Package 'pcadapt'

April 15, 2016

Type Package

Title Fast Principal Component Analysis for Outlier Detection

Version 3.0

Date 2016-04-04

Author Keurcien Luu, Michael G.B. Blum, Nicolas Duforet-Frebourg

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.

License GPL (>= 2)

Depends robust, MASS

Suggests knitr, qvalue, rmarkdown

Imports ggplot2

LazyData TRUE

RoxygenNote 5.0.1

NeedsCompilation yes

VignetteBuilder knitr

Repository CRAN

Date/Publication 2016-04-15 23:28:47

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
Date:	2016-04-04
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:

```
browseVignette("pcadapt")
```

Author(s)

Keurcien Luu, Michael G.B. Blum

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

References

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. B. (2015). Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. arXiv preprint arXiv:1504.04543.

See Also

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

Examples

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

create.pcadapt *pcadapt objects*

Description

create.pcadapt loads the numerical quantities needed to compute the test statistics, and stores them in an object of class pcadapt.

Usage

```
create.pcadapt(output.filename, K, method, data.type, min.maf)
```

Arguments

output.filename	a character string indicating which outputs from PCAdapt fast should be processed.
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. Four statistics are currently available, "mahalanobis", "communality", "euclidean" 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.

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", data.type = "genotype",
  min.maf = 0.05, ploidy = 2, output.filename = "pcadapt_output",
  clean.files = TRUE, transpose = FALSE)
```

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. Four statistics are currently available, "mahalanobis", "communality", "euclidean" 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	a logical value indicating whether the genotype matrix has to be tranposed or not. A genotype matrix should be p x n where p is the number of genetic markers and n is the number of individuals. If the data contains missing values, please encode missing values as 9 or use the function read4pcadapt to format the data.

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 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.

euclidean: the Euclidean distance between the K z-scores of each genetic marker and the mean of the K vectors of z-scores is computed.

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","euclidean". 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","communality","euclidean", 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 a pcadapt object. 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 guide in 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-valuesseparates the SNPs with the highest p-values.

Examples

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

read4pcadapt *File Converter*

Description

read4pcadapt 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

```
read4pcadapt(input.filename, type)
```

Arguments

input.filename	a character string specifying the name of the file to be converted.
type	a character string specifying the type of data to be converted to the pcadapt format. Supported formats are: ped, vcf, lfmm.

Index