Package 'pcadapt'

January 2, 2017

Type Package

Title Fast Principal Component Analysis for Outlier Detection

Version 3.0.4

Date 2016-12-20

Author Keurcien Luu [aut, cre],
Michael Blum [aut],
Nicolas Duforet-Frebourg [ctb]

Maintainer Keurcien Luu <keurcien.luu@imag.fr>

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)

Depends robust, MASS, ggplot2, vcfR

Suggests knitr, qvalue, rmarkdown

LazyData TRUE

RoxygenNote 5.0.1

NeedsCompilation yes

VignetteBuilder knitr

Repository CRAN

Date/Publication 2017-01-02 22:22:26

R topics documented:

`pcadapt-package` *Principal Component Analysis for Outlier Detection.*

Description

This package has been developed to provide statistical tools for outlier detection based on Principal Component Analysis.

Details

Package:	pcadapt
Type:	Package
Version:	3.0.4
Date:	2016-12-20
License:	(>= 2)

For an overview of how to use the package, please check the html document provided as a vignette by typing the following command in the `R` console:

```
browseVignettes("pcadapt")
```

Author(s)

Keurcien Luu, Michael G.B. Blum

Maintainer: Keurcien Luu <keurcien.luu@imag.fr>

References

K Luu, E Bazin and MGB Blum. pcadapt: an R package for performing genome scans for selection based on principal component analysis. biorXiv:10.1101/056135 (2016)

See Also

http://membres-timc.imag.fr/Michael.Blum/PCAdapt.html

Examples

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

cover.to.pool — *Simulate frequency matrix from genotype and coverage matrices*

Description

`cover.to.pool` creates a matrix of frequency estimates, given a genotype matrix and a coverage matrix.

Usage

```
cover.to.pool(data, cover.matrix, pop, ploidy = 2)
```

Arguments

`data`	a matrix with n rows and p columns where n is the number of individuals and p is the number of markers.
`cover.matrix`	a matrix with n rows and p columns where n is the number of pools and is the number of markers.
`pop`	a list of integers or strings specifying which subpopulation the individuals belong to.
`ploidy`	an integer specifying the ploidy of the individuals.

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.pool.matrix *Convert genotypes to pooled samples*

Description

get.pool.matrix creates a pooled-sequenced data out of a genotype matrix, given the labels of each individuals.

Usage

```
get.pool.matrix(data, pop, ploidy = 2)
```

Arguments

data	a matrix with n rows and p columns where n is the number of individuals and p is the number of markers.
pop	a list of integers or strings specifying which subpopulation the individuals belong to.
ploidy	an integer specifying the ploidy of the individuals.

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 = 5, method = "mahalanobis", data.type = "genotype",
  min.maf = 0.05, ploidy = 2, output.filename = "pcadapt_output",
  clean.files = TRUE, transpose, cover.matrix = NULL)
```

Arguments

input	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. Three statistics are currently available, "mahalanobis", "communality" and "componentwise".

`data.type`	a character string specifying the type of data being read, either a `genotype` matrix (`data.type="genotype"`), or a matrix of allele frequencies (`data.type="pool"`).
`min.maf`	a value between `0` and `0.45` specifying the threshold of minor allele frequencies above which p-values are computed.
`ploidy`	an integer specifying the ploidy of the individuals.
`output.filename`	a character string specifying the names of the files created by `pcadapt`.
`clean.files`	a logical value indicating whether the auxiliary files should be deleted or not.
`transpose`	deprecated argument.
`cover.matrix`	a matrix specifying the average coverage per genetic marker and per population.

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.

`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 `method="communality"`, the test statistic is first multiplied by `K` and divided by the percentage of variance explained by the first `K` PCs before accounting for genomic inflation factor. When using `method="mahalanobis"` or `"communality"`, 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", K = NULL, i = 1,
  j = 2, pop, threshold = 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.
K	an integer specifying the principal component of interest. K has to be specified only when using the loadings option.
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.
threshold	for the "qqplot" option, it displays an additional bar which shows the threshold percent of SNPs with smallest p-values and separates them from SNPs with higher p-values.

Examples

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

read.pcadapt *File Converter*

Description

read.pcadapt converts .vcf and .ped files to an appropriate type of file readable by pcadapt. You may find the converted file in the current directory.

Usage

```
read.pcadapt(input.filename, type, local.env = FALSE, ploidy = 2,
  pop.sizes = NULL, allele.sep = "/", blocksize = 10000)
```

Arguments

input.filename	a character string specifying the name of the file to be converted if local.env = FALSE. If local.env = TRUE, input.filename refers to the genotype matrix in the local environment.
type	a character string specifying the type of data to be converted to the pcadapt format. Supported formats are: ped, vcf, lfmm.
local.env	a logical value indicating whether the input has to be read from the local environment or from the working directory.
ploidy	an integer specifying the ploidy of the individuals.
pop.sizes	a vector specifying the number of individuals for each pool.
allele.sep	a character string specifying the type of allele separator used in VCF files. Set to "/" by default, but can be switched to "|" otherwise.
blocksize	an integer specifying the number of markers to be processed in the mean time.

sample.geno *Sample genotype matrix from pooled samples*

Description

sample.geno samples a genotype matrix from pooled samples.

Usage

```
sample.geno(pool.matrix = NULL, ploidy = 2, cover.matrix = NULL,
  pop.sizes = NULL, method = "per.pop")
```

Arguments

pool.matrix	a matrix with n rows and p columns where n is the number of pools and is the number of markers.
ploidy	an integer specifying the ploidy.
cover.matrix	a matrix with n rows and p columns where n is the number of pools and is the number of markers.
pop.sizes	a list specifying the number of individuals for each pool.
method	a character string indicating the method used for sampling.

Examples

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

Index