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

Fei Qin, Xizhi Luo, Guoshuai Cai

Last updated: 07/06/2023

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. 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

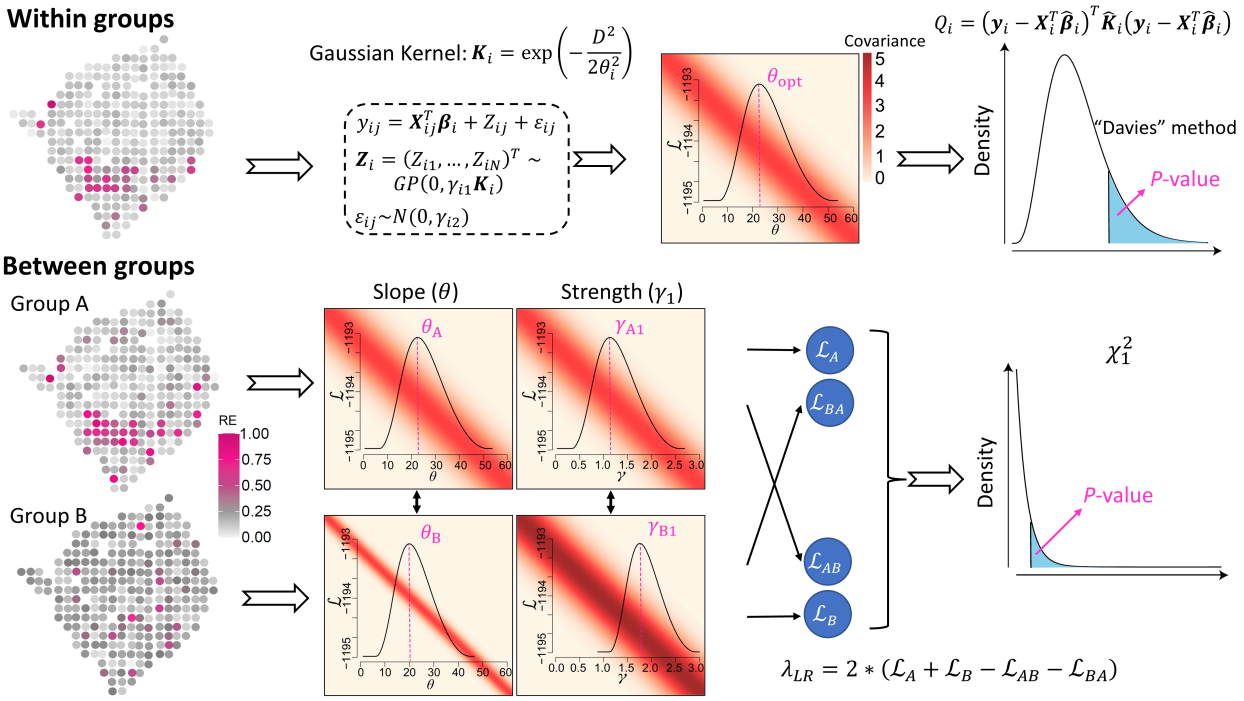


Figure 1: FLCNA framework

with uneven, aggregated, or patterned spatial distribution of gene expression magnitudes, and they play important role in biomarker discovery and drug development.

3.1 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 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
```

3.2 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)
```

3.3 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. The output of SPADE_estimate() includes 1) Gene ID; 2) theta_Gau: Optimal length-scale hyperparameter in kernel function estimated for each gene. 3) Lik_Gau: Maximum likelihood computed for each gene.

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

```
head(Est)
```

```
##   GeneID theta_Gau Gamma_hat   Lik_Gau
## 1      1  11.56163 0.8391345 -90.48894
## 2      2 152.79242 2.6691587 -85.29750
## 3      3 175.63314 5.9705723 -82.87845
## 4      4  30.80435 3.8409548 -77.50796
## 5      5 106.05900 2.8361674 -67.11764
## 6      6 145.60869 3.5697706 -63.19006
```

3.4 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. SPADE_test() is used for computing P value for each gene. The output of SPADE_test() includes
1) geneid: Gene name.
2) Q: A score statistic used to calculate P value. Q follows a mixture of independent chi-square distributions with mixing weights that depend on the eigenvalues of the kernel matrix.
3) Pvalue: Based on a chi-square mixture distribution and Q value, an exact method based on the Davies method is utilized to compute P-values for genes to define SV genes.
4) Adjusted.Pvalue: P-values across all genes are adjusted with the Benjamini and Hochberg method to correct for occurrence of false positives.

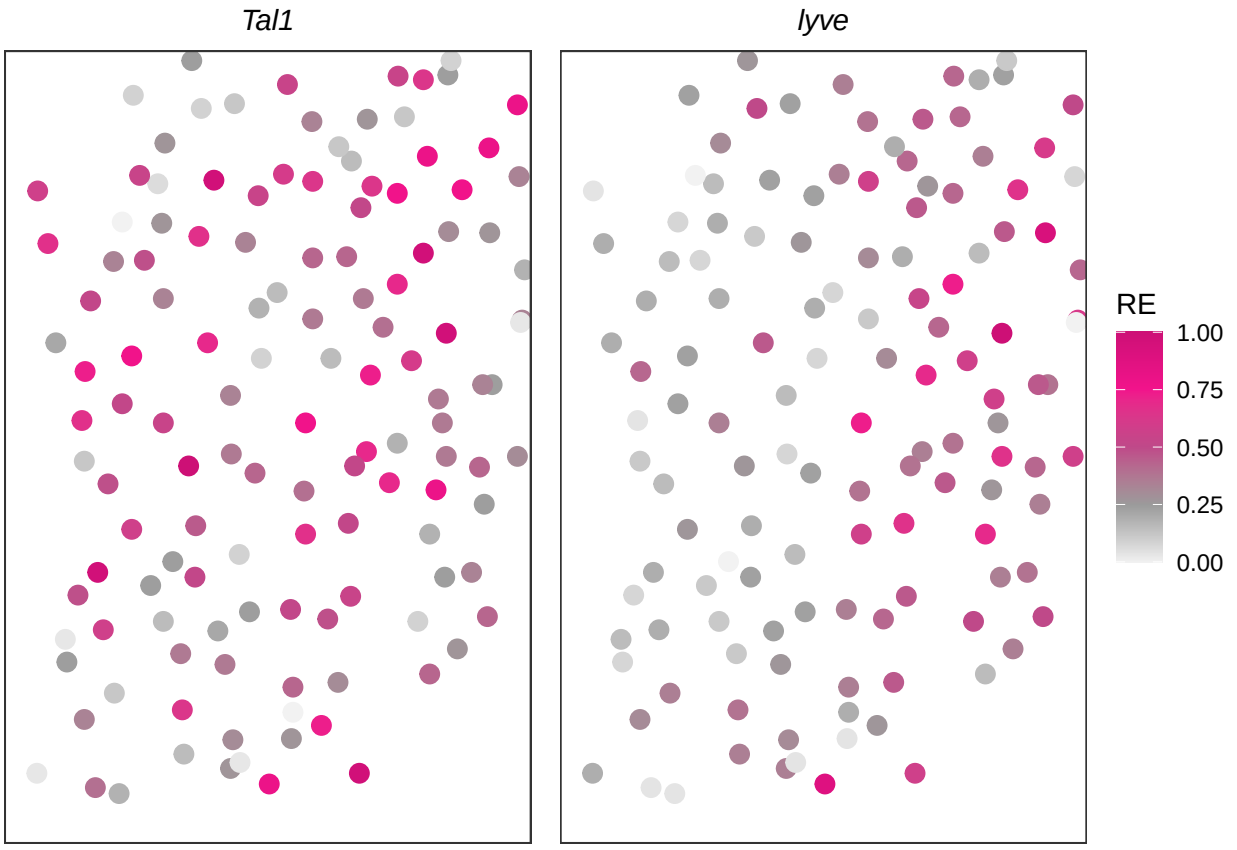
```
Test_res <- SPADE_test(object=data_norm, location=info, para=Est)
```

```
Test_res[c(1, 230),]
```

```
##     geneid         Q    Pvalue Adjust.Pvalue
## 1     Tal1  65.79099 0.4426728     0.5081204
## 230   lyve 510.53422 0.0000000     0.0000000
```

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

```
## [1] 51
```


Tal1
lyve
RE
1.00
0.75
0.50
0.25
0.00

4. 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.

4.1 Data

To illustrate how SPADE can be applied to identify SV genes between groups, a real spatial transcriptomic dataset with axolotl telencephalon (i.e., ARRISTA) was provided in the package. Two different post injury stages (i.e., 2DPI, 5DPI) were included in the dataset. This data contain three genes with 1,188 spots and 938 spots in the 2DPI group and 5DPI group, respectively.

```
data(D2_data)
data(D2_info)
data(D5_data)
data(D5_info)
dim(D2_data)
```

```
## [1]    3 1188
```

```
dim(D2_info)
```

```
## [1] 1188    2
```

```
dim(D5_data)
```

```
## [1]   3 938
```

```
dim(D5_info)
```

```
## [1] 938   2
```

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.

```
D2_norm <- SPADE_norm(readcounts=as.matrix(D2_data), info=D2_info)
D5_norm <- SPADE_norm(readcounts=as.matrix(D5_data), info=D5_info)
```

5.2 Parameter estimation and testing

SPADE identifies SV genes between groups based on a crossed likelihood ratio test in spatial transcriptomic data. SPADE first estimate the optimal hyperparameter for kernel matrix in each group, respectively. Thus, for each gene, the log likelihood in each group can be easily calculated with its optimal kernel. Then we exchange the estimated hyperparameters to compute the log likelihoods for both groups, and compare them to their optimal log likelihoods. The likelihood ratio test statistic is calculated to identify SV genes with P-values computed using F test with degree freedom of one. SPADE_DE() R function was utilized for the SV gene identification with two groups. The output of this function includes
1) geneid: Gene names.
2) theta_Gau1: Optimal length-scale hyperparameter estimated for group 1.

3) theta_Gau2: Optimal length-scale hyperparameter estimated for group 2.
4) logLik11: Log likelihood calculated for group 1 using parameters estimate from itself.
5) logLik21: Log likelihood calculated for group 2 using parameters estimate from itself.
6) logLik10: Log likelihood calculated for group 1 using parameters estimate from group 2.
7) logLik20: Log likelihood calculated for group 2 using parameters estimate from group 1.
8) Diff: Log likelihood ratio statistic calculated with 2*(logLik11+logLik21-logLik10-logLik20).
9) Pvalue: P-values computed using F test with degree freedom of one.
10) Adjust.Pvalue: P-values across all genes are adjusted with the Benjamini and Hochberg method to correct for occurrence of false positives.

```
res <- SPADE_DE(D2_norm, D5_norm, D2_info, D5_info)
```

```
## NO. Gene =  1
## NO. Gene =  2
## NO. Gene =  3
```

```
res
```

```
##                 geneid theta_Gau1 theta_Gau2           Gamma1           Gamma2
## 1 AMEX60DDU001010113   227.4917   311.8125 1.01364636932079 0.701668423918599
## 2 AMEX60DDU001038720   226.1994   283.5883  1.0164473937551 0.593835448496202
## 3 AMEX60DDU001022818   214.7249  1263.7816  2.7329907124081  5.14043020848078
##    logLik11  logLik21  logLik10  logLik20     Diff       Pvalue Adjust.Pvalue
## 1 1998.7766  565.7975 1994.6916  563.3413 13.08245 2.980747e-04  3.850513e-04
## 2 -626.1969 -407.3686 -629.5629 -410.3044 12.60337 3.850513e-04  3.850513e-04
## 3  208.6651  344.6777  196.3985  338.7330 36.42256 1.588536e-09  4.765607e-09
```