# SPADE: a spatial pattern and differential expression analysis method with spatial transcriptomic data

Fei Qin, Feifei Xiao, Guoshuai Cai

Last updated: 04/03/2023

## 1. Introduction to the SPADE method

To identify spatially variable (SV) genes with spatial transcriptomic data, the SPADE method was developed based on a Gaussian process regression (GPR) model with Gaussian kernel. GPR model can model the relationship between gene expression and other covariates (i.e., cell groups) incorporating the spatial information of multiple cells. Besides detecting SV genes within groups, SPADE also provided a framework to identify SV genes between different treatment conditions. The framework of the SPADE method is summarized and illustrated in Figure below. First, original read counts data were normalized into continuous data using a two-step normalization strategy. Second, instead of using a fixed length scale hyperparameter across genes in the kernel function of GPR model, SPADE estimated the optimal hyperparameter for each gene to improve the accuracy of SV gene identification. To identify SV genes within groups, hypothesis testing was conducted based on a quadratic score statistic with a Davies method to compute the P value. With SV gene detection between groups, SPADE exchanged the optimal hyperparameters estimated in two groups and utilized a crossed likelihood-ratio test to calculate the P value for each gene.

## 2. Installation

```
library(devtools)
install_github("thecailab/SPADE")
```

## 3. Data

To help illustrate how SPADE package can be applied, the SeqFISH dataset was provided in the package. The SeqFISH dataset was collected on the mouse hippocampus with 249 genes

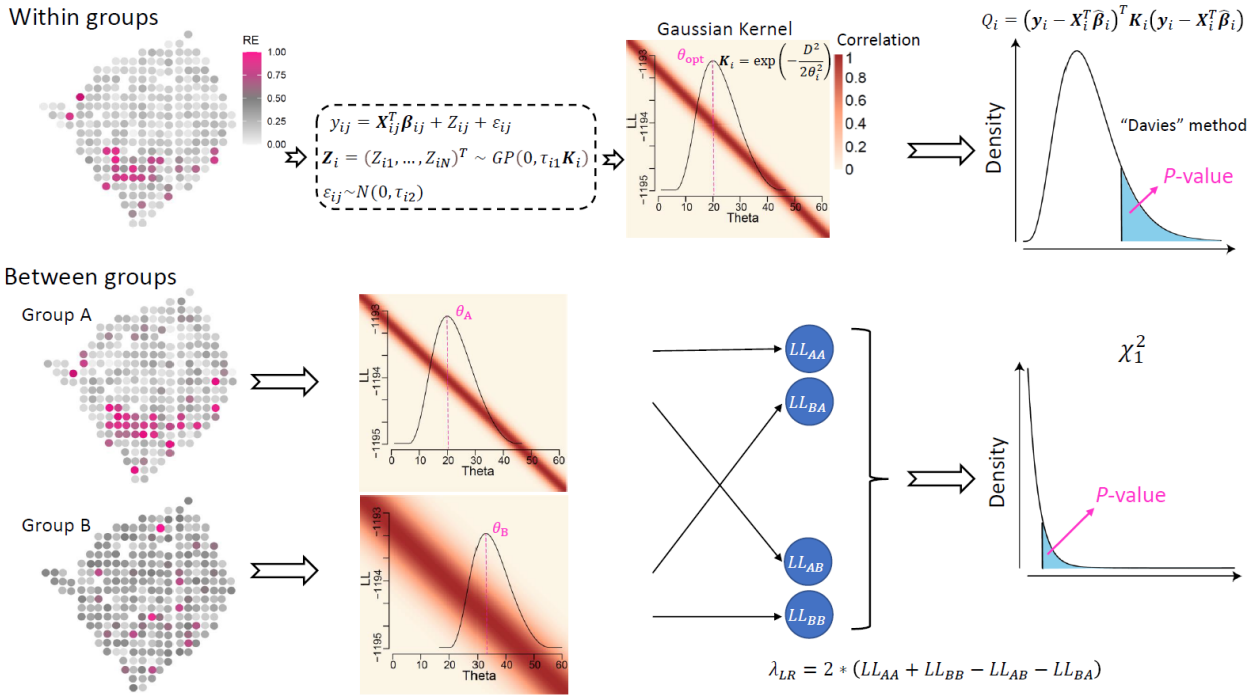
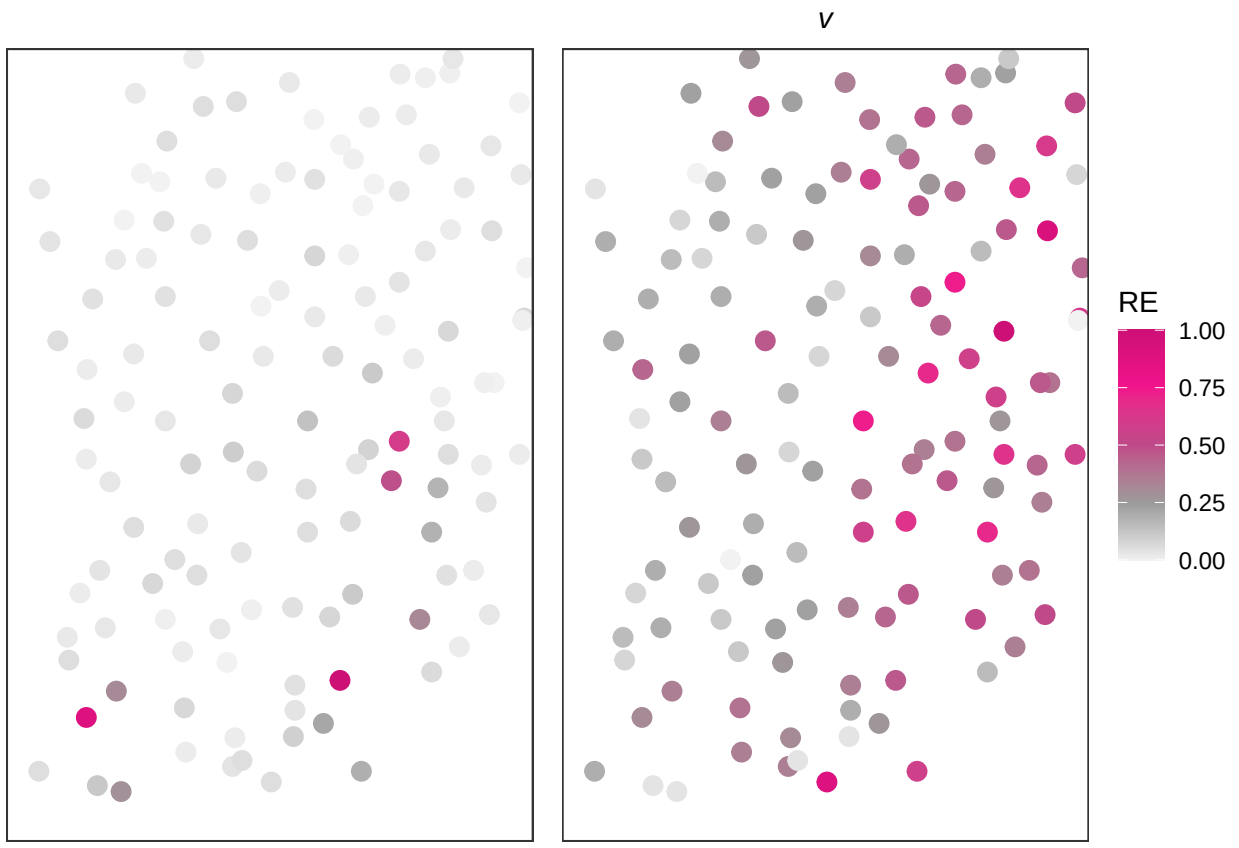



Figure 1: FLCNA framework

measured on 257 spots. After filtering out cells located at the boundary to relieve border artifacts, a final set of 249 genes measured on 131 spots were retained for the analysis. Two data files were given in the package including read counts data and coordinates information data. The application of SPADE includes steps of normalization, parameters estimation and testing.

```
library(SPADE)
data(SeqFISH)
data(info)
readcounts <- SeqFISH
```




```
dim(readcounts)
```

```
## [1] 249 131
```

```
dim(info)
```

```
## [1] 131   2
```

```
readcounts[1:5,1:5]
```

```
##        C1 C2 C3 C4 C5
## Tal1    3  3 17 20  6
## Dmbx1  11  9  1  4  0
## Emx2   13 14  2  9  0
## Uncx    5 21  3  5  4
## Paxip1 15  9 10 13  7
```

```
head(info)
```

```
##      x   y
## C1 686 352
## C2 488 572
## C3 614 698
## C4 516 726
## C5 308 208
## C6 204 225
```

# 4. Identifying SV genes within groups

To characterize complex tissues with spatial transcriptomics, the identification of SV genes within groups is an essential first step. These identified SV genes can be defined as genes with uneven, aggregated, or patterned spatial distribution of gene expression magnitudes, and they play important role in biomarker discovery and drug development.

## 4.1 Normalization

A two-step normalization strategy was implemented to transformed original read counts data into continuous data. First, the mean-variance dependency was stabilized using an Anscombe's transformation strategy. Second, a linear regression was utilized to regress out library size from the above transformed expression data. After this step, the normalized continuous data will be utilized for the parameter estimation and hypothesis testing in the SPADE method. SPADE_norm() R function is used for the normalization.

```
data_norm <- SPADE_norm(readcounts=as.matrix(readcounts), info=info)
data_norm[1:5,1:5]
```

```
##               C1        C2        C3        C4        C5
## Tal1   -5.170314 -4.796102 -3.811595 -3.792504 -3.766825
## Dmbx1  -2.487973 -2.370436 -3.743806 -3.226837 -3.427745
## Emx2   -5.239013 -4.743123 -6.342547 -5.558433 -5.908069
## Uncx   -6.217700 -4.635360 -6.399702 -6.241216 -5.194450
## Paxip1 -5.726922 -5.677565 -5.932052 -5.872189 -5.119711
```

## 4.2 Parameter estimation

The GPR model was used to model the expression of each gene across cells with different locations. Theoretically, the problem of finding SV genes in the SPADE method is to test

how well the candidate covariance matrix fits the spatial transcriptomic data. For each gene, we estimated the optimal length-scale hyperparameter in the Gaussian kernel to increase the accuracy of SV gene identification. SPADE_estimate() R function is used for the parameter estimation.

```
Est <- SPADE_estimate(expr_data=data_norm, info=info)
```

```
head(Est)
```

```
##   GeneID theta_Gau   Lik_Gau
## 1      1  11.22829 -91.34812
## 2      2  32.89034 -87.74524
## 3      3 130.70081 -83.34457
## 4      4  29.62392 -77.91255
## 5      5  44.91683 -67.52400
## 6      6  58.66151 -63.95050
```

### 4.3 Testing

After the optimal length-scale hyperparameter was estimated, P-value for each gene was computed based on a quadratic score statistic with a Davies method.

```
Test_res <- SPADE_test(object=data_norm, location=info, para=Est)
```

```
head(Test_res)
```

```
##   geneid        Q     Pvalue Adjust.Pvalue
## 1   Tal1 65.77883 0.44340157     0.5390583
## 2  Dmbx1 72.47503 0.19988947     0.4182561
## 3   Emx2 63.48979 0.26133072     0.5013248
## 4   Uncx 66.14991 0.40370188     0.5390583
## 5 Paxip1 80.50018 0.09045089     0.2618869
## 6 Ctnnb1 78.90114 0.13086639     0.3371722
```

```
sum(Test_res$Adjust.Pvalue < 0.05)
```

```
## [1] 38
```

## 5. Identifying SV genes between groups

SPADE was the first method to investigate SV genes between groups with spatial transcriptomic data. The identification of SV genes between groups was essential for understanding the changes of spatial patterns with different treatment conditions or different time phases.

## 5.1 Normalization

Still, the same two-step normalization strategy was implemented for read counts data from each group to transform original data into continuous data.

## 5.2 Parameter estimation and testing