Package 'GeneticSubsetter'

October 12, 2022

Type Package

Title Identify Favorable Subsets of Germplasm Collections

Version 0.8

Date 2016-10-25

Author Ryan C. Graebner and Alfonso Cuesta-Marcos

Maintainer Ryan C. Graebner <ryan.graebner@gmail.com>

Description Finds subsets of sets of genotypes with a high Heterozygosity, and Mean of Transformed Kinships (MTK), measures that can indicate a subset would be beneficial for rare-trait discovery and genome-wide association scanning, respectively.

License GPL (>= 2)

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2016-10-25 23:03:53

R topics documented:

`CoreSetOptimizer` *Subset Optimization*

Description

This function works to systematically improves a subset via single-genotype replacements from a larger population. This function will continue to work until no more single-genotype replacements can be made to increase the subset's value. Criteria that can be used to judge the value of subsets are Expected Heterozygosity (HET; for rare-trait discovery; called PIC in earlier versions and in the paper describing this package), and the Mean of Transformed Kinships (MTK; for GWAS). A complete comparison of these two criteria is presented in Graebner et al. (2015).

Usage

```
CoreSetOptimizer(genos=NULL, subset=NULL, criterion = "HET",
    mat = NULL, save = NULL, power = 10, print = TRUE)
```

Arguments

`genos`	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, or NA, where 0 represents a heterozygous marker, and NA represents missing data. Note that this coding is different from the earlier SubsetOptimizerPIC and SubsetOptimizerMTK, which cannot handle heterozygous markers. All data in this matrix must be numeric.
`subset`	The names of the genotypes in the starting subset.
`criterion`	The criterion to be used for comparing subsets (HET or MTK).
`mat`	A kinship matrix, if one has already been computed for the population. If a kinship matrix is included, the "genos" argument may be left empty.
`save`	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.
`power`	The transformation that should be made to the kinship matrix, if the MTK criterion is used. If power=1, the kinship matrix is not transformed, if power=2, the kinship matrix is squared, etc. When the power is higher, this function preferentially eliminates genotypes that are closely related to other genotypes in the population.
`print`	Whether to the value of intermediate subsets.

Value

Returns a list of the genotype names included in the best subset found.

Note

The ability to recogize heterozygous markers was included in CoreSetOptimizer, resulting in a slightly different genotype coding scheme than the depreciated functions SubsetOptimizerPIC and SubsetOptimizerMTK.

Author(s)

Ryan C. Graebner

References

Graebner RC, Hayes PM, Hagerty CH, Cuesta-Marcos A (2016) A comparison of polymorphism information content and mean of transformed kinships as criteria for selection informative subsets of barley (Hordeum vulgare L. s. l) from the USDA Barley Core Collection. Genet Resour Crop Evol 63:477-482.

Examples

```
data("genotypes")
CoreSetOptimizer(genotypes,subset=colnames(genotypes)[c(1,3,5,7,8,9)],
    criterion="HET",save=colnames(genotypes)[c(1,5,9)])
CoreSetOptimizer(genotypes,subset=colnames(genotypes)[c(1,3,5,7,8,9)],
    criterion="MTK",save=colnames(genotypes)[c(1,5,9)])
```

CoreSetter *Genotype Subsetting*

Description

This function systematically eliminates genotypes from a large population to arrive at a favorable subset. This method will typically return less favorable subsets than the method used by the CoreSetterCombined function if sufficient permutations are used for the later, but CoreSetter is quicker, and will rank all genotypes, as opposed to returning a single, static subset. Criteria that can be used to judge the value of subsets are Expected Heterozygosity (HET; for rare-trait discovery; called PIC in earlier versions and in the paper describing this package), and the Mean of Transformed Kinships (MTK; for GWAS). A complete comparison of these two criteria is presented in Graebner et al. (2015).

Usage

```
CoreSetter(genos=NULL, criterion = "HET", save = NULL,
    power = 10, mat = NULL)
```

Arguments

genos	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, or NA, where 0 represents a heterozygous marker, and NA represents missing data. Note that this coding is different from the earlier SubsetterPIC and SubsetterMTK, which cannot handle heterozygous markers. All data in this matrix must be numeric.
criterion	The criterion to be used for comparing subsets (HET or MTK).
save	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.

power — The transformation that should be made to the kinship matrix, if the MTK criterion is used. If power=1, the kinship matrix is not transformed, if power=2, the kinship matrix is squared, etc. When the power is higher, this function preferentially eliminates genotypes that are closely related to other genotypes in the population.

mat — A kinship matrix, if one has already been computed for the population. If a kinship matrix is included, the "genos" argument may be left empty.

Value

Returns a matrix with three columns. The first column is the rank of a particular genotype to the population's MTK, based on the order in which genotypes were eliminated (genotypes with lower rank were retained longer, genotypes with rank of 1 were not eliminated). The second column is the name of the genotype. The third column shows the value of the subset including that genotype and all genotypes with a lower rank, as judged by the specified criterion.

Note

In Graebner et al. (2015), and in their earlier functions SubsetterPIC and SubsetterMTK, heterozygous markers were counted as missing data, due to previous limitations if GeneticSubsetter. Please note that this has changed the required coding scheme for input genotype data. When using the HET criterion, this function uses the same method and criteria described in Munoz-Amatrain et al. (2014), but with a more computationally efficient approach.

Author(s)

Ryan C. Graebner and Alfonso Cuesta-Marcos

References

Graebner RC, Hayes PM, Hagerty CH, Cuesta-Marcos A (2016) A comparison of polymorphism information content and mean of transformed kinships as criteria for selection informative subsets of barley (Hordeum vulgare L. s. l) from the USDA Barley Core Collection. Genet Resour Crop Evol 64:477-482. Munoz-Amatrain M, Cuesta-Marcos A, Endelman JB, Comadran J, Bonman JM (2014) The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies. PloS One 9:e94688.

Examples

```
data("genotypes")
CoreSetter(genotypes,criterion="HET",save=colnames(genotypes)[c(1,5,9)])
CoreSetter(genotypes,criterion="MTK",save=colnames(genotypes)[c(1,5,9)])
```

`CoreSetterCombined` *Genotype Subsetting-Alternative Method*

Description

This function creates a series of random subsets. Then, each of these subsets is improved using the CoreSetOptimizer function using a series of single-genotype replacements that result in a higher value for the subset, until no more single-genotype replacements can be made to improve the subset. This process is similar to a Local Search. Criteria that can be used to judge the value of subsets are Expected Heterozygosity (HET; for rare-trait discovery; called PIC in earlier versions and in the paper describing this package), and the Mean of Transformed Kinships (MTK; for GWAS). A complete comparison of these two criteria is presented in Graebner et al. (2015).

Usage

```
CoreSetterCombined(genos=NULL, size=NULL, criterion =  "HET",
    save = NULL, power = 10, permutations = 100, print = TRUE,
    mat = NULL)
```

Arguments

`genos`	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, or NA, where 0 represents a heterozygous marker, and NA represents missing data. Note that this coding is different from the earlier SubsetterCombinedPIC and SubsetterCombinedMTK, which cannot handle heterozygous markers. All data in this matrix must be numeric.
`size`	The desired subset size.
`criterion`	The criterion to be used for comparing subsets (HET or MTK).
`save`	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.
`power`	The transformation that should be made to the kinship matrix, if the MTK criterion is used. If power=1, the kinship matrix is not transformed, if power=2, the kinship matrix is squared, etc. When the power is higher, this function preferentially eliminates genotypes that are closely related to other genotypes in the population.
`permutations`	The number of random subsets to improve.
`print`	If TRUE, this function prints the Heterozygosity or MTK of the best subset identified.
`mat`	A kinship matrix, if one has already been computed for the population. If a kinship matrix is included, the "genos" argument may be left empty.

Value

Returns a list of the genotype names included in the best subset found.

Note

The ability to recognize heterozygous markers was included in CoreSetterCombined, resulting in a slightly different genotype coding scheme than the depreciated functions SubsetterCombinedPIC and SubsetterCombinedMTK.

Author(s)

Ryan C Graebner

References

Graebner RC, Hayes PM, Hagerty CH, Cuesta-Marcos A (2016) A comparison of polymorphism information content and mean of transformed kinships as criteria for selection informative subsets of barley (Hordeum vulgare L. s. l) from the USDA Barley Core Collection. Genet Resour Crop Evol 63:477-482.

Examples

```
data("genotypes")
CoreSetterCombined(genotypes,size=6,criterion="HET",permutations=10,
    save=colnames(genotypes)[c(1,5,9)])
CoreSetterCombined(genotypes,size=6,criterion="MTK",permutations=10,
    save=colnames(genotypes)[c(1,5,9)])
```

CoreSetterPoly *Genotype Subsetting for Autopolyploids and Polymorphic Markers*

Description

This function systematically eliminates genotypes from a large population to arrive at a favorable subset, and can accommodate datasets with autopolyploids and polymorphic markers. At this time, CoreSetterPoly can only use the Expected Heterozygosity criterion to quantify the value of subsets, and Sequential Backward Selection to arrive at favorable subsets.

Usage

```
CoreSetterPoly(genos, ploidy, save = NULL)
```

Arguments

genos	A matrix of genotypes, where each row is one individual, and each set of X columns (where X is the ploidy) is one locus. At each locus, any number of alleles can be included, where each allele is referred to by a different integer. Missing data should be represented by NA. The X cells for any genotype-locus combination are the alleles known to be present at that locus for that genotype, in the frequency that they are present. If at least one but not all of the cells for a genotype-locus combination are listed as NA, that data point is imputed based on the other alleles at that locus.

ploidy	The ploidy of the organizm to be subsetted, in respect to the number of alleles that can be present at one locus.
save	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.

Value

Returns a matrix with three columns. The first column is the rank of a particular genotype to the population's MTK, based on the order in which genotypes were eliminated (genotypes with lower rank were retained longer, genotypes with rank of 1 were not eliminated). The second column is the name of the genotype. The third column shows the value of the subset including that genotype and all genotypes with a lower rank, as judged by the Expected Heterozygosity criterion.

Author(s)

Ryan C. Graebner and Alfonso Cuesta-Marcos

Examples

```
data("genotypes")
CoreSetterPoly(genotypes.poly,ploidy=2,save=rownames(genotypes.poly)[c(1,5,9)])
```

GeneticSubsetter *Genetic Subsetter*

Description

This package contains a set of tools that can be used to select a subset from a larger population, using genetic data. Two criteria are used to identify subsets, in seperate functions: Expected Heterozygosity (HET; called PIC in earlier versions and in the paper describing this package) and the Mean of Transformed Kinships (MTK).

Details

When selecting subsets of genotypes, two factors are important to consider: the criteria by which to judge subsets, and the method used to identfy the set of genotypes that best fit that criteria. Two criteria are Expected Heterozygosity (HET) and the Mean of Transformed Kinships (MTK). Tests suggest that of these two criteria, Expected Heterozygosity is better if the resulting subset will be used for rare-trait discovery, while MTK is better if the resulting subset will be used for genome-wide association scanning (Graebner et al. 2015). To reach subsets with a high Expected Heterozygosity or MTK, CoreSetter systematically removes genotypes from the full set, creating a full ranking of genotype's contributions to their respective criteria. When the HET criterion is selected, CoreSetter uses the same method and criteria described in Munoz-Amatrain et al. (2014), except CoreSetter uses a more computationally efficient approach, and CoreSetter can consider heterozygous markers. Alternatively, CoreSetterCombined works to systematically improve a user-defined number of random subsets via single-genotype replacements, until no replacement can increase the selected criteria. This later method generally returns subsets with a higher Heterozyosity or MTK, but are subset-size specific, take more time to compute, and will not always return identical results in subsequent runs.

Author(s)

Ryan C. Graebner <ryan.graebner@gmail.com> and Alfonso Cuesta-Marcos

References

Graebner RC, Hayes PM, Hagerty CH, Cuesta-Marcos A (2016) A comparison of polymorphism information content and mean of transformed kinships as criteria for selection informative subsets of barley (Hordeum vulgare L. s. l) from the USDA Barley Core Collection. Genet Resour Crop Evol 63:477-482. Munoz-Amatrain M, Cuesta-Marcos A, Endelman JB, Comadran J, Bonman JM (2014) The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies. PloS One 9:e94688.

Examples

```
data("genotypes")
CoreSetter(genotypes,criterion="HET",save=colnames(genotypes)[c(1,5,9)])
```

---

genotypes *Sample Barley Genotypes*

Description

Twenty diploid barley genotypes, with twenty markers each, for to demonstrate functions in the GeneticSubsetter package.

Usage

```
data(genotypes)
```

Format

Columns are gentoypes, and rows are markers, formatted for the CoreSetter, CoreSetterCombined, CoreSetOptimizer, HET, MTK.

Source

Triticeae Coordinated Agricultural Project (T-CAP) (http://triticeaetoolbox.org)

Examples

```
data("genotypes")
str(genotypes)
```

genotypes.poly *Sample Barley Genotypes*

Description

Twenty diploid barley genotypes, with twenty markers each, for to demonstrate functions in the GeneticSubsetter package, formatted for CoreSetterPoly.

Usage

```
data(genotypes)
```

Format

Rows are genotypes, and each set of two columns is one locus.

Source

Triticeae Coordinated Agricultural Project (T-CAP) (http://triticeaetoolbox.org)

Examples

```
data("genotypes")
str(genotypes.poly)
```

HET *Heterozygosity Calculator*

Description

This function calculates the Expected Heterozygosity (HET; called PIC in earlier versions and in the paper describing this package) of a set of genotypes.

Usage

```
HET(data)
```

Arguments

data	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, or NA, where 0 represents a heterozygous marker, and NA represents missing data. Note that this coding is different from the earlier PicCalc, which cannot handle heterozygous markers. All data in this matrix must be numeric.

Value

The mean Heterozygosity of all markers for the given set of genotypes.

Note

The ability to recogize heterozygous markers was included in HET, resulting in a slightly different genotype coding scheme than the earlier PicCalc.

Author(s)

Ryan C. Graebner

Examples

```
data("genotypes")
HET(genotypes)
```

Mat *Kinship Matrix Creator*

Description

This function creates a kinship matrix for a set of genotypes. This function is a simplified version of the function "A.mat" in the R package rrBLUP.

Usage

```
Mat(genos)
```

Arguments

genos	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, or NA, where 0 represents a heterozygous marker, and NA represents missing data. All data in this matrix must be numeric.

Value

A matrix of kinship values between genotypes.

Author(s)

Ryan C. Graebner

References

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4:250-255.

Examples

```
data(genotypes)
Mat(genotypes)
```

MTK *MTK calculator*

Description

This function calculates the Mean of Transformed Kinships (MTK) of a set of genotypes.

Usage

```
MTK(genos, subset, mat = NULL, power = 10)
```

Arguments

genos	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, or NA, where 0 represents a heterozygous marker, and NA represents missing data. Note that this coding is different from the earlier MtkCalc, which cannot handle heterozygous markers. All data in this matrix must be numeric.
subset	A vector of genotype names for which to calculate MTK.
mat	A kinship matrix, if one has already been computed for the population.
power	The transformation that should be made to the kinship matrix, if the MTK criterion is used. If power=1, the kinship matrix is not transformed, if power=2, the kinship matrix is squared, etc. When the power is higher, this function preferentially eliminates genotypes that are closely related to other genotypes in the population.

Value

The MTK of the specified subset.

Note

The ability to recognize heterozygous markers was included in MTK, resulting in a slightly different genotype coding scheme than the earlier MtkCalc.

Author(s)

Ryan C. Graebner

Examples

```
data(genotypes)
MTK(genotypes,subset=colnames(genotypes[1:5]))
```

MtkCalc *MTK calculator (Depreciated)*

Description

***This function has been superseded by MTK.

Usage

```
MtkCalc(genos, subset, power = 10)
```

Arguments

genos	A matrix of genotypes, that includes all genotypes that should be used to create the kinship matrix, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, where 0 represents missing data. All data in this matrix must be numeric.
subset	A vector of genotype names for which to calculate MTK.
power	The transformation that should be made to the kinship matrix. If power=1, the kinship matrix is not transformed, if power=2, the kinship matrix is squared, etc. When the power is higher, this function preferentially eliminates genotypes that are closely related to other specific genotypes.

Value

The MTK of the specified subset.

Author(s)

Ryan C. Graebner

Examples

```
data(genotypes)
MtkCalc(genotypes,subset=colnames(genotypes[1:5]))
```

PicCalc *PIC Calculator (Depreciated)*

Description

***This function has been superseded by PIC.

Usage

```
PicCalc(data)
```

Arguments

data	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, where 0 represents missing data. All data in this matrix must be numeric.

Value

The mean PIC of all markers included for the given set of genotypes.

Author(s)

Ryan C. Graebner and Alfonso Cuesta-Marcos

Examples

```
data("genotypes")
PicCalc(genotypes)
```

SubsetterMTK *Genotype Subsetting with PIC - Method One (Depreciated)*

Description

***This function has been superseded by CoreSetter.

Usage

```
SubsetterMTK(genos, save = NULL, power = 10, mat = NULL)
```

Arguments

genos	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, where 0 represents missing data. All data in this matrix must be numeric.
save	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.
power	The transformation that should be made to the kinship matrix. If power=1, the kinship matrix is not transformed, if power=2, the kinship matrix is squared, etc. When the power is higher, this function preferentially eliminates genotypes that are closely related to other specific genotypes.
mat	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.

Value

Returns a matrix with four columns. The first column is the importance of a particular genotype to the population's MTK, based on the order in which genotypes were eliminated. The second column is the name of the genotype. The third and fourth columns are the MTK, and the mean of untransformed kinship values, respectively, of a population that includes the corresponding genotype, plus all genotypes that are more important.

Author(s)

Ryan C. Graebner

Examples

```
data("genotypes")
SubsetterMTK(genotypes,save=colnames(genotypes)[c(1,5,9)])
```

SubsetterPIC *Genotype Subsetting with PIC - Method One (Depreciated)*

Description

***This function has been superseded by CoreSetter.

Usage

```
SubsetterPIC(genos, save = NULL)
```

Arguments

genos	A matrix of genotypes, where each column is one individual, each row is one marker, and marker values are 1, 0, or -1, where 0 represents missing data. All data in this matrix must be numeric.
save	A list of genotype names, corresponding to the column names in the genotype matrix, that will not be eliminated.

Value

Returns a matrix with three columns. The first column is the importance of a particular genotype to the population's genetic diversity, based on the order in which genotypes were eliminated. The second column is the name of the genotype, and the third column is the mean PIC of a population that includes the corresponding genotype, plus all genotypes that are more important.

Author(s)

Ryan C. Graebner and Alfonso Cuesta-Marcos

References

Munoz-Amatrain M, Cuesta-Marcos A, Endelman JB, Comadran J, Bonman JM (2014) The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies. PloS One 9:e94688.

Examples

```
data("genotypes")
SubsetterPIC(genotypes,save=colnames(genotypes)[c(1,5,9)])
```

Index