Usage Tips

Junyan Song

2017-02-10

Introduction

This is an example of using SCclust package in R. For whole genome segmented copy number profiles, SCclust implements determination of breakpoints, derivation of new feature set, hierarchical clustering on the binary matrix and identification of clone structure. This vignette aims to demonstrate the usage of SCclust through an example data.

Usage

(1) Read example data

The segmented copy number profiles `longint` and other two genome annotation files (`fullannot` and `cyto`) are input files. Note: The data files are not included in the package. You need to load the files from your local directory.

```
study<-"YL2672P9"
fullannot <- read.table("varbin.gc.content.5k.bowtie.k50.hg19.txt", header = T, as.is = T)
cyto <- read.table("cytoBandHG19.txt", header = F, as.is = T)
segfile <- paste("uber.",study,"_CORE_F.5k.seg.quantal.primary.txt", sep = "")
longint<-read.table(segfile, header = T,as.is = T)
```

(2) Preprocess input annotation files

Reformat the annotation files and identify the centromeres areas for filtering.

```
annot <- preprocess_annot(fullannot)
dropareas <- drop_areas(cyto)
```

(3) Determination of breakpoints

Determine the position of breakpoints and the sign of CN discontinuity. Output two tables `breakpoint_table` and `ploidies_table`.

```
res1 <- preprocess_segfile(longint, annot, eviltwins = NULL, ploidies = TRUE)
breakpoint_table <- res1$breakpoint_table
ploidies_table <- res1$ploidies_table
```

(4) Derivation of new feature set

The breakpoints are extended to an interval spanning 2*`smear` bins. This results in the `smear_table`.

The function `findpins` implement the procedure to find a minimum set of "piercing points" for the extended intervals as the new feature set and derive the incidence table. The piercing points are a smallest set of points such that each interval contains at least one of them. `pins` and `pinmat` provide the bin location of new feature set and the incidence table. We extract `cell_names` for stages later.

```
smear_table <- findsmears(breakpoint_table, dropareas, smear = 2, keepboundaries = TRUE, mask_XY = TRUE)

res2 <- findpins(breakpoint_table, smear_table)
pins <- res2$pins
pinmat <- res2$pinmat
cell_names <- res2$cell_names
```

(5) Perform Fisher's tests on incidence table

We perform Fisher's tests for pairwise dissimilarity on both observed incidence table and permutated incidence tables. This procedure `simFisher_parallel` generates two vectors of p-values `true_fisherPV` and `sim_fisherPV` for observed and permutated data respectively.

```
res3 <- simFisher_parallel(pins, pinmat, sim_round = 500)
true_fisherPV <- res3$true_fisherPV
sim_fisherPV <- res3$sim_fisherPV
```

(6) Significance assessment of pairwise dissimilarity

Compute the FDRs `mat_fdr` for the observed Fisher's test p-values.

```
mat_fdr <- fdr_fisherPV(true_fisherPV, sim_fisherPV, cell_names, lm_max = 0.001, graphic = TRUE)
```

(7) Identify the clone in the hierarchical clustering tree

This procedure identify the hard and soft clones and display the hierarchical tree. The clones node information can be output from `hc$softclones`.

```
hc <- hclust_tree(pinmat, mat_fdr, true_fisherPV, cell_names, hc_method = "average")
hc_clone <- find_clone(hc$hc, fdr_thresh = -2, share_min = 0.90, n_share = 3, bymax = TRUE,
                       climb_from_size = 2, climb_to_share = 3, graphic = TRUE)
```

More Information

Details for arguments and functions can be found by typing e.g. `help(package="SCclust")`, `?fdr_fisherPV` , `?findsmears`.