Package 'polyBreedR'

November 11, 2022

Title Genomics-assisted breeding for polyploids (and diploids)

Version 0.28

Author Jeffrey B. Endelman

Maintainer Jeffrey Endelman <endelman@wisc.edu>

Description Genomics-assisted breeding for polyploids (and diploids)

Depends R (>= 4.0)

License GPL-3

LazyData true

RoxygenNote 7.1.1

Encoding UTF-8

Imports AGHmatrix, ggplot2, ggrepel, pedigree, grDevices, utils, tidyr, Matrix, methods, rlang, updog

Suggests knitr, rmarkdown, asreml

VignetteBuilder knitr

R topics documented:

array2VCF *Generate pseudo-VCF file from SNP array data*

Description

Creates VCF file with GT and AD fields

Usage

```
array2VCF(VCF.file, ploidy, AD, GT = NULL, map = NULL, header = NULL)
```

Arguments

VCF.file	name of VCF file to create
ploidy	ploidy
AD	matrix of allele depths, see readXY
GT	optional, matrix of genotype dosages (0,1,2..ploidy)
map	optional, 3 column data frame (marker,chrom,pos)
header	optional, header text for the VCF file

Details

AD and GT matrices should be markers x indiv. Per the VCF standard, the "." is used for missing data.

A_mat *Additive relationship matrix from pedigree*

Description

Additive relationship matrix from pedigree

Usage

```
A_mat(ped, ploidy, order.ped = TRUE)
```

Arguments

ped	Pedigree in three column format: id, mother, father
ploidy	2 or 4
order.ped	TRUE/FALSE does the pedigree need to be ordered so that progeny follow parents

Details

This is a wrapper that prepares the pedigree in the format required for R package AGHmatrix by Amadeu et al. (2016) (cite them if you use this function). A random bivalents model for tetraploid meiosis is assumed.

Value

Additive relationship matrix (dim: indiv x indiv)

References

Amadeu et al. (2016) Plant Genome 9, doi:10.3835/plantgenome2016.01.0009

`check_ploidy` *Check ploidy*

Description

Fraction of simplex or triplex markers

Usage

```
check_ploidy(geno, map)
```

Arguments

`geno`	Genotype matrix (markers x indiv)
`map`	Data frame with marker map (Marker, Chrom, Position)

Details

For every indiv in the genotype matrix, the fraction of markers per chromosome called as simplex or triplex is calculated, which should be low for diploids. A small amount of missing genotype data can be tolerated.

Value

List containing

mat Matrix (indiv x chrom) of results

plot ggplot2 barplot

`check_trio` *Check markers for parent-offspring trio*

Description

Check markers for parent-offspring trio

Usage

```
check_trio(parentage, geno, ploidy)
```

Arguments

parentage	Data frame with three columns: id, mother, father
geno	Matrix of allele dosages: markers x indiv
ploidy	2 or 4

Details

Computes the percentage of markers at which the two parents and offspring have incompatible allele dosages (for tetraploids, the random bivalents model is used). For dihaploid offspring of a single tetraploid parent, use `ploidy = 4` and "haploid" for the father in `parentage`, as well as a diploid (0,1,2) genotype for the offspring. A small amount of missing genotype data can be tolerated.

Value

Data frame with the percentage of incompatible markers for each trio

---

dart_tag *Convert DArTag file to read count file*

Description

Convert DArTag file to read count file

Usage

```
dart_tag(
  dart.file,
  first.data.row = 9,
  first.data.col = 6,
  out.file,
  map.file = NULL,
  AB.file = NULL
)
```

Arguments

dart.file	DArTag CSV filename
first.data.row	first data row in dart.file
first.data.col	first data column in dart.file
out.file	name of output file
map.file	optional CSV file (marker, chrom, position) to integrate into the output
AB.file	optional CSV file (marker, REF) to convert to allele B dosage

Details

Designed for standard two-row format from DArT. Column 1 contains the AlleleID in format MarkerName|Haplotype. Haplotypes are named Ref,RefMatch,Alt,AltMatch,Other. Counts are combined for Ref + RefMatch, as well as Alt + AltMatch. Other haplotypes are discarded.

Output format contains the read counts for each marker x id combination in the format "count1|count2" (which is the input format for PolyOrigin).

Use `AB.file` to convert REF/ALT counts from dart.file to A/B counts from SNP array. The file must have columns named "marker" and "REF", where REF is either A or B. Only markers present in AB.file will be in the output.

Value

marker x indiv matrix of read depths

`gbs_call` *Genotype calls from allele count file*

Description

Genotype calls from allele count file

Usage

```
gbs_call(in.file, out.base, ploidy, n.core = 1)
```

Arguments

`in.file`	input file
`out.base`	base name for output files
`ploidy`	ploidy
`n.core`	number of cores

Details

Function uses R/updog package (Gerard et al. 2018), which is based on a binomial model. If a sample has no reads at a marker, the genotype is NA and posterior probability equals 0.

The input file follows the format generated by `dart_tag`, with the read counts for each marker x id combination represented as "count1|count2".

Three output files are generated with the (1) allele ratio, (2) posterior mode, and (3) posterior mean.

Value

marker x indiv matrix of read depths

`geno_call` *Genotype calls*

Description

Genotype calls based on a normal mixture model

Usage

```
geno_call(
  data,
  filename,
  model.ploidy = 4L,
  sample.ploidy = 4L,
  min.posterior = 0,
  transform = TRUE
)
```

Arguments

`data`	matrix (markers x id) of input values for the normal mixture model
`filename`	CSV filename with the model parameters
`model.ploidy`	2 or 4 (default)
`sample.ploidy`	2 or 4 (default)
`min.posterior`	minimum posterior probability (default 0) for genotype call
`transform`	TRUE (default) or FALSE whether to apply arcsin square root transformation

Details

The first column of the CSV input file should be the SNP ID, followed by columns for the normal distribution means, standard deviations, and mixture probabilities. Genotype calls are based on the maximum a posteriori (MAP) method. If the posterior probability of the MAP genotype is less than `min.posterior`, then `NA` is returned for that sample. By default, an arcsin square root transformation is applied to the input values to match the approach used by R package fitPoly. To use a tetraploid mixture model for diploid samples, set `sample.ploidy = 2` and `model.ploidy = 4`.

Value

matrix of allele dosages (0,1,2,..ploidy) with dimensions markers x individuals

get_pedigree *Generate pedigree*

Description

Generate pedigree for a set of individuals

Usage

```
get_pedigree(id, pedfile, delim = ",", na.string = "NA", trim = TRUE)
```

Arguments

id	Vector of names of individuals
pedfile	Name of pedigree file
delim	Delimiter for the pedigree file (default is "," for CSV)
na.string	String used for NA in the pedigree file (default is "NA")
trim	TRUE/FALSE whether to trim pedigree (see Details)

Details

Finds ancestors of individuals in a three-column pedigree file (id,mother,father). The id column can be the identifier for an individual or cross. String matches must be exact or based on the naming convention crossID-progenyID. The returned pedigree is ordered using R package `pedigree` so that offspring follow parents. When `trim` is TRUE (default), the pedigree is trimmed to remove ancestors with only one offspring (which are not needed to compute the pedigree relationship matrix).

Value

Data frame with columns id, mother, father

GvsA *Plot G vs. A*

Description

Plot marker-based vs. pedigree-based additive relationship coefficients

Usage

```
GvsA(
  parentage,
  G,
  A,
  filename = NULL,
  thresh.G = Inf,
  thresh.A = 0.5,
  Gmax = NULL,
  Amax = NULL
)
```

Arguments

parentage	Data frame of individuals to plot, with 3 columns: id,mother,father
G	Genomic relationship matrix
A	Pedigree relationship matrix
filename	Name of PDF file to save the results (optional for one individual)
thresh.G	Threshold above which names are displayed (default Inf)
thresh.A	Threshold above which names are displayed (default 0.5)
Gmax	Upper limit for y-axis for plotting. If NULL, maximum value in G is used.
Amax	Upper limit for x-axis for plotting. If NULL, maximum value in A is used.

Details

Useful for finding and correcting pedigree errors. If the G or A coefficient for an individual exceeds the threshold, its name is displayed in the figure. If parentage contains one individual, by default a ggplot2 variable will be returned, but the result can also be written to file. If multiple individuals are present, a filename is required.

G_mat *Additive genomic relationships*

Description

Relationship matrix for additive effects with bi-allelic markers

Usage

```
G_mat(geno, ploidy)
```

Arguments

geno	Matrix of allele dosages (markers x indiv)
ploidy	Any even integer (2,4,6,...)

Details

Additive effects are based on the traditional orthogonal decomposition of genetic variance in panmictic populations (Fisher 1918; Kempthorne 1957; Endelman et al. 2018). Missing genotype data is replaced with the population mean.

Value

G matrix

References

Fisher (1918) Trans. Roy. Soc. Edin. 52:399-433.

Kempthorne (1957) An Introduction to Genetic Statistics.

Endelman et al. (2018) Genetics 209:77-87.

impute *Impute missing marker data*

Description

Impute marker data based on the population mean or mode

Usage

```
impute(geno, method)
```

Arguments

geno	Matrix of allele dosages with dimensions markers x indiv
method	Either "mean" or "mode"

Details

Missing values are imputed with either the population mean or mode (most frequent value) for each marker

Value

Imputed genotype matrix (markers x indiv)

merge_impute *Merge two genotype matrices and impute missing data*

Description

Merge two genotype matrices and impute missing data by BLUP

Usage

```
merge_impute(geno1, geno2, ploidy)
```

Arguments

geno1	Genotype matrix (coded 0...ploidy) with dimensions markers x indiv
geno2	Genotype matrix (coded 0...ploidy) with dimensions markers x indiv
ploidy	Either 2 or 4

Details

Designed to impute from low to high density markers. The BLUP method is equivalent to Eq. 4 of Poland et al. (2012), but this function is not iterative. Additional shrinkage toward the mean is applied if needed to keep the imputed values within the range [0,ploidy]. Missing data in the input matrices are imputed with the population mean for each marker. If an individual appears in both input matrices, it is renamed with suffixes ".1" and ".2" and treated as two different individuals. Monomorphic markers are removed.

Value

Imputed genotype matrix (markers x indiv)

References

Poland et al. (2012) Plant Genome 5:103-113.

`readXY` *Read SNP array intensity data*

Description

Read SNP array intensity data

Usage

```
readXY(filename, skip = 9, output = "ratio")
```

Arguments

`filename`	filename
`skip`	number of lines to skip before the header line with the column names
`output`	One of three options: "ratio","theta","AD"

Details

The first two columns of the tab-delimited input file should be the SNP and Sample ID. Columns labeled "X" and "Y" contain the signal intensities for the two alleles. Use `output` to specify whether to return the ratio = Y/(X+Y) or theta = atan(Y/X)*2/pi. Option "AD" exports the XY data in the allele depth format for a VCF file ("X,Y"), with the X and Y values multiplied by 100 and rounded to the nearest integer.

Value

matrix with dimensions markers x individuals

`update_alias` *Update names based on alias*

Description

Update names based on data frame with alias and preferred name

Usage

```
update_alias(x, alias, remove.space = TRUE, filename = NULL)
```

Arguments

x	Vector of names to update
alias	Data frame with two columns: first is the preferred name and second is the alias
remove.space	TRUE/FALSE
filename	update names in file (variable "id")

Details

Parameter remove.space indicates whether blank spaces should be removed before string matching

Value

Vector with updated names

Index