R Package 'spatialimix' Pipeline - A Mixture Model Approach to Spatially Correlated Multi-Omics Data Integration

Ziqiao Wang

2022/08/29

Contents

1 Introduction **1**

2 Load Spatial Data **2**

3 Create Spatial IMIX Object Input **3**

4 Integrative genomics test for two omics data types **3**
4.1 Example 1: Fit Univariate Spatial IMIX model 3
4.2 Example 2: Fit Multivariate Spatial IMIX model 4

5 Summarize and Interpret the Results **5**

1 Introduction

'spatialimix' is an R package for characterizing the spatially correlated samples in high-dimensional omics datasets. This package can find biologically meaningful genes through the integration of two omic data types to investigate the associations between genes and disease subtypes. The input includes the location of each geographical region/sample, the disease subtype of each sample, and the log2ratio values of genes in the samples compared to healthy controls for each data type. spatialimix is flexible and can be applied to different data types including but not limited to gene expression data, DNA methylation data (after summarized at gene-level for integration), and copy number variation data. It considers the spatial correlation between the geographical regions and the inter-data-type correlations in data integration. Functions feature coefficient estimation using spatial linear mixed model and parameter estimation of the mixture model for the summary statistics via EM algorithm while controlling for the across-data-type false discovery rate (FDR) at a user specified level.

We introduce this pipeline using a data example here. The data is based on geographically annotated mucosal samples from a surgically removed bladder specimen from one bladder cancer patient. Each spatial sample was evaluated microscopically and classified by a pathologist into one of three categories: normal urothelium (NU), in situ precursor lesions, or urothelial carcinoma (UC). The in situ precursor lesions were further dichotomized into low-grade intraurothelial neoplasia (LG) and high-grade intraurothelial neoplasia (HG). Furthermore, each spatial sample was measured for two whole genome-wide omics data platforms, gene expression and methylation. We aim to explore the cancer-initiating events that occur in normal-appearing tissue samples that carries on to carcinoma samples in a single tissue section, i.e., discover differentially expressed and methylated genes in the spatially resolved high-dimensional datasets with respect to the sample

subtypes across the tissue, and furthermore the fundamental biological mechanisms. There are three possible outcomes that are biologically meaningful here for both data types: field effect genes (genes in LG, HG, UC samples that are all differentially expressed and methylated compared to healthy controls), HG&UC genes (genes in HG and UC labels that are differentially expressed and methylated in the same direction but LG not significant), and UC genes (genes only in UC label that are differentially expressed and methylated).

This document presents a pipeline of using the spatial IMIX model to integrate two data types of spatially-resolved omics data. The task addressed in this package is to identify significant genes associated with the disease subtypes with stringent FDR control. See help(package="spatialimix") for further details and references provided by citation("spatialimix").

```
library(spatialimix)
#> Loading required package: nlme
```

2 Load Spatial Data

Load the example data. There are 34 geographical regions/samples in this example dataset. These include 27 LG labeled samples, 3 HG labeled samples and 4 UC samples. We include two data types and each data type has 100 genes. We assume they are the gene expression value log2ratio of the geographical regions compared to healthy controls and the methylation value log2ratio of the geographical regions compared to healthy controls.

```
# Load the location information of each geographical region.
# Here both data types share the same location information for the geographical samples.
data("location")
dim(location)
#> [1] 34  2
head(location)
#>   Column Row
#> 1      3   6
#> 2      4  12
#> 3      5  12
#> 4      5  13
#> 5      6  11
#> 6      7   5

# Load the disease grade of each samples.
data("label")
head(label)
#> [1] "LG" "LG" "LG" "LG" "LG" "LG"

# Load the gene expression value log2ratio of the geographical regions
data("ratio1")
dim(ratio1)
#> [1] 100  34
ratio1[1:5,1:5]
#>            Map19.E6   Map19.F12   Map19.G12   Map19.G13    Map19.H11
#> A1BG    -3.75578501 -0.63069930 -0.61552705 -0.71396535 -3.755785006
#> A1CF    -0.77069556 -0.99956315 -4.81950872 -1.07048776 -4.819508720
#> A2M      0.37928166  0.28264109  0.36586558  0.31255579  0.209127868
#> A2ML1    0.01017054 -0.08804962  0.02078044 -0.02768376 -0.002128104
#> A3GALT2  0.47612519  0.83429511  0.07258597  0.22989052  0.797121163
```

```
## Load the methylation value log2ratio of the geographical regions
data("ratio2")
dim(ratio2)
#> [1] 100  34
ratio2[1:5,1:5]
#>            Map19.E6    Map19.F12   Map19.G12    Map19.G13     Map19.H11
#> A1BG    -0.18647250 -0.030845503 -0.06351843 -0.218543098 -0.1365310899
#> A1CF     0.07614127 -0.015561529 -0.04616940 -0.005306637  0.0877642686
#> A2M     -0.03211825  0.003490976 -0.11388014 -0.092408877 -0.0720468470
#> A2ML1   -0.05252038  0.026414519 -0.16194814 -0.104351948  0.0008624743
#> A3GALT2 -0.06168405  0.406982035  0.17954811  0.096838370  0.0385849917
```

3 Create Spatial IMIX Object Input

This step fits spatial linear models for each data type and outputs the fixed effect coefficients and LRT p-values for all the genes. It also produces the inverse normal transformed z scores that is ready for the mixture model data integration step.

```
imix_object_datatype1 <- CreateSpatialIMIXObject(ratio=ratio1,label=label,location=location)
imix_object_datatype2 <- CreateSpatialIMIXObject(ratio=ratio2,label=label,location=location)
```

4 Integrative genomics test for two omics data types

We fit two models respectively on the two data types

4.1 Example 1: Fit Univariate Spatial IMIX model

```
test_uni=fit_uni(input1=imix_object_datatype1$IMIX_Input_Zscores,
                 input2=imix_object_datatype2$IMIX_Input_Zscores)
#> number of iterations= 72
#> number of iterations= 80
#> number of iterations= 128
#> number of iterations= 56
#> number of iterations= 409
#> WARNING! NOT CONVERGENT!
#> number of iterations= 1000
fit_imix_spatial=imix_spatial(input1=imix_object_datatype1$IMIX_Input_Zscores,
                              input2=imix_object_datatype2$IMIX_Input_Zscores,
                              model_type = "univariate",input_initial_model = test_uni)
#> Successfully Done!

# Let's look at the posterior probablity after fitting the model
fit_imix_spatial$`posterior prob`[1:5,1:5]
#>           component1   component2   component3   component4   component5
#> A1BG    3.133117e-28 3.315353e-21 9.997771e-39 0.0009738674 2.182973e-02
```

```
#> A1CF    1.146929e-26 4.752172e-20 2.872575e-39 0.0350141094 2.455947e-03
#> A2M     6.937930e-27 3.606099e-20 7.273601e-39 0.0379863399 7.800980e-03
#> A2ML1   2.693424e-27 3.426700e-20 5.780720e-40 0.0034993413 1.517557e-03
#> A3GALT2 2.169552e-25 2.093453e-22 3.935894e-38 0.9535610516 7.836616e-06
```

There are in total 64 components for 3 disease grades and 2 data types after fitting the mixture model. This is the posterior probability after convergence using the EM algorithm.

4.2 Example 2: Fit Multivariate Spatial IMIX model

```
test_multi=fit_multi(input1=imix_object_datatype1$IMIX_Input_Zscores,
                     input2=imix_object_datatype2$IMIX_Input_Zscores)
#> Warning: Can't find generic `sew` in package knitr to register S3 method.
#> i This message is only shown to developers using devtools.
#> i Do you need to update knitr to the latest version?
#> Assign initial values
#> number of iterations= 65
#> number of iterations= 71
#> number of iterations= 177
#> Start IMIX-ind procedure!
#> Successfully Done!
#> Start IMIX-cor-twostep procedure!
#> Successfully Done!
#> Start Model Selection
#> Warning: No label sorting, need to identify the output groups
#> Start Adaptive FDR Control
#> Finished!
#> Assign initial values
#> number of iterations= 53
#> number of iterations= 401
#> WARNING! NOT CONVERGENT!
#> number of iterations= 1000
#> Start IMIX-ind procedure!
#> Successfully Done!
#> Start IMIX-cor-twostep procedure!
#> Successfully Done!
#> Start Model Selection
#> Warning: No label sorting, need to identify the output groups
#> Start Adaptive FDR Control
#> Finished!
fit_imix_spatial_multi=imix_spatial(input1=imix_object_datatype1$IMIX_Input_Zscores,
                      input2=imix_object_datatype2$IMIX_Input_Zscores,
                      model_type = "multivariate",input_initial_model = test_multi)
#> Successfully Done!

# Let's look at the posterior probablity after fitting the model
fit_imix_spatial_multi$`posterior prob`[1:5,1:5]
#>          component1   component2   component3   component4   component5
#> A1BG              0 1.252834e-90 8.285520e-14 1.086344e-05 3.808616e-03
#> A1CF              0 4.711246e-32 5.472233e-18 3.254500e-07 8.559195e-09
#> A2M               0 4.614433e-02 4.177098e-14 3.360665e-02 4.179341e-04
```

```
#> A2ML1            0 1.813465e-03 2.434984e-15 1.203781e-05 1.009353e-06
#> A3GALT2          0 0.000000e+00 1.200194e-11 9.999908e-01 4.931705e-40
```

This result output is similar to the univariate spatial model.

5 Summarize and Interpret the Results

We use this function to map the Spatial IMIX components to the field effect genes (genes in LG, HG, UC samples are all differentially expressed/methylated compared to healthy controls at a prespecified FDR threshold), genes in HG and UC labels that are differentially expressed in the same direction but LG not significant at the prespeficied FDR threshold, and genes only in UC label that are differentially expressed at the prespeficied FDR threshold. This example controls the FDR at 0.1.

```
# This is the summary of using the univariate spatial IMIX model
res = summary_spatial(fit_imix_spatial,threshold=0.1,
                      imix_object_datatype1 = imix_object_datatype1,
                      imix_object_datatype2 = imix_object_datatype2)

# This is the summary of using the multivariate spatial IMIX model
res_multi = summary_spatial(fit_imix_spatial_multi,threshold=0.1,
                            imix_object_datatype1 = imix_object_datatype1,
                            imix_object_datatype2 = imix_object_datatype2)

# Look at the results at FDR controlled at 0.1
head(res$results)
#>         int_imix_group              label
#> A1CF                 2 Field effect genes
#> AADACL3              2 Field effect genes
#> AAK1                 2 Field effect genes
#> ABCA9                2 Field effect genes
#> ABCB11               2 Field effect genes
#> ABCB4                2 Field effect genes
table(res$results$label)
#>
#> Field effect genes              Other
#>                 17                 83

head(res$results)
#>         int_imix_group              label
#> A1CF                 2 Field effect genes
#> AADACL3              2 Field effect genes
#> AAK1                 2 Field effect genes
#> ABCA9                2 Field effect genes
#> ABCB11               2 Field effect genes
#> ABCB4                2 Field effect genes
table(res_multi$results$label)
#>
#> Field effect genes              Other
#>                  2                 98
```

The result output for the univariate spatial IMIX model shows that there are 17 field-effect genes in both gene expression and methylation data out of 100 at FDR $\alpha = 0.1$ and for the multivariate spatial IMIX model

there are 2 field effect genes in both gene expression and methylation data detected at FDR $\alpha = 0.1$.

```
sessionInfo()
#> R version 3.6.3 (2020-02-29)
#> Platform: x86_64-apple-darwin15.6.0 (64-bit)
#> Running under: macOS Sierra 10.12.6
#>
#> Matrix products: default
#> BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] mvtnorm_1.0-12    spatialimix_0.1.0 nlme_3.1-144
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.5       pillar_1.7.0     compiler_3.6.3   mixtools_1.1.0
#>  [5] tools_3.6.3      mclust_5.4.5     digest_0.6.23    evaluate_0.14
#>  [9] tibble_3.1.6     lifecycle_1.0.1  gtable_0.3.0     lattice_0.20-38
#> [13] pkgconfig_2.0.3  rlang_1.0.1      Matrix_1.2-18    IMIX_1.1.4
#> [17] cli_3.2.0        rstudioapi_0.11  yaml_2.2.0       xfun_0.11
#> [21] dplyr_1.0.2      stringr_1.4.0    knitr_1.26       generics_0.1.0
#> [25] vctrs_0.3.8      gtools_3.8.2     tidyselect_1.1.0 segmented_1.1-0
#> [29] grid_3.6.3       glue_1.6.2       R6_2.4.1         fansi_0.4.1
#> [33] survival_3.1-8   rmarkdown_2.0    gdata_2.18.0     purrr_0.3.3
#> [37] ggplot2_3.3.5    magrittr_1.5     scales_1.0.0     htmltools_0.3.6
#> [41] ellipsis_0.3.2   MASS_7.3-51.5    splines_3.6.3    colorspace_1.4-1
#> [45] utf8_1.1.4       stringi_1.4.5    munsell_0.5.0    crayon_1.3.4
```