

Identifying Genes with Location Dependent Noise Variance in Spatial Transcriptomics Data

by

Mohammed Abid Abrar
Student ID: 20366020

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
December 2022

© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Mohammed Abid Abrar

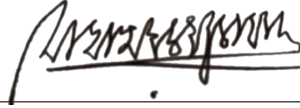
Student ID: 20366020

Approval

The thesis titled “Identifying Genes with Location Dependent Noise Variance in Spatial Transcriptomics Data” submitted by Mohammed Abid Abrar (Student ID: 20366020) has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on February 6, 2023.

Examining Committee:

Supervisor:
(Member)



Dr. Mohammad Kaykobad
Distinguished Professor
Department of Computer Science and Engineering
Brac University

Joint Supervisor:
(Member)




Dr. Md. Abul Hassan Samee
Assistant Professor
Department of Integrative Physiology
Baylor College of Medicine

External Expert Examiner:
(Member)



Dr. Atif Hasan Rahman
Associate Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)

Internal Defense Committee Member 1:
(Member)



Dr. Muhammad Iqbal Hossain
Associate Professor
Department of Computer Science and Engineering
Brac University

Internal Defense Committee Member 2:
(Member)



Dr. Farig Yousuf Sadeque
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Amitabha Chakrabarty
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Spatial transcriptomics (ST) holds the promise to identify the existence and extent of spatial variation of gene expression in complex tissues. Such analyses could help identify gene expression signatures that distinguish between physiology and disease. Existing tools to detect spatially variable genes assume a constant noise variance across location (homoscedastic). This assumption might miss important biological signals when the variance could change across locations, e.g., in the tumor microenvironment. As an alternative, we propose NoVaTeST, a novel method to identify genes with location-dependent noise variance in ST data. NoVaTeST models gene expression as a function of location with a heteroscedastic noise. It then compares the model to one with homoscedastic noise to detect genes that show significant spatial variation in noise. Our results show genes detected by NoVaTeST provide complimentary information to existing tools while providing important biological insights.

Dedication

To my beloved wife and my mother, who have been my constant source of inspiration and support.

Acknowledgement

Firstly, all praise to the Almighty Allah for whom the thesis have been completed without any major interruption.

Secondly, I would like to thank my supervisors, Dr. Mohammad Kaykobad and Dr. Md. Abul Hassan Samee, and also my mentor Dr. Mohammad Saifur Rahman for their wise guidance during my studies and research. This thesis would not have been possible without the intelligent guidance from him.

Table of Contents

| | |
|---|-----------|
| Declaration | i |
| Approval | ii |
| Abstract | iv |
| Dedication | v |
| Acknowledgment | vi |
| Table of Contents | vii |
| List of Figures | ix |
| Nomenclature | ix |
| 1 Introduction | 1 |
| 1.1 Background | 1 |
| 1.2 Motivation | 2 |
| 1.3 Research Objective | 3 |
| 1.4 Organization | 3 |
| 2 Background | 4 |
| 2.1 Cell Biology and Central Dogma | 4 |
| 2.2 Transcriptomics | 5 |
| 2.3 Spatial Transcriptomics | 6 |
| 2.4 Spatial Modeling and Gaussian Process | 7 |
| 2.5 Gaussian Process Regression | 7 |
| 2.5.1 Classical Regression vs GP Regression | 7 |
| 2.5.2 GPR Model and Training | 8 |
| 2.5.3 GPR Prediction | 9 |
| 2.6 Heteroscedastic Gaussian Process | 10 |
| 3 Literature Review | 12 |
| 3.1 Experimental Methods | 12 |
| 3.1.1 ISH Based Methods | 13 |
| 3.1.2 ISC Based Methods | 15 |
| 3.2 Computational Methods | 16 |
| 3.2.1 Spatially Expression Pattern Analysis | 17 |

| | | |
|----------|---|-----------|
| 4 | Methodology | 19 |
| 4.1 | Count Data Representation | 19 |
| 4.2 | Spatial Modeling of Gene Expression | 19 |
| 4.3 | Statistical Test to Detect Noisy Genes | 22 |
| 4.4 | Clustering | 23 |
| 4.5 | Dataset Description | 23 |
| 5 | Results from Squamous Cell Carcinoma Data | 25 |
| 6 | Results from Cutaneous Malignant Melanoma Data | 28 |
| 7 | Discussion | 30 |
| 8 | Conclusion | 33 |
| | Bibliography | 42 |

List of Figures

| | | |
|-----|---|----|
| 1.1 | Simulated example showing the importance of heteroscedastic models | 2 |
| 2.1 | Central dogma of molecular biology. | 4 |
| 2.2 | ST data acquisition pipeline for Visium. | 6 |
| 2.3 | Comparison of classical regression and GP regression. | 7 |
| 2.4 | Example of Gaussian process regression. | 9 |
| 2.5 | Example of data with heteroscedastic noise. | 10 |
| 3.1 | Brief timeline of different spatial transcriptomics methods. | 13 |
| 3.2 | Brief overview of ISH based methods. | 14 |
| 3.3 | Brief overview of ST and Visium methods. | 15 |
| 3.4 | Different types of computational methods for ST data analysis. | 16 |
| 3.5 | Graphical overview of SpatialDE and SPARK pipelines. | 18 |
| 4.1 | Graphical overview of the HGP model fitting process. | 21 |
| 4.2 | Statistical test for model selection by comparing NLPD values. | 22 |
| 4.3 | Annotated H&E stained images of the datasets used. | 23 |
| 5.1 | Results obtained from squamous cell carcinoma data using. | 26 |
| 6.1 | Results obtained from cutaneous malignant melanoma data. | 29 |
| 7.1 | Checking the existence and extent of mean-variance artifacts in the datasets. | 31 |

Chapter 1

Introduction

1.1 Background

The recent advancement in transcriptomics technologies has made it possible to profile gene-expression levels across tissue with spatial information. These include next-generation sequencing-based techniques with spatially barcoded microarray, such as Visium [1], which can profile over 18,000 genes in thousands of spots (one spot contains 1 to 100 cells) [2], and fluorescent in-situ hybridization-based techniques such as seqFISH [3]–[5].

Analyzing this spatial transcriptomics (ST) data can reveal the spatial organization of different molecules and cell types in complex tissues, which, in turn, can help us understand the mechanism of tissue function and its effect on gene expression. Furthermore, incorporating spatial information with gene expression profiles could also help identify diseases and innovate their potential treatments. For example, spatial heterogeneity of cell types and gene expressions is a defining characteristic of tumor microenvironments [6]. Variation in spatial expression may reflect communication among neighbouring cells, location-specific states, or the migration of cells to specific tissue locations to perform their functions.

Modeling the gene expression as a function of location is the first step toward the spatial variation analysis of ST data. Existing models of these data assume that gene expression noise has a constant variance across locations. Thus, a gene's expression y_i at location x_i is modeled using a function $f(x_i)$ and a Gaussian noise ϵ with mean zero and variance σ^2 :

$$y_i = f(x_i) + \epsilon; \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

These models have provided valuable biological insights by identifying genes that show significant location-dependent changes in expression [7], [8]. However, we still lack models that assess if the noise variance of a gene's expression is location-dependent, *i.e.*, heteroscedastic [9].

1.2 Motivation

To capture the extent of spatial heterogeneity of variance, the underlying model of gene expression as a function of spatial coordinates across a tissue sample must incorporate heteroscedasticity. In this case, the variance of ϵ is not an unknown constant σ^2 but a variable σ_i^2 that depends on the location x_i . Figure 1.1(A) shows an example of a simulated gene expression with smooth mean function and heteroscedastic noise, *i.e.*, σ_i^2 changing with x_i . In the presence of heteroscedastic noise, a homoscedastic (constant noise) model would not be able to capture the spatial variation of gene expression, and thus a heteroscedastic model is expected to yield a better fit. One way to measure this goodness-of-fit is to use negative log predictive density (NLPD) on a test dataset, which penalizes both over-confident and under-confident predictions [10]. As an example, the NLPDs of a homoscedastic model and a heteroscedastic model for a simulated expression are shown in Figure 1.1(B). We can clearly see that the homoscedastic model fails to capture the spatially variable noise, whereas homoscedastic model correctly predicts the noise variance, and hence results in a lower NLPD.

This premise of the existence of heteroscedastic noise is set by prior analyses of spatial data in biology [11]–[13], economics [14]–[16], and robotics [17]–[19]. For example, the magnitude of imaging noise for apparent diffusion coefficient during whole-body diffusion-weighted MRI is heteroscedastic [11]. Park *et. al.* [13] used heteroscedastic noise variance to encode confidence levels in predicting the tumor mutation burden from whole slide images. In economics, heteroscedastic regression is used for modeling time-varying volatility and stochastic volatility models in time series data [14]–[16]. Heteroscedastic regression is also used in robotics [17], [18], including where multi-modal sensor data are combined for terrain modeling [19]. Other fields where heteroscedastic noise variance is used include biophysical variable estimation [20], vehicle control [21], cosmological redshift estimation [22], etc.

In the context of ST data, spatial variation of noise variance could indicate gene

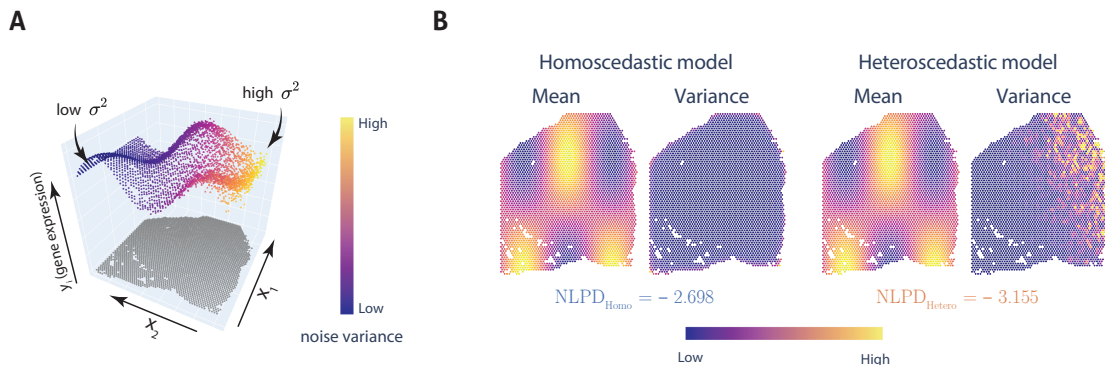


Figure 1.1: A simulated example showing the importance of heteroscedastic models. (A) Simulated gene expression with heteroscedastic noise. The noise variance increases as x_1 decreases and x_2 increases. (B) Predicted mean and noise variance along with the NLPD of the simulated data for spatial models – homoscedastic and heteroscedastic. A lower NLPD indicates a better model fitting.

expression variation due to sequencing technology, as well as variation due to biology. For example, a common source of technical noise is the mean-variance relation [23]–[25], where the noise variance typically increases with the mean expression of a gene. Several variance-stabilizing transformations are available to remove this type of technical noise [26], [27]. On the other hand, the variation in noise could also be due to underlying biologies, such as cell-type heterogeneity [28] and phenotypical variation across spatixal coordinates. Moreover, if the sample being analyzed is partially affected by some condition, the noise variance for some genes is likely to be different in the affected region compared to the non-affected region [29].

1.3 Research Objective

Spatial signals do not always indicate constant variance, and thus suggest modeling ϵ of gene expression with a location-dependent variance σ_i^2 . Moreover, as of now, there are no tools to detect genes with location-dependent noise variance in ST data, which might be important for understanding the underlying biology. Motivated by this, the main goal of this thesis are as follows:

- Develop a more generalized framework for modeling gene expression with spatial coordinates.
- Develop a method to identify genes with location-dependent noise variance in ST data.
- Develop a method to cluster genes with similar noise variance patterns, and
- Validate the method on real dataset using further downstream analysis such as pathway enrichment

With these goals in mind, in this thesis, we propose *noise variation testing in ST data (NoVaTeST)*, a pipeline to identify genes with statistically significant heteroscedasticity.

1.4 Organization

The rest of the thesis is organized as follows: Chapter 2 presents the relevant background. Chapter 3 presents the literature review and related works. Chapter 4 presents the proposed pipeline and describes the dataset used. Chapter 5 and 6 presents the experimental results. Chapter 7 presents the discussion on the results. Chapter 8 presents the conclusion and future work.

Chapter 2

Background

2.1 Cell Biology and Central Dogma

Cells are the basic units of life in all living organisms. Cells are incredibly complex and perform a wide range of functions that are essential for the survival and growth of an organism.

One of the fundamental concepts of cell biology is the central dogma of molecular biology [30], which describes the flow of genetic information within living cells (Figure 2.1). The central dogma states that the information stored in DNA is first transcribed into RNA, and then translated into protein.

DNA, or deoxyribonucleic acid, is a long, double-stranded molecule that stores the genetic information of an organism [31]. It is made up of nucleotides, which are composed of a sugar molecule, a phosphate group, and a nitrogenous base. There are four different nitrogenous bases in DNA: adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these bases determines the genetic information of an organism.

RNA, or ribonucleic acid, is a single-stranded molecule that is similar to DNA [31]. It is also made up of nucleotides, but the sugar molecule is ribose instead of deoxyribose, and the base uracil (U) is used instead of thymine. RNA plays a critical role in the central dogma because it acts as a messenger between DNA and proteins.

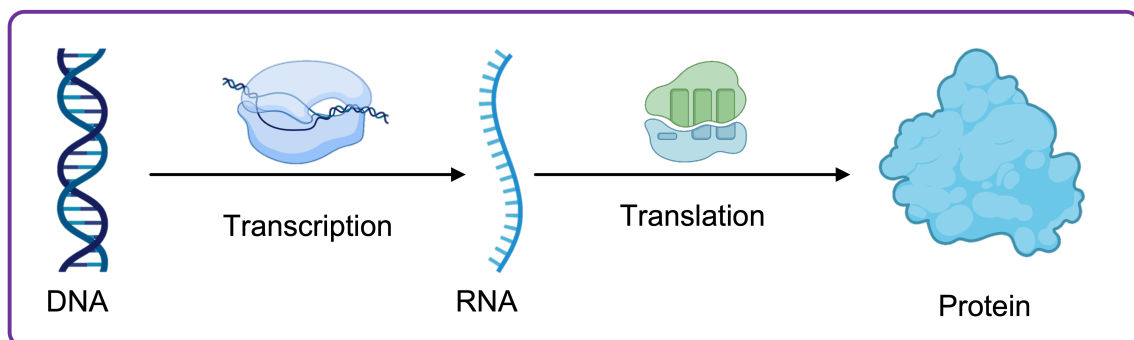


Figure 2.1: The central dogma of molecular biology. DNA is transcribed into RNA, which is then translated into protein.

Proteins are complex molecules that perform a wide range of functions in the body [31]. They are made up of smaller molecules called amino acids, which are linked together in a specific sequence determined by the genetic information stored in DNA.

The process of transcribing DNA into RNA is called transcription [31]. It involves the enzyme RNA polymerase, which reads the DNA sequence and synthesizes a complementary RNA molecule. The RNA molecule is then used as a template to create a protein during the process of translation [31].

2.2 Transcriptomics

Transcriptomics is the study of the transcriptome, which is the complete set of transcripts (RNA molecules) present in a cell or tissue at a given time [32]. Transcriptomics is a subfield of genomics, which is the study of the genome (the complete set of genetic material) of an organism [33]. The transcriptome provides a snapshot of gene expression in a cell, and can be used to understand the function of different genes and their relationships with one another.

Transcriptomics involves the use of various technologies and techniques to analyze and interpret transcriptomic data. These techniques can include Reverse transcription polymerase chain reaction (RT-PCR) [34], microarrays [35], and next-generation sequencing (NGS) [36]. RT-PCR is a laboratory technique that is used to amplify and analyze specific transcripts, while microarrays and NGS are high-throughput technologies that can be used to simultaneously analyze the expression levels of many transcripts. With the recent advancements, it is now possible to obtain transcriptomics data at single cell level using single-cell sequencing (scRNA-seq) technologies [37].

One of the key goals of transcriptomics is to gain insight into gene expression and regulation. By analyzing transcriptomic data, one can identify which genes are being expressed in a cell or tissue, and at what levels. This can help to understand the function of different genes, and how they are regulated in response to various biological and environmental factors such as cancer.

In addition to studying gene expression, transcriptomics can also be used to study the effects of different factors on gene expression. For example, researchers can use transcriptomic data to investigate how a particular drug or environmental factor affects gene expression, and how this in turn affects the function and behavior of a cell or tissue [32]. This can be useful in fields such as drug discovery and disease diagnosis, where understanding the mechanisms underlying gene expression can help to identify new targets for therapeutic intervention.

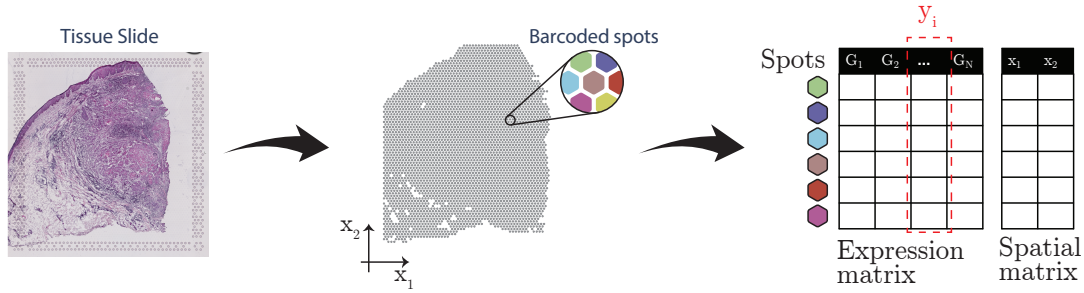


Figure 2.2: Visium technology to acquire ST data at pre-designated spatially barcoded spots from a tissue slice placed on top of a slide. Representation of ST count data is shown on the right. Each column in the Expression matrix represents the expression of a particular gene with spot locations given by the Spatial matrix. The Expression matrix is a $N \times G$ matrix, where N is the number of spots and G is the number of genes. The Spatial matrix is a $N \times 2$ matrix, where each row represents the x and y coordinates of a spot.

2.3 Spatial Transcriptomics

Spatial transcriptomics is a rapidly growing field of study that focuses on analyzing the spatial organization of the transcriptome within a tissue or cell [38]. This is in contrast to traditional transcriptomics, which typically involves analyzing the transcriptome as a whole, without considering its spatial distribution.

One of the key benefits of spatial transcriptomics is that it allows the study of gene expression and regulation with spatial information. This is important because different cells within a tissue or organ can vary greatly in their gene expression patterns, and traditional transcriptomic methods are not able to capture this heterogeneity. This can help us to gain a more detailed and accurate view of the underlying mechanisms of various biological processes.

There are several technologies to get spatial transcriptomics data from a tissue sample. This includes the use of advanced microscopy imaging techniques, such as *in situ* sequencing (ISS) [39], [40], fluorescent *in situ* hybridization (FISH) [3]–[5], [41], or laser capture microdissection (LCM) [42]. These techniques can be used to obtain spatially resolved transcriptomic data at the single cell level. However, these techniques are time-consuming and expensive, have low throughput, and are not suitable for large-scale studies. To overcome these limitations, researchers have developed a variety of next-generation sequencing (NGS)-based methods to obtain spatial transcriptomics data. These methods include spatial transcriptomics technology [38] and Visium [1], which use barcoded microarrays to capture and sequence mRNA from a tissue sample placed on top of a slide. The barcodes on the microarrays are used to identify the spatial location of each mRNA molecule. Figure 2.2 shows an overview of the Visium technology and the resulting ST data.

2.4 Spatial Modeling and Gaussian Process

One of the primary goal of thesis is spatial modeling of transcriptomics data, which involves the use of mathematical and statistical techniques to analyze and understand the spatial patterns. One approach to spatial modeling is to use Gaussian processes, which are a type of mathematical model that can be used to model complex spatial patterns and incorporate uncertainty in the data [43]. Gaussian processes have been applied in a wide range of fields, including geography [44], environmental science [45], and epidemiology [46], [47]. For example, they have been used to analyze geographical data and predict the spread of disease outbreaks.

2.5 Gaussian Process Regression

A Gaussian process (GP) is a stochastic process, which is a collection of random variables, any finite number of which have a joint Gaussian distribution [43]. Gaussian processes are a popular choice of prior over functions in Bayesian nonparametric models [43], since they are flexible and can be used to model a wide range of spatial pattern by carefully selecting the mean function and/or the covariance kernel. In addition, they have a closed-form solution for regression, meaning that training and prediction are straightforward.

2.5.1 Classical Regression vs GP Regression

For task for a regression problem is to predict the output y given some training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ and a new input x (Figure 2.3A). To model the relation from input to output, an analytical formula is forced on the training data for classical

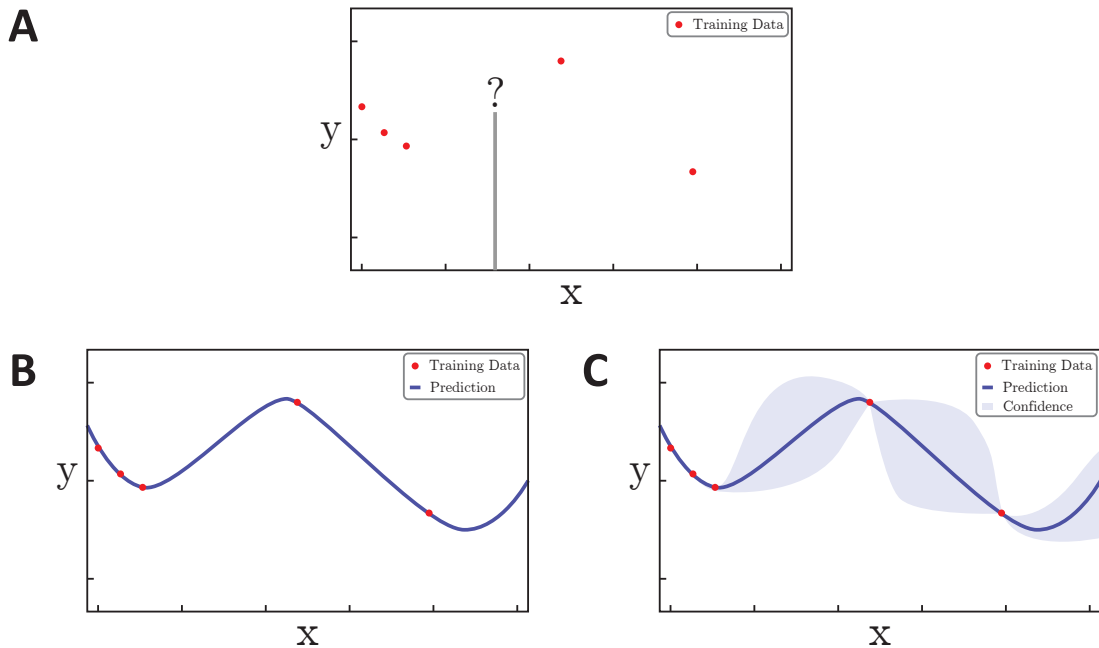


Figure 2.3: Comparison of classical regression and GP regression. (A) The main goal of regression problems. (B) Classical regression. (C) GP Regression.

methods. For example, linear regression assumes a linear relation between x and y , i.e., $y = \theta^T x + c$. However, these methods only provide a single function that it considers to fit the training data the best (Figure 2.3B).

On the other hand, instead of fitting a line or curve to the data, Gaussian process regression (GPR) models the distribution of the data, and can make predictions based on this distribution [43]. GPR is a Bayesian method, which means that it incorporates uncertainty into its predictions, and can be used to make probabilistic statements about the data (Figure 2.3C). Hence, GPR is very flexible and this able to capture more complex relationships in the data.

2.5.2 GPR Model and Training

A GP is completely specified by its mean function, $m(x)$, and covariance kernel function, $k(x, x')$, where x and x' are any two points in the input space [43]. The mean function gives the mean of the process at any point in the input space, and the covariance function gives the covariance between the values of the process at any two points in the input space. Assuming $y_i = f(x_i) + \epsilon$, where ϵ is the noise and $f(x)$ is some unknown function, we can use GP to model $f(x)$ by assigning a GP prior over $f(x)$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \quad (2.1)$$

For a finite number of inputs $x = [x_1, x_2, \dots, x_n]$, this implies that the observations $f = [f_1, f_2, \dots, f_n]$, where $f_i = f(x_i)$, is a sample of a multivariate Gaussian distribution

$$p(f(x) \mid \theta_m, \theta_k) = \mathcal{N}(\mu, \Sigma). \quad (2.2)$$

Here, $\mu = [\mu_1, \mu_2, \dots, \mu_n]$ is the mean vector, where $\mu_i = m(x_i)$ is the mean of the Gaussian distribution at x_i . Σ is the covariance matrix, where the diagonal terms $\Sigma_{(i,i)} = \sigma_s^2 = k(x_i, x_i)$ are the variance of the Gaussian distribution at input x_i , and the off-diagonal terms $\Sigma_{(i,j)} = k(x_i, x_j)$ denotes the covariance of the observations at x_i and x_j . Finally, θ_m and θ_k are the parameters of $m(x)$ and $k(x, x')$, respectively. The covariance kernel (and hence the covariance matrix) is the most important part of a GP as it controls the families of functions that can be learned by the GP. Some commonly used covariance kernels, namely, squared-exponential, periodic, and linear, and the resulting sample functions are shown in Figure 2.4A [43].

It should be noted that $\mathcal{N}(\mu, \Sigma)$ is the *prior* distribution, i.e., before the training data \mathcal{D} is incorporated. As an example, the *prior* distribution for $n = 20$ points with zero mean and squared-exponential kernel function is shown in Figure 2.4B.

Given the training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, we can find the parameters of $m(x)$ and $k(x, x')$ by minimizing the negative log-likelihood of y conditioned over the parameters:

$$\hat{\theta}_m, \hat{\theta}_k = \arg \min(-\log p(y \mid x, \theta_m, \theta_k)). \quad (2.3)$$

These parameters can then be used to obtain the posterior distribution over $f(x_*)$ at a test point x_* .

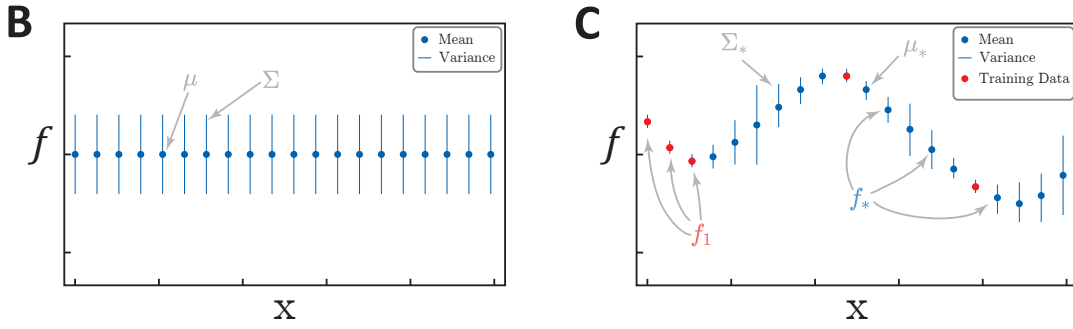
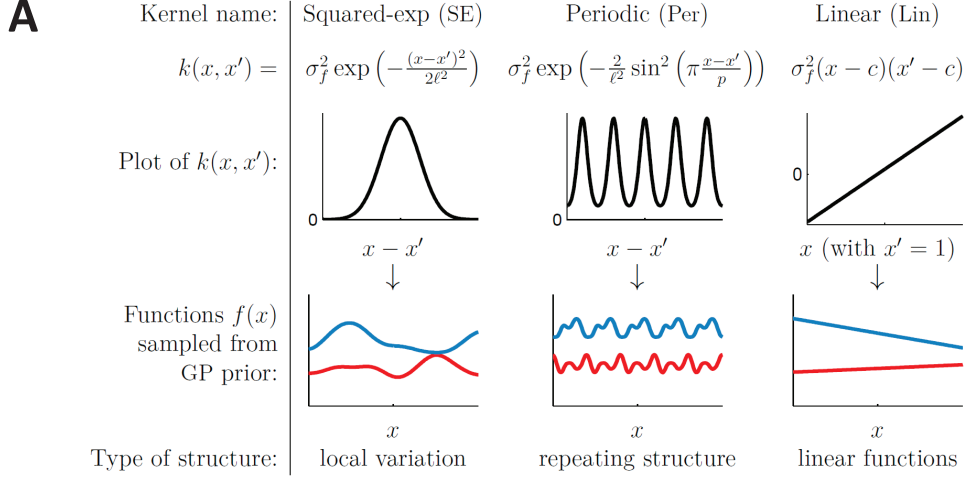


Figure 2.4: Example of Gaussian process regression. (A) Some common covariance kernel, their equations, and example functions sampled from the corresponding GP prior. (B) GPR prior for $n = 20$ points with zero mean and squared-exponential covariance kernel. (C) The posterior predictive distribution after conditioning over the training points.

2.5.3 GPR Prediction

One of the main advantages of GP over other Bayesian methods is that the predictive density has a closed form solution. In fact, the *posterior* of $f(x_*)$ is also a multivariate Gaussian distribution. To find the *posterior* predictive density of output $f_* = f(x_*)$ for a test point x_* , we first define a joint *prior* of training observations f_1 and f_* [43]:

$$p\left(\begin{bmatrix} f_1 \\ f_* \end{bmatrix} \mid \theta_m, \theta_k\right) = \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{1*} \\ \Sigma_{*1} & \Sigma_{**} \end{bmatrix}\right), \quad (2.4)$$

where $\Sigma_{11} = k(x, x)$, $\Sigma_{1*} = k(x, x_*)$, $\Sigma_{*1} = k(x_*, x)$, and $\Sigma_{**} = k(x_*, x_*)$. The parameters $\hat{\theta}_m$ and $\hat{\theta}_k$ are obtained from equation 2.3. The *posterior* predictive distribution can then be found by conditioning over the f_1 , which is a multivariate Gaussian [43]:

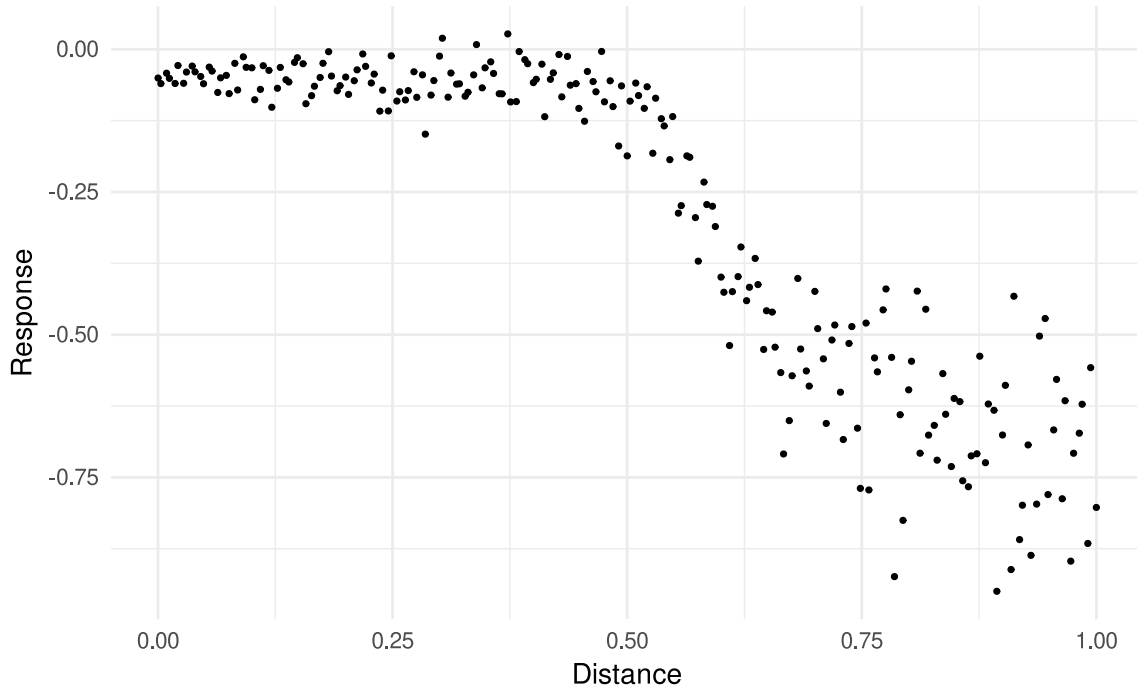


Figure 2.5: Data collected from a light detection and ranging experiment [48]. The noise variance increases with distance, i.e., heteroscedastic.

$$\begin{aligned}
 p\left(f_* \mid f_1, x, \hat{\theta}_m, \hat{\theta}_k\right) &= \mathcal{N}\left(\mu_*, \Sigma_*\right), \text{ where} \\
 \mu_* &= \mu + \Sigma_{*1} \Sigma_{11}^{-1} \left(f_1 - \mu\right) \\
 \Sigma_* &= \Sigma_{**} - \Sigma_{*1}^T \Sigma_{11}^{-1} \Sigma_{1*}
 \end{aligned}
 \tag{2.5}$$

The predictive distribution for the prior in Figure 2.4B after conditioning over the training points in Figure 2.3 is shown in Figure 2.4C. Due to the structure of the squared-exponential covariance matrix (i.e., closer points have higher correlation), the points closest to the training points have low uncertainty, while those further away have high uncertainty.

2.6 Heteroscedastic Gaussian Process

The Gaussian process model described above assumes that the variance of the noise is constant. However, in many real-world applications, the variance of the noise is not constant, and can vary across the input space. This is known as heteroscedasticity, and can be modeled using a heteroscedastic Gaussian process (HGP). Figure 2.5 shows an example data from a light detection and ranging experiment [48] where the noise variance increases with distance.

The HGP model is similar to the GP model, except that the variance of the noise is modeled as a function of the input space. As an example, we can assign a second GP prior over $\log\left(\sigma^2(x)\right)$ with zero mean and a second kernel $k_2(x, x')$, given by

$$\log\left(\sigma^2(x)\right) \sim \mathcal{GP}_2\left(0, k_2(x, x')\right)
 \tag{2.6}$$

where $\sigma^2(x)$ is the variance of the noise at point x . The HGP model can be used to model a wide range of spatial patterns, including spatially varying noise.

The posterior distribution of y_* at a test point x_* for the HGP model is a multivariate Gaussian with mean and covariance given by the following equations [43]:

$$\begin{aligned}\mu_* &= \mu + \Sigma_{*1} (\Sigma_{11} + \Sigma_N)^{-1} (f_1 - \mu) \\ \Sigma_* &= \Sigma_{**} - \Sigma_{*1}^T (\Sigma_{11} + \Sigma_N)^{-1} \Sigma_{1*}\end{aligned}\tag{2.7}$$

where $\Sigma_N = \text{diag}(\sigma^2(x))$ is the diagonal matrix of the noise variance at each point in the input space predicted by the second GP prior.

Chapter 3

Literature Review

The field of spatial transcriptomics has seen significant growth in recent years, with both experimental and computational methods being developed and refined to enable the analysis of gene expression in a spatial context. This chapter aims to review the existing literature on both experimental and computational methods related to spatial transcriptomics. The purpose of this review is to provide an overview of the current state of the field and to identify any gaps or areas in need of further research.

Experimental methods for spatial transcriptomics involve the use of specialized techniques, such as tissue sectioning and imaging, to enable the analysis of gene expression at a cellular level. These techniques have been used to study a wide range of biological systems, including both *in vitro* and *in vivo* models.

Computational methods, on the other hand, involve the use of algorithms and software tools to analyze and interpret the data generated by experimental methods. These methods are critical for the analysis and interpretation of large-scale spatial transcriptomics data sets, and have been used to identify patterns and trends in gene expression that may not be apparent from individual data points.

3.1 Experimental Methods

Experimental methods play a crucial role in this field, as they provide the means to collect and analyze the data needed to study gene expression in a spatial context. In this essay, we will review the literature on experimental methods for spatial transcriptomics while highlighting key developments.

There are five main methods for spatial transcriptomics, namely,

1. Microdissection-based methods such as LCM [49]–[51]
2. *in situ* hybridization (ISH) based methods such as single-molecule FISH (sm-FISH) [41] and multiplexed error robust FISH (MERFISH) [4]
3. *in situ* sequencing based methods such as STARmap [39] and BaristaSeq [40]
4. *in situ* capture (ISC) based methods such as Visium [1] and HDST [52]
5. *in silico* construction based methods such as DistMap [53].

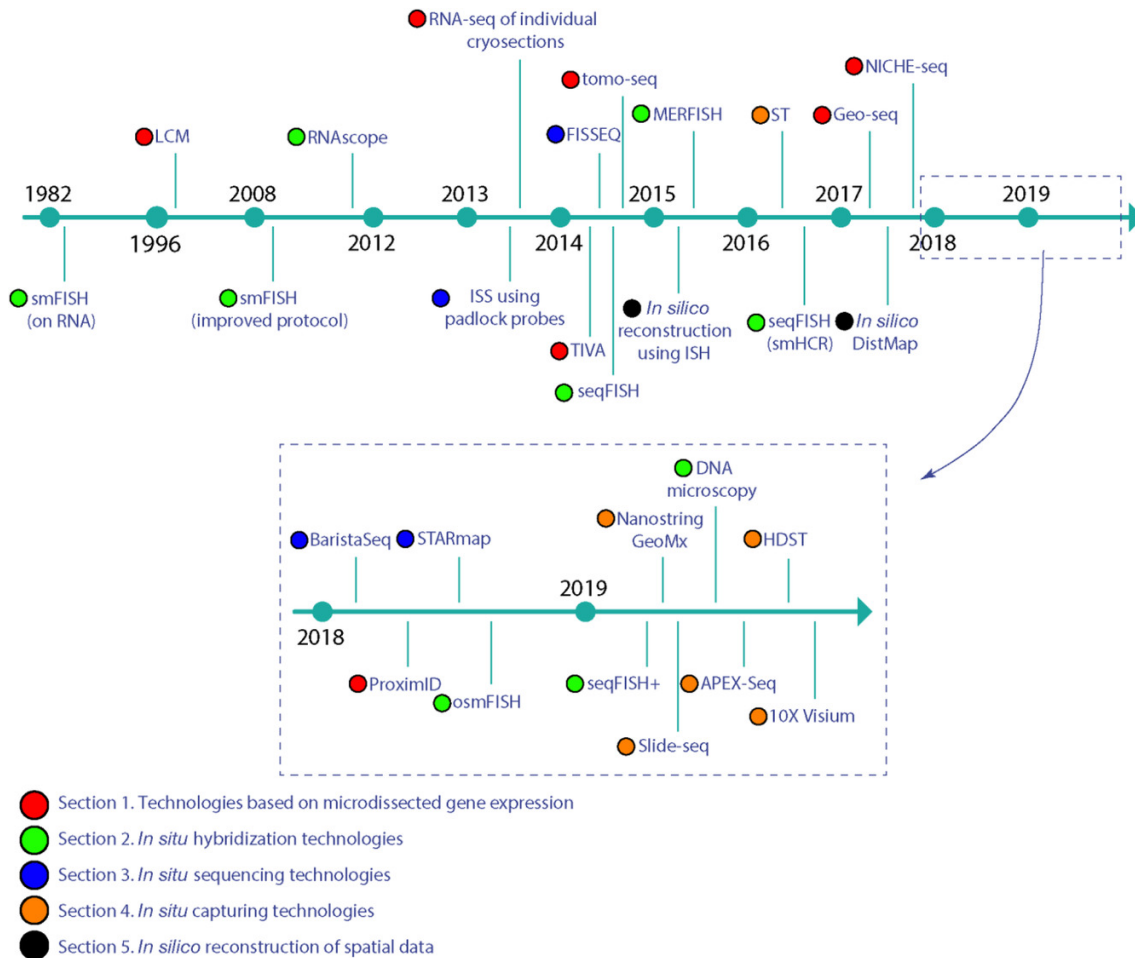


Figure 3.1: Brief timeline of different spatial transcriptomics methods (figure adapted from Asp *et al.*, 2020 [54]).

Figure 3.1 shows a brief timeline of the development of different spatial transcriptomics methods. Each of these methods has its own advantages and disadvantages, and the choice of method depends on the biological system being studied. Among these, ISH and ISC methods are the most widely used, with the former being used to study targeted gene expression and the latter being used to study complete transcriptomics and to identify cell types.

3.1.1 ISH Based Methods

ISH based methods involve the use of fluorescent probes to detect the expression of specific genes in a tissue sample. The most common type of ISH based method is FISH, which involves the use of fluorescently labeled DNA probes to detect and visualize the expression of specific genes in a tissue sample. Several variations of FISH have been developed in recent years, including single-molecule FISH (smFISH) [41], sequential FISH (seqFISH) [3], and multiplexed error-robust FISH (MERFISH) [4]. Figure 3.2 shows the brief overview of these methods [54].

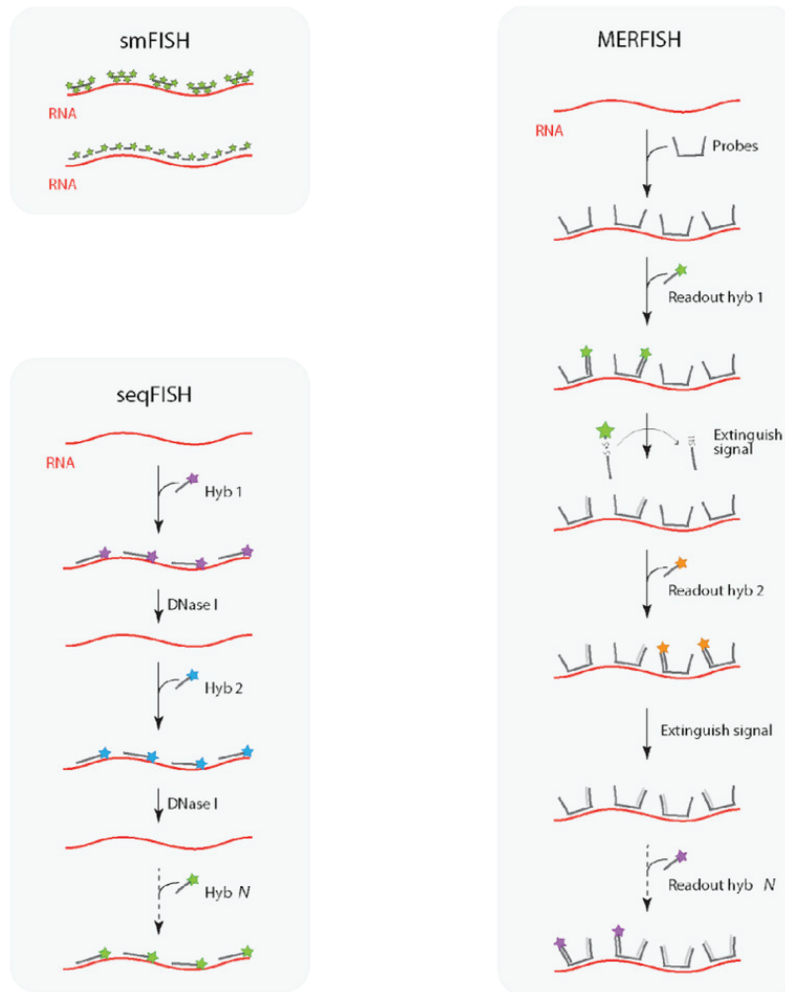


Figure 3.2: Brief overview of ISH based methods (adapted from Asp *et al* [54]).

smFISH

smFISH is one of the first methods capable of profiling individual cells while maintaining the spatial context of the tissue sample [41]. smFISH allows for the detection and quantification of individual RNA molecules within cells, enabling the analysis of gene expression at a sub-cellular single-molecule resolution.

seqFISH

seqFISH is a high-throughput FISH method that allows for the simultaneous detection and quantification of multiple RNA targets within cells [3]. seqFISH method is similar to smFISH, but instead of using a single probe to detect the expression of a single gene, it uses multiple probes to detect the expression of multiple genes. This has been used to study the expression of multiple genes in a variety of biological systems, including both in vitro and in vivo models.

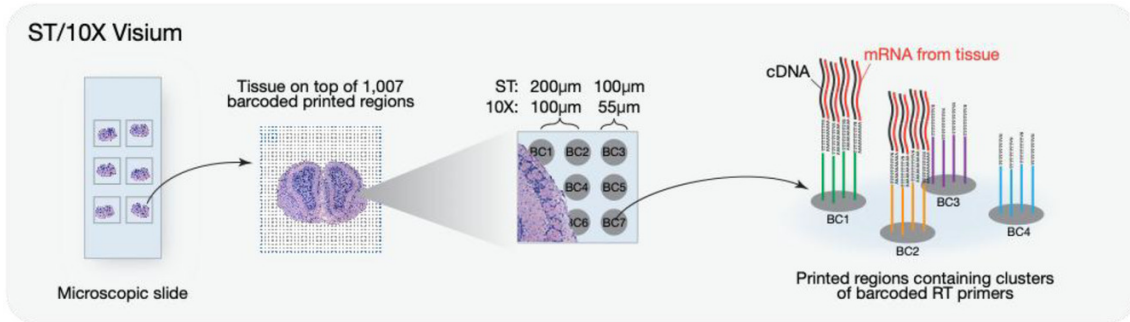


Figure 3.3: Brief overview of ST and Visium methods (adapted from Asp *et al* [54]).

MERFISH

Since there are a small number of distinct color channels available, standard FISH-based methods are constrained in terms of the number of genes that can be analyzed at once. To overcome this limitation, MERFISH was developed, which is a multiplexed FISH method that enables the simultaneous detection and quantification of multiple RNA targets within cells [4]. MERFISH uses a combination of fluorescent dyes and error-correcting codes, and employs multiple rounds of hybridization to increase the number of RNA species that can be imaged simultaneously. In fact, MERFISH is capable of profiling hundreds to tens of thousands of RNA molecules in single cells, which is significantly more than that can be imaged using standard FISH-based methods.

3.1.2 ISC Based Methods

In contrast to ISH based methods, ISC based methods involve the use of specialized probes to capture and isolate RNA molecules from a tissue sample. These probes are then used to generate a library of RNA molecules, which can then be sequenced to identify the RNA molecules present in the tissue sample. The most common type of ISC based method is 10X Visium [1], which involves the use of spatially barcoded probes to capture and isolate RNA molecules from a tissue sample. Other ISC based methods include spatial transcriptomics (ST) [38] and high definition ST (HDST) [52], as shown in Figure 3.3 [54]. The main advantage of ISC based methods is that they are capable of profiling the complete and unbiased transcriptome of a tissue sample, which is not possible using ISH based methods.

ST

ST is one of the first ISC-based technology. ST involves the use of spatially barcoded probes to capture and isolate RNA molecules from a tissue sample placed on top of a glass slide [38]. Each slide contains around 1000 spots of 100 μm diameter. The distance between adjacent spots are about 200 μm , which allows ST to capture the transcripts of 10 to 40 cells per spot.

Visium

Visium technology is a spatial transcriptomics method developed by 10x Genomics [1]. Similar to ST, it uses spotted microarrays of mRNA-capturing probes on the

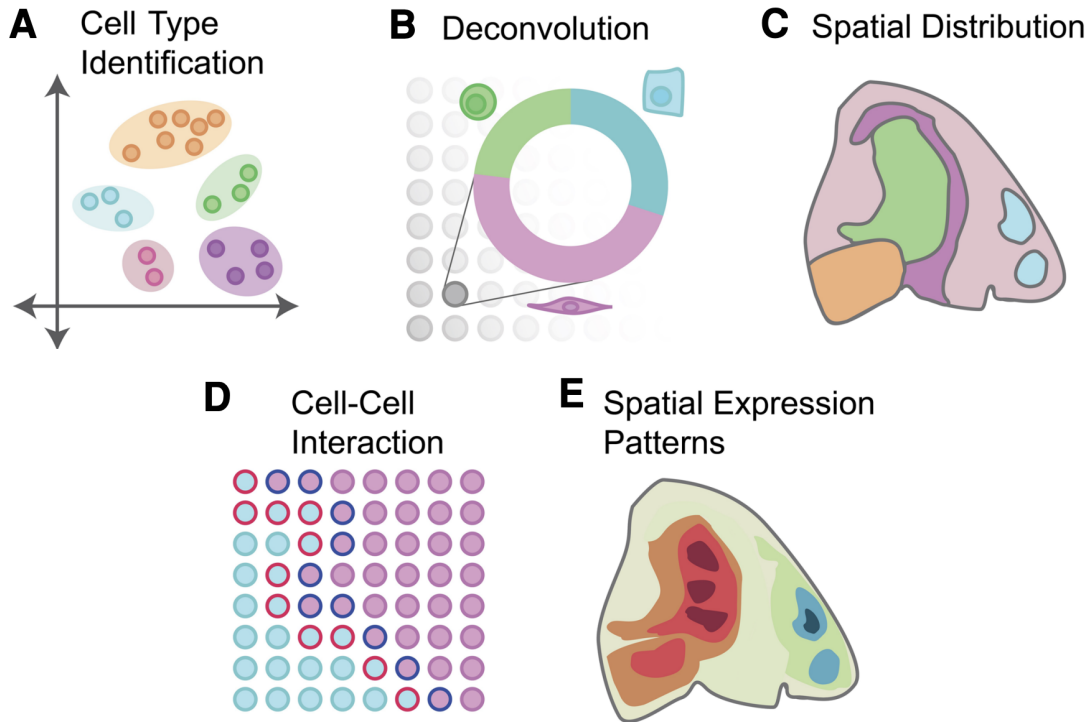


Figure 3.4: Different types of computational methods for ST data analysis (adapted from Dries *et al* [55]).

surface of glass slides, but with a greater number of spots than ST, a smaller spot size, and a greater quantity of capture probes per spot. On each slide, there are approximately 5000 barcoded spots containing millions of spatially barcoded capture oligonucleotides within each capture area of the Visium Spatial Gene Expression slides. Each barcoded spot has a $55 \mu\text{m}$ diameter, and the distance between the centers of adjacent spots is approximately $100 \mu\text{m}$. The placement of the spots is staggered to reduce the distance between them. On average, mRNA from between 1 and 10 cells is captured per spot, providing resolution close to that of a single cell [2].

3.2 Computational Methods

Although a number of high throughput spatial transcriptomics technologies have been developed, the analysis of ST data is still a challenging task. In this section, we will discuss some of the computational methods that have been developed so far to analyze spatial transcriptomics data. These tools for ST analysis connects gene expression and cellular/transcript locations, which is required for extracting meaningful biological insight, linking cell morphology, and coming up with novel hypotheses.

A number of different analyses can be performed on ST data, as shown in Figure 3.4 [55]. The most common analyses are cell type identification, spatial distribution analysis, and spatial expression pattern analysis.

For ST data with single-cell resolution, for example FISH-based methods [3]–[5], [56]–[60], the common process is similar to single cell RNA (scRNA) data – Louvain [61] or Leiden-based clustering [62], identification of marker genes for each cluster, followed by manual or automated [63] annotation (Figure 3.4A). However, since FISH-based data contains expression of targeted genes, it is not possible to identify novel cell types. To solve this, some methods have been developed that integrate scRNA data to impute whole transcriptome from the FISH data [64], [65]. For ISC-based methods, the ST data is multi-cellular level, meaning each spot contains multiple cells. Therefore, the techniques for scRNA data cannot be directly applied to ST data. A common approach to identify different cell types and their relative proportion using deconvolution techniques used in RNA-seq data analysis [66] (Figure 3.4B). However, given the distinct characteristics of spatial data (for example fewer cells per spot), the deconvolution methods need to be adapted to ST data for optimal results [67]–[74]. STdeconvolve [75], Cell2location [76], and SPOTlight [72] are the most commonly used tools for cell type distribution mapping.

Spatial organization of cell types is another important study that is only possible due to ST data (Figure 3.4C). This is crucial for the study of histology and cell-cell communication (Figure 3.4D). Pairwise enrichment analysis can identify cell type pairs that are likely to be adjacent [77], [78]. Tools such as BayesSpace [79], SPICEMIX [80] and staNMF [81] leverage the spatial information and the fact that similar cell types are likely to be physically nearby to identify spatial patterns.

3.2.1 Spatially Expression Pattern Analysis

The idea of differential gene expression, which is the most common analysis in RNA-seq and scRNA-seq data, can be expanded to ST data in the form of spatial differential gene expression, also known as SE analysis (Figure 3.4E) [82]. SE identifies genes that show significant spatial expression patterns. The most commonly used tools for SE analysis are SpatialDE [7], Trendsceek [83], and SPARK [8]. These tools can also be used to cluster genes based on their spatial expression patterns and compare with the histological images to find meaningful biological insights.

SpatialDE

SpatialDE models the expression by breaking it down into a spatial and a non-spatial (random) component. The spatial component is modeled as a sample of a Gaussian process with different kernel functions such as linear, periodic, and Gaussian. The ratio of the explained variation by these two terms is then used to determine the amount of spatial variability. It then compares the model with one with no spatial component to identify the SE genes. Figure 3.5A summarizes the SpatialDE pipeline [7].

SPARK

SPARK models count data directly using generalized linear models. It then uses Satterthwaite method and Cauchy P value combination rule to combine the p-values from the different models and identify the SE genes. It also uses a spatial smoothing term to account for the spatial correlation. SPARK provides more statistical power

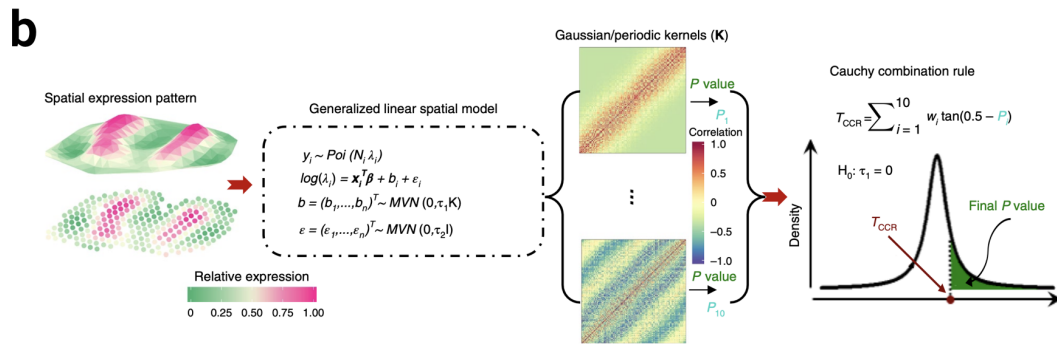
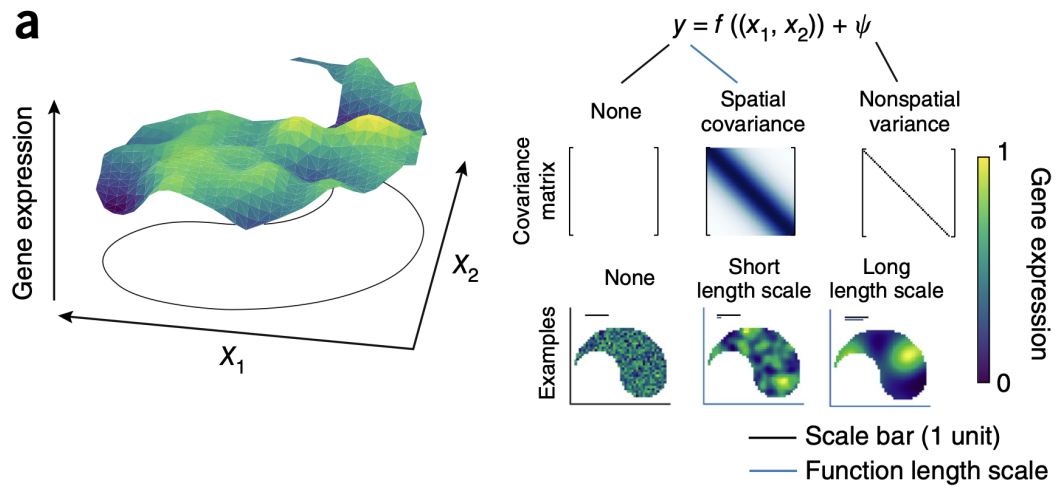


Figure 3.5: Graphical overview of the (a) SpatialDE pipeline and the (b) SPARK pipeline. Adapted from Svensson *et al* [7] and Sun *et al* [8].

than SpatialDE. However, it uses a Poisson model, for which the mean and variance are equal, which might not always be the case for ST data. Figure 3.5B summarizes the SPARK pipeline [8].

Chapter 4

Methodology

In this thesis, we present NoVaTeST, A pipeline for detecting genes with spatially variable noise. Briefly, after an initial quality check (QC) filtering, the method first transforms the gene expression count data using the Anscombe technique [26] as a variance-stabilizing transformation to reduce the effect of the mean-variance relation, followed by regressing out library size effect. Next, for each gene, a regular GP [43], and a heteroscedastic version of the GP [84] is used to get two models, one homoscedastic and one heteroscedastic, respectively. Finally, statistical model selection technique is used to identify the better fitting model and thus finding a set of genes which show significant location dependent noise.

4.1 Count Data Representation

A spatial transcriptomics data consists of N spots, with locations of the spots denoted as $X = [x_1, x_2, \dots, x_N]^T$, where $x_i = [x_{i_1}, x_{i_2}]$. Here, x_{i_1} and x_{i_2} are the horizontal and vertical coordinates of the i -th spot, respectively. We denote the relative gene expression profile (normalized and variance stabilized) for a given gene as $y = [y_1, y_2, \dots, y_N]^T$ (Figure 2.2).

4.2 Spatial Modeling of Gene Expression

Due to the effect of tissue niche and inter-cellular communication, the expression at any spot is regulated by nearby spots. Therefore, the expression at a particular spot have some dependency on the location of the spot, and the covariance of nearby spots is likely to be higher. We model this by decomposing the expression profile y in to two components – (1) $f(X)$, which captures the spatial dependency of y , and (2) ϵ , the “noise” term that captures the part of y that is not explained by $f(X)$.

$$y = f(X) + \epsilon. \quad (4.1)$$

Here, we model $f(X)$ as a sample from a GP with a constant mean function $\mu(X) = \mu_s$ and covariance kernel $K(X, X') = \sigma_s^2 \exp(-\|X - X'\|^2/2l^2)$, i.e., the radial basis function (RBF) kernel.

$$f(X) \sim \mathcal{GP}(\mu_s, K(X, X')). \quad (4.2)$$

The RBF kernel results in higher correlation for nearby spots as $\|X - X'\|^2$ is lower. The two parameters of the RBF kernel, namely the kernel variance σ_s^2 and the lengthscale parameter l , control the extent and smoothness of $f(X)$, respectively. Since the number of spots is discrete, the Gaussian process boils down to a multivariate Gaussian distribution with mean vector $\mu = [\mu_s, \mu_s, \dots, \mu_s]^T$ and covariance matrix Σ_s , where the (i, j) th entry is given by the RBF kernel, $\Sigma_{s(i,j)} = \sigma_s^2 \exp(-\|x_i - x_j\|^2/2l^2)$. Hence, we can write

$$f(X) \sim \mathcal{N}(\mu, \Sigma_s). \quad (4.3)$$

The second term, ϵ , is the “noise” term that captures the part of y that is not explained by the RBF kernel, which includes technical noise as well as biological noise, cell-type variation, phenotypical variation, etc. Existing tools to detect spatially variable genes assume the noise to be independent and identically distributed Gaussian. However, if the sample being analyzed is heterogeneous in terms of cell-type or phenotype (e.g., part of the sample being affected by a disease), this assumption might not hold for some genes. Therefore, we model the noise term to be a sample from a Gaussian with zero mean and covariance matrix $\Sigma_n = \text{diag}([\sigma_{n1}^2, \sigma_{n2}^2, \dots, \sigma_{nN}^2])$, that is,

$$\epsilon \sim \mathcal{N}(0, \Sigma_n). \quad (4.4)$$

Combining equations 4.1, 4.3 and 4.4, we get the likelihood of y in terms of the spatial coordinates X and the parameters μ_s, σ_s^2, l , and Σ_n ,

$$P(y \mid X, \mu_s, \sigma_s^2, l, \Sigma_n) = \mathcal{N}(y, \Sigma_s + \Sigma_n). \quad (4.5)$$

The above noise model where the noise variance is location-dependent is called a *heteroscedastic* noise, and the GP model with heteroscedastic noise is known as a heteroscedastic Gaussian Process. Usually, a second independent GP is used to model the location-dependent noise variance and thus calculate the model parameters. The posterior distribution of the HGP regression model can then be calculated by conditioning the prior HGP with the calculated parameters and training observations $\mathcal{D} = \{y_i, x_i\}_{i=1}^N$. This distribution will also be a multivariate Gaussian with mean μ_p and covariance Σ_p given by

$$\mu_p = \mu + \Sigma_s (\Sigma_s + \Sigma_n)^{-1} (y - \mu) \quad (4.6)$$

$$\Sigma_p = \Sigma_s - \Sigma_s^\top (\Sigma_s + \Sigma_n)^{-1} \Sigma_s. \quad (4.7)$$

In our case, we need to model tens of thousands of genes. Hence we adopt and approximate method proposed by Urban *et al.* [85] which provides a fast way to estimate the posterior mean and variance of the HGP model using two independent GPs. The method is briefly described below.

First, the parameters of a regular *homoscedastic* GP with $\Sigma_n = \sigma_n^2 \mathcal{I}_N$ is estimated by minimizing the negative log-likelihood (from equation 4.5) on the given observations \mathcal{D} ,

$$\begin{aligned} L_1 &= -\log P(y \mid X, \mu_s, \sigma_s^2, l, \sigma_n^2) \\ &= \frac{1}{2} (y - \mu)^\top (\Sigma_s + \sigma_n^2 \mathcal{I}_N)^{-1} (y - \mu) + \frac{1}{2} \log |\Sigma_s + \sigma_n^2 \mathcal{I}_N| + \frac{N}{2} \log 2\pi, \end{aligned}$$

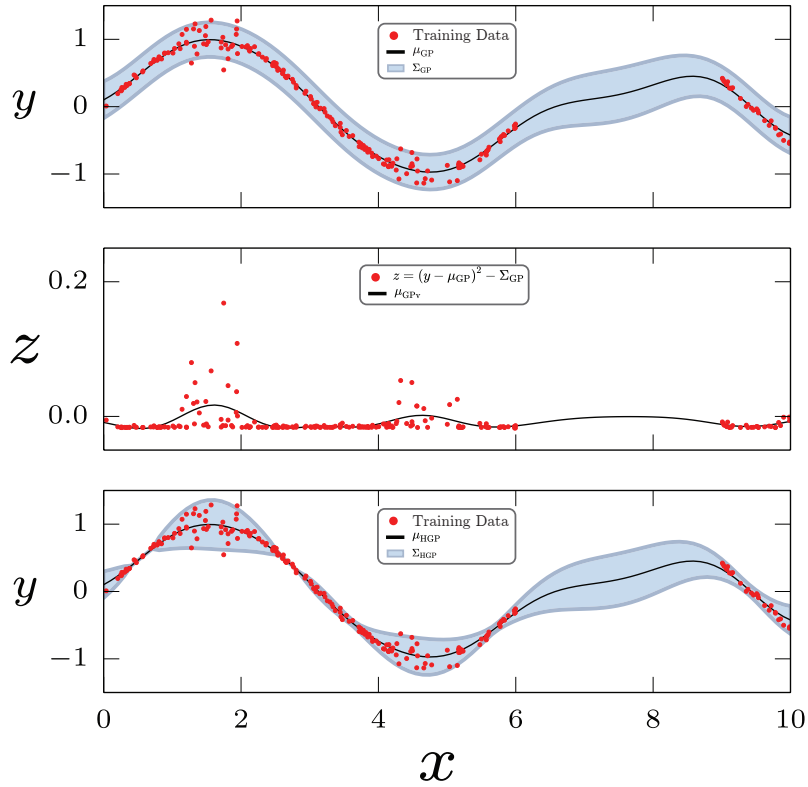


Figure 4.1: Graphical overview of the HGP model fitting process used in this thesis.

using gradient-based Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [86]. These parameters are then used to calculate the posterior mean μ_{GP} and covariance Σ_{GP} using equations 4.6 and 4.7. Next, a new dataset $\mathcal{D}_v = \{z_i, x_i\}_{i=1}^N$ is formed, where

$$z_i = (y_i - \mu_{\text{GP},i})^2 - \Sigma_{\text{GP}(i,i)}$$

is the difference between the one-sample empirical variance at the i th spot and the variance for the i th spot. After that, a second homoscedastic GP with constant zero mean and a separate covariance matrices (Σ_{sv} and $\Sigma_{nv} = \sigma_{nv}^2 \mathcal{I}_N$) is fitted on the newly formed dataset \mathcal{D}_v by minimizing the negative log-likelihood

$$L_2 = \frac{1}{2} z^\top (\Sigma_{sv} + \sigma_{nv}^2 \mathcal{I}_N)^{-1} z + \frac{1}{2} \log |\Sigma_{sv} + \sigma_{nv}^2 \mathcal{I}_N| + \frac{N}{2} \log 2\pi,$$

where $z = [z_1, z_2, \dots, z_N]^T$. Finally, denoting the posterior predictive mean of the second GP regression model as $\mu_{\text{GP}v}$, the mean μ_{HGP} and variance σ_{HGP}^2 of the HGP regression model for y is estimated by combining the two heteroscedastic GPs according to

$$\begin{aligned} \mu_{\text{HGP}} &= \mu_{\text{GP}} \text{ and} \\ \sigma_{\text{HGP}}^2 &= \max(0, \text{diag}(\Sigma_{\text{GP}}) + \mu_{\text{GP}v}). \end{aligned}$$

Note that for this method, both the homoscedastic GP and the (approximate) heteroscedastic GP models have same mean, and only the variance is refined for the HGP model. Figure 4.1 summarizes the steps for the HGP model described above.

4.3 Statistical Test to Detect Noisy Genes

To identify genes with significantly spatially variable noise variance, we compare the NLPD of the regular homoscedastic model $\text{NLPD}_{\text{Homo}}$ with that of the heteroscedastic model $\text{NLPD}_{\text{Hetero}}$. The negative log predictive densities on a test dataset $\mathcal{D}_T = \{y'_i, x'_i\}_{i=1}^{N_T}$ for the two models are calculated using the equation

$$\text{NLPD} = -\frac{1}{N} \sum_{i=1}^N \log \mathcal{N}(y'_i \mid \mu_{T,i}, \sigma_{T,i}^2) \quad (4.8)$$

$$= \frac{1}{N} \sum_{i=1}^N \left(\log 2\pi\sigma_{T,i}^2 + \frac{(y'_i - \mu_{T,i})^2}{2\sigma_{T,i}^2} \right), \quad (4.9)$$

where $\mu_{T,i}$ and $\sigma_{T,i}^2$ are the predicted mean and variance of the models at the i th spot on the test dataset \mathcal{D}_T .

Since a lower NLPD implies a better model fitting, the genes for which the NLPD for the heteroscedastic model is significantly lower than the regular model will have location-dependent noise variance and hence denoted as “*noisy*” genes in this paper. Therefore, a statistical test on the difference between the NLPDs of homoscedastic and heteroscedastic models can be used to identify such genes. However, as the background distribution of this difference is unknown, we cannot use a z-test or t-test because they assume a normal distribution [87]. Therefore, we resort to a non-parametric test. Specifically, Wilcoxon signed-rank test [88] was used on a set of paired NLPD values obtained by repeating the model fitting process ten times with a different random splitting of the spots into 90% training and 10% testing set (Figure 4.2). We also used the Benjamini-Hochberg method [89] to correct the false discovery rate (FDR).

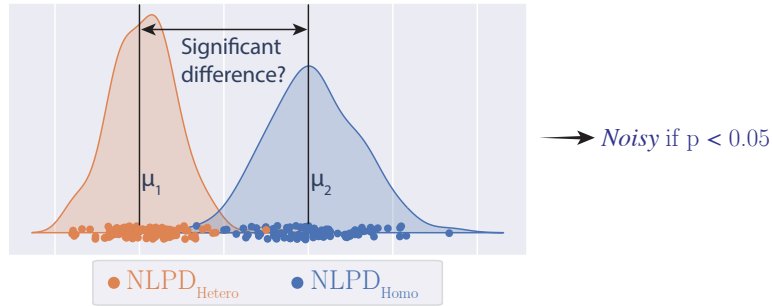


Figure 4.2: Comparing the NLPD values of the two models generated from multiple trials to see if there is a statistically significant difference between the two models. A p -value (obtained using Wilcoxon signed rank test) less than 0.05 indicates that the heteroscedastic model provides better fit than the homoscedastic model, thus the expression is *noisy*.

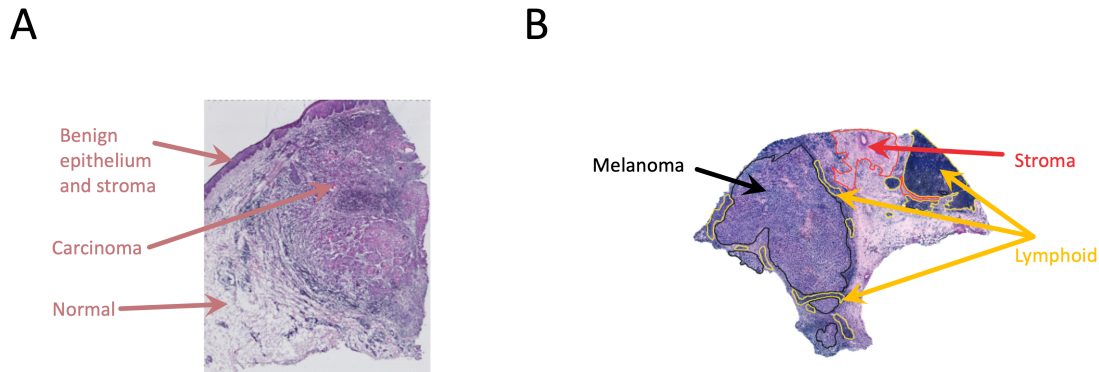


Figure 4.3: H&E stained images of the datasets used. (A) Annotated tissue H&E stained image for the 10x Visium squamous cell carcinoma dataset. (B) Tissue H&E stained image for the ST cutaneous malignant melanoma dataset. The histopathological annotations - melanoma (black), lymphoid (yellow), and stroma (red) - are provided by the authors of the dataset.

4.4 Clustering

The goal of the clustering algorithm is to cluster *noisy* genes based on similar gene expression variance patterns. For this, we defined a heuristic distance function between the predicted variance of two genes. To calculate this distance, first the predicted noise variance for the two genes are binarized, i.e., converted to 0 or 1, by comparing the values to a threshold obtained from Otsu’s method [90]. Next, the Jaccard similarity index [91], J_I , is calculated between the binarized gene variance. Finally, the distance between the variance of the two genes is defined as $J_D = 1 - J_I$. Bottom-up (agglomerative) hierarchical clustering [92] is performed using this J_D as the distance metric. The cluster representative is calculated as the average gene expression variance (averaged over cluster members).

4.5 Dataset Description

We have used two ST datasets in this study to validate our pipeline. The first one is a 10x Visium[1] data downloaded from the study of Ji *et al.* [93] (GEO: GSE144240), which contains single cell transcriptomics data (scRNA-seq) as well as ST data from ten different patients with squamous cell carcinoma. We used the filtered ST data of patient 6 replicate 1 provided by the author (from the file `GSE144239_ST_Visium_counts.txt.gz`), which contained gene expression counts from 17736 genes expressed across 3650 spots. Additionally, we filtered out the ERCC genes and mitochondrial genes, as well as practically unobservable genes that had a total count less than three across all the spots. After filtering, we were left with expression of 15733 genes across 3650 spots, i.e., an expression matrix of shape 3650×15733 and a spatial matrix of shape 3650×2 . This sample contain skin tissue slice of the squamous cell carcinoma along with patient-matched normal adjacent skin samples, as shown in the tissue H&E stained image of Figure 4.3(A).

The second dataset was downloaded from the Spatial Research website from the

study of Thrane *et al.* [94], which contains spatial transcriptomics technology [38] data from four different patients with stage III cutaneous malignant melanoma. We used the data of patient 1 replicate 1 (from the file `ST_mel1_rep1_counts.tsv`), which contained gene expression counts from 15666 genes expressed across 279 spots. Similar to the previous dataset, we filtered out the ERCC genes and mitochondrial genes, as well as practically unobservable genes that had a total count less than three. After filtering, we were left with gene expression data of 13088 genes across 279 spots, i.e., an expression matrix of shape 279×13088 and a spatial matrix of shape 279×2 . This melanoma lymph node biopsy sample contains three distinct regions – melanoma, stroma, and lymphoid, as shown in the tissue H&E stained image of Figure 4.3(B).

Chapter 5

Results from Squamous Cell Carcinoma Data

NoVaTeST detects genes enriched in cancer related pathways in squamous cell carcinoma data

The first dataset contains gene expression from the skin tissue sample from a patient with squamous cell carcinoma and patient-matched normal adjacent skin sample collected using the Visium technology [93]. After QC filtering and normalization, the dataset contained normalized count data of 15725 genes expressed across 3650 spots.

Applying our model on this data, we identified 771 *noisy* genes at an FDR level of 5%. We performed enrichment analysis of the *noisy* genes to detect all statistically significant enrichment terms and hierarchically clustered them into a tree based on similarity of their gene memberships using Metascape [95]. The resulting top cluster representative enrichment terms are shown in Figures 5.1(A). The terms associated with cancer and immuno-response are marked bold in Figure 5.1(A). Interestingly, the *noisy* genes for this dataset are mostly associated with cancer-related pathways — not only cancer development [96], tumor progression [97], [98], angiogenesis [99], but also cancer immuno-response [100]. This result makes sense since tumor micro-environment are highly heterogeneous due to their uncontrolled growth and thus the genes related to cancer and immuno-response are likely to have different noise variance in the tumor region compared to the tumor-adjacent healthy region.

The *noisy* genes were clustered based on a heuristic approach (details provided in the methods section) to find groups of genes that show visually similar noise-variance patterns. The cluster representative, which is the gene expression noise variance averaged over cluster members, for the five identified clusters in the *noisy* genes are shown in Figure 5.1(B). The distribution of the 771 *noisy* genes, shown in Figure 5.1(C), reveals that most of the genes belong to clusters 1, 2, and 3. Comparing the cluster representatives and the tissue H&E stained image, we see that for cluster 1, the variance is high mainly in the stroma and tumor-adjacent healthy region, whereas, for cluster 2, the variance is high mainly in the tumor region. For cluster 3, the variance is high along the thin line on the upper-left of the H&E image, which corresponds to the benign squamous epithelium and stroma region (see Fig. 4D of

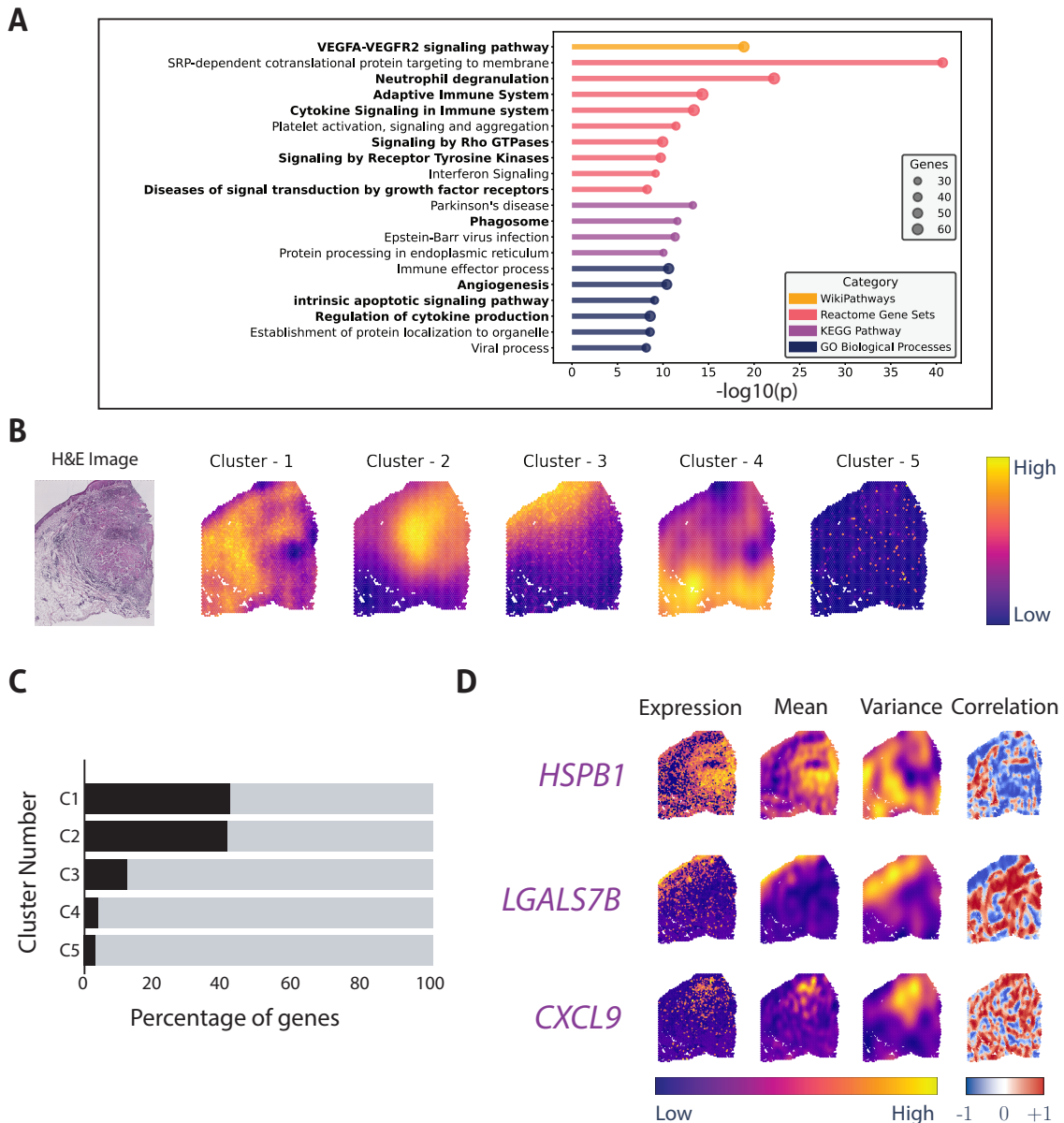


Figure 5.1: Results obtained from squamous cell carcinoma data using NoVaTeST. (A) Top cluster representative enriched terms for the detected *noisy* genes using Metascape. Terms associated with cancer and immuno-response are shown in bold font. (B) The average gene expression noise variance, averaged over the cluster members, for the five identified clusters along with the tissue H&E stained image. (C) Distribution of the *noisy* genes among the identified clusters. (D) Gene expression (log scale) and corresponding modelled mean and variance for three representative genes from the first three clusters. Also plotted is the spatial Spearman correlation between the mean and variance of the model, where the correlation for a spot is computed by considering twelve nearby spots.

Erickson *et al.*[101])

The log-expression and the mean and variance of the heteroscedastic model for three representative genes, *HSPB1*, *LGALS7B*, and *CXCL9*, selected from cluster 1, 2, and 3, respectively, are shown in Figure 5.1(D). *HSPB1*, which shows high mean

in the tumor region and high variance primarily in the adjacent healthy and the stroma region of the H&E image, is strongly associated with tumor metastasis and metastatic colonization [102]–[106]. *LGALS7B* is known to play role in several types of carcinomas [107], and the expression shows high mean in the squamous epithelium region but high variance in the stroma region. *CXCL9* from cluster 2 shows high variance in the tumor affected region and is involved in T-cell trafficking [108]. These results indicate that genes with significant spatially variable noise detected by our method indeed carry biologically significant results, and thus the importance of a generalized model.

For the selected genes mentioned above, we calculated the spatial correlation between the estimated mean and estimated variance by calculating the Spearman correlation over the twelve nearest neighbors of each spot. The results are shown in Figure 5.1(D). For *HSPB1*, we see that the correlation is overall low for almost all the spots. For *LGALS7B*, on the other hand, the correlation is low for the spots where the variance is high (and vice versa). These results are indication that these genes are not an artifact of the mean-variance relationship, and thus provide complimentary information to existing methods (see Discussion).

Chapter 6

Results from Cutaneous Malignant Melanoma Data

NoVaTeST identifies distinct noise variance patterns in cutaneous malignant melanoma data

The second dataset contains the expression of 13088 genes expressed across 279 spots from a melanoma lymph node biopsy sample collected using spatially resolved transcriptomics technology [94]. The manual annotation of the H&E image by Thrane *et. al.* [94] shows three distinct regions — melanoma, stroma, and lymphoid (Figure 6.1(A)).

Applying our model, we find 472 *noisy* genes at an FDR level of 5%. We first clustered the *noisy* genes into three clusters based on the proposed heuristic approach to elucidate the biological processes impacted by these genes. The gene expression noise variances averaged over cluster members for the three identified clusters are shown in Figure 6.1(B). The high variance regions of the cluster representatives of clusters 1 and 3 overlap with the manually annotated the lymphoid regions of the H&E image. On the other hand, genes in cluster 2 show high variance in melanoma region. These results point to the importance of the heteroscedastic model, as variance in phenotypically different regions show different patterns of noise variance.

Next, enrichment analysis was performed for each cluster to identify the top enriched terms. Then, the top enriched terms were hierarchically clustered based on gene memberships' similarity. The top enriched cluster representative terms for the melanoma dataset across the three identified clusters are shown in Figure 6.1(C). Notably, genes in cluster 1, which show high variance in the lymphoid region, are enriched in “supramolecular fiber organization”, as well as immuno-response related terms “inflammatory response” and “adaptive immune response”. Genes in cluster 3, which also show high variance in the lymphoid region, results in only two clustered enriched terms (as there are only 24 genes), one of which is “innate immune response”. Lastly, genes in cluster 2 show high noise variance in the melanoma region, and is enriched in the GO term “melanocyte differentiation”. These results indicate that genes with similar noise-variance pattern might perform similar operations.

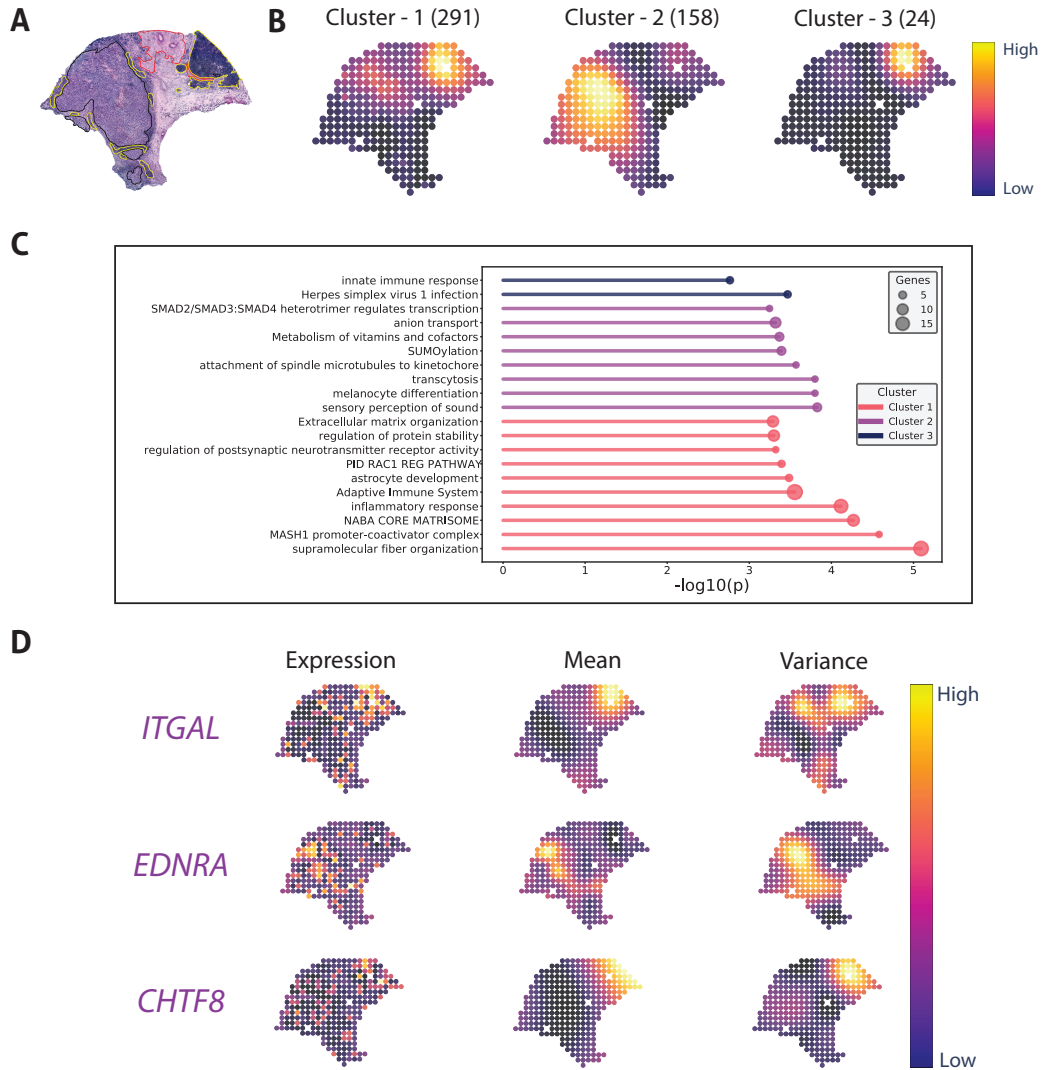


Figure 6.1: Results obtained from cutaneous malignant melanoma data using No-VaTeST. (A) Tissue H&E stained image of the sample with histopathological annotations - melanoma (black), lymphoid (yellow), and stroma (red), adapted from Thrane *et. al.* 2018 [94]. (B) The average gene expression noise variance, averaged over the cluster members, for the three identified clusters. The numbers inside the parentheses denote the number of genes in each cluster. (C) Top cluster representative enriched terms for the detected *noisy* genes in each detected cluster using Metascape. (D) Gene expression (log scale) and corresponding modelled mean and variance for three representative genes from the three clusters.

The log-expression and the mean and variance of the heteroscedastic model for three representative genes, *ITGAL*, *EDNRA*, and *CHTF8*, selected from cluster 1, 2, and 3, respectively, are shown in Figure 6.1(D). Abnormal expression of *ITGAL* is linked with immune regulation [109]. High expression of *EDNRA* is linked with metastasis [110], which in this sample is also the region where the variance is high. The results indicate that the heteroscedastic model can be used to identify genes with abnormal expression in specific regions of the tissue.

Chapter 7

Discussion

In this thesis, we have developed NoVaTeST, a generalized framework for gene expression variation analysis in ST. Specifically, the contributions of this thesis are as follows:

- We have developed a more general spatial gene expression modelling in ST data that uses heteroscedastic Gaussian process.
- We have developed a rigorous statistical testing pipeline using Wilcoxon signed rank test and FDR correction to identify genes with location-dependent noise variance.
- We have developed a method to cluster genes with similar noise variance patterns using a custom distance function/
- We have applied our method to two different cancer ST datasets and show that the detected noisy genes provide complimentary information to existing techniques, and provide important biological insights.

An important first step for ST data analysis is modeling the gene expression as a function of location. There are two main types of uncertainty to consider while modeling, namely *epistemic* uncertainty and *aleatoric* uncertainty. The epistemic uncertainty is the variability of the model output due to the randomness of the model itself. In case of modeling ST data, this refers to the uncertainty (or confidence) of gene expression prediction given spatial location. On the other hand, the aleatoric uncertainty is the variability of the model output due to the randomness or noise present in the data. In this case, aleatoric variability is the part of gene expression having no spatial variation, also referred to as noise. The aleatoric uncertainty can be further classified into *homoscedastic* or input independent constant noise, and *heteroscedastic* or input dependent noise.

While existing methods use Bayesian framework to capture the epistemic uncertainty, they use a constrained assumption of homoscedastic noise to model the aleatoric uncertainty. The proposed pipeline NoVaTeST can detect the presence of heteroscedastic uncertainty, that is, *noisy* genes that show significant spatial variation in noise variance.

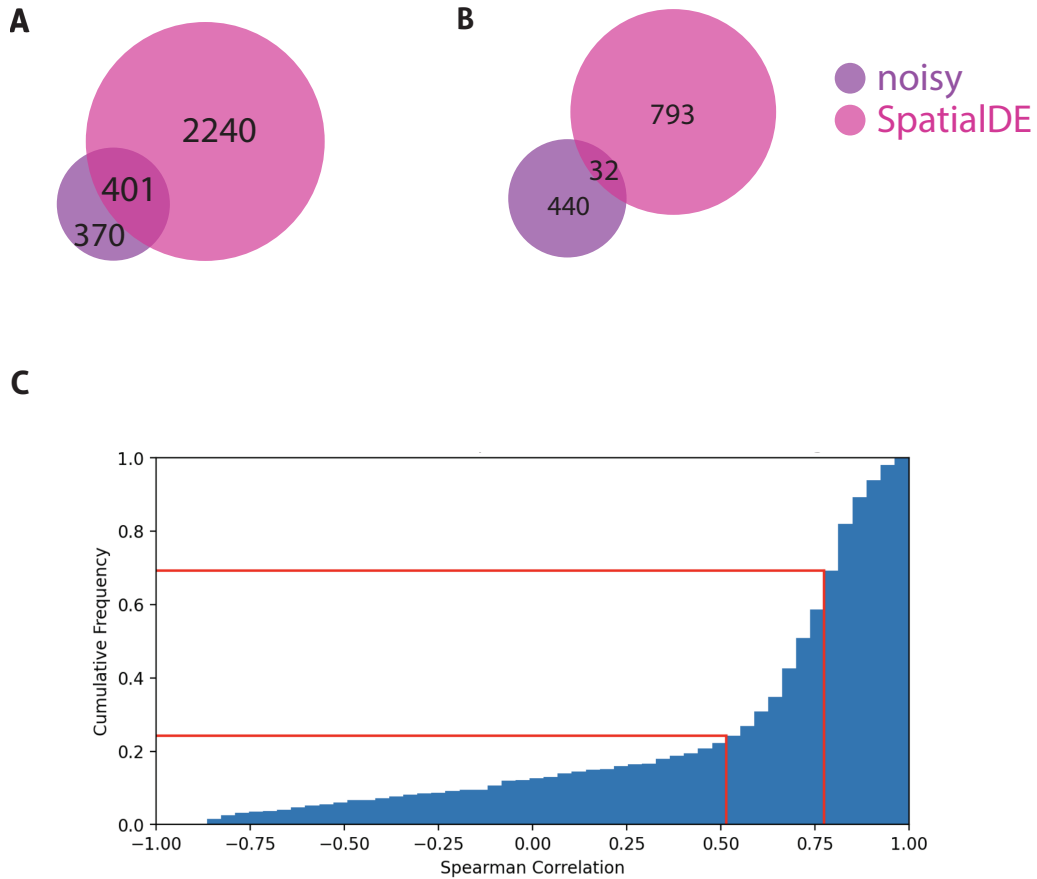


Figure 7.1: Analysis to check the existence and extent of mean-variance artifacts in the datasets. (A) Venn diagram showing the overlap between genes detected by SpatialDE and NoVaTeST for the carcinoma dataset. (B) Venn diagram showing the overlap between genes detected by SpatialDE and NoVaTeST for the melanoma dataset. (C) Cumulative frequency of Spearman correlation between the estimated mean and variance for the common genes (genes detected by both SpatialDE and NoVaTeST).

To check whether the *noisy* genes are an artifact of mean-variance relationship, we compared the *noisy* gene list to that detected by SpatialDE [7], a tool to detect genes with significant spatial mean expression patterns. If the detected *noisy* genes were indeed an artifact, then the *noisy* genes would also have been detected by SpatialDE. However, only about 50% of the *noisy* genes from the carcinoma data were common with the 2641 genes detected by SpatialDE (Figure 7.1(A)). The rest 370 *noisy* genes were not detected by SpatialDE, meaning the expression of these genes do not show any significant spatial pattern, but their noise-variance display a location-dependent pattern. For the melanoma dataset, only 32 out of the 470 *noisy* genes overlap with the 825 genes detected by SpatialDE (Figure 7.1(B)). These results demonstrate that the detected noisy genes are not an artifact of the mean-variance relation, rather they provide complimentary information to existing methods like SpatialDE.

Further analysis, however, reveal that the detected *noisy* genes are not completely

unaffected by the mean-variance artifact, especially the *overlapping* genes detected by both NoVaTeST and SpatialDE. This can be seen from the cumulative histogram of Spearman correlation between mean and variance of the the 401 *overlapping* genes of the carcinoma dataset (Figure 7.1(C)). Around 50% or 200 genes show moderate correlation (between 0.5 and 0.75), meaning the mean and variance are somewhat correlated, *i.e.*, spots with high expression have high variance. On top of that, around 25% or 100 genes show strong correlation between mean and variance at each spot, thus an artifact of mean-variance relation. These analyses suggest a more robust variance-stabilizing transformation should be adopted.

Gaussian processes are inherently computationally expensive. Moreover, a heteroscedastic GP, where the noise variance is modelled using another GP, does not have a closed form solution, and therefore we have to resort to iterative methods to fit a model. This further increases the runtime of the pipeline, which, again, is primarily due to computational expense of Gaussian processes. To combat this, we had to use an approximate GP where we assumed that the mean of the homoscedastic model is same as that of the heteroscedastic model.

Chapter 8

Conclusion

In this thesis, we propose the NoVaTeST pipeline that uses a more generalized modeling for ST data. The pipeline can also detect genes that show statistically significant spatial variability in terms of noise variance. Analysis on two different cancer ST datasets show the detected *noisy* genes (genes that show significant heteroscedastic noise) in squamous cell carcinoma data are mostly associated with cancer and immuno-response related pathways. On the other hand, the noisy genes in cutaneous malignant melanoma data form three clusters in terms of noise-variance pattern, and these pattern overlap with manual annotation of different phenotypical conditions in the H&E image. These results are consistent with our initial hypothesis regarding the noisy genes and provide evidence of the biological significance of the *noisy* genes. Moreover, we have shown that the pipeline provides complimentary information to existing techniques such as SpatialDE. In future, we want to explore more heteroscedastic models that are computationally less expensive than the one we used in this thesis, and can model the count data directly. Moreover, we want to investigate the biological significance of the detected *noisy* genes in more datasets. Additionally, we want to investigate the effect of non-uniform cellular densities on the NoVaTeST pipeline, similar to MERINGUE [111]. Finally, since ST data can be used to interpret cell-cell and gene-gene interactions, we want to investigate whether the noise variance pattern and the detected *noisy* genes reveal any novel information regarding cell-cell communications.

Bibliography

- [1] *Visium spatial gene expression*, <https://www.10xgenomics.com/products/spatial-gene-expression>, Accessed: 2022-09-01. [Online]. Available: <https://www.10xgenomics.com/products/spatial-gene-expression>.
- [2] L. Lucero, *Answering your questions about visium spatial gene expression for FFPE*, <https://www.10xgenomics.com/blog/answering-your-questions-about-visium-spatial-gene-expression-for-ffpe-spotlight-on-sample-prep>, Accessed: 2022-06-30. [Online]. Available: <https://www.10xgenomics.com/blog/answering-your-questions-about-visium-spatial-gene-expression-for-ffpe-spotlight-on-sample-prep>.
- [3] E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, and L. Cai, “Single-cell in situ rna profiling by sequential hybridization,” *Nature methods*, vol. 11, no. 4, pp. 360–361, 2014.
- [4] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang, “Spatially resolved, highly multiplexed rna profiling in single cells,” *Science*, vol. 348, no. 6233, aaa6090, 2015.
- [5] S. Shah, E. Lubeck, W. Zhou, and L. Cai, “In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus,” *Neuron*, vol. 92, no. 2, pp. 342–357, 2016.
- [6] Y. Wang, S. Ma, and W. L. Ruzzo, “Spatial modeling of prostate cancer metabolic gene expression reveals extensive heterogeneity and selective vulnerabilities,” *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.
- [7] V. Svensson, S. A. Teichmann, and O. Stegle, “Spatialde: Identification of spatially variable genes,” *Nature methods*, vol. 15, no. 5, pp. 343–346, 2018.
- [8] S. Sun, J. Zhu, and X. Zhou, “Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies,” *Nature methods*, vol. 17, no. 2, pp. 193–200, 2020.
- [9] J. H. McCulloch, “Miscellanea: On heteros*edasticity,” *Econometrica*, vol. 53, no. 2, pp. 483–483, 1985, issn: 00129682, 14680262. [Online]. Available: <http://www.jstor.org/stable/1911250> (visited on 06/30/2022).
- [10] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, “Most likely heteroscedastic gaussian process regression,” in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 393–400.

- [11] M. D. Blackledge, N. Tunariu, F. Zugni, R. Holbrey, M. R. Orton, A. Ribeiro, J. C. Hughes, E. D. Scurr, D. J. Collins, M. O. Leach, *et al.*, “Noise-corrected, exponentially weighted, diffusion-weighted mri (nicedwi) improves image signal uniformity in whole-body imaging of metastatic prostate cancer,” *Frontiers in oncology*, vol. 10, p. 704, 2020.
- [12] P. H. Bradley and K. S. Pollard, “Proteobacteria explain significant functional variability in the human gut microbiome,” *Microbiome*, vol. 5, no. 1, pp. 1–23, 2017.
- [13] S. Park, H. Xu, and T. H. Hwang, “Gaussian process based heteroscedastic noise modeling for tumor mutation burden prediction from whole slide images,” *bioRxiv*, p. 554 261, 2019.
- [14] C. Brooks, S. P. Burke, and G. Persand, *Benchmarks and the accuracy of garch model estimation*, 2001.
- [15] C. T. Brownlees, R. F. Engle, and B. T. Kelly, “A practical guide to volatility forecasting through calm and storm,” *Available at SSRN 1502915*, 2011.
- [16] J. S. Liu and J. S. Liu, *Monte Carlo strategies in scientific computing*. Springer, 2001, vol. 10.
- [17] M. Bauza and A. Rodriguez, “A probabilistic data-driven model for planar pushing,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2017, pp. 3008–3015.
- [18] A. J. Smith, M. AlAbsi, and T. Fields, “Heteroscedastic gaussian process-based system identification and predictive control of a quadcopter,” in *2018 AIAA Atmospheric Flight Mechanics Conference*, 2018, p. 0298.
- [19] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte, “Heteroscedastic gaussian processes for data fusion in large scale terrain modeling,” in *2010 IEEE International Conference on Robotics and Automation*, IEEE, 2010, pp. 3452–3459.
- [20] P. Kou, D. Liang, L. Gao, and J. Lou, “Probabilistic electricity price forecasting with variational heteroscedastic gaussian process and active learning,” *Energy Conversion and Management*, vol. 89, pp. 298–308, 2015.
- [21] M. Lázaro-Gredilla, M. K. Titsias, J. Verrelst, and G. Camps-Valls, “Retrieval of biophysical parameters with heteroscedastic gaussian processes,” *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 4, pp. 838–842, 2013.
- [22] I. A. Almosallam, M. J. Jarvis, and S. J. Roberts, “Gpz: Non-stationary sparse gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts,” *Monthly Notices of the Royal Astronomical Society*, vol. 462, no. 1, pp. 726–739, 2016.
- [23] V. Svensson, *Variance stabilizing scrna-seq counts*, <https://www.nxn.se/valent/2017/10/15/variance-stabilizing-scrna-seq-counts>, Oct. 2017.
- [24] S. Nakagawa and H. Schielzeth, “The mean strikes back: Mean–variance relationships and heteroscedasticity,” *Trends in Ecology and Evolution*, vol. 27, no. 9, 2012.

- [25] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Nature Precedings*, pp. 1–1, 2010.
- [26] F. J. Anscombe, “The transformation of poisson, binomial and negative-binomial data,” *Biometrika*, vol. 35, no. 3/4, pp. 246–254, 1948.
- [27] N. Eling, A. C. Richard, S. Richardson, J. C. Marioni, and C. A. Vallejos, “Correcting the mean-variance dependency for differential variability testing using single-cell rna sequencing data,” *Cell systems*, vol. 7, no. 3, pp. 284–294, 2018.
- [28] Q. Li, X. Zhang, and R. Ke, “Spatial transcriptomics for tumor heterogeneity analysis,” *Frontiers in Genetics*, vol. 13, 2022.
- [29] W. Dinalankara and H. C. Bravo, “Gene expression signatures based on variability can robustly predict tumor progression and prognosis,” *Cancer informatics*, vol. 14, CIN–S23862, 2015.
- [30] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.
- [31] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *et al.*, “Molecular biology of the cell,” *Scandinavian Journal of Rheumatology*, vol. 32, no. 2, 2003.
- [32] Z. Wang, M. Gerstein, and M. Snyder, “Rna-seq: A revolutionary tool for transcriptomics,” *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.
- [33] W. Klug, M. Cummings, C. Spencer, and M. Palladino, “Concepts of genetics 10th edition,” *Harlow: Pearson Education Limited*, 2014.
- [34] W. M. Freeman, S. J. Walker, and K. E. Vrana, “Quantitative rt-pcr: Pitfalls and potential,” *Biotechniques*, vol. 26, no. 1, pp. 112–125, 1999.
- [35] E. TAUB FLOYD, J. M. DeLEO, and E. B. Thompson, “Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated rnas,” *Dna*, vol. 2, no. 4, pp. 309–327, 1983.
- [36] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyren, “Real-time dna sequencing using detection of pyrophosphate release,” *Analytical biochemistry*, vol. 242, no. 1, pp. 84–89, 1996.
- [37] J. Eberwine, J.-Y. Sul, T. Bartfai, and J. Kim, “The promise of single-cell sequencing,” *Nature methods*, vol. 11, no. 1, pp. 25–27, 2014.
- [38] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, *et al.*, “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics,” *Science*, vol. 353, no. 6294, pp. 78–82, 2016.
- [39] X. Wang, W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan, J. Liu, *et al.*, “Three-dimensional intact-tissue sequencing of single-cell transcriptional states,” *Science*, vol. 361, no. 6400, eaat5691, 2018.
- [40] X. Chen, Y.-C. Sun, G. M. Church, J. H. Lee, and A. M. Zador, “Efficient in situ barcode sequencing using padlock probe-based baristaseq,” *Nucleic acids research*, vol. 46, no. 4, e22–e22, 2018.

- [41] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, “Visualization of single rna transcripts in situ,” *Science*, vol. 280, no. 5363, pp. 585–590, 1998.
- [42] S. Nichterwitz, G. Chen, J. Aguila Benitez, M. Yilmaz, H. Storvall, M. Cao, R. Sandberg, Q. Deng, and E. Hedlund, “Laser capture microscopy coupled with smart-seq2 for precise spatial transcriptomic profiling,” *Nature communications*, vol. 7, no. 1, pp. 1–11, 2016.
- [43] C. E. Rasmussen, “Gaussian processes in machine learning,” in *Summer school on machine learning*, Springer, 2003, pp. 63–71.
- [44] H. Li, A. Chowdhury, G. Terejanu, A. Chanda, and S. Banerjee, “A stacked gaussian process for predicting geographical incidence of aflatoxin with quantified uncertainties,” in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015, pp. 1–4.
- [45] T. Cui, D. Pagendam, and M. Gilfedder, “Gaussian process machine learning and kriging for groundwater salinity interpolation,” *Environmental Modelling & Software*, vol. 144, p. 105 170, 2021.
- [46] J. Vanhatalo and A. Vehtari, “Sparse log gaussian processes via mcmc for spatial epidemiology,” in *Gaussian processes in practice*, PMLR, 2007, pp. 73–89.
- [47] L. S. Canas, C. H. Sudre, J. C. Pujol, L. Polidori, B. Murray, E. Molteni, M. S. Graham, K. Klaser, M. Antonelli, S. Berry, *et al.*, “Early detection of covid-19 in the uk using self-reported symptoms: A large-scale, prospective, epidemiological surveillance study,” *The Lancet Digital Health*, vol. 3, no. 9, e587–e598, 2021.
- [48] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric regression*, 12. Cambridge university press, 2003.
- [49] M. R. Emmert-Buck, R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta, “Laser capture microdissection,” *Science*, vol. 274, no. 5289, pp. 998–1001, 1996.
- [50] N. L. Simone, R. F. Bonner, J. W. Gillespie, M. R. Emmert-Buck, and L. A. Liotta, “Laser-capture microdissection: Opening the microscopic frontier to molecular analysis,” *Trends in Genetics*, vol. 14, no. 7, pp. 272–276, 1998.
- [51] J. Chen, S. Suo, P. P. Tam, J.-D. J. Han, G. Peng, and N. Jing, “Spatial transcriptomic analysis of cryosectioned tissue samples with geo-seq,” *Nature protocols*, vol. 12, no. 3, pp. 566–580, 2017.
- [52] S. Vickovic, G. Eraslan, F. Salmén, J. Klughammer, L. Stenbeck, D. Schapiro, T. Äijö, R. Bonneau, L. Bergensträhle, J. F. Navarro, *et al.*, “High-definition spatial transcriptomics for in situ tissue profiling,” *Nature methods*, vol. 16, no. 10, pp. 987–990, 2019.
- [53] N. Karaiskos, P. Wahle, J. Alles, A. Boltengagen, S. Ayoub, C. Kipar, C. Kocks, N. Rajewsky, and R. P. Zinzen, “The drosophila embryo at single-cell transcriptome resolution,” *Science*, vol. 358, no. 6360, pp. 194–199, 2017.
- [54] M. Asp, J. Bergensträhle, and J. Lundeberg, “Spatially resolved transcriptomes—next generation tools for tissue exploration,” *BioEssays*, vol. 42, no. 10, p. 1 900 221, 2020.

- [55] R. Dries, J. Chen, N. Del Rossi, M. M. Khan, A. Sistig, and G.-C. Yuan, “Advances in spatial transcriptomic data analysis,” *Genome research*, vol. 31, no. 10, pp. 1706–1718, 2021.
- [56] S. Codeluppi, L. E. Borm, A. Zeisel, G. La Manno, J. A. van Lunteren, C. I. Svensson, and S. Linnarsson, “Spatial organization of the somatosensory cortex revealed by osmfish,” *Nature methods*, vol. 15, no. 11, pp. 932–935, 2018.
- [57] J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, *et al.*, “Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region,” *Science*, vol. 362, no. 6416, eaau5324, 2018.
- [58] C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, *et al.*, “Transcriptome-scale super-resolved imaging in tissues by rna seqfish+,” *Nature*, vol. 568, no. 7751, pp. 235–239, 2019.
- [59] J. Y. Kishi, S. W. Lapan, B. J. Beliveau, E. R. West, A. Zhu, H. M. Sasaki, S. K. Saka, Y. Wang, C. L. Cepko, and P. Yin, “Saber amplifies fish: Enhanced multiplexed imaging of rna and dna in cells and tissues,” *Nature methods*, vol. 16, no. 6, pp. 533–544, 2019.
- [60] J. J. L. Goh, N. Chou, W. Y. Seow, N. Ha, C. P. P. Cheng, Y.-C. Chang, Z. W. Zhao, and K. H. Chen, “Highly specific multiplexed rna imaging in tissues with split-fish,” *Nature methods*, vol. 17, no. 7, pp. 689–693, 2020.
- [61] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008.
- [62] V. A. Traag, L. Waltman, and N. J. Van Eck, “From louvain to leiden: Guaranteeing well-connected communities,” *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.
- [63] G. Pasquini, J. E. R. Arias, P. Schäfer, and V. Busskamp, “Automated methods for cell type annotation on scrna-seq data,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 961–969, 2021.
- [64] R. Lopez, B. Li, H. Keren-Shaul, P. Boyeau, M. Kedmi, D. Pilzer, A. Jelinski, E. David, A. Wagner, Y. Addad, *et al.*, “Multi-resolution deconvolution of spatial transcriptomics data reveals continuous patterns of inflammation,” *BioRxiv*, 2021.
- [65] T. Lohoff, S. Ghazanfar, A. Missarova, N. Koulena, N. Pierson, J. A. Griffiths, E. S. Bardot, C.-H. Eng, R. C. Tyser, R. Argelaguet, *et al.*, “Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis,” *BioRxiv*, 2020.
- [66] F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, and K. De Preter, “Benchmarking of cell type deconvolution pipelines for transcriptomics data,” *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

- [67] A. Andersson, L. Larsson, L. Stenbeck, F. Salmén, A. Ehinger, S. Wu, G. Al-Eryani, D. Roden, A. Swarbrick, Å. Borg, *et al.*, “Spatial deconvolution of her2-positive breast tumors reveals novel intercellular relationships,” *bioRxiv*, 2020.
- [68] T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, A. Segerstolpe, M. Zhang, *et al.*, “Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with tangram,” *Biorxiv*, 2020.
- [69] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, A. Lomakin, V. Kedlian, M. S. Jain, J. S. Park, *et al.*, “Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics,” *BioRxiv*, 2020.
- [70] D. M. Cable, E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, and R. A. Irizarry, “Robust decomposition of cell type mixtures in spatial transcriptomics,” *Nature Biotechnology*, vol. 40, no. 4, pp. 517–526, 2022.
- [71] R. Dong and G.-C. Yuan, “Spatialdws: Accurate deconvolution of spatial transcriptomic data,” *Genome biology*, vol. 22, no. 1, pp. 1–10, 2021.
- [72] M. Elosua-Bayes, P. Nieto, E. Mereu, I. Gut, and H. Heyn, “Spotlight: Seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes,” *Nucleic acids research*, vol. 49, no. 9, e50–e50, 2021.
- [73] R. Lopez, A. Nazaret, M. Langevin, J. Samaran, J. Regier, M. I. Jordan, and N. Yosef, “A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements,” *arXiv preprint arXiv:1905.02269*, 2019.
- [74] Q. Song and J. Su, “Dstg: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence,” *Briefings in Bioinformatics*, vol. 22, no. 5, bbaa414, 2021.
- [75] B. F. Miller, F. Huang, L. Atta, A. Sahoo, and J. Fan, “Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data,” *Nature communications*, vol. 13, no. 1, pp. 1–13, 2022.
- [76] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, *et al.*, “Cell2location maps fine-grained cell types in spatial transcriptomics,” *Nature biotechnology*, vol. 40, no. 5, pp. 661–671, 2022.
- [77] D. Schapiro, H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller, “Histocat: Analysis of cell phenotypes and interactions in multiplex image cytometry data,” *Nature methods*, vol. 14, no. 9, pp. 873–876, 2017.
- [78] R. Dries, Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, *et al.*, “Giotto: A toolbox for integrative analysis and visualization of spatial expression data,” *Genome biology*, vol. 22, no. 1, pp. 1–31, 2021.

- [79] E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uyttingco, S. E. Taylor, P. Nghiem, *et al.*, “Spatial transcriptomics at subspot resolution with bayesspace,” *Nature Biotechnology*, vol. 39, no. 11, pp. 1375–1384, 2021.
- [80] B. Chidester, T. Zhou, S. Alam, and J. Ma, “Spicemix: Integrative single-cell spatial modeling of cell identity,” *bioRxiv*, pp. 2020–11, 2022.
- [81] S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise, “Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. 4290–4295, 2016.
- [82] J. Zhu, S. Sun, and X. Zhou, “Spark-x: Non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies,” *Genome biology*, vol. 22, no. 1, pp. 1–25, 2021.
- [83] D. Edsgård, P. Johnsson, and R. Sandberg, “Identification of spatial expression trends in single-cell gene expression data,” *Nature methods*, vol. 15, no. 5, pp. 339–342, 2018.
- [84] P. Goldberg, C. Williams, and C. Bishop, “Regression with input-dependent noise: A gaussian process treatment,” *Advances in neural information processing systems*, vol. 10, 1997.
- [85] S. Urban, M. Ludersdorfer, and P. Van Der Smagt, “Sensor calibration and hysteresis compensation with heteroscedastic gaussian processes,” *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6498–6506, 2015.
- [86] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.
- [87] R. C. Sprinthal, *Basic statistical analysis*. Allyn & Bacon, 2003.
- [88] R. F. Woolson, “Wilcoxon signed-rank test,” *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.
- [89] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
- [90] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [91] P. Jaccard, “The distribution of the flora in the alpine zone. 1,” *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [92] L. Rokach and O. Maimon, “Clustering methods,” in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 321–352.
- [93] A. L. Ji, A. J. Rubin, K. Thrane, S. Jiang, D. L. Reynolds, R. M. Meyers, M. G. Guo, B. M. George, A. Mollbrink, J. Bergensträhle, *et al.*, “Multi-modal analysis of composition and spatial architecture in human squamous cell carcinoma,” *Cell*, vol. 182, no. 2, pp. 497–514, 2020.
- [94] K. Thrane, H. Eriksson, J. Maaskola, J. Hansson, and J. Lundeberg, “Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage iii cutaneous malignant melanoma,” *Cancer research*, vol. 78, no. 20, pp. 5970–5979, 2018.

- [95] Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda, “Metascape provides a biologist-oriented resource for the analysis of systems-level datasets,” *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.
- [96] F. Mollinedo, “Neutrophil degranulation, plasticity, and cancer metastasis,” *Trends in immunology*, vol. 40, no. 3, pp. 228–242, 2019.
- [97] M. Lee and I. Rhee, “Cytokine signaling in tumor progression,” *Immune network*, vol. 17, no. 4, p. 214, 2017.
- [98] M. Yamazaki, S. Maruyama, T. Abé, M. Tsuneki, H. Kato, K. Izumi, J.-i. Tanuma, J. Cheng, and T. Saku, “Rac1-dependent phagocytosis of apoptotic cells by oral squamous cell carcinoma cells: A possible driving force for tumor progression,” *Experimental Cell Research*, vol. 392, no. 1, p. 112 013, 2020.
- [99] C. S. Abhinand, R. Raju, S. J. Soumya, P. S. Arya, and P. R. Sudhakaran, “Vegf-a/vegfr2 signaling network in endothelial cells relevant to angiogenesis,” *Journal of cell communication and signaling*, vol. 10, no. 4, pp. 347–354, 2016.
- [100] S. Fulda and K.-M. Debatin, “Extrinsic versus intrinsic apoptosis pathways in anticancer chemotherapy,” *Oncogene*, vol. 25, no. 34, pp. 4798–4811, 2006.
- [101] A. Erickson, M. He, E. Berglund, M. Marklund, R. Mirzazadeh, N. Schultz, L. Kvastad, A. Andersson, L. Bergenstråhle, J. Bergenstråhle, *et al.*, “Spatially resolved clonal copy number alterations in benign and malignant tissue,” *Nature*, vol. 608, no. 7922, pp. 360–367, 2022.
- [102] B. Gibert, B. Eckel, V. Gonin, D. Goldschneider, J. Fombonne, B. Deux, P. Mehlen, A. Arrigo, P. Clezardin, and C. Diaz-Latoud, “Targeting heat shock protein 27 (hspb1) interferes with bone metastasis and tumour formation in vivo,” *British journal of cancer*, vol. 107, no. 1, pp. 63–70, 2012.
- [103] P. Lemieux, S. Oesterreich, J. Lawrence, P. Steeg, S. Hilsenbeck, J. Harvey, and S. Fuqua, “The small heat shock protein hsp27 increases invasiveness but decreases motility of breast cancer cells.,” *Invasion & metastasis*, vol. 17, no. 3, pp. 113–123, 1997.
- [104] M. A. Bausero, D. T. Page, E. Osinaga, and A. Asea, “Surface expression of hsp25 and hsp72 differentially regulates tumor growth and metastasis,” *Tumor Biology*, vol. 25, no. 5-6, pp. 243–251, 2004.
- [105] G. Nagaraja, P. Kaur, and A. Asea, “Role of human and mouse hspb1 in metastasis,” *Current molecular medicine*, vol. 12, no. 9, pp. 1142–1150, 2012.
- [106] R. V. Blackburn, S. S. Galoforo, C. M. Berns, E. P. Armour, D. McEachern, P. M. Corry, and Y. J. Lee, “Comparison of tumor growth between hsp25- and hsp27-transfected murine 1929 cells in nude mice,” *International journal of cancer*, vol. 72, no. 5, pp. 871–877, 1997.
- [107] N. Fujimoto, C. Asano, K. Ono, and S. Tajima, “Verruciform xanthoma results from epidermal apoptosis with galectin-7 overexpression.,” *Journal of the European Academy of Dermatology and Venereology: JEADV*, vol. 27, no. 7, pp. 922–923, 2012.

- [108] E. Ochiai, Q. Sa, M. Brogli, T. Kudo, X. Wang, J. P. Dubey, and Y. Suzuki, “Cxcl9 is important for recruiting immune t cells into the brain and inducing an accumulation of the t cells to the areas of tachyzoite proliferation to prevent reactivation of chronic cerebral infection with toxoplasma gondii,” *The American journal of pathology*, vol. 185, no. 2, pp. 314–324, 2015.
- [109] J. Zhang, H. Wang, C. Yuan, J. Wu, J. Xu, S. Chen, C. Zhang, and Y. He, “Itgal as a prognostic biomarker correlated with immune infiltrates in gastric cancer,” *Frontiers in Cell and Developmental Biology*, vol. 10, p. 808212, 2022.
- [110] J. R. Laurberg, J. B. Jensen, T. Schepeler, M. Borre, T. F. Ørntoft, and L. Dyrskjøt, “High expression of gem and ednra is associated with metastasis and poor outcome in patients with advanced bladder cancer,” *BMC cancer*, vol. 14, no. 1, pp. 1–10, 2014.
- [111] B. F. Miller, D. Bambah-Mukku, C. Dulac, X. Zhuang, and J. Fan, “Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities,” *Genome research*, vol. 31, no. 10, pp. 1843–1855, 2021.