Package 'pcadapt'

February 27, 2019

Type Package

Title Fast Principal Component Analysis for Outlier Detection

Version 4.1.0

Date 2019-02-26

Description Methods to detect genetic markers involved in biological adaptation. 'pcadapt' provides statistical tools for outlier detection based on Principal Component Analysis. Implements the method described in (Luu, 2016) <DOI:10.1111/1755-0998.12592>.

License GPL (>= 2)

Imports data.table, ggplot2, magrittr, mmapcharr (>= 0.3), plotly, Rcpp (>= 0.12.8), robust, RSpectra, vcfR

LinkingTo Rcpp, rmio, mmapcharr

LazyData TRUE

Suggests knitr, rmarkdown, testthat, shiny, covr

RoxygenNote 6.1.0

VignetteBuilder knitr

Encoding UTF-8

NeedsCompilation yes

Author Keurcien Luu [aut],
Michael Blum [aut, cre],
Florian Privé [aut],
Eric Bazin [ctb],
Nicolas Duforet-Frebourg [ctb]

Maintainer Michael Blum <michael.blum@univ-grenoble-alpes.fr>

Repository CRAN

Date/Publication 2019-02-27 13:20:03 UTC

R topics documented:

`bed2matrix` *Convert a bed to a matrix*

Description

Convert a bed to a matrix

Usage

```
bed2matrix(bedfile, n = NULL, p = NULL)
```

Arguments

`bedfile`	Path to a bed file.
`n`	Number of samples. Default reads it from coresponding fam file.
`p`	Number of SNPs. Default reads it from coresponding bim file.

Value

An integer matrix.

Examples

```
bedfile <- system.file("extdata", "geno3pops.bed", package = "pcadapt")
mat <- bed2matrix(bedfile)
dim(mat)
table(mat)
```

get.pc — *Get the principal component the most associated with a genetic marker*

Description

`get.pc` returns a data frame such that each row contains the index of the genetic marker and the principal component the most correlated with it.

Usage

```
get.pc(x, list)
```

Arguments

`x`	an object of class 'pcadapt'.
`list`	a list of integers corresponding to the indices of the markers of interest.

Examples

```
## see also ?pcadapt for examples
```

get_statistics — *pcadapt statistics*

Description

`get_statistics` returns chi-squared distributed statistics.

Usage

```
get_statistics(zscores, method, pass)
```

Arguments

`zscores`	a numeric matrix containing the z-scores.
`method`	a character string specifying the method to be used to compute the p-values. Two statistics are currently available, `"mahalanobis"`, and `"componentwise"`.
`pass`	a boolean vector.

Value

The returned value is a list containing the test statistics and the associated p-values.

pcadapt *Principal Component Analysis for outlier detection*

Description

pcadapt performs principal component analysis and computes p-values to test for outliers. The test for outliers is based on the correlations between genetic variation and the first K principal components. pcadapt also handles Pool-seq data for which the statistical analysis is performed on the genetic markers frequencies. Returns an object of class pcadapt.

Usage

```
pcadapt(input, K = 2, method = "mahalanobis", min.maf = 0.05,
  ploidy = 2, LD.clumping = NULL, pca.only = FALSE)

## S3 method for class 'pcadapt_matrix'
pcadapt(input, K = 2,
  method = c("mahalanobis", "componentwise"), min.maf = 0.05,
  ploidy = 2, LD.clumping = NULL, pca.only = FALSE)

## S3 method for class 'pcadapt_bed'
pcadapt(input, K = 2, method = c("mahalanobis",
  "componentwise"), min.maf = 0.05, ploidy = 2, LD.clumping = NULL,
  pca.only = FALSE)

## S3 method for class 'pcadapt_pool'
pcadapt(input, K = (nrow(input) - 1),
  method = "mahalanobis", min.maf = 0.05, ploidy = NULL,
  LD.clumping = NULL, pca.only = FALSE)
```

Arguments

input	a genotype matrix or a character string specifying the name of the file to be processed with pcadapt.
K	an integer specifying the number of principal components to retain.
method	a character string specifying the method to be used to compute the p-values. Two statistics are currently available, "mahalanobis", and "componentwise".
min.maf	a value between 0 and 0.45 specifying the threshold of minor allele frequencies above which p-values are computed.
ploidy	Number of trials, parameter of the binomial distribution. Default is 2, which corresponds to diploidy, such as for the human genome.
LD.clumping	Default is NULL and doesn't use any SNP thinning. If you want to use SNP thinning, provide a named list with parameters size and thr which corresponds respectively to the window radius and the squared correlation threshold. A good default value would be list(size = 200, thr = 0.1).

`pca.only` a logical value indicating whether PCA results should be returned (before computing any statistic).

Details

First, a principal component analysis is performed on the scaled and centered genotype data. To account for missing data, the correlation matrix between individuals is computed using only the markers available for each pair of individuals. Depending on the specified `method`, different test statistics can be used.

`mahalanobis` (default): the robust Mahalanobis distance is computed for each genetic marker using a robust estimate of both mean and covariance matrix between the `K` vectors of z-scores.

`communality`: the communality statistic measures the proportion of variance explained by the first `K` PCs. Deprecated in version 4.0.0.

`componentwise`: returns a matrix of z-scores.

To compute p-values, test statistics (`stat`) are divided by a genomic inflation factor (`gif`) when `method="mahalanobis"`. When using `method="mahalanobis"`, the scaled statistics (`chi2_stat`) should follow a chi-squared distribution with `K` degrees of freedom. When using `method="componentwise"`, the z-scores should follow a chi-squared distribution with `1` degree of freedom. For Pool-seq data, `pcadapt` provides p-values based on the Mahalanobis distance for each SNP.

Value

The returned value `x` is an object of class `pcadapt`.

`plot.pcadapt` *pcadapt visualization tool*

Description

`plot.pcadapt` is a method designed for objects of class `pcadapt`. It provides a plotting utile for quick visualization of `pcadapt` objects. Different options are currently available : `"screeplot"`, `"scores"`, `"stat.distribution"`, `"manhattan"` and `"qqplot"`. `"screeplot"` shows the decay of the genotype matrix singular values and provides a figure to help with the choice of `K`. `"scores"` plots the projection of the individuals onto the first two principal components. `"stat.distribution"` displays the histogram of the selected test statistics, as well as the estimated distribution for the neutral SNPs. `"manhattan"` draws the Manhattan plot of the p-values associated with the statistic of interest. `"qqplot"` draws a Q-Q plot of the p-values associated with the statistic of interest.

Usage

```
## S3 method for class 'pcadapt'
plot(x, ..., option = "manhattan", i = 1, j = 2,
  pop, col, chr.info = NULL, snp.info = NULL, plt.pkg = "ggplot",
  K = NULL)
```

Arguments

x	an object of class "pcadapt" generated with pcadapt.
...	...
option	a character string specifying the figures to be displayed. If NULL (the default), all three plots are printed.
i	an integer indicating onto which principal component the individuals are projected when the "scores" option is chosen. Default value is set to 1.
j	an integer indicating onto which principal component the individuals are projected when the "scores" option is chosen. Default value is set to 2.
pop	a list of integers or strings specifying which subpopulation the individuals belong to.
col	a list of colors to be used in the score plot.
chr.info	a list containing the chromosome information for each marker.
snp.info	a list containing the names of all genetic markers present in the input.
plt.pkg	a character string specifying the package to be used to display the graphical outputs. Use "plotly" for interactive plots, or "ggplot" for static plots.
K	an integer specifying the principal component of interest. K has to be specified only when using the "componentwise" method.

Examples

```
## see ?pcadapt for examples
```

---

print_convert *Summary*

Description

print_convert prints out a summary of the file conversion.

Usage

```
print_convert(input, output, M, N, pool)
```

Arguments

input	a genotype matrix or a character string specifying the name of the file to be converted.
output	a character string specifying the name of the output file.
M	an integer specifying the number of genetic markers present in the data.
N	an integer specifying the number of individuals present in the data.
pool	an integer specifying the type of data. ‘0‘ for genotype data, ‘1‘ for pooled data.

Examples

```
## see also ?pcadapt for examples
```

`read.pcadapt` *File Converter*

Description

`read.pcadapt` converts genotype matrices or files to an appropriate format readable by `pcadapt`. For a file as input, you can return either a matrix or convert it in bed/bim/fam files. For a matrix as input, this return a matrix.

Usage

```
read.pcadapt(input, type = c("pcadapt", "lfmm", "vcf", "bed", "ped",
  "pool", "example"), type.out = c("bed", "matrix"),
  allele.sep = c("/", "|"), pop.sizes, ploidy, local.env, blocksize)
```

Arguments

`input`	a genotype matrix or a character string specifying the name of the file to be converted.
`type`	a character string specifying the type of data to be converted from. Converters from 'vcf' and 'ped' formats are not maintained anymore; if you have any issue with those, please use PLINK 1.9 to convert them to the 'bed' format.
`type.out`	Either a bed file or a standard R matrix. If the input is a matrix, then the output is automatically a matrix (so that you don't need to specify this parameter). If the input is a bed file, then the output is also a bed file.
`allele.sep`	a vector of characters indicating what delimiters are used in VCF files. By default, only "|" and "/" are recognized. So, this argument is only useful for `type = "vcf"`.
`pop.sizes`	deprecated argument.
`ploidy`	deprecated argument.
`local.env`	deprecated argument.
`blocksize`	deprecated argument.

run.pcadapt *Shiny app*

Description

pcadapt comes with a Shiny interface.

Usage

```
run.pcadapt()
```

vcf2pcadapt *vcfR-based converter*

Description

vcf2pcadapt uses the package vcfR to extract the genotype information from a vcf file and exports it under the format required by pcadapt.

Usage

```
vcf2pcadapt(input, output = "tmp.pcadapt", allele.sep = c("/", "|"))
```

Arguments

input	a character string specifying the name of the file to be converted.
output	a character string indicating the name of the output file.
allele.sep	a vector of characters indicating what delimiters are used to separate alleles.

Examples

```
## see also ?pcadapt for examples
```

writeBed *Write PLINK files*

Description

Function to write bed/bim/fam files from a pcadapt or an lfmm file. Files shouldn't already exists.

Usage

```
writeBed(file, is.pcadapt)
```

Arguments

file	A [mmapchar][mmapchar-class] object associated with a pcadapt or lfmm file.
is.pcadapt	a boolean value.

Value

The input 'bedfile' path.

Index