Package ‘subtype’

February 20, 2015

Type Package
Title Cluster analysis to find molecular subtypes and their assessment
Version 1.0
Date 2013-01-14
Author Andrey Alexeyenko, Woojoo Lee and Yudi Pawitan
Maintainer Woojoo Lee <lwj221@gmail.com>
Description subtype performs a biclustering procedure on a input dataset and assess whether resulting clusters are promising subtypes. Note that the R-package rsmooth should be installed before implementing subtype. rsmooth can be downloaded from http://www.meb.ki.se/~yudpaw.
Depends penalized,ROCR
License GPL-2
Repository CRAN
Date/Publication 2013-01-14 13:53:58
NeedsCompilation no

R topics documented:

  subtype ................................................................. 1
  summary ............................................................... 3

Index

subtype Cluster analysis to find molecular subtypes and their assessment

Description

  subtype performs a biclustering procedure on a input dataset and assess whether resulting clusters are promising subtypes.
Usage

```r
subtype(GEset, outcomeLabels, treatment=NULL, Npermutes=10, Nchunks = 25, minClusterSizeB = 20, NclustersASet = 100, fdrpermutation = TRUE, nfdrperm = 50, seed = NULL, testMode = "quick", survivaltimes = NULL, method = "penalized", top_best_probes = 100, Niter = 20, showMovie = 0, redefineSubtypeMembers = 0, holdOut = 10)
```

Arguments

- **GEset**: p-by-n data matrix, where p is the number of variables (e.g. genes) and n is the number of subjects. Row and column names are necessary.
- **outcomeLabels**: n-by-1 vector. Binary prognosis labels assigned to the subjects. The order of subjects should be equalized to that of GEset.
- **treatment**: NULL.
- **Npermutes**: Number of permutations for the variables. For each permutation, the variables belong to different chunks.
- **Nchunks**: Number of chunks of the variables. When the number of variables is too large for clustering analysis, we split the variables into several (=Nchunks) chunks.
- **minClusterSizeB**: The minimum number of subjects per each selected subtype. The default is 20.
- **NclustersASet**: Cut a tree from hierarchical clustering into several groups. The default is 100.
- **FDRpermutation**: Determine whether FDR computation is based on permutation procedure. The default is TRUE.
- **nFDRperm**: Number of permutation to compute FDR. The default is 50.
- **seed**: seed number for reproducibility.
- **testMode**: the mode is fixed at "quick".
- **survivaltimes**: NULL.
- **method**: penalized is used.
- **top_best_probes**: top-ranked probes are used in t-test, and this is input for penalized. The default is 100.
- **Niter**: The number of iterations of (TrainingSet, TestSet)->training->test->recordResults. The default is 20.
- **showMovie**: display RUC/Surv curves and heatmaps. The default is 0.
- **redefineSubtypeMembers**: detect subtype members after every hold-out. The default is 0.
- **holdOut**: out of the subtype, i.e. Nsubtype - holdOut = Ntraining_set. The default is 10.

Details

This implements a biclustering algorithm to find hidden subtypes in a dataset. summary provides a measure based on FDR and its p-value for assessing the subtypes. Note that the R-package rsmooth should be installed before implementing subtype. rsmooth can be downloaded from http://www.meb.ki.se/~yudpaw. For large dataset, the computation can be heavy, so it is desirable for users to consider parallel processing in R.

Value
resultsAll: a matrix including subtypeID and summary statistics for each subtypeID. For a specific subtypeID,
GenesDefiningSubtypes: Variables in each subtypeID. This can be identified with "subtypeID".
SubtypePatients: Subjects in each subtypeID. This can be identified with subtypeID.

Author(s)

Andrey Alexeyenko, Woojoo Lee (maintainer:lwj221@gmail.com) and Yudi Pawitan

References


Examples

```r
set.seed(1234)
p<-100   #num.variables
n1<-5    #number of sample in population 1
n2<-5    #num.samples from population 2
group<-c(rep(1,length.out=n1),rep(2,length.out=n2))
data<-matrix(rnorm((n1+n2)*p),(n1+n2),p)

# making row names
dimnames(data)[[1]]<-as.character(paste("P",runif(nrow(data),0,1),sep=""))
# making column names
dimnames(data)[[2]]<-as.character(paste("G",runif(ncol(data),0,1),sep=""))

### The following procedure takes ~ 1 minute.
A=subtype(
    GESet = t(data),
    outcomeLabels = group,
    Npermutes = 2,
    Nchunks = 5,
    NclustersASet = 3,
    seed=1234
  )
summary(A,f.out=0) ### f.out can be used for filtering out uninteresting subtypes. e.g. if f.out=2, we ignore sub
```

summary

<table>
<thead>
<tr>
<th>summary</th>
<th>Summarizing the output from subtype</th>
</tr>
</thead>
</table>

Description

summary summarizes the output from subtype.
summary

Usage

summary(object,...)

Arguments

object the output from subtype
... criterion for filtering out uninteresting cases

Details

summary provides a measure based on FDR and its p-value for assessing the subtypes.

Value

NFDR01 : see the reference

Author(s)

Andrey Alexeyenko, Woojoo Lee (maintainer:lwj221@gmail.com) and Yudi Pawitan

References


Examples

### Should be DIRECTLY executable !! ----
### ==> Define data, use random,
### or do help(data=index) for the standard data sets.

set.seed(1234)
p<-100   #num.variables
n1<5    #number of sample in population 1
n2<5    #num.samples from population 2
group<-c(rep(1,length.out=n1),rep(2,length.out=n2))
data<-matrix(rnorm((n1+n2)*p),(n1+n2),p)

# making row names
dimnames(data)[[1]]<-as.character(paste("P",runif(nrow(data),0,1),sep=""))
# making column names
dimnames(data)[[2]]<-as.character(paste("G",runif(ncol(data),0,1),sep=""))

### The following procedure takes ~ 1 minute.
A=subtype(
    GEset = t(data),
    outcomeLabels = group,
    Npermutes = 2,
    Nchunks = 5,
    NclustersASet = 3,
    seed=1234
summary(
)

summary(A, f.out=0) ### f.out can be used for filtering out uninteresting subtypes. e.g. if f.out=2, we ignore sub
Index

subtype, 1
summary, 3