Package 'WSH'

January 8, 2020

Type Package

Title R package for the calculation of the following Within-Sample Heterogeneity Scores in Bisul-
fite Sequencing Data: FDRP, qFDRP, PDR, Epipolymorphism, Methylation Entropy and MHL

Version 0.1.3

Author Michael Scherer

Maintainer Michael Scherer <mscherer@mpi-inf.mpg.de>

Description R package for the calculation of the following Within-
Sample Heterogeneity Scores in Bisulfite Sequencing Data:
Fraction of Discordant Read Pairs (FDRP), quantitative FDRP (qFDRP),
Proportion of Discordant Reads (PDR, Landau et.al. (2014), Cancer Cell), Epipolymor-
phism (Landan et.al. (2012), Nature Genetics),
Methylation Entropy (Xie et.al. (2011), Nucleic Acids Research) and Methylation Haplo-
type Load (MHL, Guo et.al. (2017), Nature Genetics)

License GPL-3

Encoding UTF-8

LazyData true

Date 2020-01-08

Depends RnBeads,
GenomicAlignments,
Rsamtools,
parallel,
rtracklayer,
doParallel,
RUnit,
RnBeads.hg38

RoxygenNote 7.0.2

R topics documented:

---

calculate.entropy   *calculate.entropy*

Description

This function calculates Methylation entropy by calling the methclone software to compute epiallele
frequencies and then uses the definition of entropy for the calculation.

Usage

```
calculate.entropy(bam.file, out.folder = getwd(), out.name = "methclone")
```

Arguments

bam.file	path to bam file containing the reads of the data set
out.folder	folder to store the temporary file produced by methclone
out.name	name of temporary file

Value

a data frame containing the position in the reference genome and the Entropy scores

```
calculate.epipolymorphism
```
calculate.epipolymorphism

Description

This function calculates Epipolymorphism by calling the methclone software to compute epiallele frequencies and then uses the definition of epipolymorphism for the calculation.

Usage

```
calculate.epipolymorphism(
  bam.file,
  out.folder = getwd(),
  out.name = "methclone"
)
```

Arguments

bam.file	path to bam file containing the reads of the data set
out.folder	folder to store the temporary file produced by methclone
out.name	name of temporary file

Value

a data frame containing the position in the reference genome and the Epipolymorphism scores

```
calculate.fdrp
```
calculate.fdrp

Description

This function calculates the FDRP scores for the reads given in the bam files for the CpGs present in the annotation.

Usage

```
calculate.fdrp(
  bam.file,
  anno,
  log.path = getwd(),
  cores = 1,
  window.size = get.option("WINDOW.SIZE")
)
```

Arguments

`bam.file`	bath to the bam file to be analyzed already aligned to a reference genome
`anno`	annotation as a GRanges object with the CpG sites to be analyzed
`log.path`	location of the log file
`cores`	number of cores available for the analysis
`window.size`	window size used to restrict the concordance/discordance classification of each read pair DEFAULT: 50 as the maximum distance

Value

FDRP scores for the given annotation.

Author(s)

Michael Scherer

---

calculate.fdrp.by.chromosome

> *calculate.fdrp.by.chromosome This function calculates the FDRP scores for the reads given in the bam files for the CpGs present in the annotation.*

Description

calculate.fdrp.by.chromosome This function calculates the FDRP scores for the reads given in the bam files for the CpGs present in the annotation.

Usage

```
calculate.fdrp.by.chromosome(bam, anno)
```

Arguments

`bam`	bam file with the reads from a bisulfite sequencing technique already aligned to a reference genome
`anno`	annotation as a GRanges object with the CpG sites to be analyzed

Details

This function is called by calculate.fdrps for each chromosome separately and calls toCpG as well as calculate.fdrp.site.

Value

vector of fdrp scores for the given CpG sites in anno

Author(s)

Michael Scherer

---

calculate.fdrp.score   *calculate.fdrp.score*

Description

This function calculates the FDRPs/qFDRPs (depending on how the options are set) for all CpG sites in the annotation from the reads provided by the bam file.

Usage

```
calculate.fdrp.score(
  bam.file,
  anno,
  log.path = getwd(),
  cores = 1,
  window.size = get.option("WINDOW.SIZE")
)
```

Arguments

bam.file	bath to the bam file to be analyzed already aligned to a reference genome
anno	CpG sites to be analyzed as a GRanges object
log.path	location of the log file
cores	number of cores available for the analysis
window.size	window size used to restrict the concordance/discordance classification of each read pair DEFAULT: 50 as the maximum distance

Value

FDRP scores for the given annotation.

Author(s)

Michael Scherer

---

calculate.fdrp.site   *calculate.fdrp.site This function compute the FDRP score for a given cpg site.*

Description

calculate.fdrp.site This function compute the FDRP score for a given cpg site.

Usage

```
calculate.fdrp.site(pos, cpg, reads, site)
```

Arguments

pos   index of the cpg site

cpg   list containing the mapping from each cpg site to the reads that contain this site

reads   GRanges object containing the reads needed for the calculation of the FDRP,
             already converted into the custom representation

site   position of the CpG site of interest as an integer number

Details

This function is called by calculate.fdrps.chromosome each CpG present on the chormosome and
calls compute.discordant for each pair of reads.

Value

FDRP score for the given CpG site

Author(s)

Michael Scherer we actually calculate the FDRP as frac#discordant read pairs#all read pairs

---

calculate.mhl   *calculate.mhl*

Description

This function computed the Methylation Haplotype Load (MHL) for all sites in a given annotation
by employing the scripts downloaded from the paper's website

Usage

```
calculate.mhl(
  roi,
  bam.file,
  out.folder = getwd(),
  out.name = "mhl.txt",
  bam.type = "bismark"
)
```

Arguments

`roi`	Region of interest for which haplotype information should be computed
`bam.file`	path to bam file containing the reads
`out.folder`	folder to write the output
`out.name`	name of the output file
`bam.type`	alignment tools used to create the bam file

Value

a data frame containing the position in the reference genome and the MHL scores

---

`calculate.pdr`	*calculate.pdr*

Description

This function calculates the PDRs for all analyzed CpG sites in the bam file of the corresponding sample

Usage

```
calculate.pdr(bam.file, anno, log.path = getwd(), cores = 1)
```

Arguments

`bam.file`	bath to the bam file to be analyzed already aligned to a reference genome
`anno`	annotation as a GRanges object with the CpG sites to be analyzed
`log.path`	location of the log file
`cores`	number of cores available for the analysis

Value

PDR scores for the given annotation.

Author(s)

Michael Scherer

---

calculate.pdr.by.chromosome
                              calculate.pdr.by.chromosome

Description

This function calculates the PDR scores for the reads given in the bam files for the CpGs present in the annotation.

Usage

```
calculate.pdr.by.chromosome(bam, anno)
```

Arguments

bam   bam file with the reads from a bisulfite sequencing technique already aligned to
               a reference genome

anno   annotation as a GRanges object with the CpG sites to be analyzed

Value

vector of pdr scores for the given CpG sites in anno

Author(s)

Michael Scherer

---

calculate.pdr.site   *calculate.pdr.site This function computes the PDR score for a given*
                            cpg site.

Description

calculate.pdr.site This function computes the PDR score for a given cpg site.

Usage

```
calculate.pdr.site(cpg, reads)
```

Arguments

cpg   index of the cpg site

reads   GRanges object containing the reads needed for the calculation of the PDR,
               already converted into the custom representation

Value

PDR score for the given CpG site

Author(s)

Michael Scherer

---

calculate.qfdrp   *calculate.qfdrp*

Description

This function calculates the qFDRP scores for the reads given in the bam files for the CpGs present in the annotation.

Usage

```
calculate.qfdrp(
  bam.file,
  anno,
  log.path = getwd(),
  cores = 1,
  window.size = get.option("WINDOW.SIZE")
)
```

Arguments

bam.file	bath to the bam file to be analyzed already aligned to a reference genome
anno	annotation as a GRanges object with the CpG sites to be analyzed
log.path	location of the log file
cores	number of cores available for the analysis
window.size	window size used to restrict the concordance/discordance classification of each read pair DEFAULT: 50 as the maximum distance

Value

qFDRP scores for the given annotation.

Author(s)

Michael Scherer

---

compute.score *compute.score*

Description

Generic function to call, passes its aruguments either to `compute.score.rnb` or `compute.score.GRanges`.

Usage

```
compute.score(bam.file, ..., score = "qfdrp")
```

Arguments

`bam.file`	path to bam file containing the reads
`...`	additional arugment. Either RnBSet, GRanges or empty (only for Epipolymorphism and Entropy)
`score`	The WSH score which should be computed, needs to be one of `fdrp`,`qfdrp`,`pdr`,`epipolymorphism`, `entropy` or `mhl`

Value

data frame containing the annotation and the computed WSH scores

Author(s)

Michael Scherer

Examples

```
load(system.file(file.path("extData","example_GRanges.RData"),package="WSH"))
example.rnb.set <- system.file(file.path("extData","small_rnbSet.zip"),package="WSH")
example.bam <- system.file(file.path("extData","small_example.bam"),package="WSH")
fdrp <- compute.score(bam.file=example.bam,example.GRanges,score="fdrp")
qfdrp <- compute.score(bam.file=example.bam,example.rnb.set,score="qfdrp")
```

`compute.score.GRanges` *compute.score.GRanges*

Description

Main function to compute WSH scores based on a input GRanges object and a bam file containing the reads.

Usage

```
compute.score.GRanges(bam.file, range, score = "qfdrp")
```

Arguments

`bam.file`	path to bam file containing the reads
`range`	GRanges object containing the annotation
`score`	The WSH score which should be computed, needs to be one of `fdrp,qfdrp,pdr,epipolymorphism,` `entropy` or `mhl`

Value

data frame containing the annotation and the computed WSH scores

Author(s)

Michael Scherer

Examples

```
load(system.file(file.path("extData","example_GRanges.RData"),package="WSH"))
example.bam <- system.file(file.path("extData","small_example.bam"),package="WSH")
fdrp <- compute.score.GRanges(bam.file=example.bam,range=example.GRanges)
```

`compute.score.rnb` *compute.score.rnb*

Description

Main function to compute WSH scores based on a input RnBSet and a bam file containing the reads.

Usage

```
compute.score.rnb(bam.file, rnb.set, score = "qfdrp")
```

Arguments

`bam.file`	path to bam file containing the reads
`rnb.set`	path to RnBSet contaning methylation, coverage and sample meta information
`score`	The WSH score which should be computed, needs to be one of `fdrp,qfdrp,pdr,epipolymorphism,` `entropy` or `mhl`

Value

data frame containing the annotation and the computed WSH scores

Author(s)

Michael Scherer

Examples

```
example.rnb.set <- system.file(file.path("extData","small_rnbSet.zip"),package="WSH")
example.bam <- system.file(file.path("extData","small_example.bam"),package="WSH")
fdrp <- compute.score.rnb(bam.file=example.bam,rnb.set=example.rnb.set)
```

---

create.genomebrowser.track

create.genomebrowser.track

Description

Converts the output of compute.score und compute.score.GRanges into genome browser tracks, that can be directly imported into e.g. UCSC Genome Browser.

Usage

```
create.genomebrowser.track(
  score.output,
  sample.name = "Sample",
  store.path = getwd(),
  bin.width = 5000
)
```

Arguments

`score.output`	The output as generated through compute.score or compute.score.GRanges: a data.frame with chromosome positions and the score values for those positions.
`sample.name`	The name of the sample to be added to the file.
`store.path`	The path to which the file is to be stored.
`bin.width`	Width of bins in bp to aggregate the scores. If NULL, the scores are stored as single CpGs.

Author(s)

Michael Scherer

---

get.option   *get.option Print the value of the global option*

Description

get.option Print the value of the global option

Usage

```
get.option(names)
```

Arguments

names   string or character vector containing the names of the options to be printed

Value

the option for the specified option

Author(s)

Michael Scherer

Examples

```
get.option("coverage.threshold")
set.option(coverage.threshold=42)
get.option("coverage.threshold")
```

---

rnb.calculate.fdrp   *rnb.calculate.fdrp*

Description

This function calculates the FDRP scores for a given RnBSet and a bam file containing the reads.

Usage

```
rnb.calculate.fdrp(rnb.set, bam.path, log.path = getwd(), cores = 1)
```

Arguments

`rnb.set`	RnBSet object containing the CpG sites for which calculation should be conducted
`bam.path`	path to the bam file containing the reads
`log.path`	path to the log directory, if not existing it is created
`cores`	cores available for the analysis

Value

FDRP scores for the CpG sites in the RnBSet with a higher coverage than COVERAGE.THRESHOLD

Author(s)

Michael Scherer

Examples

```
example.rnb.set <- system.file(file.path("extData","small_rnbSet.zip"),package="WSH")
example.bam <- system.file(file.path("extData","small_example.bam"),package="WSH")
fdrp <- rnb.calculate.fdrp(rnb.set=example.rnb.set,bam.path=example.bam)
```

---

rnb.calculate.mhl   *rnb.calculate.mhl*

Description

This function calculates the MHL scores for a given RnBSet and a bam file containing the reads.

Usage

```
rnb.calculate.mhl(rnb.set, bam.path, roi.path = getwd())
```

Arguments

`rnb.set`	RnBSet object containing the CpG sites for which calculation should be conducted
`bam.path`	path to the bam file containing the reads
`roi.path`	path to a directory to which output can be added

Value

MHL scores for the CpG sites in the RnBSet with a higher coverage than COVERAGE.THRESHOLD

---

rnb.calculate.pdr   *rnb.calculate.pdr*

Description

This function calculates the PDR scores for a given RnBSet and a bam file containing the reads.

Usage

```
rnb.calculate.pdr(rnb.set, bam.path, log.path = getwd(), cores = 1)
```

Arguments

rnb.set   RnBSet object containing the CpG sites for which calculation should be con-
                 ducted

bam.path   path to the bam file containing the reads

log.path   path to the log directory, if not existing it is created

cores   cores available for the analysis

Value

PDR scores for the CpG sites in the RnBSet with a higher coverage than COVERAGE.THRESHOLD

Author(s)

Michael Scherer

Examples

```
example.rnb.set <- system.file(file.path("extData","small_rnbSet.zip"),package="WSH")
example.bam <- system.file(file.path("extData","small_example.bam"),package="WSH")
fdrp <- rnb.calculate.pdr(rnb.set=example.rnb.set,bam.path=example.bam)
```

---

rnb.calculate.qfdrp   *rnb.calculate.qfdrp*

Description

This function calculates the qFDRP scores for a given RnBSet and a bam file containing the reads.

Usage

```
rnb.calculate.qfdrp(rnb.set, bam.path, log.path = getwd(), cores = 1)
```

Arguments

rnb.set	RnBSet object containing the CpG sites for which calculation should be conducted
bam.path	path to the bam file containing the reads
log.path	path to the log directory, if not existing it is created
cores	cores available for the analysis

Value

qFDRP scores for the CpG sites in the RnBSet with a higher coverage than COVERAGE.THRESHOLD

Author(s)

Michael Scherer

Examples

```
example.rnb.set <- system.file(file.path("extData","small_rnbSet.zip"),package="WSH")
example.bam <- system.file(file.path("extData","small_example.bam"),package="WSH")
fdrp <- rnb.calculate.qfdrp(rnb.set=example.rnb.set,bam.path=example.bam)
```

---

run.haplotype.calculation

run.haplotype.calculation

Description

Runs the haplotype calculation associated with MHL's publication

Usage

```
run.haplotype.calculation(
  roi,
  bam.file,
  out.folder = getwd(),
  out.name = "hapinfo.txt",
  bam.type = "bismark"
)
```

Arguments

roi	Region of interest for which haplotype information should be computed
bam.file	path to bam file containing the reads
out.folder	folder to write the output
out.name	name of the output file
bam.type	alignment tools used to create the bam file

Author(s)

Michael Scherer

---

run.methclone *run.methclone*

Description

Runs the methclone software

Usage

```
run.methclone(bam.file, out.folder = getwd(), out.name = "methclone")
```

Arguments

bam.file	path to bam file containing the reads for defining epialleles
out.folder	path to folder where methclones output should be stored
out.name	name of the output file

Author(s)

Michael Scherer

---

run.mhl.calculation *run.mhl.calculation*

Description

Runs the MHL calculation given the haplotype information file associated with MHL's publication

Usage

```
run.mhl.calculation(hapinfo.file, out.folder = getwd(), out.name = "mhl.txt")
```

Arguments

hapinfo.file	path to haplotype information file as produced by run.haplotype.calculation
out.folder	folder to write the output
out.name	name of the output file

Author(s)

Michael Scherer

---

set.option   *set.option*

Description

Change global options for WSH score calculation

Usage

```
set.option(
  window.size = 50,
  mapq.filter = 35,
  max.reads = 40,
  min.overlap = 35,
  fdrp.type = "FDRP",
  coverage.threshold = 10,
  methclone.methylation.diff = 0,
  perl.path = "/usr/bin/perl",
  samtools.path = "/usr/bin"
)
```

Arguments

`window.size`	Window around the CpG site of interest to consider in FDRP and qFDRP calculation, the higher the value, the more likely it is to find heterogeneity
`mapq.filter`	mapq filter used to filter out low quality reads
`max.reads`	Maximum number of reads to be considered in FDRP and qFDRP calculation. The scores compute all pairs, therefore this is a crucial parameter for the running time of the calculation.
`min.overlap`	Miniumum overlap between two reads to consider it as a read pair in FDRP/qFDRP calculation in bp.
`fdrp.type`	FDRP type to be used: either FDRP or qFDRP
`coverage.threshold`	
	Coverage Threshold emloyed to select the sites in a RnBSet annotation that fullfill havin a coverage higher than this threshold
`methclone.methylation.diff`	
	Methylation difference parameter employed by the methclone software. Only sites are considered that have a methylation difference higher than this value in the methclone package.
`perl.path`	Path to an installation of perl on the machine
`samtools.path`	path to the directory where samtools is located in your machine

Author(s)

Michael Scherer

Examples

```
get.option("coverage.threshold")
set.option(coverage.threshold=42)
get.option("coverage.threshold")
```

VALID.SCORES *main.R*

Description

This scripts contains main functions to calculate the scores.

Usage

```
VALID.SCORES
```

Format

An object of class character of length 6.

---

wsh.run.example *wsh.run.example*

Description

Returns the scores for a small example data set of size 50kb

Usage

```
wsh.run.example(score = "qfdrp")
```

Arguments

score   The WSH score which should be computed, needs to be one of fdrp,qfdrp,pdr,epipolymorphism,
             entropy or mhl

Value

A data frame containing the annotation and the corresponding score

Author(s)

Michael Scherer

Examples

```
wsh.run.example("pdr")
```

Index