Package ‘perfectphyloR’

March 8, 2021

**Type** Package

**Title** Reconstruct Perfect Phylogenies from DNA Sequence Data

**Version** 0.2.1

**Date** 2021-02-28

**Author** Charith Karunarathna and Jinko Graham

**Maintainer** Charith Karunarathna <ckarunar@sfu.ca>

**Description** Reconstructs perfect phylogeny at a user-given focal point and to depict and test association in a genomic region based on the reconstructed partitions. Charith B Karunarathna and Jinko Graham (2019) <bioRxiv:10.1101/674523>.

**Depends** R (>= 3.4.0)

**License** GNU General Public License

**Imports** ape, phytools, Rcpp (>= 0.12.16)

**LinkingTo** Rcpp, RcppArmadillo

**RoxygenNote** 7.1.1

**Suggests** HHG, dendextend, vcfR, R.rsp

**VignetteBuilder** R.rsp

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2021-03-08 05:30:02 UTC

**R topics documented:**

perfectphyloR-package .................................................. 2
createHapMat ......................................................... 2
dCorTest ............................................................. 3
ex_hapMatSmall_data .................................................. 4
ex_hapMat_data ........................................................ 5
HHGtest ............................................................... 5
MantelTest ............................................................ 6
phenoDist .............................................................. 7
Description

Functions to reconstruct perfect phylogeny underlying a sample of DNA sequences, at a focal single-nucleotide variant (SNV) and to depict and test association in a genomic region based on the reconstructed partitions.

Author(s)

Charith Karunarathna and Jinko Graham

createHapMat

Create an object of class hapMat

Description

This function creates a hapMat data object, a required input for reconstructPP.

Usage

createHapMat(hapmat, snvNames, hapNames, posns)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hapmat</td>
<td>A matrix of 0’s and 1’s, with rows representing haplotypes and columns represent-</td>
</tr>
<tr>
<td></td>
<td>ing single-nucleotide variants (SNVs).</td>
</tr>
<tr>
<td>snvNames</td>
<td>A vector of names of SNVs for the columns of hapmat.</td>
</tr>
<tr>
<td>hapNames</td>
<td>A vector of names of haplotypes for the rows of hapmat.</td>
</tr>
<tr>
<td>posns</td>
<td>A numeric vector specifying the genomic positions (e.g. in base pairs) of</td>
</tr>
<tr>
<td></td>
<td>SNVs in the columns of hapmat.</td>
</tr>
</tbody>
</table>
**dCorTest**

**Value**

An object of class `hapMat`.

**Examples**

```r
dcorTest
```

```r
hapmat = matrix(c(1,1,1,0,
0,0,0,0,
1,1,1,1,
1,0,0,0,
1,1,0,0,
1,0,0,1,
1,0,0,1), byrow = TRUE, ncol = 4)
snvnames = c(paste("SNV", 1:4, sep = ""))
allhaps = c("h1", "h2", "h3", "h4", "h5", "h6", "h7")
# Physical positions
posns = c(1000, 2000, 3000, 4000)

# Create hapMat data object
ex_hapMat <- createHapMat(hapmat = hapmat,
snvNames = snvnames,
hapNames = allhaps,
posns = posns)
```

**Description**

This function performs dCor test for association between two distance matrices and computes permutation P value. Permutation P value is computed by randomly permuting rows and columns of the second distance matrix.

**Usage**

```r
dCorTest(Dx, Dy, nperm)
```

**Arguments**

- **Dx**
  A numeric matrix of pairwise distances.

- **Dy**
  A second numeric matrix of pairwise distances.

- **nperm**
  The number of times to permute the rows and columns of Dy.

**Value**

A list contains RV coefficient and permutation P value.
References


Examples

```r
x <- runif(8)
y <- runif(8)
# Distance matrices
distX = as.matrix(dist(x, upper = TRUE, diag = TRUE))
distY = as.matrix(dist(y, upper = TRUE, diag = TRUE))

dCorTest(Dx = distX, Dy = distY, nperm = 1000)
```

---

ex_hapMatSmall_data  Example small dataset

Description

A subset of ex_hapMat_data, containing 10 sequences (haplotypes) with 20 SNVs.

Usage

data(ex_hapMatSmall_data)

Format

A list of ten haplotypes with the physical positions of each SNV.

- **hapmat** A matrix of 0’s and 1’s, with rows representing haplotypes and columns representing SNVs.
- **snvNames** A vector of names of SNVs for the columns of hapmat.
- **hapNames** A vector of names of haplotypes for the rows of hapmat.
- **posns** a numeric vector specifying the genomic positions (e.g. in base pairs) of SNVs in the columns of hapmat.
ex_hapMat_data

---

**Example dataset**

**Description**

A hapMat data object containing 200 sequences (haplotypes) with 2747 SNVs.

**Usage**

```r
data(ex_hapMat_data)
```

**Format**

A list of 200 haplotypes with the physical positions of each SNV.

- **hapmat**: A matrix of 0’s and 1’s, with rows representing haplotypes and columns representing SNVs.
- **snvNames**: A vector of names of SNVs for the columns of hapmat.
- **hapNames**: A vector of names of haplotypes for the rows of hapmat.
- **posns**: A numeric vector specifying the genomic positions (e.g., in base pairs) of SNVs in the columns of hapmat.

---

**HHGtest**

**HHG test for association of two distance matrices**

**Description**

This function performs HHG test to find the association between two distance matrices. It permutes rows and columns of the second matrix randomly to calculate P value.

**Usage**

```r
HHGtest(Dx, Dy, nperm)
```

**Arguments**

- **Dx**: A numeric matrix of pairwise distances.
- **Dy**: A second numeric matrix of pairwise distances.
- **nperm**: The number of times to permute the rows and columns of Dy.

**Value**

A list contains HHG coefficient and permutation P value.
References


Examples

```r
x <- runif(8)
y <- runif(8)
# Distance matrices
distX = as.matrix(dist(x, upper = TRUE, diag = TRUE))
distY = as.matrix(dist(y, upper = TRUE, diag = TRUE))

HHGtest(Dx = distX, Dy = distY, nperm = 1000)
```

### MantelTest

**Mantel test for association of two distance matrices**

#### Description

This function performs Mantel test for correlation between two distance matrices. It computes P value by randomly permuting rows and columns of the second matrix.

#### Usage

```r
MantelTest(Dx, Dy, nperm)
```

#### Arguments

- **Dx**
  A numeric matrix of pairwise distances.
- **Dy**
  A second numeric matrix of pairwise distances.
- **nperm**
  The number of times to permute the rows and columns of Dy.

#### Value

A list contains Mantel statistic and permutation P value.

#### References

**Examples**

```r
x <- runif(8)
y <- runif(8)
# Distance matrices
distX = as.matrix(dist(x, upper = TRUE, diag = TRUE))
distY = as.matrix(dist(y, upper = TRUE, diag = TRUE))

MantelTest(Dx = distX, Dy = distY, nperm = 1000)
```

---

**PhenoDist**

**Phenotypic distances**

**Description**

This is the pairwise phenotypic distances described in Karunarathna and Graham (2018).

**Usage**

```r
data(phenoDist)
```

**Format**

An object of class `matrix`.

**References**


---

**PlotDend**

**Plot reconstructed dendrogram**

**Description**

This function plots reconstructed dendrogram in a genomic region.

**Usage**

```r
plotDend(dend, direction = "downwards")
```
RandIndexTest

Arguments

- **dend**
  An object of class *phylo* or of class *multiPhylo* returned from *reconstructPP* or *reconstructPPregion*.

- **direction**
  A character string specifying the direction of the dendrogram. Four values are possible: "downwards" (the default), "upwards", "leftwards" and "rightwards".

Examples

```r
data(ex_hapMat_data)
ex_dend <- reconstructPP(hapMat = ex_hapMat_data, focalSNV = 3, minWindow = 1, sep = "-")
plotDend(dend = ex_dend, direction = "downwards")
```

RandIndexTest

Rand Index Test

Description

This function performs Rand index test for association between two *phylo* objects.

Usage

```r
RandIndexTest(dend1, dend2, k = 2, nperm)
```

Arguments

- **dend1**
  An object of type *phylo*.

- **dend2**
  A second object of type *phylo*.

- **k**
  An integer that specifies the number of clusters that the dendrogram should be cut into. The default is $k = 2$. Clusters are defined by starting from the root of the dendrogram and cutting across.

- **nperm**
  The number of times to permute tips of the *dend2*.

Value

A numeric value between 0 and 1 and permutation P value.

References

Examples

```r
data(ex_hapMat_data)
d1 <- reconstructPP(ex_hapMat_data, focalSNV = 1, minWindow = 1)
d2 <- reconstructPP(ex_hapMat_data, focalSNV = 5, minWindow = 1)
RandIndexTest(dend1 = d1, dend2 = d2, k = 5, nperm = 100)
```

**rdistMatrix**

*Rank-based distances between haplotypes in a given partition*

**Description**

This function computes the pairwise distances between haplotypes (tips) of the dendrogram based on the ranking of the nested partitions in the dendrogram. See the details.

**Usage**

```
rdistMatrix(dend, sep = "-")
```

**Arguments**

- `dend` A list of nodes that represents the nested partition of haplotypes.
- `sep` A character string separator for concatenating haplotype labels in the dendrogram if they are undistinguishable in the window around the focal SNV. See the arguments in `reconstructPP`.

**Details**

We code the distance between two haplotypes of a dendrogram as the number of inner nodes that separate the haplotypes plus one. That is, we assign the distance between two internal neighbouring nodes as one, and the distance between an internal node and its neighbouring tip as one. To illustrate, consider the following figure of a dendrogram. In the figure, the distance between the haplotypes 2931 and 454 is 3; the distance between other haplotypes are given in the table below.
Value

A matrix of pairwise distances between haplotypes.

Examples

```r
data(ex_hapMat_data)
rdend <- reconstructPP(hapMat = ex_hapMat_data, focalSNV = 2, minWindow = 1, sep = "-" )
rdistMatrix(rdend)
```

---

### `reconstructPP`

*Reconstruct the perfect phylogeny at a given focal SNV*

**Description**

This function reconstructs the perfect phylogeny at a given focal SNV using the recursive partitioning algorithm of Gusfield (1991) on compatible SNVs, and the modification of Mailund et al. (2006) to include incompatible SNVs that are nearby.

**Usage**

```r
reconstructPP(hapMat, focalSNV, minWindow = 1, sep = "-")
```

**Arguments**

- `hapMat` A data structure of class `hapMat`. Eg: created by the `createHapMat` function.
- `focalSNV` The column number of the focal SNV at which to reconstruct the reconstructed partitions.
**reconstructPP**

- **minWindow**: Minimum number of SNVs around the focal SNV in the window of SNVs used to reconstruct the partitions (default is the maximum of one and 2% of the total number of the SNVs).

- **sep**: Character string separator to separate haplotype names for haplotypes that cannot be distinguished in the window around the focal point. For example, if a tip is comprised of haplotypes "h1" and "h3", and sep = "-", then the tip label will be "h1-h3". The default value is "-". See details.

**Details**

To reconstruct the perfect phylogeny from sequence data, these two steps are followed: (1) Select a window of SNVs at a given focal SNV. (2) Build the perfect phylogey for the window of SNVs. More details can be found in the references.

The following figure shows the reconstructed partitions at the tenth SNV position of ex_hapMatSmall_data.

![Reconstructed partitions](image)

**Value**

An object of class `phylo` with indices of the column boundaries of the `hapMat` object that were used to reconstruct the partition in the window of SNVs.

**References**


**Examples**

```r
data(ex_hapMatSmall_data)
rdend <- reconstructPP(hapMat = ex_hapMatSmall_data, focalSNV = 10, minWindow = 1, sep = "-")
```
# Plot the reconstructed perfect phylogeny.
plotDend(rdend, direction = "down")

# Extract the positions of the lower and upper limits of a window of SNVs in hapMat object # to reconstruct the partition, rdend.
ex_hapMatSmall_data$posns[rdend$snvWinIndices]

reconstructPPregion  
*Reconstruct perfect phylogeny sequence across a region*

**Description**
This function reconstructs perfect phylogenies on each possible focal SNV across a genomic region.

**Usage**

```r
reconstructPPregion(hapMat, minWindow, posn.lb = NULL, posn.ub = NULL)
```

**Arguments**

- `hapMat`: A data structure of class `hapMat`. See the arguments in `reconstructPP`.
- `minWindow`: Minimum number of SNVs around the focal SNV in the window of SNVs used to reconstruct the perfect phylogeny.
- `posn.lb`: Lower bound of the subregion of `hapMat` (in base pairs) within which to consider SNVs.
- `posn.ub`: Upper bound of the subregion of `hapMat` (in base pairs) within which to consider SNVs.

**Value**
An object of class `multiPhylo` that contains multiple `phylo` objects.

**Examples**

```r
data(ex_hapMatSmall_data)

# Reconstruct partitions across the region of ex_hapMatSmall_data.
rdends <- reconstructPPregion(hapMat = ex_hapMatSmall_data, minWindow = 1)

# Reconstruct partitions between a given range SNV positions.
rdends_range <- reconstructPPregion(hapMat = ex_hapMatSmall_data, minWindow = 1, posn.lb = 2000, posn.ub = 7000)
```
RVtest

RV test for association of two distance matrices

Description

This function performs RV test for similarity of two distance matrices. It permutes rows and columns of the second matrix randomly to calculate P value.

Usage

RVtest(Dx, Dy, nperm)

Arguments

Dx A numeric matrix of pairwise distances.
Dy A second numeric matrix of pairwise distances.
nperm The number of times to permute the rows and columns of Dy.

Value

A list contains RV coefficient and permutation P value.

References


Examples

x <- runif(8)
y <- runif(8)
# Distance matrices
distX = as.matrix(dist(x, upper = TRUE, diag = TRUE))
distY = as.matrix(dist(y, upper = TRUE, diag = TRUE))

RVtest(Dx = distX, Dy = distY, nperm = 1000)
**tdend**  
*True dendrogram object*

**Description**

A phylo object containing attributes of the comparator true dendrogram for the example data at SNV position 975 kilo base pairs.

**Usage**

```r
data(tdend)
```

**Format**

A phylo object from the ape package containing four attributes:

- **edge** A matrix containing the node labels and their child nodes.
- **Nnode** The number of nodes.
- **tip.label** A character vector containing the haplotype labels of the true dendrogram.
- **edge.length** A numeric vector giving the lengths of the branches given by edge.

---

**testAssoDist**  
*Test the association between a comparator distance matrix, and the reconstructed dendrograms across a genomic region*

**Description**

This function calculates and tests the association between a comparator distance matrix, based on any pairwise distance measure, and the reconstructed dendrograms across a genomic region of interest using association measures such as the dCor statistic, HHG statistic, Mantel statistic, and RV coefficient. See the section Applications in vignette("perfectphyloR") for the detailed example.

**Usage**

```r
testAssoDist(rdend, cdmat, method, hapMat, nperm = 0, xlab = "", ylab = "", main = "")
```
**Arguments**

rdend  
A multiPhylo object of reconstructed dendrograms at each focal SNV.

cdmat  
A comparator matrix of pairwise distances (e.g. pairwise distances between haplotypes of a comparator dendrogram).

method  
Association measures. Use "dCor" for dCor test, "HHG" for HHG test, "Mantel" for mantel test, and "RV" for RV test.

hapMat  
An object of class hapMat containing SNV haplotypes.

nperm  
Number of permutations for the test of any association across the genomic region of interest. The default is nperm = 0; i.e., association will not be tested.

xlab  
An optional character string for the label on the x-axis in the plot that is returned (none by default).

ylab  
An optional character string for the label on the y-axis in the plot that is returned (none by default).

main  
An optional character string for title in the plot that is returned (none by default).

**Value**

A list with the following components:

Stats  
A vector of observed statistics computed from the user-provided distance association method.

OmPval  
A permutation-based omnibus P value for the test of any association across the genomic region using the maximum statistic over the genomic region as the test statistic.

mPval  
A vector of marginal P values at each SNV position.

plt  
A plot of the association profile over SNV locations in the region of interest.

**See Also**

HHGtest, dCorTest, RVtest, MantelTest

---

**Description**

This function performs the Rand Index between a user-supplied comparator dendrogram and the reconstructed dendrograms at each focal SNV position in a genomic region. See the section Applications in vignette(“perfectphyloR”) for the detailed example.

**Usage**

testDendAssoRI(rdend, cdend, hapMat, k = 2, nperm = 0, xlab = "", ylab = "", main = "")
**Arguments**

- `rdend` A `multiPhylo` object of reconstructed dendrograms at each focal SNV.
- `cdend` A `phylo` object of the comparator dendrogram.
- `hapMat` An object of class `hapMat` containing SNV haplotypes.
- `k` An integer that specifies the number of clusters that the dendrogram should be cut into. The default is `k=2`. Clusters are defined by starting from the root of the dendrogram and moving towards the tips, cutting horizontally at any given point in the dendrogram.
- `nperm` Number of permutations for the test of any association across the genomic region of interest. The default is `nperm = 0`; i.e., association will not be tested.
- `xlab` An optional character string for the label on the x-axis in the plot that is returned (none by default).
- `ylab` An optional character string for the label on the y-axis in the plot that is returned (none by default).
- `main` An optional character string for title in the plot that is returned (none by default).

**Value**

A list with the following components:

- `Stats` A vector of observed Rand indices.
- `OmPval` A permutation-based omnibus P value for the test of any association across the genomic region using the maximum Rand index over the genomic region as the test statistics.
- `mPval` A vector of marginal P values at each SNV position.
- `plt` A plot of the association profile of Rand indices over SNV locations in the region of interest.

---

**vcftohapMat**  
Create a `hapMat` object from variant call format (`vcf`) file.

**Description**

This function creates a `hapMat` object from variant call format (`vcf`) file.

**Usage**

`vcftohapMat(vcf_file_path)`

**Arguments**

- `vcf_file_path` File path to the `vcf` file.
Value

An object of class `hapMat`.

Examples

```r
## Not run:
# Specify the file path.
vcf_file_path <- "C:/vcfData/vcfData.vcf.gz"
# Create a hapMat object from the vcf file.
ex_vcf_hapMat <- vcftohapMat(vcf_file_path)

## End(Not run)
```
Index

* datasets
  ex_hapMat_data, 5
  ex_hapMatSmall_data, 4
  phenoDist, 7
  tdend, 14

createHapMat, 2, 10

dCorTest, 3, 15

ex_hapMat_data, 5
ex_hapMatSmall_data, 4

HHGtest, 5, 15

MantelTest, 6, 15

perfectphyloR (perfectphyloR-package), 2
perfectphyloR-package, 2
phenoDist, 7
plotDend, 7

RandIndexTest, 8
rdistMatrix, 9
reconstructPP, 2, 9, 10, 12
reconstructPPregion, 12
RVtest, 13, 15

tdend, 14
testAssoDist, 14
testDendAssoRI, 15

vcftohapMat, 16