Package 'PedCNV'

February 19, 2015

Type Package

Title An implementation for association analysis with CNV data.

Version 0.1

Date 2013-08-03

Author Meiling Liu, Sungho Won and Weicheng Zhu

Maintainer Meiling Liu <meiling.sta@gmail.com>

Description An implementation for association analysis with CNV data in R. It provides two methods for association study: first, the observed probe intensity measurement can be directly used to detect the association of CNV with phenotypes of interest. Second, the most probable copy number is estimated with the proposed likelihood and the association of the most probable copy number with phenotype is tested. This method can be applied to both the independent and correlated population.

License MIT + file LICENSE

URL https://github.com/rksyouyou/PedCNV

LazyData true

Depends Rcpp (>= 0.10.4), RcppArmadillo (>= 0.3.900.0), ggplot2

LinkingTo Rcpp, RcppArmadillo

Collate 'AI_EM_mixed.R' 'ClusProc.R' 'cnvlmm_plot.R' 'silWidth.R' 'docs.R' 'score_test.R' 'AssoTestProcCS_mix.R'

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-01-02 19:18:18

R topics documented:

`PedCNV-package` *CNV association implementation*

Description

A package to perform robust quantitative traits association testing of copy number variants. It provides two methods for association study: first, the observed probe intensity measurement can be directly used to detect the association of CNV with phenotype of interest. Second, the most probable copy number is estimated with the proposed likelihood and the association of the most probable copy number with phenotype is tested. Also, it can be used to determine the optimal clustering number and clustering assignment for each individuals. This method can be applied to both the independent and correlated population.

Details

Package:	PedCNV
Type:	Package
Version:	0.1
Date:	2013-09-03
License:	MIT
Main functions:	AssoTestProc
	ClusProc
	STE
	STIM
	print.asso
	print.clus
	plot.clus

Author(s)

Meiling Liu, Sungho Won and Weicheng Zhu

References

On the association analysis of CNV data: fast and efficient method with family-based samples

---

`AssoTestProc` *CNV association test procedure*

Description

This function tests the association of CNV with continuous trait of interest. Two statistics are provided for different strategies with the intensity measurement.

Usage

```
AssoTestProc(signal, fam, envirX, phi, N,
  varSelection = c("PC1", "RAW", "PC.9", "MEAN"),
  H0 = TRUE, threshold = 1e-05, itermax = 8,
  thresEM = 0.005, thresAI = 1e-05)
```

Arguments

`signal`	The matrix of intensity measurements. The row names must be consistent with the Individual ID in fam file.
`fam`	The FAM file which follows the format defined in PLINK.
`envirX`	The matrix of environmental variables. The intercept is automatically included and it does not need to be in this matrix.
`phi`	The correlation matrix between individuals. It can be built with the kinship coefficient or the estimated correlation matrix with SNP data. Free software that builds this matrix is available, and one of them can be downloaded at http://biostat.ac.kr/fqls/ The default is an identity matrix and it is for independent samples.
`N`	Number of clusters one wants to fit to the data. N needs to be larger than 1 and if it is 1, error will be returned. It can be estimated with the function `ClusProc`.
`varSelection`	Factor. For specifying how to handle the intensity values. It must take value on 'RAW', 'PC.9', 'PC1'and 'MEAN'. If the value is 'RAW', then the raw intensity value will be used. If it is 'PC.9', then the first several PCA scores which account for 90% of all the variance will be used. If the value is 'PC1', then the first PCA scores will be used. If the value is 'MEAN', the mean of all the probes will be used. The default method is 'PC1'.
`H0`	Logicals. If it is TRUE (the default), all parameters are estimated under the assumption that there is no genetic association between CNV and phenotypes. If it is FALSE, parameters are estimated under the null or alternative hypothesis.
`threshold`	Optional number of convergence threshold. The iteration stops if the absolute difference of log likelihood between successive iterations is less than it. The default threshold 1e-05 will be used if it's missing.

itermax	Optional. The iteration stops if the times of iteration is large than this value. The default number 8 will be used if it's missing.
thresEM	Optional number of convergence threshold in the EM (expectation-maximization method) procedure. The default threshold 0.005 will be used if it's missing.
thresAI	Optional number of convergence threshold in the AI (average information method) procedure. The default threshold 1e-05 will be used if it's missing.

Value

It returns object of class 'asso'. The result is obtained under the null hypothesis if H0 is TRUE, otherwise the result is obtained under null or alternative hypothesis.

para	The parameter estimations for the best fit.
clusRes	The clustering assignment for each individual.

Author(s)

Meiling Liu, Sungho Won and Weicheng Zhu

Examples

```
# Fit the data under the assumption that there are 3 clusters
fit.pc <- AssoTestProc(signal=signal,fam=fam,envirX=envirX,phi=phi,N=3,varSelection='PC.9')
```

ClusProc *CNV clustering Procedure*

Description

This function chooses the optimal number of clusters and provides the assignments of each individuals under the optimum clustering number.

Usage

```
ClusProc(signal, N = 2:6,
  varSelection = c("PC1", "RAW", "PC.9", "MEAN"),
  threshold = 1e-05, itermax = 8, adjust = TRUE,
  thresMAF = 0.01, scale = FALSE, thresSil = 0.01)
```

Arguments

signal	The matrix of intensity measurements. The row names must be consistent with the Individual ID in fam file.
N	Number of clusters one wants to fit to the data. N needs to be larger than 1 and if it is 1, error will be returned. The default value 2,3,...,6 will be used if it is missing.

varSelection	Factor. For specifying how to handle the intensity values. It must take value on 'RAW', 'PC.9', 'PC1'and 'MEAN'. If the value is 'RAW', then the raw intensity value will be used. If it is 'PC.9', then the first several PCA scores which account for 90% of all the variance will be used. If the value is 'PC1', then the first PCA scores will be used. If the value is 'MEAN', the mean of all the probes will be used. The default method is 'PC1'.
threshold	Optional number of convergence threshold. The iteration stops if the absolute difference of log likelihood between successive iterations is less than it. The default threshold 1e-05 will be used if it's missing.
itermax	Optional. The iteration stops if the time of iteration is large than this value. The default number 8 will be used if it's missing.
adjust	Logicals, If TRUE (default), the result will be adjusted by the silhouette score. See details.
thresMAF	The minor allele frequency threshold.
thresSil	The abandon threshold. The individual whose silhouette score is smaller than this value will be abandoned.
scale	Logicals. If TRUE, the signal will be scale by using sample mean and sample variance by columns before further data-processing.

Details

- adjustIf adjust is TRUE, the result will be adjusted by the silhouette score in the following criterion. For each individual, the silhouette scores are calculated for each group. The individual will assigned forcefully to the group which maximize the silhouette scores.

Value

It returns object of class 'clust'. 'clust' is a list containing following components:

clusNum	The optimal number of clusters among give parameter N.
silWidth	Silhouette related results.

Author(s)

Meiling Liu

Examples

```
# Fit the data under the given clustering numbers
clus.fit <- ClusProc(signal=signal,N=2:6,varSelection='PC.9')
```

envirX *CNV simulated environmental variables*

Description

The simulated environmental file which contains the possible environmental variables. The order of the row in this file must consistent with the second column in FAM file.

Author(s)

Meiling Liu

fam *CNV simulated data*

Description

The simulated FAM file. The first six columns of FAM file are mandatory: Family ID, Individual ID, Paternal ID, Maternal ID, Sex (1=male; 2=female; other=unknown) and Phenotype.

Author(s)

Meiling Liu

phi *Empirical correlation matrix*

Description

Empirical/kinship correlation matrix between individuals. This correlation matrix can be calculated based on the familial relationship between individuals or large-scale SNP data by omic data analysis toolkit FQLS. The free software FQLS can be downloaded from http://biostat.cau.ac.kr/fqls/. If correlation matrix is estimated with the large-scale SNP data, the proposed method becomes robust under the presence of population substructure.

Author(s)

Meiling Liu

References

FQLS http://biostat.cau.ac.kr/fqls/

Examples

```
data(phi)
```

plot.clust *Plots clustering result*

Description

Makes formatted plots from the clustering result returned from ClusProc.

Usage

```
 ## S3 method for class 'clust'
plot(x,
   type = c("histo", "scat", "sil"), adjust = TRUE, ...)
```

Arguments

x	The clustering results obtained from ClusProc.
type	Factor. For specifying the plot type. It must be one of 'histo', 'scat' and 'sil'. If it is 'histo', the histogram is obtained with the first PC score of the intensity measurement. For 'scat', the first PC score of the intensity measurement is plotted against the mean of the intensity measurement. For 'sil', the silhouette score is plotted. See details.
adjust	Logicals. If TRUE (default), the silhouette-adjusted clustering result will be used. If FALSE, the initial clustering result will be used. See details in ClusProc.
...	Usual arguments passed to the qplot function.

Details

- typeWe provide three types of plots: 'hist', 'scat' and 'sil'. The first two plots are used to visually check the performance of clustering. Different clusters are represented by using different colors. The 'sil' plot is the the overview of the silhouette value for all the individuals, the silhouettes of the different clusters are printed below each other. The higher silhouettes value means the better performance.

Author(s)

Meiling Liu

Examples

```
# Fit the data under the given clustering numbers
clus.fit <- ClusProc(signal=signal,N=2:6,varSelection='PC.9')
plot(clus.fit,type='histo')
```

print.asso *Prints association study results*

Description

Prints formatted results from the association study returned by AssoTestProc.

Usage

```
 ## S3 method for class 'asso'
print(x, ...)
```

Arguments

x	The association study results obtained from the AssoTestProc.
...	Usual arguments passed to the print function.

Author(s)

Meiling Liu

Examples

```
# Fit the data under the assumption that there are 3 clusters
asso.fit <- AssoTestProc(signal=signal,fam=fam,envirX=envirX,phi=phi,N=3,varSelection='PC.9')
print(asso.fit)
```

print.clust *Prints clustering results*

Description

Prints formatted results returned by ClusProc.

Usage

```
 ## S3 method for class 'clust'
print(x, ...)
```

Arguments

x	The clustering results obtained from the ClusProc.
...	Usual arguments passed to the print function.

Author(s)

Meiling Liu and Sungho Won

Examples

```
# Fit the data under the given clustering numbers
clus.fit <- ClusProc(signal=signal,N=2:6,varSelection='PC.9')
print(clus.fit)
```

---

signal *CNV simulated intensity measurements*

Description

The simulated intensity measurements. The order of the row in this file must consistent with the second column in FAM file.

Author(s)

Meiling Liu

STE *Score test with the most probable CNV*

Description

Calculates the score test statistics with the most probable CNV.

Usage

```
STE(envirX, clusRes, fam, alpha, phi, sig2g, sig2)
```

Arguments

envirX	The matrix of environmental variables. The intercept should be included if it's needed.
fam	The FAM file which follows the format defined in PLINK.
clusRes	The clustering group which is signed to each individual.
alpha	The estimated parameters for environmental variables under null hypothesis. This value can be calculated by using function AssoTestProc.
phi	The matrix of correlation between individuals.
sig2g	The estimated standard error for polygenic effect under null hypothesis. This value can be calculated by using function AssoTestProc.
sig2	The estimated standard error for environmental effect under null hypothesis. This value can be calculated by using function AssoTestProc.

Value

It returns the statistic value and pvalue of the score test.

STEs	The statistic value of score test with the most probable CNV.
STEp	The pvalue of score test with the most probable CNV.

Author(s)

Meiling Liu, Sungho Won

Examples

```
# Fit the data under the assumption that there are 3 clusters
asso.fit <- AssoTestProc(signal=signal,fam=fam,envirX=envirX,phi=phi,N=3,varSelection='PC.9')
cnv_e <- asso.fit$clusRes
alpha <- asso.fit$para$alpha
sig2g <- asso.fit$para$sig2g
sig2 <- asso.fit$para$sig2
STE(envirX=envirX,clusRes=cnv_e,fam=fam,alpha=alpha,phi=phi,sig2g=sig2g,sig2=sig2)
```

STIM *Score test with the intensity value*

Description

Calculates the score test statistics with the intensity value.

Usage

```
STIM(envirX, signal, fam, alpha, phi, sig2g, sig2)
```

Arguments

envirX	The matrix of environmental variables. The intercept should be included if it's needed.
fam	The FAM file which follows the format defined in PLINK.
signal	The matrix of intensity measurements. The row names must be consistent with the Individual ID in fam file.
alpha	The estimated parameters for environmental variables under null hypothesis. This value can be calculated by using function AssoTestProc.
phi	The matrix of correlation between individuals.
sig2g	The estimated standard error for polygenic effect under null hypothesis. This value can be calculated by using function AssoTestProc.
sig2	The estimated standard error for environmental effect under null hypothesis. This value can be calculated by using function AssoTestProc.

Value

It returns the statistic value and pvalue of the score test.

STIMs	The statistic value of score test with the intensity value under null hypothesis.
STIMp	The pvalue of score test with the intensity value under null hypothesis.
df	The degree of freedom of score test with the intenstiy value under null hypothesis.

Author(s)

Meiling Liu, Sungho Won

Examples

```
# Fit the data under the assumption that there are 3 clusters
asso.fit <- AssoTestProc(signal=signal,fam=fam,envirX=envirX,phi=phi,N=3,varSelection='PC.9')
alpha <- asso.fit$para$alpha
sig2g <- asso.fit$para$sig2g
sig2 <- asso.fit$para$sig2
STIM(envirX=envirX,signal=signal,fam=fam,alpha=alpha,phi=phi,sig2g=sig2g,sig2=sig2)
```

Index