Package 'Rbec'

November 6, 2024

Type Package

Title Rbec: a tool for analysis of amplicon sequencing data from synthetic microbial communities

Version 1.14.0

Description Rbec is a adapted version of DADA2 for analyzing amplicon sequencing data from synthetic communities (SynComs), where the reference sequences for each strain exists. Rbec can not only accurately profile the microbial compositions in SynComs, but also predict the contaminants in SynCom samples.

License LGPL-3

Imports Rcpp (>= 1.0.6), dada2, ggplot2, readr, doParallel, foreach, grDevices, stats, utils

LinkingTo Rcpp

RoxygenNote 7.1.1

biocViews Sequencing, MicrobialStrain, Microbiome

Suggests knitr, rmarkdown

VignetteBuilder knitr

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

git_branch RELEASE_3_20

git_last_commit 5acc5f9

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-05

Author Pengfan Zhang [aut, cre]

Maintainer Pengfan Zhang <pzhang@mpipz.mpg.de>

Contents

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

Examples

```
#log_file <- system.file("extdata", "rbec_test.list", package = "Rbec")
log_path <- list.files(paste(path.package("Rbec"),
 "extdata/contamination_test", sep="/"),
 recursive=TRUE, full.names=TRUE)
log_file <- tempfile()
writeLines(log_path, log_file)
Contam_detect(log_file, tempdir())
```

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_obs_abd=200, min_
```

Arguments

`fastq`	the path of the fastq file containg 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_obs_abd`	the minimum oberseved abundace of unique tags for detecting contamination sequences, default 200
`min_cont_abd`	the relative abundance of unique tgas for detecting contamination sequences that can't be corrected by any of the references, default 0.03
`min_E`	the minimum expectation of the Possion distribution for the identification of paralogues, default 0.05
`min_P`	the minimum P value threshold of the Possion distribution to correct a read, default 1e-40
`ref_seeker`	the method for finding the candidate error-producing reference sequence for a tag showing identical lowest K-mer distance to multiple references. 1 for the abundance-based method; 2 for the transition probability-based method, default 1.
`cn`	the copy number table documenting the copy number of the marker gene in each strain. Rbec will normalize the strain abundance if the copy number is available

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

strain_table_normalized.txt the copy-number-normalized strain composition of the sample if the copy number table is provided

contamination_seq.fna the potential sequences generated by contaminants

rbec.log percentage of corrected reads, which can be used to predict contaminated samples

paralogue_seq.fna paralogue sequences found in each strain except for the reference provided

Author(s)

Pengfan Zhang

Examples

```
fastq <- system.file("extdata", "test_raw_merged_reads.fastq.gz", package = "Rbec")

ref <- system.file("extdata", "test_ref.fasta", package = "Rbec")

Rbec(fastq=fastq, reference=ref, outdir=tempdir(), threads=1, sampling_size=500, ascii=33)
```

Index