

Application of SPADE and Seurat to Between Group Expression

Yingzhen Wang

June 2025

1 Introduction

To identify differential expression among genes with spatial transcriptomic data from normal vs. Alzheimer's disease mouse brains, spatial pattern and differential expression (SPADE) analysis [1] was planned to be used. To make the whole research more robust, a cluster analysis was performed using the BASS package [2, 4], and the Seurat package [3] was used to achieve differential expression analysis, then compare the results from two different methods to determine the difference in the detection of spatial variance (SV) genes.

2 Method and Implementation

2.1 SPADE

This method implements a Gaussian process regression (GPR) machine learning model with a gene-specific Gaussian kernel that is capable of capturing the relationship between gene expression and spot groups by combining the spatial coordinates of spots [1]. In the research, SPADE offers a framework for detecting SV genes between groups using a crossed likelihood-ratio test.

2.1.1 Parameter Estimation and Test

Gene: $i = 1, \dots, P$ and Spatial Spot: $j = 1, \dots, N$. \mathbf{x}_{ij} is a k -covariate vector such that $\mathbf{x}_{ij} = [1, \text{batch}_j, \text{celltype}_j, \dots] \in \mathbb{R}^k$, and the intercept indicates the mean log-expression of the gene across all spots. γ_{i1} is used to measure the amount of expression variation explained by spatial patterns. γ_{i2} captures the random noise, which indicates the random variation not explained by covariates. Therefore, the residual error is i.i.d. from $\mathcal{N}(0, \gamma_{i2})$, and $\mathbf{z}_i = (z_{i1}, \dots, z_{iN})^\top \sim \text{MVN}(0, \gamma_{i1} \mathbf{K}_i)$ [1].

The Gaussian Kernel $K_i = \exp\left(-\frac{D^2}{2\theta_i^2}\right)$ is used to model the covariance matrix among spots, where D is the Euclidean distance matrix between all pairs of

spatial spots with spatial coordinate s and s' , θ_i is Length-scale hyperparameter controlling the curvature of the covariance to distance between spots (larger θ_i corresponds to a smoother covariance change). $\mathbf{K}_i \in \mathbb{R}^{N \times N}$ is a symmetric positive-definite matrix where each entry $K_i(j, j')$ measures spatial similarity between spot j and spot j' . Specifically, if two spots are close in space, then $D_{jj'} \rightarrow 0$, so $K_i(j, j') \approx 1$, and if two spots are far apart in space, then $D_{jj'} \rightarrow \infty$, so $K_i(j, j') \approx 0$ [1].

Based on the original model

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}^\top \beta_i + z_{ij} + \varepsilon_{ij} \\ z_{ij} &\sim \text{GP}(0, \gamma_{i1} \mathbf{K}_i) \Rightarrow \mathbf{z}_i = (z_{i1}, \dots, z_{iN})^\top \sim \text{MVN}(0, \gamma_{i1} \mathbf{K}_i) \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \gamma_{i2}) \Rightarrow \boldsymbol{\varepsilon}_i \sim \text{MVN}(0, \gamma_{i2} \mathbf{I}) \end{aligned}$$

Then, we derive the vector form of the original model as:

$$\mathbf{y}_i = \mathbf{X}_i \beta_i + \mathbf{z}_i + \boldsymbol{\varepsilon}_i$$

Therefore,

$$\Sigma_i = \text{Cov}(\mathbf{y}_i) = \text{Cov}(\mathbf{z}_i) + \text{Cov}(\boldsymbol{\varepsilon}_i) = \gamma_{i1} \mathbf{K}_i + \gamma_{i2} \mathbf{I}$$

where \mathbf{I} is an identity matrix of n .

The way to obtain the estimated $\Theta_i = (\beta_i, \theta_i, \gamma_{i1}, \gamma_{i2})$ follows the same procedure as to obtain the maximum likelihood estimate (MLE) [1]. The marginal likelihood function is

$$\begin{aligned} p(\mathbf{y}_i | \mathbf{x}_i, \Theta_i) &= \text{MVN}(\mathbf{y}_i | \mathbf{x}_i^\top \beta_i, \Sigma_i) \\ &= \frac{1}{(2\pi)^{N/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \mu_i)^\top \Sigma_i^{-1} (\mathbf{y}_i - \mu_i)\right) \end{aligned}$$

And the log marginal likelihood function is

$$\mathcal{L}(\mathbf{y}_i | \mathbf{x}_i, \Theta_i) = -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{y}_i - \mu_i)^\top \Sigma_i^{-1} (\mathbf{y}_i - \mu_i),$$

where μ_i is the mean gene expression level modeled by $\mu_i = \mathbf{x}_i^\top \beta_i$.

Then, we take the derivative of the log marginal likelihood with respect to one of the $\beta_i, \theta_i, \gamma_{i1}, \gamma_{i2}$ and get its corresponding MLE $\hat{\beta}_i, \hat{\theta}_i, \hat{\gamma}_{i1}, \hat{\gamma}_{i2}$. Maybe we need to check the second derivative and Hessian matrix to prove the Global Maximum. Next, obtaining the constrained MLE $\hat{\beta}_i$ under $H_0: \gamma_{i1} = 0$ and $\hat{\theta}_i$. Then, using SKAT (using a kernel matrix and tests whether the variance explained by these variants is significant) to get the

$$Q_i = (y_i - \mathbf{X}_i^\top \hat{\beta}_i)^\top \hat{\mathbf{K}}_i (y_i - \mathbf{X}_i^\top \hat{\beta}_i), \quad \text{where } \hat{\mathbf{K}}_i = \exp\left(-\frac{D^2}{2\hat{\theta}_i^2}\right)$$

Since Q_i follows a mixture of independent chi-square distributions, using the Davies method and the BH method to obtain the adjusted P-values of FDR, where the BH method adjusts the p-values to control the expected proportion of false positives among the rejected hypotheses, and FDR is defined as $\text{FDR} = \mathbb{E} \left[\frac{V}{R} \right]$, where V = number of false positives and R = number of total rejections. If adjusted p-value $< 0.05 \Rightarrow$ the spatial variation is significant \Rightarrow SV gene [1].

For identification between groups, SPADE first optimized $K_i = \exp \left(-\frac{D^2}{2\theta_i^2} \right)$, $\text{Cov}(\mathbf{y}_i) = \hat{\gamma}_{i1} K_i + \hat{\gamma}_{i2} \mathbf{I}$ [1]. Then for each gene in group A or group B, there are

$$\hat{\mathcal{L}}_A = -\frac{N_A}{2} \log(2\pi) - \frac{1}{2} \log \left| \hat{\Sigma}_i^{(A)} \right| - \frac{1}{2} \left(\mathbf{y}_i^{(A)} - \mathbf{X}_i^{(A)\top} \hat{\beta}_i^{(A)} \right)^\top \left(\hat{\Sigma}_i^{(A)} \right)^{-1} \left(\mathbf{y}_i^{(A)} - \mathbf{X}_i^{(A)\top} \hat{\beta}_i^{(A)} \right)$$

and

$$\hat{\mathcal{L}}_B = -\frac{N_B}{2} \log(2\pi) - \frac{1}{2} \log \left| \hat{\Sigma}_i^{(B)} \right| - \frac{1}{2} \left(\mathbf{y}_i^{(B)} - \mathbf{X}_i^{(B)\top} \hat{\beta}_i^{(B)} \right)^\top \left(\hat{\Sigma}_i^{(B)} \right)^{-1} \left(\mathbf{y}_i^{(B)} - \mathbf{X}_i^{(B)\top} \hat{\beta}_i^{(B)} \right)$$

where the optimal kernel matrix is $\hat{\Sigma}_i^{(A)} = \hat{\gamma}_{i1}^{(A)} \hat{K}_i^{(A)} + \hat{\gamma}_{i2}^{(A)} \mathbf{I}$, and no further optimization is required.

Then, for instance, evaluating the log likelihood for group B under the group A parameters estimations, and the log likelihood for group A under the group B parameters estimations [1].

$$\hat{\mathcal{L}}_{AB} = \log p \left(\mathbf{y}_i^{(B)} \mid \mathbf{X}_i^{(B)}, \hat{\theta}_i^{(A)} \right)$$

$$= -\frac{N_B}{2} \log(2\pi) - \frac{1}{2} \log \left| \hat{\Sigma}_i^{(A)} \right| - \frac{1}{2} \left(\mathbf{y}_i^{(B)} - \mathbf{X}_i^{(B)\top} \hat{\beta}_i^{(A)} \right)^\top \left(\hat{\Sigma}_i^{(A)} \right)^{-1} \left(\mathbf{y}_i^{(B)} - \mathbf{X}_i^{(B)\top} \hat{\beta}_i^{(A)} \right)$$

$$H_0: (\gamma_{i1}^{(A)}, \theta_i^{(A)}) = (\gamma_{i1}^{(B)}, \theta_i^{(B)}) \text{ and } (\gamma_{i2}^{(A)}, \theta_i^{(A)}) = (\gamma_{i2}^{(B)}, \theta_i^{(B)}) \\ \iff \text{no differential spatial patterns}$$

$$H_1: (\gamma_{i1}^{(A)}, \theta_i^{(A)}) \neq (\gamma_{i1}^{(B)}, \theta_i^{(B)}) \text{ or } (\gamma_{i2}^{(A)}, \theta_i^{(A)}) \neq (\gamma_{i2}^{(B)}, \theta_i^{(B)}) \\ \iff \text{there are differential spatial patterns}$$

Using Likelihood Ratio Test (LRT) to compare the likelihood under the null ($\hat{\mathcal{L}}_{AB}$) with the maximum likelihood ($\hat{\mathcal{L}}_B$) [1], and similar for comparison be-

tween $\hat{\mathcal{L}}_{BA}$ and $\hat{\mathcal{L}}_A$

$$\begin{aligned}\lambda(\mathbf{y}_i) &= \frac{\sup_{\theta_i \in \Theta_0} \mathcal{L}(\theta_i | \mathbf{y}_i^{(B)}, \mathbf{X}_i^{(B)})}{\sup_{\theta_i \in \Theta} \mathcal{L}(\theta_i | \mathbf{y}_i^{(B)}, \mathbf{X}_i^{(B)})} = \frac{\mathcal{L}(\hat{\theta}_i^{(A)} | \mathbf{y}_i^{(B)}, \mathbf{X}_i^{(B)})}{\mathcal{L}(\hat{\theta}_i^{(B)} | \mathbf{y}_i^{(B)}, \mathbf{X}_i^{(B)})} \iff -2 \log \lambda(\mathbf{y}_i) \\ &= -2 \log \left(\frac{p(\mathbf{y}_i^{(B)} | \mathbf{X}_i^{(B)}, \hat{\theta}_i^{(A)})}{p(\mathbf{y}_i^{(B)} | \mathbf{X}_i^{(B)}, \hat{\theta}_i^{(B)})} \right) \iff 2 [\hat{\mathcal{L}}_B - \hat{\mathcal{L}}_{AB}] = \lambda_{LR}^{(A)} \xrightarrow{d} \chi_1^2\end{aligned}$$

Similarly,

$$2 [\hat{\mathcal{L}}_A - \hat{\mathcal{L}}_{BA}] = \lambda_{LR}^{(B)} \xrightarrow{d} \chi_1^2$$

$$\iff \lambda_{LR} = 2 [(\hat{\mathcal{L}}_A + \hat{\mathcal{L}}_B) - (\hat{\mathcal{L}}_{AB} + \hat{\mathcal{L}}_{BA})] \stackrel{H_0}{\sim} \chi_1^2$$

Finally, we calculated the BH adjusted P-value using F-test,

$$F_i = \frac{\hat{\gamma}_{i1}^{(A)}}{\hat{\gamma}_{i1}^{(B)}} \sim F(d_A, d_B),$$

where both $d_A = d_B = 1$ are used as a conservative approximation when each variance estimate is based on independent spatial models.

2.1.2 Data Preprocessing

As the dimensions of two datasets are 32285×3541 and 32285×3476 and SPADE is a single-threaded algorithm, the reduction of computational cost is needed to improve the efficiency. Seurat [3] and a capping algorithm are utilized to achieve feature selection, spot filtering, and spot merging.

To initialize the Seurat object for normal and AD datasets [3], the number of spots in which the a gene expressed are selected to determine the minimum number of spots for a gene to be classified as significant. In the normal dataset, 9 genes are randomly selected that expressed exactly in 5, 10, 15, 20, 25, 30, 35, 40 spots, then 9 plots for normal and AD datasets are generated to determine the range of minimum number of spots. After comparing the two plots Figure 1, when spot counts greater or equal to 20, almost all gene expressions are significant different between normal slice and AD slice. Additionally, spot counts equal to 10 is the last time that the difference in gene expression is insignificant. Therefore, 11 genes are randomly selected which expressed in 10 to 20 spots, and 11 plots are created respectively to determine the robust number of minimum spots. After comparing the two plots Figure 2, genes expressed in more than 11 spots are continuously significant different between normal and AD. Finally, the minimum spot counts are deicide to be 11.

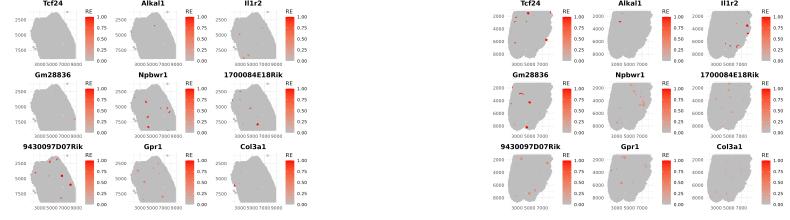


Figure 1: Selected genes expressed in 5 to 40 spots in normal and AD datasets

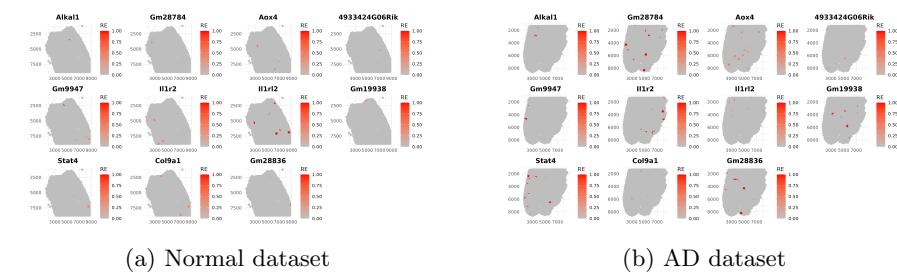


Figure 2: Selected genes expressed in 10 to 20 spots in normal and AD datasets

After the procedure of feature selection, 16279 and 17116 unique genes are remained in normal and AD objects, respectively. After normalizing the data using log normalization [3], filtering spots is also a fancy way to reduce the computational costs. However, directly remove a large number of spots will compromise the accuracy and reliability of the study. Based on Figure 3, in order to avoid a large number of removal spots, spots in normal dataset are filtered that have unique genes (nFeature RNA) detected over 6500 or less than 1000; total number of read counts (nCount RNA) detected over 32000 or less than 1000; greater than 40 percent mitochondrial counts (percent.mt). Based on Figure 4, similarly, spots in AD dataset are filtered that have unique genes (nFeature RNA) detected over 8000 or less than 500; total number of read counts (nCount RNA) detected over 42000 or less than 1000; greater than 35 percent mitochondrial counts (percent.mt). Finally, the dimension of normal count matrix reduced to 16279×3455 and 86 spots are removed and 17116×3441 and 35 spots are removed for AD count matrix.

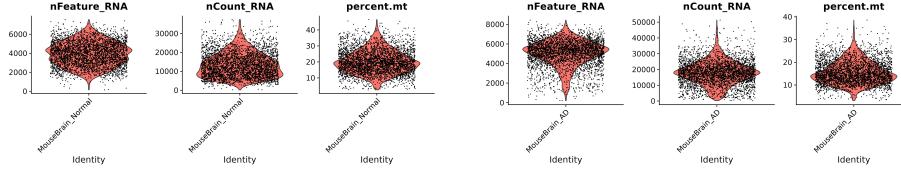


Figure 3: Violin plot for normal normalized data

Figure 4: Violin plot for AD normalized data

Finally, a binning algorithm is created to merge the spots to further improve the computational efficiency. Specifically, the binning algorithm aggregates spatial expression data by dividing the tissue area into square bins of size 250×250 units. For each spot, a bin ID is created using its spatial coordinates by applying floor to $x / \text{bin size}$ and $y / \text{bin size}$. These bin IDs are used to construct a sparse design matrix where each row corresponds to a spot and each column to a bin, with entries indicating bin affiliation. The expression matrix is then multiplied by this design matrix to sum the expression values within each bin. To obtain the average expression per bin, the resulting matrix is normalized by the number of spots in each bin. Finally, the coordinates of each bin are calculated as the mean x and y coordinates of the spots. After performing the algorithm, the dimension of normal count matrix is reduced to 16279×659 , and 17116×666 for AD count matrix.

2.2 Cluster Analysis

BASS (Bayesian Analytics for Spatial Segmentation) is used to achieve the joint cell type clustering and spatial domain detection simultaneously from multiple spatial transcriptomic data within a Bayesian hierarchical modeling framework [2].

2.2.1 Model and Inference

Assume there are S samples. For each sample $s \in \{1, \dots, S\}$, n_s spots and G genes are observed. Let $\mathbf{Y}_s \in \mathbb{R}^{n_s \times G}$ be the log-normalized gene expression matrix, $Z_s = (z_{si})_{i=1}^{n_s}$ be the latent cluster assignments for sample s , $D_s = (d_{si})_{i=1}^{n_s}$ be the domain labels for sample s , and x_{si} denote the spatial coordinates of spot i in sample s . Each domain $d \in \{1, \dots, D\}$ is shared across samples and contains a set of mixture components (clusters) that are modeled using a Gaussian mixture model [2].

The likelihood for spot i in sample s is:

$$y_{sig} | z_{si} = k, d_{si} = d \sim \mathcal{N}(\mu_{kg}^{(d)}, \sigma_g^2)$$

where $\mu_{kg}^{(d)}$ is the mean expression for gene g in cluster k and domain d , and σ_g^2 is a gene-specific variance shared across all clusters and domains [2].

The domain labels D_s are modeled with a Potts model to encourage spatial smoothness:

$$P(D_s) \propto \exp \left(\beta \sum_{i \sim j} \mathbf{1}(d_{si} = d_{sj}) \right)$$

where the sum is over neighboring spot pairs $i \sim j$, and β is a spatial smoothness parameter controlling domain contiguity [2]. Given the domain d_{si} , the cluster assignment z_{si} is modeled as:

$$z_{si} \mid d_{si} = d \sim \text{Categorical}(\pi_d)$$

where $\pi_d = (\pi_{d1}, \dots, \pi_{dK})$ is the domain-specific mixing proportion [2]. The mean parameters $\mu_{kg}^{(d)}$ and variances σ_g^2 follow conjugate priors, and the mixing proportions π_d have Dirichlet priors:

$$\begin{aligned} \mu_{kg}^{(d)} &\sim \mathcal{N}(m_g, \tau_g^2), \quad \sigma_g^2 \sim \text{InvGamma}(a_0, b_0) \\ \pi_d &\sim \text{Dirichlet}(\alpha_d) \end{aligned}$$

The joint posterior distribution is:

$$\begin{aligned} P\left(\{Z_s\}, \{D_s\}, \{\pi_d\}, \{\mu_{kg}^{(d)}\}, \{\sigma_g^2\} \mid \{Y_s\}\right) &\propto P\left(\{Y_s\} \mid \{Z_s\}, \{\mu_{kg}^{(d)}\}, \{\sigma_g^2\}\right) \\ &\cdot P(\{Z_s\} \mid \{\pi_d\}, \{D_s\}) \cdot P(\{D_s\}) \\ &\cdot P(\{\mu_{kg}^{(d)}\}) \cdot P(\{\pi_d\}) \cdot P(\{\sigma_g^2\}) \end{aligned}$$

BASS uses Gibbs sampling to infer the posterior distributions of the latent variables and parameters [2]. Of the procedures, first, domain labels \mathbf{D}_s are updated using a Metropolis-Hastings step based on the Potts prior and expression likelihood. Second, cluster assignments z_{si} are sampled from their full conditional posterior based on the observed data and current π_{kd} . Third, the cluster proportions π_{kd} are updated using a Dirichlet distribution. Fourth, the cluster-specific means $\mu_{kg}^{(d)}$ are updated using a conjugate Gaussian posterior. Finally, the gene-specific variances σ_g^2 are sampled from an inverse-gamma posterior [2].

2.2.2 Parameter Selection

Number of cell types and spatial domains are two important hyper-parameters that need to be specified in BASS [2]. As mentioned in the article, miss-specified number of spatial domains and increased rare cell types do not have much influence on performance of cell type clustering, an over-specified number of cell type clusters did not influence the performance of spatial domains either. Only under-specified number of cell type clusters reduced the performance of spatial domain detection, because the algorithm merged multiple true cell types into the same cluster, which led to worse cell type classification [2]. Therefore, larger values of these two hyper-parameters are preferred to avoid the reduction of performance.

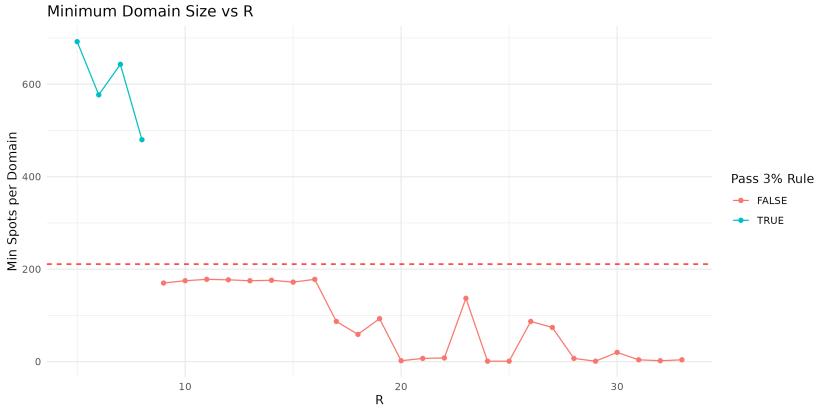


Figure 5: Domain sizes evaluated by the minimum spots baseline

First, two named lists of count matrix and coordinate information are created as the inputs for the BASS functions. Then, the maximum number of spatial domains (with greater than 3 percent of cells per domain) is calculated as: Total $3,541 + 3,476 = 7,017$ spots, and then $7017 \times 0.03 \approx 211$ minimum spots per domain, finally the maximum number of domains is $7017/211 \approx 33$. To ensure that the heterogeneity in gene expression can be fully accounted and avoid noise from small domains, we determine the range of $R = (5, 33)$ and the number of cell types to 20 [4]. As shown in Figure 5, domain size equal to 8 is the largest one among those who satisfy the minimum spots baseline that with greater than 3 percent of cells per domain.

Additionally, the hierarchical modeling structure with an intermediate layer was used to explicitly model distinct gene expression of different cell types, avoiding the problem of insufficient ability of using same distribution to capture the gene expression heterogeneity across cell types [2]. The Swendsen-Wang algorithm is known to have a much better mixing rate than the Gibbs sampling algorithm using by BayesSpace, hence BASS achieve greater accuracy [2].

3 Results and Interpretation

3.1 Visualizations

Several plots are generated to represent the gene expression spatially. In order to better display the figures, only the top genes are selected to show the expression in both slices based on the adjusted p-values from the SPADE test or the adjusted p-values from the Seurat test. For clearly show the distribution of spatial domains, the domain labels and spatial coordinates are merged based on matching group labels and spot names. Additionally, the expression plot of the top 15 of genes that are detected as significant different by one of the method

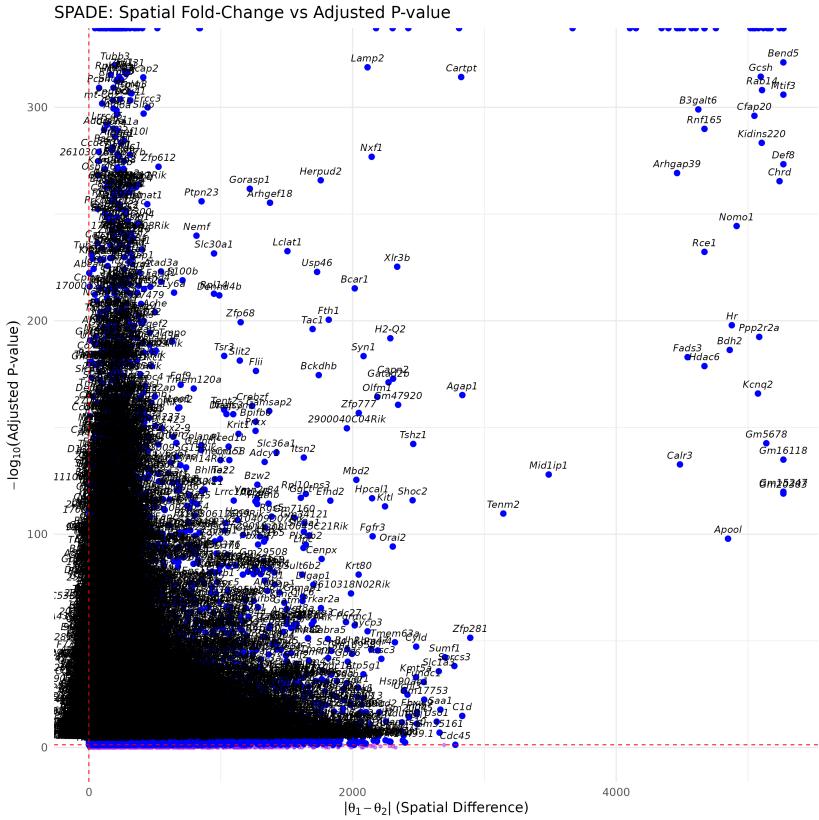


Figure 6: Between group gene expression evaluation

while as nonsignificant different by the other method are created.

3.1.1 SPADE

As shown in Figure 6, the x axis indicates the magnitude of spatial difference between normal and AD group. Additionally, since most of the significant differences are along with very low adjusted p-values that hard to visualize directly, taking the negative base-10 logarithm is able to turn smaller p-values into larger, positive values. Finally, genes distributed at the top right corner have both large spatial difference and strong significance, which are the most significantly spatially differentially expressed genes between the two groups. As contrast, genes around bottom left corner have small between group difference and low significance. Moreover, the horizontal red dashed line at $-\log_{10}(0.05) \approx 1.3$ is the 0.05 significance threshold, and the vertical line near 0 indicates no spatial difference.

As shown in Figure 7, the relative expression (RE) is a normalized measure of

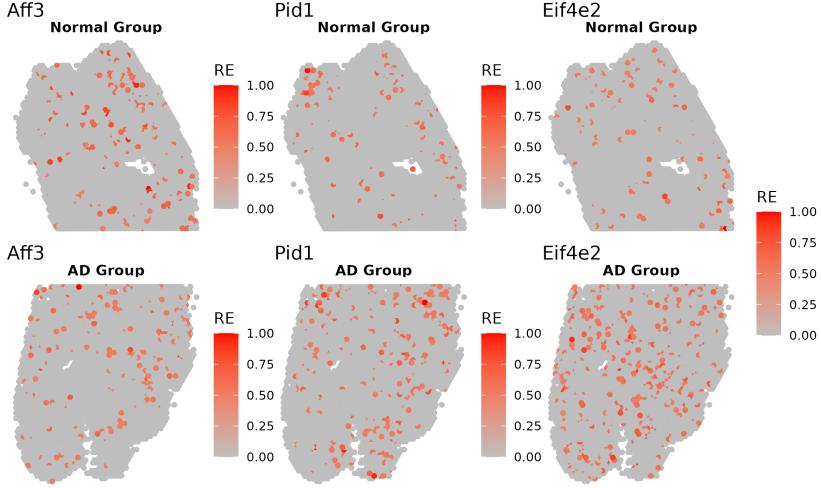


Figure 7: Spatial patterns for the top three genes

gene expression at each spatial spot and is calculated as $RE_{ij} = x_{ij}/\max_j(x_{ij})$, where x_{ij} refers to the raw counts of gene i at spot j and $\max_j(x_{ij})$ the maximum expression of that gene across all spots in the group. RE is used to directly and clearly visualize the comparison of spatial patterns within each gene, even if genes have vastly different expression scales, and highlights relative spatial localization rather than absolute expression magnitude. For gene "Aff3", the expression is localized, especially around the top-right region in the normal slice with concentrated and isolated RE values. As contrast, the expression is more widespread across the AD slice with stronger RE signals in various regions. For gene "Pid1", the expression appears very sparse, with scattered and weak RE signals in the normal slice, but is much stronger, with many more spots showing higher RE values in the AD slice. For gene "Eif4e2", the expression is moderate and uniformly distributed across the tissue with wide-spread but low-intense RE values in the normal slice, while the expression becomes more intense and covers a larger fraction of the tissue with clearly dense clusters of high RE in the AD slice.

3.1.2 Cluster Analysis

As shown in Figure 8, the domain structure differs between AD and WT. For instance, Domain 4 and Domain 7 appear more expanded in AD compared to WT. Additionally, some domains, such as Domain 3 and Domain 8, may shrink or relocate, indicating potential degeneration. These results suggest that spatial

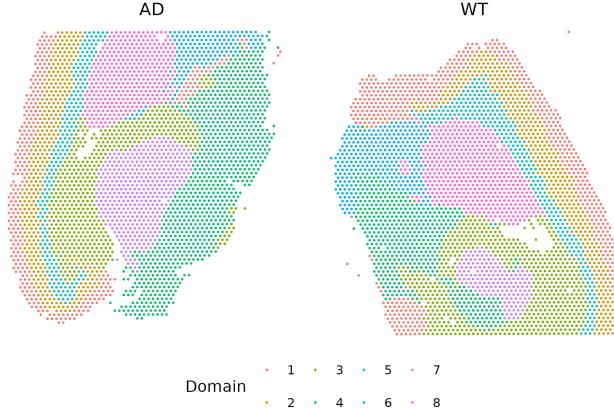


Figure 8: Spatial domains distribution

reorganization of transcriptional domains occurs in AD pathology. Such reorganization may reflect neuronal loss, glial proliferation, or regional dysfunction. BASS’s domain alignment makes it easier to detect domain-specific gene dysregulation and to localize pathological changes more accurately, offering valuable insights into region-specific vulnerability in Alzheimer’s disease [2, 4].

Additionally, to achieve the differential expression analysis for the domain-specific genes, a Wilcoxon test from Seurat is used to compare the distribution of expression counts in AD vs. normal slice, ranks the counts across both slices, and finally produce the FDR-adjusted p-value to assess whether the distributions differ significantly [3]. Since there are totally 8 domains and each domain contained 32285 unique genes, the final results have 258234 rows and classified by domain labels.

3.1.3 Comparison

As shown in Table 1, despite processing with fewer input genes, SPADE detected a higher number and proportion of significant genes compared to Seurat. Such outcomes suggests that SPADE is more effective at identifying spatially varying signals than cluster-based comparison using Seurat, when applied to raw gene expression data with appropriate pre-filtering. The higher detection rate from SPADE may reflect its better sensitivity to raw spatial data. However, as mentioned before, since the possible low expression genes were filtered by Seurat before tested by SPADE [3], such results in the table may be influenced. Therefore, a combination of SPADE and Seurat is more preferred to test the significance of spatial data.

The Next goal is to check overlapping and conflict genes, which means a gene detected as significantly expressed by both methods or as significantly expressed by one of the method. As shown in Table 2, 9639 gene expression are classified as

Table 1: Comparison of two methods in terms of significant gene expression

Method	Total Genes	Significant Genes	Percent (%)
Seurat	32282	11812	36.59
SPADE (Capping)	16279	13085	80.38

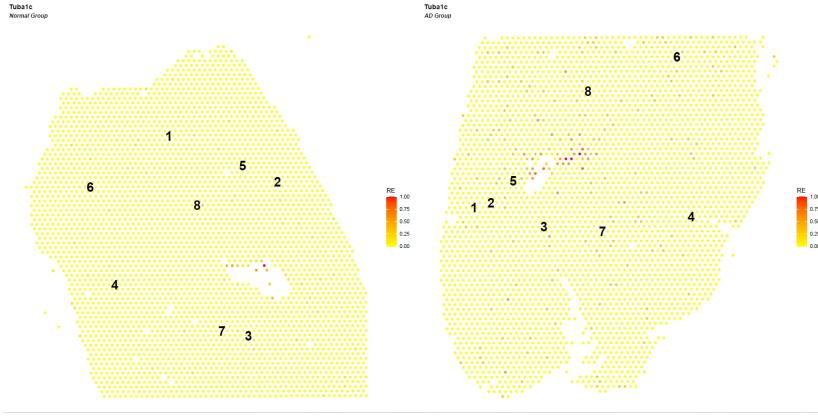


Figure 9: Conflict gene expression "Tuba1c"

significant by both two methods. Among the conflict genes, 2173 detections of significant gene expression are made by Seurat but not by SPADE. Additionally, almost 37% of all the significant genes are conflict, which may indicates that there are some disagreements between the results from two methods.

Table 2: Comparison of Significant Gene Results between Seurat and SPADE

Overlapping	Conflict	Only in Seurat	Only in SPADE
9639	5619	2173	3446

To more intuitively present the conflict gene expression and better infer why the other method categorized it as non-significant gene, a combination of paired plots are created to show the spatial expression of the top 15 genes that selected by the the lowest 15 adjusted p-values from the two methods, and only several representative plots were shown as illustrations. Figure 9 represents the spatial expression patterns of the gene "Tuba1c" with relative expression (RE) intensity. Although "Tuba1c" appears largely lowly expressed across most of the tissue in both groups, there is a localized increase in expression in a small region of the AD sample. This spatially confined upregulation was captured by SPADE, thus classified the gene as significant. Along with the Figure 8 and the domain labels in Figure 9, the highly expressed spots are belong to domain 3 and have similar relative expression intensity. Therefore, cluster-based analysis using Seurat did not detect "Tuba1c" as significant.

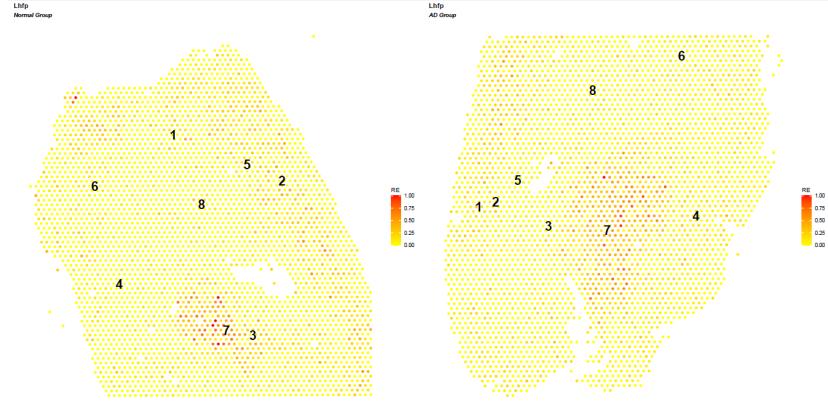


Figure 10: Conflict gene expression "Lhfp"

Figure 10 represents the spatial expression patterns of the gene "Lhfp" with relative expression (RE) intensity. Based on the DE analysis, the lowest adjusted p-value is from domain 7, and cluster-based analysis detects "Lhfp" as significant. Along with the Figure 8 and the domain labels in Figure 10, the possible reason is that the relative gene expression in domain 7 is significantly different as shown by the density degree. However, it is not classified as SV gene by SAPDE, which may because that the RE intensity is similar across several specific corresponding regions in the two tissues.

Figure 11 represents that the gene expression intensity of "Wbp1" is uniformly and evenly spread across each region and domains. However, this gene is detected as SV gene by SAPDE and non-SV gene by cluster analysis. SPADE identifies spatially structured expression variation across the tissue, capturing the localized differences in gene expression. In contrast, the domain-based method performs comparisons of average expression levels within each domain between conditions. If a domain contains both high and low expressing region, these variations may cancel out in the averaging process, leading to a nonsignificant result.

4 Conclusion

This research compared the effectiveness of SPADE and Seurat-based clustering in identifying spatially variable gene expression differences between normal and Alzheimer's disease (AD) mouse brain tissues. SPADE, utilizing Gaussian process models, identified a larger proportion of significant genes from a smaller input gene set, indicating high sensitivity to spatial expression patterns that are not spread across several regions [1]. In contrast, the Seurat-based clustering approach, combined BASS for domain segmentation, captured broader domain-specific differences in gene expression [3,4].

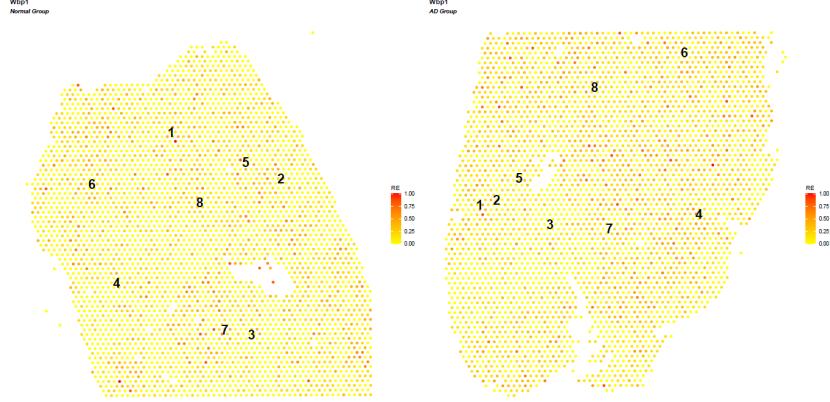


Figure 11: Conflict gene expression "Wbp1"

The significant number of conflict genes between the two methods highlights their methodological differences. SPADE tended to detect genes with spatially gene expression without regional restriction, while Seurat captured genes with consistent expression differences across broader regions [1, 3]. Integrating SPADE with filtered data by Seurat allows for improved resolution and interpretability in spatial transcriptomics. Such combined framework indicating high sensitivity to local spatial expression heterogeneity, especially relevant in studies of complex diseases such as Alzheimer's, where genes may show abnormal expression only in very specific spots. In contrast, the Seurat-based approach identifies transcriptional domains using clustering and tests for average expression differences within these domains, which ensure particularly eligibility for detecting domain-specific difference [3,4]. However, it may overlook genes that are differentially expressed in only subpart of domains, especially when those domains include heterogeneous expression patterns that average out.

The overlap analysis revealed that 37% of all significant genes were discordant between the two methods, reflecting these fundamental differences. Visualization of conflict genes confirmed that SPADE frequently detects localized difference that do not align with the Seurat domains [1], while Seurat may detect domain-specific difference that non-significant from SPADE due to shared spatial variance patterns[3].

5 Limitations

In the method and comparison section, the selection the number of domains only based on the empirical criteria, and does not account for the influence of effect size when evaluating the significance of between-group gene expression. Under such conditions, some potentially conflicting results, like Figure 12 and a large number of conflict genes, may arise. As shown in Figure 12, only a

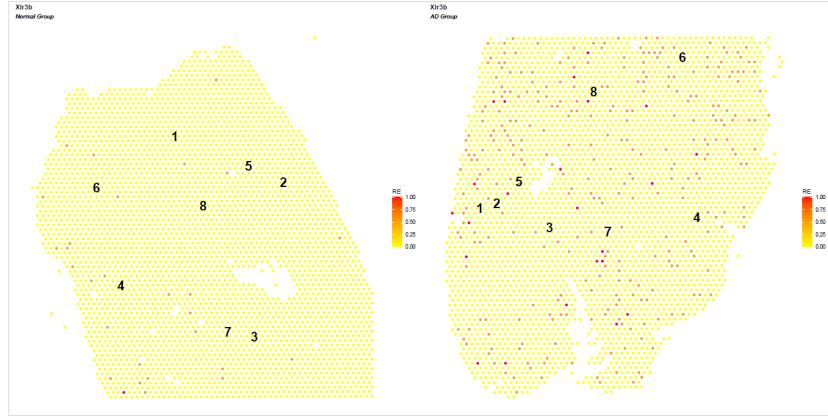


Figure 12: Conflict gene expression "Xlr3b"

very small part of domains 4,6, and 8 in normal tissue is highly expressed. As comparison, the highly expressed spots are almost across every domains of the AD tissue. Therefore, the result is supposed to indicate that the gene "Xlr3b" expression is significantly different between normal and AD tissue. However, since the effect size was not set up previously, the current domain-based result cannot detect such between-group difference.

To improve the future research, volcano plots should be created to determine the most appropriate effect size to the evaluation of between-group difference in gene expression. Additionally, to diagnose the disagreement of two methods, one method is to using an external validation set (known AD genes) to compare the reliability accuracy of the results from SPADE versus domain-based differential expression methods.

6 Reference

- [1] Qin, F., Luo, X., Lu, Q., Cai, B., Xiao, F., & Cai, G. (2024). Spatial pattern and differential expression analysis with spatial transcriptomic data. *Nucleic Acids Research*, 52(21), e101. <https://doi.org/10.1093/nar/gkae962>
- [2] Li, Z., Zhou, X. (2022). BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biology*, 23, 168. <https://doi.org/10.1186/s13059-022-02734-7>
- [3] Satija Lab. (2023, October 31). Seurat – Guided clustering tutorial. https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

- [4] Li, Z. (2023, April 2). *Software tutorial*. BASS. <https://zhengli09.github.io/BASS-Analysis/about.html>