

R documentation

of ‘/Users/zhangpengfan/Documents/phd’ etc.

February 22, 2021

R topics documented:

abd_prob	1
consis_err	2
Contam_detect	3
error_m	3
kmer_dist	4
loessErr	4
Rbec	5
trans_m	6

Index	7
--------------	----------

abd_prob	<i>Reference-based error correction of amplicon sequencing data</i>
----------	---

Description

This function calculates the abundance probabilities for each reads using poisson distribution

Usage

```
abd_prob(derep, ref, error_matrix)
```

Arguments

derep	dereplicated reads (Ns are not allowed in the reads)
ref	the unique reference sequences of the reference sequences, each sequence must be in one line (Ns are not allowed in the sequences)
error_matrix	The error matrix from the former iteration

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

Returns the lambda value and pvalue for each reads

Author(s)

Pengfan Zhang

consis_err	<i>Reference-based error correction of amplicon sequencing data</i>
------------	---

Description

This function iterates the error matrix till reaching the stable stage

Usage

```
consis_err(fq, derep, ref, lambda_out, sampling_size, ascii, min_E=0.05, min_P=1e-40, max_diff_abs=
```

Arguments

fq	the path of the fastq file (Ns are not allowed in the reads)
derep	the dereplicated reads by dada2 function
ref	the reference sequences, each sequence should take up one line (Ns are not allowed in the reads)
lambda_out	the matrix containg lambda value and pvalue from the former iteration
sampling_size	the subsampling size of the reads
ascii	ascii characters used to encode phred scores
min_E	the minimum expectation value of the Poission distribution for detecting paralogs within the same strain
min_P	the P value cutoff for identifying erroneous reads
max_diff_abs	the maximum absolute difference in number of corrected reads between two iterations, together with max_diff_ratio, before jumping out of the iteration
max_diff_ratio	the maximum difference in the percentages of corrected reads between two iterations

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

Returns the final files

Author(s)

Pengfan Zhang

Contam_detect

*Reference-based error correction of amplicon sequencing data***Description**

This function is designed for predicting the contaminated samples

Usage

```
Contam_detect(log_file, outdir, outlier_constant=1.5)
```

Arguments

log_file	the file contains a list of log files of each sample outputted with Rbec function
outdir	output directory
outlier_constant	the multiplier of variance to define the outlier

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

Returns a plot showing the distribution of percentage of corrected reads across the whole sample set and a summary file recording which samples might be contaminated

Author(s)

Pengfan Zhang

error_m

*Reference-based error correction of amplicon sequencing data***Description**

This function calculate the error matrix

Usage

```
error_m(fq, ref, sample_size, threads, ascii)
```

Arguments

fq	the path of merged amplicon sequencing reads in fastq format (Ns are not allowed in the reads)
ref	the unique reference sequences of the reference sequences, each sequence must be in one line (Ns are not allowed in the sequences)
sample_size	the sampling size of reads to generate the transition matrix
threads	the number of threads used to align the query reads to reference sequences
ascii	ascii characters used to encode phred scores

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

The output is a 20 by 43 transition probability matrix

Author(s)

Pengfan Zhang

kmer_dist	<i>DADA2</i>
-----------	--------------

Description

Calculate the kmer distance between two sequences. This function directly copied from DADA2.

Usage

```
kmer_dist(s1, s2, kmer_size)
```

Arguments

s1	A character(1) of DNA sequence 1.
s2	A character(1) of DNA sequence 2.
kmer_size	Kmer size.

Value

The kmer distance between two sequences

loessErr	<i>Reference-based error correction of amplicon sequencing data</i>
----------	---

Description

This function fits the loess regression to the error matrix

Usage

```
loessErr(trans, min_err_rate=1e-07)
```

Arguments

trans	the transition matrix
min_err_rate	the minimum transition probability for each substitution or insertion case

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

Returns the loess fitted error matrix

Author(s)

Pengfan Zhang

Rbec

Reference-based error correction of amplicon sequencing data

Description

This function corrects the amplicon sequencing data from synthetic communities where the reference sequences are known a priori

Usage

```
Rbec(fastq, reference, outdir, threads=1, sampling_size=5000, ascii=33, min_cont_abs=0.03)
```

Arguments

fastq	the path of the fastq file containing merged amplicon sequencing reads (Ns are not allowed in the reads)
reference	the path of the unique reference sequences, each sequence must be in one line (Ns are not allowed in the sequences)
outdir	the output directory, which should be created by the user
threads	the number of threads used, default 1
sampling_size	the sampling size for calculating the error matrix, default 5000
ascii	ascii characters used to encode phred scores (33 or 64), default 33
min_cont_abs	the relative abundance of unique tags for detecting contamination sequences that can't be corrected by any of the references

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

lambda_final.out the lambda value and pvalue of the Poisson distribution for each read
 error_matrix_final.out the error matrix in the final iteration
 strain_table.txt the strain composition of the sample
 contamination_seq.fna the potential sequences generated by contaminants
 rbec.log percentage of corrected reads, which can be used to predict contaminated samples

Author(s)

Pengfan Zhang

Examples

```
fastq <- system.file("extdata", "test_raw_merged_reads.fastq", package = "Rbec")  
ref <- system.file("extdata", "test_ref.fasta", package = "Rbec")  
Rbec(fastq=fastq, reference=ref, outdir=".", threads=1, sampling_size=500, ascii=33)
```

trans_m

Reference-based error correction of amplicon sequencing data

Description

This function count the transition matrix

Usage

```
trans_m(query, ascii)
```

Arguments

query	list containing subsampled amplicon sequencing reads, quality scores, and reference sequences showing the highest identity for each read (Ns are not allowed in the reads)
ascii	ascii characters used to encode phred scores

Details

Ruben Garrido-Oter's group, Plant-Microbe interaction, Max Planck Institute for Plant Breeding Research

Value

The output is a 20 by 43 matrix containing the counts for different kinds of transitions

Author(s)

Pengfan Zhang

Index

abd_prob, [1](#)

consis_err, [2](#)

Contam_detect, [3](#)

error_m, [3](#)

kmer_dist, [4](#)

loessErr, [4](#)

Rbec, [5](#)

trans_m, [6](#)