Package 'QSutils'

October 13, 2019

Type Package

Title Quasispecies Diversity

Version 1.3.0

Date 2018-04-06

Author Mercedes Guerrero-Murillo and Josep Gregori i Font

Maintainer Mercedes Guerrero-Murillo <mergumu@gmail.com>

Depends R (>= 3.5), Biostrings, BiocGenerics,methods

Imports ape, stats, psych

Encoding UTF-8

Description Set of utility functions for viral quasispecies analysis with NGS data. Most functions are equally useful for metagenomic studies. There are three main types: (1) data manipulation and exploration—functions useful for converting reads to haplotypes and frequencies, repairing reads, intersecting strand haplotypes, and visualizing haplotype alignments. (2) diversity indices—functions to compute diversity and entropy, in which incidence, abundance, and functional indices are considered. (3) data simulation—functions useful for generating random viral quasispecies data.

License file LICENSE

biocViews Software, Genetics, DNASeq, GeneticVariability, Sequencing, Alignment, SequenceMatching, DataImport

VignetteBuilder knitr

Suggests BiocStyle, knitr, rmarkdown, ggplot2

NeedsCompilation no

RoxygenNote 6.0.1

git_url https://git.bioconductor.org/packages/QSutils

git_branch master

git_last_commit 862558e

git_last_commit_date 2019-05-02

Date/Publication 2019-10-12

R topics documented:

`QSutils-package` *Quasispecies Diversity*

Description

Set of utility functions for viral quasispecies analysis with NGS data. Most functions are equally useful for metagenomic studies. There are three main types: (1) data manipulation and exploration—functions useful for converting reads to haplotypes and frequencies, repairing reads, intersecting strand haplotypes, and visualizing haplotype alignments. (2) diversity indices—functions to compute diversity and entropy, in which incidence, abundance, and functional indices are considered. (3) data simulation—functions useful for generating random viral quasispecies data.

Details

The DESCRIPTION file: This package was not yet installed at build time.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori i Font

Maintainer: Mercedes Guerrero-Murillo <mergumu@gmail.com>

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83.doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

Collapse *Collapse reads into haplotypes and frequencies*

Description

Collapse summarizes aligned reads into haplotypes with their frequencies. Recollapse is used to update the collapse after some type of manipulation may have resulted in duplicate haplotypes.

Usage

```
Collapse(seqs)
Recollapse(seqs,nr)
```

Arguments

seqs	DNAStringSet or AAStringSet object with the sequences to collapse.
nr	Vector with the haplotype counts.

Details

Recollapse is used when haplotypes may become equivalent after some type of manipulation. It removes duplicate sequences and updates their frequencies.

Value

Collapse and Recollapse return a list with two elements.

nr	Vector of the haplotype counts.
hseqs	DNAStringSet or AAStringSet with the haplotype sequence.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

Examples

```
# Load raw reads.
filepath<-system.file("extdata","Toy.GapsAndNs.fna", package="QSutils")
reads <- readDNAStringSet(filepath)

# Collapse this reads into haplotypes
lstCollapsed <- Collapse(reads)
lstCorrected<-CorrectGapsAndNs(lstCollapsed$hseqs[2:length(lstCollapsed$hseqs)],
               lstCollapsed$hseqs[[1]])
#Add again the most abundant haplotype.
lstCorrected<- c(lstCollapsed$hseqs[1],lstCorrected)
lstCorrected
# Recollapse the corrected haplotypes
lstRecollapsed<-Recollapse(lstCorrected,lstCollapsed$nr)
lstRecollapsed
```

ConsSeq *Consensus sequence given an alignment and frequencies*

Description

ConsSeq determines the consensus sequence from a set of haplotypes.

Usage

```
ConsSeq(seqs, w=NULL)
```

Arguments

seqs	DNAStringSet or AAStringSet object with the haplotype sequences.
w	An optional numeric vector with the haplotype counts.

Details

The most frequent nucleotide or amino acid at each position is taken. No IUPAC ambiguity codes are considered; in the case of ties, the consensus nucleotide is decided randomly.

Value

Character vector with the consensus sequence.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

ReadAmplSeqs

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

ConsSeq(lst$hseqs,lst$nr)
```

`CorrectGapsAndNs` *Function to correct an alignment with gaps and Ns*

Description

Corrects positions in a DNAStringSet or AAStringSet of aligned haplotypes, replacing gaps and Ns (indeterminates) with the nucleotide or amino acid from the corresponding position in the reference sequence.

Usage

```
CorrectGapsAndNs(hseqs, ref.seq)
```

Arguments

`hseqs`	DNAStringSet or AAStringSet object with the alignment to correct.
`ref.seq`	Character vector with the reference sequence of the alignment.

Value

DNAStringSet or AAStringSet object with the sequences corrected. Duplicate haplotypes may arise as a consequence of this operation. See `Recollapse`.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

See Also

`Recollapse`

Examples

```
# Create a random reference sequence.
ref.seq <-GetRandomSeq(50)
ref.seq

# Create an alignment with gaps and Ns.
symb <- c(".","-","N")
nseqs <- 12
```

```
p <- c(0.9,0.06,0.04)
hseqs <- matrix(sample(symb,50*nseqs,replace=TRUE,prob=p),ncol=50)
hseqs <- apply(hseqs,1,paste,collapse="")
hseqs
hseqs <- DNAStringSet(hseqs)

# Apply the function and visualize the result.
cseqs <- CorrectGapsAndNs(hseqs,as.character(ref.seq))
c(ref.seq,as.character(cseqs))
```

`DBrule` *Genotyping by the DB rule*

Description

Computes the nearest cluster to a given sequence.

Usage

```
DBrule(grpDist, hr, oDist, g.names = NULL)
```

Arguments

`grpDist`	Distances between reference sequences.
`hr`	Factor or a vector of integers that contains the type or subtype for each reference sequence.
`oDist`	Distance from the sequence to be classified to the reference sequences.
`g.names`	Type or subtype names to classify the sequence.

Value

List with three elements:

`Phi2`	Vector with the distances to each cluster.
`DB.rule`	The index of the nearest cluster.
`Type`	Name of the nearest cluster.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Caballero A, Gregori J, Homs M, Tabernero D, Gonzalez C, Quer J, Blasi M, Casillas R, Nieto L, Riveiro-Barciela M, Esteban R, Buti M, Rodriguez-Frias F. Complex Genotype Mixtures Analyzed by Deep Sequencing in Two Different Regions of Hepatitis B Virus. PLoS One. 2015 Dec 29;10(12):e0144816. doi: 10.1371/journal.pone.0144816. eCollection 2015. PubMed PMID: 26714168; PubMed Central PMCID: PMC4695080.

Examples

```
# Load haplotype to be genotyped.
filepath<-system.file("extdata","Unknown-Genotype.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")
hseq <- lst$hseq[1]

# Load genotype references.
filepath_geno<-system.file("extdata","GenotypeStandards_A-H.fas",
package="QSutils")
RefSeqs <- readDNAStringSet(filepath_geno)

# Compute pairwise distances.
dm <- as.matrix(DNA.dist(c(hseq,RefSeqs),model="K80"))

# Distances between genotype RefSeqs
dgrp <- dm[-1,-1]
grp <- factor(substr(rownames(dgrp),1,1))
hr <- as.integer(grp)

# Distance of the query haplotype to the reference sequence.
d <- dm[1,-1]

# Genotyping by the DB rule.
dsc <- DBrule(dgrp,hr,d,levels(grp))
dsc
```

Diverge *Generate a set of diverging haplotypes*

Description

Generates a set of diverging haplotypes from the given DNA sequence. The haplotypes produced share a pattern of divergence with an increasing number of mutations.

Usage

```
Diverge(vm, seq)
```

Arguments

vm	Vector with number of diverging mutations to be generated.
seq	Reference sequence from which to generate the variants.

Details

`max(vm)` Positions in the given sequence are randomly generated. A substitution is also randomly produced for each of these positions. A haplotype is generated for each element in `vm`, so that it contains `vm[i]` substitutions of those previously generated.

Value

Character string vector with the segregating haplotypes generated.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

GetRandomSeq

Examples

```
set.seed(123)
m1 <- GetRandomSeq(50)
hpl <- Diverge(3:6,m1)
DottedAlignment(DNAStringSet(hpl))
```

DNA.dist *Matrix of DNA distances given an alignment*

Description

Function to compute a matrix of pairwise distances from DNA sequences using a model of DNA evolution. It relies on the `dist.dna()` function in the APE package.

Usage

```
DNA.dist(seqs, model = "raw", gamma = FALSE, pairwise.deletion = FALSE)
```

Arguments

`seqs`	DNAStringSet object with the aligned haplotypes.
`model`	Evolutionary model to compute genetic distance by default "raw", but "N", "TS", "TV", "JC69", "K80", "F81", "K81", "F84", "BH87", "T92","TN93", "GG95", "logdet", "paralin", "indel", or "indelblock" can also be used.
`gamma`	Gamma parameter possibly used to apply a correction to the distances or FALSE (by default).
`pairwise.deletion`	A logical indicating whether to delete sites with missing data (gaps) in a pairwise manner. The default is to delete sites with at least one missing datum in all sequences.

Value

Object of class "dist" with pairwise distances.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Paradis E., Claude J. and Strimmer K., APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004, 20, 289-290

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

dist.dna

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

dst <- DNA.dist(lst$hseqs,model="N")
dst
```

DottedAlignment *Align haplotypes into a dotted alignment*

Description

Given an alignment, it takes the first sequence as reference, and depicts all equivalences in the alignment as dots, leaving only the variants with respect to the reference.

Usage

```
DottedAlignment(hseqs)
```

Arguments

hseqs	DNAStringSet or AAStringSet with haplotype sequences.

Value

A character string vector of the alignment, with dots in the conserved positions.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

ReadAmplSeqs

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")
strs <- DottedAlignment(lst$hseqs)

# Create a data frame to visualize the result.
data.frame(Hpl=strs,stringsAsFactors=FALSE)
```

DSFT *Downsampling followed by fringe trimming*

Description

Diversity indices are influenced to a greater or lesser degree by the sample size on which they are computed. This function helps to minimize the bias inherent to sample size. First the vector of abundances is scaled to a smaller sample size, then all haplotypes with abundances below a given threshold are excluded with high confidence.

Usage

```
DSFT(nr, size, p.cut = 0.002, conf = 0.95)
```

Arguments

nr	Vector of observed haplotype counts.
size	Size to downsample.
p.cut	Abundance threshold.
conf	Confidence in trimming.

Value

Vector of logicals, with false the haplotypes to be excluded.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Examples

```
# Generate viral quasispecies abundance data.
set.seed(123)
n <- 2000
y <- geom.series(n,0.8)+geom.series(n,0.0004)
nr.pop <- round(y*1e7)
# Get a sample of 10000 reads from this population.
sz2 <- 10000
nr.sz2 <- table(sample(length(nr.pop),size=sz2,replace=TRUE,p=nr.pop))
# Filter out haplotypes below 0.1%.
thr <- 0.1
fl <-  nr.sz2>=sz2*thr/100
nr.sz2 <- nr.sz2[fl]
Shannon(nr.sz2) #0.630521
# DSFT to 5000 reads.
sz1 <- 5000
fl <- DSFT(nr.sz2,sz1)
nr.sz2 <- nr.sz2[fl]
# Compute size corrected Shannon entropy.
Shannon(nr.sz2) #0.6189798
```

FAD *Functional attribute diversity*

Description

Computes the Functional Attribute Diversity as the sum of elements in the pairwise distance matrix.

Usage

```
FAD(dst)
```

Arguments

dst	A "dist" object or a symmetrical matrix with pairwise distances.

Value

A value that corresponds to the Functional Attribute Diversity. The sum of matrix elements.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

See Also

DNA.dist

Examples

```
# Create the object.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Compute the DNA distance matrix.
dst <- DNA.dist(lst$hseqs,model="N")

FAD(dst)
```

`fn.ab` *Vector of abundances with different methods*

Description

Function to simulate haplotype abundances in the quasispecies.

Usage

```
fn.ab(n, h = 10000, r = 0.5,fn="pcf")
```

Arguments

`n`	Number of counts to compute.
`h`	Highest abundance value.
`r`	A number to compute the abundance. See details.
`fn`	Character indicating which function to use to compute the abundances "pf","pcf" or "dfp", see details. By default "pcf".

Details

The abundances computed as a power of fractions, when fn is "pf", are computed according to the following equation, taking the integer part:

$$max(hr^{(i-1)}, 1);\ \ 0 < r < 1;\ \ i = 1..n$$

The lower `r`, the faster the decrease in abundance, `r` is in the range 0 < r < 1.

With "pcf" the abundances are computed by a power of decreasing fractions, as counts, according to the following equation, taking the integer part:

$$max(h\ \left(\frac{1}{i}\right)^{r}, 1);\ \ r > 0;\ \ i = 1..n$$

The higher `r`, the faster the decrease in abundances. In this case `r` corresponds to the power of the function, a value larger than 0, usually in the range 0.5 < r < 4.

If fn is equal to "dfp", the abundances are computed by increasing root powers according to the following equation,taking the integer part:

$$max(h^{(1/i)}, 1);\ \ i = 1..n$$

Value

Numeric vector with n decreasing counts, where the first element equals h, and no element is lower than 1.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

geom.series,GetRandomSeq, GenerateVars,Diverge

Examples

```
# Simulate a quasispecies alignment.
m1 <- GetRandomSeq(50)
v1 <- GenerateVars(m1,50,2,c(10,1))
qs <- c(m1,v1)
w_pf <- fn.ab(length(qs),h=1000,r=1.5,fn="pf")
w_pf
w_pcf <- fn.ab(length(qs),h=1000,r=1.5,fn="pcf")
w_pcf
w_dfp <- fn.ab(length(qs),h=1000,fn="dfp")
w_dfp
```

FreqMat *Matrix of nucleotide or amino acid frequencies in alignment by position*

Description

Computes the nucleotide or amino acid frequency at each position in the alignment.

Usage

```
FreqMat(seqs,nr=NULL)
```

Arguments

seqs	DNAStringSet or AAStringSet with the aligned haplotype sequences.
nr	An optional numeric vector with the haplotype counts.

Value

Matrix with the frequency of each nucleotide or amino acid in each position. A (4 x n) or (20 x n) matrix, where n is the alignment length.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Frequencies only in the alignment.
FreqMat(lst$hseqs)
# Also taking into account haplotype frequencies.
FreqMat(lst$hseqs,lst$nr)
```

GenerateVars *Generate variants of a given haplotype*

Description

Function to generate a set of variants for a given DNA sequence.

Usage

```
GenerateVars(seq, nhpl, max.muts, p.muts)
```

Arguments

seq	A character string with a DNA sequence from which to generate the variants.
nhpl	Number of haplotypes to generate.
max.muts	Maximum number of mutations in each sequence.
p.muts	Vector of length max.muts with the probability of each number of mutations, some of which may be 0.

Details

Given a DNA sequence, nhpl variant haplotypes are generated at random, with a maximum of max.muts substitutions each. The probability of the number of mutations in each haplotype generated is given by the vector p.muts. The positions of the mutations in each haplotype are independent and random.

Value

A character vector with nhpl haplotype variants of the seq sequence.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

GetRandomSeq, Diverge

Examples

```
set.seed(123)
m1 <- GetRandomSeq(50)

GenerateVars(m1,50,2,c(10,1))
```

GenotypeStandards_A-H.fas *Genotype standards of hepatitis B virus*

Description

Fasta file with a set of well characterized sequences belonging to each HBV genotype. See the QSutils vignette: `vignette("QSutils",package = "QSutils")`.

Format

Fasta file format. Each sequence starts with the symbol ">" followed by the sequence ID. Subsequent lines correspond to the nucleotide sequences or peptide sequences.

See Also

DBrule

Examples

```
filepath<-system.file("extdata","GenotypeStandards_A-H.fas", package="QSutils")
lstRefs <- ReadAmplSeqs(filepath,type="DNA")

RefSeqs <- lstRefs$hseq
```

---

geom.series *Geometric series*

Description

Function to simulate haplotype abundances in the quasispecies by geometric series.

Usage

```
geom.series(n,p=0.001)
```

Arguments

n	Number of frequencies to compute.
p	Numeric parameter of the geometric function.

Details

The abundances, as counts, are computed according to the following equation:

$$p\,(1-p)^{i-1},\;\; i = 1..n$$

The lower r, the faster the decrease in abundances.

Value

Numeric vector with n decreasing counts.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

GetRandomSeq, GenerateVars, Diverge

Examples

```
# Simulate a quasispecies alignment.
m1 <- GetRandomSeq(50)
v1 <- GenerateVars(m1,50,2,c(10,1))
qs <- c(m1,v1)
w <- geom.series(100,0.8)
```

GetInfProfile *Information content profile of an alignment*

Description

GetInfProfile computes the information content at each position of an alignment.

Usage

```
GetInfProfile(seqs,nr=NULL)
```

Arguments

seqs	DNAStringSet or AAStringSet with the haplotype alignment.
nr	An optional numeric vector with the haplotype counts to take into account the information content of each position in the alignment.

Value

Returns a numeric vector whose length is equal to the length of the alignment. Each value corresponds to the information content of each position in the alignment.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Examples

```
# Load the alignment.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Compute the alignment's IC profile.
GetInfProfile(lst$hseqs)
# Also taking into account haplotype frequencies.
GetInfProfile(lst$hseqs,lst$nr)
```

`GetQSData` *Read the aligned sequences, filter at minimum abundance, and sort the sequences*

Description

Reads aligned amplicon sequences with abundance data, filters at a given minimum abundance, and sorts by mutations and abundance.

Usage

```
GetQSData(flnm,min.pct=0.1,type="DNA")
```

Arguments

`flnm`	Fasta file with haplotype sequences and their frequencies. The header of each haplotype in the fasta file is composed of an ID followed by a vertical bar "|" followed by the read counts, and eventually followed by another vertical bar and additional information (eg, Hpl.2.0001|15874|25.2).
`min.pct`	Minimum abundance, in %, to filter the reads. Defaults to 0.1%.
`type`	Character string specifying the type of the sequences in the fasta file. This must be one of "DNA" or "AA". It is "DNA" by default.

Details

The fasta file is loaded and the haplotype abundances, as counts, are taken from the header of each sequence. Haplotypes with abundances below `min.pct` % are filtered out. The haplotypes are then sorted: first, by decreasing order of the number of mutations with respect to the dominant haplotype, and second, by decreasing order of abundances. The haplotypes are then renamed according to the pattern `Hpl.n.xxxx`, where `n` represents the number of mutations, and `xxxx` the abundance order within the mutation number.

Value

Returns a list with three elements.

`bseqs`	DNAStringSet or AAStringSet with the haplotype sequences.
`nr`	Vector of haplotype counts.
`nm`	Vector of number of mutations of each haplotype with respect to the dominant (most frequent) haplotype.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

See Also

ReadAmplSeqs

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst<-GetQSData(filepath,min.pct=0.1,type="DNA")
lst
```

GetRandomSeq *Generate a random sequence*

Description

Creates a random DNA sequence of a given length.

Usage

```
GetRandomSeq(seq.len)
```

Arguments

seq.len	The sequence length.

Value

A character string representing a DNA sequence.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

See Also

GenerateVars, Diverge

Examples

```
set.seed(123)
GetRandomSeq(50)
```

`GiniSimpson` *Functions to calculate the GiniSimpson index*

Description

GiniSimpson calculates the unbiased estimator, GiniSimpsonVar computes Gini-Simpson asymptotic variance, and GiniSimpsonMVUE calculates the minimum variance unbiased estimator of the Gini-Simpson index.

Usage

```
GiniSimpson(w)
GiniSimpsonMVUE(w)
GiniSimpsonVar(w)
```

Arguments

`w`	Vector of observed counts or frequencies.

Value

A value that corresponds to the Gini-Simpson diversity index.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Examples

```
# A vector of haplotype counts.
nr <- c(464, 62, 39, 27, 37, 16, 33, 54, 248, 20)

# Gini-Simpson index.
GiniSimpson(nr)

# Gini-Simpson variance.
GiniSimpsonVar(nr)

# MVUE Gini-Simpson index.
GiniSimpsonMVUE(nr)
```

HCq *Set of functions to compute the Havrda-Charvat estimator*

Description

HCq computes the Havrda-Charvat estimator, and HCqVar computes the Havrda-Charvat asymptotic variance for a given exponent. By using HCqProfile, a Havrda-Charvat estimator is calculated for a predefined vector of exponents to obtain the full profile in the range, 0 to Inf.

Usage

```
HCq(w, q)
HCqVar(w, q)
HCqProfile(w, q = NULL)
```

Arguments

w	Vector of observed counts or frequencies.
q	Exponent. By default, a vector of values 1, 2, 3, 4 and Inf.

Details

In HCq only the first element in q is considered. HCqProfile is vectorized and considers all elements in q. When q is NULL: in this case, a default vector is taken to obtain the full profile in the range 0 to Inf.

Value

A value that corresponds to the Havrda-Charvat estimator when HCq or HCqVar is used. A data frame with the Havrda-Charvat estimator for each exponent when HCqProfile is used.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Pavoine, S. (2005). M?thodes statistiques pour la mesure de la biodiversit?. UMR CNRS 5558 Biometrie et Biologie Evolutive.

See Also

Hill, Renyi

Examples

```
# A vector of observed counts.
nr<-c(464, 62, 39, 27, 37, 16, 33, 54, 248, 20)

# Havrda-Charvat estimator for q=4.
HCq(nr,4)

# Havrda-Charvat estimator variance for q=4.
HCqVar(nr,4)

# Prolife of Havrda-Charvat estimator for 0:4 and Inf.
HCqProfile(nr,c(0:4,Inf))

# Full profile.
HCqProfile(nr)
```

`Hill` *Hill numbers*

Description

Functions to compute Hill numbers. `Hill` computes the Hill number of a single q value. `HillProfile` computes Hill numbers for all elements in vector q.

Usage

```
Hill(w, q)
HillProfile(w, q = NULL)
```

Arguments

`w`	Vector of observed counts or frequencies.
`q`	Exponent.

Details

In `Hill` , only the first element in `q` is considered. `HillProfile` is vectorized and considers all elements in `q`.When `q` is NULL: in this case, a default vector is taken to obtain the full profile in the range, 0 to Inf.

Value

A value or vector of values corresponding to the Hill number estimators of passed exponents.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

HCq, Renyi

Examples

```
# Vector of observed counts.
nr<-c(464, 62, 39, 27, 37, 16, 33, 54, 248, 20)

# Hill numbers of order 2.
Hill(nr,2)

# Set of most common values.
HillProfile(nr,q=c(0:4,Inf))

# Full Hill numbers profile.
HillProfile(nr)
```

IntersectStrandHpls *Forward and reverse strand haplotype intersections*

Description

Computes the intersection of forward and reverse strand haplotypes after a previous abundance filter that removes strand haplotypes below a given frequency threshold or unique to a single strand.

Usage

```
IntersectStrandHpls(nrFW, hseqsFW, nrRV, hseqsRV, thr = 0.001)
```

Arguments

nrFW	Numeric vector with forward strand haplotype counts.
hseqsFW	DNAStringSet object with the forward strand haplotypes.
nrRV	Numeric vector with forward reverse strand haplotypes.
hseqsRV	DNAStringSet object with the reverse strand haplotypes.
thr	Threshold to filter haplotypes at minimum abundance.

Value

List object with this elements:

hseqs	DNAStringSet object with the forward and reverse strand intersected.
nr	Numeric vector with the abundance of each haplotype.
pFW	Vector of abundances of aligned forward strand.
pRV	Vector of abundances of aligned reverse strand.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

See Also

ReadAmplSeqs

Examples

```
# Load objects.
filepath_FW<-system.file("extdata","ToyData_FWReads.fna", package="QSutils")
FW<- ReadAmplSeqs(filepath_FW,type="DNA")
filepath_RV<-system.file("extdata","ToyData_RVReads.fna", package="QSutils")
RV<- ReadAmplSeqs(filepath_RV,type="DNA")

# Intersect the two objects, with a default threshold.
IntersectStrandHpls(FW$nr,FW$hseqs,RV$nr,RV$hseqs)
```

MutationFreq *Mutation frequency with respect to the dominant haplotype*

Description

MutationFreq computes the mutation frequency given a vector of counts, and the genetic distances of each haplotype to the dominant haplotype. MutationFreqVar returns the variance of the mutation frequency.

Usage

```
MutationFreq(dst=NULL,nm=NULL,nr=NULL,len=1)
MutationFreqVar(nm,nr=NULL,len=1)
```

Arguments

dst	A "dist" object or a symmetric matrix with pairwise distances.
nm	Vector of distances or differences with respect to the dominant haplotype including itself (eg, nm[1] is 0 if w[1]==max(w)).
nr	An optional numeric vector with the haplotype counts.
len	The alignment width when nm is the number of differences, otherwise 1. Defaults to 1.

Value

A value corresponding to the mutation frequency for MutationFreq or its variance for MutationFreqVar. When nr is NULL, the same weight is given to each haplotype and the computed value corresponds to the mutation frequency by entity.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

DNA.dist, GetQSData, ReadAmplSeqs

Examples

```
# Load alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- GetQSData(filepath,type="DNA")

# Mutation frequency.
dst <- DNA.dist(lst$seqs,model="raw")
MutationFreq(dst=dst,len=width(lst$seqs)[1])

# Mutation frequency with abundances.
MutationFreq(nm=lst$nm,nr=lst$nr,len=width(lst$seqs)[1])

# Variance of the mutation frequency.
MutationFreqVar(nm=lst$nm,nr=lst$nr,len=width(lst$seqs)[1])
```

MutsTbl *Table of mutation frequencies by position*

Description

Computes the table of mutation frequencies by position with respect to the alignment consensus.

Usage

```
MutsTbl(hseqs,nr=NULL)
```

Arguments

hseqs	DNAStringSet or AAStringSet with the aligned haplotype sequences.
nr	An optional numeric vector with the haplotype counts. When nr is NULL, the same weight is given to each haplotype.

Value

Matrix of mutation counts by position. A (4 x n) or (20 x n) matrix, where n is the alignment length.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

ReadAmplSeqs

Examples

```
# Load the haplotypes alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Table of mutations in the alignment, regardless of haplotype abundance.
MutsTbl(lst$hseqs)

# Table of mutations taking into account abundance.
MutsTbl(lst$hseqs,lst$nr)
```

NucleotideDiversity *Nucleotide diversity*

Description

Computes the mean pairwise genetic distance between sequences in the alignment.

Usage

```
NucleotideDiversity(dst,w=NULL)
```

Arguments

dst	A "dist" object or a symmetrical matrix with haplotype pairwise distances (ie, the output of DNA.dist).
w	An optional numeric vector with the haplotype counts. When w is NULL, the same weight is given to each haplotype, and nucleotide diversity is computed at the entity level.

Value

A value that corresponds to the nucleotide diversity, either by entity or abundance, depending on argument w.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

DNA.dist, ReadAmplSeqs

Examples

```
# Load haplotype alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Compute the DNA distance matrix.
dst <- DNA.dist(lst$hseqs,model="K80")

NucleotideDiversity(dst, lst$nr)
NucleotideDiversity(dst)
```

PolyDist *Fraction of substitutions by polymorphic site*

Description

Computes the fraction of substitutions at each polymorphic site. The wild-type base is taken as the most abundant at each site, taking into account the weights, w.

Usage

```
PolyDist(seqs,w=NULL)
```

Arguments

seqs	DNAStringSet or AAStringSet with the haplotype sequences.
w	An optional numeric vector with the haplotype counts. When w is NULL, the same weight is given to each haplotype.

Value

Vector of numbers corresponding to the fraction of substitutions at polymorphic sites. Note that the wild type also depends on w.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

ReadAmplSeqs

Examples

```
# Load haplotype alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

PolyDist(lst$hseqs)
PolyDist(lst$hseqs,lst$nr)
```

Rao *Functions to compute Rao's entropy*

Description

Set of functions to estimate Rao's functional entropy. `Rao` calculates the Rao entropy, `RaoVar` the variance of the Rao estimator, `RaoPow` the Rao entropy of order q, and `RaoPowProfile` the functional Rao entropy profile for the given set of exponents.

Usage

```
Rao(dst, w=NULL)
RaoVar(dst,w=NULL)
RaoPow(dst,q,w=NULL)
RaoPowProfile(dst,w=NULL,q=NULL)
```

Arguments

`dst`	A "dist" object, output of the DNA.dist function.
`w`	An optional numeric vector with the haplotype counts. When `w` is NULL the same weight is given to each haplotype.
`q`	Exponent. A single value for `Rao`, `RaoVar` and `RaoPow`. A vector of values for `RaoPowProfile`. The default value for `RaoPowProfile` is a set of exponents to obtain a smooth profile.

Value

A single value for `Rao`, `RaoVar` and `RaoPow`. A vector of values for `RaoPowProfile` corresponding to each exponent in vector q.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Pavoine, S. (2005). Méthodes statistiques pour la mesure de la biodiversité. UMR CNRS 5558 «Biométrie et Biologie Evolutive».

See Also

`DNA.dist`, `ReadAmplSeqs`

Examples

```
# Load haplotype alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")
# DNA pairwise distances.
dst <- DNA.dist(lst$hseqs,model="N")

Rao(dst,lst$nr)
RaoVar(dst,lst$nr)
RaoPow(dst,2,lst$nr)
RaoPowProfile(dst,lst$nr,c(0:4,Inf))
RaoPowProfile(dst,lst$nr)
```

ReadAmplSeqs *Read a fasta file with haplotypes and frequencies*

Description

Loads an alignment of haplotypes and their frequencies from a fasta file.

Usage

```
ReadAmplSeqs(flnm,type="DNA")
```

Arguments

flnm	File name of a fasta file with haplotype sequences and their frequencies. The header of each haplotype in the fasta file is composed of an ID followed by a vertical bar "|" followed by the read count, and eventually followed by another vertical bar and additional information (eg, Hpl.2.0001|15874|25.2).
type	Character string specifying the types of sequences in the fasta file. This must be either "DNA" or "AA". It is "DNA" by default.

Value

Returns a list with two elements:

nr	Vector of the haplotype counts.
hseqs	DNAStringSet or AAStringSet with the haplotype DNA sequences or amino acid sequences.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

See Also

GetQSData

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")
lst
```

`Renyi` *Rényi profiles*

Description

Functions to compute the Rényi entropy given a vector of counts `RenyiProfile` computes the Rényi number for a set of exponents.

Usage

```
Renyi(w, q)
RenyiProfile(w, q = NULL)
```

Arguments

`w`	Vector of observed counts or frequencies.
`q`	Exponent. A single value for `Renyi`, a vector of values or NULL for `RenyiProfile`.

Value

A single value for `Renyi`. A data frame with exponents and entropies for `RenyiProfile`.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Pavoine, S. (2005). Méthodes statistiques pour la mesure de la biodiversité. UMR CNRS 5558 «Biométrie et Biologie Evolutive».

See Also

Hill, HCq

Examples

```
# A vector of observed counts.
nr<-c(464, 62, 39, 27, 37, 16, 33, 54, 248, 20)

Renyi(nr,2)

RenyiProfile(nr,c(0:4,Inf))

RenyiProfile(nr)
```

ReportVariants *Report variants*

Description

Reports the variants of a DNAStringSet or AAStringSet of haplotypes given a reference sequence.

Usage

```
ReportVariants(hseqs,ref.seq,nr=NULL,start=1)
```

Arguments

hseqs	DNAStringSet or AAstringSet object of the aligned haplotypes.
ref.seq	Character vector with the reference sequence of the alignment.
nr	Numeric vector with the abundances of each haplotype in hseqs. When nr is NULL, a vector of ones is taken as default.
start	Position of the first nucleotide in the alignment

Value

A dataframe with 4 columns: the nucleotide in the reference sequence, the position, the variant nucleotide, and its abundance.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

Examples

```
# Load objects.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Report the variants in these haplotypes,
# taking as a reference the most abundant haplotype.
ReportVariants(lst$hseqs[-1], ref.seq= as.character(lst$hseqs[1]),
lst$nr[-1], start = 1)
```

SegSites *Compute the number of segregating sites*

Description

Computes the number of segregating (polymorphic) sites in a given alignment. That is, the number of sites with more than a single nucleotide or amino acid in the alignment.

Usage

```
SegSites(seqs)
```

Arguments

seqs	DNAStringSet or AAStringSet with the haplotype sequences.

Value

A value corresponding to the number of polymorphic sites.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

ReadAmplSeqs

Examples

```
# Create the object.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

SegSites(lst$hseqs)
```

Shannon *Set of functions to compute Shannon entropy*

Description

`Shannon` computes the Shannon entropy. `NormShannon` returns the normalized Shannon entropy. `ShannonVar` computes the Shannon entropy asymptotic variance. `NormShannonVar` computes the normalized Shannon entropy asymptotic variance.

Usage

```
Shannon(w)
ShannonVar(w)
NormShannon(w)
NormShannonVar(w)
```

Arguments

`w`	Vector of observed counts or frequencies.

Value

A single value with the result of the computations.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

Examples

```
# Create a vector of observed counts.
nr<-c(464, 62, 39, 27, 37, 16, 33, 54, 248, 20)

# Shannon entropy.
Shannon(nr)

# Shannon entropy variance.
ShannonVar(nr)

# Normalized Shannon entropy.
NormShannon(nr)

# Normalized Shannon entropy variance.
NormShannonVar(nr)
```

`SortByMutations` *Sort haplotypes by mutations and abundance*

Description

Sorts and renames haplotypes by the number of mutations with respect to the dominant haplotype, and by abundance.

Usage

```
SortByMutations(bseqs, nr)
```

Arguments

`bseqs`	DNAStringSet or AAStringSet object with the haplotype alignment.
`nr`	Vector with the haplotype counts.

Details

The haplotypes are pairwise-aligned to the dominant haplotype and then sorted: first, by decreasing order of the number of differences with respect to the dominant haplotype, and second, by decreasing order of abundance. As a result, haplotypes are renamed according to the pattern `Hpl.n.xxxx`, where `n` represents the number of differences, and `xxxx` the abundance order within the mutation number.

Value

Returns a list with three elements.

bseqs	DNAStringSet or AAStringSet with the haplotype sequences.
nr	Vector of the haplotype counts.
nm	Vector of the number of differences of each haplotype with respect to the dominant haplotype.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

ReadAmplSeqs

Examples

```
# Load haplotype alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

SortByMutations(lst$hseq,lst$nr)
```

SummaryMuts *Distribution of nucleotides or amino acids in polymorphic sites*

Description

Computes the nucleotide or amino acid frequencies at all polymorphic sites in the alignment.

Usage

```
SummaryMuts(seqs, w = NULL, off = 0)
```

Arguments

seqs	DNAStringSet or AAStringSet with the haplotype sequences.
w	An optional numeric vector with the haplotype counts. When w is NULL, a vector of ones is taken as default.
off	Offset of first position in the alignment.

Value

Data frame with the polymorphic positions and nucleotide or amino acid frequencies.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One. 2013 Dec 31;8(12):e83361. doi: 10.1371/journal.pone.0083361. eCollection 2013. PubMed PMID: 24391758; PubMed Central PMCID: PMC3877031.

Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodriguez-Frías F. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Res. 2013 May;98(2):273-83. doi: 10.1016/j.antiviral.2013.03.007. Epub 2013 Mar 20. PubMed PMID: 23523552.

See Also

ReadAmplSeqs

Examples

```
# Load haplotype alignment with abundances.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

# Distribution of nucleotides at polymorphic sites.
SummaryMuts(lst$hseqs,lst$nr,off=0)
```

TotalMutations *Number of Mutations*

Description

TotalMutations computes the number of mutations in the alignment.

Usage

```
TotalMutations(hseqs,w)
```

Arguments

hseqs	DNAStringSet or AAStringSet with the haplotype sequences.
w	An optional numeric vector with the haplotype counts used to compute the total number of mutations in the population, that is, taking into account haplotype abundances. When w is NULL, a vector of ones is taken as default.

Value

A value corresponding to the number of mutations. Note that the wild-type is decided taking `w` into account.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

`SegSites`

Examples

```
# Create the object.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

TotalMutations(lst$hseqs)
TotalMutations(lst$hseqs,lst$nr)
```

`Toy.GapsAndNs.fna` *Fasta file with raw reads with gaps and Ns*

Description

Fasta file of sequenced data with some missing information. This is toy data to illustrate some functions of the package `QSutils` package.

Format

Fasta file format. Each sequence starts with the symbol ">" followed by the sequence ID. Subsequent lines correspond to the nucleotide sequences or peptide sequences.

See Also

`Collapse`, `CorrectGapsAndNs` and `Recollapse`

Examples

```
filepath<-system.file("extdata","Toy.GapsAndNs.fna", package="QSutils")
reads <- readDNAStringSet(filepath)

lstCollapsed <- Collapse(reads)
DottedAlignment(lstCollapsed$hseqs)
lstCorrected<-CorrectGapsAndNs(lstCollapsed$hseqs[2:length(lstCollapsed$hseqs)],
               lstCollapsed$hseqs[[1]])
lstCorrected<- c(lstCollapsed$hseqs[1],lstCorrected)
lstCorrected
lstRecollapsed<-Recollapse(lstCorrected,lstCollapsed$nr)
lstRecollapsed
```

ToyData_10_50_1000.fna

Fasta file with 10 haplotypes, 50 basepairs in size.

Description

Fasta file that contains the sequence of 10 haplotypes used as examples in the `QSutils` package.

Format

Fasta file format. Each sequence starts with the symbol ">" followed by the sequence ID. Subsequent lines correspond to the nucleotide sequences or peptide sequences.

Examples

```
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")
lst
```

ToyData_FWReads.fna *Fasta file with forward reads*

Description

Fasta file with forward strand reads. Toy data used to illustrate the intersections of forward and reverse haplotypes with the function `IntersectStrandHpls`.

Format

Fasta file format. Each sequence starts with the symbol ">" followed by the sequence ID. Subsequent lines correspond to the nucleotide sequences or peptide sequences.

See Also

`ToyData_RVReads.fna`, `IntersectStrandHpls`

Examples

```
filepath_FW<-system.file("extdata","ToyData_FWReads.fna", package="QSutils")
lstFW <- ReadAmplSeqs(filepath_FW,type="DNA")
filepath_RV<-system.file("extdata","ToyData_RVReads.fna", package="QSutils")
lstRV <- ReadAmplSeqs(filepath_RV,type="DNA")

lstI <- IntersectStrandHpls(lstFW$nr,lstFW$hseqs,lstRV$nr,lstRV$hseqs)
lstI
```

ToyData_RVReads.fna *Fasta file with reverse reads.*

Description

Fasta file with reverse strand reads. Toy data used to illustrate the intersections of forward and reverse haplotypes with the function `IntersectStrandHpls`.

Format

Fasta file format. Each sequence starts with the symbol ">" followed by the sequence ID. Subsequent lines correspond to the nucleotide sequences or peptide sequences.

See Also

`ToyData_FWReads.fna`, `IntersectStrandHpls`

Examples

```
filepath_FW<-system.file("extdata","ToyData_FWReads.fna", package="QSutils")
lstFW <- ReadAmplSeqs(filepath_FW,type="DNA")
filepath_RV<-system.file("extdata","ToyData_RVReads.fna", package="QSutils")
lstRV <- ReadAmplSeqs(filepath_RV,type="DNA")

lstI <- IntersectStrandHpls(lstFW$nr,lstFW$hseqs,lstRV$nr,lstRV$hseqs)
lstI
```

UniqueMutations *Number of unique mutations*

Description

`UniqueMutations` computes the number of unique mutations in the alignment.

Usage

```
UniqueMutations(hseqs)
```

Arguments

`hseqs`	DNAStringSet or AAStringSet with the haplotype sequences.

Value

A value corresponding to the number of mutations.

Author(s)

Mercedes Guerrero-Murillo and Josep Gregori

References

Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016 Jun;493:227-37. doi: 10.1016/j.virol.2016.03.017. Epub 2016 Apr 6. Review. PubMed PMID: 27060566.

Gregori J, Salicrú M, Domingo E, Sanchez A, Esteban JI, Rodríguez-Frías F, Quer J. Inference with viral quasispecies diversity indices: clonal and NGS approaches. Bioinformatics. 2014 Apr 15;30(8):1104-1111. Epub 2014 Jan 2. PubMed PMID: 24389655.

See Also

TotalMutations

Examples

```
# Create the object.
filepath<-system.file("extdata","ToyData_10_50_1000.fna", package="QSutils")
lst <- ReadAmplSeqs(filepath,type="DNA")

UniqueMutations(lst$hseqs)
```

Unknown-Genotype.fna *Fasta file with reads of unknown genotype*

Description

Fasta file with hepatitis B virus sequences of unknown genotype. This is used to illustrate the genotyping of HBV sequences with the QSutils package.

Format

Fasta file format. Each sequence starts with the symbol ">" followed by the sequence ID. Subsequent lines correspond to the nucleotide sequences or peptide sequences.

See Also

DBrule

Examples

```
filepath<-system.file("extdata","Unknown-Genotype.fna", package="QSutils")
lst2Geno <- ReadAmplSeqs(filepath,type="DNA")
hseq <- lst2Geno$hseq[1]
hseq
```

Index