Package 'longreadvqs'

April 14, 2024

Title Viral Quasispecies Comparison from Long-Read Sequencing Data

Version 0.1.2

Description Performs variety of viral quasispecies diversity analyses [see Gregori et al. (2016) <doi:10.1016/j.virol.2016.03.017>] based on long-read sequence alignment. Main functions include 1) sequencing error minimization and read sampling, 2) Single nucleotide variant (SNV) profiles comparison, and 3) viral quasispecies profiles comparison and visualization.

License GPL-3

URL https://github.com/NakarinP/longreadvqs

BugReports https://github.com/NakarinP/longreadvqs/issues

Encoding UTF-8

RoxygenNote 7.3.1

biocViews

Imports ape, Biostrings, cowplot, dplyr, ggplot2, ggpubr, grDevices, magrittr, plyr, purrr, QSutils, RColorBrewer, reshape2, scales, seqinr, stats, stringdist, stringr, tibble, tidyr

Depends R (>= 2.10)

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation no

Author Nakarin Pamornchainavakul [aut, cre] (<https://orcid.org/0000-0003-0378-0316>)

Maintainer Nakarin Pamornchainavakul <pamornakarin@gmail.com>

Repository CRAN

Date/Publication 2024-04-14 19:20:02 UTC

R topics documented:

filtfast . . . 2
gapremove . . . 3
otucompare . . . 4
pctopt . . . 5
snvcompare . . . 6
vqsassess . . . 6
vqscompare . . . 7
vqscustompct . . . 9
vqsout . . . 10
vqsresub . . . 11
vqssub . . . 12

Index **14**

`filtfast` *Filtering highly dissimilar reads/sequences out of the alignment*

Description

Removes reads/sequences of which Hamming similarity to the consensus of all reads/sequences in the alignment is less than the specified quantile (qt) of the similarity distribution.

Usage

```
filtfast(fasta, qt = 0.25, fastaname = "filteredfast.fasta")
```

Arguments

`fasta`	Input as a read or multiple sequence alignment in FASTA format
`qt`	If Hamming similarity score of a read/sequence to the consensus of all reads/sequences is less than the specified quantile (qt) of the similarity distribution, that read/sequence will be removed.
`fastaname`	Output file name in FASTA format

Value

FASTA read or multiple sequence alignment written out to the input directory

Examples

```
## Locate input FASTA file-----------------------------------------------------------------------
fastafilepath <- system.file("extdata", "dissimfast.fasta", package = "longreadvqs")

## Indicate output directory and file name-------------------------------------------------------
outfast <- tempfile()

## Remove reads/sequences that the similarity < 1st quartile (0.25 quantile)---------------------
```

```
filtfast(fastafilepath, qt = 0.25, fastaname = outfast)
```

---

gapremove *Removing gap-rich positions and/or reads/sequences*

Description

Removes nucleotide positions (vertical) and/or reads/sequences (horizontal) that contain gaps more than the specified cut-off percentage from the alignment.

Usage

```
gapremove(fasta, vgappct = 70, hgappct = 70, fastaname = "filteredfast.fasta")
```

Arguments

fasta	Input as a read or multiple sequence alignment in FASTA format
vgappct	The percent cut-off of vertical gap (-), i.e., if a position in the alignment has %gap >= vgappct, that position will be removed.
hgappct	The percent cut-off of horizontal gap (-), i.e., if a sequence or read in the alignment has %gap >= hgappct, that sequence or read will be removed.
fastaname	Output file name in FASTA format

Value

FASTA read or multiple sequence alignment written out to the input directory

Examples

```
## Locate input FASTA file-----------------------------------------------------------------------
fastafilepath <- system.file("extdata", "gaprichfast.fasta", package = "longreadvqs")

## Indicate output directory and file name-------------------------------------------------------
outfast <- tempfile()

## Remove positions with gap >= 60% and reads/sequences with gap >= 10%--------------------------
gapremove(fastafilepath, vgappct = 60, hgappct = 10, fastaname = outfast)
```

otucompare *Comparing operational taxonomic unit (OTU) by k-means clustering between samples*

Description

Pools error-minimized down-sampled read samples and compares their diversity based on operational taxonomic unit (OTU) classified by k-means clustering of single nucleotide variant (SNV) distance. This function is a subset of "vqscompare" function.

Usage

```
otucompare(samplelist = list(BC1, BC2, BC3), kmeans.n = 20)
```

Arguments

samplelist	List of samples, i.e., name of resulting objects from "vqsassess" or "vqscustompct" functions, for example list(BC1, BC2, BC3).
kmeans.n	Number of clusters or operational taxonomic units (OTUs) needed from k-means clustering on multidimensional scale (MDS) of all samples' pairwise genetic distance.

Value

Comparative table of OTU diversity metrics between listed samples calculated from consensus sequence of each OTU by QSutils package

Examples

```
## Locate input FASTA files--------------------------------------------------------------------------
sample1filepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")
sample2filepath <- system.file("extdata", "s2.fasta", package = "longreadvqs")

## Prepare data for viral quasispecies comparison between two samples----------------------------
sample1 <- vqsassess(sample1filepath, pct = 10, samsize = 20, label = "sample1")
sample2 <- vqsassess(sample2filepath, pct = 10, samsize = 20, label = "sample2")

## Compare OTU (4 clusters) diversity metrics  between two samples-------------------------------
otucompare(samplelist = list(sample1, sample2), kmeans.n = 4)
```

pctopt *Optimizing cut-off percentage for error minimization*

Description

Finds an optimal cut-off percentage for error minimization (in vqssub, vqsassess, and vqscustompct functions) that can decrease the number of singleton haplotypes to less than the desired percentage of the total reads.

Arguments

fasta	Input as a read alignment in FASTA format
pctsing	The desired percentage of singleton haplotypes relative to the total reads in the alignment.
method	Sequencing error minimization methods that replace low frequency nucleotide base (less than the "pct" cut-off) with consensus base of that position ("conbase": default) or with base of the dominant haplotype ("domhapbase").
samplingfirst	Downsampling before (TRUE) or after (FALSE: default) the error minimization.
gappct	The percent cut-off particularly specified for gap (-). If it is not specified or less than "pct", "gappct" will be equal to "pct" (default).
ignoregappositions	Replace all nucleotides in the positions in the alignment containing gap(s) with gap. This will make such positions no longer single nucleotide variant (SNV). The default is "FALSE".
samsize	Sample size (number of reads) after down-sampling. If it is not specified or more than number of reads in the original alignment, down-sampling will not be performed (default).
label	String within quotation marks indicating name of read alignment (optional).

Value

An optimal cut-off percentage for error minimization of an input sample and parameter settings. If label is specified, the output will be a data frame with percentage of singleton haplotypes at each cut-off percentage from zero to the optimal cut-off percentage.

Examples

```
## Locate input FASTA file--------------------------------------------------------------------------
fastafilepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")

## Find an cut-off percentage that creates singleton haplotypes less than 50% of the alignment.----
pctopt(fastafilepath, pctsing = 50, label = "s1")
```

snvcompare *Plotting single nucleotide variant (SNV) frequency in read alignment across different samples*

Description

Compares single nucleotide variant (SNV) profile between error-minimized down-sampled read samples using cowplot's "plot_grid" function. The resulting plot may help evaluating the optimal cut-off percentage of low frequency nucleotide base used in "vqsassess", "vqscustompct", or "vqssub" functions.

Arguments

samplelist	List of samples, i.e., name of resulting objects from "vqsassess" or "vqscustompct" functions, for example list(BC1, BC2, BC3).
ncol	Number of columns for multiple plots (see cowplot's "plot_grid" function)

Value

Comparative plot of SNV frequency in read alignment across different samples

Examples

```
## Locate input FASTA files-----------------------------------------------------------------------
sample1filepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")
sample2filepath <- system.file("extdata", "s2.fasta", package = "longreadvqs")

## Prepare data for viral quasispecies comparison between two samples-----------------------------
sample1 <- vqsassess(sample1filepath, pct = 10, label = "sample1")
sample2 <- vqsassess(sample2filepath, pct = 10, label = "sample2")

## Compare SNV profile between two listed samples-------------------------------------------------
snvcompare(samplelist = list(sample1, sample2), ncol = 1)
```

vqsassess *Sequencing error minimization, read down-sampling, and data preparation for viral quasispecies comparison*

Description

Minimizes potential long-read sequencing error based on the specified cut-off percentage of low frequency nucleotide base and down-samples read for further comparison with other samples. The output of this function is a list of several objects representing diversity of each sample that must be used as an input for other functions such as "snvcompare" or "vqscompare".

Arguments

fasta	Input as a read alignment in FASTA format
method	Sequencing error minimization methods that replace low frequency nucleotide base (less than the "pct" cut-off) with consensus base of that position ("conbase": default) or with base of the dominant haplotype ("domhapbase").
samplingfirst	Downsampling before (TRUE) or after (FALSE: default) the error minimization.
pct	Percent cut-off defining low frequency nucleotide base that will be replaced (must be specified).
gappct	The percent cut-off particularly specified for gap (-). If it is not specified or less than "pct", "gappct" will be equal to "pct" (default).
ignoregappositions	Replace all nucleotides in the positions in the alignment containing gap(s) with gap. This will make such positions no longer single nucleotide variant (SNV). The default is "FALSE".
samsize	Sample size (number of reads) after down-sampling. If it is not specified or more than number of reads in the original alignment, down-sampling will not be performed (default).
label	String within quotation marks indicating name of read alignment (optional). Please don't use underscore (_) in the label.

Value

list of 1) "dat": viral quasispecies diversity metrics calculated by QSutils package (similar to "vqssub" function's output), 2) "snvhap": SNV profile of each haplotype with frequency and new label for "vqscompare" function, 3) "snv": plot of SNV frequency for "snvcompare" function, 4) "hapre": DNAStringSet of read alignment of each haplotype for "vqscompare" function, 5) "lab": name of sample or read alignment

Examples

```
## Locate input FASTA file-----------------------------------------------------------------------
sample1filepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")

## Prepare data for viral quasispecies comparison ------------------------------------------------
sample1 <- vqsassess(sample1filepath, pct = 10, samsize = 20, label = "sample1")

## For more examples on other choices of arguments, please see "vqssub" function's examples-------
```

---

vqscompare *Comparing viral quasispecies profile and operational taxonomic unit (OTU) classified by k-means clustering between samples*

Description

Pools error-minimized down-sampled read samples and compares their diversity by 1) viral quasispecies profile (haplotype and metrics from QSutils package), 2) operational taxonomic unit (OTU) classified by k-means clustering of single nucleotide variant (SNV) distance, and 3) visualization of different comparative method, i.e., haplotype, OTU, phylogenetic tree, MDS plot.

Arguments

samplelist	List of samples, i.e., name of resulting objects from "vqsassess" or "vqscustompct" functions, for example list(BC1, BC2, BC3).
lab_name	Name of variable or type of sample for instance "barcode", "sample", "dpi", or "isolate" (optional).
kmeans.n	Number of clusters or operational taxonomic units (OTUs) needed from k-means clustering on multidimensional scale (MDS) of all samples' pairwise SNV distance.
showhap.n	Number of largest haplotypes (default = 30) labeled in the top five OTUs' MDS plot (optional).

Value

list of 1) "hapdiv": comparative table of viral quasispecies diversity metrics between listed samples calculated by QSutils package, 2) "otudiv": comparative table of OTU diversity metrics between listed samples calculated from consensus sequence of each OTU (similar to "otucompare" function's output), 3) "sumsnv_hap": frequency and SNV profile (by position in the alignment) of haplotypes that are not singleton (number of reads > 1), 4) "sumsnv_otu": frequency and SNV profile of all haplotypes grouped into different operational taxonomic unit (OTU), 5) "fullseq": complete read sequence of haplotypes that are not singleton, 6) "fulldata": complete read sequence of all haplotypes in every sample with frequency and OTU classification, 7) "summaryplot": visualization of viral quasispecies comparison between samples including 7.1) "happlot": proportion of haplotypes (top left), 7.2) "otuplot": proportion of OTUs (bottom left), and 7.3) multidimensional scale (MDS) plots (right) of k-means OTU ("top5otumds": 5 largest groups with major haplotypes labeled and "allotumds": all groups)

Examples

```
## Locate input FASTA files-----------------------------------------------------------------------
sample1filepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")
sample2filepath <- system.file("extdata", "s2.fasta", package = "longreadvqs")

## Prepare data for viral quasispecies comparison between two samples-----------------------------
set.seed(123)
sample1 <- vqsassess(sample1filepath, pct = 5, samsize = 50, label = "sample1")
sample2 <- vqsassess(sample2filepath, pct = 5, samsize = 50, label = "sample2")

## Compare viral quasispecies and OTU (4 clusters) diversity between two samples------------------
out <- vqscompare(samplelist = list(sample1, sample2),
           lab_name = "Sample", kmeans.n = 4, showhap.n = 5)
out$summaryplot
```

vqscustompct *Sequencing error minimization with customized % cut-off at particular nucleotide region, read down-sampling, and data preparation for viral quasispecies comparison*

Description

Minimizes potential long-read sequencing error based on the specified cut-off percentages of low frequency nucleotide base and down-samples read for further comparison with other samples. In this function, the cut-off percentage can be specifically adjusted for different ranges of nucleotide positions which is very useful when sequencing error heavily occurs in a particular part of reads. The output of this function is a list of several objects representing diversity of each sample that must be used as an input for other functions such as "snvcompare" or "vqscompare".

Arguments

`fasta`	Input as a read alignment in FASTA format
`method`	Sequencing error minimization methods that replace low frequency nucleotide base (less than the "pct" cut-off) with consensus base of that position ("conbase": default) or with base of the dominant haplotype ("domhapbase").
`samplingfirst`	Downsampling before (TRUE) or after (FALSE: default) the error minimization.
`pct`	Percent cut-off defining low frequency nucleotide base that will be replaced (must be specified).
`brkpos`	Ranges of nucleotide positions with different % cut-off specified in "lspct" for example c("1:50","51:1112") meaning that the first and the second ranges are nucleotide positions 1 to 50 and 51 to 1112, respectively.
`lspct`	List of customized % cut-off applied to nucleotide ranges set in "brkpos" for example c(15,8) meaning that 15% and 8% cut-offs will be applied to the first and the second ranges, respectively.
`gappct`	The percent cut-off particularly specified for gap (-). If it is not specified or less than "pct", "gappct" will be equal to "pct" (default).
`ignoregappositions`	Replace all nucleotides in the positions in the alignment containing gap(s) with gap. This will make such positions no longer single nucleotide variant (SNV). The default is "FALSE".
`samsize`	Sample size (number of reads) after down-sampling. If it is not specified or more than number of reads in the original alignment, down-sampling will not be performed (default).
`label`	String within quotation marks indicating name of read alignment (optional). Please don't use underscore (_) in the label.

Value

list of 1) "dat": viral quasispecies diversity metrics calculated by QSutils package (similar to "vqssub" function's output), 2) "snvhap": SNV profile of each haplotype with frequency and new label for "vqscompare" function, 3) "snv": plot of SNV frequency for "snvcompare" function, 4) "hapre": DNAStringSet of read alignment of each haplotype for "vqscompare" function, 5) "lab": name of sample or read alignment

Examples

```
## Locate input FASTA file--------------------------------------------------------------------------
fastafilepath <- system.file("extdata", "badend.fasta", package = "longreadvqs")

## Prepare data for viral quasispecies comparison using 10% cut-off across all positions----------
nocustom <- vqsassess(fastafilepath, pct = 10, label = "nocustom")

## Prepare data using 10% cut-off for the first 74 positions and 30% cut-off for the rest---------
custom <- vqscustompct(fastafilepath, pct = 10,
                       brkpos = c("1:74","75:84"), lspct = c(10,30), label = "custom")

## Use "snvcompare" function to check whether SNV profile looks better or not-------------------
snvcompare(samplelist = list(nocustom, custom), ncol = 1)
```

`vqsout` *Exporting viral quasispecies profile comparison results*

Description

Writes out resulting objects from "vqscompare" function as tables (TSV files) and alignment (FASTA file) to the working directory.

Usage

```
vqsout(vqscompare.obj, directory = "path/to/directory")
```

Arguments

`vqscompare.obj`	A resulting object from "vqscompare" function.
`directory`	Path to desired directory (location) for output files. If it is not specified, the directory will be the current working directory.

Value

TSV files of viral quasispecies profile comparison results and FASTA file of unique haplotype alignment.

Examples

```
## Locate input FASTA files-------------------------------------------------------------------
sample1filepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")
sample2filepath <- system.file("extdata", "s2.fasta", package = "longreadvqs")

## Prepare data for viral quasispecies comparison between two samples-----------------------------
set.seed(123)
sample1 <- vqsassess(sample1filepath, pct = 0, samsize = 50, label = "sample1")
sample2 <- vqsassess(sample2filepath, pct = 0, samsize = 50, label = "sample2")

## Compare viral quasispecies and OTU (4 clusters) diversity between two samples------------------
comp <- vqscompare(samplelist = list(sample1, sample2),
           lab_name = "Sample", kmeans.n = 4, showhap.n = 5)

## Export Key outputs from "vqscompare" function--------------------------------------------------
notrun <- vqsout(comp, directory = tempdir())
```

---

vqsresub *Computing viral quasispecies diversity metrics of error-minimized repeatedly down-sampled read alignments*

Description

Minimizes potential long-read sequencing error based on the specified cut-off percentage of low frequency nucleotide base and repeatedly down-samples read for sensitivity analysis of the diversity metrics varied by different sample sizes. The output of this function is a summary of viral quasispecies diversity metrics per each iteration of down-sampling calculated by QSutils package's functions. This function is an extension of "vqssub" function.

Arguments

`fasta`	Input as a read alignment in FASTA format
`iter`	Number of iterations for downsampling after error minimization.
`method`	Sequencing error minimization methods that replace low frequency nucleotide base (less than the "pct" cut-off) with consensus base of that position ("conbase": default) or with base of the dominant haplotype ("domhapbase").
`pct`	Percent cut-off defining low frequency nucleotide base that will be replaced (must be specified).
`gappct`	The percent cut-off particularly specified for gap (-). If it is not specified or less than "pct", "gappct" will be equal to "pct" (default).
`ignoregappositions`	Replace all nucleotides in the positions in the alignment containing gap(s) with gap. This will make such positions no longer single nucleotide variant (SNV). The default is "FALSE".

samsize	Sample size (number of reads) after down-sampling. If it is not specified or more than number of reads in the original alignment, down-sampling will not be performed (default).
label	String within quotation marks indicating name of read alignment (optional).

Value

Data frame containing all viral quasispecies diversity metrics calculated by QSutils package, error minimization, and down-sampling information per each downsampling iteration.

Examples

```
## Locate input FASTA file-----------------------------------------------------------------------
fastafilepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")

## Summarize viral quasispecies diversity metrics from five downsampling iterations.--------------
vqsresub(fastafilepath, iter = 5, pct = 10, samsize = 20, label = "sample1")
```

---

vqssub *Computing viral quasispecies diversity metrics of error-minimized down-sampled read alignment*

Description

Minimizes potential long-read sequencing error based on the specified cut-off percentage of low frequency nucleotide base and down-samples read for further comparison with other samples. The output of this function is a summary of viral quasispecies diversity metrics calculated by QSutils package's functions. This function is a subset of "vqsassess" function.

Arguments

fasta	Input as a read alignment in FASTA format
method	Sequencing error minimization methods that replace low frequency nucleotide base (less than the "pct" cut-off) with consensus base of that position ("conbase": default) or with base of the dominant haplotype ("domhapbase").
samplingfirst	Downsampling before (TRUE) or after (FALSE: default) the error minimization.
pct	Percent cut-off defining low frequency nucleotide base that will be replaced (must be specified).
gappct	The percent cut-off particularly specified for gap (-). If it is not specified or less than "pct", "gappct" will be equal to "pct" (default).
ignoregappositions	Replace all nucleotides in the positions in the alignment containing gap(s) with gap. This will make such positions no longer single nucleotide variant (SNV). The default is "FALSE".

samsize	Sample size (number of reads) after down-sampling. If it is not specified or more than number of reads in the original alignment, down-sampling will not be performed (default).
label	String within quotation marks indicating name of read alignment (optional).

Value

Data frame containing all viral quasispecies diversity metrics calculated by QSutils package, error minimization, and down-sampling information.

Examples

```
## Locate input FASTA file----------------------------------------------------------------------
fastafilepath <- system.file("extdata", "s1.fasta", package = "longreadvqs")

## Summarize viral quasispecies diversity metrics-----------------------------------------------
# From error-minimized unsampled reads (10% cut-off):
vqssub(fastafilepath, pct = 10, label = "sample1")
# From error-minimized sampled reads (n = 20):
vqssub(fastafilepath, pct = 10, samsize = 20, label = "sample1")
# From error-minimized sampled reads with 50% cut-off for gap:
vqssub(fastafilepath, pct = 10, gappct = 50, samsize = 20, label = "sample1")
# From error-minimized sampled reads but ignore positions with gap:
vqssub(fastafilepath, pct = 10, ignoregappositions = TRUE, samsize = 20, label = "sample1")
# From reads that were down-sampled before error minimization:
vqssub(fastafilepath, pct = 10, samplingfirst = TRUE, samsize = 20, label = "sample1")
```

Index

filtfast, 2

gapremove, 3

otucompare, 4

pctopt, 5

snvcompare, 6

vqsassess, 6
vqscompare, 7
vqscustompct, 9
vqsout, 10
vqsresub, 11
vqssub, 12