Package ‘ohtadstats’

November 15, 2019

Version 2.1.1

Date 2019-10-12

Title Tomoka Ohta D Statistics

Description Calculate's Tomoka Ohta's partitioning of linkage disequilibrium, deemed D-statistics, for pairs of loci. Petrowski et al. (2019) <doi:10.5334/jors.250>.

Author Paul F. Petrowski <pfpetrowski@gmail.com> & Timothy M. Beissinger <timbeissinger@gmail.com>

Maintainer Paul F. Petrowski <pfpetrowski@gmail.com>

Depends R (>= 3.0.0)

Imports lattice, grDevices, stats, utils

License MIT + file LICENSE

URL https://github.com/pfpetrowski/OhtaDStats

RoxygenNote 6.1.1

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2019-11-15 11:20:02 UTC

R topics documented:

`beissinger_data` *Chicken Genotype Data*

Description

This file is a matrix of genotypes from 96 chickens encompassing 5 breeds, genotyped as part of the Synbreed Project. Individuals are in rows. Marker genotypes are in columns, coded as 0, 1, and 2. Row names are a breed index so all rows named "1" are from breed 1, all rows named "2" are from breed 2, and so on. Column names are marker names. These data are a subset of the data used by Beissinger et al. (2016). The full dataset is hosted on Figshare at the which can be at the link below.

Usage

```
data(beissinger_data)
```

Format

A matrix with 1417 rows and 100 columns.

Source

(https://figshare.com/articles/Synbreed_Biodiversity_Panel_Genotypes/1497961)

References

Beissinger et al. (2016) Heredity. (https://www.nature.com/articles/hdy201581)

`dfilter` *Filtering datasets for subpopulations with low sample sizes*

Description

Simplifies the process of eliminating subpopulations with low sample sizes.

Usage

```
dfilter(data, minsample)
```

Arguments

`data`	Matrix containing genotype data with individuals as rows and loci as columns. Genotypes should be coded as 0 (homozygous), 1 (heterozygous), or 2 (homozygous). Rownames must be subpopulation names and column names should be marker names.
`minsample`	An integer representing the smallest number of individuals a subpopulation must contain to be included in analysis.

Value

filtered_data The original dataset minus the subpopulations that fail to meet the sample size threshold.

Examples

```
test <- matrix(round(runif(400,1,2)), nrow = 100)
rownames(test) <- c(rep(c('A','B','C'),each=25), rep(c('D','E'), each=5), rep('F', 15))
dim(test)

#The 'D' and 'E' subpopulations have only five members each and should be removed
filtered_test <- dfilter(test,12)

dim(filtered_test) # New dataset is reduced by 10 rows (five for 'D' and five for 'E')
```

---

dheatmap *Heatmap Plot*

Description

Plots a matrix of D statistics, output from dwrapper, as a heatmap.

Usage

```
dheatmap(d_matrix, colors = c("white", "lightblue", "blue", "darkblue",
  "black"), mode = "linear", tick.labels = TRUE, nbins = 5)
```

Arguments

d_matrix	A matrix of D statistics or a matrix of D statistic ratios.
colors	An optional color vector. Optionally modify the color scheme of the heatmap. If mode = 'binned', must be of length 5.
mode	A string indicating desired coloring scheme. The option "linear" scales colors linearly, "truncated" truncates values greater than 1, and "binned" returns a discretedistribution of colors.
tick.labels	A logical indicating whether or not marker labels should be drawn.
nbins	An integer specifying the number of bins to be used. Only relevent if mode is "binned".

Details

The d_matrix input should be one of the matrices output by dwrapper. Options are d2it_mat, d2is_mat, d2st_mat, dp2st_mat, dp2is_mat, npops_mat, ratio1, and ratio2. More customized plots can be developed using the "levelplot" package.

Value

A color plot

Examples

```
data(miyashita_langley_data)
miyashita_langley_subset <- miyashita_langley_data[,1:15]
ml_results <- dwrapper(miyashita_langley_subset)
dheatmap(ml_results[["d2it_mat"]], mode = 'linear')

## Not run:
data(miyashita_langley_data)
ml_results <- dwrapper(miyashita_langley_data)
dheatmap(ml_results[["d2it_mat"]], mode = 'linear')

## End(Not run)
```

`dparallel` *Compute Ohta's D Statistics in a manner optimized for parallelization*

Description

Infers the comparisons that this instance of the function is supposed to perform given job_id and comparisons_per_job. Returns the results of those comparisons to an SQL database.

Usage

```
dparallel(data_set, tot_maf = 0.1, pop_maf = 0.05, comparisons_per_job,
  job_id, outfile = "Ohta")
```

Arguments

`data_set`	The data set that is to be analysed.
`tot_maf`	Minimum minor allele frequency across the total population for a marker to be included in the analysis.
`pop_maf`	Minimum minor allele frequency across a subpopulation for that subpopulation to be included in analysis.
`comparisons_per_job`	The number of comparisons that each instance of dparallel will compute.
`job_id`	A number indicating that this is the nth instance of this function.
`outfile`	Prefix for the file name that results will be written to. May be a path. Do not include extension.

Examples

```
data(beissinger_data)
dparallel(data_set = beissinger_data,
                    comparisons_per_job = 300,
                    job_id = 1,
               outfile = tempfile(pattern = "beissinger_comparison", tmpdir = tempdir()))
```

dstat *Tomoka Ohta's D Statistics*

Description

Implements Ohta's D statistics for a pair of loci. Statistics are returned in a vector in the following order: Number of populations, D2it, D2is, D2st, D'2st, D'2is.

Usage

```
dstat(index, data_set, tot_maf = 0.1, pop_maf = 0.05)
```

Arguments

`index`	A two-element vector of column names or numbers for which Ohta's D Statistics will be computed.
`data_set`	Matrix containing genotype data with individuals as rows and loci as columns. Genotypes should be coded as 0 (homozygous), 1 (heterozygous), or 2 (homozygous). Rownames must be subpopulation names and column names should be marker names.
`tot_maf`	Minimum minor allele frequency across the total population for a marker to be included in the analysis.
`pop_maf`	Minimum minor allele frequency across a subpopulation for that subpopulation to be included in analysis.

Details

When the loci being evaluated fail to pass the filtering thresholds determined by tot_maf and pop_maf, NAs are returned.

Value

nPops Number of subpopulations used for computation, after filtering.

D2it A measure of the correlation of alleles at two loci on the same gametes in a subpopulation relative to their expectation according to allele frequencies in the total population.

D2is Expected variance of LD for subpopulations.

D2st Expected correlation of alleles in a subpopulation relative to their expected correlation in the total population.

Dp2st Variance of LD for the total population computed over alleles only.

Dp2is Correlation of alleles at two loci on the same gamete in subpopulations relative to their expected correlation in the total population.

References

Beissinger et al. (2016) Heredity. (https://www.nature.com/articles/hdy201581) & Ohta. (1982) Proc. Natl. Acad. Science. (http://www.pnas.org/content/79/6/1940)

Examples

```
data(beissinger_data)
dstat(index = c(5,6), data_set = beissinger_data)
```

dwrapper *Ohta D Statistic Wrapper*

Description

Pairwise computation of Ohta's D Statistics for each pair of polymorphisms in a given dataset.

Usage

```
dwrapper(data_set, tot_maf = 0.1, pop_maf = 0.05)
```

Arguments

`data_set`	Matrix containing genotype data with individuals as rows and loci as columns. Genotypes should be coded as 0 (homozygous), 1 (heterozygous), or 2 (homozygous). Rownames must be subpopulation names and column names should be marker names.
`tot_maf`	Minimum minor allele frequency across the total population for a marker to be included in the analysis.
`pop_maf`	Minimum minor allele frequency across a subpopulation for that subpopulation to be included in analysis.

Details

This wrapper implements the dstat function for all pairs of loci in a genotype matrix. If the input matrix includes n loci, choose(n,2) pairs are evaluated. Therefore, the computaiton time scales quadratically, and is not feasible for large datasets. We suggest manual parallelization across computational nodes for a large-scale (ie thousands of markers) implementation.

Value

A list of matrices containing the pairwise comparisons for each D statistic. Also included is the number of subpopulations evaluated in each comparison and the ratio of d2is_mat to d2st_mat (ratio1) and dp2st_mat to dp2is_mat (ratio2). The result of a comparison between marker M and marker N will be found in the Mth row at the Nth column.

Examples

```
data(beissinger_data)
beissinger_subset <- beissinger_data[,1:15]
dwrapper(beissinger_subset, tot_maf = 0.05, pop_maf = 0.01)

## Not run:
data(beissinger_data)
dwrapper(beissinger_data, tot_maf = 0.05, pop_maf = 0.01)

## End(Not run)
```

miyashita_langley_data

Drosophila melanogaster genotypes

Description

Genotype data obtained from Miyashita & Langley (1988). A matrix representing 85 loci in 64 individuals. Individuals are in rows. Rownames "RL", "TX", or "FK", indicate the subpopulation from which the sample was taken.

Usage

```
data(miyashita_langley_data)
```

Format

A matrix with 64 rows and 85 columns.

References

Miyashita & Langley (1988) Genetics 120:199-212 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1203490/)

Index