Package 'SNVLFDR'

January 20, 2025

Title Empirical Bayes Single Nucleotide Variant Calling

Version 1.0.1

Author Ali Karimnezhad [aut, cre, ctb]

Maintainer Ali Karimnezhad <ali.karimnezhad@gmail.com>

Description Identifies single nucleotide variants in next-generation sequencing data by estimating their local false discovery rates. For more details, see Karimnezhad, A. and Perkins, T. J. (2024) <doi:10.1038/s41598-024-51958-z>.

Encoding UTF-8

License GPL (>= 3)

RoxygenNote 7.2.3

NeedsCompilation no

Repository CRAN

Date/Publication 2024-01-25 13:30:02 UTC

Contents

`get_LFDRs` *Estimates LFDR values per genomic site*

Description

Based on a given read count matrix, identifies single nucleotide variants (SNVs) by estimating local false discovery rates (LFDRs). Users can set an initial value for the proportion of non-mutant sites and specify thresholds for allele frequency, read depth and LFDR cut-off value.

Usage

```
get_LFDRs(
  bam_input,
  bedfile,
  BQ.T,
  MQ.T,
  pi0.initial,
  AF.T,
  DP.T,
  LFDR.T,
  error,
  method,
  epsilon
)
```

Arguments

bam_input	Path to an input BAM file. The file must be in the format of a csv file generated by bam-readcount (https://github.com/genome/bam-readcount). See Examples.
bedfile	Path to a bed file containing genomic regions of interest. This file has to have three columns with chr# in the first column and start and end positions in the second and third columns respectively. No headers.
BQ.T	Minimum base call quality threshold. It can be set to 0 to include all sites. Otherwise, this threshold excludes sites with a base call quality below the specified threshold. It is recomended to set it to 20.
MQ.T	Minimum mapping quality threshold. It can be set to 0 to include all sites. Otherwise, this threshold excludes sites with a mapping quality below the specified threshold. It is recomended to set it to 20.
pi0.initial	Initial value for the proportion of non-mutant sites. It can be any number between 0 and 1. However it is recommended to set it to a number between 0.95 and 0.99 for more accuracy. If no value is specified, it will be set to 0.95 by default.
AF.T	Allele frequency threshold. It can be set to 0 to include all sites. Otherwise, this threshold excludes sites with an allele frequency below the specified threshold. If no value is specified, it will be set to 0.01 by default.
DP.T	Read depth threshold. It can be set to 0 to include all sites. Otherwise, this threshold excludes sites with a read depth below the specified threshold.
LFDR.T	A number between 0 and 1. It can be set to 0 to include all sites. Otherwise, this threshold excludes sites with an estimated LFDR below the specified threshold. If no value is specified, it will be set to 0.01 by default.
error	Error rate between 0 and 1. If it is set to NULL, a weighted average of average base call quality and average mapping quality per site will be calculated. Otherwise, it may be set to 0.01 or a desired error vector can be introduced by the user.

method	Method used to estimate pi0 and LFDRs. It can be "empirical", "uniform_empirical" or "uniform". If no method is specified, it will be set to "empirical" by default (recommended).
epsilon	The difference between old and new estimates of pi0 used for convergence. If no value is specified, it will be set to 0.01 by default.

Value

A list. Slot **estimated.pi0** returns estimated proportion of non-mutant sites. Slot **estimated.LFDRs** returns estimated LFDRs for genomic sites that were not filtered out. Slot **filtered.bam** adds estimated LFDRs, model errors and a mutant variable (indicating whether each site is detected to be a mutant (1) or non-mutant (0) site) to the filtered input file .

References

Karimnezhad, A. and Perkins, T.J. (2024). Empirical Bayes single nucleotide variant calling for next-generation sequencing data. Scientific Reports 14, 1550, <doi:10.1038/s41598-024-51958-z>

Examples

```
bam_input <- system.file("extdata", "bam_input.csv", package="SNVLFDR")
bedfile <- system.file("extdata", "regions.bed", package="SNVLFDR")
BQ.T=20
MQ.T=20
pi0.initial=0.95
AF.T=0.01
DP.T=10
LFDR.T=0.01
error=NULL
method='empirical'
epsilon=0.01
output=get_LFDRs(bam_input,bedfile,BQ.T,MQ.T,pi0.initial,AF.T,DP.T,LFDR.T,error,method,epsilon)
```

---

get_LFDRs_given_caller

Estimates LFDR values per mutant site detected by a desired variant caller

---

Description

Based on a given read count matrix and a list of calls made by a desired variant caller, estimates LFDRs that corresponds to each genomic site. It further classifies sites to either mutant or non-mutant sites by comparing their estimated LFDRs with an LFDR cut-off value. The cut-off value as well as error rates can be defined by users.

Usage

```
get_LFDRs_given_caller(bam_input, calls, LFDR.T, error)
```

Arguments

bam_input	Path to an original BAM file used to call variants. The file must be in the format of a csv file generated by bam-readcount (https://github.com/genome/bam-readcount). See Examples.
calls	Path to a vcf file generated by a variant caller. The first and second columns of this file have to be CHR names and positions, respectively.
LFDR.T	A number between 0 and 1. It can be set to 0 to include all sites. Otherwise, this threshold excludes sites with an LFDR below the specified threshold. If no value is specified, it will be set to 0.01 by default.
error	Error rate between 0 and 1. If it is set to NULL, a weighted average of average base call quality and average mapping quality per site will be calculated. Otherwise, it may be set to 0.01 or a desired error vector can be introduced by the user.

Value

A list. Slot **estimated.LFDRs** returns estimated LFDRs for all sites in the input file. Slot **updated.bam** adds estimated LFDRs, model errors and a mutant variable (indicating whether each site is detected to be a mutant (1) or non-mutant (0) site) to the input bam file.

References

Karimnezhad, A. and Perkins, T.J. (2024). Empirical Bayes single nucleotide variant calling for next-generation sequencing data. Scientific Reports 14, 1550, <doi:10.1038/s41598-024-51958-z>

Examples

```
bam_path <- system.file("extdata", "bam_input.csv", package="SNVLFDR")
calls_path <- system.file("extdata", "calls.vcf", package="SNVLFDR")
output=get_LFDRs_given_caller(bam_input=bam_path,calls=calls_path,LFDR.T=0.01,error=NULL)
```

Index