in K}\Pi_{j\in I_{p,k}} P(y_{p,j}|\theta_k)P(\theta_k))d\theta_k P(x_p|\beta)P(\beta) \\
&= \Pi_{k\in K}\int P(y_{p,I_{p,k}}|\theta_k)P(\theta_k)d\theta_k P(x_{p,i}|x_{p,-i},\beta)P(\beta) \\
&= \Pi_{k\in K} Z_k P(x_{p,i}|x_{p,-i},\beta)P(\beta)
\end{aligned}$$

in which $P(x_p) \propto P(x_{p,i}|x_{p,-i})$.

Using the above conjugacy, we can derive a closed form solution for the likelihood, which is the marginal likelihood of the Normal-Gamma distribution, as described below:

$$Z_k(\mu_{n_k}, \lambda_{n_k}, \alpha_{n_k}, \beta_{n_k}) = \frac{\Gamma(\alpha_{n_k})}{\beta_{n_k}^{\alpha_{n_k}}}\Big(\frac{2\pi}{\lambda_{n_k}}\Big)^{1/2}$$

where

$$\begin{aligned}
\mu_{n_k} &= \frac{\lambda_0\mu_0 + n_k\bar{y}}{\lambda_0 + n_k} \\
\lambda_{n_k} &= \lambda_0 + n_k \\
\alpha_{n_k} &= \alpha_0 + n_k/2 \\
\beta_{n_k} &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n_k}(y_i - \bar{y})^2 + \frac{n\lambda_0(\bar{y} - \mu_0)^2}{2(\lambda_0 + n_k)}
\end{aligned}$$

in which n_k refers to the number of cells in cluster k and $\bar{y}$ refers to the mean of the expression y.

With the described likelihood derivation, we are left with the prior term from the MRF. Let the MRF prior be the product of the energy function for each neighbour v of the node of interest u. We can write it as,

$$\sum_{v\in neighbour(u)} \frac{1}{2}\cdot\omega\cdot\beta_{x_{p,u},[x_{p,u}\neq x_{p,v}]}\cdot\beta_{x_{p,v},[x_{p,u}\neq x_{p,v}]}$$

where ω is a strength parameter that scales up the influence of the prior which is with default set to the number of markers M in the expression data.

Recall that β is a two-column matrix where each cluster label has an affinity value where there is an interaction of the same type of cells (referred to as same-same interactions), and also an affinity value where there is an interaction between cells of different types (referred to as diff-diff interactions). The two affinities also have the property of summing up to one. Applying the betas here in this function, we will have two cases. First is the case of the same-same interaction, the energy function is simply the beta term of the same-same interactions of cluster $x_{p,u}$ scaled by ω. Second is the case of the diff-diff interaction, the energy function then is the average between the beta terms of the diff-diff interactions of cluster $x_{p,u}$ and $x_{p,v}$ scaled by ω. In our notation, the type of interactions present is specified by the Iverson bracket, which it takes the value of 1 if the statement in the bracket is true and else 0.

With the likelihood and the prior for each a cell at all possible configurations in an array Q, we formulate a probability function to calculate the probability of finding the node with a particular label k:

$$\exp\left(\frac{Q_{p,u,k}}{\log \sum_{k=1}^{K} \exp(Q_{p,u,k})}\right)$$

A new label for the node is sampled from this distribution and the labels are updated accordingly. Below is a summary of the algorithm to update cell labels in the MRF.

Algorithm 1 Updating Labels: MRF Inference

```
for $p$ = $1$ to $P$ do
    for $u$ = $1$ to $N_p$ do
        for $k$ = $1$ to $K$ do
            $x_{p,u} \leftarrow k$
            $H_{p,u,k} \leftarrow \sum_{v \in neighbour(u)} \frac{1}{2} \cdot \omega \cdot \beta_{x_{p,u},[x_{p,u} \neq x_{p,v}]} \cdot \beta_{x_{p,v},[x_{p,u} \neq x_{p,v}]}$
            $L_{p,u,k} \leftarrow$ Log Marginal Likelihood
            $Q_{p,u,k} \leftarrow H_{p,u,k} + L_{p,u,k}$
        end
        Sample a new label $x_{p,u}$ with the probability function $\exp(\frac{Q_{p,u,k}}{\log \sum_{k=1}^{K} \exp(Q_{p,u,k})})$
        Update the labels for $x_{p,u}$
    end
end
```

2.4.2 Updating X: Swendsen-Wang Algorithm

Another optional step to improve the updates of the cell type labels would be to apply a generalization of the Swendsen-Wang (SW) algorithm. The SW algorithm was initially proposed as a method of simulation for large systems near criticality [45]. The algorithm is valid for discrete Ising and Potts models and slows down when applied to Bayesian inference.

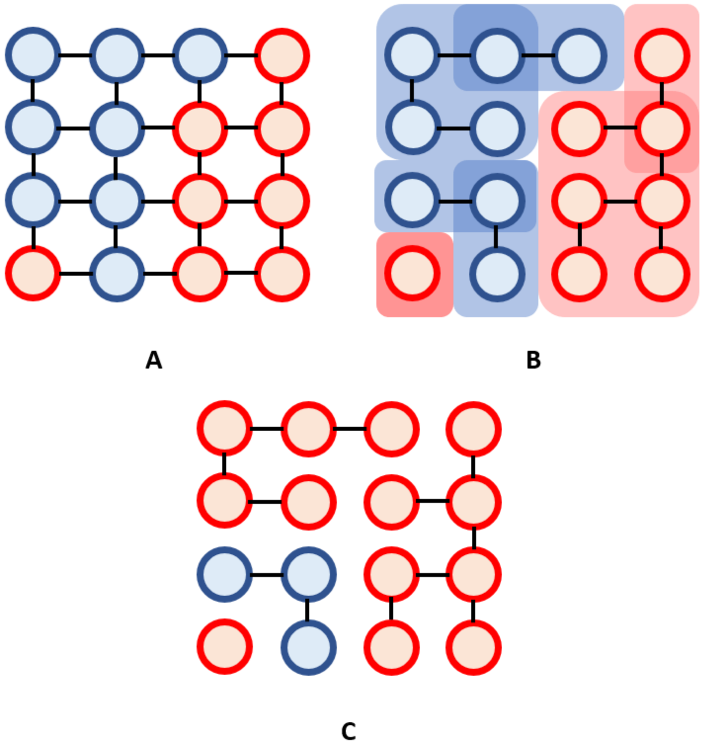


Figure 2.5: An overview of the generalized Swendsen-Wang algorithm. A is a connected graph that is binary labeled. In B, we draw edges with probability for each connected component that has the same label. We result in four components. C is the result of updating labels for each component.

A generalization of the SW [8] was introduced for arbitrary posterior probabilities which we implement in this model as an addition to the inference on the MRF. The SW algorithm allows for splitting, merging, and regrouping components of a graph at each step, which is more efficient than per node updates.

The generalized SW is often thought of as a Metropolis Hastings (MH) step, where it takes a reversible move between two graph partition states. We describe the implementation of our version of the SW algorithm for moving from a state to its proceeding state in three procedural steps.

First, in our graph an edge is drawn with a local discriminative probability for each edge that have nodes with the same label.

$$Uniform(0, \exp(\omega \cdot \beta_{x_u} \mathbb{I}(x_u = x_v))) > 1$$

where I is an indicator function that gives the value 1 if the statement in parenthesis is true, otherwise gives 0.

Second, we identify all the connected components and the graph cuts in the graph. The connected components are considered an island in the graph where the updates are performed as a whole. The cuts between the connected components serve as a similar concept to nodes and its neighbours which are used in calculating the prior of the MRF.

Third, for each connected component, we update the labels for all the nodes at once in a similar fashion to the Algorithm 1. We describe the algorithm in the following:

Algorithm 2 Updating Labels: Swendsen-Wang

```
for $p$ = 1 to $P$ do
    for every edge (u,v) in $E(G_p)$ do
        Draw an edge with if $Uniform(0, \exp(\omega \cdot \beta_{x_u} \mathbb{I}(x_u = x_v))) > 1$
    end
    Identify all connected components $C$ in $G_p$
    Identify cuts $C'$ for all connected components $C$ in $G_p$
    for $i$ = 1 to $|C|$ do
        for $k$ = 1 to $K$ do
            $x_{p,n \in N(C_i)} \leftarrow k$
            $H_{p,i,k} \leftarrow \sum_{(u,v) \in E(C'_i)} \frac{1}{2} \cdot \omega \cdot \beta_{x_{p,u}} \mathbb{I}(x_{p,u} \neq x_{p,v})$
            $L_{p,i,k} \leftarrow$ Log Marginal Likelihood
            $Q_{p,i,k} \leftarrow H_{p,i,k} + L_{p,i,k}$
        end
        Sample a new label $x_{p,i}$ with the probability function $\exp(\frac{Q_{p,i,k}}{\log \sum_{k=1}^{K} \exp(Q_{p,i,k})})$
        Update the labels for $x_{p,n \in N(C_i)}$ where $N(C_i)$ is the cells in connected component $i$
    end
end
```

2.4.3 Updating Beta: Double Metropolis Hastings

The estimation of the cell-cell interactions —the β term— is a major challenge in the model due to the need to draw samples from a doubly intractable distribution. Methods from the Markov Chain Monte Carlo (MCMC) family draw samples from a distribution of:

$$P(X = x|\beta) = f(x,\beta)/\int f(x',\beta)dx'$$

in which the partition function is intractable. Suppose we sample the β parameter, we would need to evaluate the full conditional probability:

$$\begin{aligned} P(\beta \mid X) &= \frac{P(X \mid \beta)P(\beta)}{P(X)} \\ P(X) &= \int P(X \mid \beta)P(\beta)d\beta \end{aligned}$$

where $p(\beta)$ is the prior.

The condition in inference where the partition function includes the parameter of interest is referred to as doubly intractable distributions. This is a common problem in the inference of spatial models. The MH algorithm cannot be employed to simulate from this distribution because the Metropolis-Hastings acceptance ratio involves the unknown ratio of $\frac{C_{\beta_0}}{C_{\beta'}}$. We provide the derivation in the following:

$$\begin{aligned} \alpha(\beta_0,\beta') &= \frac{P(X \mid \beta')P(\beta')}{P(X \mid \beta_0)P(\beta_0)} \\ &= \frac{C_{\beta_0} f(X,\beta')P(\beta')}{C_{\beta'} f(X,\beta_0)P(\beta_0)} \end{aligned}$$

where

$$C_\beta = \int f(x',\beta)dx'.$$

In order to address this problem, we use a method called the Double Metropolis Hastings (DMH) algorithm [30]. The DMH algorithm enables inference on doubly intractable models by introducing an auxiliary variable that approximately has a density function with an intractable normalizing constant as shown above. In order to circumvent the evaluation of the partition function, Double Metropolis Hastings (DMH) introduces an auxiliary variable Y on the same space as X, also with an identical distribution family as X. However, as we cannot draw from the distribution exactly, Y is drawn approximately through an application of Metropolis Hastings (MH).

The DMH algorithm removes the need for exact sampling and allows sampling from distributions with intractable partition functions. In summary, a new β_p is proposed from a simplex distribution S^k starting from β_t, where it is defined as:

$$S^k = \{(\beta_{c,0},\ldots,\beta_{c,k}) \in \mathbb{R}^k \colon \beta_{c,1} > 0,\ldots,\beta_{c,k} > 0, \sum_{i=0}^{K-1} \beta_{c,i} = 1\}$$

where K is the total number of cell type cluster labels.

In general, given the current β value, say β_0, the algorithm proceeds in two steps: 1) Sample β' from the prior $P(\beta)$, 2) Generate $y \sim P^{(m)}(\cdot|x,\beta')$, and, 3) Accept β' with probability

$$\frac{f(y|\beta_0)f(x|\beta')}{f(x|\beta_0)f(y|\beta')},$$

where m denotes the number of MH iterations used to generate y. Assuming the MH procedure for drawing y mixes sufficiently well to draw exactly from its distribution, the partition functions $C_{\beta'}, C_{\beta_0}$ in both the numerator and denominator cancel, yielding a tractable solution as in the acceptance probability above.

The core part of DMH is that it has two types of MH updates: for which one is to draw a realization of the auxiliary variable y and one for acceptance of β. To be precise on the implementation strategies, we perform all operations in log space to prevent underflow. The algorithm for the sampler is described below:

Algorithm 3 Double Metropolis Hastings

```
Sample a new $\beta_p \sim Simplex$ starting from $\beta_t$
$T \leftarrow 0$
$T_y \leftarrow 0$
for $p$ = 1 to P do
    $T \leftarrow T + T_p$.
    Generate an auxiliary variable $y_p \sim P_{\beta_p}(y_p|x)$, in this case $HotPotts(\beta_p)$
    $T_y \leftarrow T_y + T_{y,p}$.
end
$\log f(y|\beta_t) \leftarrow \beta_t \cdot T_y$
$\log f(x|\beta_p) \leftarrow \beta_p \cdot T$
$\log f(x|\beta_t) \leftarrow \beta_t \cdot T$
$\log f(y|\beta_p) \leftarrow \beta_p \cdot T_y$
$\alpha \leftarrow \min(1, \log f(y|\beta_t) + \log f(x|\beta_p) - \log f(x|\beta_t) - \log f(y|\beta_p) + f(\beta_p) - f(\beta_t))$
$u \sim Uniform(0,1)$
if $u < \alpha$ then
    $\beta_{t+1} \leftarrow \beta_p$
end
else
    $\beta_{t+1} \leftarrow \beta_t$
end
```

2.4.4 Gibbs Sampling and Cluster Estimation

With the procedure for updating X and β described in the previous subsections, we now want to perform a joint inference to sample from the multivariate distribution. We use another MCMC algorithm, Gibbs sampling, as it is an effective sampler when it is possible to simulate from the conditional distributions. In this algorithm, each variable is updated sequentially for a total of T iterations until convergence.

We define the algorithm below using the algorithms previously mentioned:

Algorithm 4 Gibbs Sampling for Joint Inference

```
for $k$ = $t$ to $T$ do
    Update interaction terms $\beta$ with Algorithm 3
    Optionally update cell labels $X$ with Algorithm 2
    Update cell labels $X$ with Algorithm 1
end
```

After finishing all the iterations of the samplers, the output is a trace of all of Markov Chain Monte Carlo samples of both the cell labels X for all patients' cells and the interaction term β for patient subtype $C_p = c$.

Given the traces for each variable of interest, the next step is to obtain point estimates for each variable.

For X, we use a method that optimizes a criteria called Maximization of Posterior Expected Adjusted Rand (MPEAR) [22] that returns point estimates given a matrix of MCMC samples that has a burn-in portion removed. This method starts by constructing a distance matrix using hamming distance and applying hierarchical clustering. Using the resulting clustering, we maximize MPEAR to get a consensus labeling of cell labels X.

For β, we use a much simpler approach by taking the mean of the same-same affinity values of all clusters in β across T iterations excluding the burnin phase. Using the values, an interaction matrix I is made by having each cell being:

$$I_{k1,k2} = \frac{1}{2}\left(\beta_{k1,[k1 \neq k2]} + \beta_{k2,[k1 \neq k2]}\right)$$

In this case, each cell of the interaction matrix will imply the affinity between cell types and provides a measure of how likely are cell types to be spatial proximal and interact in the TME of a patient with a certain patient subtype.

A common problem observed in the results from clustering is label switching. Label switching means that cell labels can switch between cell cluster configurations in each iteration and influence the estimation of the interaction term, β.

We propose two methods to alleviate this problem. The first being a 2-stage procedure in which we run the full joint inference and then run only inference on β. The first round of inference can result in a good estimate of cell labels X, however the consistent label switching can lead to a poor estimate β. The second solution being to include a prior matrix and/or running a semi-supervised experiment. Prior matrices give clusters a prior belief of what the expression value should be like. The semi-supervised approach anchors cells with known specific cluster numbers and allows for other unlabeled cells to group to the right cluster.

2.4.5 MCMC Convergence

A common problem in MCMC algorithms is the method to determine the number of iterations T required to reach convergence. In addition, it is difficult to determine whether MCMC algorithms have mixed well at where we claim it to converged as we do not know if all possible modes have been visited by the sampler. There exists diagnostics and heuristics that give confidence that a sampler has converged [11, 13, 18]. A notable method is to assess the trace plots.

As a preliminary way to investigate convergence, we created a function to plot values from each iteration in a histogram as well as plot the trace of the values. We check if the trace samples to the peak of the histogram.

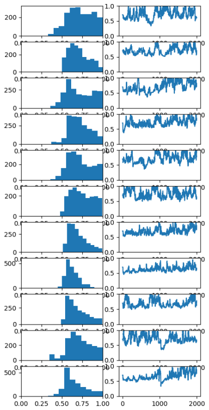


Figure 2.6: Example of investigating convergence of beta.

For real-valued parameters, we see the samplers mix sufficiently well around its modes. For discrete parameters test functions are required. In our case of latent cluster labels, a test function is to map labels to colours, and plot colours over time. A well-mixing sampler would exhibit changes in colour as shown below.

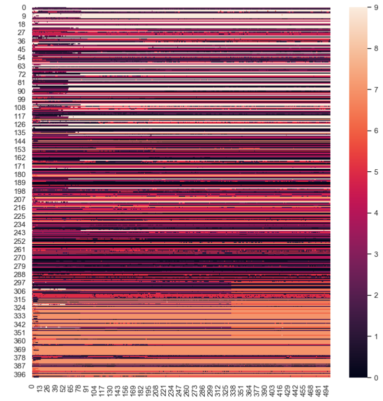


Figure 2.7: Example of investigating convergence of cell label clusters.

An additional trace plot summarizing the mixing of models is to map the model parameters to its likelihood value. As we have an intractable partition function, we plot the unnormalized likelihood against iterations. A plateau of values should signify potential convergence.

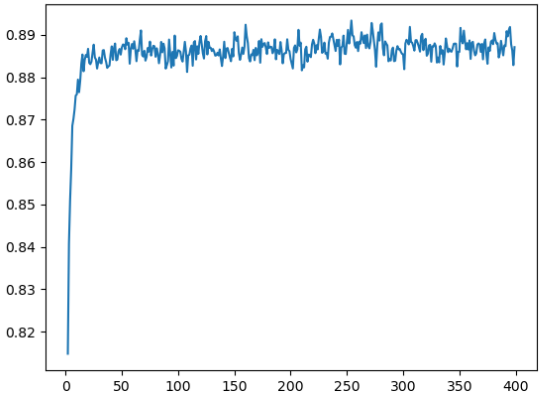


Figure 2.8: Example of likelihood trace plot.

Chapter 3

Experiments

In this chapter, we discuss the results from experiments fitting our model to three different datasets: synthetic, semi-real, and real-world. First, we show a proof of concept by fitting the synthetic dataset that is forward simulated from our model and also simulated through the package MixSim. Next, we demonstrate the utility of our model by performing inference on a semi-real dataset. This semi-real dataset is composed of disaggregated mass cytometry data with simulated spatial coordinates and topology. Lastly, we apply our model to the METABRIC imaging mass cytometry data of breast cancer for biological analysis. The experiments of each dataset have varying parameters to objectively demonstrate the performance of the method. Details of the implementation of each experiment as well as the biological analysis of each output figure is explained in this chapter.

3.1 Workflow

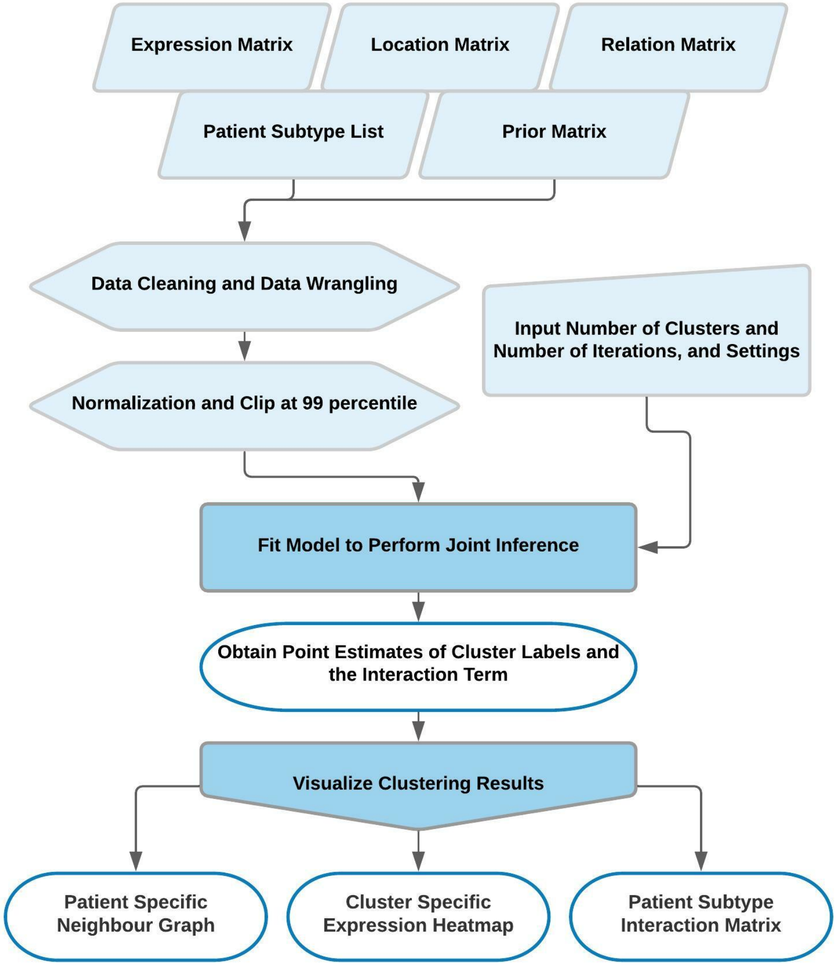

Expression Matrix
Location Matrix
Relation Matrix
Patient Subtype List
Prior Matrix
Data Cleaning and Data Wrangling
Input Number of Clusters and Number of Iterations, and Settings
Normalization and Clip at 99 percentile
Fit Model to Perform Joint Inference
Obtain Point Estimates of Cluster Labels and the Interaction Term
Visualize Clustering Results
Patient Specific Neighbour Graph
Cluster Specific Expression Heatmap
Patient Subtype Interaction Matrix

Figure 3.1: The workflow for spatially aware clustering analysis using SpatialSort. Elements shaded in light blue are inputs and other steps performed prior to running SpatialSort. Elements shaded in dark blue are steps in SpatialSort. Elements without shade are user outputs.

All experiments performed in this chapter will be using the designed workflow shown in Figure 3.1 to obtain outputs.

The main numerical output from SpatialSort are cluster labels and an interaction matrix. Cluster labels are labels for cells to indicate being a certain cell type. The downstream analysis from SpatialSort will be the identification of cell types by looking at overall cellular expression. The interaction matrix is outputted along to observe interactions between clusters.

In our analysis, there will be a repetition of analysis using cluster-specific expression matrices, patient-specific neighbour graphs, and subtype-specific interaction matrices, in which these three graphs are the main visualization outputs of SpatialSort.

Also, an important factor that affects the point estimates of the cluster labels and interaction affinities is the number of clusters that the user inputs to SpatialSort. The choice of this input parameter may be based on the user's prior knowledge or can selected by attempting different number of clusters and visualizing the heat maps and graphs to determine an appropriate cluster count. It is also important to acknowledge that the number of iterations will affect whether the output has reached convergence. There are diagnostic functions in the SpatialSort documentation that can assist in interpreting a good parameter.

3.2 Synthetic Experiments

3.2.1 Forward Simulating Synthetic Spatial Data

Our experiments start with generating synthetic datasets that are forward simulated from our proposed model. This step evaluates the ability of the inference engine to recover the known true parameters. The simulation of the data starts with the sampling of the interaction matrix for the 2k beta model, patient classes, the mean and variance of the expression matrix. The hyperparameters chosen for this specific experiment are detailed in Algorithm 5. For each patient, we simulate a topology by subsetting a part of a real breast cancer imaging mass cytometry topology using breadth first search starting at a random location. The labels of the cells in the topology are then forward simulated from the HotPotts model using the sampled beta matrix. Expression values for each label are sampled from a Gaussian distribution with parameters from the previously sampled means and variances.

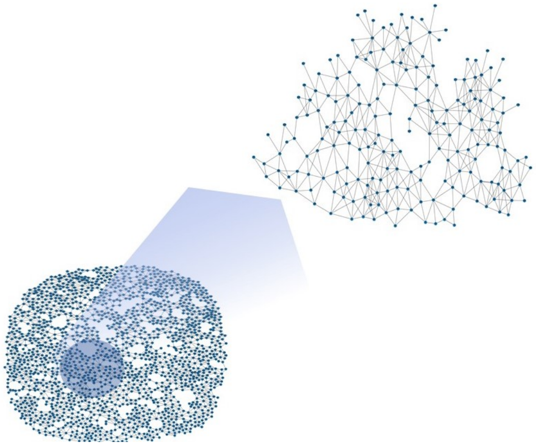


Figure 3.2: Topology simulation by subsetting a part of a real topology from spatial profiling technologies.

Algorithm 5 Synthetic Data Simulation

```
Sample $\beta \sim Beta(1,1)$ for $K$ clusters
Sample $C \sim Categorical(1/K)$ for $P$ patients
Sample $\mu \sim Gaussian(0,1)$ for $M$ markers of $K$ clusters
Sample $\sigma^2 \sim Gamma(1,1)$ for $M$ markers of $K$ clusters
for p=1 to P do
    Simulate a topology by Breadth First Search through a real topology structure for patient $p$
    Sample $x_p \sim HotPotts(\beta_{C_p})$
    Sample $y_p \sim Gaussian(\mu_{x_p}, \sigma^2_{x_p})$
end
```

Using Algorithm 5, we simulated expression data matrices with dimensions 500 cells by 20 protein markers. Each matrix includes a dataset of 10 patients of a single patient class. The beta model used to simulated data follows the 2k model. The number of Gibbs sampling iterations performed for sampling cell labels was set at 5000. The initial number of clusters was set randomly at 12 to perform label sampling, yet it is important to note that the resulting number of clusters will not always match 12 as it depends on the sampled beta value. To ensure that the results are not of a single instance, a total of 9 seeds were used to generate replicates of the data.

A variant of the algorithm was also used to generate another synthetic dataset with the same parameters described in the previous paragraph. The only change made was to the first line of Algorithm 5 where an extra step was added. The extra step is to swap the same-same interaction affinity and the diff-diff interaction affinity if the latter is greater than the former. This step ensures that cells that are of the same type prefer to be with each other than of another type.

Despite the fact that this variant contradicts with the biological behavior of some types of cell, this is an assumption with the Potts model and not our modified HotPotts model. We use this to check on the behavior of our model in contrast with the Potts model.

For all synthetic and semi-real experiments performed in this chapter, we generate datasets from both two conditions of interest: having random beta affinity values sampled from Beta(1,1) or having beta affinity values with a stronger affinity for same-same interactions. Let us name the former dataset as ”uniform” and the latter as ”biased” for ease of describing the datasets in this section. All of the datasets have replicates from using different seeds.

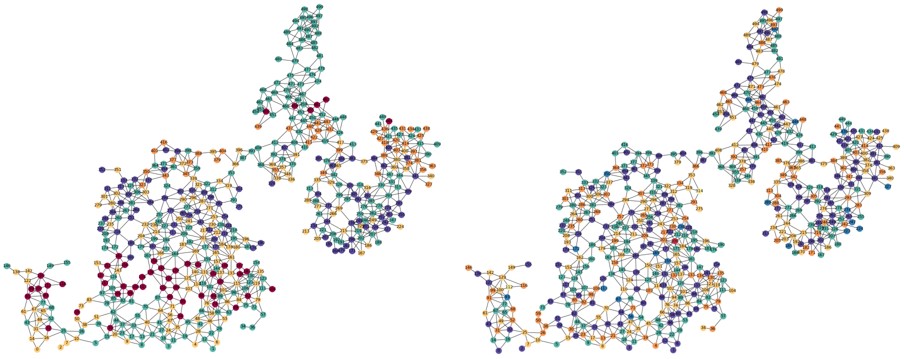


Figure 3.3: Visualization of different labeled topologies using different betas. Left is a labeled configuration from the biased dataset with a higher likelihood of cells of the same type to be near each other. Right is a labeled configuration from the uniform dataset with cells of the same type having a affinity randomly sampled from Beta(1,1).

3.2.2 Forward Simulation Dataset Clustering Performance

We fitted the model to each dataset by running SpatialSort without using the Swendsen-Wang algorithm. The number of iterations were set to 800 and the number of clusters was set to the true number of clusters that was sampled in the dataset. 3 random seeds were picked to run replicates.

Five different models or methods were employed. As for notation: 0p is the Potts model, 1p is our one parameter beta model, 2k is our 2k beta model, GMM is the Gaussian mixture model, and Phenograph is the current popular method to perform clustering.

To evaluate the accuracy of the output cluster labels, we use v-measure [40], an external entropy-based cluster evaluation measure to assess our result. V-measure is the harmonic mean between homogeneity and completeness of each cluster. Homogeneity is a measure of whether all of its clusters contain only data points of a single truth label. Completeness is a measure to evaluate if points with the same truth label are assigned to the same cluster.

The evaluation of significance between the scores of different models employ the Friedman test. When the p-value is less than 0.001, we applied the post-hoc Nemenyi test to all pairs of scores to determine if specific models showed significance [20].

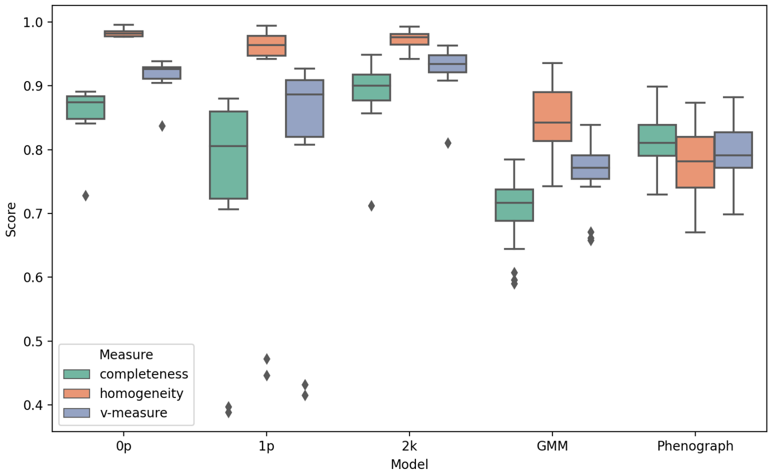


Figure 3.4: Performance of model fitting on forward simulation dataset with betas having stronger same-same interactions.

In Figure 3.4, we show the result from fitting to the biased dataset where betas are biased to have stronger same-same interactions than diff-diff interactions. A table of mean and standard deviations are shown in Table A.1. It can be seen that the 0p, 1p, 2k models have very similar performances, except the 1p had a larger variance in its scores across replicates. For these models, there is a high homogeneity score in each cluster that led the overall v-measure score to be above 0.9. On the other hand, GMM and Phenograph yielded a lower range around 0.8.

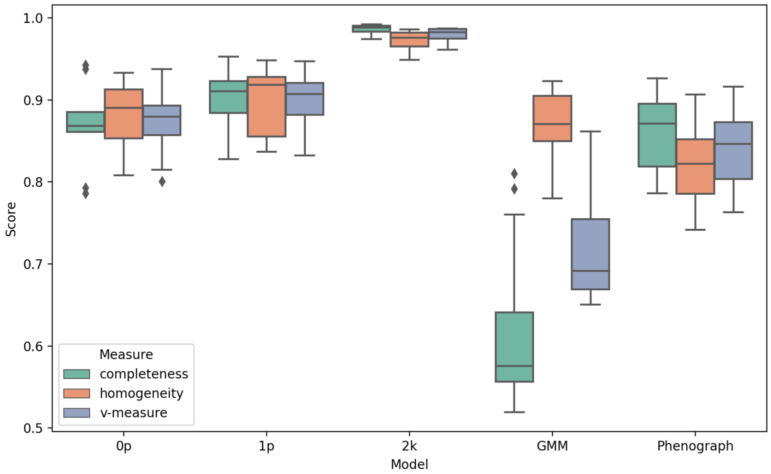


Figure 3.5: Performance of model fitting on forward simulation dataset with betas sampled from Beta(1,1).

In Figure 3.5, we show the result of the other condition in which we fit to the uniform dataset where betas are directly sampled from a Beta(1,1) distribution. A table of mean and standard deviations are shown in Table A.4.

Differing from Figure 3.4, the 2k model here stands out by having a very high performance with a score of 0.95 in all measurement categories. The Potts and the 1p models have similar performances that are lower than 2k. The 2k model has significantly better results from 0k, GMM, and Phenograph with p-values of 0.009, 0.001, 0.009 respectively. Phenograph performed lower than the other 3 models, yet it seemed to have similar performances regardless of beta, as opposed to GMM which dropped a v-measure score between 0.6 and 0.7.

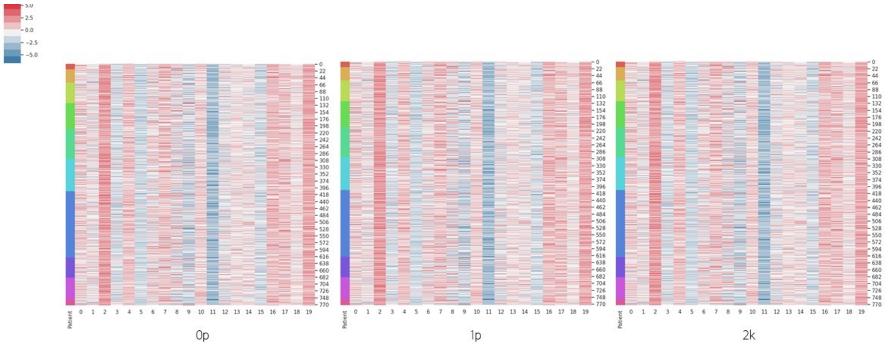


Figure 3.6: Single cluster-specific expression heat map for the forward simulation dataset with strong same-same interactions.

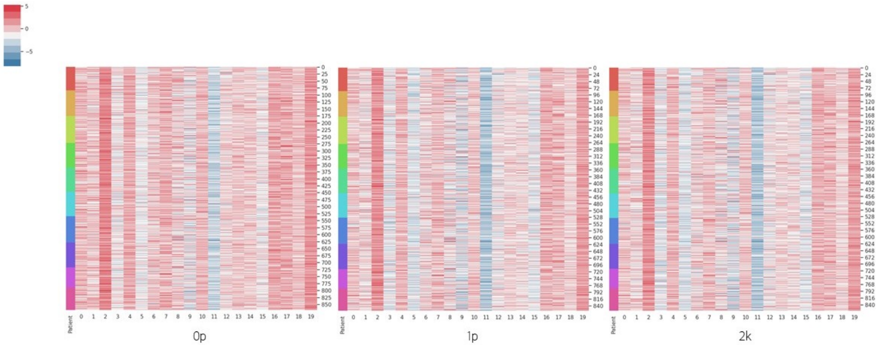


Figure 3.7: Single cluster-specific expression heat map for the forward simulation dataset with betas sampled from Beta(1,1).

We can visualize single cluster-specific expression heat maps through the visualization functions in SpatialSort. In Figure 3.6 and Figure 3.7 we show a facet wrapped result of fitting three models 0p, 1p, 2k to the two different datasets. As there are as many heat maps as cluster, we display only a single instance here.

We can see that all graphs show a consistent pattern of expressions across rows. Very minor differences are present in the expression heat maps across different models. This observation agrees with Figures 3.4 and 3.5 in which the homogeneity score is high for all models.

The important use case of this heat map in theory is the ability to use the expression patterns across columns to determine a cluster identity.

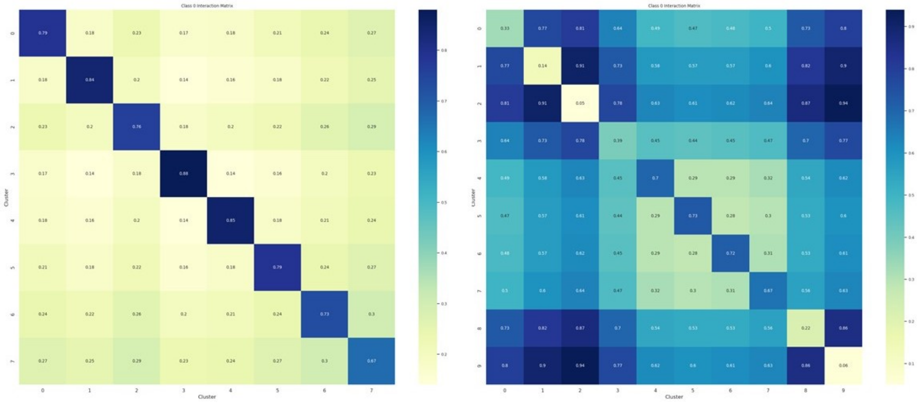


Figure 3.8: Interaction matrix for forward simulation dataset with strong same-same interactions and betas sampled from Beta(1,1).

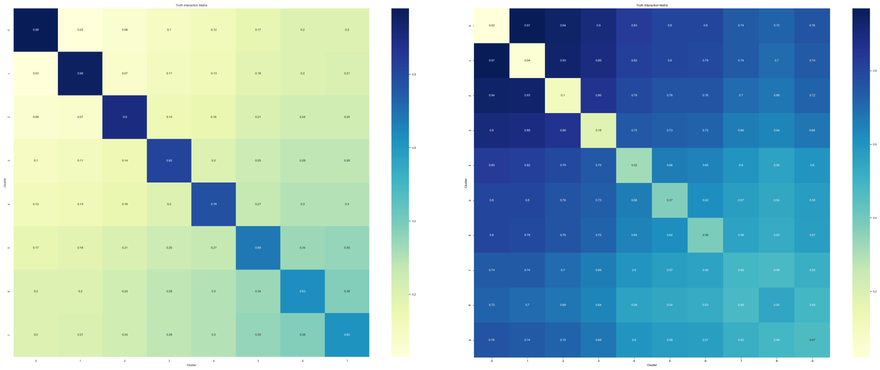


Figure 3.9: The ground truth interaction matrix for forward simulation dataset.

Another major visualization to analyze clusters spatially is through the beta interaction matrices shown in Figure 3.8.

In the left matrix, we can see a strong diagonal which refers to a strong same-same interaction for all the clusters. This matches with how the beta was simulated for that specific dataset.

The right matrix shows an interaction matrix with strong affinities distributed randomly between different clusters. This also follows with how the betas was sampled through a Beta(1,1) uniform distribution.

Table 3.1: Comparision of point estimate interaction terms and ground truth

	0	1	2	3	4	5	6	7	8	9
Biased	0.79	0.84	0.76	0.88	0.85	0.79	0.73	0.67	-	-
Biased Truth	0.98	0.96	0.90	0.82	0.78	0.68	0.63	0.62	-	-
Uniform	0.33	0.14	0.05	0.39	0.70	0.73	0.72	0.67	0.22	0.06
Uniform Truth	0.02	0.04	0.10	0.18	0.32	0.37	0.38	0.49	0.55	0.47

We can assess how closely they recover from the betas by comparing the point estimates and the truth in Figure 3.9. Table 3.1 shows a comparison of our inferred same-same interaction terms with the ground truth values. We can see that within the beta interaction terms, the patterns follow how beta should behave but the accuracy is not high.

Overall, it seems difficult to recover the actual betas with the Double Metropolis Hastings algorithm. However, the betas is still valuable in cell label cluster assignment and can be shown from Figure 3.8 that it may still contain important information regarding the interactions between clusters.

3.2.3 Gaussian Mixture Synthetic Spatial Data

The next step to evaluate the model involves the generation of mixtures of Gaussians to systematically evaluate our clustering algorithm. The MixSim package [34] provides a method of simulating Gaussian mixtures with a defined level of overlap. Input parameters include sample count, channel count, component count, and average overlap.

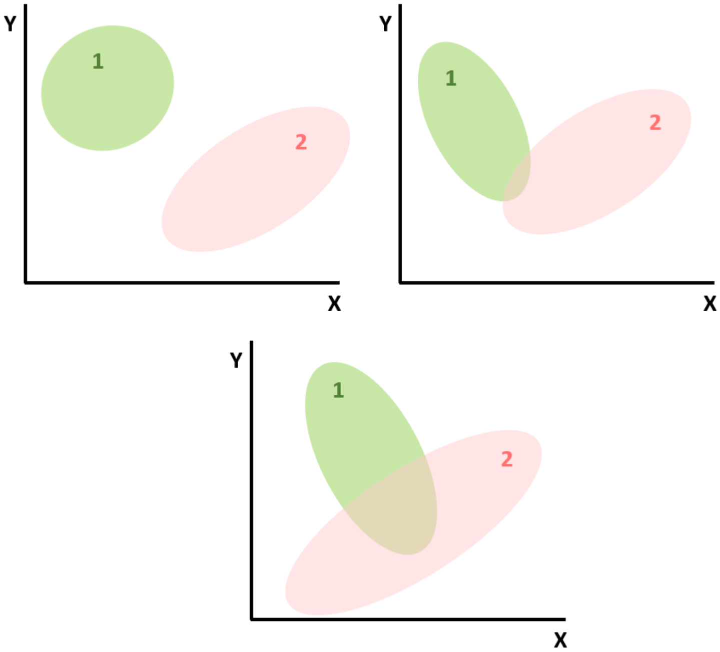


Figure 3.10: A 2D representation of overlap between 2 separate clusters modeled by Gaussian distributions. Top left is the case of no overlap. Top right is with small overlap. Bottom is with a large overlap.

As the overlap increases between the Gaussian distributions, the interaction between the components, in our case, the clusters, will increase as well thus allowing for assessment of clustering algorithms. As our model is based of Gaussian expression matrices, fitting to this dataset will further evaluate the ability of our algorithm. We detail the simulation of the algorithm below in Algorithm 6.

Algorithm 6 MixSim Synthetic Data Simulation

```
Simulate an expression matrix from MixSim by defining average overlap, component count, channels, and cell count
Sample $\beta \sim Beta(1,1)$ for $K$ clusters
Sample $C \sim Categorical(1/K)$ for $P$ patients
for p=1 to P do
    Simulate a topology by Breadth First Search through a real topology structure for patient $p$
    Sample $x_p \sim HotPotts(\beta_{C_p})$
end
for k=1 to K do
    Assign the expression of the cells labeled $k$ in the non-spatial data to the cells labeled $k$ in the simulated topology
end
```

Applying this algorithm, we simulated expression data matrices with dimensions 250 cells by 20 protein markers. The MixSim settings were set to an average overlap of 0.05. A medium level of overlap at 0.05 was chosen by referencing Mixsim [34] which described that the value of average overlap should vary from extreme (0.4) to very low (0.001) for practical use cases. All the rest of the settings were the same as how forward simulation was described in subsection 3.2.1.

3.2.4 MixSim Dataset Clustering Performance

We fitted the model to each dataset by running SpatialSort with the same parameters, models, and evaluation method as described in subsection 3.2.2.

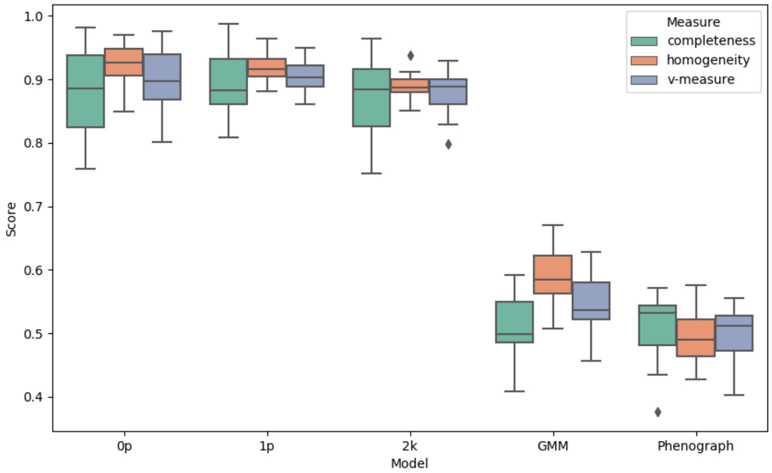


Figure 3.11: Performance of model fitting on MixSim dataset with betas having stronger same-same interactions.

In Figure 3.11, we demonstrate the performance of fitting five different models to the biased MixSim dataset where betas are biased to have stronger same-same interactions. A table of mean and standard deviations are shown in Table A.2.

Similar to Figure 3.4, the 0p, 1p, 2k models have similar scores for all measures. We also observe a high v-measure score above 0.9. However, GMM and Phenograph both dropped in clustering performance to around 0.5 when fitted to a mixture of Gaussian distributions. This suggests that SpatialSort is able to make better cluster assignments probabilistically when the dataset consists of overlapping Gaussian distributions.

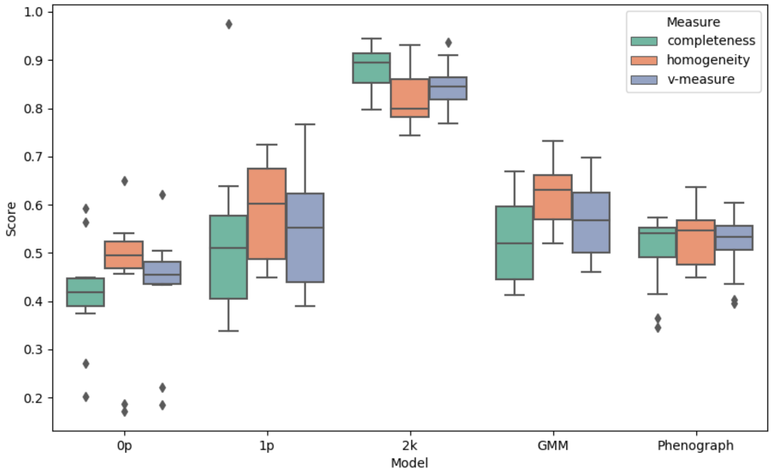


Figure 3.12: Performance of model fitting on MixSim dataset with betas sampled from Beta(1,1)

In Figure 3.12, we show the performance of fitting to the uniform dataset where betas are directly sampled from a Beta(1,1) distribution. A table of mean and standard deviations are shown in Table A.5.

A strong distinction from Figure 3.11 can be seen as the 2k model here stands outs in all measures compared the other models with a score around 0.85. Using the Nemenyi test, the 2k model is significant from the 0p, 1p, and Phenograph models with p-values of 0.001, 0.004, 0.001. Despite the differences in performance of 2k to GMM in the boxplot, the Nemenyi test gave a p-value of 0.255, which is not significant from 2k.

The relative performance of the 2k model is in agreement with the experiment shown in Figure 3.5 where the 2k model seems to be the preferred method when interaction terms can take on any value between 0 and 1.

On the contrary, the other models performed poorly and have similar performances having a v-measure score around 0.5. This demonstrates the 2k model's utility in having a relatively superior ability to classify data points to the correct components.

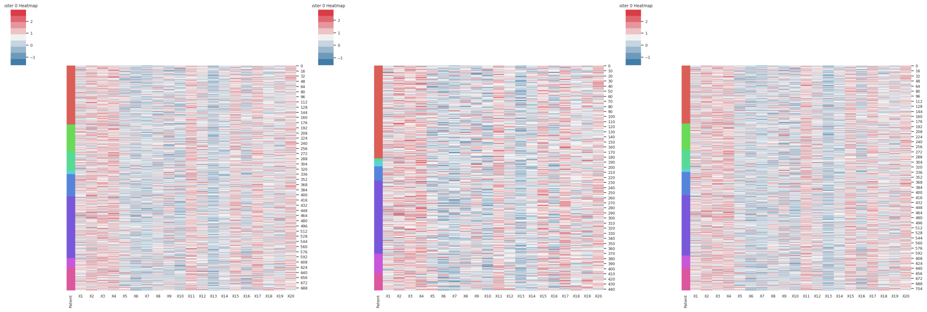


Figure 3.13: Single cluster-specific expression heat map for the MixSim dataset with strong same-same interactions.

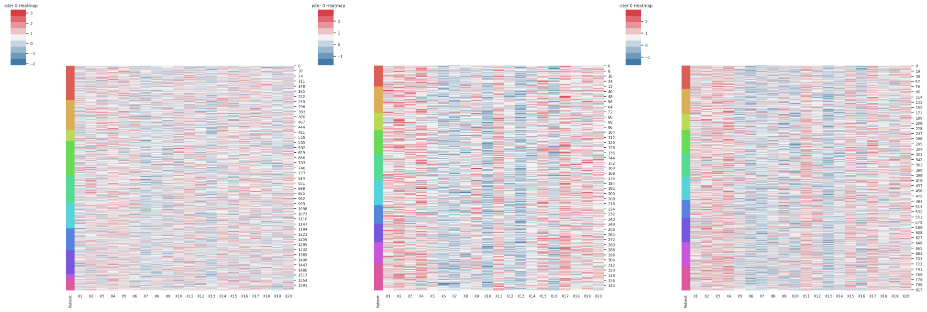


Figure 3.14: Single cluster-specific expression heat map for the MixSim dataset with betas sampled from Beta(1,1).

We visualize the single cluster-specific expression heat maps by outputting Figure 3.13 and Figure 3.14.

As the performance levels for 0p, 1p, 2k are the similar on the biased dataset as shown in Figure 3.11, we see that all models output very consistent clusters.

However, there is quite some differences between the 0p, 1p, and 2k for the uniform dataset. As the performance for the 2k is superior to that of 0p and 1p, we can see a much smoother and more homogeneous cluster for 2k. 0p has a very mixed cluster and also has high variance in columns of expression. 1p is similar to 2k, but 2k is much more smoother in the pattern of expressions.

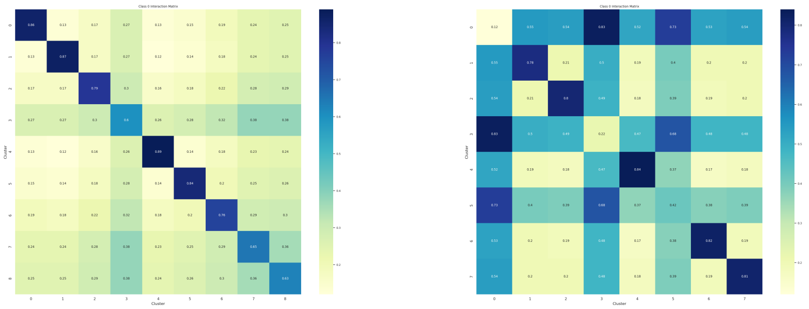


Figure 3.15: Interaction matrix for MixSim dataset with strong same-same interactions and betas sampled from Beta(1,1).

Similar to Figure 3.8, we can observe that the left matrix has a strong diagonal and the right matrix has random strong affinities between different clusters in Figure 3.15. Tables for comparisons between the ground truth and the inferred interaction terms like Table 3.1 will available in the supplementary Github files for viewing, but is not posted here for table size reasons. The key observation is the same: although it remains difficult to recover betas accurately, betas still serve an important function in assigning cells to clusters and a relative measure to affinity between clusters.

3.3 Semi-Real Experiments: Mass Cytometry Dataset

3.3.1 Simulating Semi-Real Spatial Data

The next set of experiments proceeds with generating semi-real datasets. Semi-real datasets are composed of two main elements: (1) expression data from non-spatial expression profiling and (2) simulated topology with sampled labels that are generated by the method as described in subsection 3.2.1.

The source of expression data used here is by Levine et al.. It is a 13 dimensional Cytometry by Time-of-Flight dataset of a single patient. The 13 surface markers are: CD45, CD45RA, CD19, CD11b, CD4, CD8, CD34, CD20, CD33, CD123, CD38, CD90, and CD3.

We used a subset (49%) of the dataset that consists of 81,747 cells of 24 assigned cell type labels from manual gating. The other half of the dataset was not labeled and was not used. We show summarize the dataset in the following z-scored mean marker expression heat map.

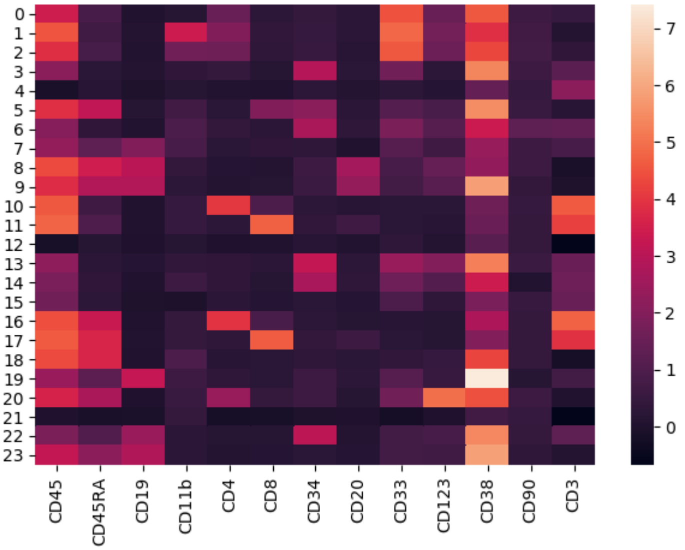


Figure 3.16: Heat map showing the z-scored mean marker expression for CYTOF expression data used for constructing the semi-real dataset.

The data available was arcsin transformed and stated to be the properly normalized by the original publication, therefore no further modification was done to the expression data prior to inference.

The simulation of the data is detailed in Algorithm 7, in which it is similar to Algorithm 6, except for swapping out the mixture of Gaussians from MixSim with mass cytometry data.

Algorithm 7 Semi-Real Data Simulation

```
Sample $\beta \sim Beta(1,1)$ for $K$ clusters
Sample $C \sim Categorical(1/K)$ for $P$ patients
for $p$=1 to $P$ do
    Simulate a topology by Breadth First Search through a real topology structure for patient $p$
    Sample $x_p \sim HotPotts(\beta_{C_p})$
end
for $k$=1 to $K$ do
    Assign mass cytometry expression of the cells labeled $k$ in the non-spatial data to the cells labeled $k$ in the simulated topology
end
```

Through the use of Algorithm 7, we generated expression data matrices with dimensions 500 cells by 13 protein markers. The rest of the settings were the same as how forward simulation was described in subsection 3.2.1.

For inference, we constructed a prior expression matrix by searching for the markers in the public human protein databases [47] as well as inferring through Figure 3.16.

3.3.2 Semi-Real Dataset Clustering Results

We applied the same parameters, models, and evaluation methods described in subsection 3.2.2 to perfrom inference on the dataset by running SpatialSort.

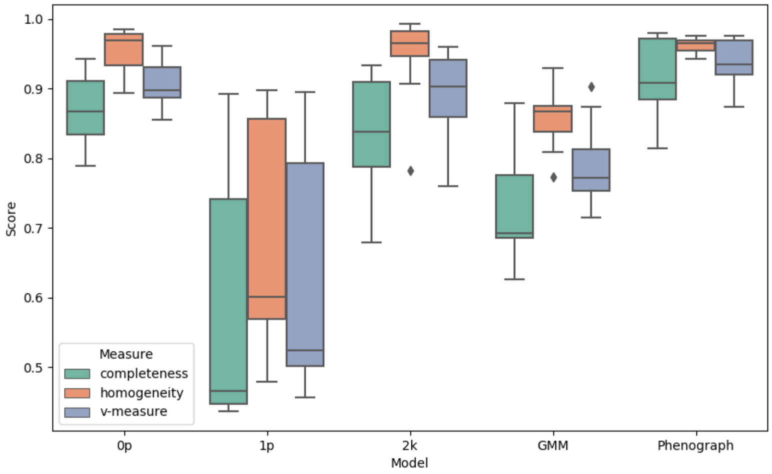


Figure 3.17: Performance of model fitting on the semi-real dataset with betas having stronger same-same interactions.

In Figure 3.17, we show the result from fitting to the biased dataset where betas are biased to have stronger same-same interactions than diff-diff interactions. The inference is done without a prior expression matrix. A table of mean and standard deviations are shown in Table A.3 and also A.6 for the uniform dataset.

Contrary to the synthetic biased dataset experiments, we can observe the 2k model having similar performance to 0p and Phenograph, with a high v-measure at about 0.95. GMM can be seen to be performing around 0.8 and 1p having relatively poor performance at around 0.55. The Potts model performs well when the interaction term is biased serving as a good smoothing prior. The 2k model is competitive to Phenograph and yields similarly high performance for the biased dataset.

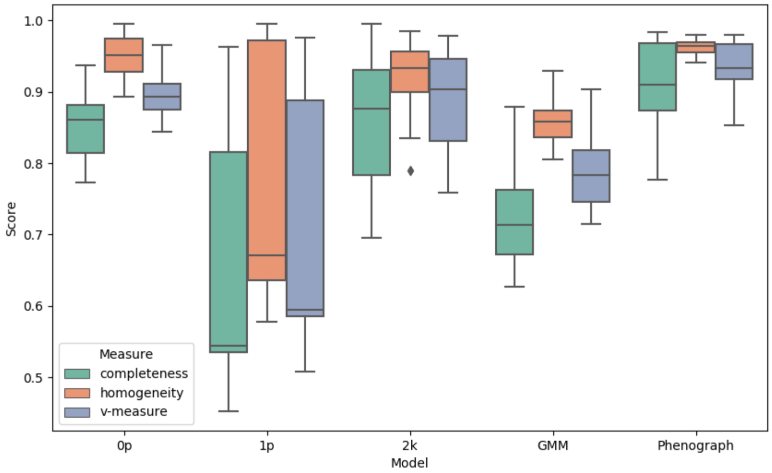


Figure 3.18: Performance of model fitting on the semi-real dataset with a prior expression matrix, and with betas having stronger same-same interactions.

Figure 3.18 shows a result similar to Figure 3.17 in which it differs by introducing an additional prior expression matrix in the inference.

Observing both figures, we do not observe a strong difference between the two. This can potentially indicate that the prior expression matrix does not improve on the datasets with betas having stronger same-same interactions. However, an alternative explanation will be that the performance has already reached a peak at about 0.9 and the prior expression matrix can not raise the performance further.

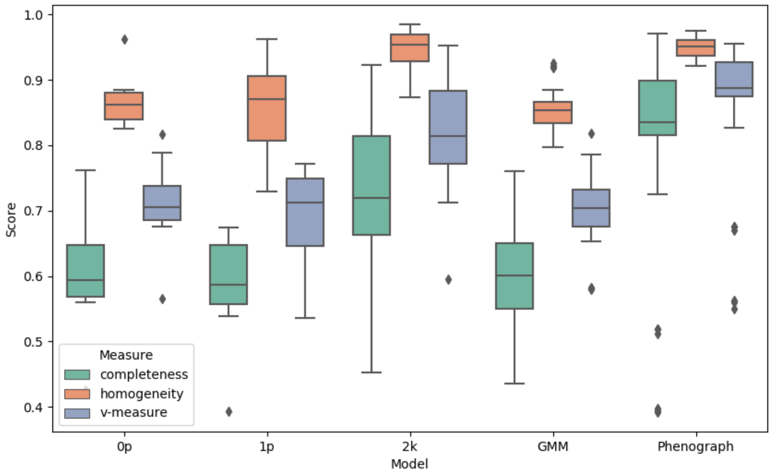


Figure 3.19: Performance of model fitting on semi-real dataset with betas sampled from Beta(1,1)

In the uniform dataset, we can see a contrast from Figure 3.17, where it is difficult to spot a model with a superior level of performance. Using the Nemenyi test, we find the 2k model is significant to 1p and GMM (0.03, 0.01) but not to 0p and Phenograph (0.05, 0.90). From the boxplot, we can see that the median performance of Phenograph to be slightly higher than that of 2k.

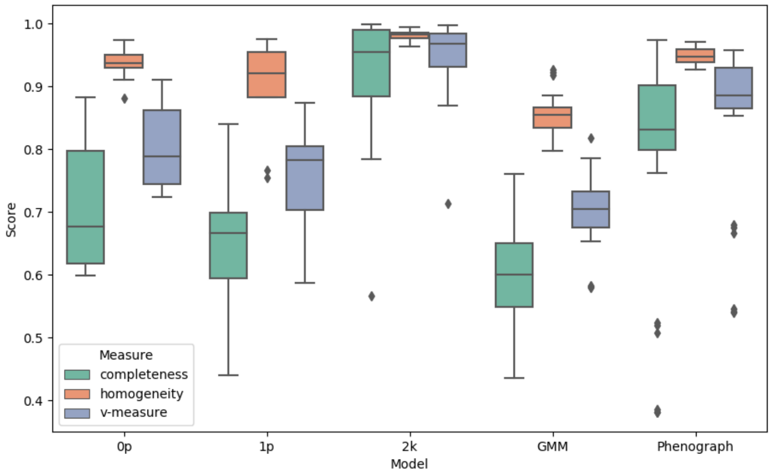


Figure 3.20: Performance of model fitting on semi-real dataset with a prior expression matrix, and with betas sampled from Beta(1,1)

We show the result of adding an prior expression matrix in inference here in Figure 3.20. Contrasting from the difference in performance between the two figures of the biased dataset, Figures 3.20 has some noticeable improvements from Figure 3.20.

The 2k model has an increase in mean v-measure performance from 0.816 to 0.937, while the 0p and 1p both have around a gain of 0.08 in performance.

The 2k model is also found to be significant to 0p, 1p and GMM (0.007, 0.002, 0.001) but not to Phenograph (0.13). The values in parentheses are computed using the Nemenyi test. From the boxplot, we can see that the median performance of Phenograph to be higher than that of 2k. We can observe that including a prior expression matrix brings the performance up by a good margin and is assists cell type assignment in inference.

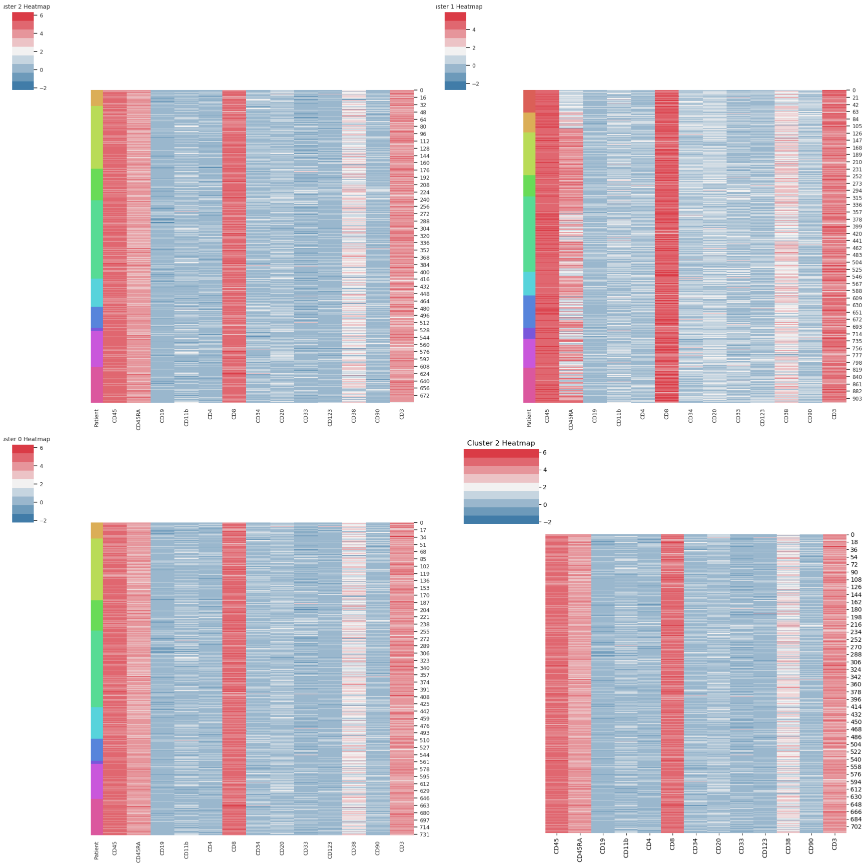


Figure 3.21: Single cluster-specific expression heat map for the semi-real biased dataset fitted by 0p (upper left), 1p (upper right), 2k (lower left), and Phenograph (lower right).

We visualize the single cluster-specific expression heat maps for the biased dataset in Figure 3.21. The cluster shown here is the Naive CD8+ T cell cell type. Inference here is performed without a prior expression matrix.

Echoing the analysis in Figure 3.17, we can see a much more homogeneous and smooth heat map with fitting the 0p, 2k, and Phenograph model. In constrast, we can observe an erroneous introduction of CD45RA- cells into this cluster, which constitutes the loss of the white band in the CD45RA column. This reflects the lower performance of the 1p model.

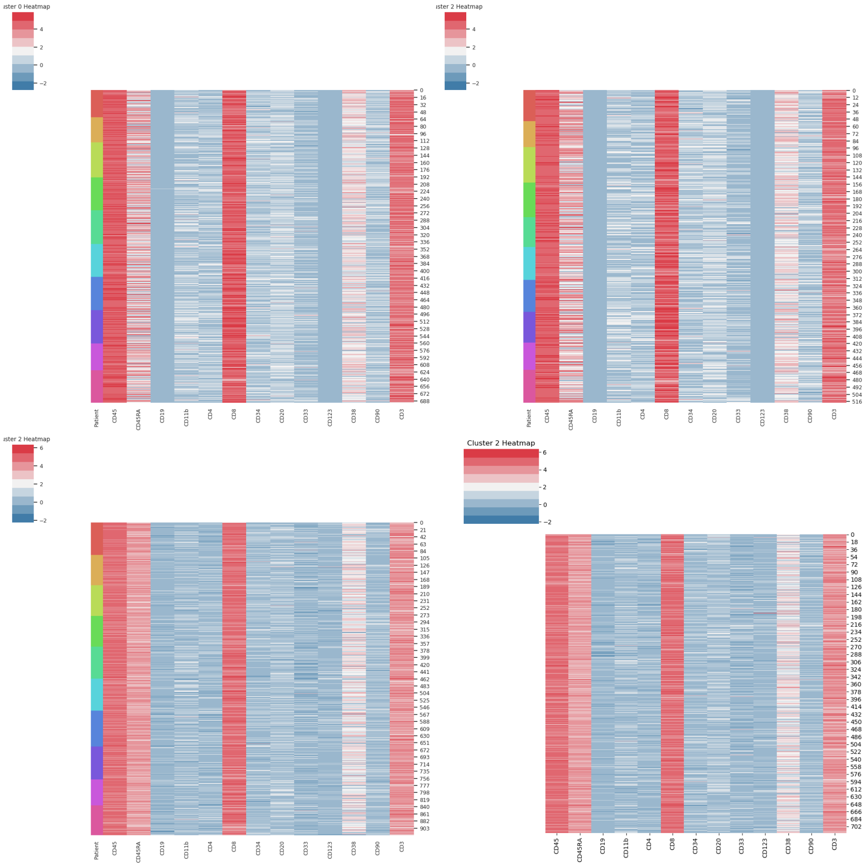


Figure 3.22: Single cluster-specific expression heat map for the semi-real uniform dataset fitted by 0p (upper left), 1p (upper right), 2k (lower left), and Phenograph (lower right).

Similarly, we can visualize the heat maps for the uniform dataset in Figure 3.22. The cluster shown here is also the Naive CD8+ T cell cell type.

It can be seen that all heat maps show consistent patterns of expressions. Differences in smoothness of the patterns of expression are present in the expression heat maps. We can observe that 0p and 1p are less consistent in the CD45RA marker. In contrast, 2k and Phenograph have have a more consistent pattern of CD45RA+. Despite the minor differences, the cluster is homogeneous across the different heat maps in general. This observation agrees with Figure 3.19 where the performance is relatively similar across all models.

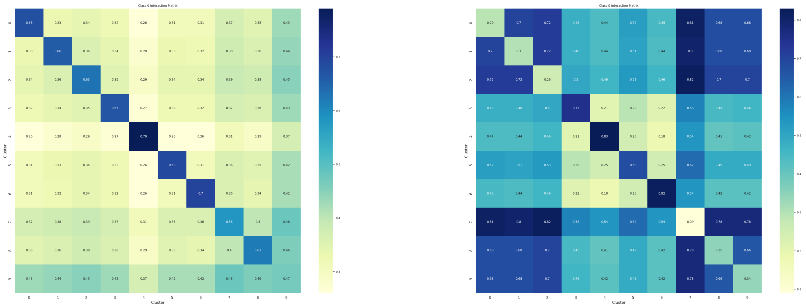


Figure 3.23: Interaction matrix for semi-real dataset with strong same-same interactions and betas sampled from Beta(1,1).

We can observe that the left matrix has a strong diagonal and the right matrix has random strong affinities between different clusters in Figure 3.23 similar to the experiments done using the synthetic datasets. For the strong same-same interaction, we were not able to obtain close enough results as the point estimates fell to 0.5 which is the mean of the uniform distribution. Overall, this could possibly mean that the beta term is acting as a smoothing prior but not much of an interpretable parameter.

3.3.3 Semi-Real Dataset Anchored with Disaggregate CyTOF Data Clustering Results

In Section 3.3.2, we explored the performances of fitting our model to biased and uniform datasets with and without the presence of a prior expression matrix. We have seen increases in performance when the prior matrix is introduced in the uniform dataset but not in the biased dataset.

To further increase the performance of our clustering, we experimented adding in labeled disaggregate CyTOF data in the model to assist in anchoring the unlabelled cells to the correct cluster.

Although this approach can be seen as purely a method to improve the accuracy of clustering, this is also an additional feature of SpatialSort to perform labeling transferring between disaggregate and spatial omic datasets. For the case of CyTOF and IMC, labeled CyTOF datasets are much more ubiquitous in data banks than IMC datasets. When combining public or previously obtained data with spatial data for new experiments, tasks such as relabeling data, cluster interpretation, manual changes in labels are often necessary. Using SpatialSort, we eliminate the need of re-performing clustering and other tasks described above. We include the labeled disaggregate data in the inference, which we here on refer to as anchors, to lead new data to be probabilistically assigned to labels as to how the disaggregate data is labeled.

In the following experiments, we perform inference again on the biased and uniformed datasets anchored with CyTOF data under conditions of with and without the prior expression matrix. It is important to note that the measurement of performance does not include cells that are already labeled. V-measure is performed on all cells that do not have labels.

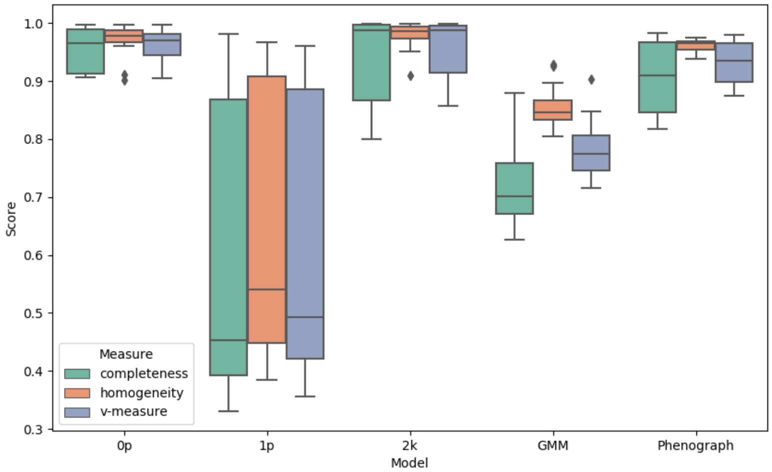


Figure 3.24: Performance of model fitting on the semi-real dataset anchored by CyTOF data, without a prior expression matrix, and with betas having stronger same-same interactions.

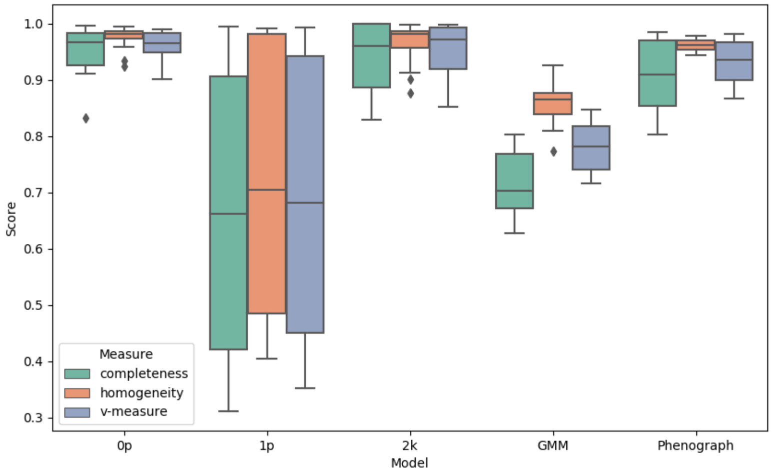


Figure 3.25: Performance of model fitting on the semi-real dataset anchored by CyTOF data, with a prior expression matrix, and with betas having stronger same-same interactions.

In Figures 3.24 and 3.25, we show the performance of the biased dataset anchored with CyTOF data without and with the prior expression matrix.

Contrasting the inference without prior expression matrix in the presence and absence of anchors, we can observe an increase of mean v-measure performance for the 2k model from 0.889 to 0.957. In the case of having a prior expression matrix, the performance increased from 0.884 to 0.950. 0p has also seen an improvement in v-measure when having anchors by an average of 0.06. The performances of 0p and 2k are similar and both are more effective than Phenograph in clustering cells. However, it is to note that 1p did not see improvement in performance.

Although we could not see improvement in performance by adding in a prior expression matrix for the 2k model, we can see that the inclusion of anchors increases the performance by a margin.

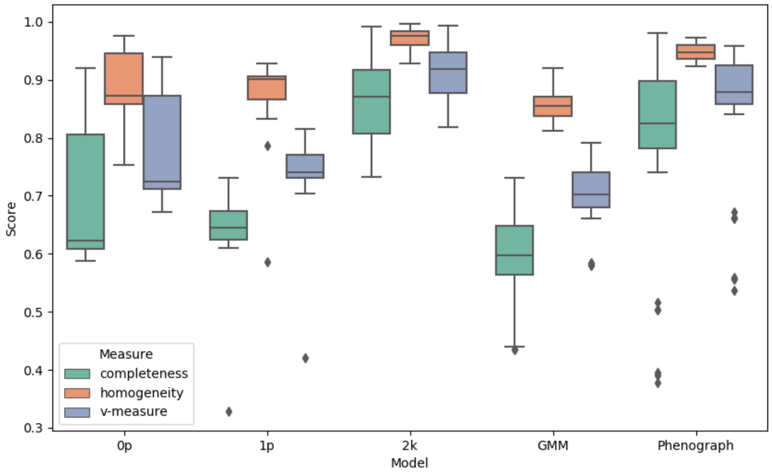


Figure 3.26: Performance of model fitting on semi-real dataset anchored by CyTOF data, without a prior expression matrix, and with betas sampled from Beta(1,1)

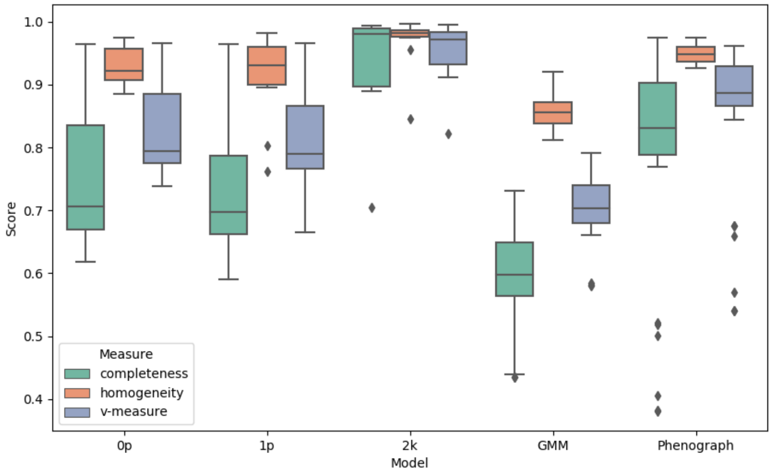


Figure 3.27: Performance of model fitting on semi-real dataset anchored by CyTOF data, with a prior expression matrix, and with betas sampled from Beta(1,1)

Similarly, we show the performance of the uniform dataset anchored with CyTOF data without and with the prior expression matrix here in Figures 3.26 and 3.27.

To evaluate the performances, we again contrasted the results of inference without prior expression matrix in the presence and absence of anchors. Here, we can observe an increase of mean v-measure performance from 0.816 to 0.914 for the 2k model. The performance went from 0.937 to 0.951 when having a prior expression matrix. The 0p and 1p also increased in performance when introducing labeled data into the model. The interpretation of the effect of anchors is the same between both datasets.

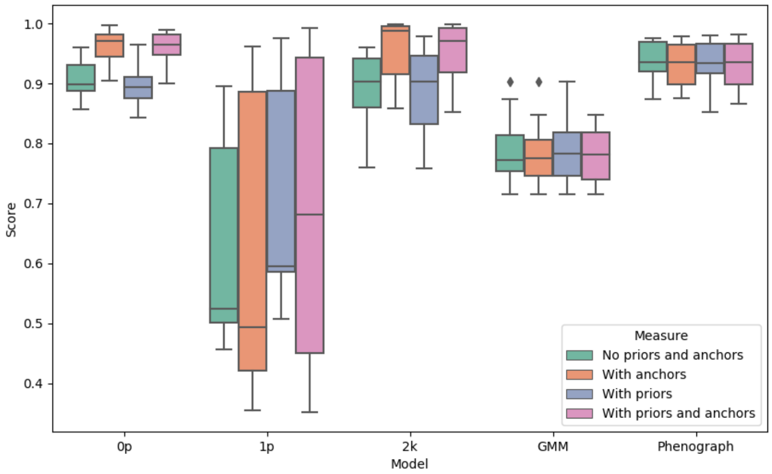


Figure 3.28: Summary of performance of model fitting on semi-real dataset with betas having stronger same-same interactions.

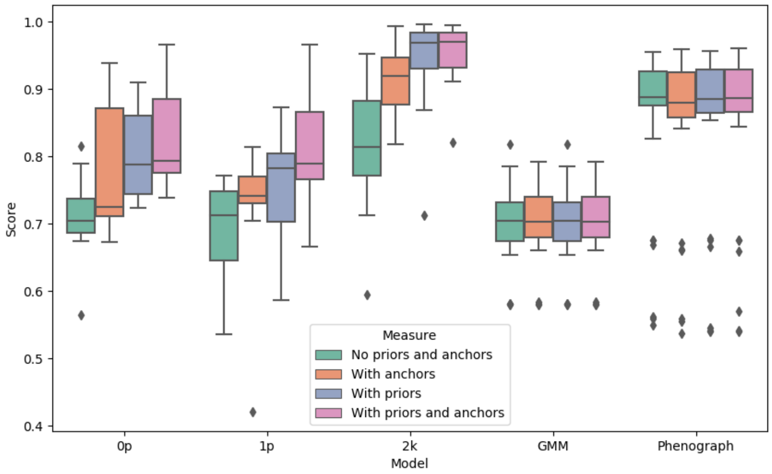


Figure 3.29: Summary of performance of model fitting on semi-real dataset with betas sampled from Beta(1,1).

In Figures 3.28 and 3.29, we summarize our findings under the four different measurement conditions. The boxplots are plotting only the v-measures.

For the biased dataset, we can see that having the prior expression matrix in the inference does not improve performance, yet the introduction of anchors led to an increase in performance. Conversely, the prior expression matrix and the anchors have equal importance in increasing total performance.

In both cases, we can conclude that introducing a prior expression matrix in the inference will lead to a better clustering performance than Phenograph.

3.4 Real-World Application: METABRIC IMC Dataset

For the last set of experiments, we use SpatialSort to cluster cells and evaluate the cluster interactions in a real-world Imaging Mass Cytometry dataset.

3.4.1 Introduction to the METABRIC IMC Dataset

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project is a joint research initiative in the targeted sequencing of breast cancer samples funded by Canada and the UK. Data in the form of expression profiles, CNV profiles, SNP genotypes, etc. have been publically released for reproducible research [16].

In our experiments, we use the IMC dataset from Ali et al. (2020) consisting of 483 spatially profiled breast tumour samples from 448 patients. 37 protein markers were profiled for 479,845 cells from the patient cohort.

Spatial coordinates of single cells and the neighbour relations between cells were provided and were pre-processed by the CellProfiler pipeline [1].

Cell type labels given were inferred by a combination of self-organizing maps, Phenograph, as well as manual inspection of location and morphology according to the publication.

PAM50 subtype labels were obtained through British Columbia Cancer Research Centre as they were not publicly accessible. PAM50 subtypes are molecular intrinsic subtypes with specific clinical properties that are determined by a signature of 50 genes [37]. Breast cancer samples are classified into one of: Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like [29, 39]. A subset of 386 patients that have identifiable PAM50 subtype class labels were used as our input.

3.4.2 Pre-processing IMC Data

A pre-processing pipeline for this dataset was built to clean, wrangle, and normalize the data. We cleaned the data as there were cells without expression data, spatial coordinates, or do not share mutual neighbour relations with other cells. Indexing of the relations were inconsistent and were fixed to 0-indexing.

We removed patient data that are composed of less than 100 cells due to low cell count compared to the average of over 1000 cells per patient.

Single cell expression data were quantile normalized to remove batch effects, arcsinh transformed with a cofactor of 0.8 and z-score transformed to normalize the data to a Gaussian-like distribution. Expression data that exceeded the 99th percentile were clipped to avoid outliers.

Following the original publication of the data, the authors suggested that clustering should use markers that have a good signal-to-noise profile. We subset the 37 markers to the following 22 markers: CK8/CK18, CK19, CK5, CD68, CD3, CD20, ER, PR, CD45, GATA3, CK7, Ki67, SMA, HER2, pan-CK, EGFR, TP53, β-catenin, vWF/CD31, CAIX, Slug and vimentin.

A prior expression matrix was manually created by searching for the protein markers in the public human protein literature and databases [47] as well as inferring through visualizing the z-score mean expression profile.

After viewing the overall expression heat map, we chose the following 20 cell types to have prior expressions: T cells, B cells, Fibroblasts, Fibroblasts Slug+, Fibroblasts CD68+, Myofibroblasts, Myofibroblasts Slug+, Macrophages, Macrophages Slug+, Endothelial, Myoepithelial, Hypoxia, HER2+, HRlow CKlow, CK5+, HR+ CK7-, Ki67+, HR+ CK7- Slug+, HR- CK7+, HR- CK7-. We show the z-score mean expression heat map in Figure 3.30 after manually merging each cluster together.

It is also important to note that some cells do not have clear expression markers present in the data or are naturally diverse in their surface markers, such as endothelial cells.

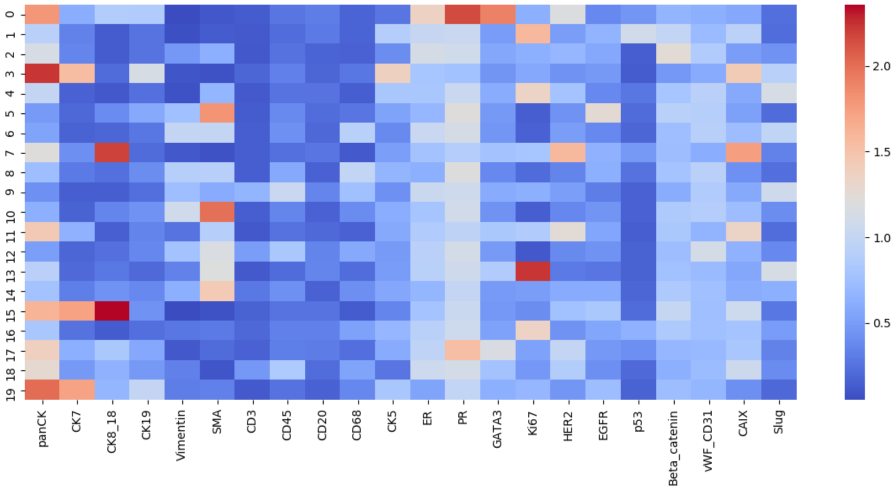


Figure 3.30: Heat map showing the z-scored mean marker expression for the METABRIC IMC expression data.

3.4.3 METABRIC IMC Dataset Clustering Settings

We ran SpatialSort on a subset portion of the whole 386 patient dataset. Inference was performed separately for patients of same PAM50 subtypes. 15 patients from each subtype were selected to be in the dataset. Subsets of the data were used to show the ability of SpatialSort to classify cells correctly when data is limited, as well as to speed up the generation of replicates for each separate experiment.

We fitted to each patient subtype expression matrix using the optional Swendsen-Wang algorithm. The number of iterations was set to 1000 and the number of clusters was set to 20. Although the publication indicated that there seems to be a total 55 total clusters and 2 erroneous clusters, we follow the observation of 20 clear cellular populations when constructing the prior expression matrix. 3 random seeds were picked to run replicates.

Considering the results from fitting models to the forward simulation dataset, the MixSim dataset, and the semi-real CYTOF dataset, we chose to apply the higher performed 2k model to perform spatially aware clustering.

3.4.4 METABRIC IMC Dataset Clustering Results

In this section, we demonstrate two major use cases of the visualization outputs of SpatialSort. We will also analyze the plots and graphs and provide interpretation of the results. As we don't have ground truth data for the cell type labels, we will compare and contrast our results with the labels from the original publication objectively.

We first demonstrate the case of analyzing T cell distributions in patients of Normal-Like breast cancer subtype. From Figures 3.31 to 3.34, we show five different kinds of plots and graphs that can be outputted using functions from SpatialSort.

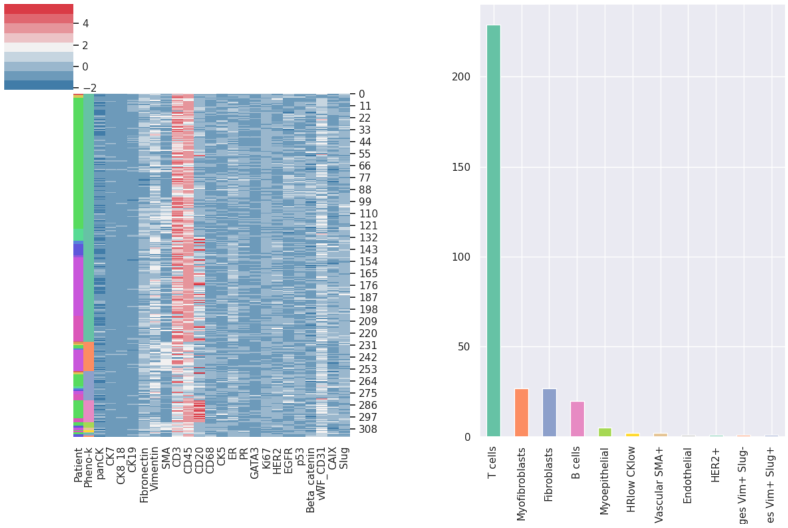


Figure 3.31: T cell cluster specific expression heat map of a 15 Normal-like breast cancer patient subset.

In Figure 3.31, we show a T cell specific expression heat map clustered by the cell type labels from the original publication which is shown on the second column. The output of the expression graph allows the user to observe what the expression levels look like in our clusters. Here, we can see that we have a strong CD45 marker which indicates hematopoietic cells except erythrocytes and platelets, and also a strong CD3 marker which indicates a T cell identity. In between we also have cells that have a strong CD20 marker which is found on B cells. Here, we plot the cell type distribution of this cluster according to the labels from the publication on the right. Note that this bar graph is not an output of SpatialSort's visualizations. We can see that it agrees with the publication except for a few that are labeled as myofibroblasts, fibroblasts, and B cells, etc.

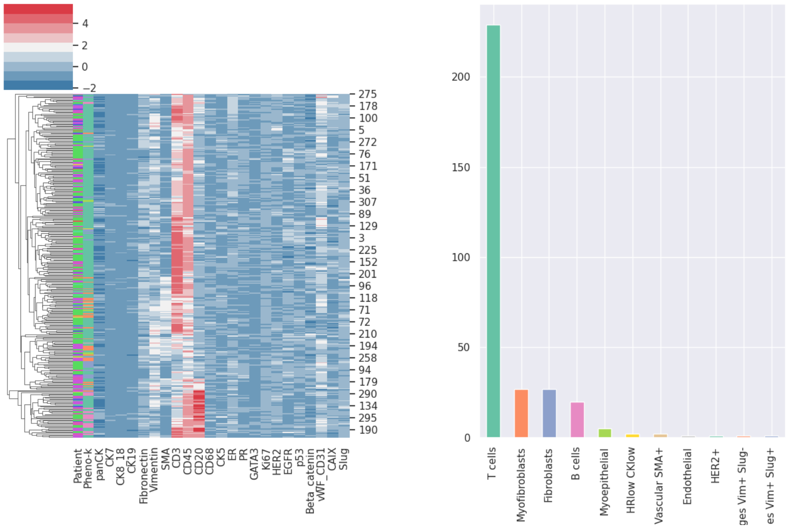


Figure 3.32: T cell cluster specific row-clustered expression heat map of a 15 Normal-like breast cancer patient subset.

To further examine the cluster, we changed to row cluster in Figure 3.32. In the bottom parts of the heat map where CD20 is strong, we see a mixed assignment to both B cells and T cells in the original publication. As CD20 is a marker for B cells in general, the expectation of CD20 is to be in its own cluster. However, due to low cell count for this population, it was assigned into this cluster possibly due to a strong CD45 signal.

Referring back to Figure 3.31, we can know how this cell type cluster is distributed across patients by the first column. We can look at the patient with the highest portion of T cells by matching the colours with Figure 3.33. Here we can see that it is bright green which refers to patient sample MB0128_1_71.

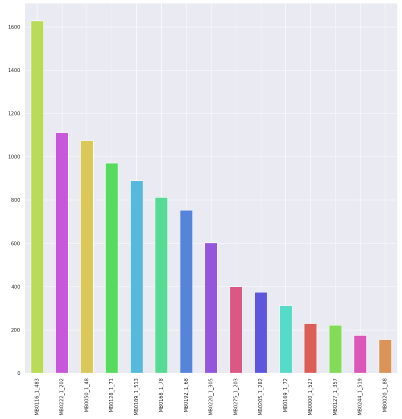


Figure 3.33: Cell count per patient sample bar graph of a 15 breast cancer patient subset.

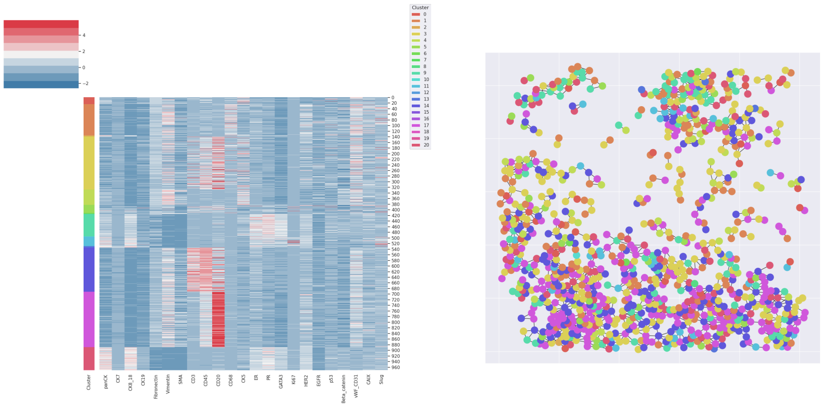


Figure 3.34: Expression heat map of patient MB0128 clustered by patient number and neighbour graph of patient MB0128 colour coded by cluster label.

We show the expression heat map and neighbour graph of patient sample MB0128_1_71 in Figure 3.34. We can investigate the proportion of different cell types in each patient. Here we can see that there seems to be a strong CD20 cluster indicating a B cell population. This allows us to hypothesize that we did not classify the cells previously incorrectly, and our clustering may be more accurate.

3.4.5 Semi-Supervised METABRIC IMC Dataset Clustering Results

To explore the performance of SpatialSort even further, we conducted experiments to explore the behavior of SpatialSort when small amounts of labeled data from patients are introduced into the inference.

This approach is a weak form of supervised learning, because we provide some ground truth labels for testing. However, it still remains a form of unsupervised learning as the majority of the labels are to be inferred. We often refer to this type of task as semi-supervised learning [55].

Semi-supervised learning requires a small number of known labels, in this case we experiment by adding in some labelled data that was originally in the publication to see behavior of SpatialSort.

Note that this approach is a biased form of semi-supervised learning, because we provide none ground truth labels for testing. We are testing whether small amounts of data will support the cell type assignment to move towards a more homogeneous result.

We evaluate the clusters between the previous unsupervised runs to the semi-supervised runs with either 1, 3, or 5 patients (Figure 3.35) having labels from the publication inputted into inference. The evaluation for labeling accuracy changes for this approach since we will have known the cell types for certain patients. We will only compare the inferred cell types to the cell types from the publication.

We picked out Myofibroblasts to demonstrate as an example since it has the greatest number of cells in this subset. We can see that as we add in more data the total number of cells increases and it becomes more homogeneous and similar to the results from the publication. Note that the publication relies on the whole dataset from over 400 patients to form accurate clusters. With SpatialSort, it seems like we could leverage some labeled data from patients to assist in accurate labelling of new patient data.

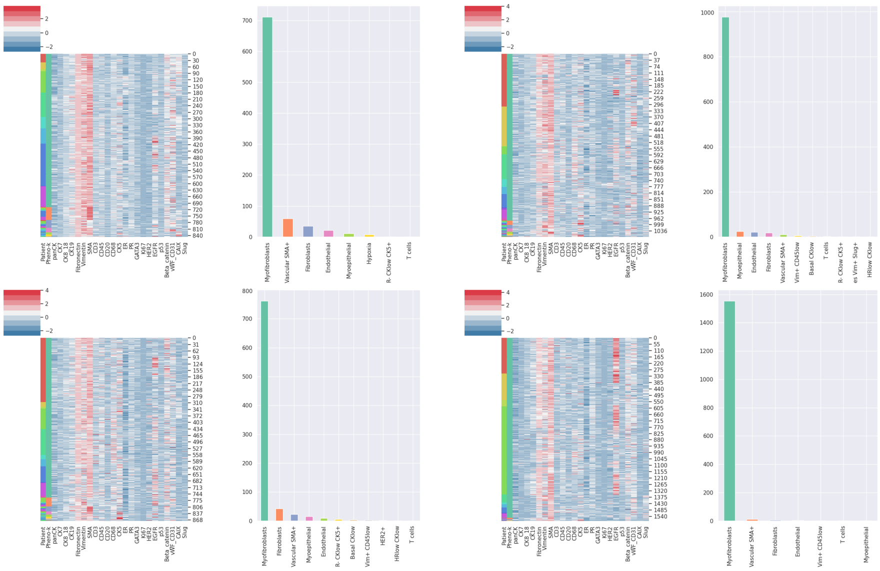


Figure 3.35: Four clusters of Myofibroblasts using different clustering methods are shown. Top left shows the cluster from an unsupervised run of SpatialSort with only prior matrix. Bottom left shows the cluster from a semi-supervised run with labels of a single patient known. Top right and bottom right are also results from semi-supervised runs with 3 and 5 patients respectively.

3.4.6 METABRIC IMC Dataset Interaction Matrix

In Figures 3.36 and 3.37, we show the interaction matrices for 2 of the 5 subtypes: Basal and Her2 under different computational approaches. The left is the original SpatialSort approach, and the right shows the semi-supervised SpatialSort approach with 30% of the patients' labels known.

It can be observed that the interaction matrices between subtypes and with or without semi-supervised approaches are not significantly different. The affinity values for values on the diagonal of all matrices average to around 0.52. We could not find significant differences in affinity between cell clusters according to the plots generated. A hypothesis can be that spatial structure does not affect the affinity values of cell type clusters between different subtypes.

Future work may be to investigate the interaction matrices and modifications that can be made to the beta parameter to explore cell-cell interactions more effectively.

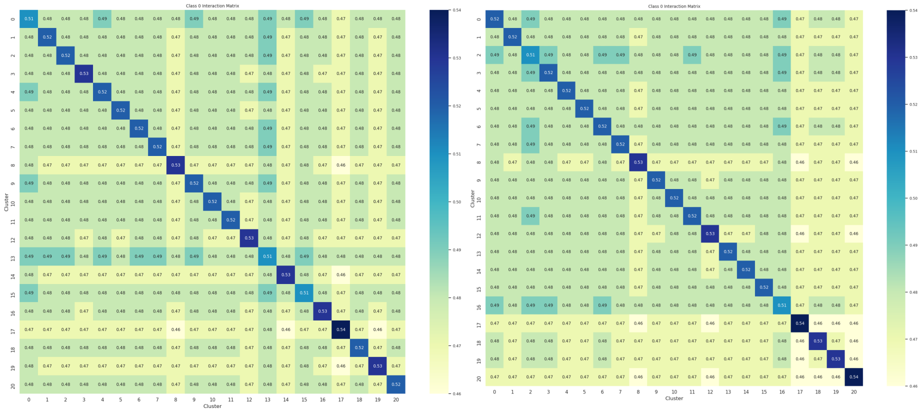


Figure 3.36: Interaction matrices for Basal subtype of the METABRIC dataset. The left is running SpatialSort without a semi-supervised approach. The right is semi-supervised with 30% of patients with known labels.

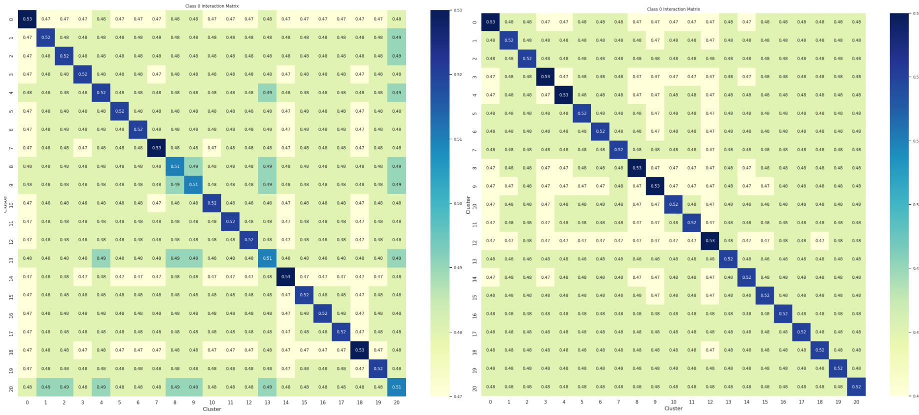


Figure 3.37: Interaction matrices for Her2 subtype of the METABRIC dataset. The left is running SpatialSort without a semi-supervised approach. The right is semi-supervised with 30% of patients with known labels.

3.4.7 Computational Performance

The complexity of SpatialSort largely depends on the the DMH step in the updates of the beta interaction term. Time complexity of SpatialSort in terms of big O is $O(CK^2PNE\varepsilon\delta + 2KPNE\delta)$, where C is the number of patient cells, K is the number of clusters, P is the number of patients, N is the number of cells, E is the number of edges, ε is the number of iterations for DMH, and δ is the number of total iterations.

Using a package called Lineprofiler, we profiled the run time of each major parameter updating function. The ratio between updating X, Swendsen-Wang, and updating beta is 0.29:0.34:0.37.

The total run time for fitting the model to the synthetic and semi-real datasets took about 0.04 minutes per iteration, and the real dataset took about 1.5 minutes per iteration. The settings of all parameters were explained in each of the sections.

Chapter 4

Conclusion

4.1 Summary of Contributions

In this thesis, we present SpatialSort, a statistical framework to perform spatial-aware clustering of single cells and estimation of cell-cell interactions.

As new high-throughput spatial profiling technologies emerge, spatial information can be viewed as a direct measurement of cell-cell interaction. The novelty of SpatialSort lies in the incorporation of spatial structure with expression data to perform probabilistic cell type cluster assignment.

SpatialSort addresses the problems of ignoring spatial context in current methods that rely on solely disaggregate data, and also address the uncertainty of cell type assignment using a Bayesian probabilistic model.

We have also shown in our experiments that spatial context does influence the cell type assignment. In our synthetic experiments, we were able to provide a proof of concept fitting to a forward simulated dataset as well as a mixed Gaussian simulated dataset. In the semi-real experiments, we were able to find that the introduction of a prior expression matrix and/or labeled disaggregate data as anchors, we could yield a high clustering performance using SpatialSort. Our results have shown that we perform better than the current state-of-the-art. In our real experiments, we begin to see differences between methods that incorporate spatial context and those that do not. Despite infrequently having non-homogeneous clusters, we are able to spot out cells that seem to have clustered incorrectly. This can indicate the value in incorporating spatial context which is a variable in cell type assignment.

4.2 Limitations

Various limitations are present in this thesis, in which we will discuss here.

The model proposed is of a Gaussian distribution and is very sensitive to the shape of the data. As we can observe through real data studies, we can observe non-smooth clusters. This implies that SpatialSort relies highly on the pre-processing and normalization.

The proposed inference of beta is done through Double Metropolis Hastings. DMH is a realistic approximation although being an asymptotically inexact algorithm. The results for our interaction matrices as shown in the results do not have good accuracy between inferred and truth.

It is also to note that our method is an unsupervised method that treats each the expressions of each marker as independent and identically distributed. This will mean that we assume that the markers are mutually independent despite possible violation of biological properties of proteins.

In addition, the approach of introducing labeled disaggregate data in the model as cell type anchors rely on accurate patient cell type labels that are considered as gold standard for the inference. As the cell type labels are anchors in the assignment inference, the overall cell type assignments may be biased if the known labels are incorrectly inputted.

4.3 Future Directions

Many directions can be explored beyond on the current proposed method in this thesis.

A possible improvement to inference would to explore parallel tempering or variational inference to improve on the dynamic MCMC sampling performed for inference of parameters. We may not have the most efficient implementation of the samplers, potentially we could look at probabilistic programming languages to refactor the code base.

Important work can be also done to improve beta to become an interpretable parameter for describing cell-cell interactions. An alternative design of the HotPotts model and thus the beta term can possibly improve the accuracy of cell type assignments and possibly bring more information regarding the spatial structure.

Lastly, to explore an automated way of determining the input number of clusters. Deciding the proper cluster count for unsupervised learning tasks is a difficult task and, in our case, relies on prior knowledge and testing. An automated method would improve on obtaining better results.

Bibliography

[1] H. R. Ali, H. W. Jackson, V. R. T. Zanotelli, E. Danenberg, J. R. Fischer, H. Bardwell, E. Provenzano, C. I. G. C. Team, O. M. Rueda, S.-F. Chin, S. Aparicio, C. Caldas, and B. Bodenmiller. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nature Cancer*, 1(1):163–175, 2020. → page 64

[2] M. Allam, S. Cai, and A. F. Coskun. Multiplex bioimaging of single-cell spatial profiles for precision cancer diagnostics and therapeutics. *npj Precision Oncology*, 4:11, 2020. → page 3

[3] V. Almendro, H. J. Kim, Y.-K. Cheng, M. Gönen, S. Itzkovitz, P. Argani, A. van Oudenaarden, S. Sukumar, F. Michor, and K. Polyak. Genetic and phenotypic diversity in breast tumor metastases. *Cancer Research*, 74(5): 1338–48, 2014. → page 2

[4] M. Angelo, S. C. Bendall, R. Finck, M. B. Hale, C. Hitzman, A. D. Borowsky, R. M. Levenson, J. B. Lowe, S. D. Liu, S. Zhao, Y. Natkunam, and G. P. Nolan. Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, 20:436–442, 2014. → page 6

[5] E. Azizi, A. J. Carr, G. Plitas, A. E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R. M.Fromme, P. Dao, P. T. McKenney, R. C. W. K. Kadaveru, L. Mazutis, and A. Y. Rudensky. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell*, 174(5):1293–1308, 2018. → page 2

[6] R. Baghban, L. Roshangar, R. Jahanban-Esfahlan, K. Seidi, A. Ebrahimi-Kalan, M. Jaymand, S. Kolahian, T. Javaheri, and P. Zare. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Communication and Signaling*, 18(1):59, 2020. → page 2

[7] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner. Mass

Cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, 81(16):6813–6822, 2009. → page 5

[8] A. Barbu and S.-C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1239–1253, 2005. → page 24

[9] E. Berglund, J. Maaskola, N. Schultz, S. Friedrich, M. Marklund, J. Bergenstråhle, F. Tarish, A. Tanoglidi, S. Vickovic, L. Larsson, F. Salmén, C. Ogris, K. Wallenborg, J. Lagergren, P. Ståhl, E. Sonnhammer, T. Helleday, and J. Lundeberg. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications*, 9:2419, 2018. → page 7

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738. → page 16

[11] A. Bouchard-Côté, K. Chern, D. Cubranic, S. Hosseini, J. Hume, M. Lepur, Z. Ouyang, and G. Sgarbi. Blang: Bayesian declarative modelling of arbitrary data structures. *arXiv preprint arXiv:1912.10396*, 2019. → page 30

[12] P. C. Boutros, M. Fraser, N. J. Harding, R. de Borja, D. Trudel, E. Lalonde, A. Meng, P. H. Hennings-Yeomans, A. McPherson, V. Y. Sabelnykova, A. Zia, N. S. Fox, J. Livingstone, Y.-J. Shiah, J. Wang, T. A. Beck, C. L. Have, T. Chong, M. Sam, J. Johns, L. Timms, N. Buchner, A. Wong, J. D. Watson, T. T. Simmons, C. P'ng, G. Zafarana, F. Nguyen, X. Luo, K. C. Chu, S. D. Prokopec, J. Sykes, A. D. Pra, A. Berlin, A. Brown, M. A. Chan-Seng-Yue, F. Yousif, R. E. Denroche, L. C. Chong, G. M. Chen, E. Jung, C. Fung, M. H. W. Starmans, H. Chen, S. K. Govind, J. Hawley, A. D'Costa, M. Pintilie, D. Waggott, F. Hach, P. Lambin, L. B. Muthuswamy, C. Cooper, R. Eeles, D. Neal, B. Tetu, C. Sahinalp, L. D. Stein, N. Fleshner, S. P. Shah, C. C. Collins, T. J. Hudson, J. D. McPherson, T. van der Kwast, and R. G. Bristow. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nature Reviews Clinical Oncology*, 47 (7):736–45, 2015. → page 6

[13] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical software*, 76(1): 1–32, 2017. → page 30

[14] Q. Chang, O. I. Ornatsky, I. Siddiqui, A. Loboda, V. I. Baranov, and D. W. Hedley. Imaging mass cytometry. *Cytometry Part A*, 91(2):160–169, 2017. → page 7

[15] J. M. Connors, C. S. Wendy Cozen, A. Carbone, R. T. Hoppe, H.-H. Flechtner, and N. L. Bartlett. Hodgkin lymphoma. *Nature Reviews Disease Primers*, 6:61, 2020. → page 3

[16] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, M. Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–52, 2012. → page 64

[17] I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15:81–94, 2018. → page 4

[18] C. Davidson-Pilon. *Bayesian methods for hackers: probabilistic programming and Bayesian inference*. Addison-Wesley Professional, 2015. → page 30

[19] D. de Jong, A. Koster, A. Hagenbeek, J. Raemaekers, D. Veldhuizen, S. Heisterkamp, J. P. de Boer, and M. van Glabbeke. Impact of the tumor microenvironment on prognosis in follicular lymphoma is dependent on specific treatment protocols. *Haematologica*, 94(1):70–77, 2009. → page 2

[20] J. Demˇsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2016. → page 38

[21] M. Egeblad, E. S. Nakasone, and Z. Werb. Tumors as organs: complex tissues that interface with the entire organism. *Developmental Cell*, 18(6): 884–901, 2010. → page 2

[22] A. Fritsch and K. Ickstadty. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–392, 2009. → page 29

[23] C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. G. andJoachim M Buhmann, S. Brandt,

Z. Varga, P. J. Wild, D. Günther, and B. Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, 2014. → page 6

[24] Y. Goltsev, N. Samusik, J. Kennedy-Darling, S. Bhate, M. Hale, G. Vazquez, S. Black, and G. P. Nolan. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174(4):968–981, 2018. → page 6

[25] B. S. Hill, A. Sarnella, G. D'Avino, and A. Zannetti. Recruitment of stromal cells into tumour microenvironment promote the metastatic spread of breast cancer. *Seminars in Cancer Biology*, 60:202–213, 2020. → page 2

[26] H. W. Jackson, J. R. Fischer, V. R. T. Zanotelli, H. R. Ali, R. Mechera, S. D. Soysal, H. Moch, S. Muenst, Z. Varga, W. P. Weber, and B. Bodenmiller. The single-cell pathology landscape of breast cancer. *Nature*, 578(1): 615–620, 2020. → page 7

[27] Y. Kashima, Y. Togashi, S. Fukuoka, T. Kamada, T. Irie, A. Suzuki, Y. Nakamura, K. Shitara, T. Minamide, T. Yoshida, N. Taoka, T. Kawase, T. Wada, K. Inaki, M. Chihara, Y. Ebisuno, S. Tsukamoto, R. Fujii, A. Ohashi, Y. Suzuki, K. Tsuchihara, H. Nishikawa, and T. Doi. Potentiality of multiple modalities for single-cell analyses to evaluate the tumor microenvironment in clinical specimens. *Nature Scientific Reports*, 11:341, 2021. → page 5

[28] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E. ad D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, and G. P. Nolan. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015. → pages 5, 49

[29] X. Li, J. Yang, L. Peng, A. A. Sahin, L. Huo, K. C. Ward, R. O'Regan, M. A. Torres, and J. L. Meisel. Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. *Breast Cancer Research and Treatment*, 161(2):279–287, 2017. → page 64

[30] F. Liang. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2009. → page 26

[31] E. Lundberg and G. H. H. Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20:285–302, 2019. → page 6

[32] N. M.Anderson and M. C. Simon. The tumor microenvironment. *Current Biology*, 30(16):921–925, 2020. → page 2

[33] J. A. McQuerry, J. T. Chang, D. D. L. Bowtell, A. Cohen, and A. H. Bild. Mechanisms and clinical implications of tumor heterogeneity and convergence on recurrent phenotypes. *Journal of Molecular Medicine*, 95 (1):1167–1178, 2017. → page 4

[34] V. Melnykov, W.-C. Chen, and R. Maitra. Mixsim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, 51:12, 2012. → pages 43, 44

[35] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. 1: 1–29, 2007. → page 20

[36] J. C. Nolz. Molecular mechanisms of CD8+ T cell trafficking and localization. *Cellular and Molecular Life Sciences*, 72(13):2461–2473, 2015. → page 3

[37] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(1): 1160–1167, 2009. → page 64

[38] P. R. Prasetyanti and J. P. Medema. Intra-tumor heterogeneity from a cancer stem cell perspective. *Molecular Cancer*, 16:41, 2017. → page 2

[39] A. Prat, G. Bianchini, M. Thomas, A. Belousov, M. C. Cheang, A. Koehler, P. Gómez, V. Semiglazov, W. Eiermann, S. Tjulandin, M. Byakhow, B. Bermejo, M. Zambetti, F. Vazquez, L. Gianni, and J. Baselga. Research-based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the NOAH study. *Clinical Cancer Research*, 20(2):511–521, 2014. → page 64

[40] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 12:410–420, 2007. → page 37

[41] F. Runa, S. Hamalian, K. Meade, P. Shisgal, P. Gray, and J. Kelber. Tumor microenvironment heterogeneity: challenges and opportunities. *Current Molecular Biology Reports*, 3(4), 2017. → page 2

[42] D. W. Scott and R. D. Gascoyne. The tumour microenvironment in B cell lymphomas. *Nature Reviews Cancer*, 14(1):517–534, 2014. → page 3

[43] S. P. Shah, R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, and S. Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(1):809–813, 2009. → page 2

[44] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. M. 3rd, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019. → page 5

[45] R. H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulation. *Physical Review Letters*, 58(2):86–88, 1987. → page 23

[46] W. C. C. Tan, S. N. Nerurkar, H. Y. Cai, H. H. M. Ng, D. Wu, Y. T. F. Wee, J. C. T. Lim, J. Yeong, and T. K. H. Lim. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*, 40(4):135–153, 2017. → page 6

[47] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Åsa Sivertsson, C. Kampf, E. Sjöstedt, A. AsplundIng, M. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén. Tissue-based map of the human proteome. *Science*, 347:6220, 2015. → pages 50, 65

[48] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1 (2):1–305, 2008. → page 17

[49] N. R. West, S. McCuaig, F. Franchini, and F. Powrie. Emerging cytokine networks in colorectal cancer. *Nature Reviews Immunology*, 15(10):615–29, 2015. → page 3

[50] T. Whiteside. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, 27(1):5904–5912, 2008. → page 3

[51] T. Yan, H. Cui, Y. Zhou, B. Yang, P. Kong, Y. Zhang, Y. Liu, B. Wang, Y. Cheng, J. Li, S. Guo, E. Xu, H. Liu, C. Cheng, L. Zhang, L. Chen, X. Zhuang, Y. Qian, J. Yang, Y. Ma, H. Li, F. Wang, J. Liu, X. Liu, D. Su, Y. Wang, R. Sun, S. Guo, Y. Li, X. Cheng, Z. Liu, Q. Zhan, and Y. Cui. Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma. *Nature Communications*, 10:1670, 2019. → page 2

[52] Y. Yan, A. A. Leontovich, M. J. Gerdes, K. Desai, J. Dong, A. Sood, A. Santamaria-Pang, A. S. Mansfield, C. Chadwick, R. Zhang, W. K. Nevala, T. J. Flotte, F. Ginty, and S. N. Markovic. Understanding heterogeneous tumor microenvironment in metastatic melanoma. *PLoS ONE*, 14(6), 2019. → page 3

[53] K. E. Yost, A. T. Satpathy, D. K. Wells, Y. Qi, C. Wang, R. Kageyama, K. L. McNamara, J. M. Granja, K. Y. Sarin, R. A. Brown, R. K. Gupta, C. Curtis, S. L. Bucktrout, M. M. Davis, A. L. S. Chang, and H. Y. Chang. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nature Medicine*, 25(1):1251–1259, 2019. → page 3

[54] E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. B. Taylor, P. Nghiem, J. H. Bielas, and R. Gottardo. Spatial transcriptomics at subspot resolution with Bayesspace. *Nature Biotechnology*, 2021. → page 16

[55] X. J. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005. → page 70

Appendix A

Supplementary Tables

A.1 Table of Results

We provide the tables of results of the performance of model fitting on different datasets with betas having stronger same-same interactions or sampled from Beta(1,1).

Table A.1: Mean and standard deviation of model fitting performance of biased forward simulation dataset.

Model	Measure	Mean	Std
0p	completeness	0.857187	0.048560
0p	homogeneity	0.983157	0.007022
0p	v-measure	0.915144	0.029284
1p	completeness	0.735447	0.188198
1p	homogeneity	0.868191	0.216154
1p	v-measure	0.795606	0.199292
2k	completeness	0.883850	0.065996
2k	homogeneity	0.971594	0.016790
2k	v-measure	0.924713	0.043516
GMM	completeness	0.705593	0.048127
GMM	homogeneity	0.842986	0.049120
GMM	v-measure	0.767450	0.042649
Phenograph	completeness	0.814558	0.043042
Phenograph	homogeneity	0.784184	0.059775
Phenograph	v-measure	0.798461	0.047262

Table A.2: Mean and standard deviation of model fitting performance of biased MixSim dataset.

Model	Measure	Mean	Std
0p	completeness	0.875236693	0.075659425
0p	homogeneity	0.921128581	0.037124119
0p	v-measure	0.896914581	0.055081612
1p	completeness	0.892359215	0.054872263
1p	homogeneity	0.918600488	0.024575641
1p	v-measure	0.904120785	0.026149414
2k	completeness	0.869725684	0.066449448
2k	homogeneity	0.890028726	0.023868577
2k	v-measure	0.878754192	0.04013768
GMM	completeness	0.50772849	0.048283077
GMM	homogeneity	0.586571789	0.043093081
GMM	v-measure	0.544109815	0.04558028
Phenograph	completeness	0.510982017	0.046239169
Phenograph	homogeneity	0.493967119	0.042074589
Phenograph	v-measure	0.501615384	0.039890154

Table A.3: Mean and standard deviation of model fitting v-measure performance of biased semi-real dataset.

Model	Has Prior?	Has Anchors?	Mean	Std
0p	0	0	0.90822527	0.031673053
0p	0	1	0.960859876	0.030996736
0p	1	0	0.89789876	0.033694059
0p	1	1	0.96127564	0.027232366
1p	0	0	0.632109566	0.17067411
1p	0	1	0.628925145	0.255461044
1p	1	0	0.696573952	0.175336741
1p	1	1	0.690310272	0.266144896
2k	0	0	0.888648131	0.068104716
2k	0	1	0.956780217	0.052743099
2k	1	0	0.884343444	0.078135827
2k	1	1	0.95009383	0.051585232
GMM	-	-	0.785940718	0.045673875
Phenograph	-	-	0.936051601	0.032967326

Table A.4: Mean and standard deviation of model fitting performance of uniform forward simulation dataset.

Model	Measure	Mean	Std
0p	completeness	0.867941	0.053820
0p	homogeneity	0.880434	0.041366
0p	v-measure	0.873984	0.046574
1p	completeness	0.900424	0.038062
1p	homogeneity	0.899736	0.041870
1p	v-measure	0.899641	0.034050
2k	completeness	0.985440	0.006708
2k	homogeneity	0.973067	0.012521
2k	v-measure	0.979203	0.009442
GMM	completeness	0.612733	0.080277
GMM	homogeneity	0.867231	0.044680
GMM	v-measure	0.716034	0.062209
Phenograph	completeness	0.859643	0.044780
Phenograph	homogeneity	0.825205	0.050460
Phenograph	v-measure	0.841972	0.047024

Table A.5: Mean and standard deviation of model fitting performance of uniform MixSim dataset.

Model	Measure	Mean	Std
0p	completeness	0.413696449	0.111553408
0p	homogeneity	0.457907526	0.146299594
0p	v-measure	0.431120036	0.1237403
1p	completeness	0.521230963	0.178908829
1p	homogeneity	0.588238326	0.101380625
1p	v-measure	0.545777682	0.122047162
2k	completeness	0.879518108	0.047495373
2k	homogeneity	0.818225678	0.059007305
2k	v-measure	0.846960739	0.046674122
GMM	completeness	0.523863427	0.08101238
GMM	homogeneity	0.625963909	0.05946859
GMM	v-measure	0.569637994	0.071668829
Phenograph	completeness	0.513937401	0.058683697
Phenograph	homogeneity	0.534124028	0.052772978
Phenograph	v-measure	0.522920606	0.052424961

Table A.6: Mean and standard deviation of model fitting v-measure performance of uniform semi-real dataset.

Model	Has Prior?	Has Anchors?	Mean	Std
0p	0	0	0.709633825	0.063060025
0p	0	1	0.778817134	0.091896866
0p	1	0	0.802928888	0.064422264
0p	1	1	0.8292731	0.075974591
1p	0	0	0.69354209	0.069579469
1p	0	1	0.727117676	0.10180895
1p	1	0	0.7525763	0.085215506
1p	1	1	0.809881181	0.083977301
2k	0	0	0.816013177	0.102819054
2k	0	1	0.91427773	0.050941199
2k	1	0	0.937497633	0.07967313
2k	1	1	0.950588104	0.049260237
GMM	-	-	0.701186096	0.051789961
Phenograph	-	-	0.858827987	0.116769618

A.2 Software Packages Used

We provide a list of software packages that were used in the implementation, testing, and benchmarking of SpatialSort.

Table A.7: Table of Software Packages and Versions

Name	Version
NumPy	1.21.2
Pandas	1.3.3
SciPy	1.7.1
Matplotlib	3.4.3
Seaborn	0.11.2
Scikit-learn	1.0
NetworkX	2.6.3
Numba	0.53.1
Phenograph	1.5.7
Scikit-Posthocs	0.6.7