

Package ‘phyloscannerR’

November 9, 2017

Title Phylogenetics between and within hosts at once, all along the genome

Version 1.3.0

Description An R package for the second half of phyloscanner (tree analysis).

Depends R (>= 3.4.0)

Imports ape, data.table (>= 1.10.4-3), dplyr, dtplyr, ff, GGally, ggplot2, ggtree, grid, gridExtra, gtable, kimisc, network, pegas, phangorn, phytools, prodlim, RColorBrewer, reshape2, scales, sna

License GPL

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

R topics documented:

draw.summary.statistics	1
gather.summary.statistics	2
multipage.summary.statistics	3
phyloscanner.analyse.tree	3
phyloscanner.analyse.trees	6
simplified.transmission.summary	9
transmission.summary	10
write.annotated.tree	10

Index	12
--------------	-----------

draw.summary.statistics

Graph summary statistics for a single host

Description

Graph summary statistics for a single host

Usage

```
draw.summary.statistics(phyloscanner.trees, sum.stats, host, verbose = F)
```

Arguments

phyloscanner.trees	A list of class phyloscanner.trees
sum.stats	The output of a call to gather.summary.statistics.
host	The host to obtain graphs for.
verbose	Verbose output

gather.summary.statistics

Make a data.table of per-window host statistics

Description

This function collects per-window statistics on hosts

Usage

```
gather.summary.statistics(phyloscanner.trees,
  hosts = all.hosts.from.trees(phyloscanner.trees),
  tip.regex = "^(.*)_read_([0-9]+)_count_([0-9]+)$", verbose = F)
```

Arguments

phyloscanner.trees	A list of class phyloscanner.trees
hosts	A list of hosts to record statistics for. If not specified, every identifiable host in phyloscanner.trees
tip.regex	Regular expression identifying tips from the dataset. This expects up to three capture groups, for host ID, read ID, and read count (in that order). If the latter two groups are missing then read information will not be used. The default matches input from the phyloscanner pipeline where the host ID is the BAM file name.
verbose	Produce verbose output

Value

A data.table

multipage.summary.statistics

Draw summary statistics to file for many hosts as a multipage file

Description

Draw summary statistics to file for many hosts as a multipage file

Usage

```
multipage.summary.statistics(phyloscanner.trees, sum.stats,
  hosts = all.hosts.from.trees(phyloscanner.trees), file.name,
  height = 11.6929, width = 8.26772, verbose = F)
```

Arguments

phyloscanner.trees	A list of class phyloscanner.trees
sum.stats	The output of a call to gather.summary.statistics.
hosts	A vector of hosts to obtain graphs for. By default, all hosts detected in phyloscanner.trees.
file.name	Output file name (should have a .pdf file extension)
height	The height of each page of the output file in inches (defaults to A4 size)
width	The width of each page of the output file in inches (defaults to A4 size)
verbose	Verbose output

phyloscanner.analyse.tree

Perform a phyloscanner analysis on a single tree

Description

This function performs a parsimony reconstruction and classification of pairwise host relationships.

Usage

```
phyloscanner.analyse.tree(tree.file.name, splits.rule = c("s", "r", "f"),
  sankoff.k = 0, sankoff.unsampled.switch.threshold = 0,
  continuation.unsampled.proximity.cost = 1000, outgroup.name = NULL,
  multifurcation.threshold = 0, guess.multifurcation.threshold = F,
  user.blacklist.file.name = NULL, duplicate.file.name = NULL,
  recombination.file.name = NULL,
  tip.regex = "^(.*)_read_[0-9+]_count_[0-9+]$",
  file.name.regex = "^\\D*[0-9+]_to_[0-9+]\\D*$",
  seed = sample(1:1e+07, 1), norm.ref.file.name = NULL,
  norm.standardise.gag.pol = F, norm.constants = NULL,
  parsimony.blacklist.k = 0, raw.blacklist.threshold = 0,
  ratio.blacklist.threshold = 0, do.dual.blacklisting = F,
  max.reads.per.host = Inf, blacklist.underrepresented = F, use.ff = F,
  prune.blacklist = F, read.counts.matter.on.zero.length.tips = T,
  verbose = F, no.progress.bars = F)
```

Arguments

<code>tree.file.name</code>	The name of the tree file (Newick or NEXUS format).
<code>splits.rule</code>	The rules by which the sets of hosts are split into groups in order to ensure that all groups can be members of connected subgraphs without causing conflicts. Options: s=Sankoff with optional within-host diversity penalty (slow, rigorous, recommended), r=Romero-Severson (quick, less rigorous with >2 hosts), f=Sankoff with continuation costs (experimental).
<code>sankoff.k</code>	For <code>splits.rule</code> = s or f only. The <i>k</i> parameter in the Sankoff reconstruction, representing the within-host diversity penalty.
<code>sankoff.unsampled.switch.threshold</code>	For <code>splits.rule</code> = s only. Threshold at which a lineage reconstructed as infecting a host will transition to the unsampled state, if it would be equally parsimonious to remain in that host.
<code>continuation.unsampled.proximity.cost</code>	For <code>splits.rule</code> = f only. The branch length at which an node is reconstructed as unsampled if all its neighbouring nodes are a greater distance away. The default is 1000, intended to be effectively infinite, such a node will never normally receive the unsampled state.
<code>outgroup.name</code>	The name of the tip in the phylogeny/phylogenies to be used as outgroup (if unspecified, trees will be assumed to be already rooted). This should be sufficiently distant to any sequence obtained from a host that it can be assumed that the MRCA of the entire tree was not a lineage present in any sampled individual.
<code>multifurcation.threshold</code>	If specified, branches shorter than this in the input tree will be collapsed to form multifurcating internal nodes. This is recommended; many phylogenetics packages output binary trees with short or zero-length branches indicating multifurcations.
<code>guess.multifurcation.threshold</code>	Whether to guess the multifurcation threshold from the branch lengths of the trees and the width of the genomic window (if that information is available). It is recommended that trees are examined by eye to check that they do appear to have multifurcations if using this option.
<code>user.blacklist.file.name</code>	The path of a text file containing the user-specified list of tips to be blacklisted
<code>duplicate.file.name</code>	The path of a .csv file specifying which tree tips are from duplicate reads. Normally this is produced by <code>phyloscanner_make_trees.py</code> .
<code>recombination.file.name</code>	The path for file containing the results of the <code>phyloscanner_make_trees.py</code> recombination metric analysis.
<code>tip.regex</code>	Regular expression identifying tips from the dataset. This expects up to three capture groups, for host ID, read ID, and read count (in that order). If the latter two groups are missing then read information will not be used. The default matches input from the phyloscanner pipeline where the host ID is the BAM file name.
<code>file.name.regex</code>	Regular expression identifying window coordinates. Two capture groups: start and end; if the latter is missing then the first group is a single numerical identifier for the window. The default matches input from the phyloscanner pipeline.

seed	Random number seed; used by the downsampling process, and also ties in some parsimony reconstructions can be broken randomly.
norm.ref.file.name	Name of a file giving a normalisation constant for every genome position. Cannot be used simultaneously with norm.constants. If neither is given then no normalisation will be performed.
norm.standardise.gag.pol	Use only if norm.ref.file.name is given. An HIV-specific option: if true, the normalising constants are standardised so that the average on gag+pol equals 1. Otherwise they are standardised so the average on the whole genome equals 1.
norm.constants	Either the path of a CSV file listing the file name for each tree (column 1) and the respective normalisation constant (column 2) or a single numerical normalisation constant to be applied to every tree. Cannot be used simultaneously with norm.ref.file.name. If neither is given then no normalisation will be performed.
parsimony.blacklist.k	The k parameter of the single-host Sankhoff parsimony reconstruction used to identify probable contaminants. A value of 0 is equivalent to not performing parsimony blacklisting.
raw.blacklist.threshold	Used to specify a read count to be used as a raw threshold for duplicate or parsimony blacklisting. Use with parsimony.blacklist.k or duplicate.file.regex or both. Parsimony blacklisting will blacklist any subgraph with a read count strictly less than this threshold. Duplicate blacklisting will blacklist any duplicate read with a count strictly less than this threshold. The default value of 0 means nothing is blacklisted.
ratio.blacklist.threshold	Used to specify a read count ratio (between 0 and 1) to be used as a threshold for duplicate or parsimony blacklisting. Use with parsimony.blacklist.k or duplicate.file.regex or both. Parsimony blacklisting will blacklist a subgraph if the ratio of its read count to the total read count from the same host is strictly less than this threshold. Duplicate blacklisting will blacklist a duplicate read if the ratio of its count to the count of the duplicate (from another host) is strictly less than this threshold.
do.dual.blacklisting	Blacklist all reads from the minor subgraphs for all hosts established as dual by parsimony blacklisting (which must have been done for this to do anything).
max.reads.per.host	Used to turn on downsampling. If given, reads will be blacklisted such that read counts (or tip counts if no read counts are identified) from each host are equal (although see blacklist.underrepresented).
blacklist.underrepresented	If TRUE and max.reads.per.host is given, blacklist hosts from trees where their total tip count does not reach the maximum.
use.ff	Use the ff package to store parsimony reconstruction matrices. Use if you run out of memory.
prune.blacklist	If TRUE, all blacklisted and reference tips (except the outgroup) are pruned away before starting parsimony-based reconstruction.

`read.counts.matter.on.zero.length.tips`
 If TRUE, read counts on tips will be taken into account in parsimony reconstructions at the parents of zero-length terminal branches. Not applicable for the Romero-Severson-like reconstruction method.

`verbose`
 Give verbose output.

`no.progress.bars`
 Hide the progress bars from verbose output.

Value

A list of class `phyloscanner.trees` with a single item of class `phyloscanner.tree`.

`phyloscanner.analyse.trees`

Perform a phyloscanner analysis on a set of trees

Description

This function performs a parsimony reconstruction and classification of pairwise host relationships.

Usage

```
phyloscanner.analyse.trees(tree.directory,
  tree.file.regex = "^RAXML_bestTree.InWindow_([0-9]+_to_[0-9]+)\\.tree$",
  splits.rule = c("s", "r", "f"), sankoff.k = 0,
  sankoff.unsampled.switch.threshold = 0,
  continuation.unsampled.proximity.cost = 1000, outgroup.name = NULL,
  multifurcation.threshold = 0, guess.multifurcation.threshold = F,
  user.blacklist.directory = NULL, user.blacklist.file.regex = NULL,
  duplicate.file.directory = NULL,
  duplicate.file.regex = "^DuplicateReadCountsProcessed_InWindow_([0-9]+_to_[0-9]+).csv$",
  recombination.file.directory = NULL,
  recombination.file.regex = "^RecombinantReads_InWindow_([0-9]+_to_[0-9]+).csv$",
  tip.regex = "^(.*)_read_([0-9]+)_count_([0-9]+)$",
  file.name.regex = "^\\D*([0-9]+)_to_([0-9]+)\\D*$",
  seed = sample(1:1e+07, 1), norm.ref.file.name = NULL,
  norm.standardise.gag.pol = F, norm.constants = NULL,
  parsimony.blacklist.k = 0, raw.blacklist.threshold = 0,
  ratio.blacklist.threshold = 0, do.dual.blacklisting = F,
  max.reads.per.host = Inf, blacklist.underrepresented = F, use.ff = F,
  prune.blacklist = F, read.counts.matter.on.zero.length.tips = T,
  verbose = F, no.progress.bars = F)
```

Arguments

`tree.directory` The directory containing all input trees.

`tree.file.regex`
 A regular expression identifying every file in `tree.directory` that is to be included in the analysis. The first capture group, if present, gives a unique string identifying each tree. If this is NULL then phyloscanner will attempt to open every file in `tree.directory`.

splits.rule	The rules by which the sets of hosts are split into groups in order to ensure that all groups can be members of connected subgraphs without causing conflicts. Options: s=Sankoff with optional within-host diversity penalty (slow, rigorous, recommended), r=Romero-Severson (quick, less rigorous with >2 hosts), f=Sankoff with continuation costs (experimental).
sankoff.k	For splits.rule = s or f only. The <i>k</i> parameter in the Sankoff reconstruction, representing the within-host diversity penalty.
sankoff.unsampled.switch.threshold	For splits.rule = s only. Threshold at which a lineage reconstructed as infecting a host will transition to the unsampled state, if it would be equally parsimonious to remain in that host.
continuation.unsampled.proximity.cost	For splits.rule = f only. The branch length at which an node is reconstructed as unsampled if all its neighbouring nodes are a greater distance away. The default is 1000, intended to be effectively infinite, such a node will never normally receive the unsampled state.
outgroup.name	The name of the tip in the phylogeny/phylogenies to be used as outgroup (if unspecified, trees will be assumed to be already rooted). This should be sufficiently distant to any sequence obtained from a host that it can be assumed that the MRCA of the entire tree was not a lineage present in any sampled individual.
multifurcation.threshold	If specified, branches shorter than this in the input tree will be collapsed to form multifurcating internal nodes. This is recommended; many phylogenetics packages output binary trees with short or zero-length branches indicating multifurcations.
guess.multifurcation.threshold	Whether to guess the multifurcation threshold from the branch lengths of the trees and the width of the genomic window (if that information is available). It is recommended that trees are examined by eye to check that they do appear to have multifurcations if using this option.
user.blacklist.directory	An optional path for a folder containing pre-existing blacklist files. These tips are specified by the user to be excluded from the analysis.
user.blacklist.file.regex	A regular expression identifying every file in user.blacklist.directory that contains a blacklist. If a capture group is specified then its contents will uniquely identify the tree it belongs to, which must matches the IDs found by user.blacklist.regex. If these IDs cannot be identified then matching will be attempted using genome window coordinates.
duplicate.file.directory	An optional path for a folder containing information on duplicate reads, to be used for duplicate blacklisting. Normally this is produced by phyloscanner_make_trees.py.
duplicate.file.regex	A regular expression identifying every file in duplicate.file.directory that contains a duplicates file. If a capture group is specified then its contents will uniquely identify the tree it belongs to, which must matches the IDs found by duplicate.file.regex. If these IDs cannot be identified then matching will be attempted using genome window coordinates.
recombination.file.directory	An optional path for a folder containing results of the phyloscanner_make_trees.py recombination metric analysis.

<code>recombination.file.regex</code>	A regular expression identifying every file in <code>recombination.file.directory</code> that contains a recombination file. If a capture group is specified then its contents will uniquely identify the tree it belongs to, which must match the IDs found by <code>recombination.file.regex</code> . If these IDs cannot be identified then matching will be attempted using genome window coordinates.
<code>tip.regex</code>	Regular expression identifying tips from the dataset. This expects up to three capture groups, for host ID, read ID, and read count (in that order). If the latter two groups are missing then read information will not be used. The default matches input from the phyloscanner pipeline where the host ID is the BAM file name.
<code>file.name.regex</code>	Regular expression identifying window coordinates. Two capture groups: start and end; if the latter is missing then the first group is a single numerical identifier for the window. The default matches input from the phyloscanner pipeline.
<code>seed</code>	Random number seed; used by the downsampling process, and also ties in some parsimony reconstructions can be broken randomly.
<code>norm.ref.file.name</code>	Name of a file giving a normalisation constant for every genome position. Cannot be used simultaneously with <code>norm.constants</code> . If neither is given then no normalisation will be performed.
<code>norm.standardise.gag.pol</code>	Use only if <code>norm.ref.file.name</code> is given. An HIV-specific option: if true, the normalising constants are standardised so that the average on gag+pol equals 1. Otherwise they are standardised so the average on the whole genome equals 1.
<code>norm.constants</code>	Either the path of a CSV file listing the file name for each tree (column 1) and the respective normalisation constant (column 2) or a single numerical normalisation constant to be applied to every tree. Cannot be used simultaneously with <code>norm.ref.file.name</code> . If neither is given then no normalisation will be performed.
<code>parsimony.blacklist.k</code>	The k parameter of the single-host Sankhoff parsimony reconstruction used to identify probable contaminants. A value of 0 is equivalent to not performing parsimony blacklisting.
<code>raw.blacklist.threshold</code>	Used to specify a read count to be used as a raw threshold for duplicate or parsimony blacklisting. Use with <code>parsimony.blacklist.k</code> or <code>duplicate.file.regex</code> or both. Parsimony blacklisting will blacklist any subgraph with a read count strictly less than this threshold. Duplicate blacklisting will blacklist any duplicate read with a count strictly less than this threshold. The default value of 0 means nothing is blacklisted.
<code>ratio.blacklist.threshold</code>	Used to specify a read count ratio (between 0 and 1) to be used as a threshold for duplicate or parsimony blacklisting. Use with <code>parsimony.blacklist.k</code> or <code>duplicate.file.regex</code> or both. Parsimony blacklisting will blacklist a subgraph if the ratio of its read count to the total read count from the same host is strictly less than this threshold. Duplicate blacklisting will blacklist a duplicate read if the ratio of its count to the count of the duplicate (from another host) is strictly less than this threshold.
<code>do.dual.blacklisting</code>	Blacklist all reads from the minor subgraphs for all hosts established as dual by parsimony blacklisting (which must have been done for this to do anything).

<code>max.reads.per.host</code>	Used to turn on downsampling. If given, reads will be blacklisted such that read counts (or tip counts if no read counts are identified) from each host are equal (although see <code>blacklist.underrepresented</code>).
<code>blacklist.underrepresented</code>	If TRUE and <code>max.reads.per.host</code> is given, blacklist hosts from trees where their total tip count does not reach the maximum.
<code>use.ff</code>	Use the <code>ff</code> package to store parsimony reconstruction matrices. Use if you run out of memory.
<code>prune.blacklist</code>	If TRUE, all blacklisted and reference tips (except the outgroup) are pruned away before starting parsimony-based reconstruction.
<code>read.counts.matter.on.zero.length.tips</code>	If TRUE, read counts on tips will be taken into account in parsimony reconstructions at the parents of zero-length terminal branches. Not applicable for the Romero-Severson-like reconstruction method.
<code>verbose</code>	Give verbose output.
<code>no.progress.bars</code>	Hide the progress bars from verbose output.

Value

A list of class `phyloscanner.trees`.

`simplified.transmission.summary`

Simplify and visually display the pairwise host relationships across all trees

Description

Simplify and visually display the pairwise host relationships across all trees

Usage

```
simplified.transmission.summary(phyloscanner.trees, transmission.summary,
  arrow.threshold, plot = F)
```

Arguments

<code>phyloscanner.trees</code>	A list of class <code>phyloscanner.trees</code>
<code>arrow.threshold</code>	The proportion of trees in which a pair of hosts need to show a direction of transmission for that direction to be indicated as an arrow. If both directions meet this threshold, the arrow is in the direction with the larger proportion of trees.
<code>plot</code>	If TRUE, the returned list has an item called <code>simp.diagram</code> , a <code>ggplot</code> object plotting the simplified relationship diagram.
<code>trans.summary</code>	The output of <code>transmission.summary</code> ; a <code>data.table</code> .

transmission.summary	<i>Summarise the pairwise host relationships across all trees</i>
----------------------	---

Description

Summarise the pairwise host relationships across all trees

Usage

```
transmission.summary(phyloscanner.trees, win.threshold = 0,
  dist.threshold = Inf, allow.mt = T, close.sib.only = F, verbose = F)
```

Arguments

phyloscanner.trees	A list of class phyloscanner.trees
win.threshold	The proportion of windows that a pair of hosts need to be related (adjacent and within dist.threshold of each other) in order for them to appear in the summary.
dist.threshold	The patristic distance within which the subgraphs from two hosts need to be in order for them to be declared related (default is infinity, so adjacent hosts are always related).
allow.mt	If FALSE, directionality is only inferred between pairs of hosts where a single clade from one host is nested in one from the other; this is more conservative.
close.sib.only	If TRUE, then the distance threshold applies only to hosts on sibling clades. Any ancestry is automatically a relationship.
verbose	Give verbose output

Value

A data.table, every line of which counts the number of pairwise relationships of a particular type between a pair of hosts

write.annotated.tree	<i>Write the phylogeny with reconstructed host annotations to file</i>
----------------------	--

Description

Write the phylogeny with reconstructed host annotations to file

Usage

```
write.annotated.tree(phyloscanner.tree, file.name, format = c("pdf", "nexus"),
  pdf.scale.bar.width = 0.01, pdf.w = 50, pdf.hm = 0.15)
```

Arguments

<code>phyloscanner.tree</code>	A list of class <code>phyloscanner.tree</code> (usually an item in a list of class <code>phyloscanner.trees</code>)
<code>file.name</code>	The name of the output file
<code>format</code>	The format - PDF or NEXUS - in which to write the output.
<code>pdf.scale.bar.width</code>	The width, in substitutions per site, of the scale bar in PDF output
<code>pdf.w</code>	The width of the output PDF file, in inches
<code>pdf.hm</code>	The height, in inches per tip, of the output PDF file

Index

`draw.summary.statistics`, [1](#)
`gather.summary.statistics`, [2](#)
`multipage.summary.statistics`, [3](#)
`phyloscanner.analyse.tree`, [3](#)
`phyloscanner.analyse.trees`, [6](#)
`simplified.transmission.summary`, [9](#)
`transmission.summary`, [10](#)
`write.annotated.tree`, [10](#)