IDENTIFICATION OF SPATIALLY VARIABLE GENES WITH WEIGHTED LEVERAGE SCORES

by

Ran Wei

A Thesis Submitted to the Faculty of the

DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

WITH A MAJOR IN BIOSTATISTICS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2023

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by: Ran Wei
titled:

and recommend that it be accepted as fulfilling the thesis requirement for the Master's Degree.

Yiwen Liu (Jul 7, 2023 15:27 PDT) Date: Jul 7, 2023
Yiwen Liu

Edward John Bedrick Date: Jul 8, 2023
Edward Bedrick

Deepta Bhattacharya Date: Jul 7, 2023
Deepta Bhattacharya

Date: Jul 7, 2023
Xiaoxiao Sun

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement.

Yiwen Liu (Jul 7, 2023 15:27 PDT) Date: Jul 7, 2023
Yiwen Liu
Epidemiology and Biostatistics

ACKNOWLEDGEMENTS

I would like to express my gratitude to my thesis advisor, Dr. Yiwen Liu, for her patient guidance on navigating statistical algorithms. She has not only assisted me in planning this thesis but also helped me revise this thesis paper. I am also grateful to Dr. Xiaoxiao Sun for his valuable suggestions regarding my statistical approaches. Further, my sincere thanks go to Dr. Edward Bedrick and Dr. Deepta Bhattacharya for their insightful perspectives that enrich the depth and breadth of the study.

CONTENTS

LIST OF FIGURES 5
1. ABSTRACT 6
2. INTRODUCTION 7
2.1. Introduction to Spatial Transcriptomics 7
2.2. Introduction to SV genes screening 7
2.3. Introduction to the WLS Screening Method 8
3. METHODS 10
3.1. General Index Model 10
3.2. The WLS Variable Screening Method 10
3.3. BayesSpace 11
3.4. ScGCO and SPARK-X 12
4. SIMULATION ANALYSIS 13
4.1. Data Generation 13
4.2. SV Genes Identification with True Labels 14
4.3. SV Genes Identification without Label Information 14
5. ANALYSIS OF MOUSE OLFACTORY BULB DATA 17
5.1. Data Overview 17
5.2. SV Gene Identification 17
6. ANALYSIS OF HUMAN BREAST CANCER DATA 19
6.1. Data Overview 19
6.2. SV Gene Identification 19
6.3. Disease Gene Set Enrichment Analysis 19
7. CONCLUSION 22
7.1. Strengths 22
7.2. Limitations 22
7.3. Outlook 22
BIBLIOGRAPHY 25

LIST OF FIGURES

4.1 Three authentic biological tissue patterns from mouse olfactory bulb (MOB) 13

4.2 Box plots showing accuracy, sensitivity, and F1 score for SV gene identification with true cluster label using various methods with increasing Gaussian noises across 3 tissue patterns 15

4.3 Illustrative examples of clustering results obtained from BayesSpace . . . 16

4.4 Box plots showing accuracy, sensitivity, and F1 score for SV gene identification on simulated datasets using scGCO, SPARK-X, and the combined application of the WLS and BayesSpace 16

5.1 a. Histology image of the MOB tissue[4]. b. Annotation of the known MOB tissue architecture displaying the five distinct layers 17

5.2 a. Spatial clusters derived from BayesSpace. b. Venn diagram illustrating the set relationship among SV genes identified by WLS, scGCO, and SPARK-X. c. Reconstructed MOB tissue structures using identified SV genes from different methods. d. Reconstructed MOB tissue structures using SV genes uniquely identified by WLS. 18

6.1 a. Histology image with annotated regions of the breast cancer biopsy[4] b. Venn diagram illustrating the set relationship among SV genes identified by WLS, scGCO, and SPARK-X. c. Reconstructed tissue domains using identified SV genes from different methods. d. Reconstructed tissue domains using SV genes uniquely identified by each method. 20

6.2 Dot plot showing enriched DOs and their evaluation metrics: p.adjusted denotes the adjusted statistical significance of the enriched DO; count refers to the number of genes associated with each DO term; gene ratio is the ratio of differentially expressed gene number enriched in the DO to the total gene number in the DO. 21

1. ABSTRACT

Spatially variable (SV) gene detection plays a crucial role in analyzing spatial transcriptomics data for dissecting complex tissue microenvironments. Recently, a novel weighted leverage score (WLS) variable screening method was proposed to efficiently identify important features. Herein, this research introduced a new approach applying the WLS method together with a clustering algorithm, such as BayesSpace, to effectively identify SV genes. Through rigorous testing on simulated and rea-world datasets against established methods, the combination of the WLS method and BayesSpace consistently demonstrated superior accuracy in SV gene detection. As a robust, efficient, and flexible tool, the WLS method can be used to significantly improves the results in studying complex relationships between molecular cell functions and tissue phenotypes in spatial transcriptomics analysis.

2. INTRODUCTION

2.1. Introduction to Spatial Transcriptomics

Transcriptomics technologies such as bulk and single-cell RNA sequencing (scRNA-seq) are designed to measure the complete set of RNA transcripts in a biological sample, allowing us to study the expression of an organism's genes under different tissues, conditions, or time points[1]. Despite the ongoing success of RNA-seq in biomedical field, the technique requires cells to be dissociated from whole tissue which leads to the loss of spatial context of gene expression. The isolation of cells may also induce stress and death that alter the expression profile. Researchers studying biological structures such as lymph nodes and nervous tissues could greatly benefit from the location patterns of gene expression, which primarily drives the development of spatially resolved transcriptomics, or simply 'spatial transcriptomics'[2].

There are two main types of spatial transcriptomics: imaging-based and sequencing-based. Imaging-based techniques, such as in situ hybridization (ISH) and in situ sequencing (ISS), visualize mRNA molecules in their native tissue context via microscopy. While they offer high spatial resolution often at cellular or even subcellular level, they are time-consuming and challenging at capture efficiency as the number of profiled genes increase[3]. Sequencing-based spatial transcriptomics profile gene expression by mounting tissue on arrayed reverse transcription primers with unique positional barcodes, which can profile the whole transcriptome of an organism with next-generation sequencing (NGS) library, at a cost of lower spatial resolution due to the limited size of capture spot[4]. The lack of single-cell resolution poses challenge of distinguishing cellular identities because multiple cells or fractions of multiple cells could contribute to the mRNA measurements. Since my focus is predominantly on analyzing data generated from sequencing-based spatial transcriptomics, I will refer to sequencing-based spatial transcriptomics simply as 'spatial transcriptomics' unless otherwise noted.

2.2. Introduction to SV genes screening

Much like the process of cell type annotation in scRNA-seq studies, a crucial downstream step in spatial transcriptomics involves identifying spatial structures based on gene expression patterns. Cells are organized within tissue structures, and their contribution to various biological functions are often revealed through their interactions with neighboring cells. By investigating spatial structures and their corresponding gene expression profiles, we can gain insights into cell-cell interactions and cell functions within their native environment[5]. This objective is typically achieved by initially identifying spatially variable (SV) genes, which exhibit significant variation in expression profiles across distinct tissue

regions. SV genes play a vital role in unraveling the biological relationships between spatial regions and cellular functions, thus offering valuable insights into cell type mapping and spatial heterogeneity within a tissue.

In the past, several methods have been developed to identify SV genes, such as trendScreek[6], spatialDE[7], and SPARK[8]. While these methods have achieved success in SV gene identification, they are computationally intensive. As the volume of spatial transcriptomics data continues to grow exponentially, these methods may require days or even months of processing time and consume extensive amounts of physical RAM memory. To address the computational challenge, two recently introduced scalable algorithms have emerged: scGCO[9] and SPARK-X[10]. ScGCO combines the power of hidden Markov random field (HMRF) and graph cuts to identify SV genes efficiently. SPARK-X utilizes a non-parametric modeling approach based on a robust covariance test framework, which enables the identification of gene expression variations across different spatial regions while being scalable to handle large datasets. These new algorithms offer improved scalability, enabling faster and more memory-efficient SV gene identification, even in the face of increasingly large spatial transcriptomics datasets. However, they have certain limitations that could affect their efficacy. For example, the graph cuts algorithm used by scGCO may not efficiently identify structures with thin layers due to shrinking bias[11]. SPARK-X, as a nonparametric model, have a greater risk of overfitting the data, leading to the identification of excessive numbers of SV genes that may not be biologically meaningful.

2.3. Introduction to the WLS Screening Method

An innovative model-free variable screening method has recently been developed based on the weighted leverage scores (WLS)[12]. The leverage score is a metric used in regression analysis to evaluate the fitness of a model, measuring how well a random subsample can represent the entire sample. Random subsamples selected based on leverage scores are often effective surrogates of the full sample. Leverage score-based sampling methods have demonstrated significant success in reducing sample sizes for big data analysis. This is crucial to statistical analysis today, as the computing resources are lagging far behind the exponential growth of dataset[13], [14]. The WLS screening method extends the leverage score-based sampling algorithm to predictors screening by incorporating both the left and right singular vectors to evaluate a predictor's importance. It has been shown that WLS enjoys a nice ranking property and achieves sure screening consistency. That is, important predictors have higher WLSs than redundant predictors, and predictors selected using the WLS screening method encompasses true predictors with probability approaching one.

The WLS screening method offers several notable advantages. Firstly, the method was developed under the general index model assumption, which can be considered as a model-free approach. It mitigates the potential biases associated with incorrect model

specification, which is a common issue in variable screening methods relying on parametric models. This feature enhances the reliability of the screening results. Secondly, unlike nonparametric models, the WLS screening method does not require the estimation of the unknown link function between the response and predictors, resulting in analytical and computational benefits. Thirdly, the WLS method leverages the widely available singular value decomposition (SVD) technique, which is commonly supported in various computing software. This accessibility ensures that researchers can readily apply the method without significant barriers. Additionally, the use of SVD contributes to the efficiency and scalability of the method, particularly in high-dimensional settings, as it employs a one-pass screening algorithm. Lastly, the WLS screening method incorporates a unique BIC-type criteria to consistently decide the number of selected variables. This criterion aids in selecting an appropriate subset of predictors, enhancing the interpretability and practicality of the method's outcomes.

To demonstrate the robustness and effectiveness of the WLS method in SV gene identification, I designed a series of experiments that employed simulated spatial transcriptomics data with different spatial patterns and noise levels, as well as two real-world data obtained from previous studies. For simulated dataset with provided spatial labels, the performance of the WLS method was compared to the common feature detection algorithms from the Seurat package [15]. For data without spatial labels, the WLS method was coupled with BayesSpace, which is currently considered one of the best methods for distinguishing tissue structures into clusters in spatial transcriptomics [16]. These cluster labels served as input for the WLS method to detect SV genes. The combined WLS and BayesSpace approach was compared against scGCO and SPARK-X. For the simulated dataset, the performance was measured by the true SV genes accurately detected by each method. In real data analysis, spatial clusters were reconstructed based on the identified SV genes, and subsequently compared against annotated patterns from histology image. Gene set enrichment analysis (GSEA) was also performed to evaluate the biological relevance of the identified SV genes. Through these extensive comparisons, the capability of WLS method could be examined, both as a standalone tool and in combination with BayesSpace.

The rest of the paper is arranged as follows. Section 3 provides a brief overview of the WLS method for variable screening, along with an introduction to other methods employed in the analysis, including BayesSpace, scGCO, and SPARK-X. This section sets the foundation for understanding the subsequent discussions. Section 4 presents the results of simulation studies. The results obtained from these simulated patterns provide insights into the performance and effectiveness of the WLS screening method in identifying SVGs, allowing for a full evaluation of its capabilities. Two real-world data analyses are reported in Section 5 and Section 6. Finally, the paper is concluded in Section 7 with a comprehensive discussion, summarizing the key findings and implications of the study.

3. METHODS

3.1. General Index Model

To understand the model-free nature of the WLS method, we can first look at the general index model (GIM)[17]. General index model is a statistical model used to describe the relationship between a response variable $\boldsymbol{y}$ and a $\boldsymbol{p}$-dimensional vector of predictor variables $\boldsymbol{x}$ in the following form:

$$\boldsymbol{y} = \boldsymbol{f}(\beta_1^T \boldsymbol{x}, ..., \beta_k^T \boldsymbol{x}, \boldsymbol{\epsilon}), \tag{1}$$

where $\boldsymbol{f}(\cdot)$ is an unknown link function, $\beta_1, ..., \beta_k$ are $\boldsymbol{p}$-dimensional parameter vector, $\boldsymbol{k}$ is an integer that is much smaller than $\boldsymbol{p}$, and ϵ is the error term. The key idea is that, instead of assessing the relationship between the response and each predictor variable separately, the GIM can condense this information into a linear combination of predictor variables, $\{\beta_m^T \boldsymbol{x} | \boldsymbol{m} = 1, ..., \boldsymbol{k}\}$. By sufficiently reducing the $\boldsymbol{p}$-dimensional predictor vector, $\boldsymbol{x}$, into the $\boldsymbol{k}$-dimensional index, we can handle high-dimensional datasets more efficiently, while still capturing the important relationship between the predictors and the response variable. And since the GIM allows for a nonlinear relationship between the predictor variables and the response variable via the use of a link function, $\boldsymbol{f}(\cdot)$, it has the flexibility to model complex relationships like nonparametric models.

Although the GIM offers a balance between the adaptability of nonparametric models and the simplicity of linear models, its efficiency in estimating the link function can be compromised when an excess of irrelevant features are present[18]. NGS technologies such as spatial transcriptomics often generate data that includes over tens of thousands of gene features, yet only a few thousands of spatial spots, which complicates the identification of important genes using the GIM. Therefore, in the case of $\boldsymbol{p} >> \boldsymbol{n}$, it becomes especially beneficial to detect and remove irrelevant features while modeling GIM.

3.2. The WLS Variable Screening Method

The estimated coefficients $\beta_1, ..., \beta_k$ in the GIM could be understood as the individual contribution of each gene to predict spatial domain $\boldsymbol{y}$. Based on this idea, the WLS method was developed to rank genes based on their significance in predicting spatial domains, hence selecting important SV genes over the redundant ones.

Given the transcriptomics data $(\mathbf{x}_i^T, \boldsymbol{y}_i)$ for spatial location $i = 1, \ldots, n$ where the cluster label $\boldsymbol{y}_i$ are associated with the $\boldsymbol{k}$ linear combination of the gene expression count $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. As a sparse vector, $\mathbf{x}_i$ can be approximated using singular value decomposition (SVD): $\mathbf{x}_i \approx V \Lambda \mathbf{u}_i^T$, where $V \in \mathbb{R}^{p \times d}$ is the right singular matrix, $\Lambda \in \mathbb{R}^{d \times d}$

is a diagonal matrix with with sigular values as its diagonal elements, and $\mathbf{u}_i \in \mathbb{R}^{d\times 1}$ is the left singular vector. The weighted leverage score is derived based on the slicing scheme and inverse regression idea[17]. Initially, the range of $\boldsymbol{y}_i$ is partitioned into H slices $S_1, \ldots, S_H$. If $\boldsymbol{y}_i$ is a categorical variable, it naturally divides into H slices. The mean of $\mathbf{u}_i$ within each slice is calculated and denoted as $\bar{\mathbf{u}}_h = \frac{1}{n_h}\sum_{i=1}^{I} \mathbf{u}_i I(y_i \in S_h)$, for $h = 1, \ldots, H$, where $I(\cdot)$ is the indicator function, and $n_h = \sum_{i=1}^{I} I(y_i \in S_h)$. Additionally, the sample variance from each slice is obtained as:

$$M = \sum_{h=1}^{H} \frac{n_h}{n} \bar{\mathbf{u}}_h \bar{\mathbf{u}}_h^T, \tag{2}$$

This variance captures the association between the left singular vector $\mathbf{u}_i$ and the response $\boldsymbol{y}_i$, and consequently, the information contained in the link function $\boldsymbol{f}(\cdot)$. The right singular vector $\mathbf{v}_j$, the jth row of V, reflects the jth predictor information. Accordingly, the weighted leverage score is constructed as:

$$w_j = \mathbf{v}_j^T M \mathbf{v}_j, \tag{3}$$

for $j = 1, \ldots, \boldsymbol{p}$. It has been shown that the WLS processes a nice ranking property: important predictors have higher WLS than redundant predictors[12].

After sorting the genes from high to low WLS, a modified BIC-type of criterion is introduced to determine the cutoff point. Suppose the genes are arranged such that their WLSs satisfy $w_1 > \ldots > w_p$, and let r be the number of selected SV genes:

$$G(r) = -\log(\sum_{j=1}^{r} w_j) + r(\log n + c_n \log p)/\max(n, p), \tag{4}$$

where c_n is a pre-specified positive constant that acts as a control variable for the threshold level for SV gene selection. It has been demonstrated that $G(r)$ is able to effectively filter out redundant genes, hence ensuring the robustness and efficiency of the WLS method for SV gene identification.

3.3. BayesSpace

BayesSpace is a fully Bayesian statistical method developed to analyze spatial transcriptomics data, with a primary goal to identify distinct spatial regions that have similar gene expression profiles within a tissue sample. The data is modelled as follows:

$$(y_i|z_i = k, w_i) \sim \mathcal{N}(y_i; \mu_k, w_i^{-1}\Lambda^{-1}), \tag{5}$$

y_i is a low d-dimensional representation of the gene expression profile for the i spot. $z_i \in \{1, ..., q\}$ is the latent cluster that spot i belongs to. μ_k is the mean vector for cluster k. Λ denotes the precision matrix, and w_i denotes an unknown scaling factor. Prior distributions were given to μ_k, Λ, and w_i. The parameters of the model are then estimated using

a Markov Chain Monte Carlo (MCMC) approach, which tend to explore the parameter space more efficiently than alternatives. A t-distributed error model is used to make the method more robust to outlier caused by technical noise. To encourage neighboring spots to belong to the same cluster, a Markov random field prior is given to smooth the clustering. Lastly, the number of clusters q in a given sample is determined either by prior biological knowledge when available, or by the elbow of the pseudo-log-likelihood plot.

BayesSpace is a robust, efficient, and highly integrative tool for detecting spatial patterns and providing precise clustering label that can then serve as informative input for the WLS variable screening method on SV genes identification. The combination of BayesSpace and the WLS method can provide an efficient and powerful solution for dissecting the complex spatial gene expression patterns.

3.4. ScGCO and SPARK-X

To benchmark the performance of the WLS method in combination with BayesSpace, I used scGCO and SPARK-X as comparative methods.

Single-cell graph cuts optimization (scGCO) is primarily developed to identify SV genes in spatial transcriptomics. It utilizes a Hidden Markov Random Field (HMRF) model to screen candidate regions with distinct spatial expression patterns, which were subsequently optimized through a graph cuts algorithm to identify the best segmentation based on signal-to-noise ratio. Each gene is evaluated using the optimal segmentation under the Complete Spatial Randomness (CSR) framework, providing an adjusted p-value that represents its importance.

SPARK-X is another scalable non-parametric method for SV gene identification. It is built upon a robust covariance test framework. Covariance matrix measures the similarity between pairs of locations in a sample. In essence, SPARK-X constructs an expression covariance matrix and a distance covariance matrix, which were then compared with each other in terms of correlation. If the two covariance matrices are correlated, it suggests that the spatial coordinates are not independent with the gene expression measurement, indicating a potential SV gene.

Wile both BayesSpace and scGCO utilize Markov random field (MRF) models in their analysis, they differ in the way that BayesSpace uses MRF as a spatial prior for the MCMC modeling, whereas scGCO use HMRF, a variant of MRF, coupled with an optimization method (graph cuts) to identify clusters. Moreover, BayesSpace does not test for SV genes. It only provides spatial clustering information that can be used for the WLS variable screening method to identify SV genes. On the other hand, scGCO directly test for SV genes using the CSR framework. Therefore, it's worth to compare the combined approach of BayesSpace and the WLS method with scGCO as well as SPARK-X.

4. SIMULATION ANALYSIS

4.1. Data Generation

To mimic the spatial transcriptomic data, I generated three authentic biological patterns extracted from the histology images of mouse olfactory bulb (MOB) using the scGCO package(Zhang et al., 2022). Based on the reference annotation, these patterns were designated as T1 representing the outer plexiform layer (OPL), T2 corresponding to the granular cell layer (GCL), and T3 indicating the olfactory nerve layer (ONL) (Figure 4.1). These authentic biological patterns serve as representative examples in our study, allowing us to analyze and explore the properties and behavior of proposed methods in identifying SV genes of the spatial transcriptomic data.

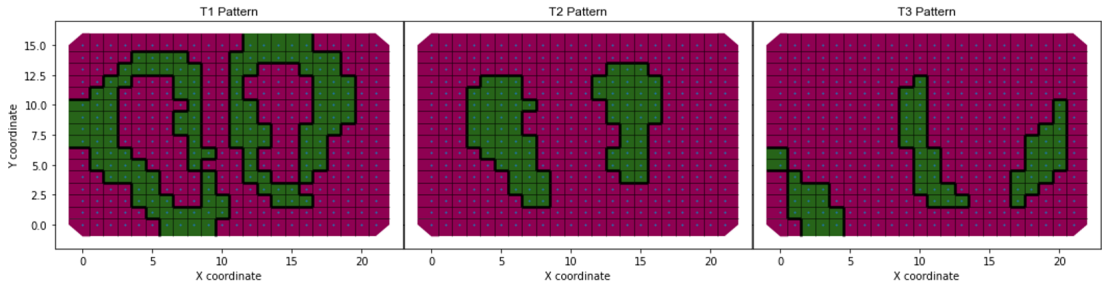

Figure 4.1: Three authentic biological tissue patterns from mouse olfactory bulb (MOB)

All the simulated datasets were generated in Python using the UofA HPC system. Each simulated data matrix consists of 352 spatial locations and 9200 genes, organized in a matrix of 352×9200 (locations × gene). Among these 9200 genes, 200 are designated as ground truth SV genes related to the spatial pattern, and the remaining 9000 genes serve as the background/redundant genes. For SV genes, the expression counts are randomly drawn from a normal distribution S_1 if the corresponding location falls within the pattern (Figure 4.1, the green area), or from a normal distribution S_2 if it falls outside the pattern (Figure 4.1, the red area). Background genes receive values drawn from a distribution B. These distribution are defined as follows:

$$
\begin{aligned}
S_1 &\sim N(\mu_v, \sigma_v^2), \\
S_2 &\sim N(\mu_v - 1, \sigma_v^2), \\
B &\sim N(\mu_b, \sigma_b^2),
\end{aligned}
$$

where μ_v is drawn from the set $\{2, 3, 4\}$ for each SV gene. This setup ensures that the mean expression of SV genes varies by a value of 1 depending on whether the location is inside or outside the simulated spatial pattern. The noise level σ_v is selected from a range of 0.5 to 3.0, incremented by 0.5. For background genes, the mean expression μ_b is a random integer between 0 to 4, and the corresponding σ_b is randomly chosen from the interval

$[0.1, \sigma_v]$. Each expression count was rounded to the nearest integer. Any values fall below zero were set to zero. A total of 1800 data matrices were generated, containing 3 tissue patterns, 6 noise levels (σ_v) for each pattern, and 100 replicates for each combination of tissue pattern and noise level.

4.2. SV Genes Identification with True Labels

To evaluate the performance of the WLS method alone, I initially provided the true spatial labels associated with each patterns to the WLS algorithm for SV gene identification. Since scGCO and SPARK-X do not accommodate user-specified spatial labels, I instead used Seurat package's marker gene identification methods[15] to determine SV genes among predetermined clusters. The outcomes of these methods were then compare with the results by the WLS method. These conventional methods include the Wilcoxon Rank Sum test ("wilcox"), the likelihood-ratio test for single cell gene expression ("bimod")[19], the Student's t-test ("t"), the logistic regression test ("LR"), and the Model-based Analysis of Single-cell Transcriptomics ("MAST")[20].

For all alternative methods, genes with adjusted p-values of 0.05 or less were selected as candidate SV genes. As the WLS method uses a modified BIC-type criterion to decide the number of selected genes, I set the cutoff at $c_n = 0.5$ to have a moderate screening threshold. Accuracy, selectivity (i.e., true positive rate), and F1 score were used to evaluate the performance of methods. Accuracy is the ratio of correctly selected genes to the whole gene pool in the sample. Selectivity measures the proportion of the true SV genes that are correctly identified. F1 socre is the harmonic mean of precision and sensitivity. All these performance metrics reach their best value at 1 and worst at 0.

Figure 4.2 shows that, across all three tissue patterns, and especially when the noise level exceeds 1.5, the WLS method outperforms all the other methods. This indicates that the WLS method can accurately identify a larger number of true SV genes while minimizing the selection background genes. The result also suggests the potential effectiveness of the WLS method in subsequent analysis when it is combined with the BayesSpace clustering algorithm.

4.3. SV Genes Identification without Label Information

In this analysis, I did not provide prior label information for each tissue pattern. The WLS method was paired with BayesSpace to detect SV genes. The results were compared with those from the scGCO and SPARK-X. Accuracy, selectivity, and F1 scores were computed to evaluate the performance of these methods.

Figure 4.3 provides examples of clustering results obtained from BayesSpace across the three tissue patterns, at a noise level of 3.0. Even at such the elevated noise level, BayesSpace is able to decompose the spatial transcriptomics profile into clusters that

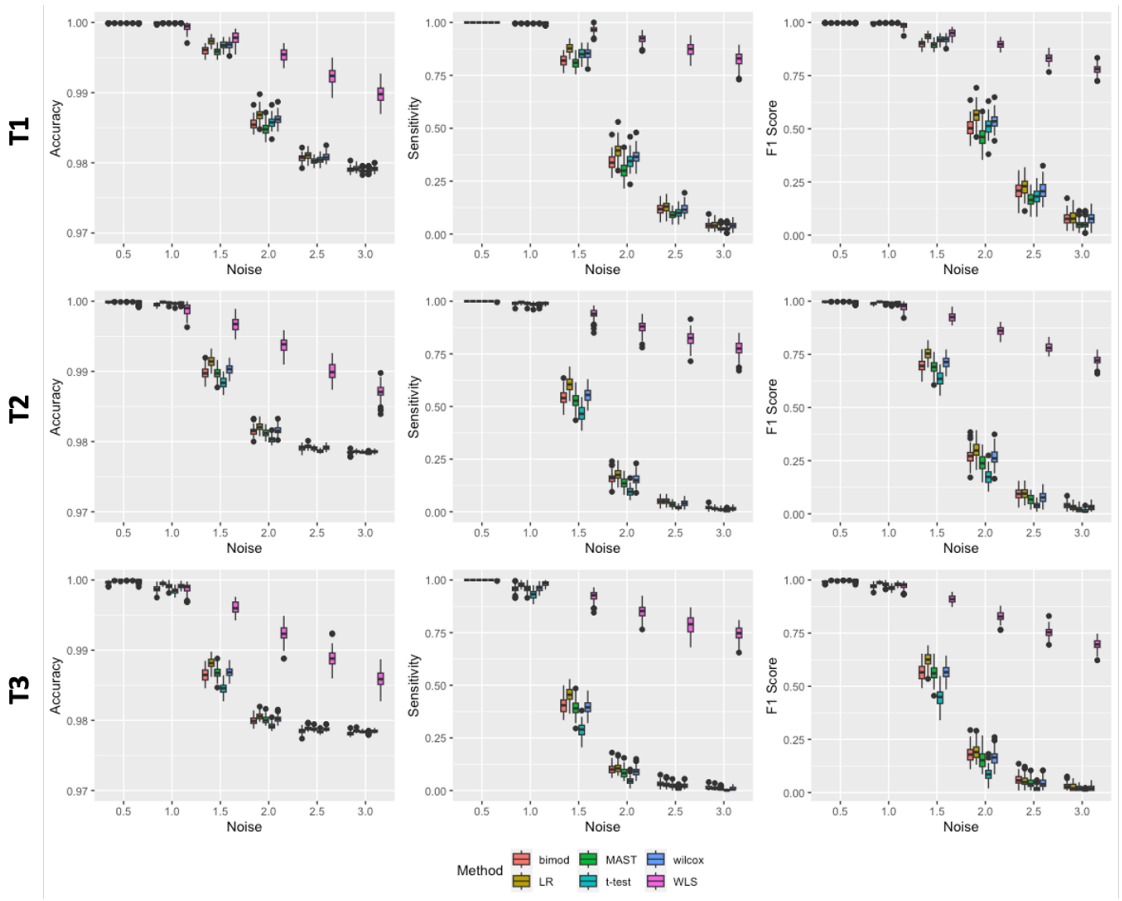

Figure 4.2: Box plots showing accuracy, sensitivity, and F1 score for SV gene identification with true cluster label using various methods with increasing Gaussian noises across 3 tissue patterns

closely align with the authentic patterns, with only a small number of misaligned spots. By combining BayesSpace with WLS methods, I create a pipeline that can accurately select SV genes and outperform the scGCO and SPARK-X methods across all tissue patterns and noise levels (Figure 4.4). It is shown that the WLS method with correct label information exhibits superior performance in detecting SV genes (Figure 4.2). Thus, the effective clustering by BayesSpace greatly contribute to the exceptional results achieved by the WLS method.

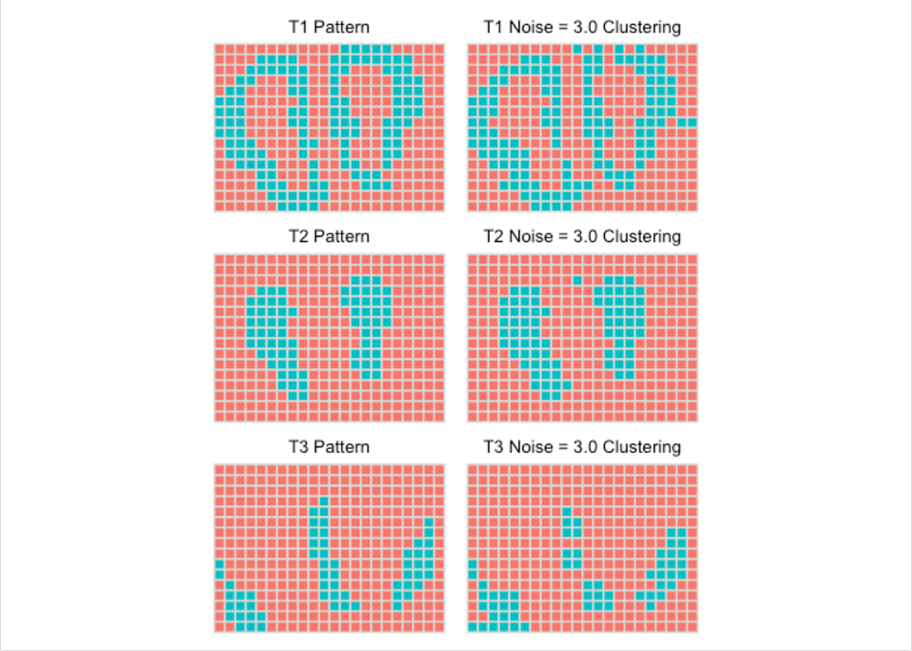

Figure 4.3: Illustrative examples of clustering results obtained from BayesSpace

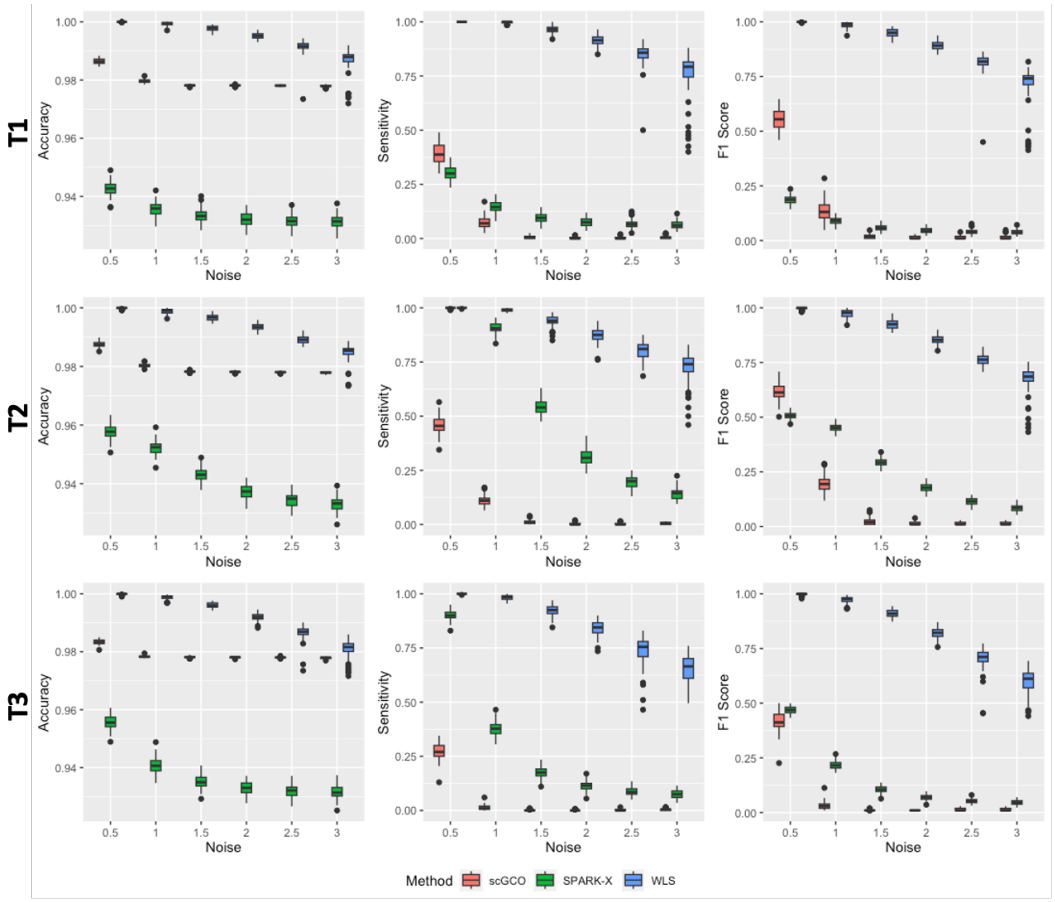

Figure 4.4: Box plots showing accuracy, sensitivity, and F1 score for SV gene identification on simulated datasets using scGCO, SPARK-X, and the combined application of the WLS and BayesSpace

5. ANALYSIS OF MOUSE OLFACTORY BULB DATA

5.1. Data Overview

The spatial transcriptomics data from mouse olfatory bulb (MOB) comprises expression profiles for 16,218 genes over 260 spatial spots. Based on the reference from known MOB tissue structures[4], [9], this dataset can be annotated with five distinct layers within the MOB tissue according to the histology image: granular cell layer (GCL), mitral cell layer (MCL), outer plexiform layer (OPL), glomerular layer (GL), and olfactory nerve layer (ONL) (Figure 5.1)

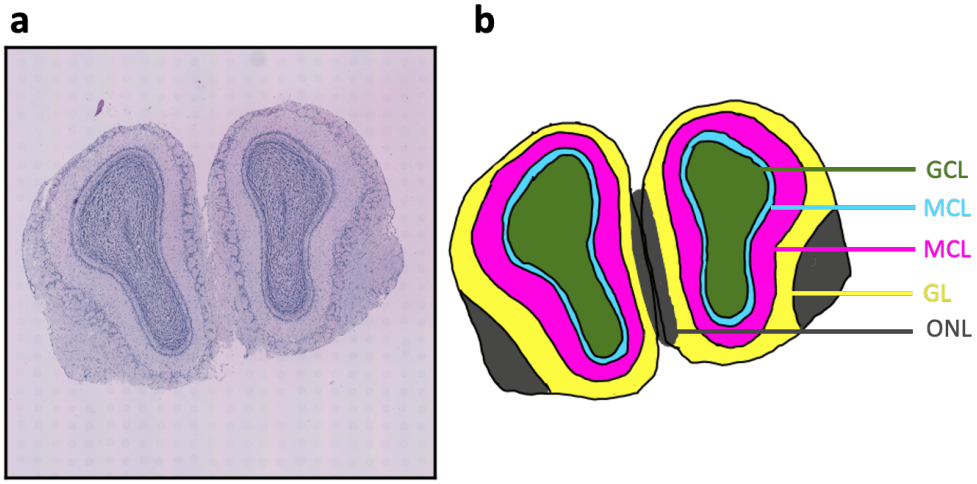

Figure 5.1: a. Histology image of the MOB tissue[4]. b. Annotation of the known MOB tissue architecture displaying the five distinct layers

5.2. SV Gene Identification

SV genes in the MOB data were identified using the combined application of the WLS method and BayesSpace, along with scGCO and SPARK-X. I first applied BayesSpace to decompose the data into clusters, resulting in five distinct regions that closely align with the annotated layers (Figure 5.2a). Based on the cluster labels, the WLS method was able to identify 260 SV genes at a moderate level of screening threshold. ScGCO identified 502 SV genes (FDR < 0.01), while SPARK-X identified 1,329 SV genes (adjusted p-value < 0.01).

Because the expression profile of biologically meaningful SV genes should correlate strongly with annotated tissue structures, I then investigated whether the identified SV genes could accurately delineate all five annotated layers of the MOB tissue structure.

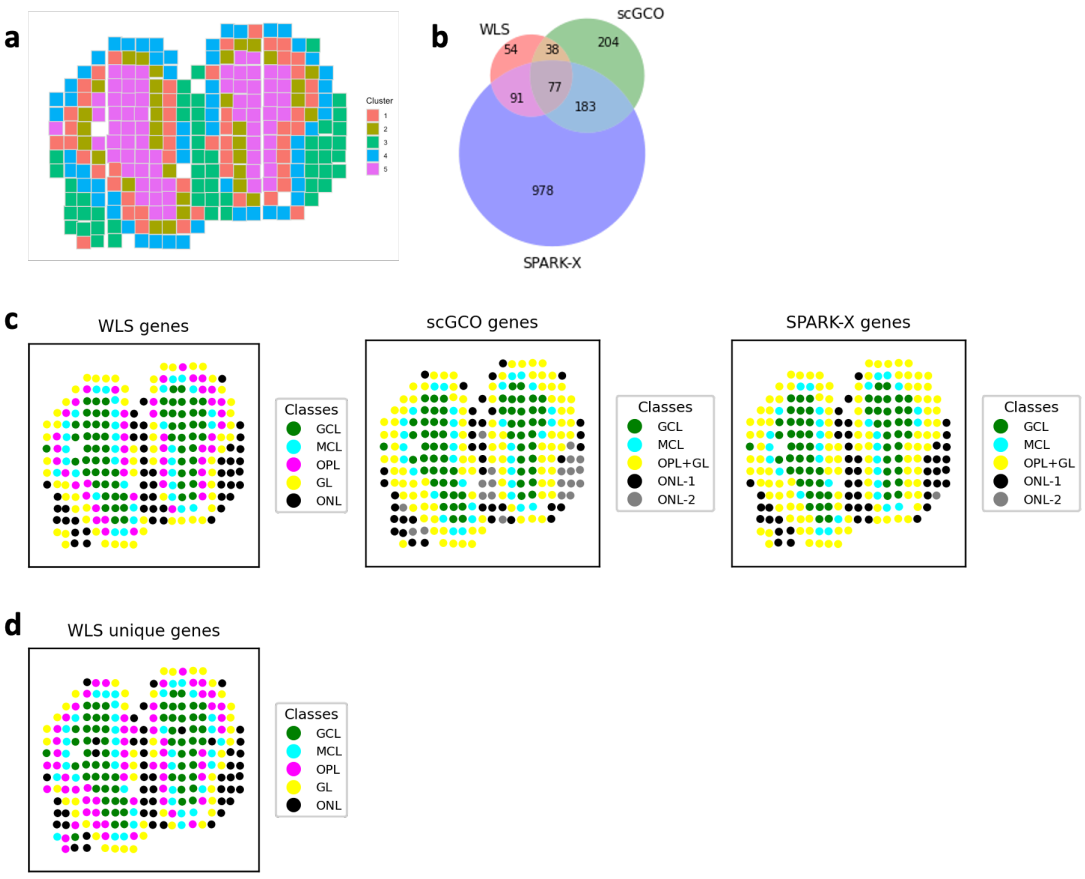

Figure 5.2: a. Spatial clusters derived from BayesSpace. b. Venn diagram illustrating the set relationship among SV genes identified by WLS, scGCO, and SPARK-X. c. Reconstructed MOB tissue structures using identified SV genes from different methods. d. Reconstructed MOB tissue structures using SV genes uniquely identified by WLS.

This was performed using principle component analysis (PCA) and a k-means clustering algorithm. Figure 5.2c shows that the SV genes identified by the WLS method successfully reconstruct all five distinct layers of the MOB tissue, which was not achieved by scGCO nor SPARK-X. Specifically, scGCO and SPARK-X failed to differentiate between the OPL and GL layers, and instead tried dissect sub-layers of ONL. Importantly, the MOB tissue structure resolved using the 54 SV genes uniquely identified by WLS also displays a five-layer pattern (Figure 5.2d). This indicates that the combined approach of BayesSpace and the WLS method demonstrates superior performance in identifying SV genes compare to both scGCO and SPARK-X.

6. ANALYSIS OF HUMAN BREAST CANCER DATA

6.1. Data Overview

I further extended the application of the BayesSpace + WLS technique to spatial transcriptomics data from a human breast cancer biopsy[4]. Based on morphological criteria, this biopsy comprises three types of tissues: invasive ductal cancer (INV), ductal cancer in situ (DC), and normal tissues (NT) (Figure. 6.2a)

6.2. SV Gene Identification

With the cluster labels provided by BayesSpace, the WLS method identified 251 SV genes at moderate level of screening threshold, while scGCO and SPARK-X identified 252 (FDR < 0.01) and 211 (adjusted p-value < 0.01) SV genes, respectively. Spatial domains were reconstructed based on these SV genes to assess their ability to differentiate INV and DC regions from normal tissues. SV genes from all three methods were able to distinguish the three spatial domains within breast cancer biopsy (Figure 6.2c). Then, I evaluated the SV genes uniquely identified by each method. Figure 6.2d shows that the uniquely identified SV genes from all three methods retained the ability to identify INV and DC regions, although with slightly less precision compared to the results from the complete sets of genes. It could be difficult to directly compare the reconstructed patterns across the three methods because cancer development is a continuous process where different SV genes selected by these methods might indicate subtle differences that are not captured by the morphological pattern. Therefore, the comparison of the quality of SV genes could be conducted quantitatively using gene enrichment analysis.

6.3. Disease Gene Set Enrichment Analysis

Since the breast cancer biopsy contains three different stages of cancer development, the SV genes are not only expected to spatially correlate with the histological domains, but also to be related to cancer development. Thus, I performed disease gene set enrichment analysis (DSEA) on the SV genes identified by the three methods. As a variation of the gene set enrichment analysis (GSEA), DSEA is a computational technique that determines whether a predefined set of genes shows statistically significant differences between groups of cells/spots in relation to the disease ontology (DO). As illustrated in Figure 6.3, enriched DOs are ranked from top to bottom at the y-axis. The common two DOs enriched by all three lists of SV genes are "cancer" and "disease of cellular proliferation" which are also highly related to cancer development. Although both DOs are significantly enriched (shown by the color scale of adjusted p-value), only the SV gene

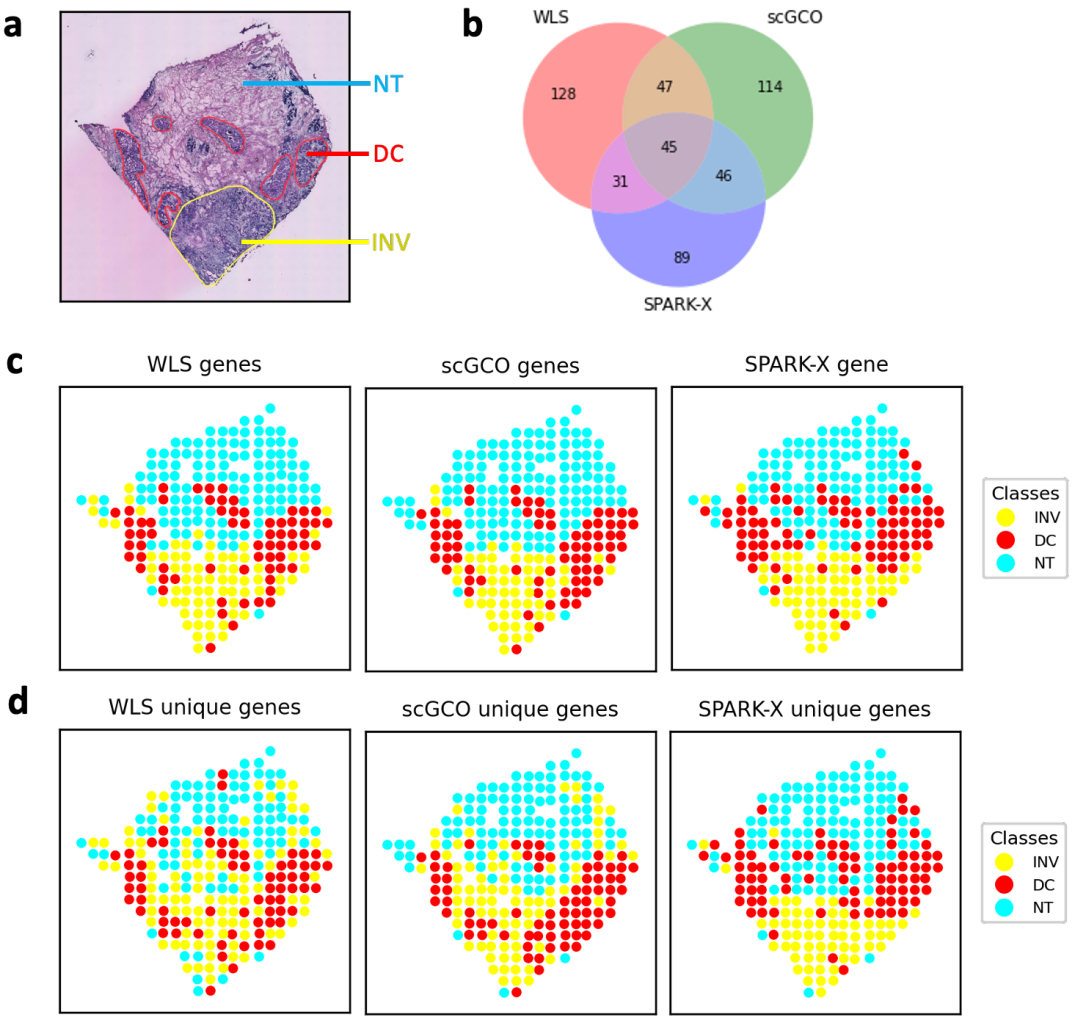

Figure 6.1: a. Histology image with annotated regions of the breast cancer biopsy[4] b. Venn diagram illustrating the set relationship among SV genes identified by WLS, scGCO, and SPARK-X. c. Reconstructed tissue domains using identified SV genes from different methods. d. Reconstructed tissue domains using SV genes uniquely identified by each method.

lists from WLS and SPARK-X are able to present them as the top two enriched terms. In addition, the number of genes associated with cancer and proliferation disease are higher in the list from WLS compared to the results from both scGCO's and SPARK-X's lists, as indicated by count and gene ratio. Therefore, although all three SV gene lists demonstrated their qualities by successfully enriching cancer-related DOs, the list selected by WLS has more genes associated with cancer-related DOs. This further provies that the WLS method is capable of consisitently identifying SV genes of high quality.

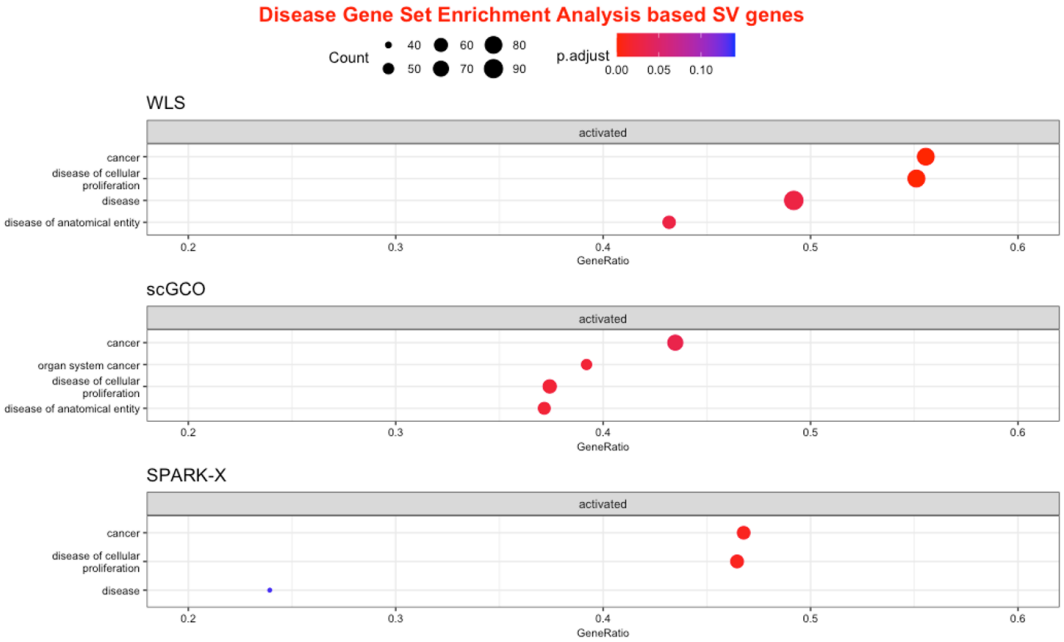

Figure 6.2: Dot plot showing enriched DOs and their evaluation metrics: p.adjusted denotes the adjusted statistical significance of the enriched DO; count refers to the number of genes associated with each DO term; gene ratio is the ratio of differentially expressed gene number enriched in the DO to the total gene number in the DO.

7. CONCLUSION

7.1. Strengths

The WLS variable screening method, in combination with BayesSpace, has demonstrated high accuracy and effectiveness in identifying SV genes, outperforming other methods such as scGCO and SPARK-X across different datasets. It has versatility and adaptability such that other clustering approaches could also be coupled with the WLS method to potentially further improve its capability to detect SV genes. For example, cell2location is a Bayesian model that uses scRNA-seq reference to decompose spatial transcriptomics data into fine-grained cellular maps and, thus, can cluster spots based on their cellular compositions[21]. Without using any prior clustering algorithm, the WLS method alone can be supplied with user-specified spatial labels to detect SV genes among any area-of-interest in a tissue sample.

Additionally, the WLS method is built upon SVD, a common statistical tool available in many programming languages, making it readily available to any users with minor modifications required. In fact, the scripts for WLS method will soon be released in both R and Python - two of the most widely-used languages in data science. Moreover, to address computational concerns related to large-scale spatial transcriptomics data, I have added the option in WLS to compute the truncated SVD instead of full SVD. This extension offers significant improvements in terms of computational speed and memory efficiency without compromising the integrity of the data.

7.2. Limitations

The performance of the WLS method can be largely influenced by the quality of input labels and thus are highly dependent on the performance of clustering algorithms. Although the BayesSpace has consistently provided accurate labels, inaccuracies could still randomly occur and potentially impact the performance of the WLS method. The WLS method also relies on a BIC-type criterion to determine the number of selected predictors, requiring the specification of an appropriate threshold. This could potentially pose challenges in balancing sensitivity and specificity.

7.3. Outlook

One of the potential improvement for refining the WLS method is to integrate various clustering algorithms into the WLS script. This could streamline the process of SV genes identification. Tailoring the choice of clustering algorithms to the particular context of the

spatial transcriptomics analysis can ensure optimal cluster labels for the WLS method. For example, BayesSpace could be used to quickly deconvolute the spatial expression patterns in tissues with simple and distinct compartments such as lymph nodes and nerve tissues. However, in complex structures like bone marrow, a reference-based algorithm such as cell2location might be preferred as it offers greater accuracy in pattern decomposition using scRNA-seq reference. This flexible approach ensures that the WLS method can consistently yield good results across different scenarios in spatial transcriptomics analysis.

The combined approach of BayesSpace and the WLS method requires careful consideration of the number of clusters (Q) in a tissue sample. This parameter directly influences the resolution of clusters and subsequently the quality of selected SV genes. An underestimation of Q might result in SV genes lacking the power to effectively dissect the biological structures within a sample. On the other hand, an overestimation could introduce trivial SV genes that contribute little meaningful information about the genuine spatial profiles. An optimal scenario would be to use biological context such as a clear histology image or number of cell types known in the sample to determine the Q. But when biological context is not readily available, methods such as the elbow method offered by BayesSpace, can be employed to infer Q based on pseudo-log-likelihood[16]. However if the elbow point is ambiguous, it can be subjective and unreliable[22]. When careful estimation of the Q is required, Bayesian methods might be a worthwhile alternative, albeit at a slight computational cost. Specifically, Hierarchical Dirichlet Process (HDP) and Bayesian Hierarchical Clustering (BHC) are specialized to tackle hierarchical structures often present in biological tissues, providing a robust, probabilistic framework for determining Q.

As the affordability of spatial transcriptomics technologies improves over time, it will become increasingly common to create 3D gene expression profiles from a series of tissue slides, opening up new dimensions for understanding the complexity of biological tissues. Here, 3D clustering algorithms, such as 3D Gaussian Mixture Models (GMMs) or 3D K-means clustering, could potentially be used to identify 3D clusters in spatial transcriptomics data. These cluster labels could then serve as input for the WLS method to identify SV genes in 3D coordinates. While the field of 3D spatial transcriptomics is advancing rapidly, the inherent adaptability of the WLS method ensures its continued relevance in SV genes identification.

Another advancement happening in spatial transcritpomics is the ability to achieve single-cell resolution where the spatial spot is reduced to the average size of a cell or smaller, allowing precise mapping of each cell's location and corresponding gene expression profile. One of the primary challenges for single-cell spatial transcriptomics is the accurate definition of single-cell boundaries. Even with smaller spot sizes, expresssion signals from neighboring cells may contribute to a single spot due to misalignment, blurring the lines of cell boundaries. This obstacle could potentially be solved by using scRNA-seq reference data[23], or by further improving the resolution to subcellular

scales. Once these challenges are addressed, traditional marker gene identification methods used in scRNA-seq or the above-mentioned clustering algorithms could be employed to cluster single cells. Subsequently, the WLS method can be used for SV genes identification based on these clusters, allowing for in-depth analysis in the field of single-cell spatial transcriptomics.

BIBLIOGRAPHY

[1] D. Edsgärd, 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, Mar. 2018. DOI: 10.1038/nmeth.4634. [Online]. Available: https://doi.org/10.1038/nmeth.4634.

[2] V. Marx, "Method of the year: Spatially resolved transcriptomics," *Nature Methods*, vol. 18, no. 1, pp. 9–14, Jan. 2021. DOI: 10.1038/s41592-020-01033-y. [Online]. Available: https://doi.org/10.1038/s41592-020-01033-y.

[3] V. Petukhov *et al.*, "Cell segmentation in imaging-based spatial transcriptomics," *Nature Biotechnology*, vol. 40, no. 3, pp. 345–354, Oct. 2021. DOI: 10.1038/s41587-021-01044-w. [Online]. Available: https://doi.org/10.1038/s41587-021-01044-w.

[4] P. L. Ståhl *et al.*, "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics," *Science*, vol. 353, no. 6294, pp. 78–82, Jul. 2016. DOI: 10.1126/science.aaf2403. [Online]. Available: https://doi.org/10.1126/science.aaf2403.

[5] B. Walker, Z. Cang, H. Ren, E. Bourgain-Chang, and Q. Nie, "Deciphering tissue structure and function using spatial transcriptomics," *Communications Biology*, vol. 5, no. 1, Mar. 2022. DOI: 10.1038/s42003-022-03175-5. [Online]. Available: https://doi.org/10.1038/s42003-022-03175-5.

[6] D. Edsgärd, 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, Mar. 2018. DOI: 10.1038/nmeth.4634. [Online]. Available: https://doi.org/10.1038/nmeth.4634.

[7] V. Svensson, S. A. Teichmann, and O. Stegle, "Spatialde: Identification of spatially variable genes," *Nature Methods*, vol. 15, no. 5, pp. 343–346, Mar. 2018. DOI: 10.1038/nmeth.4636. [Online]. Available: https://doi.org/10.1038/nmeth.4636.

[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, Jan. 2020. DOI: 10.1038/s41592-019-0701-7. [Online]. Available: https://doi.org/10.1038/s41592-019-0701-7.

[9] K. Zhang, W. Feng, and P. Wang, "Identification of spatially variable genes with graph cuts," *Nature Communications*, vol. 13, no. 1, Sep. 2022. DOI: 10.1038/s41467-022-33182-3. [Online]. Available: https://doi.org/10.1038/s41467-022-33182-3.

[10] 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, Jun. 2021. DOI: 10.1186/s13059-021-02404-0. [Online]. Available: https://doi.org/10.1186/s13059-021-02404-0.

[11] S. Candemir and Y. S. Akgul, "Statistical significance based graph cut segmentation for shrinking bias," in *Lecture Notes in Computer Science*. Springer Science+Business Media, Jan. 2011, pp. 304–313. DOI: 10.1007/978-3-642-21593-3_31. [Online]. Available: https://doi.org/10.1007/978-3-642-21593-3_31.

[12] W. Zhong, Y. Liu, and P. Zeng, "A model-free variable screening method based on leverage score," *Journal of the American Statistical Association*, vol. 118, no. 541, pp. 135–146, Jun. 2021. DOI: 10.1080/01621459.2021.1918554. [Online]. Available: https://doi.org/10.1080/01621459.2021.1918554.

[13] P. Ma and X. Sun, "Leveraging for big data regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 1, pp. 70–76, Sep. 2014. DOI: 10.1002/wics.1324. [Online]. Available: https://doi.org/10.1002/wics.1324.

[14] P. Ma, M. W. Mahoney, and B. Yu, "A statistical perspective on algorithmic leveraging," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 861–911, Jan. 2015. [Online]. Available: https://jmlr.org/papers/volume16/ma15a/ma15a.pdf.

[15] Y. Hao *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, 2021. DOI: 10.1016/j.cell.2021.04.048. [Online]. Available: https://doi.org/10.1016/j.cell.2021.04.048.

[16] E. Zhao *et al.*, "Spatial transcriptomics at subspot resolution with bayesspace," *Nature Biotechnology*, vol. 39, no. 11, pp. 1375–1384, Jun. 2021. DOI: 10.1038/s41587-021-00935-2. [Online]. Available: https://doi.org/10.1038/s41587-021-00935-2.

[17] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, Jun. 1991. DOI: 10.1080/01621459.1991.10475035. [Online]. Available: https://doi.org/10.1080/01621459.1991.10475035.

[18] L. Zhu, B. Miao, and H. Peng, "On sliced inverse regression with high-dimensional covariates," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 630–643, Jun. 2006. DOI: 10.1198/016214505000001285. [Online]. Available: https://doi.org/10.1198/016214505000001285.

[19] A. McDavid *et al.*, "Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments," *Bioinformatics*, vol. 29, no. 4, pp. 461–467, Dec. 2012. DOI: 10.1093/bioinformatics/bts714. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts714.

[20] G. Finak *et al.*, "MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data," *Genome Biology*, vol. 16, no. 1, Dec. 2015. DOI: 10.1186/s13059-015-0844-5. [Online]. Available: https://doi.org/10.1186/s13059-015-0844-5.

[21] V. Kleshchevnikov *et al.*, "Cell2location maps fine-grained cell types in spatial transcriptomics," *Nature Biotechnology*, vol. 40, no. 5, pp. 661–671, Jan. 2022. DOI: 10.1038/s41587-021-01139-4. [Online]. Available: https://doi.org/10.1038/s41587-021-01139-4.

[22] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," en, *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, Feb. 2021. DOI: 10.1186/s13638-021-01910-w. [Online]. Available: https://doi.org/10.1186/s13638-021-01910-w.

[23] M. R. Vahid *et al.*, "High-resolution alignment of single-cell and spatial transcriptomes with cytospace," *Nature Biotechnology*, Mar. 2023. DOI: 10.1038/s41587-023-01697-9. [Online]. Available: https://doi.org/10.1038/s41587-023-01697-9.