Package 'GENESIS'

October 9, 2015

Type Package

Title GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness

Version 1.0.0

Date 2015-03-23

Author Matthew P. Conomos and Timothy Thornton

Maintainer Matthew P. Conomos <mconomos@uw.edu>

Description The GENESIS package provides methodology for estimating, inferring, and accounting for population and pedigree structure in genetic analyses. The current implementation provides functions to perform PC-AiR (Conomos et al., 2015): a Principal Components Analysis with genome-wide SNP genotype data for robust population structure inference in samples with related individuals (known or cryptic).

License GPL-3

Depends GWASTools

Suggests gdsfmt, SNPRelate, RUnit, BiocGenerics, knitr

VignetteBuilder knitr

biocViews SNP, GeneticVariability, Genetics, StatisticalMethod, DimensionReduction, PrincipalComponent, GenomeWideAssociation, QualityControl, BiocViews

NeedsCompilation no

R topics documented:

GENESIS-package	*GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness*

Description

The GENESIS package provides methodology for estimating, inferring, and accounting for population and pedigree structure in genetic analyses. The current implementation performs PC-AiR (Conomos et al., 2015): a Principal Components Analysis on genome-wide SNP data for the detection of population structure in a sample that may contain known or cryptic relatedness. Unlike standard PCA, PC-AiR accounts for relatedness in the sample to provide accurate ancestry inference that is not confounded by family structure.

Details

Package:	GENESIS
Type:	Package
Version:	0.99.4
Date:	2015-03-23
License:	GPL-3
Depends:	GWASTools
Suggests:	gdsfmt, SNPRelate, RUnit, BiocGenerics, knitr
VignetteBuilder:	knitr
biocViews:	SNP, GeneticVariability, Genetics, StatisticalMethod, DimensionReduction, PrincipalComponent, Genome

The main function in this package is `pcair`. This function takes genotype data and pairwise measures of kinship and ancestry divergence as input and returns PC-AiR PCs as the ouput. The function `pcairPartition` is called within `pcair` and uses the PC-AiR algorithm to partition the sample into an ancestry representative 'unrelated subset' and 'related subset'. The function `plot.pcair` can be used to plot pairs of PCs from a class 'pcair' object returned by the function `pcair`. The function `king2mat` can be used to convert output text files from the KING software (Manichaikul et al., 2010) into an R matrix of pairwise kinship coefficient estimates in a format that can be used by the functions `pcair` and `pcairPartition`.

Author(s)

Matthew P. Conomos and Timothy Thornton

Maintainer: Matthew P. Conomos <mconomos@uw.edu>

References

Conomos M.P., Miller M., & Thornton T. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. (Accepted to Genetic

Epidemiology).

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., ... & Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of Genome-Wide Association Studies. Bioinformatics, 28(24), 3329-3331.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., & Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867-2873.

`HapMap_ASW_MXL_KINGmat`

Matrix of Pairwise Kinship Coefficient Estimates for the combined HapMap ASW and MXL Sample found with the KING-robust estimator from the KING software.

Description

KING-robust kinship coefficient estimates for the combined HapMap African Americans in the Southwest U.S. (ASW) and Mexican Americans in Los Angeles (MXL) samples.

Usage

```
data(HapMap_ASW_MXL_KINGmat)
```

Format

The format is: num [1:173, 1:173] 0 0.00157 -0.00417 0.00209 0.00172 ...

Value

A matrix of pairwise kinship coefficient estimates as calculated with KING-robust for the combined HapMap African Americans in the Southwest U.S. (ASW) and Mexican Americans in Los Angeles (MXL) samples.

Source

http://hapmap.ncbi.nlm.nih.gov/

References

International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature, 467(7311), 52-58.

king2mat *Convert KING text output to an R Matrix*

Description

king2mat is used to extract the pairwise kinship coefficient estimates or IBS0 values from the output text files of KING and put them into an R object of class matrix that can be read by the functions pcair and pcairPartition.

Usage

```
king2mat(file.kin0, file.kin = NULL, iids = NULL,
            type = "kinship", verbose = TRUE)
```

Arguments

file.kin0	File name of the .kin0 text file output from KING.
file.kin	Optional file name of the .kin text file output from KING.
iids	An optional vector of individual IDs in the same order as desired for the output matrix. See 'Details' for more information.
type	Character string taking the values "kinship" (default) or "IBS0", to inform the function to read in kinship coefficeints or IBS0 values from the KING output.
verbose	A logical indicating whether or not to print status updates to the console; the default is TRUE.

Details

When using the function pcair, it is important that the order of individuals in the kinMat matrix matches the order of individuals in genoData. The KING software has a tendency to reorder individuals. If iids = NULL, the default is for the order to be taken from the KING output text file. By specifying iids the user can control the order of individuals in the output matrix. The IDs used for iids must be the same set of character IDs that are output as columns 'ID1' and 'ID2' in the KING output text files; all of the IDs specified in iids must be in the KING output, and all IDs in the KING output must be specified in iids.

Value

An object of class 'matrix' with pairwise kinship coefficients or IBS0 values as estimated by KING for each pair of individuals in the sample. The estimates are on both the upper and lower triangle of the matrix, and the diagonal is arbitrailly set to 0.5. Individual IDs are set as the column and row names of the matrix.

Author(s)

Matthew P. Conomos

References

Conomos M.P., Miller M., & Thornton T. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. (Accepted to Genetic Epidemiology).

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., & Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867-2873.

See Also

pcair and pcairPartition for functions that use the output matrix.

Examples

```
file.kin0 <- system.file("extdata", "MXL_ASW.kin0", package="GENESIS")
file.kin <- system.file("extdata", "MXL_ASW.kin", package="GENESIS")
KINGmat <- king2mat(file.kin0 = file.kin0, file.kin = file.kin, type="kinship")
```

---

pcair *PC-AiR: Principal Components Analysis in Related Samples*

Description

pcair is used to perform a Principal Components Analysis using genome-wide SNP data for the detection of population structure in a sample. Unlike a standard PCA, PC-AiR accounts for sample relatedness (known or cryptic) to provide accurate ancestry inference that is not confounded by family structure.

Usage

```
pcair(genoData, v = 10, MAF = 0.05, kinMat = NULL, kin.thresh = 0.025,
            divMat = NULL, div.thresh = -0.025, unrel.set = NULL,
            scan.include = NULL, snp.include = NULL, Xchr = FALSE,
            block.size = 10000, verbose = TRUE)
## S3 method for class 'pcair'
print(x, ...)
## S3 method for class 'pcair'
summary(object, ...)
## S3 method for class 'summary.pcair'
print(x, ...)
```

Arguments

genoData	An object of class GenotypeData from the package GWASTools containing the genotype data for SNPs and samples to be used for the analysis. This object can easily be created from a matrix of SNP genotype data, PLINK files, or GDS files.

v	The number of principal components to be returned; the default is 10. If v = NULL, then all the principal components are returned.
MAF	Minor allele frequency filter; any SNPs with MAF less than this value will be excluded from the analysis; the default value is 0.05.
kinMat	An optional symmetric matrix of pairwise kinship coefficients for every pair of individuals in the sample (the values on the diagonal do not matter, but the upper and lower triangles must both be filled) used for partitioning the sample into the 'unrelated' and 'related' subsets. See 'Details' for how this interacts with kin.thresh and unrel.set. IDs for each individual must be set as the row and column names of the matrix.
kin.thresh	Threshold value on kinMat used for declaring each pair of individuals as related or unrelated. The default value is 0.025. See 'Details' for how this interacts with kinMat.
divMat	An optional symmetric matrix of pairwise divergence measures for every pair of individuals in the sample (the values on the diagonal do not matter, but the upper and lower triangles must both be filled) used for partitioning the sample into the 'unrelated' and 'related' subsets. See 'Details' for how this interacts with div.thresh. IDs for each individual must be set as the row and column names of the matrix.
div.thresh	Threshold value on divMat used for deciding if each pair of individuals is ancestrally divergent. The default value is -0.025. See 'Details' for how this interacts with divMat.
unrel.set	An optional vector of IDs for identifying individuals that are forced into the unrelated subset. See 'Details' for how this interacts with kinMat.
scan.include	A vector of IDs for samples to include in the analysis. If NULL, all samples are included.
snp.include	A vector of SNP IDs to include in the analysis. If NULL, all SNPs are included (see Xchr for further details).
Xchr	Logical indicator for whether the analysis is of X chromosome SNPs; the default is FALSE. If snp.include is NULL: when FALSE only autosomal SNPs are analyzed; when TRUE only X chromosome SNPs are analyzed.
block.size	The number of SNPs to read-in/analyze at once. The default value is 10000.
verbose	Logical indicator of whether updates from the function should be printed to the console; the default is TRUE.
object	An object of class 'pcair', i.e. output from the pcair function.
x	An object of class 'pcair', i.e. output from the pcair function.
...	Further arguments passed to or from other methods.

Details

The basic premise of PC-AiR is to partition the entire sample of individuals into an ancestry representative 'unrelated subset' and a 'related set', perform standard PCA on the 'unrelated subset', and predict PC values for the 'related subset'.

We recommend using software that accounts for population structure to estimate pairwise kinship coefficients to be used in kinMat. Any pair of individuals with a pairwise kinship greater than

kin.thresh will be declared 'related.' Kinship coefficient estimates from the KING-robust software are used as measures of ancestry divergence in divMat. Any pair of individuals with a pairwise divergence measure less than div.thresh will be declared ancestrally 'divergent'. Typically, kin.thresh and div.thresh are set to be the amount of error around 0 expected in the estimate for a pair of truly unrelated individuals.

If divMat = NULL and kinMat is specified, the kinship coefficient estimates in kinMat will also be used as divergence measures in place of divMat.

It is important that the order of individuals in the matrices kinMat and divMat match the order of individuals in the genoData.

There are multiple ways to partition the sample into an ancestry representative 'unrelated subset' and a 'related subset'. If kinMat is specified and unrel.set = NULL, then the PC-AiR algorithm is used to find an 'optimal' partition (see 'References' for a paper describing the algorithm). If kinMat = NULL and unrel.set is specified, then the individuals with IDs in unrel.set are used as the 'unrelated subset'. If both kinMat and unrel.set are specified, then all individuals with IDs in unrel.set are forced in the 'unrelated subset' and the PC-AiR algorithm is used to partition the rest of the sample; this is especially useful for including reference samples of known ancestry in the 'unrelated subset'. If kinMat = NULL and unrel.set = NULL, then a standard principal components analysis that does not account for relatedness is performed.

Value

An object of class 'pcair'. A list including:

vectors	A matrix of the top v principal components; each column is a principal component. Sample IDs are provided as rownames.
values	A vector of eigenvalues matching the top v principal components. These values are determined from the standard PCA run on the 'unrelated subset'.
sum.values	The sum of all the eigenvalues from the standard PCA run on the 'unrelated subset' (regardless of how many were returned).
rels	A vector of IDs for individuals in the 'related subset'.
unrels	A vector of IDs for individuals in the 'unrelated subset'.
kin.thresh	The threshold value used for declaring each pair of individuals as related or unrelated.
div.thresh	The threshold value used for determining if each pair of individuals is ancestrally divergent.
nsamp	The total number of samples in the analysis.
nsnps	The total number of SNPs used in the analysis, after filtering on MAF.
MAF	The minor allele frequency (MAF) filter used on SNPs.
call	The function call passed to pcair.
method	A character string. Either "PC-AiR" or "Standard PCA" identifying which method was used for computing principal components.

Note

The GenotypeData function in the GWASTools package should be used to create the input genoData. Input to the GenotypeData function can easily be created from an R matrix or GDS file. PLINK .bed, .bim, and .fam files can easily be converted to a GDS file with the function snpgdsBED2GDS in the SNPRelate package.

Author(s)

Matthew P. Conomos

References

Conomos M.P., Miller M., & Thornton T. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. (Accepted to Genetic Epidemiology).

Gogarten, S.M., Bhangale, T., Conomos, M.P., Laurie, C.A., McHugh, C.P., Painter, I., ... & Laurie, C.C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of Genome-Wide Association Studies. Bioinformatics, 28(24), 3329-3331.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., & Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867-2873.

See Also

pcairPartition for a description of the function used by pcair that can be used to partition the sample into 'unrelated' and 'related' subsets without performing PCA. plot.pcair for plotting. king2mat for creating a matrix of pairwise kinship coefficient estimates from KING output text files that can be used for kinMat or divMat. GWASTools for a description of the package containing the following functions: GenotypeData for a description of creating a GenotypeData class object for storing sample and SNP genotype data, MatrixGenotypeReader for a description of reading in genotype data stored as a matrix, and GdsGenotypeReader for a description of reading in genotype data stored as a GDS file. Also see snpgdsBED2GDS in the SNPRelate package for a description of converting binary PLINK files to GDS. The generic functions summary and print.

Examples

```
# file path to GDS file
gdsfile <- system.file("extdata", "HapMap_ASW_MXL_geno.gds", package="GENESIS")
# read in GDS data
HapMap_geno <- GdsGenotypeReader(filename = gdsfile)
# create a GenotypeData class object
HapMap_genoData <- GenotypeData(HapMap_geno)
# load saved matrix of KING-robust estimates
data("HapMap_ASW_MXL_KINGmat")
# run PC-AiR
mypcair <- pcair(genoData = HapMap_genoData, kinMat = HapMap_ASW_MXL_KINGmat,
                 divMat = HapMap_ASW_MXL_KINGmat)
close(HapMap_genoData)
```

pcairPartition *Partition a sample into an ancestry representative 'unrelated subset' and a 'related subset'*

Description

pcairPartition is used to partition a sample from a genetic study into an ancestry representative 'unrelated subset' and a 'related subset'. The 'unrelated subset' contains individuals who are all mutually unrelated to each other and representative of the ancestries of all individuals in the sample, and the 'related subset' contains individuals who are related to someone in the 'unrealted subset'.

Usage

```
pcairPartition(kinMat, kin.thresh = 0.025, divMat = NULL,
                    div.thresh = -0.025, unrel.set = NULL)
```

Arguments

kinMat	A symmetric matrix of pairwise kinship coefficients for every pair of individuals in the sample (the values on the diagonal do not matter, but the upper and lower triangles must both be filled) used for partitioning the sample into the 'unrelated' and 'related' subsets. See 'Details' for how this interacts with kin.thresh and unrel.set. IDs for each individual must be set as the row and column names of the matrix.
kin.thresh	Threshold value on kinMat used for declaring each pair of individuals as related or unrelated. The default value is 0.025. See 'Details' for how this interacts with kinMat.
divMat	A symmetric matrix of pairwise ancestry divergence measures for every pair of individuals in the sample (the values on the diagonal do not matter, but the upper and lower triangles must both be filled) used for partitioning the sample into the 'unrelated' and 'related' subsets. See 'Details' for how this interacts with div.thresh. IDs for each individual must be set as the row and column names of the matrix.
div.thresh	Threshold value on divMat used for deciding if each pair of individuals is ancestrally divergent. The default value is -0.025. See 'Details' for how this interacts with divMat.
unrel.set	An optional vector of IDs for identifying individuals that are forced into the unrelated subset. See 'Details' for how this interacts with kinMat.

Details

We recommend using software that accounts for population structure to estimate pairwise kinship coefficients to be used in kinMat. Any pair of individuals with a pairwise kinship greater than kin.thresh will be declared 'related.' Kinship coefficient estimates from the KING-robust software are typically used as measures of ancestry divergence in divMat. Any pair of individuals with a pairwise divergence measure less than div.thresh will be declared ancestrally 'divergent'.

Typically, kin.thresh and div.thresh are set to be the amount of error around 0 expected in the estimate for a pair of truly unrelated individuals. If unrel.set = NULL, the PC-AiR algorithm is used to find an 'optimal' partition (see 'References' for a paper describing the algorithm). If unrel.set and kinMat are both specified, then all individuals with IDs in unrel.set are forced in the 'unrelated subset' and the PC-AiR algorithm is used to partition the rest of the sample; this is especially useful for including reference samples of known ancestry in the 'unrelated subset'.

Value

A list including:

rels	A vector of IDs for individuals in the 'related subset'.
unrels	A vector of IDs for individuals in the 'unrelated subset'.

Note

pcairPartition is called internally in the function pcair but may also be used on its own to partition the sample into an ancestry representative 'unrelated' subset and a 'related' subset without performing PCA.

Author(s)

Matthew P. Conomos

References

Conomos M.P., Miller M., & Thornton T. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. (Accepted to Genetic Epidemiology).

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., & Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867-2873.

See Also

pcair which uses this function for finding principal components in the presence of related individuals. king2mat for creating a matrix of kinship coefficent estimates or pairwise ancestry divergence measures from KING output text files that can be used as kinMat or divMat.

Examples

```
# load saved matrix of KING-robust estimates
data("HapMap_ASW_MXL_KINGmat")
# partition the sample
part <- pcairPartition(kinMat = HapMap_ASW_MXL_KINGmat,
divMat = HapMap_ASW_MXL_KINGmat)
```

plot.pcair *PC-AiR: Plotting PCs*

Description

plot.pcair is used to plot pairs of principal components contained in a class 'pcair' object obtained as output from the pcair function.

Usage

```
## S3 method for class 'pcair'
plot(x, vx = 1, vy = 2, pch = NULL, col = NULL,
        xlim = NULL, ylim = NULL, main = NULL, sub = NULL,
        xlab = NULL, ylab = NULL, ...)
```

Arguments

x	An object of class 'pcair' obtained as output from the pcair function.
vx	An integer indicating which principal component to plot on the x-axis; the default is 1.
vy	An integer indicating which principal component to plot on the y-axis; the default is 2.
pch	Either an integer specifying a symbol or a single character to be used in plotting points. If NULL, the default is dots for the 'unrelated subset' and + for the 'related subset'.
col	A specification for the plotting color for points. If NULL, the default is black for the 'unrelated subset' and blue for the 'related subset'.
xlim	The range of values shown on the x-axis. If NULL, the default shows all points.
ylim	The range of values shown on the y-axis. If NULL, the default shows all points.
main	An overall title for the plot. If NULL, the default specifies which PC-AiR PCs are plotted.
sub	A sub title for the plot. If NULL, the default is none.
xlab	A title for the x-axis. If NULL, the default specifies which PC-AiR PC is plotted.
ylab	A title for the y-axis. If NULL, the default specifies which PC-AiR PC is plotted.
...	Other parameters to be passsed through to plotting functions, (see par).

Details

This function provides a quick and easy way to plot principal components obtained with the function pcair to visualize the population structure captured by PC-AiR.

Value

A figure showing the selected principal components plotted against each other.

Author(s)

Matthew P. Conomos

See Also

pcair for obtaining principal components that capture population structure in the presence of relatedness. par for more in depth descriptions of plotting parameters. The generic function plot.

Examples

```
# file path to GDS file
gdsfile <- system.file("extdata", "HapMap_ASW_MXL_geno.gds", package="GENESIS")
# read in GDS data
HapMap_geno <- GdsGenotypeReader(filename = gdsfile)
# create a GenotypeData class object
HapMap_genoData <- GenotypeData(HapMap_geno)
# load saved matrix of KING-robust estimates
data("HapMap_ASW_MXL_KINGmat")
# run PC-AiR
mypcair <- pcair(genoData = HapMap_genoData, kinMat = HapMap_ASW_MXL_KINGmat,
                 divMat = HapMap_ASW_MXL_KINGmat)
# plot top 2 PCs
plot(mypcair)
# plot PCs 3 and 4
plot(mypcair, vx = 3, vy = 4)
close(HapMap_genoData)
```

Index