Package ‘DEMEtics’

October 19, 2018

Type Package

Title Evaluating The Genetic Differentiation Between Populations Based on Gst and D Values

Version 0.8-8

Date 2018-02-03

Author Alexander Jueterbock, Philipp Kraemer, Gabriele Gerlach and Jana Deppermann

Maintainer Alexander Jueterbock <Alexander-Jueterbock@web.de>

Depends R (>= 2.8.0)

Description Allows to calculate the fixation index Gst
(Nei, 1973) and the differentiation index D (Jost, 2008) pairwise
between or averaged over several populations. P-values, stating the
significance of differentiation, and 95 percent confidence intervals
can be estimated using bootstrap resamplings. In the case that more
than two populations are compared pairwise, the p-values are
adjusted by bonferroni correction and in several other ways due to
the multiple comparison from one data set.

License GPL (>= 2)

RoxygenNote 6.0.1

R topics documented:

DEMEtics-package *Evaluating The Genetic Differentiation Between Populations Based on Gst and D Values*

Description

Allows to calculate the fixation index Gst (Nei, 1973) and the differentiation index D (Jost, 2008) pairwise between or averaged over several populations. P-values, stating the significance of differentiation, and 95 percent confidence intervals can be estimated using bootstrap resamplings. In the case that more than two populations are compared pairwise, the p-values are adjusted by bonferroni correction and in several other ways due to the multiple comparison from one data set.

Details

Package:	DEMEtics
Type:	Package
Version:	0.8-8
Date:	2018-02-02
Depends:	R (>= 2.8.0)
License:	GPL (>= 2)
Literature:	Gerlach G., Jueterbock A., Kraemer P., Deppermann J. and Harmand P. 2010 Calculations of population differentiation based on Gst and D: forget Gst but not all of statistics! *Molecular Ecology* **19**, p. 3845–3852. Goudet J., Raymond M., deMeeues T. and Rousset F. 1996 Testing differentiation in diploid populations. *Genetics* **144**, 4, p. 1933–1940. Jost, L. 2008 Gst and its relatives do not measure differentiation. *Molecular Ecology* **17**, 18, p. 4015–4026. Manly, B.F.J. 1997 *Randomization, bootstrap and Monte Carlo methods in biology* Chapman & Hall. Nei, M. 1973 Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 12, p. 3321–3323. Nei M., Chesser R. 1983 Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**, 253–259. Wright, S.P. 1992

	Adjusted p-values for simultaneous inference. *Biometrics* **48**, 1005–1013.
LazyLoad:	yes
Packaged:	2010-12-18 13:48:33 UTC; alexj
Built:	R 2.10.1; ; 2010-12-18 13:48:39 UTC; unix

Index:

Citation

To cite the package 'DEMEtics' in publications use:

Gerlach, G., Jueterbock, A., Kraemer, P., Deppermann, J. and Harmand, P. 2010 Calculations of population differentiation based on G(ST) and D: forget G(ST) but not all of statistics! *Molecular Ecology* **19**, 3845-3852

Author(s)

Alexander Jueterbock, Philipp Kraemer, Gabriele Gerlach and Jana Deppermann

Maintainer: Alexander Jueterbock <Alexander-Jueterbock@web.de>

`D.Jost-and-Gst.Nei` *Comparing Populations - Differentiation and Fixation Indices*

Description

The degree of genetic differentiation between populations is often measured by the fixation index Gst (Nei, 1973). However, differentiation at polymorphic loci with more than 2 alleles is much better reflected by the D value (Jost, 2008; Gerlach et al., 2010). The functions of this package allow to estimate locus by locus (and averaged over loci) pairwise Gst and D values for codominant markers between populations and their averages over all populations. P-values (indicating the strength of evidence against the null hypothesis of no genetic differentiation) and 95% confidence limits are obtained from bootstrap methods. Depending on whether or not all populations are in Hardy Weinberg Equilibrium for a given locus, either alleles or genotypes are randomized over populations, respectively (see Goudet, 1996).

Usage

```
D.Jost(filename, bias = "correct", object = FALSE, format.table = TRUE,
pm = "pairwise", statistics = "CI", bt = 1000)
Gst.Nei(filename, bias = "correct", object = FALSE, format.table = TRUE,
pm = "pairwise", statistics = "CI", bt = 1000)
```

Arguments

filename	Its syntax depends on the setting of the argument object. If object=FALSE (default), the filename has to be a combination of (1) the name of the data file (.txt format) in which the raw data are saved and (2) the extension .txt. It has to be enclosed in quotes (“filename.txt”). If object=TRUE, the filename has to be the name of the object under which the data table was assigned to the R workspace, enclosed in quotes (“filename”).
bias	An argument providing two options (correct (default) and uncorrected). When using the correct option, Hs and Ht are transformed into nearly unbiased estimators Hs.est and Ht.est derived by Nei & Chesser (1993), thus reducing the bias of D and Gst values (now named Dest and Gst.est) as they are estimated from a small sample that intends to represent the whole population (see Jost, 2008, p. 4022).
object	This argument can be set as TRUE or FALSE, depending on the format of the argument filename.
format.table	A logical argument either set as TRUE (default) or FALSE that defines if the format of the table has to be transformed before analysis (see details).
pm	A two-level argument providing the opportunity to compare populations pairwise (pm="pairwise", default) or otherwise to average the D or Gst values over all populations (pm="overall").
statistics	A four-level argument to select whether no statistics (statistics="none"), 95% confidence intervals (statistics="CI"), p-values (statistics= "p", testing against the null hypothesis of no genetic differentiation) or both, CI intervals and p-values (statistics="all") shall be provided. Be aware, that the bootstrapping to obtain these statistics can take long, especially when both shall be evaluated.
bt	A numeric argument (default=1000) that defines the amount of bootstrap resamplings, that the p-values and/or the 95% confidence intervals are based on.

Details

The input format

The input data can be of two different formats. Both of them should be tab-delimited. The information that has to be provided are names or numbers for each individual, the according population they were sampled from and the alleles (length in base pairs, rounded) at each locus. Two alleles have to be defined for each diploid individual. Haplotype data can not be evaluated with this package. Missing alleles have to be set to zero (possible: 0, 00, 000).

The data table that has to be transformed by choosing `format.table=TRUE`, can be provided in the following format:

individual	population	locus1.allele.a	locus1.allele.b	locus2.allele.a	locus2.allele.b
P1.1	P1	175	183	110	110
P1.2	P1	183	183	123	126
P2.1	P2	230	225	110	110
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

The number of populations and loci are not restricted. The column names `individual` and `population` must be included. The other columns listing the fragment lengths in base pairs can be named arbitrarily. It is recommended name the two columns that refer to the same locus, equally (e.g. `locus1.allele.a` and `locus1.allele.b` should both be named `Locus1`). Mathematical signs, like `+` or `-` should be avoided and spaces are not allowed in column names.

Alternatively, when the input data are given in the following format, they do not have to be transformed (`format.table=FALSE`):

individual	population	fragment.length	locus
P1.1	P1	175	L1
P1.1	P1	183	L1
P1.2	P1	183	L1
P1.2	P1	183	L1
P2.1	P2	230	L1
P2.1	P2	225	L1
.	.	.	.
.	.	.	.
.	.	.	.
P1.1	P1	110	L2
P1.1	P1	110	L2
P1.2	P1	123	L2
P1.2	P1	126	L2
P2.1	P2	110	L2
P2.1	P2	110	L2
.	.	.	.
.	.	.	.
.	.	.	.

The data in the column `fragment.length` represent numbers of base pairs.

Details on confidence interval calculation

95% confidence intervals of the D or Gst values are based on the range of these values from re-allocated data sets that are obtained by bootstrapping alleles (or genotypes) of one locus within populations. Hardy Weinberg Equilibrium (HWE) is tested for each locus and each population. If all of the tested populations are in HWE, the alleles of a single locus, are randomized within

populations. Otherwise, alleles are not inherited independently from each other and genotypes are randomized within populations (Goudet, 1996). The upper and lower 95% confidence limits are evaluated as the lower (0.025) and upper (0.975) bounds of the quantiles of D or Gst values from the resampled data using the function `quantile`:
Empirical D or Gst +(-) upper(lower) quantile bound

Details on p-value calculation

To be able to test the null hypothesis of absence of genetic differentiation between populations, a bootstrap method is performed. Thereby, alleles (or genotypes) of one locus are randomized over all compared populations. Hardy Weinberg Equilibrium HWE is tested for each locus and each population. If all of the tested populations are in HWE, the alleles of a single locus, are randomized over all populations. Otherwise, alleles are not inherited independently from each other and genotypes are randomized over all populations (Goudet, 1996). Reallocating alleles or genotypes simulates populations that share a common gene pool and are not differentiated. Since the empirical value of genetic differentiation is expected to be larger than a value obtained from within a panmictic population when the tested populations are significantly differentiated, a one tailed test is carried out. The null hypothesis (panmictic populations) can be rejected at a 95% significance level (p<0.05) when the empirical value is larger than 95% of the bootstrapped test statistics. The p-value is calculated according to Manly (1997, p. 62).

When more than two populations are compared with one another, using the option `pm="pairwise"`, the p-values are adjusted in order to account for the multiple comparison from one data set, using the function `p.adjust` of the package `stats`. They represent the smallest overall significance levels, at which the hypothesis would be rejected (Wright, 1992). Those p-values giving the significance levels for different loci, are adjusted independently from each other. Those p-values giving the significance levels for the averaged differentiation over all loci, are adjusted to one another. The adjustment is performed by Bonferroni correction, by Holm's method, by Hommel's method and by a method provided by Benjamini and Hochberg. See the help file of the function `p.adjust` for further information on these methods.

Test for Hardy Weinberg Equilibrium HWE

Before bootstrapping, populations are automatically tested for being in HWE by comparing the empirical numbers of genotypes and those expected under HWE using the function `chisq.test` with the arguments: `simulate.p.value=TRUE`, `b=10000`. This means, that the p-value is obtained from a Monte Carlo method with 10000-fold resampling. The null hypothesis of HWE is rejected when p is smaller than 0.05.

Value

Results are saved as .txt files (space-delimited) in the actual working directory, which is normally the one your input data were loaded from. The path of the working directory can be requested by typing `getwd()` and changed by using the function `setwd()`. During the calculation, the output is printed in the R console where the kind of data is also shortly described and how the respective .txt files are named. The filenames include the argument `filename` and the actual date.

In case that you are comparing more than two populations pairwise and are calculating p-values and/or confidence intervals, you will be informed about the estimated end of the analysis after completion of the first pairwise comparison.

If the same analysis is carried out more than once at the same day on a single dataset, the results will all be found, one written below the other, separated by a row of column names, in the same file (if the working directory was not changed).

The output files are described in the following paragraphs:

`allelefrequencies`
A data table comprising the following columns:

- `allele` a factor with each fragment length of an allele representing one level
- `number` a numeric vector listing how often the actual allele of the actual locus occurred in the actual population
- `population` a factor with each population representing one level
- `locus` a factor with each locus representing one level
- `proportion` a numeric vector giving the proportion of the actual allele and the actual locus in each population

`sample sizes` A data table comprising the following columns:

- `population` a factor with each population representing one level
- `sample.size` a numeric vector listing the number of individuals providing data for the actual locus
- `locus` a factor with each locus name representing one level

`heterozygosities`
A data table that lists heterozygosites which are calculated according to the formulas given in Jost (2008).

- `locus` a factor with each locus name representing one level
- `Hs` a numeric vector providing the mean non-bias corrected heterozygosities for each locus within the populations
- `Hs.est` a numeric vector providing the bias corrected heterozygosities for each locus within the populations
- `Ht` a numeric vector providing the non-bias corrected total heterozygosities for each locus over all populations
- `Ht.est` a numeric vector providing the bias corrected total heterozygosities for each locus over all populations

Depending on whether populations are compared pairwise `pm="pairwise"` or differentiation / fixation is estimated over all populations `pm="overall"`, the result tables comprising the D/Gst values differ slightly. When overall D or Gst values are evaluated, the output comprises the following two data tables (X stands for D, Dest, Gst or Gst.est values):

`X.loci.over.all.populations`

- `X.locus` A numeric vector providing the D, Dest, Gst or Gst.est values for each locus and pairwise comparison
- `Locus` A factor with the locus names as levels, sorted alphabethically or numerically (depending on how the loci are named)
- (`Lower.0.95.CI`) A numerical vector giving the lower 95% confidence limit for the empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting `statistics="CI"` or `statistics="all"`

(`Upper.0.95.CI`) A numerical vector giving the upper 95% confidence limit for the empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting `statistics="CI"` or `statistics="all"`

(`P.value`) A numerical vector comprising the p-values that state the level of significance of genetic differentiation. This output is only obtained when setting `statistics="p"` or `statistics="all"`

`X.mean.over.all.populations`

`X.mean` The mean D, Dest, Gst or Gst.est value averaged over all populations and loci (by calculating the arithmetic mean)

(`Lower.0.95.CI`) The lower 95% confidence limit for the empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting `statistics= "CI"` or `statistics="all"`

(`Upper.0.95.CI`) The upper 95% confidence limit for the empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting `statistics= "CI"` or `statistics="all"`

(`P.value`) The p-value that states the level of significance of genetic differentiation. This output is only obtained when setting `statistics="p"` or `statistics="all"`

When populations are compared pairwise, INTERMEDIATE RESULTS are printed and saved after each comparison. automatically. The next INTERMEDIATE RESULT is printed to the same file, separated from the preceding result by a row of column names. When the whole analysis is completed, the END RESULT containing the information of all the INTERMEDIATE RESULTs in a single data frame is printed and saved to the same file, separated from the preceding INTERMEDIATE RESULTs by a row of column names. Appending the results one below the other avoids loss of data. But you have to be careful. If you want to work with the INTERMEDIATE RESULTs that have already been saved, it is recommended to copy the respective file and work with the copy. Otherwise, problems can arise, when you work with the original file and R tries to write new results to it. This could cause interruption of the analysis.

If an analysis is carried out more than once at the same day, the results will all be found, one written below the other, separated by a row of column names in the same file (if the working directory was not changed).

If an analysis runs more than one day, the INTERMEDIATE RESULTs will be saved in different files, according to the date, they had been analysed on. But all the INTERMEDIATE RESULTs will be included in the END RESULT in which all INTERMEDIATE RESULTs are finally saved together.

The output comprises data tables with the following information (X stands for D, Dest, Gst or Gst.est values):

`X.loci.pairwise.comparison`

`X.locus` A numeric vector providing the D, Dest, Gst or Gst.est values for each locus and pairwise comparison

`Locus` A factor with the locus names as levels, sorted alphabethically or numerically (depending on how the loci are named)

`Population1` A factor listing the first of the pairwise compared populations

`Population2` A factor listing the second of the pairwise compared populations

(Lower.0.95.CI) A numerical vector giving the lower 95% confidence limit for the empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting statistics="CI" or statistics="all"

(Upper.0.95.CI) A numerical vector giving the upper 95% confidence limit for the empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting statistics="CI" or statistics="all"

(P.value) A numerical vector comprising the p-values that state the level of significance of genetic differentiation. This output is only obtained when setting statistics="p" or statistics="all"

X.mean.pairwise.comparison

X.mean A numerical vector giving mean D, Dest, Gst or Gst.est values for the respective populationpair averaged over all loci (by calculating the arithmetic mean)

Population1 A factor listing the first of the pairwise compared populations

Population2 A factor listing the second of the pairwise compared populations

(Lower.0.95.CI) A numerical vector listing the lower 95% confidence limit for the respective empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting statistics="CI" or statistics="all"

(Upper.0.95.CI) A numerical vector listing the upper 95% confidence limit for the respective empirical D, Dest, Gst or Gst.est value. This output is only obtained when setting statistics="CI" or statistics="all"

(P.value) A numerical vector containing the p-values that state the level of significance of genetic differentiation for the respective pairwise comparison. This output is only obtained when setting statistics="p" or statistics="all"

When you choose the option format.table=TRUE, a data file called “Output-Inputformat.txt” is created that is needed by the functions of this package to analyze the data.

Warning

Depending on the size of your data set and the performance of your computer, the bootstrapping process for calculating p-values and confidence intervals, can take very long so that you might want to run the analysis over night.

When you carry out pairwise population comparisons, you will be informed after evaluation of the data for the first population pair, when the whole analysis is estimated to finish.

Author(s)

Alexander Jueterbock, <Alexander-Jueterbock@web.de>
Philipp Kraemer, <philipp.kraemer@mail.uni-oldenburg.de>

References

Gerlach G., Jueterbock A., Kraemer P., Deppermann J. and Harmand P. 2010
Calculations of population differentiation based on Gst and D:
forget Gst but not all of statistics!
Molecular Ecology **19**, p. 3845–3852.

Goudet J., Raymond M., deMeeues T. and Rousset F. 1996
Testing differentiation in diploid populations.
Genetics **144**, 4, p. 1933–1940.

Jost, L. 2008
Gst and its relatives do not measure differentiation.
Molecular Ecology **17**, 18, p. 4015–4026.

Manly, B.F.J. 1997
Randomization, bootstrap and Monte Carlo methods in biology
Chapman & Hall.

Nei, M. 1973
Analysis of gene diversity in subdivided populations.
Proceedings of the National Academy of Sciences of the United States of America
70, 12, p. 3321–3323.

Nei M., Chesser R. 1983
Estimation of fixation indices and gene diversities.
Annals of Human Genetics **47**, 253–259.

Wright, S.P. 1992
Adjusted p-values for simultaneous inference.
Biometrics **48**, 1005–1013.

Examples

```
# loading data from the example files of this package

data(Example.transformed)
Example.t <- Example.transformed

data(Example.untransformed)
Example.u <- Example.untransformed

# Calculating mean Dest values (averaged over all populations) with
# p-values and confidence intervals using only 10 bootstrap resamplings

D.Jost("Example.t", bias="correct", object=TRUE, format.table=FALSE,
pm="overall", statistics="all", bt=10)

# Calculating pairwise Gst values without any statistics

Gst.Nei("Example.u", bias="uncorrected", object=TRUE, format.table=TRUE,
pm="pairwise", statistics="none")

# If you do not know where the results of these example tables have been
# saved, type getwd()
```

`Example.transformed` *Allelic Data of Three Populations for Three Loci*

Description

The data set gives the fragment lengths (in base pairs) for four loci and three populations.

Usage

```
data(Example.transformed)
```

Format

A data frame with 76 observations on the following 5 variables.

`individual` a factor with each individual representing one level

`population` a factor with levels `population1`, `population2`, `population3`

`allele` a factor with the two alleles of one individual for one locus as levels

`fragment.length` a numeric vector listing the base pairs for the actual locus and allele

`locus` a factor with the loci names as levels

Examples

```
data(Example.transformed)
Example.transformed
```

`Example.untransformed` *Allelic Data of Three Populations for Three Loci*

Description

The data set gives the fragment lengths (in base pairs) for three loci and three populations.

Usage

```
data(Example.untransformed)
```

Format

A data frame with 216 observations on the following 8 variables.

`individual` a factor with each individual representing one level

`population` a factor with levels `population1`, `population2`, `population3`

`locus1.allele1` a numeric vector listing the base pairs of the first allele of locus1

`locus1.allele2` a numeric vector listing the base pairs of the second allele of locus1

`locus2.allele1` a numeric vector listing the base pairs of the first allele of locus2

`locus2.allele2` a numeric vector listing the base pairs of the second allele of locus2

`locus3.allele1` a numeric vector listing the base pairs of the first allele of locus3

`locus3.allele2` a numeric vector listing the base pairs of the second allele of locus3

Examples

```
data(Example.untransformed)
Example.untransformed
```

Index