Package 'MHCtools'

September 16, 2020

Type Package

Title Analysis of MHC Data in Non-Model Species

Version 1.3.0

Description Ten tools for bioinformatical processing and analysis of major histocompatibility complex (MHC) data. The functions are tailored for amplicon data sets that have been filtered using the 'dada2' method (for more information on 'dada2', visit <https://benjjneb.github.io/dada2/>), but even other types of data sets can be analyzed.
The DistCalc() function calculates Grantham, Sandberg, or p-distances from pairwise comparisons of all sequences in a data set, and mean distances of all pairwise comparisons within each sample in a data set. The function additionally outputs five tables with physico-chemical z-descriptor values (based on Sandberg et al. 1998) for each amino acid position in all sequences in the data set. These tables may be useful for further downstream analyses, such as estimation of MHC supertypes.
The HpltFind() function infers putative haplotypes from families in the data set.
The GetHpltTable() and GetHpltStats() functions evaluate the accuracy of the haplotype inference.
The PapaDiv() function compares parent pairs in the data set and calculate their joint MHC diversity, taking into account sequence variants that occur in both parents.
The ReplMatch() function matches replicates in data sets in order to evaluate genotyping success.
The GetReplTable() and GetReplStats() functions perform such an evaluation.
The CreateFas() function creates a fasta file with all the sequences in the data set.
The CreateSamplesFas() function creates individual fasta files for each sample in the data set.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports stats, utils

RoxygenNote 7.1.1

NeedsCompilation no

Author Jacob Roved [aut, cre]

Maintainer Jacob Roved <jacob.roved@biol.lu.se>

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2020-09-16 08:10:15 UTC

R topics documented:

`CreateFas` *CreateFas() function*

Description

`CreateFas` creates a FASTA file with all the sequences in a 'dada2' sequence table.

Usage

```
CreateFas(seq_table, path_out)
```

Arguments

`seq_table`	seq_table is a sequence table as output by the 'dada2' pipeline, which has samples in rows and nucleotide sequence variants in columns.
`path_out`	is a user defined path to the folder where the output files will be saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A FASTA file with all the sequences in a 'dada2' sequence table. The sequences are named in the FASTA file by an index number corresponding to their column number in the sequence table.

See Also

CreateSamplesFas; for more information about 'dada2' visit <https://benjjneb.github.io/dada2/>

Examples

```
seq_table <- sequence_table_fas
path_out <- tempdir()
CreateFas(seq_table, path_out)
```

CreateSamplesFas *CreateSamplesFas() function*

Description

CreateSamplesFas creates a set of FASTA files with the sequences present in each sample in a 'dada2' sequence table.

Usage

```
CreateSamplesFas(seq_table, path_out)
```

Arguments

seq_table	seq_table is a sequence table as output by the 'dada2' pipeline, which has samples in rows and nucleotide sequence variants in columns.
path_out	is a user defined path to the folder where the output files will be saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A set of FASTA files with the sequences present in each sample in the sequence table.The sequences are named in the FASTA files by an index number corresponding to their column number in the sequence table, thus identical sequences will have identical sample names in all the FASTA files.

See Also

CreateFas; for more information about 'dada2' visit <https://benjjneb.github.io/dada2/>

Examples

```
seq_table <- sequence_table_fas
path_out <- tempdir()
CreateSamplesFas(seq_table, path_out)
```

DistCalc *DistCalc() function*

Description

DistCalc calculates Grantham distances, Sandberg distances, or p-distances from pairwise comparisons of aligned sequences.

Usage

```
DistCalc(
  seq_file,
  path_out,
  input_fasta = NULL,
  input_seq = "aa",
  aa_dist = NULL,
  codon_pos = NULL,
  dist_type = "G"
)
```

Arguments

seq_file	is a sequence occurrence table as output by the 'dada2' pipeline, which has samples in rows and nucleotide sequence variants in columns. Optionally, a fasta file can be supplied as input in the format rendered by read.fasta() from the package 'seqinr'.
path_out	is a user defined path to the folder where the output files will be saved.
input_fasta	optional, a logical (TRUE/FALSE) that indicates whether the input file is a fasta file (TRUE) or a 'dada2'-style sequence table (NULL/FALSE). The default is NULL/FALSE.
input_seq	defines the type of sequences in seq_file. It may take the values 'nucl' or 'aa'.

`aa_dist` is optional, a logical (TRUE/FALSE) that determines whether nucleotide sequences should be translated to amino acid sequences before distance calculation, default is NULL/FALSE. Note that aa_dist must be set to TRUE, if Grantham or Sandberg distances are calculated from an alignment of nucleotide sequences.

`codon_pos` is optional, a vector of comma separated integers specifying which codon positions to include in distance calculations. If omitted, distance calculations are made using all codons.

`dist_type` is used to specify which kind of distances that are calculated. It takes the values 'G' for Grantham distances, 'S' for Sandberg distances, or 'P' for p-distances. The argument is optional with 'G' as default setting.

Details

The DistCalc() function takes a fasta file or a 'dada2'-style sequence occurrence table (with aligned sequences as column names and samples in rows) as input and produces a matrix with pairwise distances for all sequences in the data set. If calculation of Sandberg distances is specified, the function additionally outputs five tables with physico-chemical z-descriptor values (based on Sandberg et al. 1998) for each amino acid position in all sequences in the data set. These tables may be useful for further downstream analyses, such as estimation of MHC supertypes. If a sequence occurrence table is provided as input, the DistCalc() function furthermore produces a table with the mean distances from all pairwise comparisons of the sequences in each sample in the data set.

Grantham distances and Sandberg distances are calculated as described in Pierini & Lenz 2018. The Grantham distances produced by DistCalc() are simply the mean Grantham distances (Grantham 1974) between all amino acid codons in sequence pairs. When calculating Sandberg distances, DistCalc() first computes Euclidian distances between all amino acid pairs based on the five physico-chemical z-descriptors defined in Sandberg et al. 1998. Sandberg distances are then calculated as the mean Euclidian distances between all amino acid codons in sequence pairs. P-distances calculated by DistCalc() are simply the proportion of varying codons between pairs of sequences.

The DistCalc() function includes an option for the user to specify which codons to compare, which is useful e.g. if conducting the analysis only on codon positions involved in specific functions, such as peptide binding of an MHC molecule. It also accepts calculating amino acid distances directly from protein-coding DNA sequences using the standard genetic code.

The DistCalc() function accepts the following characters in the sequences: Nucleotide sequences: A,T,G,C Amino acid sequences: A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V

It accepts gaps defined by '-'. Nucleotide triplets containing gaps are translated to 'X', if amino acid distances are calculated directly from DNA nucleotide sequences. Please note that '-' or 'X' are treated as unique characters in p-distance calculations. The function will not accept 'X' or gaps in Grantham or Sandberg distance calculations. If you wish to exclude codons with 'X' or gaps from distance calculations, please use the codon_pos option to specify which codons to compare.

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

If you calculated Grantham or Sandberg distances, please additionally cite: Pierini, F., Lenz, T.L. 2018. Divergent allele advantage at human MHC genes: Signatures of past and ongoing selection. Mol. Biol. Evol. 35, 2145–2158.

...and either of the following references: Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S. 1998. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. JMed Chem. 41(14):2481–2491.

Value

The function returns a matrix with distances from all pairwise sequence comparisons, where n is the number of sequences. If a sequence occurrence table is given as input file, the function additionally returns a table with the mean distance for each sample in the data set. If a sequence occurrence table is given as input file, the sequences are named in the output matrix by an index number that corresponds to their column number in the input file. If calculation of Sandberg distances is specified, the function additionally outputs five tables with physico-chemical z-descriptor values for each amino acid position in all sequences in the data set. All output tables are saved as .csv files in the output path.

See Also

For more information about 'dada2', visit <https://benjjneb.github.io/dada2/>

Examples

```
seq_file <- sequence_table_fas
path_out <- tempdir()
DistCalc(seq_file, path_out, input_fasta=NULL, input_seq="nucl", aa_dist=NULL,
codon_pos=c(1,2,3,4,5,6,7,8), dist_type="P")
```

`GetHpltStats` *GetHpltStats() function*

Description

`GetHpltStats` uses the output files produced by the HpltFind() function to calculate the mean of the mean proportion of incongruent sequences across all nests in the data set.

Usage

```
GetHpltStats(filepath)
```

Arguments

`filepath`	is a user defined path to the folder where the output files from the HpltFind() function have been saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A mean of the mean proportion of incongruent sequences for each nest.

See Also

HpltFind; GetHpltTable

Examples

```
filepath <- system.file("extdata/HpltFindOut/", package="MHCtools")
GetHpltStats(filepath)
```

GetHpltTable *GetHpltTable() function*

Description

GetHpltTable uses the output files produced by the HpltFind() function to produce a table with the mean proportion of incongruent sequences for each nest. If the mean proportion of incongruent sequences is generally low, but certain nests have many incongruent sequences, biological reasons may be causing the mismatches, e.g. extra-pair fertilizations or recombination events.

Usage

```
GetHpltTable(filepath)
```

Arguments

filepath	is a user defined path to the folder where the output files from the HpltFind() function have been saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A table with the mean proportion of incongruent sequences for each nest.

See Also

HpltFind; GetHpltStats

Examples

```
filepath <- system.file("extdata/HpltFindOut/", package="MHCtools")
GetHpltTable(filepath)
```

GetReplStats *GetReplStats function*

Description

GetReplStats uses the output files produced by the ReplMatch() function to calculate statistics on the agreement between replicated samples in the sequencing experiment.

Usage

```
GetReplStats(filepath)
```

Arguments

filepath	is a user defined path to the folder where the output files from the ReplMatch() function have been saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A list containing the number of replicate sets with zero incongruent sequences, the proportion of replicate sets with zero incongruent sequences, the mean of the mean proportion of incongruent sequences across all replicate sets, and the repeatability of the sequencing experiment.

See Also

ReplMatch; GetReplTable

Examples

```
filepath <- system.file("extdata/ReplMatchOut/", package="MHCtools")
GetReplStats(filepath)
```

`GetReplTable` *GetReplTable function*

Description

`GetReplTable` uses the output files produced by the ReplMatch() function to produce a table with the replicate sets and their respective mean proportion of incongruent sequences.

Usage

```
GetReplTable(filepath)
```

Arguments

`filepath`	is a user defined path to the folder where the output files from the ReplMatch() function have been saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A table with the mean proportion of incongruent sequences for each replicate set.

See Also

`ReplMatch`; `GetReplStats`

Examples

```
filepath <- system.file("extdata/ReplMatchOut/", package="MHCtools")
GetReplTable(filepath)
```

HpltFind *HpltFind() function*

Description

HpltFind is designed to automatically infer major histocompatibility complex (MHC) haplotypes from the genotypes of parents and offspring in families (defined as nests) in non-model species, where MHC sequence variants cannot be identified as belonging to individual loci. The functions GetHpltTable() and GetHpltStats() are designed to evaluate the output files.

Usage

```
HpltFind(nest_table, seq_table, path_out)
```

Arguments

nest_table	is a table containing the sample names of parents and offspring in each nest. This table should be organized so that the individual names are in the first column (Sample_ID), and the nest number is in the second column (Nest). For each nest, the first two rows should be the parents, followed immediately by the offspring in the subsequent rows, and then followed by the next nest, and so on. It is assumed that nests are numbered consecutively beginning at 1.
seq_table	seq_table is a sequence table as output by the 'dada2' pipeline, which has samples in rows and nucleotide sequence variants in columns.
path_out	is a user defined path to the folder where the output files will be saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A set of R lists containing for each nest the putative haplotypes, the names of sequences that could not be resolved with certainty in each parent, the names of the sequences that were incongruent in the genotypes of the nest, and the mean proportion of incongruent sequences (which is a measure of the haplotype inference success and largely influenced by the exactness of the genotyping experiment). The sequences are named in the output by an index number corresponding to their column number in the sequence table, thus identical sequences will have identical sample names in all the output files. These files can be reopened in R e.g. using the readRDS() function in the base package.

See Also

GetHpltTable; GetHpltStats; for more information about 'dada2' visit <https://benjjneb.github.io/dada2/>

Examples

```
nest_table <- nest_table
seq_table <- sequence_table
path_out <- tempdir()
HpltFind(nest_table, seq_table, path_out)
```

`nest_table` *Data nest_table*

Description

nest_table, parents_table, and sequence_table comprise a randomized dataset derived from a real major histocompatibility complex (MHC) genotyping experiment with data from parents and offspring. The nucleotide sequences have been replaced with randomly generated sequences, and sample names have been anonymized.

Usage

```
nest_table
```

Format

nest_table is a data frame with 213 samples in rows and 2 columns:

Sample_ID Sample ID

Nest Nest index number

Source

original data.

`PapaDiv` *PapaDiv() function*

Description

`PapaDiv` calculates the joint major histocompatibility complex (MHC) diversity in parent pairs, taking into account alleles that are shared between the parents. The joint diversity in parent pairs is useful for heritability analyses in non-model species, where one wants to estimate the heritability of MHC diversity. The number of unique alleles in offspring may not be directly derived from the parental genotypes if some alleles are shared between the parents.

Usage

```
PapaDiv(parents_table, seq_table, path_out)
```

Arguments

parents_table	is a table containing the sample names of the parents in each nest. This table should be organized so that each row represents one nest, with the individual names of the mothers in the first column (Mother), and the individual names of the fathers in the second column (Father).
seq_table	seq_table is a sequence table as output by the 'dada2' pipeline, which has samples in rows and nucleotide sequence variants in columns.
path_out	is a user defined path to the folder where the output files will be saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

a set of R lists containing for the joint diversity of each parent pair, the proportion of sequences that are shared between the parents, the diversity of each of the parents, the observed sequence variants in each parent, the matched sequence variants, and the incongruent sequence variants in each parent. The sequences are named in the output by an index number corresponding to their column number in the sequence table, thus identical sequences will have identical sample names in all the output files. These files are saved in a sub folder in the output path called Parent_pairs (created by PapaDiv()) and can be reopened in R e.g. using the readRDS() function in the base package. For downstream data analysis, the PapaDiv() function also produces a summary table with the names of the parents in a pair, their respective MHC diversities, and the joint parent pair diversity. This table is saved as a .csv file in the output path.

See Also

For more information about 'dada2' visit <https://benjjneb.github.io/dada2/>

Examples

```
parents_table <- parents_table
seq_table <- sequence_table
path_out <- tempdir()
PapaDiv(parents_table, seq_table, path_out)
```

parents_table *Data parents_table*

Description

nest_table, parents_table, and sequence_table comprise a randomized dataset derived from a real major histocompatibility complex (MHC) genotyping experiment with data from parents and offspring. The nucleotide sequences have been replaced with randomly generated sequences, and sample names have been anonymized.

Usage

```
parents_table
```

Format

parents_table is a data frame with 57 parent pairs in rows and 2 columns:

Mother Mother ID

Father Father ID

Source

original data.

replicates_table *Data replicates_table*

Description

replicates_table and sequence_table_repl comprise a randomized dataset derived from a real major histocompatibility complex (MHC) genotyping experiment with technical replicates. The nucleotide sequences have been replaced with randomly generated sequences, and sample names have been anonymized.

Usage

```
replicates_table
```

Format

replicates_table is a data frame with 111 technical replicate samples in rows and 2 columns:

Sample_ID Technical replicate sample ID

Replic_set Index number of replicate set

Source

original data.

`ReplMatch` *ReplMatch() function*

Description

In amplicon filtering it is sometimes valuable to compare technical replicates in order to estimate the accuracy of a genotyping experiment. This may be done both to optimize filtering settings and to estimate repeatability to report in a publication. `ReplMatch` is designed to automatically compare technical replicates in an amplicon filtering data set and report the proportion of mismatches. The functions GetReplTable() and GetReplStats() are designed to evaluate the output files.

Usage

```
ReplMatch(repl_table, seq_table, path_out)
```

Arguments

`repl_table`	is a table containing the sample names of technical replicates in the data set. This table should be organized so that the individual names are in the first column (Sample_ID), and the index number of the replicate set is in the second column (Replic_set). Replicate sets are allowed to contain more than two replicates. It is assumed that replicate sets are numbered consecutively beginning at 1.
`seq_table`	seq_table is a sequence table as output by the 'dada2' pipeline, which has samples in rows and nucleotide sequence variants in columns.
`path_out`	is a user defined path to the folder where the output files will be saved.

Details

If you publish data produced with MHCtools, please cite: Roved, J. 2020. MHCtools: Analysis of MHC data in non-model species. Cran. Roved, J., Hansson, B., Stervander, M., Hasselquist, D., Westerdahl, H. 2020. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. bioRxiv.

Value

A set of R lists containing for each replicate set the observed sequence variants, the names of the sequences that were incongruent in the replicates, and the mean proportion of incongruent sequences (if 100 matches are expected between the replicates, this is equivalent of an error rate in the sequencing process). The sequences are named in the output by an index number corresponding to their column number in the sequence table, thus identical sequences will have identical sample names in all the output files. These files can be reopened in R e.g. using the readRDS() function in the base package.

See Also

GetReplTable; GetReplStats; for more information about 'dada2' visit <https://benjjneb.github.io/dada2/>

Examples

```
repl_table <- replicates_table
seq_table <- sequence_table_repl
path_out <- tempdir()
ReplMatch(repl_table, seq_table, path_out)
```

---

sequence_table *Data sequence_table*

Description

nest_table, parents_table, and sequence_table comprise a randomized dataset derived from a real major histocompatibility complex (MHC) genotyping experiment with data from parents and offspring. The nucleotide sequences have been replaced with randomly generated sequences, and sample names have been anonymized.

Usage

```
sequence_table
```

Format

sequence_table is a data frame with 334 samples in rows and 329 DNA sequence variants in columns.

Source

original data.

---

sequence_table_fas *Data sequence_table_fas*

Description

sequence_table_fas is a randomized dataset derived from a real major histocompatibility complex (MHC) genotyping experiment. The nucleotide sequences have been replaced with randomly generated sequences, and sample names have been anonymized.

Usage

```
sequence_table_fas
```

Format

sequence_table_fas is a data frame with 100 samples in rows and 166 DNA sequence variants in columns.

Source

original data.

`sequence_table_repl` *Data sequence_table_repl*

Description

replicates_table and sequence_table_repl comprise a randomized dataset derived from a real major histocompatibility complex (MHC) genotyping experiment with technical replicates. The nucleotide sequences have been replaced with randomly generated sequences, and sample names have been anonymized.

Usage

```
sequence_table_repl
```

Format

sequence_table_repl is a data frame with 412 samples in rows and 511 DNA sequence variants in columns.

Source

original data.

Index