SPADE: a spatial pattern and differential expression analysis method, to identify spatially expressed genes within groups and between

Fei Qin, Feifei Xiao, Guoshuai Cai

Last updated: 03/31/2023

1. Introduction to the SPADE method

To identify spatially expressed (SE) genes with spatially resolved 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 SE genes within groups, SPADE also provided a framework to identify SE 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 SE gene identification. To identify SE genes within groups, hypothesis testing was conducted based on a quadratic score statistic with a Davies method to compute the P value. With SE gene detection between groups, SPADE exchanged the optimal hyperparameters estimated in two groups and then utilized a crossed likelihood ratio test to calculate the P value for each gene.

2. Installation

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

3. Data

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

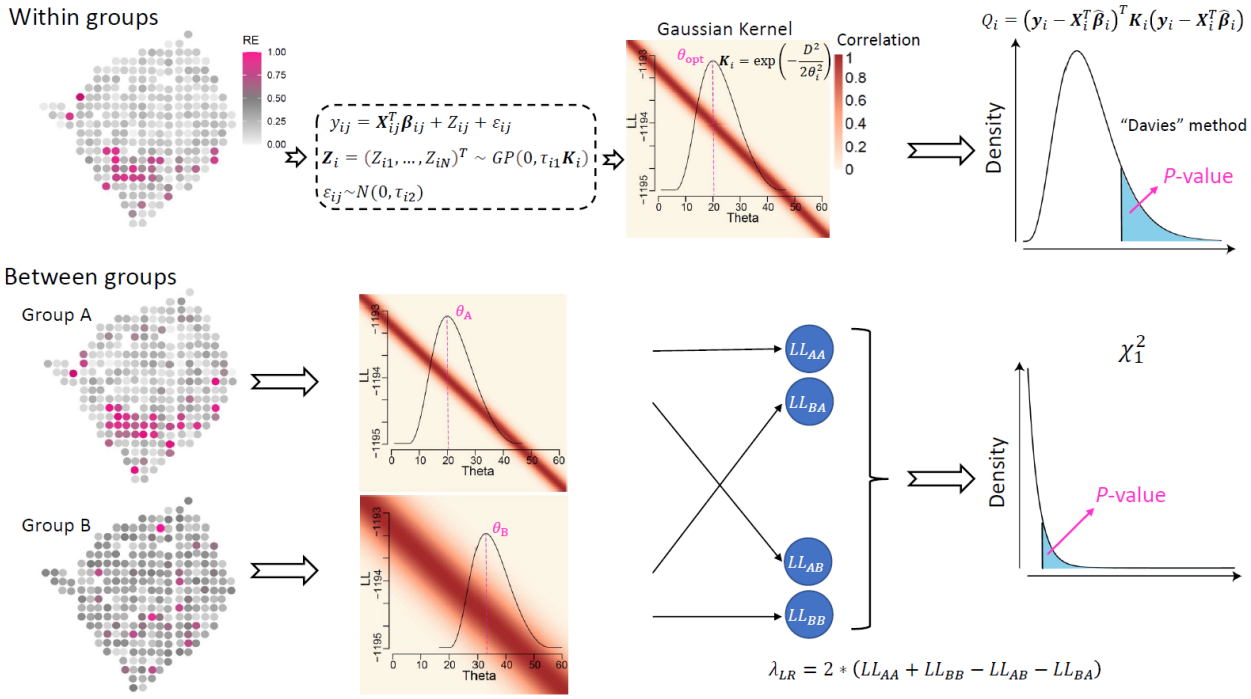

Figure 1: FLCNA framework

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 application.

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

```
# readcounts[1:5,1:5]
# head(info)
```

4. Identifying SE genes within groups

To characterize complex tissues with spatial transcriptomics, the identification of spatially expressed (SE) genes within groups is an essential first step. These identified SE 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.

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

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 SE genes in the SPADE method is to test how well the candidate covariance matrix fits the spatial transcriptomic data. For each gene, we selected the optimal length-scale hyperparameter in the Gaussian kernel to increase the accuracy of SE gene identification. The kernel with optimal hyperparameter will be used in covariance matrix to test whether candidate gene is the significant gene with differential spatial pattern. SPDE_estimate() R function was used for the parameter estimation.

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

4.3 Testing

```
# Test_res <- SPADE_test(object=data_norm, location=info, para=Est)
# head(Test_res)
# sum(Test_res$Adjust.Pvalue < 0.05)
```

5. Identifying SE genes between groups

SPADE was the first method to investigate SE between groups with spatially resolved transcriptomic data. The identification of SE 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