Package 'apparent'

March 22, 2019

Version 1.1

Date 2019-03-18

Title Accurate Parentage Analysis in the Absence of Guiding Information

Author Arthur Melo, Iago Hale

Maintainer Arthur Melo <arthurmelobio@gmail.com>

Description Performs parentage analysis based on a test of genetic identity between expected progeny (EPij), built using Single Nucleotide Polymorphism (SNP) homozygous loci from all pairs of possible parents (i and j), and all potential offspring (POk). Using the Gower Dissimilarity metric (GD), genetic identity between EPij and POk is taken as evidence that individuals i and j are the true parents of offspring k. Evaluation of triad (two parents + offspring) significance is based on the distribution of all GD (EPij|k) values. Specifically, a Dixon test is used to identify a gap-based threshold that separates true triads and from spurious associations. For any offspring not successfully assigned to a pair of parents, perhaps due to the absence of one parent from the test population, a non-mandatory Dyad analysis can be employed to identify a likely single parent for a given offspring. In this analysis, a two-stage test is applied to discriminate an offspring's true parent from its other close relatives (e.g. siblings) that may also be present in the population. In the first stage, 'apparent' calculates the mean GD (GDM) between a POk and all expected progeny arising from the j possible triads involving potential parent i. In the second stage, it calculates a coefficient of variation (GDCV) among the pairwise GD's between POk and each expected progeny arising from the j triads involving potential parent i. An individual that is simultaneously a low outlier in the first test and a high outlier in the second is identified as a likely parent of POk. In an effort to facilitate interpretation, results of both the triad and optional dyad analyses are presented in tabular and graphical form.

Depends R (>= 3.0.2), outliers

License GPL (>= 2)

LazyData true

NeedsCompilation no

Repository CRAN

RoxygenNote 6.0.1

Date/Publication 2019-03-22 13:33:30 UTC

R topics documented:

`apparent` *Accurate parentage analysis in the absence of guiding information*

Description

Performs parentage analysis based on a test of genetic identity between expected progeny (EPij), built using Single Nucleotide Polymorphism (SNP) homozygous loci from all pairs of possible parents (i and j), and all potential offspring (POk). Using the Gower Dissimilarity metric (GD), genetic identity between EPij and POk is taken as evidence that individuals i and j are the true parents of offspring k. Evaluation of triad (two parents + offspring) significance is based on the distribution of all GD (EPij|POk) values. Specifically, a Dixon test is used to identify a gap-based threshold that separates true triads and from spurious associations. For any offspring not successfully assigned to a pair of parents, perhaps due to the absence of one parent from the test population, a non-mandatory Dyad analysis can be employed to identify a likely single parent for a given offspring. In this analysis, a two-stage test is applied to discriminate an offspring's true parent from its other close relatives (e.g. siblings) that may also be present in the population. In the first stage, 'apparent' calculates the mean GD (GDM) between a POk and all expected progeny arising from the j possible triads involving potential parent i. In the second stage, it calculates a coefficient of variation (GDCV) among the pairwise GD's between POk and each expected progeny arising from the j triads involving potential parent i. An individual that is simultaneously a low outlier in the first test and a high outlier in the second is identified as a likely parent of POk. In an effort to facilitate interpretation, results of both the triad and optional dyad analyses are presented in tabular and graphical form.

Usage

```
apparent(InputFile, MaxIdent=0.10, alpha=0.01, nloci=300,
self=TRUE, plot=TRUE, Dyad=FALSE)
```

Arguments

`InputFile` A tab-delimited text file with n rows (n = number of individuals in the population) and m+2 columns (m = number of SNP loci). Column one contains the ID's of the individuals in the population. Column two contains a classification key which assigns each individual to one of five possible classes from analysis (see below). The third and all subsequent columns contain genotype calls, with one SNP per column and the alleles separated by "/". A missing genotype is represented by "-/-". An example input file with 5 individuals and 5 SNP loci is shown below:

Geno1 All A/A A/T C/A C/C T/C
Geno2 Pa T/A A/A -/- C/C T/T

Geno3 Mo A/A A/A C/A C/G T/T
Geno4 Fa T/T A/T A/A G/G -/-
Geno5 Off A/A T/T C/C C/C T/C

The keys values allowed in the second column are:
All: "All" - The individual will be tested as a potential Mother, Father, and Offspring.
Pa: "Parent" - The individual will be tested as potential Mother and Father.
Mo: "Mother" - The individual will be tested only as a potential Mother.
Fa: "Father" - The individual will be tested only as a potential Father.
Off: "Offspring" - The individual will be tested only as a potential Offspring.

`MaxIdent` Sets the maximum triad GDij|k to be considered for outlier significance testing. This parameter directly impacts computation time. By default, MaxIdent is set to 0.1.

`alpha` The alpha level for all significance testing (triad and dyad analyses).

`nloci` The minimum acceptable number loci to be used when computing the pairwise GDij|k. The default value of 300 is suggested, based on previous investigations. All triads for which the number of usable SNPs falls below nloci will be excluded from the analysis.

`self` Logical value for instructing 'apparent' whether or not to consider self-crossing (parent i = parent j). The default value is TRUE.

`plot` Logical value for create the plots of both Triad and Dyad analysis. By default, 'apparent' creates only the Triad analysis plot. The default value is TRUE, meaning 'apparent' creates only the Triad analysis plot. The Dyad analysis plot can't be printed, only saved (it is a pdf file with multiple plots for each significant offspring). To save Dyad analysis plot, make sure plot, files and Dyad flags are TRUE.

`Dyad` Logical value for instructing 'apparent' to perform a Dyad analysis, following the Triad analysis. The default value is FALSE.

Details

See Melo ATO and Hale I (2019) 'apparent': a simple and flexible R package for accurate SNP-based parentage analysis in the absence of guiding information. BMC Bioinformatics. doi: doi.org/10.1186/s12859-019-2662-3. Also, find useful information at https://github.com/halelab/apparent.

Value

Returns a list object with multiple results; i.e., four outputs from the mandatory Triad analysis and the result of the Dyad analysis, if the analysis was done. The list of results are: Triad_all: contains the full table of results from the triad analysis, including the cross type (self or out-cross), the number of usable SNPs, and the GDij|POk, of all tested triads. Triad_sig: reports only those triads found to be significant. Triad_summary_pop: useful summary statistics from the Triad analysis. For the population as a whole: Overall mean GDij|POk and its standard deviation, as well as overall mean number of usable SNPs and its standard deviation. Triad_summary_geno: useful summary statistics from the Triad analysis. For each individual genotype in the analysis: Mean GDij|POk, GDij|POk range, and mean usable number of SNPs for all comparisons involving that genotype.

Genotypes exhibiting a mean GDij|POk or number of usable SNPs less than 2 SD's below the population means are considered outliers. Dyad_Sig: reports only those parent-offspring dyads found to be significant. Only if the Dyad analysis was done. Along with the data frame outputs, if plot is TRUE, apparent will create both the Triad and Dyad analysis plots. The Triad analysis plot has the distribution of GDij|POk values, annotated with the gap-based threshold that separates true triads from spurious associations. The plot is useful in interpretating the results of the triad analysis , the full details of which are found in Triad_all Triad_sig.txt. The Dyad analysis plot (only if plot=TRUE and Dyad=TRUE) is a two-paneled figure showing the distributions of GDM and GDCV values upon which the dyad analysis is based. These plots are useful in interpretating the results of the dyad analysis, the full details of which are found in Dyad_sig.

Note

As the size of the exploratory triad space inflates (i.e. as the population size grows), the likelihood of low-probability instances of low GD values, occurring purely by chance, increases. Such outlier values can "bleed" into the gap between true and spurious triads, in effect making the gap disappear and leading to a null result, even if true triads exist in the population (Type II error). Such a dilution of a true gap is generally due to the presence of one or more problematic genotypes in the population, problematic due to the fact that the available number of loci for their analysis is much for some reason lower than for other individuals, leading to artificially low GDij|POk values. To address such problems the following two "best practices" are recommended to users: 1. Always include at least one known (true) triad in the analysis, as a reference/control. 2. After running the triad analysis, inspect the summary file "apparent-Triad-Summary.cvs" to see if any individuals exhibit unusually low mean GDij|k values or mean number of loci usable for analysis. If such individuals exist, remove them and re-run the analysis.

Author(s)

Arthur T. O. Melo and Iago Hale Department of Agriculture Nutrition and Food Systems, University of New Hampshire, Durham, NH, USA.

References

Dixon W.J. (1950) Analysis of extreme values. Ann. Math. Stat. 21(4):488-506. Dixon W.J. (1951) Ratios involving extreme values. Ann. Math. Stat. 22(1):68-78. Gower JC. (1971) A General Coefficient of Similarity and Some of Its Properties. Biometrics. 27(4):857-71.

See Also

outlier

Examples

```
# Load the input file
InputFile <- apparent_TestData
# Run the apparent
apparentOUT <- apparent(InputFile)
# Check the Triad analysis output
apparentOUT$Triad_all
apparentOUT$Triad_sig
```

`apparent_TestData` *Example of input file*

Description

An example of input file for apparent.

Format

a data frame with the following columns:

- Name: Genotype name
- Key: Key ID for analysis
- Loci: The third and all subsequent columns contain genotype calls, with one SNP per column and the alleles separated by "/"

Index